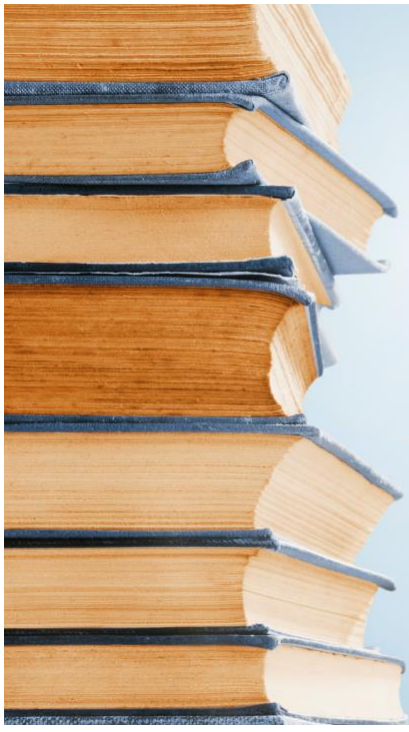


Gen AI / LLM Training

Rules for Today

- Participate. Get engaged.
- Make mistakes.
- Learn from mistakes. Learn from each other.
- Raise your hands.
- Ask questions. There are no dumb questions.
- Bring your own questions.
- If you need a break, please take a break.
- Bring notepad so you can do exercise.
- Discuss.
- Laugh. Have a fun.

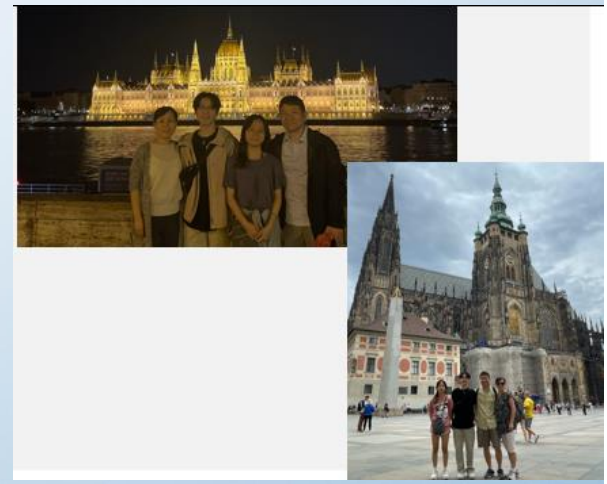
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Introduction

Name:
Kevin Lee

What you do:
Managing member and CTO at ClinVia

Something special about you:
Love to travel with the family



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Agenda

- Introduction of Large Language Model (LLM), Generative AI (Gen AI)
- ChatGPT Introduction and Application
- ChatGPT Use Cases
- Prompts, not Search
- Prompt Engineering in Gen AI (ChatGPT)
- Gen AI functionality – ChatGPT, Copilot, Gemini, Sonnet, DeepSeek
- Application development using Gen AI API (SAS & Python)
- LangChain & RAG LLM application development
- AI Agent and MCP
- Enterprise level Gen AI (ChatGPT) Implementation
- Future of Gen AI (ChatGPT)
- Discussion

Confidential – Do not duplicate or distri[illegible] of Kevin Lee

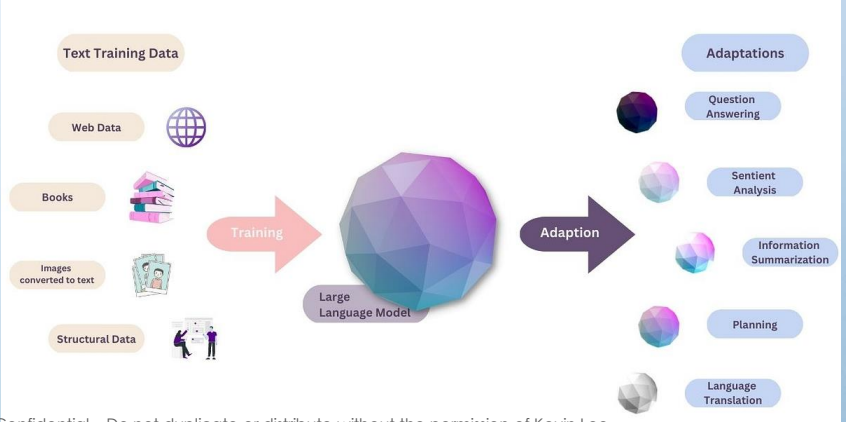
Large Language Model

- Introduction
- LLM in AI Landscape
- LLM Architecture
- History of LLM Development

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

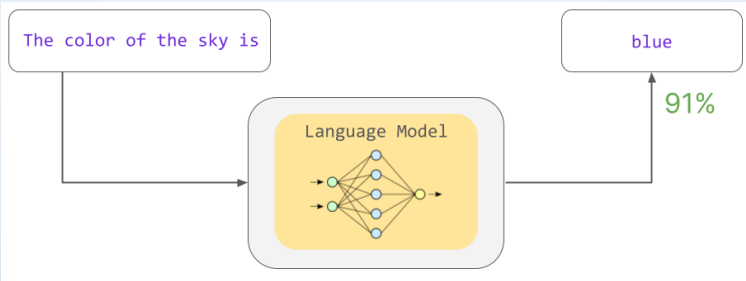
What is LLM?

- Large Language Model (LLM) – a trained Machine Learning model that understands and generates text in a human-like fashion.
- LLM has been trained on a huge corpus of data (a large numbers of texts)
- Highly efficient in generating language-based outputs.
- Popular LLM
  - BERT
  - ChatGPT3.5
  - ChatGPT4
  - GPTs
  - Llama
  - Copilot
  - Bard / Gemini

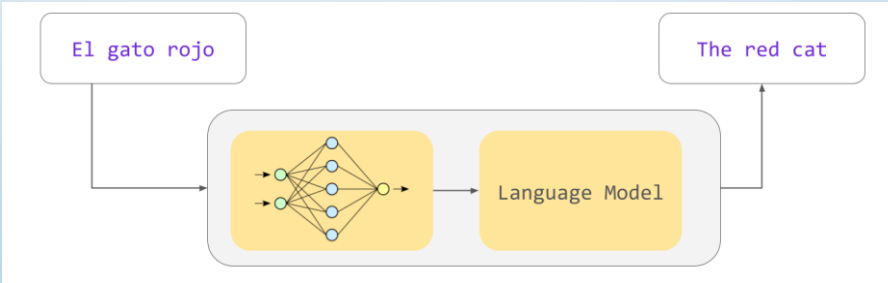


Confidential – Do not duplicate or distribute without the permission of Kevin Lee

How LLM Works



- Large Languages are used to train LLM
- So, when 'input texts' are prompted into LLM, LLM provide 'response', which show a highest probability.



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

History of LLM Model Development

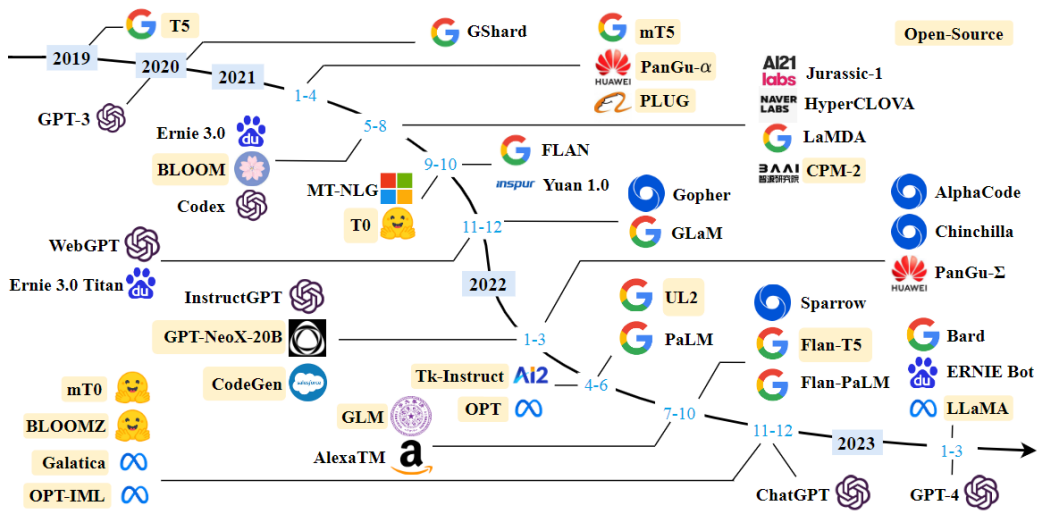


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Gen AI (Generative AI)

- Introduction
- Gen AI Market Trend

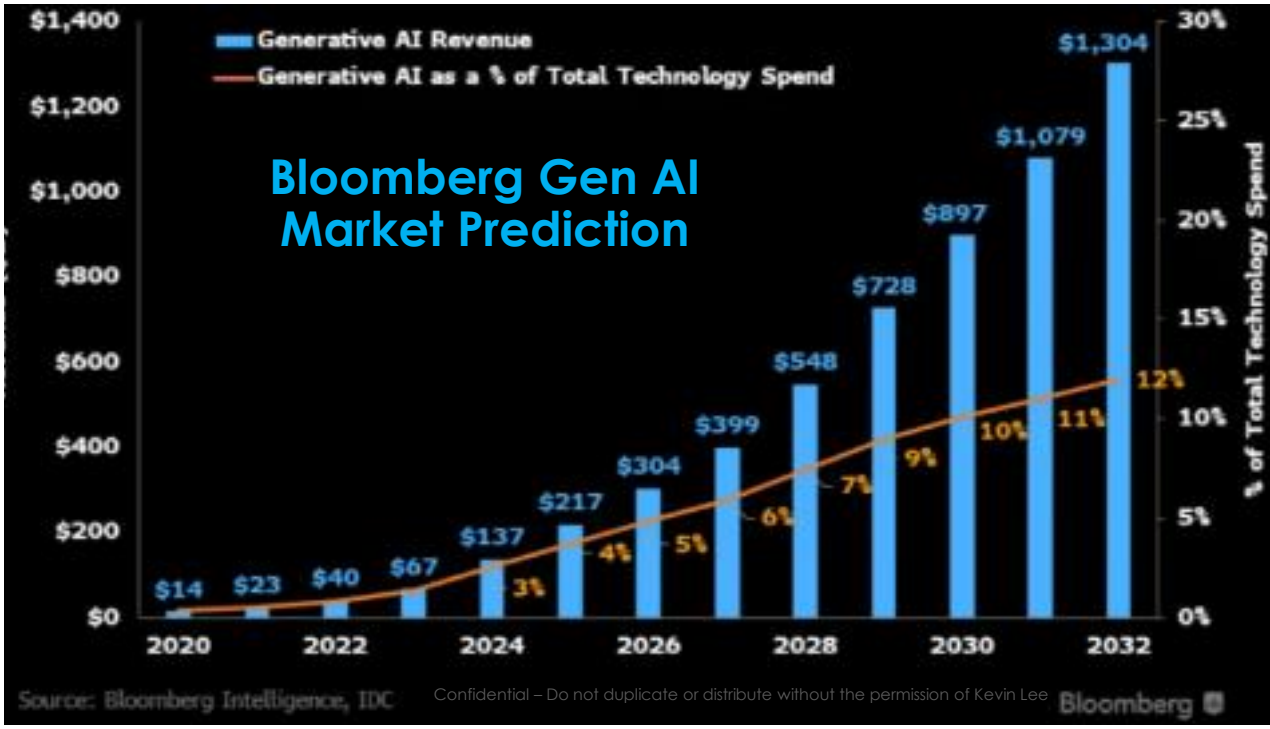
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

What is Gen AI?

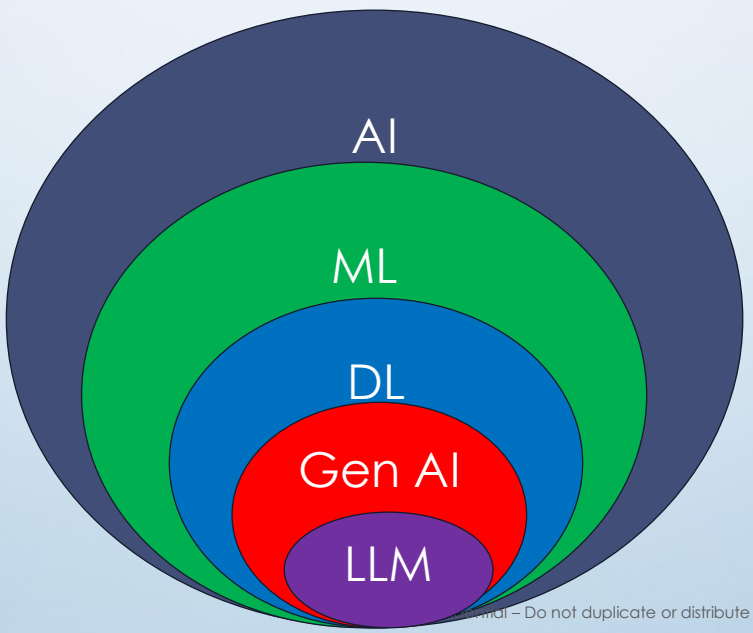
Gen AI – a trained Machine Learning model that generate new contents with a simple prompt.

- Text : Content Writing, Chatbots, Assistants, Search
- Code : Code Generation, Data Set Generation
- Image : Image Generation, Image Edit
- Audio : Voice Generation/Edit, Sound creation, Audio Translation
- Video : Video Creation/Edit, Voice Translation, Deepfake

Confidential – Do not duplicate or distribute without the permission of Kevin Lee



Gen AI & LLM in AI Landscape



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

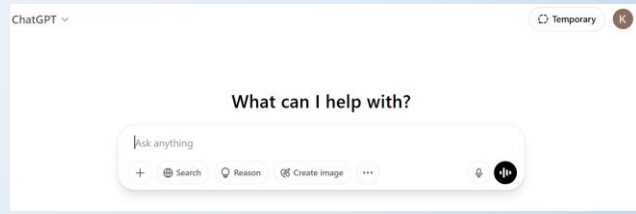
ChatGPT

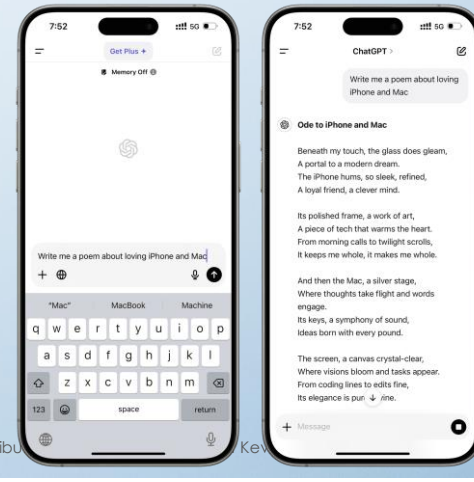
- Introduction
- Development
- Plan
- Popularity
- Why Should I use it & Why not

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

What is ChatGPT?

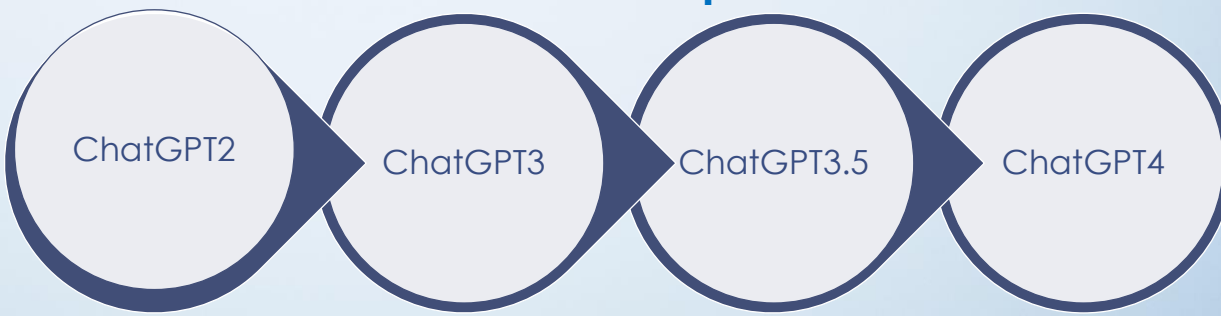
- ChatGPT is an advanced LLM developed by OpenAI.
- ChatGPT is trained on large corpus of text (~300B tokens).
- Its main strength
  - the ability to generate human-like responses in various contexts.
  - The ability to understand and generate text in a wide range of domains.
- Its application
  - Inquiry/prompt (rather than search)
  - Content Generation
  - Customer support
  - Interactive Storytelling
  - Coding
  - Image Analysis
  - Art Development





Confidential – Do not duplicate or distribu

ChatGPT Development



- Feb'2019
- 768 word
- 1.5B parameters

- June'2020
- 1,536 words
- 175B parameters

- Sep'2021
- 4K tokens (3,072 words) – Both input & output
- 16K tokens (ChatGPT3.5-turbo-16k)
- 175B parameters

- April'2023
- 8K tokens
- 32k tokens (ChatGPT4-32k)
- 128k tokens (GPT-4 Turbo)
- ~1,760B parameters

Reference Points
- Tweet : 55 words
- NY Times Article : 622 words
- Standard Novel : 90K words
- Kings James Bible : 783,137 words

Each prompt include the current tokens and previous tokens.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

ChatGPT model

- GPT-4.5: This model is currently a research preview that is OpenAI's largest model to date. The organization claims it excels at tasks involving creativity and open-ended thinking.
- GPT-4o: OpenAI's flagship model, capable of performing complex tasks requiring multiple steps. omni model with text + image + video capabilities + top reasoning
- GPT-4o mini: A smaller and more affordable version of GPT-4o that's meant for speedy, low-lift tasks.
- o1-preview and o1-mini: Models specially trained with reinforcement learning that can perform more complex reasoning tasks.
- GPT-4 Turbo and GPT-4: OpenAI's former flagship models equipped with more general knowledge and reasoning abilities than previous models.
- GPT-3.5 Turbo: A relatively inexpensive model used for straightforward, simple tasks.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Bar Exam Score Performance :

ChatGPT3.5 vs ChatGPT4

10% vs 90%

Confidential – Do not duplicate or distribute without the permission of Kevin Lee 17

CPA Exam Performance:

ChatGPT3.5 vs ChatGPT4

35.1% vs 85.1%

Confidential – Do not duplicate or distribute without the permission of Kevin Lee 18

ChatGPT Plan

- ChatGPT
  - Free Plan
  - Access to GPT-4o mini model
  - Use custom GPTs
  - Standard voice mode
- ChatGPT Plus
  - Paid service : $20/month
  - Access to o3-mini, o1, advanced voice mode
  - Faster Response Speed
  - Exclusive access to beta features like GPT4.5
- ChatGPT Pro
  - Paid service : $200/month
  - Unlimited Access to all reasoning models and GTP 4o
  - Extended access to Sora Video

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

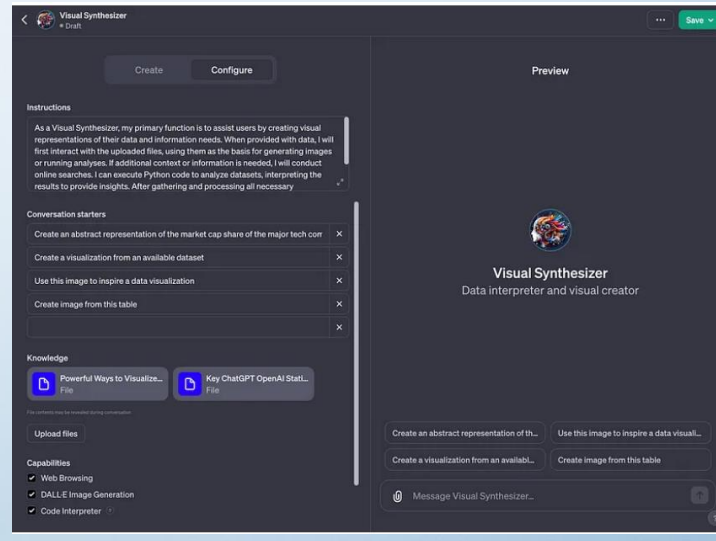
ChatGPT Plan for Enterprise

- ChatGPT API
  - Pay as you go :
    - GPT3.5-Turbo : $0.5/ 1m tokens (input) and $1.5 / 1m tokens (output)
    - GPT-4o : $2.5/ 1m tokens (input) and $10 / 1m tokens (output)
  - An interface that allows developers to integrate the ChatGPT model into their own applications, software, or platforms.
- ChatGPT Team
  - Paid service : $25/month
  - Access to GPT-4 with 32k context window, DALL-E, GPT-4 with Vision
  - No training your business data
  - Secure workspace for your team
  - Create and share custom GPTS with your workspace
- ChatGPT Enterprise
  - Paid service :
  - The most Enterprise-grade Security & Privacy
  - Unlimited access to GPT-4, Code Interpreter
  - Free credits to use GPT API

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

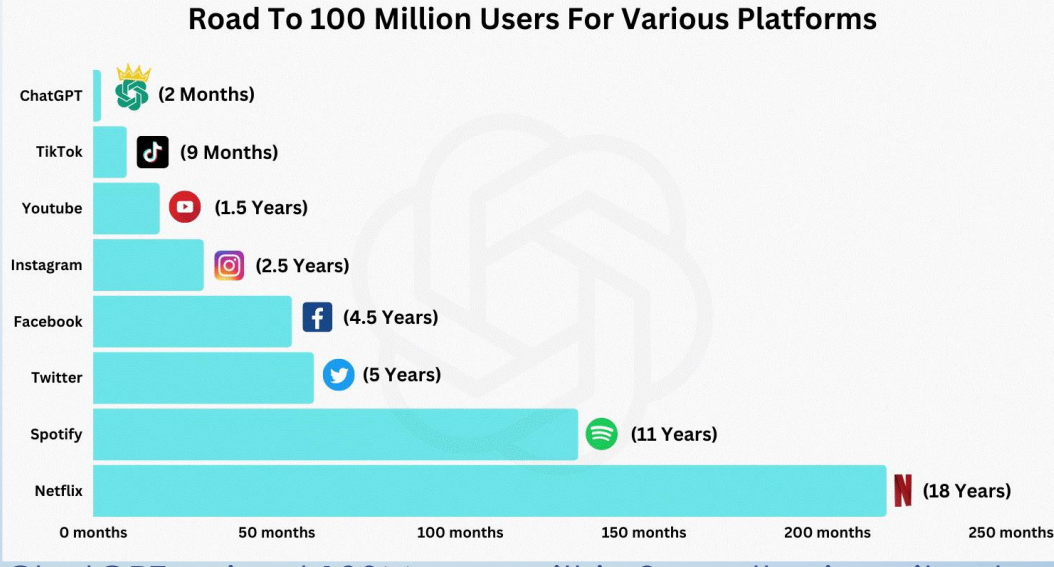
GPTs (GPT store)

- ChatGPT user platform
- Build your own ChatGPT without coding
- GPT Store / GPT Apps
- ChatGPT Plus Subscription
- Configuration
  - Instructions : Prompt
  - Knowledge : contents
- Use Case
  - Regulatory Response
  - Drug Compound Information



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

ChatGPT Popularity



ChatGPT gained 100M users within 2 months since its release.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Reasons that ChatGPT is not being used

- Concerns/Risks using ChatGPT
- A lack of understanding of ChatGPT Use Cases
- A lack of imagination of impact
- A lack of understanding of Prompt

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

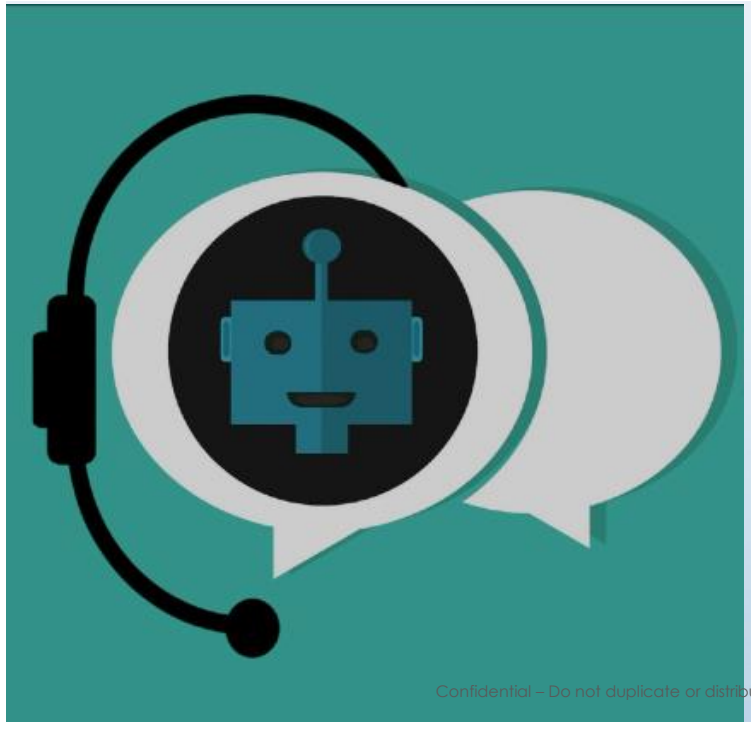
Potential Benefits using ChatGPT

- Higher Productivity / Efficiency
- Better Patient Engagement
- Patient's Recruitment
- Enhancement in Image
- Market Research & Insights

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

ChatGPT Users vs Non-Users Comparison in Boston Consulting

- Results of ChatGPT Users
  - finished 12.2% more tasks on average
  - completed tasks 25.1% more quickly
  - produced 40% higher quality results



Patients & HCP Support and Engagement

- Better Support to Patients & HCP : Instant Responses to frequently asked questions
  - New Medicine
  - Dosage
  - Side effects
  - Usage instructions
  - Mechanism
  - Ongoing Research / Clinical Trials
- Better Patient's engagement thru digital platforms (Chat Bot)

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Better Patient Recruitments

- Better Patients Recruitment finding through LLM(ChatGPT) from Large EHR using Inclusion/Exclusion criteria
- User-friendly AI interface
  - More efficient and cost-effective recruitment process using ChatGPT
  - Higher awareness and conception about ongoing and future clinical trials
  - Better education
  - Higher participation

Enhancement in Brand Image

- Implementing cutting-edge technology like ChatGPT
- Launching new medicine along with new services and technology like ChatGPT
- Commitment to innovation and patient-centric services using ChatGPT

Market Research and Insights

- Customer Feedback Collection and Sentiment Analysis
- Analysis in Patients Interaction and Preferences
- Valuable Insights into Market Trends and Optimizing Marketing Strategies
- Competitor Analysis
- Forecasting Demand

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Drug Discovery

- Quicker Scientific Literature Review and Information Retrieval
- New ideas for Drug Targets, Compounds and Experimental Approaches
- Identification of the binding affinity of compounds to targets
- Drug-Target Pairing Predictions
- Personalized medicine through biomarkers and genetic factors

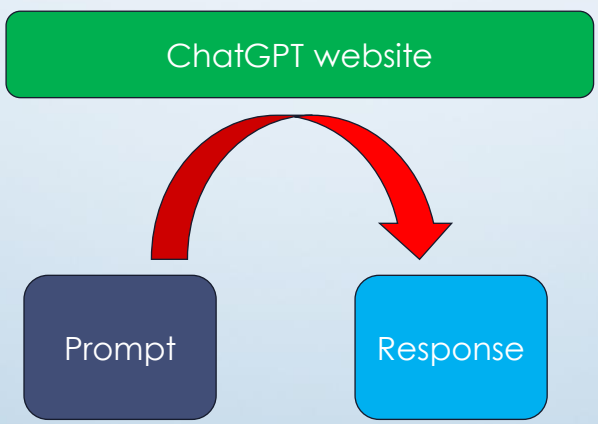
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

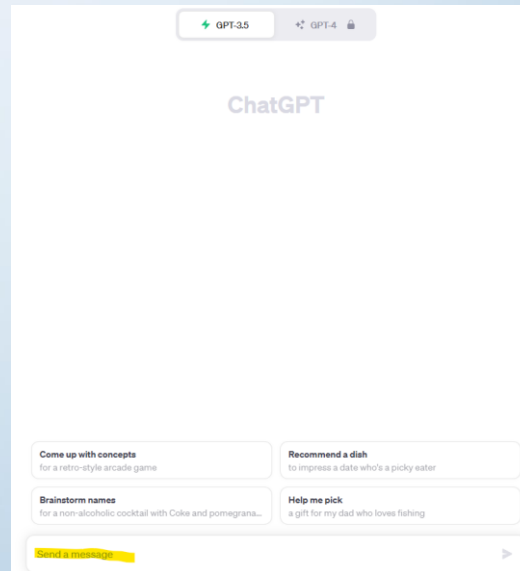
ChatGPT Application Guideline

- Simple Prompt
- Application Development using ChatGPT API
- Application Development using fine-tuned ChatGPT API

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

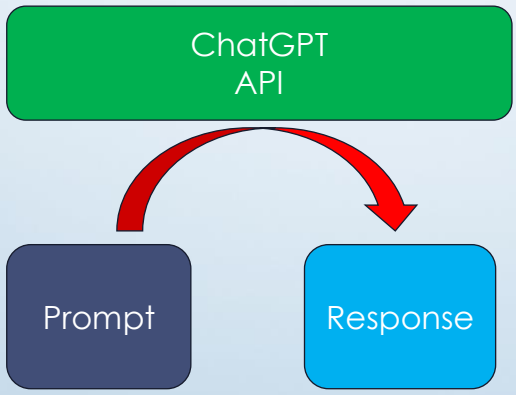
Simple Prompt in OpenAI ChatGPT website





Confidential – Do not duplicate or distribute without the permission of Kevin Lee

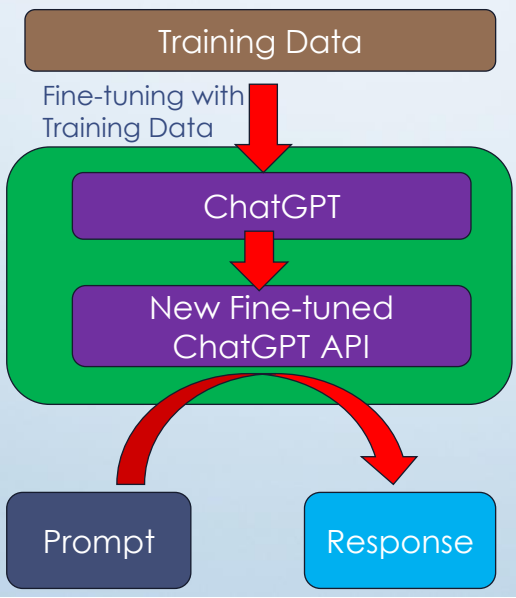
Application using ChatGPT API




```
def get_completion(prompt, model="gpt-3.5-turbo"):
    messages = [{"role": "user", "content": prompt}]
    response = openai.ChatCompletion.create(
        model=model,
        messages=messages,
        temperature=0
    )
    return response.choices[0].message["content"]

text = f"""You should express what you want a model to do by
providing instructions that are as clear and specific as you
can possibly make them.
This will guide the model towards the desired output,
and reduce the chances of receiving irrelevant or incorrect
responses. Don't confuse writing a clear prompt with writing a
short prompt. """

prompt = f"""Summarize the text delimited by triple backticks
into a single sentence.
```{text}```
"""
response = get_completion(prompt)
print(response)
```

33

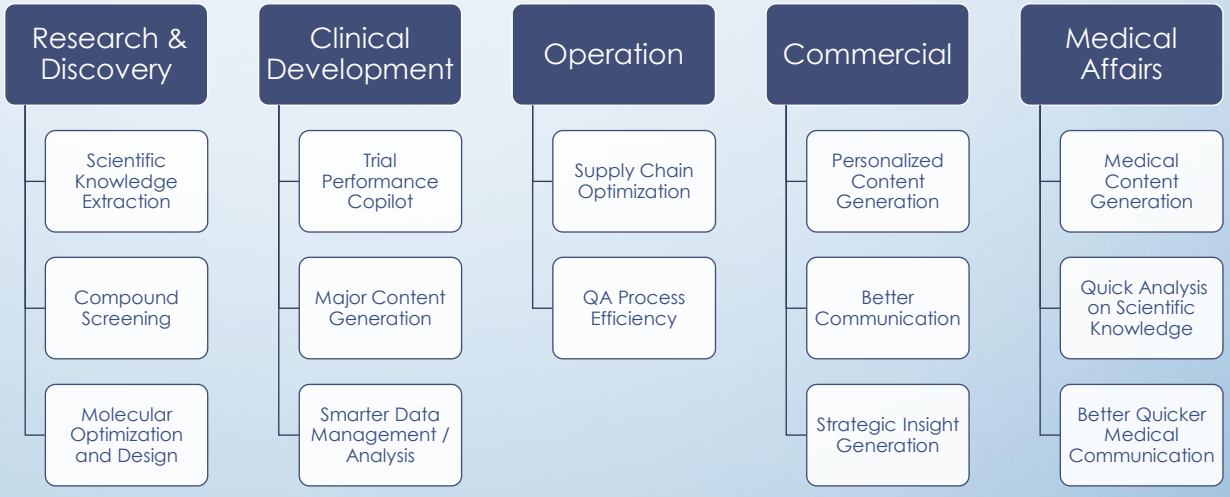
Application using fine-tuned ChatGPT API




```
def get_completion(prompt,
                   model="gpt-3.5-turbo:personal::xxxxxx"):
    messages = [{"role": "user", "content": prompt}]
    response = openai.ChatCompletion.create(
        model=model,
        messages=messages,
        temperature=0
    )
    return response.choices[0].message["content"]

text = f"""You should express what you want a model to do by
providing instructions that are as clear and specific as you
can possibly make them.
This will guide the model towards the desired output,
and reduce the chances of receiving irrelevant or incorrect
responses. Don't confuse writing a clear prompt with writing a
short prompt. """

prompt = f"""Summarize the text delimited by triple backticks
into a single sentence.
```{text}```
"""
response = get_completion(prompt)
print(response)
```

34

Gen AI Impact in Each Function



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

How Gen AI can help Pharma Function

Research & Discovery
- New Treatment Generation

Clinical Development
- Smarter Trials, Better Data, Quicker Results

Operation
- Better Operation Performance

Commercial
- Better Relationship with Partners, Providers and Patients

Medical Affairs
- Smarter Providers, Healthier Patients

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

ChatGPT Use Cases

- Simple Inquiry
- Content Development
- Coding
- Comparison
- Patient Profiles
- Safety Cases
- Image Analysis
- Art Generation

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Simple Inquiry 1

Prompt :
What is schizophrenia? Tell me in 200 words.

Output :
Schizophrenia is a complex and chronic mental disorder that profoundly affects a person's thinking, emotions, and behavior. It typically emerges in late adolescence or early adulthood, and its exact cause remains unclear, though it is thought to involve a combination of genetic, environmental, and neurobiological factors.

Individuals with schizophrenia often experience a range of symptoms, including hallucinations (false sensory perceptions), delusions (false beliefs), disorganized thinking, and reduced emotional expression. They may also struggle with cognitive impairments, such as difficulty with memory, attention, and decision-making.

Schizophrenia can have a profound impact on a person's daily life and functioning, making it challenging to maintain relationships, hold a job, or pursue personal goals. It can lead to significant distress and impairment in social, occupational, and educational domains.

Treatment for schizophrenia typically involves a combination of antipsychotic medications, psychotherapy, and psychosocial support. While these interventions can help manage symptoms and improve the quality of life for individuals with schizophrenia, the condition is often chronic and requires ongoing care and support.

Early intervention and a comprehensive, individualized treatment plan are crucial for managing schizophrenia and helping affected individuals lead fulfilling lives while minimizing the impact of their symptoms.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Simple Inquiry 2

Prompt :
Provide the information about CDISC. Tell me in 200 words.

Output :
CDISC, or Clinical Data Interchange Standards Consortium, is a global non-profit organization focused on developing and advancing data standards for the clinical research and healthcare industries. Founded in 1997, CDISC aims to facilitate the efficient exchange and integration of clinical research data to enhance collaboration, streamline processes, and ultimately improve patient outcomes.

CDISC standards provide a common framework for organizing and formatting data, ensuring consistency and interoperability across diverse research studies and healthcare settings. The organization's flagship standards include the Study Data Tabulation Model (SDTM), which standardizes the structure and content of clinical trial data, and the Analysis Data Model (ADaM), which defines dataset and metadata standards for analysis.

By adopting CDISC standards, researchers, pharmaceutical companies, and regulatory agencies can expedite the regulatory submission process, enhance data quality, and foster data sharing and reuse. CDISC collaborates with stakeholders globally, including academia, industry, and regulatory bodies, to develop and maintain these standards, promoting transparency and efficiency in the generation and analysis of clinical data. The widespread adoption of CDISC standards contributes to advancing clinical research practices, enabling more robust and interoperable data exchange in the pursuit of improved healthcare outcomes.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Personal Question

Prompt :

Based on the interaction with ChatGPT, what is my personality type?

Use MBTI, and provide the reasons

Output :
Personality Type: INTJ (The Architect)

Introversion (I): vs Extraversion(E)
You prefer working independently and are energized by deep, focused tasks. Your conversations show a strong inclination toward thoughtful reflection rather than social interaction.

Intuition (N): vs Sensing (S)
You focus on long-term goals, systems, and innovative ideas. You consistently explore how to improve processes using advanced tools like AI and automation, indicating a forward-thinking mindset.

Thinking (T): vs Feeling (F)
You rely on logic and structure in your decision-making. Your work in programming, system validation, and analytics demonstrates a preference for objective analysis over personal values.

Judging (J): vs Perceiving (P)
You show a high level of organization and planning. Whether you're creating templates, managing clinical trial workflows, or mentoring others, your approach is methodical and goal-driven.

Summary:
You are likely an INTJ—strategic, independent, and analytical. You enjoy solving complex problems, optimizing systems, and leading through expertise rather than authority.

Brain Storming

Prompt :

I want to generate creative ideas for promoting a new health supplement for weight loss.

Please suggest 10 different marketing ideas that are low-budget, eye-catching, and easy to test online.

Consider the audience is mostly women aged 30-50 who care about fitness and natural remedies.

Think outside the box and include social media, influencers, and offline ideas too

Output :

- "Glow-Up Challenge" on Social Media
  Create a 21-day transformation challenge where participants document their journey using the supplement, combined with simple fitness or wellness routines.
  Hashtag idea: #NaturalGlowUp
  Incentive: Free products for best transformation or progress.
- "Fridge Confessionals" Mini-Series
  Invite real users or micro-influencers to share raw, short videos from their kitchen talking about their weight loss struggles and how the supplement fits into their daily routine.
  Authenticity sells in this demographic.
  Easy to produce with a phone camera.
- Partner with Yoga or Barre Studios
  Offer free sample packs or mini pop-up tastings at local fitness studios frequented by your target audience.
  Leave behind QR codes linking to your website and testimonials.

.....

Content Development - Email

Prompt :
I am salesperson.

I need to contact the doctors about new medicine, which is 3 times more effective to lower blood pressure.

write marketing emails within 100 words.

Output :
Subject: Revolutionize Blood Pressure Management with Our New Medicine!
Dear Dr. [Doctor's Last Name],

I hope this message finds you well. I am excited to introduce our breakthrough medication, [Medicine Name], designed to transform blood pressure management. With its cutting-edge formula, it is proven to be three times more effective in lowering blood pressure compared to existing solutions.

[Medicine Name] offers a unique opportunity for your patients to achieve better health outcomes. Its advanced technology ensures quicker results and fewer side effects, leading to enhanced patient compliance.

We would be honored to provide you with more information and samples for your evaluation. Together, we can make a significant impact on improving patients' lives.

Let's revolutionize hypertension treatment together. Please reply to this email, and I'll be happy to arrange a meeting at your convenience.

Best Regards,

[Your Name]
[Your Title]
[Your Contact Information]

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Content Development - Email

Prompt :

I am stat programmer.

I am waiting for the final raw data from data managers. I have communicated with them couple times already, but I have not received the response yet.

Could you write the email for me? Write it in politely and professionally.

Output :

Dear [Data Manager's Name],

I hope this email finds you well. I wanted to follow up on the status of the final raw data for [Study Name/Project]. As we are preparing for the next steps in our analysis, I would appreciate any updates on the expected timeline for data availability.

Please let me know if there are any issues or if further clarification is needed from my end. I appreciate your support and look forward to your response.

Best regards,
[Your Name]
[Your Position]
[Your Company]
.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Content Development - Simple protocol synapsis

- Prompt :
  Write clinical trial protocol synapsis based on below "title of study".

  "title of study" : A phase 3, randomized, double-blind, placebo controlled study to evaluate the efficacy of drug A in schizophrenia patients.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

- Output :

Clinical Trial Protocol Synopsis: Phase 3 Study on the Efficacy of Drug A in Schizophrenia Patients

1. Study Title:
A phase 3, randomized, double-blind, placebo-controlled study to evaluate the efficacy of drug A in schizophrenia patients.

2. Objective:
The primary objective of this clinical trial is to assess the efficacy of drug A in treating schizophrenia in comparison to a placebo. Secondary objectives include evaluating the safety, tolerability, and overall clinical benefit of drug A.

3. Study Design:
This study will adopt a randomized, double-blind, placebo-controlled design. Participants will be randomly assigned to either the treatment group receiving drug A or the control group receiving a placebo. The double-blind nature of the study ensures that both investigators and participants are unaware of the assigned treatment.

4. Participants:
The study will enroll adult participants diagnosed with schizophrenia according to standardized diagnostic criteria. Inclusion and exclusion criteria will be defined to ensure the homogeneity of the study population.

5. Interventions:
- Treatment Group: Participants will receive drug A in the prescribed dosage.
- Control Group: Participants will receive a placebo identical in appearance to drug A.

6. Primary Outcome Measure:
The primary efficacy endpoint is the reduction in the severity of schizophrenia symptoms, as measured by a standardized assessment tool, at the end of the treatment period.

7. Secondary Outcome Measures:
Secondary endpoints include:
- Changes in cognitive function.
- Improvement in overall functioning and quality of life.
- Incidence and severity of adverse events.
- Assessment of treatment adherence.

8. Study Duration:
The trial will be conducted over a specified period, with defined time points for participant recruitment, treatment administration, and follow-up assessments. The duration of the study will be determined based on the anticipated rate of symptom improvement and the need for robust data collection.

9. Statistical Analysis:
Statistical methods, including intention-to-treat analysis, will be employed to evaluate the primary and secondary endpoints. Sample size calculations will ensure the study's power to detect meaningful differences between the treatment and control groups.

10. Ethical Considerations:
The study will be conducted in accordance with ethical principles outlined in the Declaration of Helsinki and local regulatory requirements. Informed consent will be obtained from all participants, and measures will be taken to ensure participant confidentiality and safety.

11. Data Monitoring and Safety:
An independent Data Monitoring Committee (DMC) will oversee the trial's progress, ensuring participant safety and data integrity. Adverse events will be monitored and reported according to regulatory guidelines.

12. Conclusion:
This phase 3 clinical trial aims to contribute valuable insights into the efficacy and safety of drug A in the treatment of schizophrenia. The results of this study may have implications for improving the management of schizophrenia and enhancing the overall well-being of affected individuals.

45

Codes conversion

Prompt :

Convert below SAS codes to R codes:

```
data dm;
   merge demo rand;
   by usubjid ;
run;
```

Output :

You can convert the SAS code to R using the `merge` function in R to perform the equivalent operation. Assuming you have two data frames `demo` and `rand` that you want to merge by the `usubjid` variable, here's the equivalent R code:

```
# Assuming you have loaded your data into two data frames: demo and rand
# Merge the data frames by usubjid
dm <- merge(demo, rand, by="usubjid")
```

In this code:

- `demo` and `rand` are your input data frames.
- We use the `merge` function to merge the data frames.
- The `by` parameter specifies the key variable by which to merge the data frames, which in this case is `usubjid`.

After running this code, the merged data frame `dm` will contain the result of the merge operation, similar to what you would achieve with the SAS code you provided.

46

Codes Creation 1

Prompt :
Write a **R script** that groups data by "trta" and "avisitn" columns in "advs.xpt" and calculates the mean of the "aval" column for each group at parmcd ='WEIGHT'

Output :

```R
# Load necessary libraries
library(haven)
library(dplyr)

# Read the XPT file
advs <- read_xpt("path/to/advs.xpt")

# Filter the dataset for parmcd == 'WEIGHT'
advs_weight <- advs %>% filter(parmcd == 'WEIGHT')

# Group by 'trta' and 'avisitn' and calculate the mean of 'aval' for each group
mean_aval <- advs_weight %>%
  group_by(trta, avisitn) %>%
  summarise(mean_aval = mean(aval, na.rm = TRUE))

# Print the result
print(mean_aval)
```

Make sure to replace `"path/to/advs.xpt"` with the actual file path of your `advs.xpt` dataset. This script will output a data frame with the mean `aval` for each combination of `trta` and `avisitn` where `parmcd` is 'WEIGHT'.

47

Codes Creation 2

Prompt :
Write a **SAS script** that groups data by "trta" and "avisitn" columns in "advs.xpt" and calculates the mean of the "aval" column for each group at parmcd ='WEIGHT'

Output :

```sas
/* Step 1: Import the advs.xpt file */
libname mylib xport "path_to_your_advs.xpt";

/* Step 2: Create a SAS dataset from the imported data */
data advs;
   set mylib.advs;
run;

/* Step 3: Filter the data for parmcd = 'WEIGHT' */
data advs_weight;
   set advs;
   where parmcd = 'WEIGHT';
run;

/* Step 4: Group the data by trta and avisitn, and calculate the mean of aval */
proc means data=advs_weight noprint;
   class trta avisitn;
   var aval;
   output out=mean_aval (drop=_TYPE_ _FREQ_) mean=mean_aval;
run;

/* Step 5: Display the result */
proc print data=mean_aval;
run;
```

48

Codes Creation 3

Prompt :

Write a **Python script** that groups data by "trta" and "avisitn" columns in "advs.xpt" and calculates the mean of the "aval" column for each group at parmcd ='WEIGHT'

Output :

```bash
pip install pandas xport
```

```python
import pandas as pd
import xport

# Load the .xpt file
with open("advs.xpt", "rb") as file:
    data = xport.to_dataframe(file)

# Filter the data where parmcd is 'WEIGHT'
weight_data = data[data['parmcd'] == 'WEIGHT']

# Group by 'trta' and 'avisitn' and calculate the mean of 'aval'
grouped_data = weight_data.groupby(['trta', 'avisitn'])['aval'].mean().reset_index()

# Rename the column for clarity
grouped_data = grouped_data.rename(columns={'aval': 'mean_aval'})

# Display the result
print(grouped_data)
```


Metadata driven Codes Development

Prompt :

Write a SAS script that reads input data and generates an output dataset based on the following #metadata: Ensure that the script correctly derives each variable from the input data according to the metadata specifications.

- Set up two libname
  - libname raw " " for input data folder path
  - Libname SDTM " " for output data folder path
- Refer to #reference to create each variables;
- Find metadata of the same input data. For your reference, in demo.study_id, demo is input data and study_id is variable from demo input data, so all the identical input data will be processed in the same data procedure.
- Create the new "Variables_Name" from the same input data.
- Merge all the new variables to create the final output dataset.
- Assign "Variable_Label" to "Variable_Name"
- Keep only "Variable_Name" in #metadata.

Assumption:

- We are using sas datasets as input data, so we will use libname raw and directly read from the folder.
- We will write back to

```
#metadata:
[ {   "Variable_Name":"STUDYID",
      "Variable_Label":"Study Identifier",
      "Type":"Char",
      "Derivation":"demo.study_id",
      "example":101},
  {  },,,,,,
]
```

```
#reference:
Libname raw " ";
Libname sdtm " ";

Data _dm;
Set raw.demo;
STUDYID = study_id;
DOMAIN = "DM"
Run;

Data _rand;
Set raw.rado;
ARMCD = trt_cd;
Run;

Data dm_derived;
Merge _dm (in=a) _rand (in=b);
By subjid;
Run;

Data sdtm.dm;
Set dm_derived;
Keep STUDYID ...;
Label ....;
Run;
```

Metadata driven Codes Development

Output :

```
/* Replace with actual input data folder path */
libname raw "path_to_input_data";
/* Replace with actual output data folder path */
libname sdtm "path_to_output_data";

/* Extract and derive variables from raw.demo */
data _dm;
   set raw.demo;
   STUDYID = study_id;
   DOMAIN = "DM";
   USUBJID = cats(study_id, '-', site_id, '-', subjid);
   SUBJID = subjid;
   SITEID = site_id;
   BRTHDTC = birth_dt;
   SEX = sex;
   RACE = race;
run;

/* Extract and derive variables from raw.expo */
data _expo;
   set raw.expo;
   retain RFSTDTC RFXSTDTC;
   RFSTDTC = exdtc;  /* First exposure date */
   RFXSTDTC = exdtc; /* First exposure date */
   RFENDTC = exdtc;  /* Last exposure date */
   RFXENDTC = exdtc; /* Last exposure date */
run;

/* Extract and derive variables from raw.desp */
data _desp;
   set raw.desp;
   RFICDTC = cons_dt;
   DTHDTC = death_dt;
   if not missing(death_dt) then DTHFL = "Y";
Run;

/* Extract and derive variables from raw.rand */
data _rand;
   set raw.rand;
   ARMCD = trt_cd;
   ARM = trt;
   ACTARMCD = trt_cd;
   ACTARM = trt;
   ARMNRSN = trt_reas;
   ACTARMUD = trt_reas2;
run;

/* Merge all derived datasets */
data dm_merged;
   merge
      _dm (in=a)
      _expo (in=b keep=subjid RFSTDTC RFENDTC
RFXSTDTC RFXENDTC)
      _desp (in=c keep=subjid RFICDTC DTHDTC
DTHFL)
      _rand (in=d keep=subjid ARMCD ARM
ACTARMCD ACTARM ARMNRSN ACTARMUD);
   by subjid;
   if a;
run;

/* Finalize the dataset */
data sdtm.dm;
   set dm_merged;
   COUNTRY = "USA";
   AGE = year(input(RFICDTC, yymmdd10.)) -
year(input(BRTHDTC, yymmdd10.));
   AGEU = "YEARS";
   keep STUDYID DOMAIN USUBJID SUBJID RFSTDTC RFENDTC
RFXSTDTC RFXENDTC RFICDTC REPENDTC DTHDTC DTHFL SITEID
BRTHDTC AGE AGEU SEX RACE ARMCD ARM ACTARMCD
ACTARM ARMNRSN ACTARMUD COUNTRY;
   label
      STUDYID = "Study Identifier"
      DOMAIN = "Domain Abbreviation"
      USUBJID = "Unique Subject Identifier"
      SUBJID = "Subject Identifier for the Study"
      RFSTDTC = "Subject Reference Start Date/Time"
      RFENDTC = "Subject Reference End Date/Time"
      RFXSTDTC = "First Date/Time of Exposure"
      RFXENDTC = "Last Date/Time of Exposure"
      RFICDTC = "Informed Consent Date/Time"
      REPENDTC = "Date/Time of End of Participation"
      DTHDTC = "Date/Time of Death"
      DTHFL = "Subject Death Flag"
      SITEID = "Study Site Identifier"
      BRTHDTC = "Date/Time of Birth"
      AGE = "Age"
      AGEU = "Age Units"
      SEX = "Sex"
      RACE = "Race"
      ARMCD = "Planned Arm Code"
      ARM = "Description of Planned Arm"
      ACTARMCD = "Actual Arm Code"
      ACTARM = "Description of Actual Arm"
      ARMNRSN = "Reason Arm and/or Actual Arm is Null"
      ACTARMUD = "Description of Unplanned Actual Arm"
```

Codes / Comments Creation

Prompt :

See below comments and convert them in SAS codes.
Show only SAS codes comments:

Merge demo and rand dataset by usubjid and name it as dm

Output :

```
data dm;
  merge demo rand;
  by usubjid;
run;
```

Prompt :

See below SAS codes and provide SAS comments. Refer to examples. Show only SAS comments.

SAS codes:
data dm; merge demo rand; by usubjid; run;

Examples:
SAS codes:
proc sort data=dm; by usubjid; run;

SAS comments:
Sort dm by usubjid

ChatGPT

kotlin Copy code

```
data dm;
  merge demo rand;
  by usubjid;
run;
## Merge datasets demo and rand by variable usubjid
```

Confidential - Do not duplicate or distribute without the permission of Kevin Lee

Comparison between SAS and R

Prompt :
I want you to compare SAS and R programming. Provide me with differences, present in a table format

Output :

Certainly! Here is a comparison of SAS and R programming in a table format highlighting the key differences between the two:

Aspect	SAS	R
Type of Language	Proprietary language	Open-source language
Cost	Commercial software with licensing fees	Free and open-source software
Data Manipulation	Excellent data manipulation and transformation capabilities	Strong data manipulation capabilities
Data Visualization	Basic data visualization features; often requires external tools like SAS Visual Analytics	Rich data visualization libraries (e.g., ggplot2)
Statistical Analysis	Comprehensive built-in statistical procedures	Extensive statistical packages and libraries
Machine Learning	Limited machine learning capabilities	Extensive machine learning libraries (e.g., scikit-learn, TensorFlow)
Community Support	Smaller user community and less open-source collaboration	Large and active user community with extensive online resources
Customization	Limited customization options compared to open-source alternatives	Highly customizable with the ability to create packages and functions

Patient Profile Development

- Prompt :
  Use below data to provide patient profiles:
  { 'patient' : '01-001', 'sex' : 'male', 'age' : 25, 'race' : 'White'}

- Output :
  Based on the provided data, here is a patient profile for '01-001':

  - Patient ID: 01-001
  - Sex: Male
  - Age: 25
  - Race: White

  This patient is a 25-year-old male of White race. Please let me know if you need any additional information or if you would like to create profiles for more patients.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Data Analysis (1)

- Prompt :

  Select patients whose sex is male. Use below "Patient Data". Provide me the answers only.

  "Patient Data":
  { 'patient' : '01-001', 'sex' : 'male', 'age' : 25, 'race' : 'White'},
  { 'patient' : '01-002', 'sex' : 'female', 'age' : 28, 'race' : 'Asian'},
  { 'patient' : '02-001', 'sex' : 'female', 'age' : 50, 'race' : 'White'},
  { 'patient' : '02-002', 'sex' : 'male', 'age' : 40, 'race' : 'Black'}

- Output :

  The selected patients whose sex is male are:

  1. {'patient': '01-001', 'sex': 'male', 'age': 25, 'race': 'White'}
  2. {'patient': '02-002', 'sex': 'male', 'age': 40, 'race': 'Black'}

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Data Analysis (2)

- Prompt :

  Select patients whose age is older than or equal to 40. Use below "Patient Data". Provide me the answers only.

  "Patient Data":
  { 'patient' : '01-001', 'sex' : 'male', 'age' : 25, 'race' : 'White'},
  { 'patient' : '01-002', 'sex' : 'female', 'age' : 28, 'race' : 'Asian'},
  { 'patient' : '02-001', 'sex' : 'female', 'age' : 50, 'race' : 'White'},
  { 'patient' : '02-002', 'sex' : 'male', 'age' : 40, 'race' : 'Black'}

- Output :

  Answers:
  - {'patient': '02-001', 'sex': 'female', 'age': 50, 'race': 'White'}
  - {'patient': '02-002', 'sex': 'male', 'age': 40, 'race': 'Black'}

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Data Analysis (3)

- Prompt :

  Sort "Patient Data“ by age. Provide me the answers only.

  “Patient Data”:
  { 'patient' : '01-001', 'sex' : 'male', 'age' : 25, 'race' : 'White'},
  { 'patient' : '01-002', 'sex' : ‘female', 'age' : 28, 'race' : ‘Asian'},
  { 'patient' : ‘02-001', 'sex' : ‘female', 'age' : 50, 'race' : ‘White'},
  { 'patient' : ‘02-002', 'sex' : 'male', 'age' : 40, 'race' : ‘Black'}

- Output :

  Sorted "Patient Data" by age:

  1. {'patient': '01-001', 'sex': 'male', 'age': 25, 'race': 'White'}
  2. {'patient': '01-002', 'sex': 'female', 'age': 28, 'race': 'Asian'}
  3. {'patient': '02-002', 'sex': 'male', 'age': 40, 'race': 'Black'}
  4. {'patient': '02-001', 'sex': 'female', 'age': 50, 'race': 'White'}

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Data Analysis (4)

- Prompt :

  Count below “Patient Data” by race

  “Patient Data”:
  { 'patient' : '01-001', 'sex' : 'male', 'age' : 25, 'race' : 'White'},
  { 'patient' : '01-002', 'sex' : ‘female', 'age' : 28, 'race' : ‘Asian'},
  { 'patient' : ‘02-001', 'sex' : ‘female', 'age' : 50, 'race' : ‘White'},
  { 'patient' : ‘02-002', 'sex' : 'male', 'age' : 40, 'race' : ‘Black'}

- Output :

  Certainly! Let's count the number of patients by race in the given "Patient Data":

  - White: 2
  - Asian: 1
  - Black: 1

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Data Analysis (5)

- Prompt :

  Sort "Patient Data" by age. Provide me the answers only.

  "Patient Data":
  patient,sex,age,race
  01-001,male,25,White
  01-002,female,28,Asian
  02-001,female,50,White
  02-002,male,40,Black

- Output :

  Sorted "Patient Data" by age:

  1. 01-001, male, 25, White
  2. 01-002, female, 28, Asian
  3. 02-002, male, 40, Black
  4. 02-001, female, 50, White

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

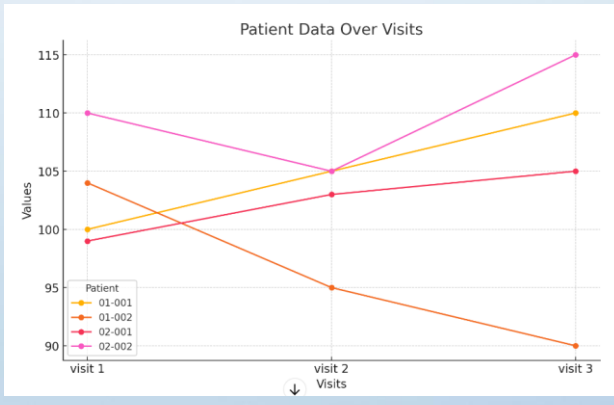
Data Visualization

- Prompt :

  Create line graph of "Patient Data"

  "Patient Data":
  { 'patient' : '01-001', 'visit 1' : 100, 'visit 2' : 105, 'visit 3' : 110},
  { 'patient' : '01-002', 'visit 1' : 104, 'visit 2' : 95, 'visit 3' : 90},
  { 'patient' : '02-001', 'visit 1' : 99, 'visit 2' : 103, 'visit 3' : 105,
  { 'patient' : '02-002', 'visit 1' : 110, 'visit 2' : 105, 'visit 3' : 115}

- Output :



Safety Cases Identification

- Prompt :

  Classify the safety cases of the text.
  Follow these steps in the classification.
  1. Identify adverse event in text.
  2. Identify patient information in text
  3. Identify medication information in text

  If text answers to "Yes" on all the steps, classify as 'Safety Issues' Otherwise 'No Safety Issues'

  Let's think step by step for classification of the text.

  Text: Patient 001 has adverse events after taking the medication.

- Output :

  Let's follow the steps you provided to classify the safety issues in the text:

  Step 1: Identify adverse event in text
  - The text mentions "Patient 001 has adverse events after taking the medication." This indicates that there is an adverse event mentioned in the text.

  Step 2: Identify patient information in text
  - The text mentions "Patient 001," which is patient information.

  Step 3: Identify medication information in text
  - The text mentions "taking the medication," which is medication information.

  Now, let's evaluate the results of each step:
  - Step 1: Adverse event identified
  - Step 2: Patient information identified
  - Step 3: Medication information identified

  Since all three steps have been answered "Yes," we can classify this text as 'Safety Issues' because it contains information about a patient experiencing adverse events after taking medication.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Mental Health Assessment 1 – Prompt

- Prompt :

I am a psychiatrist, and I am interviewing a patient for his PANSS score based on answers from the patient.

Use 'Anxiety Answers from patient' and match the 'PANSS Score Anxiety Score' of Absent, Minimal, Mild, Moderate, Moderate severe, Severe and Extreme based on its definition.

Anxiety Answers from patient:
1. Question : Have you been feeling worried or nervous in the past week? Answer : No
2. Question : Would you say that you are usually calm and relaxed? Answer : Yes
3. Question : What has been making you feel nervous (worried, not clam, not relaxed) Answer : I have not felt nervous.
4. Question : Just how nervous (worried, etc.) have you been feeling? Answer : I have not felt nervous
5. Question : Have you been shaking at times, or has your heart been racing? Answer : No
6. Question : Do you get into a state of panic? Answer : No
7. Question : Has your sleep, eating, or participation in activities been affected? Answer : No

'PANSS Score Anxiety Score' : Subjective experience of nervousness, worry, apprehension, or restlessness, ranging from excessive concern about the present or future to feelings of panic.
1. Absent - Definition does not apply
2. Minimal - Questionable pathology; may be at the upper extreme of normal limits
3. Mild - Expresses some worry, overconcern, or subjective restlessness, but no somatic and behavioral consequences are reported or evidence.
4. Moderate - Patient reports distinct symptoms of nervousness, which are reflected in mild physical manifestations such as fine hand tremor and excessive perspiration.
5. Moderate severe - Patient reports serious problems of anxiety which have significant physical and behavioral consequences, such as marked tension, poor concentration, palpitations, or impaired sleep.
6. Severe - Subjective state of almost constant fear associated with phobias, marked restlessness, or numerous somatic manifestations.
7. Extreme - Patient's life is seriously disrupted by anxiety, which is present almost constantly and at times reaches panic proportion or is manifested in actual panic attacks.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Mental Health Assessment 1 – Answer

- Answer :

Based on the patient's answers to the anxiety-related questions, you can match the PANSS Score Anxiety Score as follows:

1. **Question: Have you been feeling worried or nervous in the past week?**
   - Answer: No
   - **PANSS Score Anxiety Score: Absent**

2. **Question: Would you say that you are usually calm and relaxed?**
   - Answer: Yes
   - **PANSS Score Anxiety Score: Absent**

3. **Question: What has been making you feel nervous (worried, not calm, not relaxed)?**
   - Answer: I have not felt nervous.
   - **PANSS Score Anxiety Score: Absent**

4. **Question: Just how nervous (worried, etc.) have you been feeling?**
   - Answer: I have not felt nervous.
   - **PANSS Score Anxiety Score: Absent**

5. **Question: Have you been shaking at times, or has your heart been racing?**
   - Answer: No
   - **PANSS Score Anxiety Score: Absent**

6. **Question: Do you get into a state of panic?**
   - Answer: No
   - **PANSS Score Anxiety Score: Absent**

7. **Question: Has your sleep, eating, or participation in activities been affected?**
   - Answer: No
   - **PANSS Score Anxiety Score: Absent**

Based on the patient's responses, it seems that the subjective experience of anxiety is absent, and there are no signs or symptoms that warrant a PANSS Score Anxiety Score other than "Absent."

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Mental Health Assessment 2 – Prompt

- Prompt :

I am a psychiatrist, and I am interviewing a patient for his PANSS score based on answers from the patient.

Use 'Anxiety Answers from patient' and match the 'PANSS Score Anxiety Score' of Absent, Minimal, Mild, Moderate, Moderate severe, Severe and Extreme based on its definition. Final PANSS score will be the most severe assessment.

Anxiety Answers from patient:
1. Question : Have you been feeling worried or nervous in the past week? Answer : Yes
2. Question : Would you say that you are usually calm and relaxed? Answer : Not Applicable
3. Question : What has been making you feel nervous (worried, not clam, not relaxed) Answer : worried
4. Question : Just how nervous (worried, etc.) have you been feeling? Answer : a little
5. Question : Have you been shaking at times, or has your heart been racing? Answer : No
6. Question : Do you get into a state of panic? Answer : No
7. Question : Has your sleep, eating, or participation in activities been affected? Answer : No

'PANSS Score Anxiety Score' : Subjective experience of nervousness, worry, apprehension, or restlessness, ranging from excessive concern about the present or future to feelings of panic.
1. Absent - Definition does not apply
2. Minimal - Questionable pathology; may be at the upper extreme of normal limits
3. Mild - Expresses some worry, overconcern, or subjective restiessness, but no somatic and behavioral consequences are reported or evidence.
4. Moderate - Patient reports distinct symptoms of nervousness, which are reflected in mild physical manifestations such as fine hand tremor and excessive perspiration.
5. Moderate severe - Patient reports serious problems of anxiety which have significant physical and behavioral consequences, such as marked tension, poor concentration, palpitations, or impaired sleep.
6. Severe - Subjective state of almost constant fear associated with phobias, marked restlessness, or numerous somatic manifestations.
7. Extreme - Patient's life is seriously disrupted by anxiety, which is present almost constantly and at times reaches panic proportion or is manifested in actual panic attacks.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Mental Health Assessment 2 – Answer

- Answer :

Based on the patient's answers and the PANSS Score Anxiety definitions, we can match the responses to the corresponding scores:

1. Question: Have you been feeling worried or nervous in the past week?
   Answer: Yes
   PANSS Score Anxiety: Mild

2. Question: Would you say that you are usually calm and relaxed?
   Answer: Not Applicable
   PANSS Score Anxiety: Absent

3. Question: What has been making you feel nervous (worried, not calm, not relaxed)
   Answer: worried
   PANSS Score Anxiety: Moderate

4. Question: Just how nervous (worried, etc.) have you been feeling?
   Answer: a little
   PANSS Score Anxiety: Minimal

5. Question: Have you been shaking at times, or has your heart been racing?
   Answer: No
   PANSS Score Anxiety: Absent

6. Question: Do you get into a state of panic?
   Answer: No
   PANSS Score Anxiety: Absent

7. Question: Has your sleep, eating, or participation in activities been affected?
   Answer: No
   PANSS Score Anxiety: Absent

The final PANSS score would be the most severe assessment, which in this case is Moderate (Question 3). Therefore, the overall PANSS Score Anxiety for this patient based on the provided answers is Moderate.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Mental Health Assessment 3 – Prompt

- Prompt :

I am a psychiatrist, and I am interviewing a patient for his PANSS score based on answers from the patient.

Use 'Anxiety Answers from patient' and match the 'PANSS Score Anxiety Score' of Absent, Minimal, Mild, Moderate, Moderate severe, Severe and Extreme based on its definition. Final PANSS score will be the most severe assessment.

Anxiety Answers from patient:
1. Question : Have you been feeling worried or nervous in the past week? Answer : Yes
2. Question : Would you say that you are usually calm and relaxed? Answer : Not Applicable
3. Question : What has been making you feel nervous (worried, not clam, not relaxed) Answer : worried
4. Question : Just how nervous (worried, etc.) have you been feeling? Answer : a lot
5. Question : Have you been shaking at times, or has your heart been racing? Answer : Yes
6. Question : Do you get into a state of panic? Answer : Yes
7. Question : Has your sleep, eating, or participation in activities been affected? Answer : Yes

'PANSS Score Anxiety Score' : Subjective experience of nervousness, worry, apprehension, or restlessness, ranging from excessive concern about the present or future to feelings of panic.
1. Absent - Definition does not apply
2. Minimal - Questionable pathology; may be at the upper extreme of normal limits
3. Mild - Expresses some worry, overconcern, or subjective restiessness, but no somatic and behavioral consequences are reported or evidence.
4. Moderate - Patient reports distinct symptoms of nervousness, which are reflected in mild physical manifestations such as fine hand tremor and excessive perspiration.
5. Moderate severe - Patient reports serious problems of anxiety which have significant physical and behavioral consequences, such as marked tension, poor concentration, palpitations, or impaired sleep.
6. Severe - Subjective state of almost constant fear associated with phobias, marked restlessness, or numerous somatic manifestations.
7. Extreme - Patient's life is seriously disrupted by anxiety, which is present almost constantly and at times reaches panic proportion or is manifested in actual panic attacks.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Mental Health Assessment 3 – Answer

- Answer :

Based on the patient's answers, let's match them to the corresponding PANSS Score Anxiety Scores:

1. Question: Have you been feeling worried or nervous in the past week?
   Answer: Yes
   PANSS Score: Mild

2. Question: Would you say that you are usually calm and relaxed?
   Answer: Not Applicable
   PANSS Score: Absent

3. Question: What has been making you feel nervous (worried, not calm, not relaxed)?
   Answer: Worried
   PANSS Score: Mild

4. Question: Just how nervous (worried, etc.) have you been feeling?
   Answer: I was worried a lot
   PANSS Score: Moderate

5. Question: Have you been shaking at times, or has your heart been racing?
   Answer: Yes
   PANSS Score: Moderate Severe

6. Question: Do you get into a state of panic?
   Answer: Yes
   PANSS Score: Severe

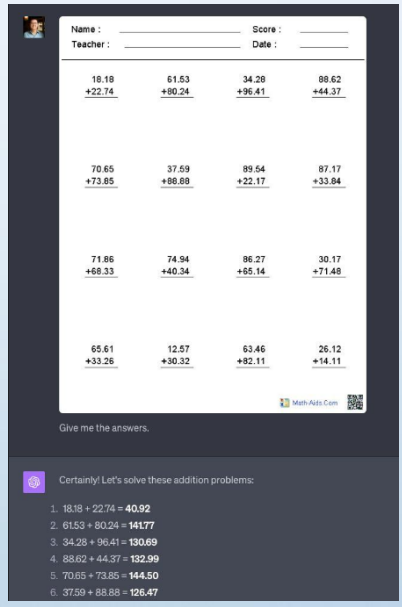
7. Question: Has your sleep, eating, or participation in activities been affected?
   Answer: Yes
   PANSS Score: Severe

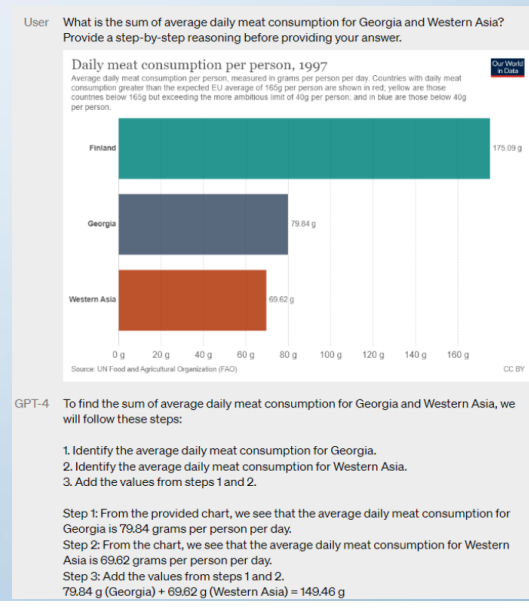
Based on the most severe assessment, the final PANSS Score for anxiety would be "Severe" as it is the highest level mentioned in the responses.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Image and Graphic recognition using ChatGPT4

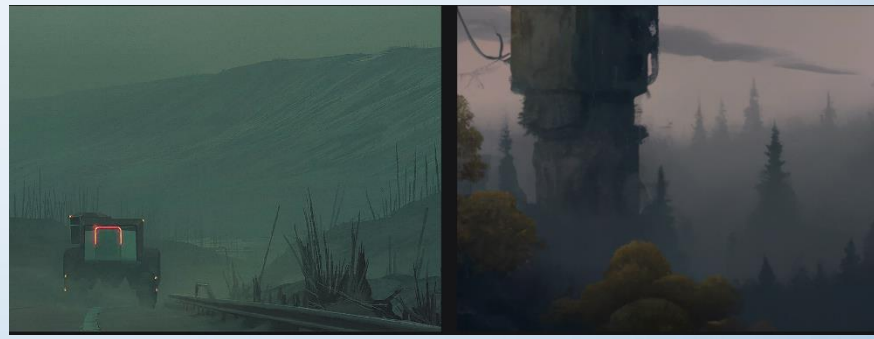
- ChatGPT 4 can analyze and comment on images and graphics while ChatGPT3.5 mainly focuses on text
- ChatGPT4
  - Describe the content of phone,
  - Identify trends in graph
  - Generate captions for images.





Art Generation using DALL.E 2

- DALL.E can generate creative, realistic visuals and art from a simple prompt by mixing concepts, characteristics and styles.



- A left drawing is the original artwork by Simon Stalenhag
- A right drawing from DALL.E generation for the prompt:
  Monumental old ruins tower of a dark misty forest, overcast, sci-fi digital painting by Simon Stalenhag

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

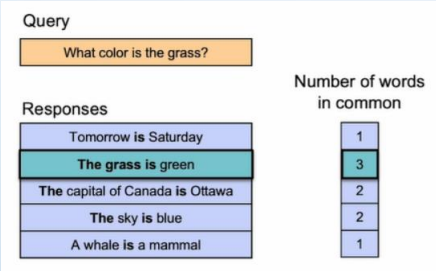
Video Generation using Sora

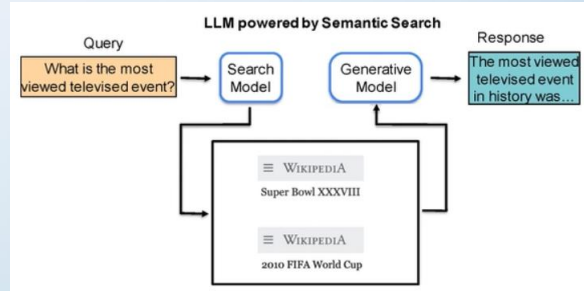
- OpenAI Sora can generate short video(60 second) based on simple prompt.

- Prompt :
  A bicycle race on ocean with different animals as athletes riding the bicycles with drone camera view

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Keyword Search (Google) vs Semantic Search (ChatGPT/Gemini/Copilot)





Confidential – Do not duplicate or distribute without the permission of Kevin Lee

How to use ChatGPT

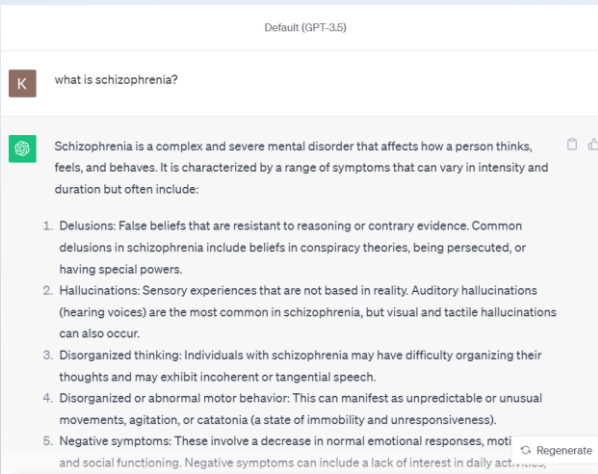
- What is Prompt?
- Introduction of Prompt Engineering

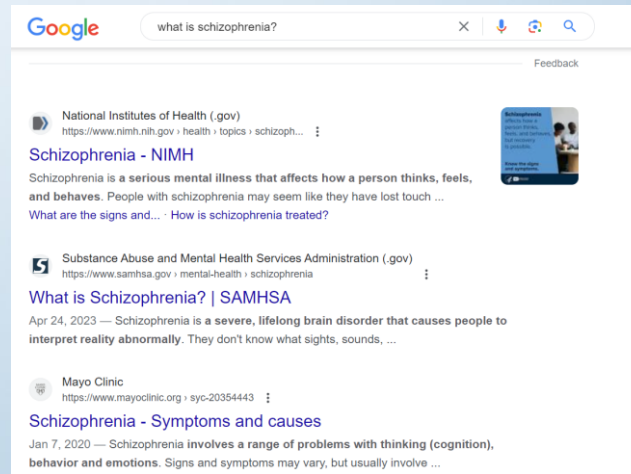
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt it is the new ~~Google it~~

What is Prompt?

- Prompt – A tool to communicate with Gen AI (ChatGPT)





How to prompt better

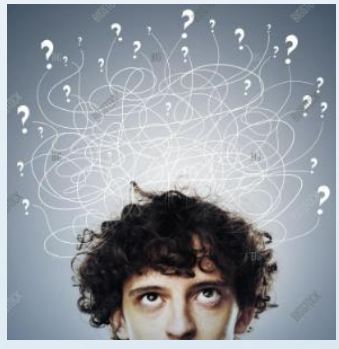
Prompt Engineering

75

What is Prompt Engineering?

- Prompt Engineering – the art of creating effective communication/ prompt that guide Gen AI (ChatGPT) to generate desired response.

Bad Prompt – Bad Response (e.g., Hallucination)



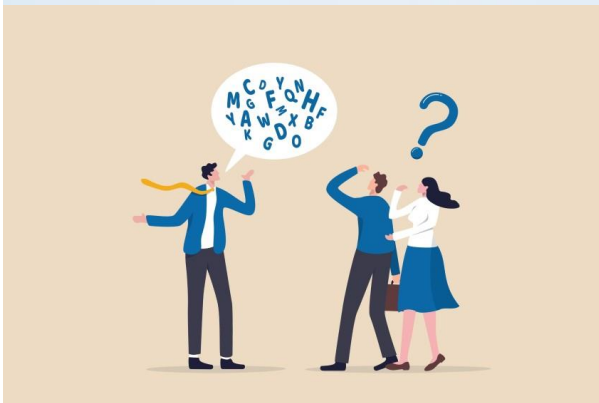
Good Prompt – Good Response

ChatGPT is like a very smart, capable, recently graduated intern/new hire.

=

77

How to work with the recently graduated, smart intern at work?
How to make an intern competent or incompetent?



78

How to prompt better in ChatGPT (How to communicate better)

- Get Specific
- Add Constraints
- Add more contexts
- Set the role
- Coach with feedback
- Probe further
- Step by step introduction

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Get specific

- Generic inquiry will provide generic (and less impressive) answers

Generic Prompt	Specific Prompt
Write an essay about flowers.	Write an essay about flowers at 9th grade level
Write me a rap song	Write me a rap song using references to SpongeBob Square Pants.
Write me a promotion request.	Write me a promotion request as Manager of Clinical Data Manager.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Add Parameter or Constraints

- Restriction and constraints breed a creativity
- Example of parameters
  - Under 100 words
  - Include word of "creative" / do not use a word of "creative"
  - Make it under $100
  - Make it work for 15 people

Generic Prompt	Prompt with parameter
Write an essay about flowers.	Write an essay about flowers at 9th grade level within 200 words
Tell me the best places to visit in Prague.	Tell me the best places to visit in Prague within 4 days.
Tell me the best places to visit in Prague.	Tell me the best places to visit in Prague with $500 budget.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Add more context

- If one adds more background, ChatGPT can provide more relevant, customized answers.
- Provide the reason that I am using ChatGPT

Generic Prompt	Prompt with more context
Write a greeting email.	I have not seen my friend for a long time. I want to know how he is doing. Write a greeting email.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Set the role

- If one gives a role to ChatGPT and ask, ChatGPT play a role given and provide the answers.
- Example of roles : boss, parent, salesperson,

Generic Prompt	Prompt with role
Write a greeting email.	I am a salesperson who markets new medicine. Write a greeting email.
Review this proposal and provide your feedback.	Pretend you are the head of IT. Review this proposal and provide your feedback.
How do you respond below proposal?	You are Steve Jobs. How do you respond below proposal?

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Coach ChatGPT with feedback

- After ChatGPT provide an answer, one can provide a feedback and ask again.
- By iterating, one can get the better answer

Generic Prompt	Prompt with feedback
Write a greeting email.	
	Make it in 50 words.
	Make it more formal.
	Make it more persuasive.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Provide a reference

- This technique is called 'Few-shot Learning'
- By providing reference/ examples, one could receive customized response.
- It will help ChatGPT to mimic writing style and format.

Generic Prompt	Prompt with feedback
Write a greeting email.	Read the following passage and analyze its style. I want you to copy this style to write a greeting email.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Probe further

- After receiving the answers, one could probe further in the current chat.
- One can find more detail information from the certain part of previous response.

Generic Prompt	Prompts with probing
	Tell me more about paragraph 2.
	Tell me more about option 1and 2.
	Please explain the attraction in paragraph 2.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Provide step by step instruction

- One can give step by step instruction on how ChatGPT think the answer "step by step"
- This will help ChatGPT to perform better.

Generic Prompt	Prompt with step by step introduction
	Perform the following actions: 1. Summarize the below text 2. Translate in French 3. Convert into JSON format

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt Engineering Technique

- Zero Shot
- Few Shot
- Reflection
- Chain of Thought
- Tree of Thought

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt Technique : Zero-Shot

- Purpose : One time Query
- Method : Straight-up Query

- Examples
  - Prompt : What is 2 + 8?
  - Output : 2 + 8 equals 10.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt Technique : Few-Shot (1)

- Purpose : Teaching AI models to follow specific patterns
- Method : Feeding model with examples before asking the desired questions

- Examples
  - Prompt :
    2@4 = 8
    3@3 = 9
    4@2 = 8
    5@3 = 15
    4@2 =
  - Output :
    It seems like you have a pattern where you're using "@" as an operator. Based on the examples you provided:
    2@4 = 8 (2 multiplied by 4)
    3@3 = 9 (3 multiplied by 3)
    4@2 = 8 (4 multiplied by 2)
    5@3 = 15 (5 multiplied by 3)
    So, following this pattern, 4@2 would be equal to 4 multiplied by 2, which is 8.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt Technique : Few-Shot Examples (2)

- Prompt :

  Classify these conversations:

  Text: Patients has Adverse event.
  Label: Safety Issue

  Text: Patients does not have Adverse event.
  Label: No Safety Issue

  Text: Patients has a headache.
  Label: Safety Issue

  Text: Patients feel sick after taking the medication
  Label: Safety Issue

  Text: Patients does not have any Adverse Event
  Label: No Safety Issue

  Text: Patients feel good after taking the medication
  Label:

- Output :

  Based on the provided texts and labels, here is the classification of these conversations:
  1. Text: Patients has Adverse event.
     Label: Safety Issue
  2. Text: Patients does not have Adverse event.
     Label: No Safety Issue
  3. Text: Patients has a headache.
     Label: Safety Issue
  4. Text: Patients feel sick after taking the medication.
     Label: Safety Issue
  5. Text: Patients does not have any Adverse Event.
     Label: No Safety Issue
  6. Text: Patients feel good after taking the medication.
     Label: No Safety Issue

  It appears that the classification is based on whether the patient experienced an adverse event (Safety Issue) or not (No Safety Issue) after taking the medication or as described in the text.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt Technique : Chain of Thought

- Purpose :
  - Teaching AI model the thought process it should use.
  - It is a subset of few-shot.
- Method : Breaking the prompt into simpler steps – step 1, step 2, step 3, step 4 and so on
- Examples
  - Prompt :

    Provide the answers
    Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can have 3 tennis balls. How many tennis balls does he have now?
    A: Rogers started with 5 balls.
    2 cans of 3 tennis balls each is 6 tennis balls.
    5 + 6 = 11. The answer is 11.
    Q: The cafeteria has 23 oranges. If they use 20 and buy 6 more. How many oranges do they have?
  - Output :

    A: The cafeteria started with 23 oranges.
    They used 20, so they have 23 - 20 = 3 oranges left.
    Then, they bought 6 more.
    So, they now have 3 + 6 = 9 oranges. The answer is 9.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt Technique : Chain of Thought 2

- **Prompt :**

Explain why SAS programmers need to learn prompt engineering.

Follow below steps:
1. Provide the introduction
2. Provide the two examples
3. Provide the conclusion with introduction and two examples

- **Output :**

Introduction:
SAS (Statistical Analysis System) programmers play a crucial role in extracting meaningful insights from data through programming and analytics. One essential skill for SAS programmers is prompt engineering, which involves creating interactive prompts to enhance the flexibility and usability of their programs. This skill is particularly important as it allows users to customize their analyses, making the programming process more dynamic and user-friendly.

Two Examples:

Example 1: Enhancing User Interaction

Consider a scenario where a SAS programmer develops a program to analyze sales data. Without prompt engineering, the program might be fixed to analyze a specific time period or product category. However, by incorporating prompts, the programmer can allow users to input their preferred date range or choose specific product categories. This enhances user interaction, making the program adaptable to different business scenarios and reducing the need for multiple versions of the same code.

Example 2: Streamlining Data Selection

In another example, a SAS programmer is tasked with creating a report that summarizes employee performance metrics. Instead of hardcoding specific employee IDs or performance criteria, prompt engineering enables the programmer to incorporate prompts for users to input the desired employee IDs, performance thresholds, or other relevant parameters. This ensures that the program can be reused for various teams or departments without the need for extensive code modification, thereby streamlining the data selection process.

Conclusion:

In conclusion, prompt engineering is a valuable skill for SAS programmers due to its ability to enhance user interaction and streamline data selection. By incorporating interactive prompts into their programs, SAS programmers empower users to customize analyses based on specific requirements, making the programs more versatile and user-friendly. The examples provided illustrate how prompt engineering enables adaptability, reducing the need for repeated code development and maintenance. Ultimately, mastering prompt engineering is essential for SAS programmers aiming to create robust and flexible solutions for data analysis and reporting.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt Technique : Chain of Thought 3

- Prompt :

Classify the safety issues of the text.
Follow these steps in the classification.
1. Identify adverse event in text.
2. Identify patient information in text
3. Identify medication information in text

If text answers to "Yes" on all the steps, classify as 'Safety Issues' Otherwise 'No Safety Issues'

Let's think step by step for classification of the text.

Text: Patient 001 has adverse events after taking the medication.

- Output :

Let's follow the steps you provided to classify the safety issues in the text:

Step 1: Identify adverse event in text
- The text mentions "Patient 001 has adverse events after taking the medication." This indicates that there is an adverse event mentioned in the text.

Step 2: Identify patient information in text
- The text mentions "Patient 001," which is patient information.

Step 3: Identify medication information in text
- The text mentions "taking the medication," which is medication information.

Now, let's evaluate the results of each step:
- Step 1: Adverse event identified
- Step 2: Patient information identified
- Step 3: Medication information identified

Since all three steps have been answered "Yes," we can classify this text as 'Safety Issues' because it contains information about a patient experiencing adverse events after taking medication.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt Technique : Reflection

- Purpose : To understand how ChatGPT4 generates the output.
- Method : Ask the question to ChatGPT, and ask 'Helper Agent' to explain the output / process of ChatGPT

- Examples
  - Prompt :

    ChatGPT and Helper Agent, I'd like to ask both of you. ChatGPT, provide answers to my questions, and Helper Agent, reflect on ChatGPT's responses to help optimize the conversation. After the Helper Agent's reflection, ChatGPT, provide a final response to the initial question that incorporates the insights from the Helper Agent's analysis. Let's begin. \[Your question here\]"
  - Output :
    - ChatGPT:
    - Helper Agent:

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt Template

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt Template

Categories	Explanation	Examples
# Identity	Identity or the roles that AI will play	A short sentence. I am SAS programmer.
# Task	Main task that AI will do for me.	A short sentence. Create SAS codes.
# Sub tasks	Subtasks, Step by Step Tasks that AI will follow to do main task. Refer to #Context and #Guidance and examples for tasks.	Multiple sentences. First read ## Guidance and examples for the reference. Do Step 1, 2 and 3.
# Context	Background of this task. Who you are State of projects Goal of projects	Multiple sentences. This is SAS codes that create SDTM DM domain from EDC raw data.
# Guidance and examples	Guidance or Examples that AI use or refer to for the task	Current SOP examples for the new SOP development task SDTM and raw metadata for SDTM development task

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt Template Examples

- Prompt :

Identity

Task

Sub Tasks

Context

Guidance and Example

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt Template Example - Prompt

Identity
I am SOP developer.

Task
I am writing SOP for TFL development

Sub Tasks
Refer to SOP for SDTM Development in #Guidance.
Write SOP for TFL development

Context
TFL is the abbreviation of Table, Figure and Listing. SAS programmers create TFL from ADaM and SDTM datasets based on SAP. TFL should be validated by independent SAS programmers.

Guidance
SOP for SDTM Development
1. Purpose
This Standard Operating Procedure (SOP) outlines the process for developing Study Data Tabulation Model (SDTM) datasets for clinical trials, ensuring data consistency, regulatory compliance, and efficient analysis.

2. Scope
This SOP applies to all personnel involved in SDTM development, including SAS programmers, data managers, and clinical trial leads.

3. References
•ICH E6 Good Clinical Practice (https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf)

4. Responsibilities
•SDTM Developer:
- Develops SDTM mapping specification.
- Develops SDTM datasets based on the approved mapping document.

•SDTM Validator:
- Validate SDTM datasets based on SDTM specification.
- Reviews and approves the final SDTM datasets.

5. Procedure
5.1. Pre-development Activities
•SDTM Developer reviews the clinical trial protocol and data collection forms.
•Identify data elements for SDTM mapping.
•Select the appropriate CDISC SDTM Implementation Guide based on the study type.

5.2. SDTM Mapping
•Map each data element from the clinical trial forms to the corresponding SDTM domain and variable.
•Utilize controlled terminologies (e.g., MedDRA) for coded data elements.
•Document variable labels, formats, definitions, and value ranges in SDTM mapping specification.

5.3. SDTM Dataset Development
•SDTM Developer creates SDTM datasets using SAS programming techniques.
•Ensure adherence to SDTM variable naming conventions and data types.
•Include variable labels, formats, and definitions within the SAS datasets.
•Implement data cleaning and validation procedures to ensure data quality.

5.4. Review and Approval
•SDTM validator reviews the developed SDTM datasets for accuracy and completeness.

6. Training
•Train personnel involved in data analysis and reporting on the structure and content of the SDTM datasets.

7. Records
•Maintain records of all SDTM development documents, including mapping documents, programming code, and data cleaning logs.

8. Deviations
•Any deviations from this SOP must be documented with justification and approval from the Programming Lead.

9. Revision History
•Maintain a revision history of this SOP, including the date of each revision and a brief description of the changes made.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt Template Examples - Output

SOP for TFL Development

1. Purpose
This Standard Operating Procedure (SOP) outlines the process for developing Table, Figure, and Listing (TFL) outputs from Analysis Data Model (ADaM) and Study Data Tabulation Model (SDTM) datasets for clinical trials. The objective is to ensure the creation of accurate, consistent, and regulatory-compliant TFLs that support efficient data analysis.

2. Scope
This SOP applies to all personnel involved in TFL development, including SAS programmers, data managers, and clinical trial leads.

3. References
• ICH E6 Good Clinical Practice (https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf)
• SDTM Development SOP

4. Responsibilities
TFL Developer:
• Develops TFL mapping specification based on the Analysis Plan (SAP).
• Creates TFL outputs using SAS programming techniques.
• Implements validation checks to ensure TFL accuracy and consistency.
Independent SAS Programmer (TFL Validator):
• Validates TFL outputs based on TFL mapping specification.
• Reviews and approves the final TFL outputs.

5. Procedure

5.1. Pre-development Activities
• TFL Developer reviews the SAP and identifies variables for TFL mapping.
• Identify any specific requirements or considerations for TFL development outlined in the SAP.

5.2. TFL Mapping
• Map each variable identified in the SAP to the corresponding TFL output.
• Utilize controlled terminologies when applicable.
• Document output labels, formats, definitions, and any statistical methodology used in the TFL mapping specification.

5.3. TFL Development
• TFL Developer creates TFL outputs using SAS programming techniques.
• Ensure adherence to TFL output naming conventions and formatting.
• Include output labels, formats, and definitions within the output files.
• Implement data quality checks and validation procedures.

5.4. Review and Approval
• Independent SAS Programmer (TFL Validator) reviews the developed TFL outputs for accuracy and completeness.
• Any discrepancies or issues identified during the validation process are documented and addressed by the TFL Developer.

6. Training
• Train personnel involved in TFL development and validation on the structure and content of TFL outputs.

7. Records
• Maintain records of all TFL development documents, including mapping documents, programming code, and validation logs.

8. Deviations
• Any deviations from this SOP must be documented, providing justification and obtaining approval from the Programming Lead.

9. Revision History
• Maintain a revision history of this SOP, including the date of each revision and a brief description of the changes made.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

The Anatomy of an o1 Prompt

I want a list of the best medium-length hikes within two hours of San Francisco.

Each hike should provide a cool and unique adventure, and be lesser known.

For each hike, return the name of the hike as I'd find it on AllTrails, then provide the starting address of the hike, the ending address of the hike, distance, drive time, hike duration, and what makes it a cool and unique adventure.

Return the top 3.

Be careful to make sure that the name of trail is correct, that it actually exists, and that the time is correct.

--

For context: my girlfriend and i hike a ton! we've done pretty much all of the local SF hikes, whether that's presidio or golden gate park. we definitely want to get out of town -- we did mount tam pretty recently, the whole thing from the beginning of the stairs to stinson -- it was really long and we are definitely in the mood for something different this weekend! ocean views would still be nice. we love delicious food. one thing i loved about the mt tam hike is that it ends with a celebration (Arriving in town to breakfast!) The old missile silos and stuff near Discovery point is cool but I've just done that hike probably 20x at this point. We won't be seeing eachother for a few weeks (she has to stay in LA for work) so the uniqueness here really counts.

Goal

Return Format

Warnings

Context Dump

Prompt Template Examples

- Prompt 1: Pareto Principle (80/20 Rule)
  "Using the Pareto Principle, help me identify the 20% of activities in my [personal life/professional life/business goals] that yield 80% of my results. Provide a step-by-step guide for analyzing my current efforts, pinpointing high-impact tasks, and creating an action plan to prioritize these tasks. Additionally, suggest ways to delegate, automate, or minimize the remaining 80% of lower-impact activities to free up time for what truly matters."

- Prompt 2: SMART Goals
  "Help me create three SMART goals for [specific time frame] that align with my [personal/professional/fitness/etc.] objectives. For each goal, provide a clear breakdown of the goal itself, milestones to track progress, specific metrics to measure success, potential challenges I may face, and strategies to overcome these challenges and stay on track."

- Prompt 3: The 5 Whys Method
  "Using the 5 Whys technique, help me identify the root cause of a recurring issue I face in [specific area, e.g., productivity, personal relationships, project management]. Provide a detailed step-by-step approach to uncover the core problem by repeatedly asking 'Why?' until the root cause is clear. Then, offer actionable solutions to address the root issue and prevent similar challenges in the future."

- Prompt 4: Atomic Habits
  "Using the principles from Atomic Habits by James Clear, help me build a sustainable habit for [specific goal]. Provide a comprehensive framework that includes habit stacking ideas to make this habit easier to integrate into my daily routine, clear cues and triggers to initiate the habit, strategies for tracking progress and staying motivated, and a reinforcement plan to make the habit stick over the next [time frame]."

- Prompt 5: Ikigai Framework
  "Using the Ikigai framework, help me find clarity about how to align [my career/personal life/business] with my purpose by identifying what I love doing, what I'm skilled at, what the world needs, and what I can be paid for. Based on this alignment, provide actionable steps for creating a fulfilling [career/life plan] that incorporates all four elements of Ikigai."

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

What can we (Stat Programmers and Biostatisticians) accomplish when we are fully utilizing with prompt engineering?

Prompt Formula

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

6 Prompt Components

More important

1. Task
2. Context
3. Example
4. Persona
5. Format
6. Tone

Optional

Task:
Generate an open source strategy for the biometrics department.
Context:
The user is a senior biometrics professional working in a regulated clinical research setting, aiming to reduce reliance on commercial tools and promote innovation through open source adoption.
Example:
For instance, the strategy could include evaluating tools like R for statistical programming, Git for version control, and defining contribution guidelines for internal package development.
Persona:
You are an experienced open source strategist familiar with clinical data workflows, regulatory environments, and open science principles.
Format:
Provide a structured outline in bullet points, organized into key sections: Goals, Current State, Tool Evaluation, Governance, Implementation Plan, and Recommendations.
Tone:
Professional, clear, and practical—suitable for presentation to department leadership.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Task

- Always start “Task sentence” with an Action Verb
  - Generate
  - Give
  - Write
  - Analyze…
- One Simple Task
  - Generate Open Source strategy for biometrics department.
- Multi-Task Request
  - Assess current tools and needs for biometrics department
  - Research and Evaluate Open Source Solutions
  - Draft Open Source Strategy Document

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Context

- To generate a great context, ask three questions to self.
  1. What is the user's background?
  2. What does success look like?
  3. What environment are they in?

- Example context
  1. background : I am the head of Biometrics Department.
  2. Success Criteria : I want to build a clear, compliant, practical open source strategy that improves efficiency and encourages the innovation within 1 or 2 years.
  3. Environment : The work should be done in a regulated clinical trial setting with a mix of proprietary tools like SAS and open source tools like R and Python.

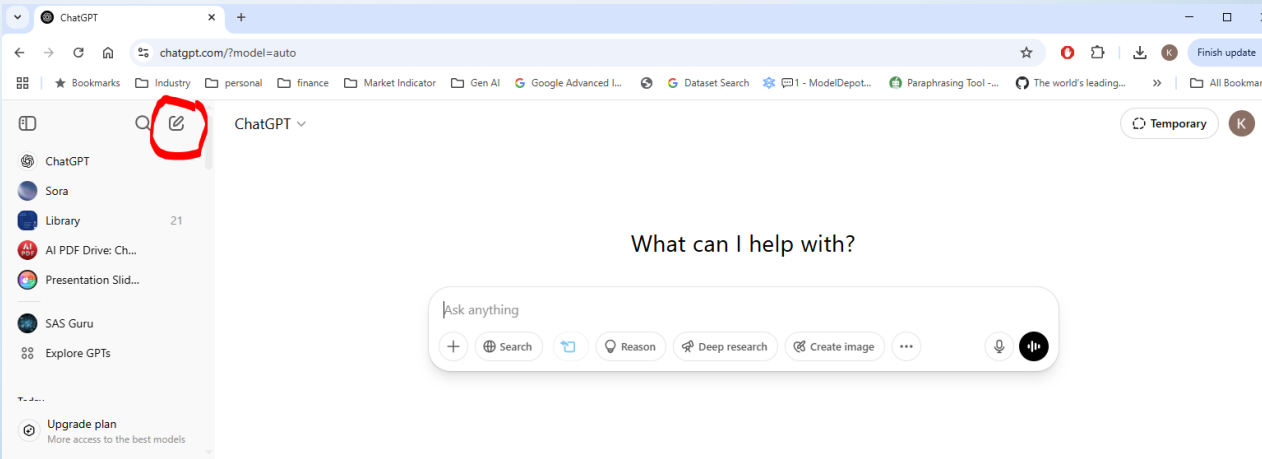
  - Then, add "Generate Open Source strategy for biometrics department" to the end.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

ChatGPT Functionality

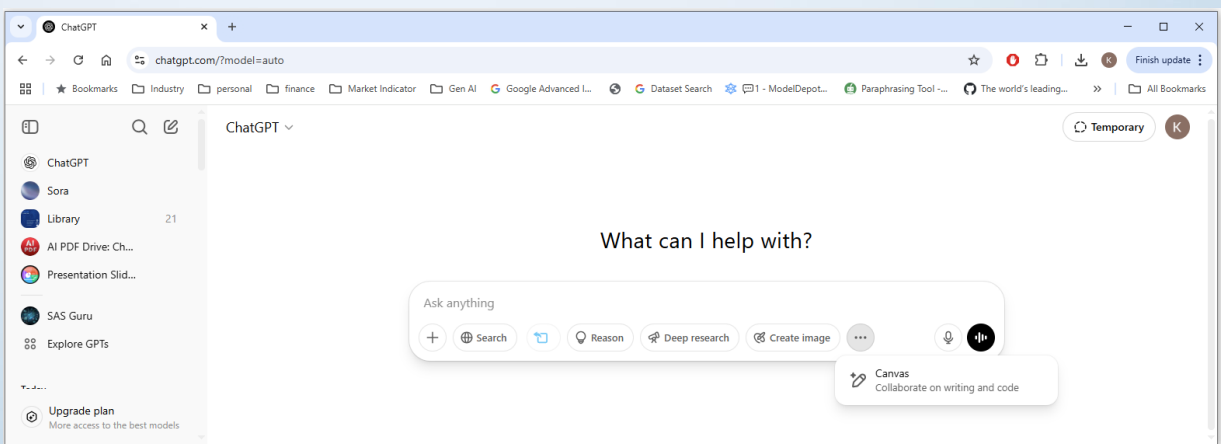
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Start new chat



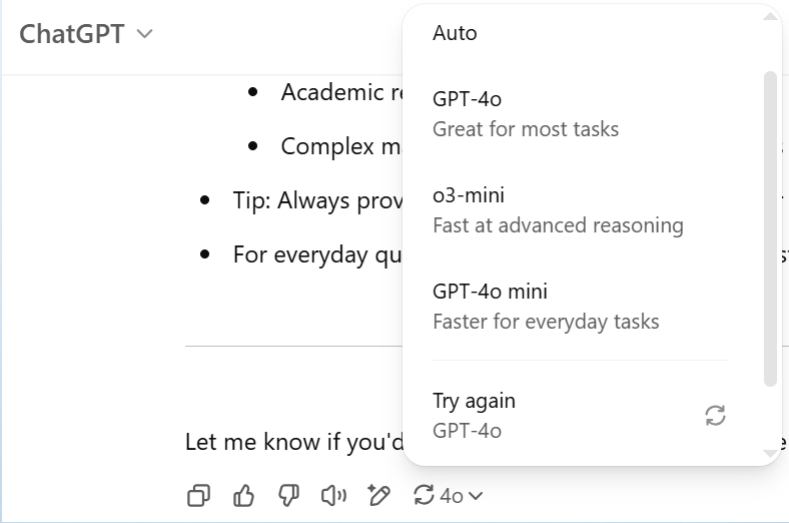
- Go to https://chatgpt.com/?model=auto
- Click "New Chat" to start the new chat.

different function



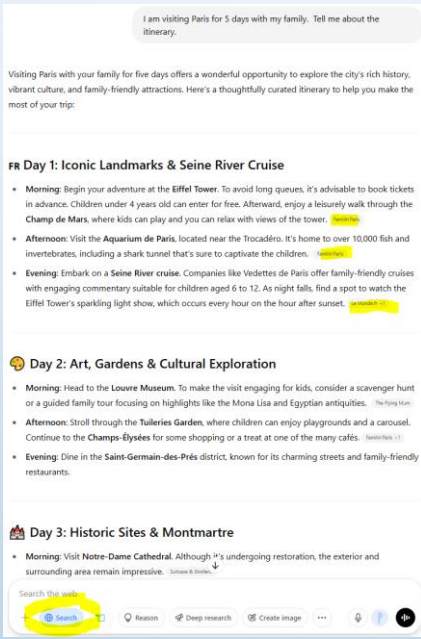
- "+" upload the files.

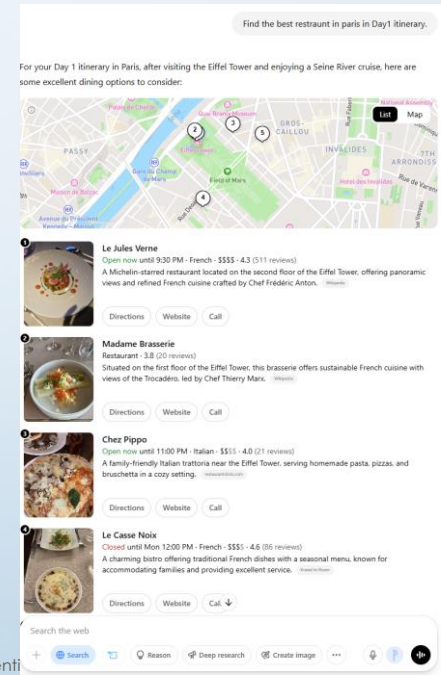
Select different mode;



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Search





- Click "Search" and prompt to have response and search link together.

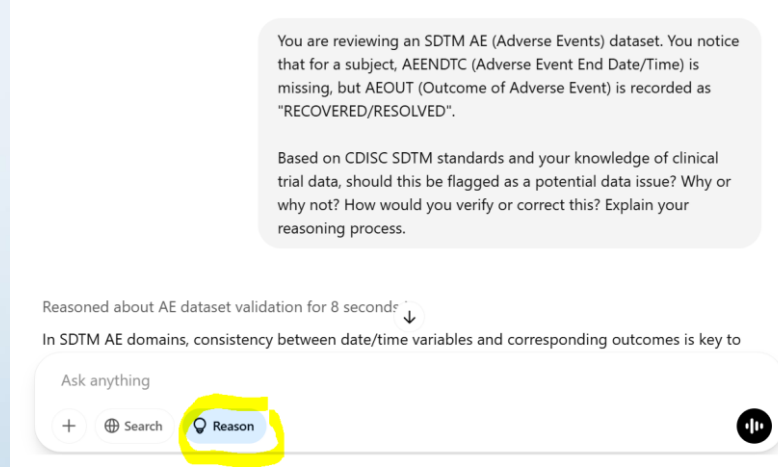
nfidenti n of Kevin Lee

Reason

Reasoning Prompt :

You are reviewing an SDTM AE (Adverse Events) dataset. You notice that for a subject, AEENDTC (Adverse Event End Date/Time) is missing, but AEOUT (Outcome of Adverse Event) is recorded as "RECOVERED/RESOLVED".

Based on CDISC SDTM standards and your knowledge of clinical trial data, should this be flagged as a potential data issue? Why or why not? How would you verify or correct this? Explain your reasoning process.



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Reason

Output :

In SDTM AE domains, consistency between date/time variables and corresponding outcomes is key to accurately characterizing the event. Here's a detailed breakdown of the issue and the recommended steps:

Identifying a Potential Data Issue

•Expectation from Standards:
According to CDISC SDTM standards, if an adverse event is recorded as "RECOVERED/RESOLVED" (via AEOUT), it implies that the adverse event has ended. Therefore, there should typically be a corresponding AEENDTC (Adverse Event End Date/Time) value that marks when the event resolved.

•Flagging the Inconsistency:
When AEENDTC is missing despite AEOUT indicating that the event resolved, it can be flagged as a potential data quality issue because:

- •It may signal incomplete or inaccurately captured event timelines.
- •It could affect the analyses that depend on the duration of events, such as time-to-resolution or incidence rates.
- •It may lead to challenges in regulatory review, where precise event timing is necessary for safety assessments.

Verification and Correction Process

1.Review Source Documentation:

- **•Case Report Forms (CRFs) or Medical Records:** Verify the original source documents to see if an end date was recorded at the time of data entry.
- **•Adverse Event Narratives:** Check if there is additional descriptive information or notes indicating when the event resolved, even if the structured field is missing.

2.Consult Data Management or Clinical Teams:

- •Engage with the data management or clinical safety team to confirm if the absence of AEENDTC is a data entry oversight, a coding anomaly, or if there's an approved rationale (e.g., if the end date wasn't captured due to a protocol-specific condition).

3.Data Review and Cross-Checks:

- **•Cross-Reference with Other Variables:** Look at related variables such as AESTDTC (start date/time of the adverse event) and any flags or comments that might hint at the timeline of the event.
- **•Consistency Rules:** Apply internal consistency checks used in your study's data review procedures to see if similar issues occur across other records.

4.Correction of Data:

- **•Updating Missing Dates:** If source documentation confirms an event resolution date, update AEENDTC accordingly.
- **•Documentation of the Change:** Ensure that any changes are well-documented in your data review logs, and that a query is raised (and resolved) through your study's quality control processes.
- **•Data Cleaning SOPs:** Follow your sponsor's or regulatory guidance for data correction in SDTM datasets, ensuring compliance with validation and audit trail requirements.

Reasoning Process Recap

•Standards Compliance:
The SDTM standards stipulate that an adverse event's timeline is essential for its characterization. Since AEOUT indicates the event resolved, logically, AEENDTC should not be missing.

•Impact on Analysis:
Missing resolution dates can affect safety analyses, signal detection, and regulatory reviews, warranting further investigation.

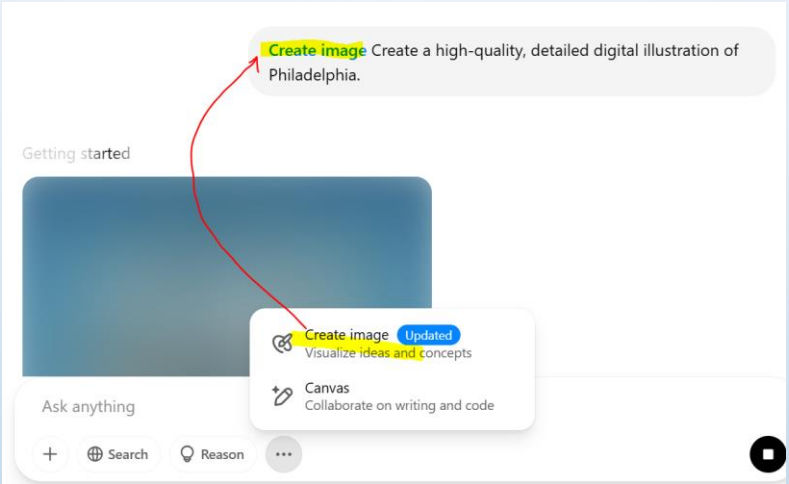
•Data Verification:
The process involves cross-referencing source documents, consulting with study teams, and applying consistency checks—all designed to ensure that the dataset accurately reflects the clinical scenario.

Conclusion

Yes, the situation should be flagged as a potential data issue because the absence of AEENDTC contradicts the indication that the event resolved. Verifying the source data and discussing the discrepancy with data management are necessary steps to either confirm an error or understand the rationale behind the missing date. Correcting it—if necessary—by updating the dataset based on authenticated source documentation will improve the dataset's integrity and ensure compliance with CDISC SDTM standards.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

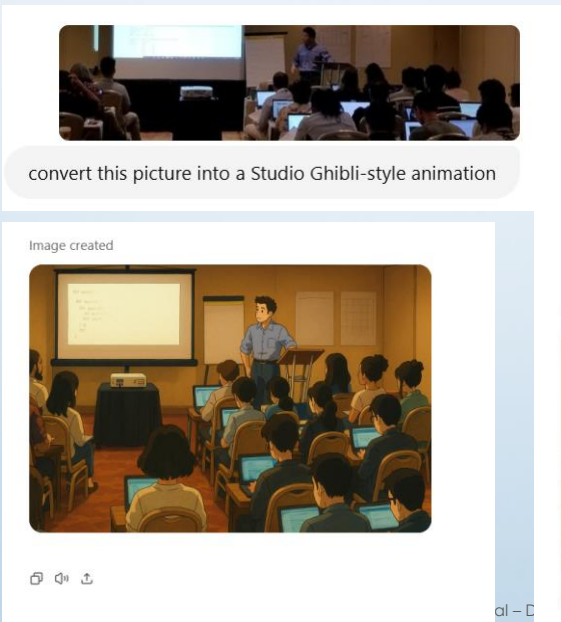
Create Image



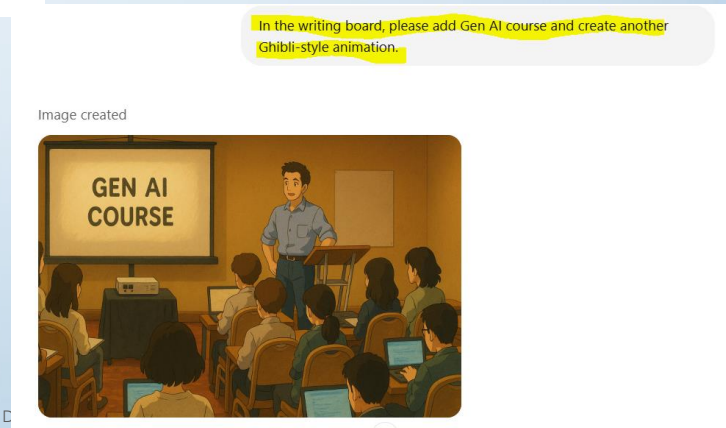
This will take some time to create a image

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

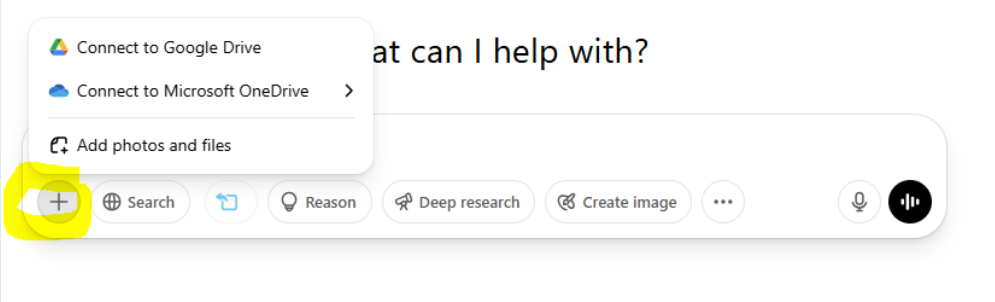
Create Image



This is a famous Ghibli-style animation prompt.



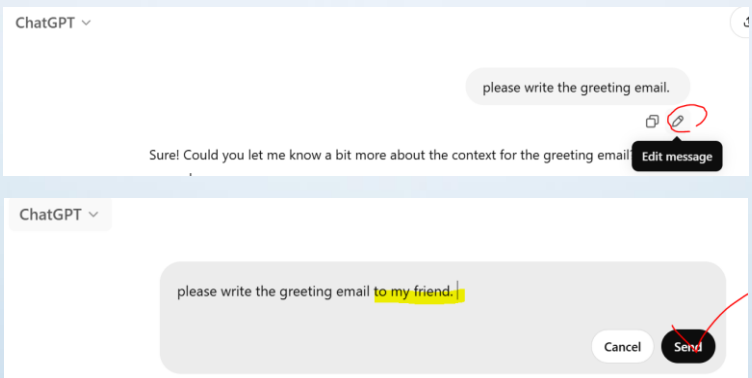
Upload the document



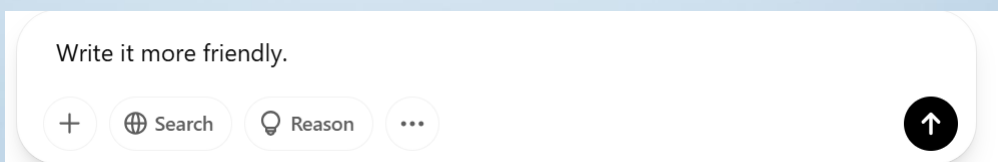
- Click + sign,or paste over prompt panel.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

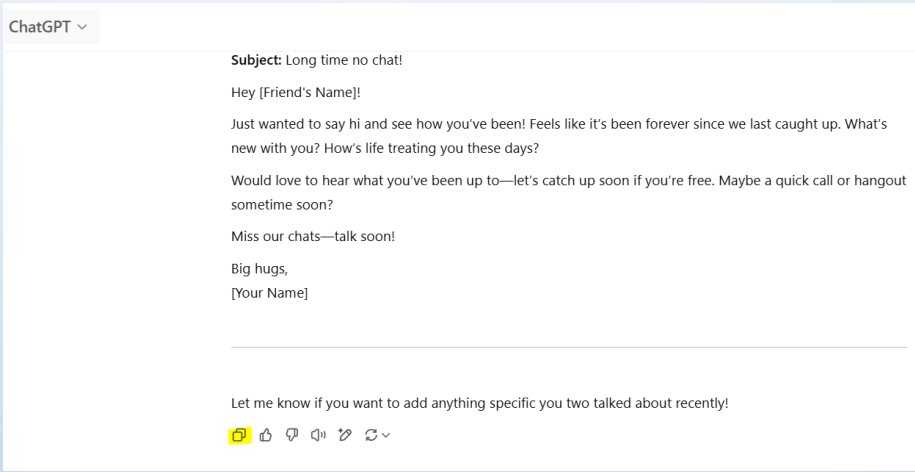
Modify and resubmit the inquiry



- We can also re-prompt to make it better.



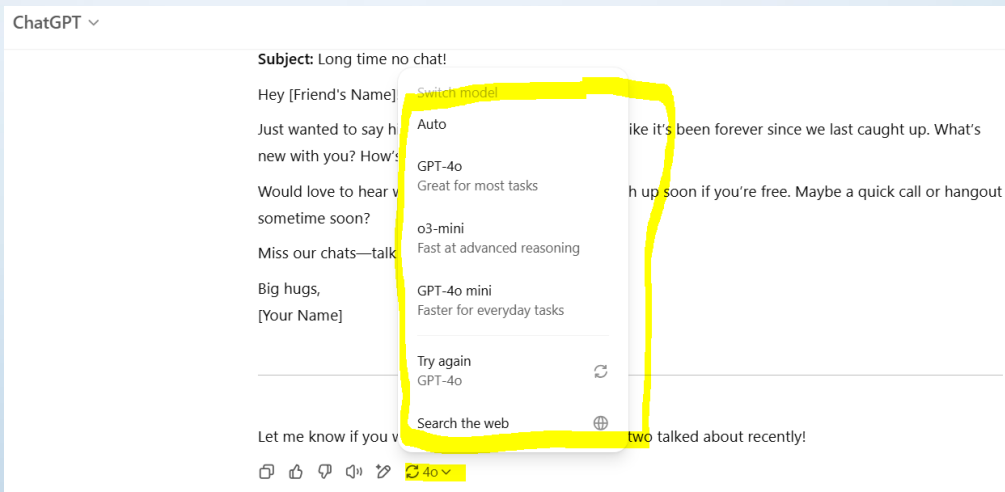
Copy the answer



- Click "Copy" icon.
- Paste the saved texts in other application.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

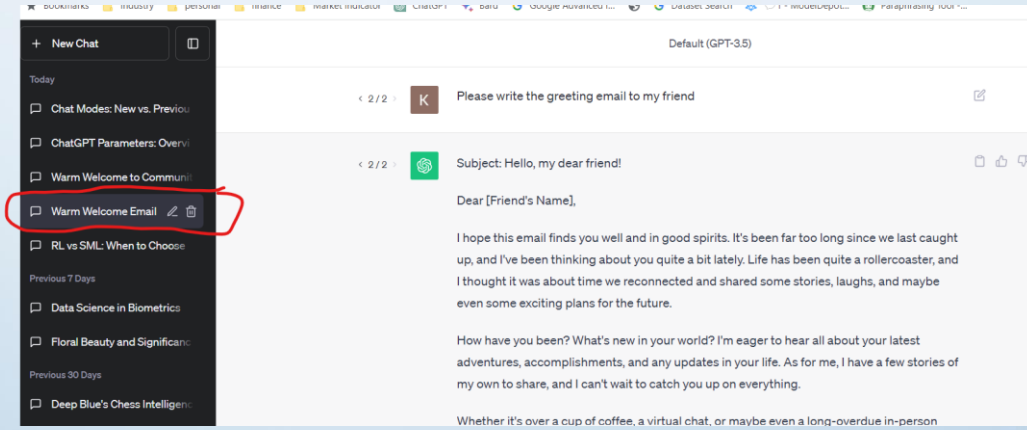
Regenerate the answer



- Click "Regenerate" icon.
- We can get the same contents with different model, but a little different answers.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

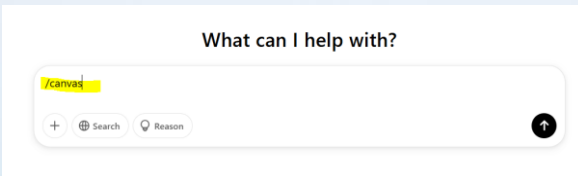
Continue the previous Chat

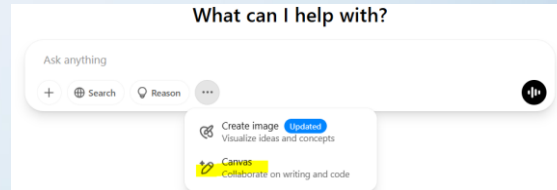


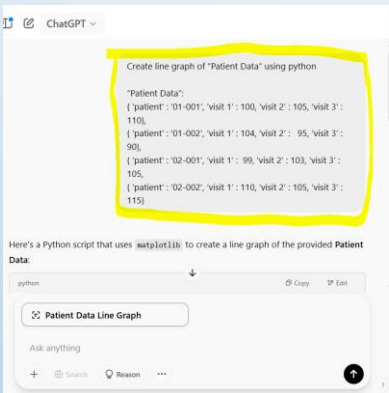
- Click previous chat.
- In summary, "new chat" starts a conversation from scratch, while "previous chat" continues an ongoing conversation within the same session, leveraging the context and information established in earlier messages.

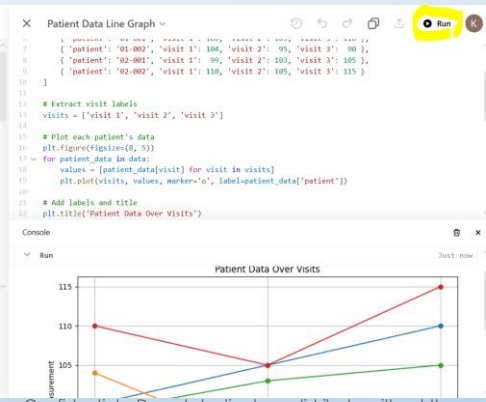
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Using OpenAI Canvas (writing and coding)









- Start Canvas.
- prompt for codes
- Run the code, and you will see the results in Console.
- Note: There is a limit to use Canvas in free version (few times a day)

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Using OpenAI Canvas (prototyping in html)

ChatGPT

Canvas
Create line graph of "Patient Data" in html.

"Patient Data":
{ 'patient' : '01-001', 'visit 1' : 100, 'visit 2' : 105, 'visit 3' : 110},
{ 'patient' : '01-002', 'visit 1' : 104, 'visit 2' : 95, 'visit 3' : 90},
{ 'patient' : '02-001', 'visit 1' : 99, 'visit 2' : 103, 'visit 3' : 105,
{ 'patient' : '02-002', 'visit 1' : 110, 'visit 2' : 105, 'visit 3' : 115)

Patient Data Linegraph

Answer in chat instead >

Your HTML file with a line graph of the patient data is ready. It's using Chart.js to plot the measurements over three visits. Let me know if you'd like to add interactivity, export features, or style changes!

Ask anything

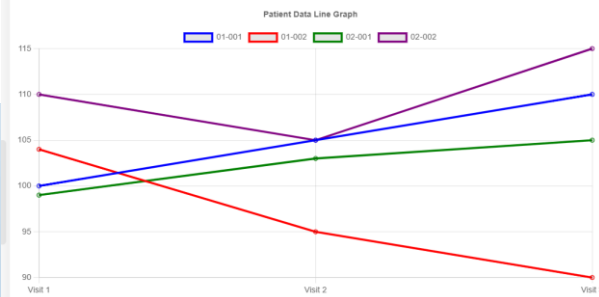
Search Reason

Patient Data Linegraph Preview

```
<!DOCTYPE html>
<html>
<head>
  <meta charset="UTF-8">
  <title>Patient Data Line Graph</title>
  <script src="https://cdn.jsdelivr.net/npm/chart.js"></script>
</head>
<body>
  <h2>Patient Data Over Visits</h2>
  <canvas id="patientChart" width="800" height="400"></canvas>
  <script>
    const ctx = document.getElementById('patientChart').getContext('2d');

    const data = {
      labels: ['Visit 1', 'Visit 2', 'Visit 3'],
      datasets: [
        {
          label: '01-001',
          data: [100, 105, 110],
          borderColor: 'blue',
          fill: false
        },
        {
          label: '01-002',
          data: [104, 95, 90],
          borderColor: 'red',
          fill: false
        },
        {
          label: '02-001',
          data: [99, 103, 105],
```

Patient Data Linegraph Stop

Patient Data Over Visits



Note : The limited trials in free version (a couple of times per a day

Confidential – Do not duplicate or distribute without the permission of Kevin Lee 123

Deep Research

What are you researching?

Build AI Strategy for Biostatistical department in Pharma Industry.

Search Reason Deep research Create image

Copy

To tailor the AI strategy effectively, could you please clarify a few points:

1. What is the size and structure of the biostatistics department (e.g., global vs. local, number of team members)?
2. What are the primary responsibilities of the department (e.g., clinical trial data analysis, regulatory submissions, RWE)?
3. What types of data are typically handled (e.g., clinical trial data, omics data, EHRs)?
4. Are there existing AI initiatives or tools in use?
5. What are the key goals or challenges you want the AI strategy to address (e.g., automation, predictive modeling, regulatory compliance)?

This will help in designing a strategy that aligns with your operational needs and long-term vision.

Please use below answer.
1. 500 stat programmers, biostatisticians, data managers
2. clinical trial data analysis, submission, efficcay and safety analysis, RWE
3. clinical trial data
4. Not yet
5. automation, faster time of drug development, regulatory compliance

Thanks for the detailed context. I'll develop an AI strategy tailored for a biostatistics department in the pharma industry with 500 professionals, focusing on clinical trial data analysis, regulatory submissions, and RWE, aiming to improve automation, speed up drug development, and maintain compliance.

I'll structure this strategy around key pillars such as data infrastructure, model deployment, regulatory considerations, and team upskilling. I'll get back to you shortly with a comprehensive strategy document.

Research completed in 6m · 34 sources · 79 searches

AI Strategy for Biostatistics in Pharma

AI Strategy for Biostatistics Department in Pharma

Executive Summary: This strategy outlines how the biostatistics department can embed AI across the clinical trial lifecycle to automate manual tasks, speed development, and maintain compliance. Key elements include deploying AI use cases (e.g. automated SDTM/ADaM mapping, anomaly detection, trial simulation), building a secure, GxP-ready AI infrastructure, upskilling staff and recruiting data scientists, and instituting a governance framework aligned with FDA/EMA guidelines. Short-term pilots will demonstrate quick wins in data quality and reporting; medium-term goals focus on expanding to design optimization and RWE analytics; and long-term goals aim for full AI integration in data analysis and regulatory submissions. The proposed roadmap and table below summarize prioritized use cases, impact, and feasibility.

AI Use Cases Across Clinical Trial Lifecycle

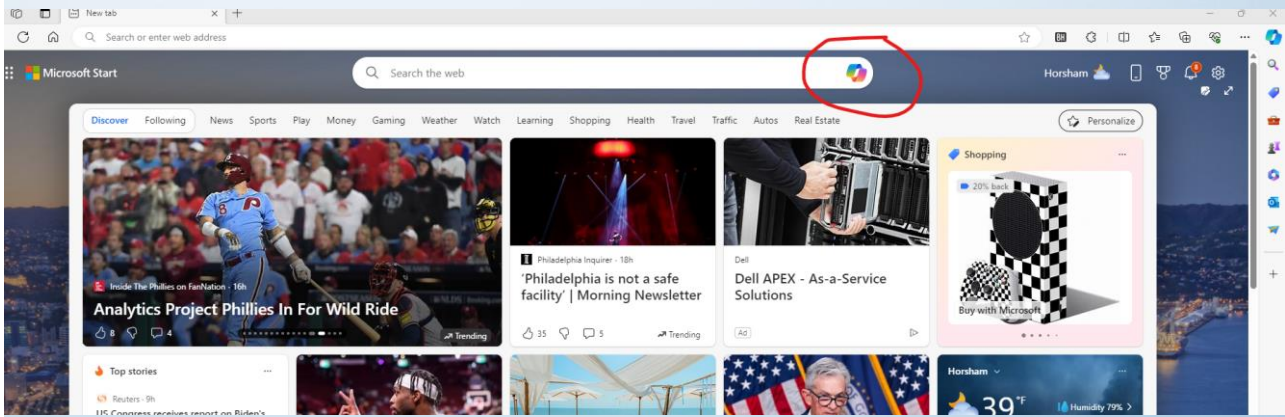
- **SDTM/ADaM Automation:** Use machine learning to auto-map clinical data into CDISC standards. AI + human-in-the-loop approaches can **clean, map and validate** source data to SDTM, dramatically reducing manual effort and error appliedclinicaltrialsonline.com . AbbVie's pilots show such AI frees scientists for interpretation and cuts SDTM conversion timelines/costs appliedclinicaltrialsonline.com , Extending to ADaM (analysis datasets) can further streamline analysis-ready data generation.
- **Data Quality & Anomaly Detection:** AI can continuously scan incoming trial data to flag outliers, inconsistencies, or protocol deviations. Algorithms trained on historical data detect subtle anomalies (e.g. implausible lab values or site irregularities) that human review might miss eclipsesol.com clinion.com . This real-time monitoring improves data integrity, reduces rework, and shortens queries, leading to **faster, more reliable results** eclipsesol.com clinion.com ,
- **Protocol Design & Simulation:** Machine learning and simulation models can optimize trial parameters before launch. For example, discrete-event simulation of enrollment and dropout scenarios helps refine inclusion criteria and sample size coherentsolutions.com . Reinforcement learning can adapt designs mid-trial to balance efficacy and risk coherentsolutions.com . An FDA-endorsed Alzheimer's trial simulator showed AI-driven

plicate or distribute

MS Bing Copilot Functionality

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

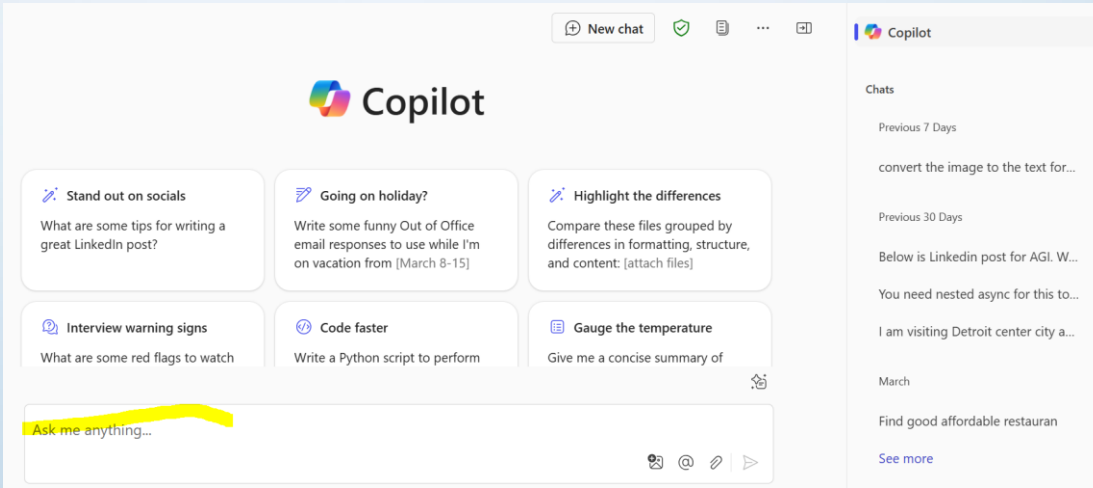
Start Bing Copilot (free version – Copilot with GPT-4)



- Start "Edge" and Click "Open Copilot"
- https://copilot.Microsoft.com

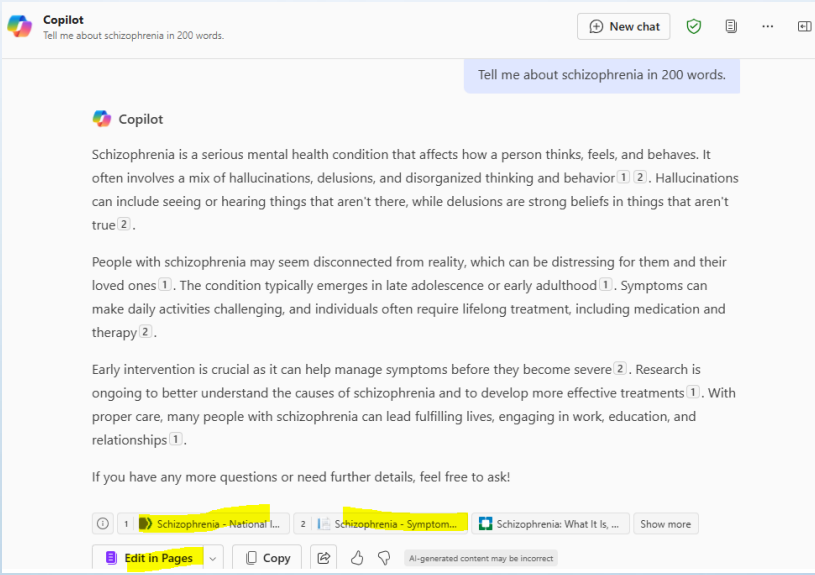
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Start new chat



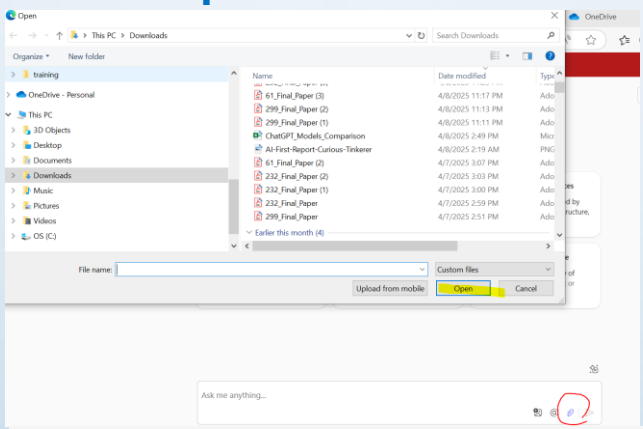
- Start your prompt in Copilot

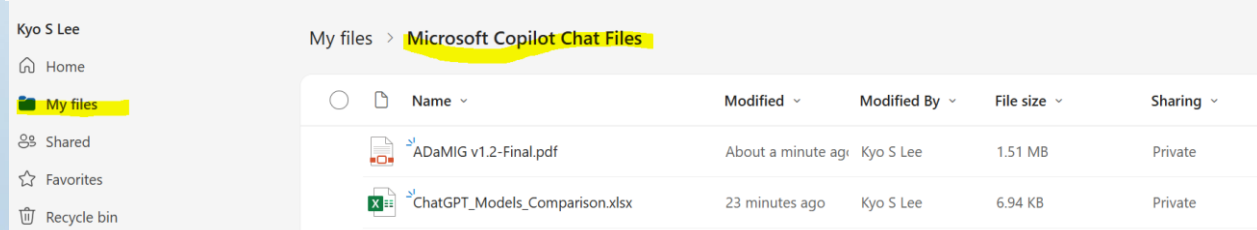
Results from prompt



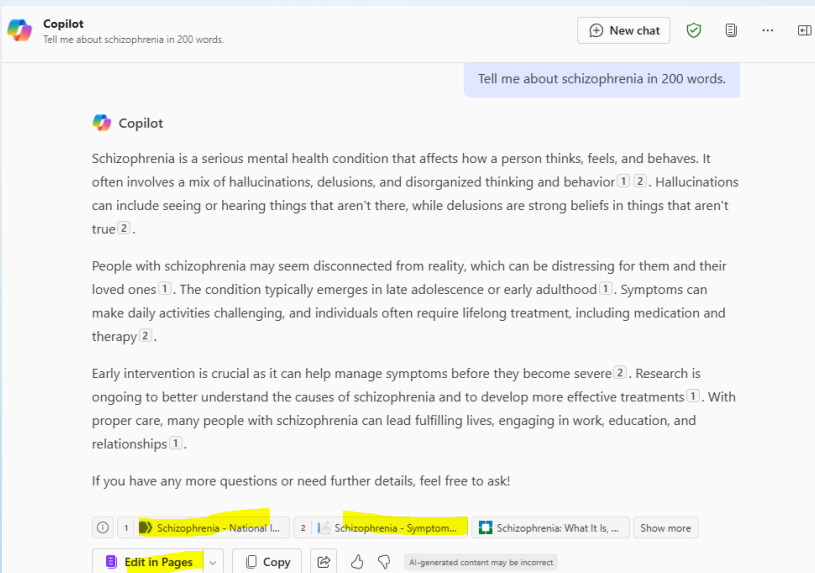
- Answers from the prompt
- Links for the prompt

Upload the files for the analysis





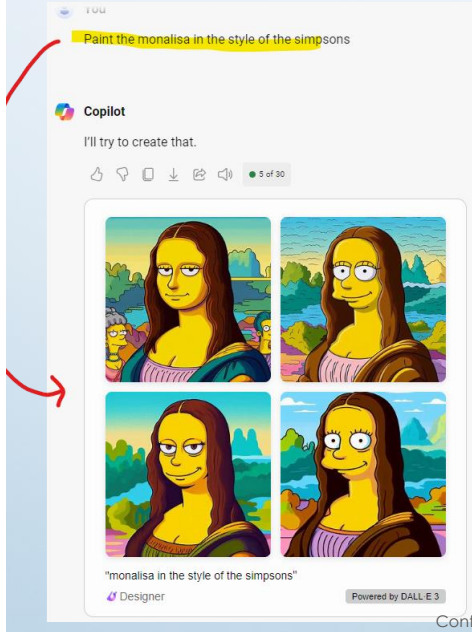
Results from prompt

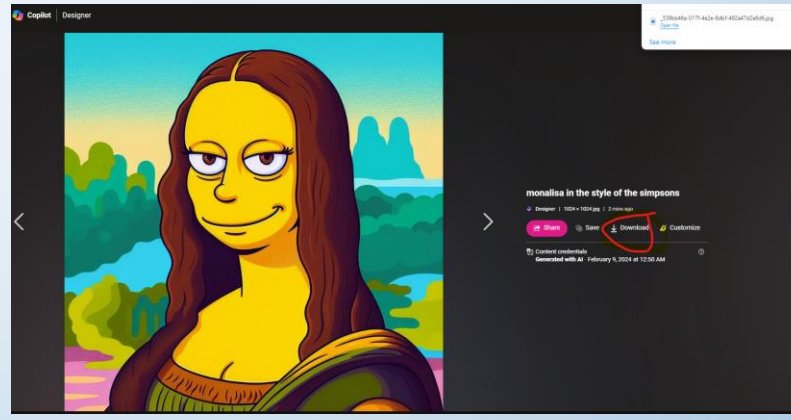


- Answers from the prompt
- Links for the prompt

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

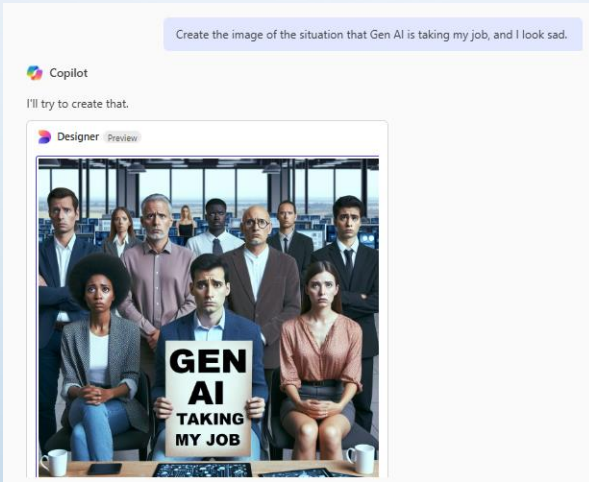
Create the image

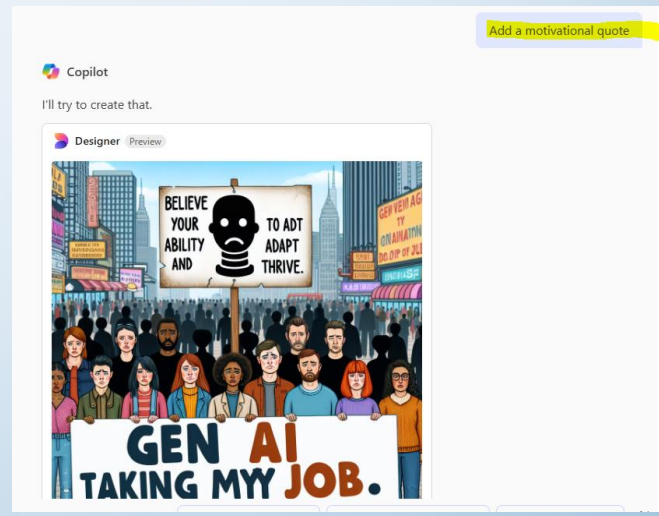




- Create the image with DALL-E 3

Create the image (2)





- Create the image with DALL-E 3

Summarize the article in link

You

Summarize the article in https://www.fda.gov/news-events/fda-voices/fda-releases-two-discussion-papers-spur-conversation-about-artificial-intelligence-and-machine with two paragraphs.

Copilot

The U.S. Food and Drug Administration (FDA) has recently released two discussion papers to foster dialogue about the integration of **artificial intelligence (AI)** and **machine learning (ML)** in drug development and manufacturing. These papers aim to engage stakeholders, including pharmaceutical companies, ethicists, academia, patients, and global regulatory authorities, in a thoughtful exploration of how AI/ML can revolutionize the field.

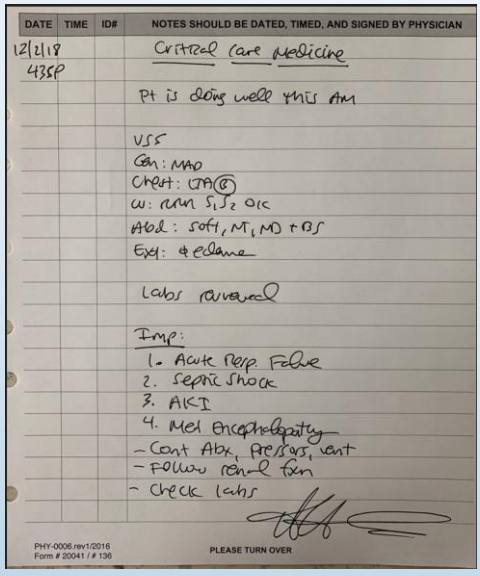
The first paper, titled "Using Artificial Intelligence and Machine Learning in the Development of Drug and Biological Products," highlights the potential of AI/ML to accelerate the delivery of safe and effective treatments. For instance, AI/ML could analyze medical literature to identify relevant findings, predict individual responses to therapies, and even facilitate patient engagement through conversational agents. However, challenges such as ethical considerations, data security, and algorithmic transparency must be addressed to harness AI/ML's full potential. The second paper, "Artificial Intelligence in Drug Manufacturing," delves into how AI/ML can enhance drug production processes. By fostering informed discussions, the FDA aims to promote equitable and responsible adoption of these transformative technologies [1] [2].

Learn more [1] fda.gov [2] bing.com [3] fda.gov [4] biospace.com [5] prnewswire.com [6] gettyimages.com

1 of 30

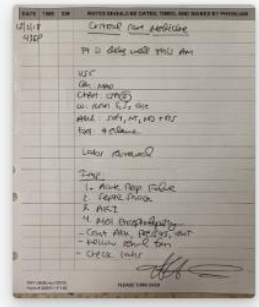
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Convert the image to the text



You

convert the image to text



Copilot

Certainly! Here's the text extracted from the image:

```
12/21/18
435P
Critical Care Medicine
Pt is doing well this am
VSS
Gen: MAO
Chest: LTG
W: cm 5'2 91c
Abd: soft, NT, ND +BS
Eyes: +edema
Labs reviewed

Imp:
1. Acute Rep Failure
2. Septic Shock
3. AKI
4. Met Encephalopathy

Cont ABX per resp/vent
Follow renal fxn
Check labs
```

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Create the spreadsheet and export it to excel

convert this in table format.

ORDER	DOMAIN	OBSERVATION CLASS	VARIABLE NAME	VARIABLE LABEL	DATA TYPE	LENGTH	ROLE
1	DM	Special-Purpose	STUDYID	Study Identifier	Text	21	Identifier
2	DM	Special-Purpose	DOMAIN	Domain Abbreviation	Text	8	Identifier
3	DM	Special-Purpose	USUBJID	Unique Subject Identifier	Text	30	Identifier
4	DM	Special-Purpose	SUBJID	Subject Identifier for the Study	Text	8	Topic
5	DM	Special-Purpose	RFSTDTC	Subject Reference Start Date/Time	Datetime	19	Record Qualifier
6	DM	Special-Purpose	RFENDTC	Subject Reference End Date/Time	Datetime	19	Record Qualifier

ORDER	OBSERVATION CLASS	VARIABLE NAME	VARIABLE LABEL
1	Special-Purpose	STUDYID	Study Identifier
2	Special-Purpose	DOMAIN	Domain Abbreviation
3	Special-Purpose	USUBJID	Unique Subject Identi
4	DM	SUBJID	Subject Identifier for t Study
5	DM	RFSTDTC	Subject Reference Sta Date/Time
6	DM	RFENDTC	Subject Reference En Date/Time
7	DM	RFXSTDTC	Date/Time of First Stu Treatment
8	DM	RFPENDTC	Date/Time of Last Stu Treatment
9	DM	DTHDTC	Date/Time of Death
10	DM	DTHFL	Death Flag
11	DM	SITEID	Study Site Identifier

Is there anything else you need help with?

Sure! Here's the information from the image converted into a table format:

Export to Excel
Copy page link to here
Move block
Delete

		OBSERVATI...	VARIABLE ...
		ecial-Purpose	STUDYID
		ecial-Purpose	DOMAIN
3	3	Special-Purpose	USUBJID
4	4	DM	SUBJID
5	5	DM	RFSTDTC

Use the editor

The table is metadata for CDISC DM. I want to add the sample data. Can you create the sample data?

Copilot

Got it! Here's a sample dataset based on the CDISC DM metadata:

STUDYID	DOMAIN	USUBJID	SUBJID
STUDY001	DM	USUBJ001	SUBJ001
STUDY002	DM	USUBJ002	SUBJ002
STUDY003	DM	USUBJ003	SUBJ003

Feel free to adjust the sample data as needed. Is there anything else you need help with?

Add to page | Copy | AI-generated content may be incorrect

Can you explain the significance of each field? | Add more rows to this sample data.

sk me anything...

convert this in table format

Just start typing...

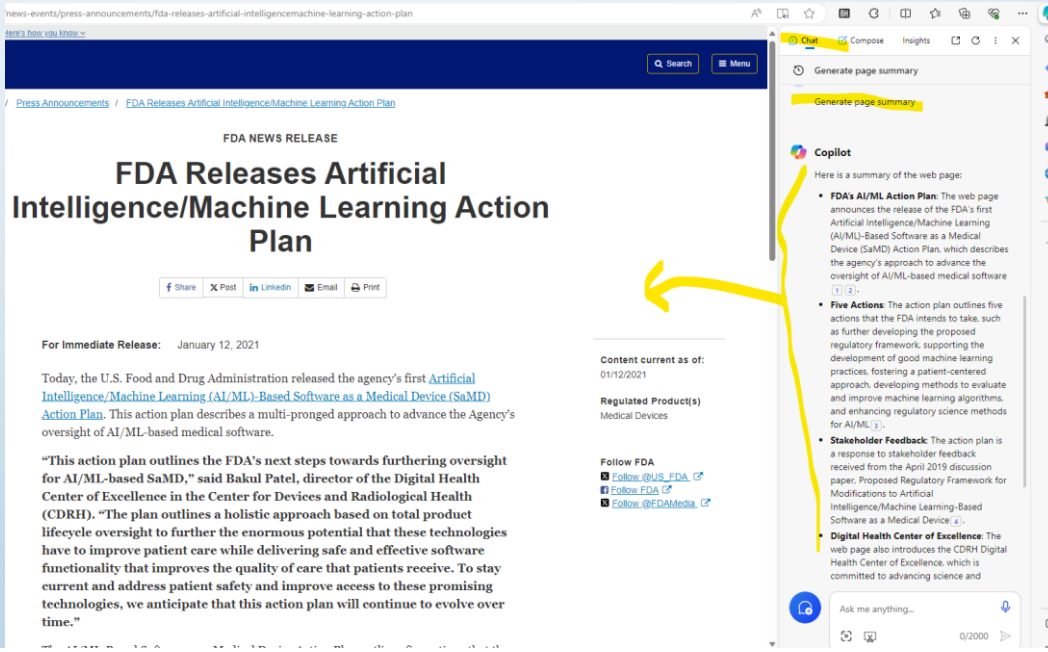
Got it! Here's a sample dataset based on the CDISC DM metadata:

	STUDYID	DOMAIN	USUBJID
1	STUDY001	DM	USUBJ001
2	STUDY002	DM	USUBJ002
3	STUDY003	DM	USUBJ003

New

Feel free to adjust the sample data as needed. Is there anything else you need help with?

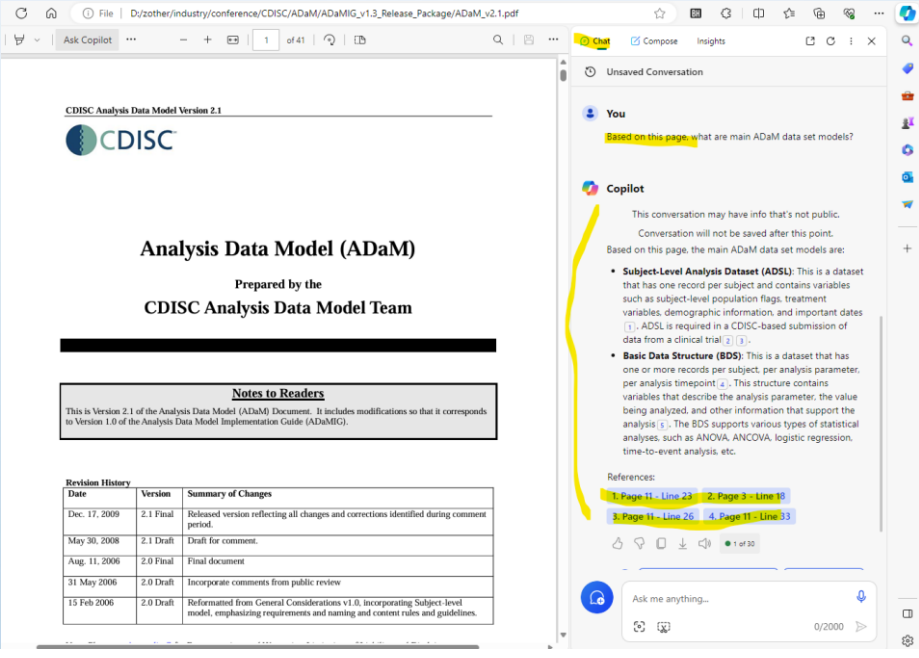
Prompt for search results



- Your search in Bing Search
- Prompt (e.g. summary) for your search results
- In Copilot

137

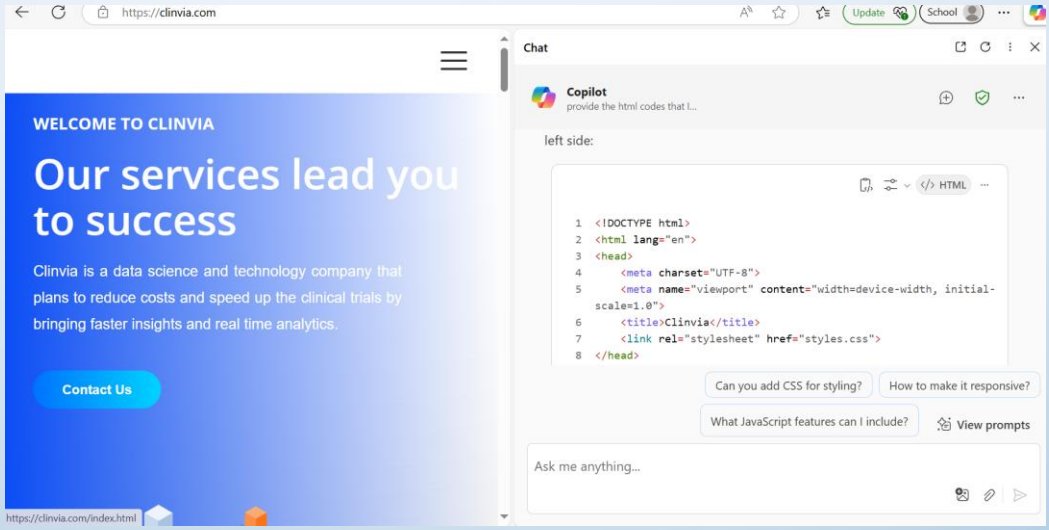
Upload document in Edge and prompt in Copilot



- Upload/ Open document (e.g pdf) in Edge
- Prompt and get the answers from the document

138

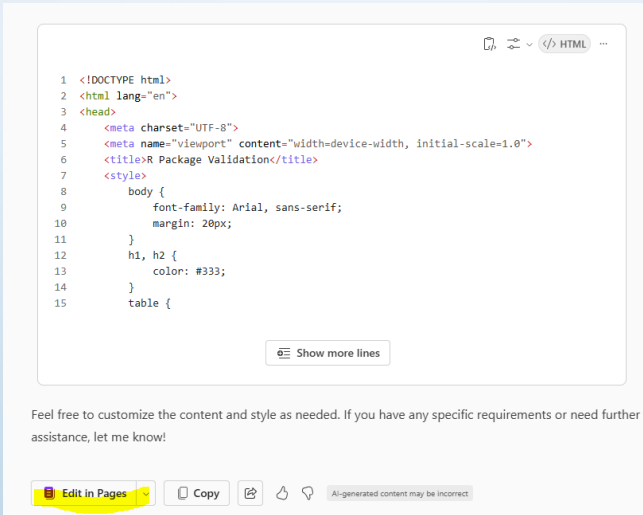
Find source codes in the website



- We could find html source codes for the web-page.

139

Find source codes in the website



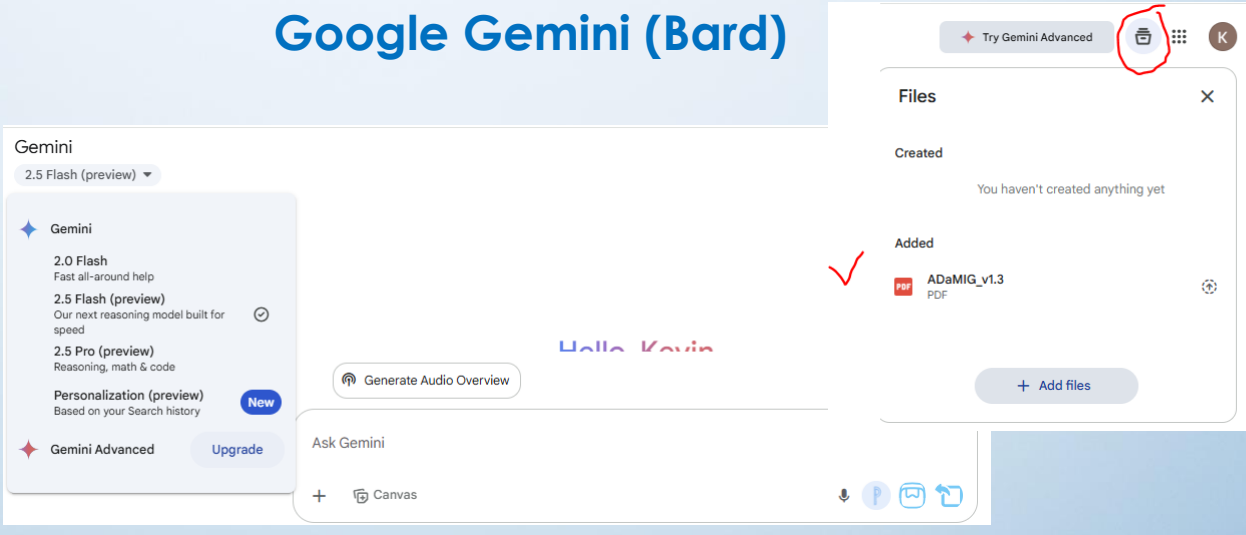
- We could find html source codes for the web-page.

140

Google Gemini Functionality

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Google Gemini (Bard)



- https://gemini.google.com/app
- You could sign in with current google account.

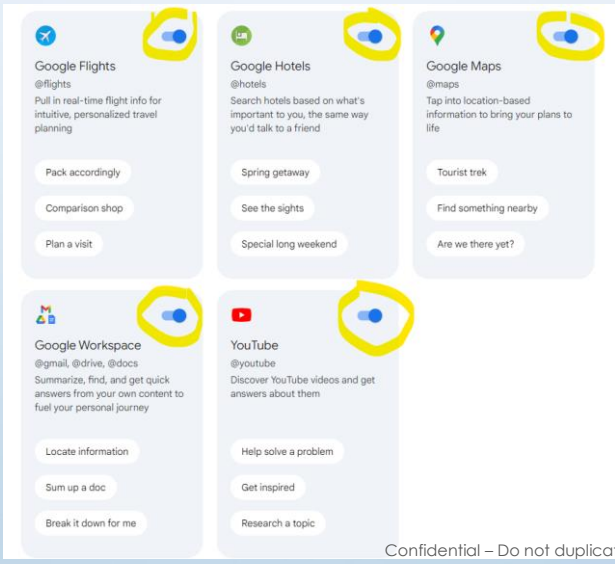
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Strengths in Gemini

1. Can connect to other Google tools.
2. Faster than other Gen AI
3. Usually free

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

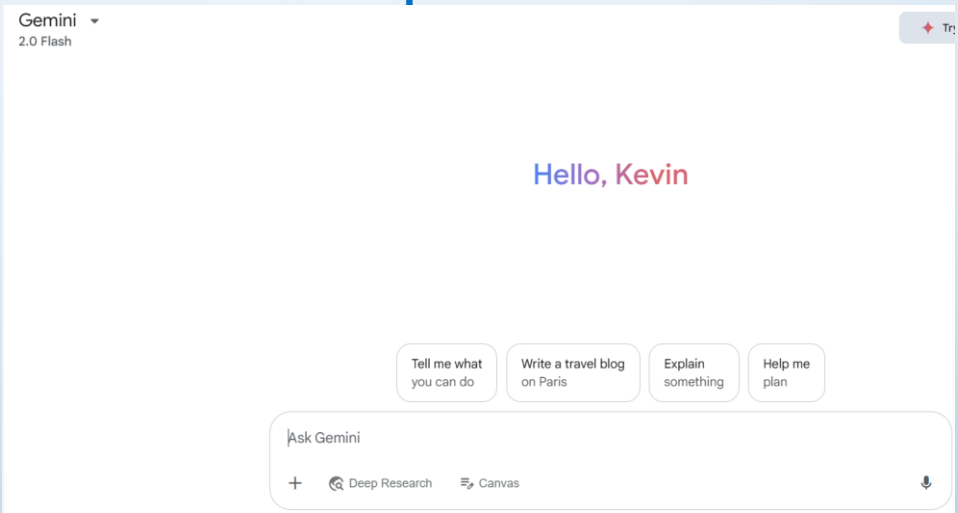
Add extensions in gemini.google.com/extensions



- Pick five extensions to add apps functionality.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

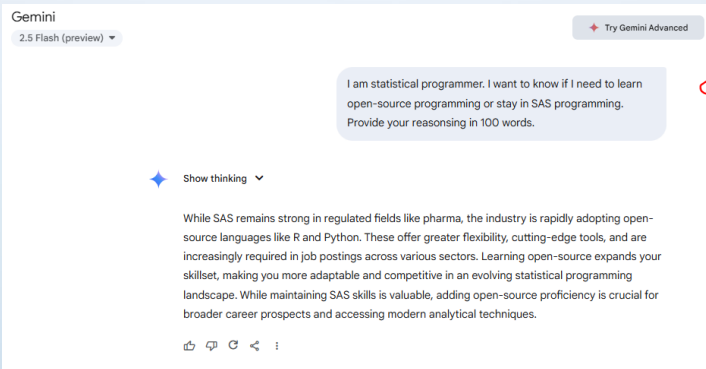
Prompt – 2.0 Flash

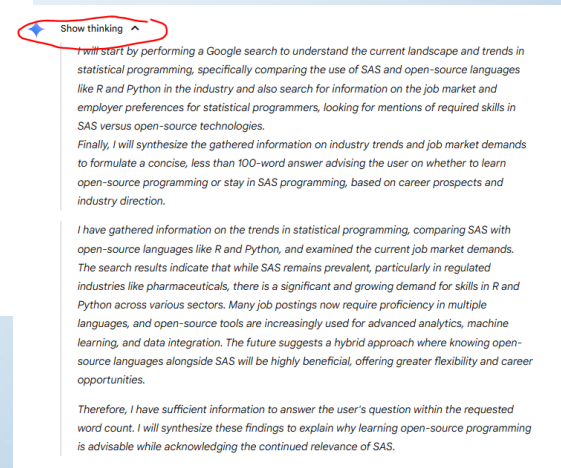


- You can search sample prompts or simply ask Gemini.
- Click the sample, and it will give you prompt example

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompts – 2.5 Flash (Reasoning)





We can get the answer.

We can also get the reasoning behind.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt – enhancing prompt

As a statistical programmer, evaluate the necessity of transitioning from SAS programming to open-source programming. Consider factors such as:
- Industry demand and trends
- Career advancement opportunities
- Cross-industry compatibility and adaptability
- Learning curve and resource availability
- Potential impact on current and future projects

+ Canvas

- We can edit plan or start research
- "Deep Search" take some time – 2 to 3 minutes.
- Prompt:

I am statistical programmer. I want to know if I need to learn open-source programming or stay in SAS programming. Provide your reasoning in 100 words. Explain your reasoning step by step.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Prompt – Deep Research in "2.0 Flash"

I am statistical programmer. I want to know if I need to learn open-source programming or stay in SAS programming. Provide your reasoning in 100 words. Explain your reasoning step by step.

+ Deep Research Canvas

I've put together a research plan. If you need to update it in any way, let me know.

SAS vs. Open-Source Skills

Research Websites

(1) Search for current job market trends for statistical programmers in Maple Glen, Pennsylvania and the broader United States, specifically comparing the demand for SAS versus open-source skills like Python and R.
(2) Investigate the adoption rate and growth of open-source tools in statistical programming across various industries.
(3) Compare the capabilities, flexibility, cost-effectiveness, and community support of SAS and...

More

Analyze Results

Create Report

Ready in a few mins

Edit plan Start research

- We can edit plan or start research
- "Deep Search" take some time – 2 to 3 minutes.
- Prompt:

I am statistical programmer. I want to know if I need to learn open-source programming or stay in SAS programming. Provide your reasoning in 100 words. Explain your reasoning step by step.

ibute without the permission of Kevin Lee

Prompt – Deep Research (thinking)

Gemini
Deep Research

Try Gemini Advanced

Start research

I'm on it. I'll let you know when your research is done. In the meantime, you can leave this chat.

SAS vs. Open-Source Skills
Researching 104 websites...

Ask Gemini
Deep Research

SAS vs. Open-Source Skills | Show thinking

Initial Research Focus

I'm starting my research by looking into the current job market to see how strong the demand is for both SAS and open-source programming skills in the field of statistical programming. This will help me understand the immediate relevance of learning new tools.

Researching websites

handsonsystem.com Clini... | ziprecruiter.com $34-$76/... | mindmajix.com Python vs ... | ziprecruiter.com The 9 Hi...
sas.com SAS 2025 predict... | m.economicpoint.com Pyt... | datasciencecentral.com P... | mobilunity.com Hire SAS ...
index.dev Programming L... | communities.sas.com SA... | optiontrain.com Demand ... | r4stats.com The Populari...
reddit.com [C] Does the c... | edvancer.in R, Python or S... | reddit.com Is it worth it to... | bacancytechnology.com P...
communities.sas.com Wh... | scholar.smu.edu Data Sci... | bomberbot.com R vs Pyth... | reddit.com [D] How popul...
reddit.com Should I learn ... | syntelli.com Python & R v... | datacamp.com Top 12 Pro... | reddit.com SAS vs Python...
coursera.org 7 In-Demand... | coursera.org Python or R f... | youtube.com Comparing ... | reddit.com What are some...
stxnext.com R vs. Python: ...

Job Market Insights

I've found that SAS is still a key player, especially in big organizations like those in finance and pharmaceuticals, often due to existing systems and regulations. There are definitely jobs out there for SAS programmers, and they can pay well. However, it seems fewer new professionals are learning SAS. the other hand, Python is becoming increasingly popular and is in high demand for data science, machine learning, and even web development. It's also considered easier to learn, which is a plus.

Open Source Growth

Prompt – Deep Research (final output)

Gemini
Deep Research

Try Gemini Advanced

Start research

I've completed your research. Feel free to ask me follow-up questions or request changes.

SAS vs. Open-Source Skills
Apr 14, 12:32 PM

Ask Gemini
Deep Research

SAS vs. Open-Source Skills | Export to Docs

Navigating the Landscape of Statistical Programming: SAS and the Rise of Open-Source Alternatives

The field of statistical programming is dynamic, with continuous shifts in technology and industry demands. For professionals skilled in SAS, a long-established and powerful tool, the question of whether to embrace open-source programming languages like Python and R is increasingly pertinent. This report examines the current job market trends, the capabilities of these languages, their cost implications, flexibility, integration potential, learning curves, and the impact on career growth to provide a reasoned perspective on this critical decision.

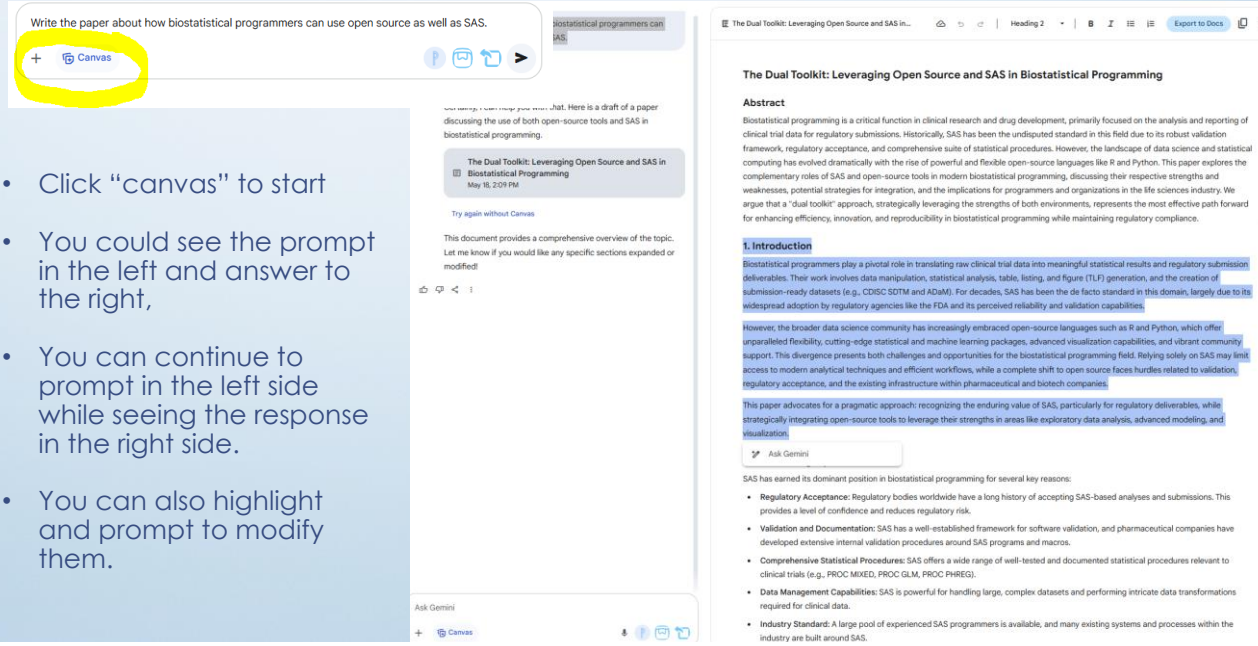
I. Analyzing the Current Job Market for Statistical Programmers

An examination of the present job market reveals a complex landscape where SAS maintains a significant foothold while open-source languages are experiencing rapid growth. Despite the increasing adoption of open-source alternatives, SAS skills remain in considerable demand. This is evidenced by its strong ranking in the "Jobs" category of the IEEE Spectrum 2024 list. Furthermore, numerous job postings on platforms such as Indeed.com consistently seek professionals with SAS expertise. This sustained demand is particularly notable in specialized areas like clinical SAS programming, driven by ongoing developments in pharmaceutical research and clinical trials. SAS's long-standing reputation as a global leader in data science, known for its extensive statistical functions and reliable performance, contributes to its continued prevalence in many organizations. Large corporations often favor SAS due to the robust customer service and technical support it offers, which are crucial in enterprise settings where operational reliability is paramount. This indicates that despite the growing popularity of open-source languages, SAS maintains a strong and active presence in the current job market, especially within established and regulated industries with substantial investments in SAS infrastructure and validated processes. Professionals with specialized SAS knowledge in fields like clinical trials or financial modeling may continue to find numerous opportunities.

Conversely, the job market is also witnessing a significant surge in demand for open-source programming skills, particularly in Python and R. Surveys and job market analyses highlight a tremendous increase in job openings requiring these skills, signifying a strong competitive

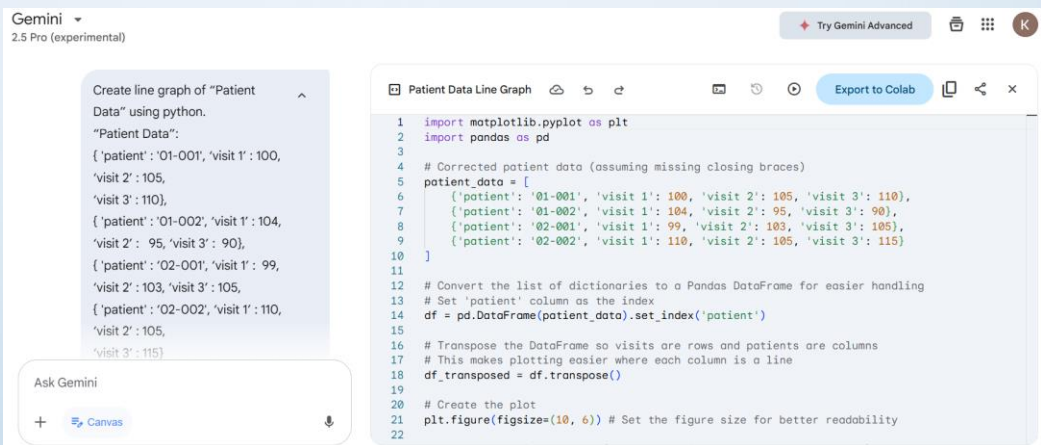
Reasoning

Using Canvas (writing)



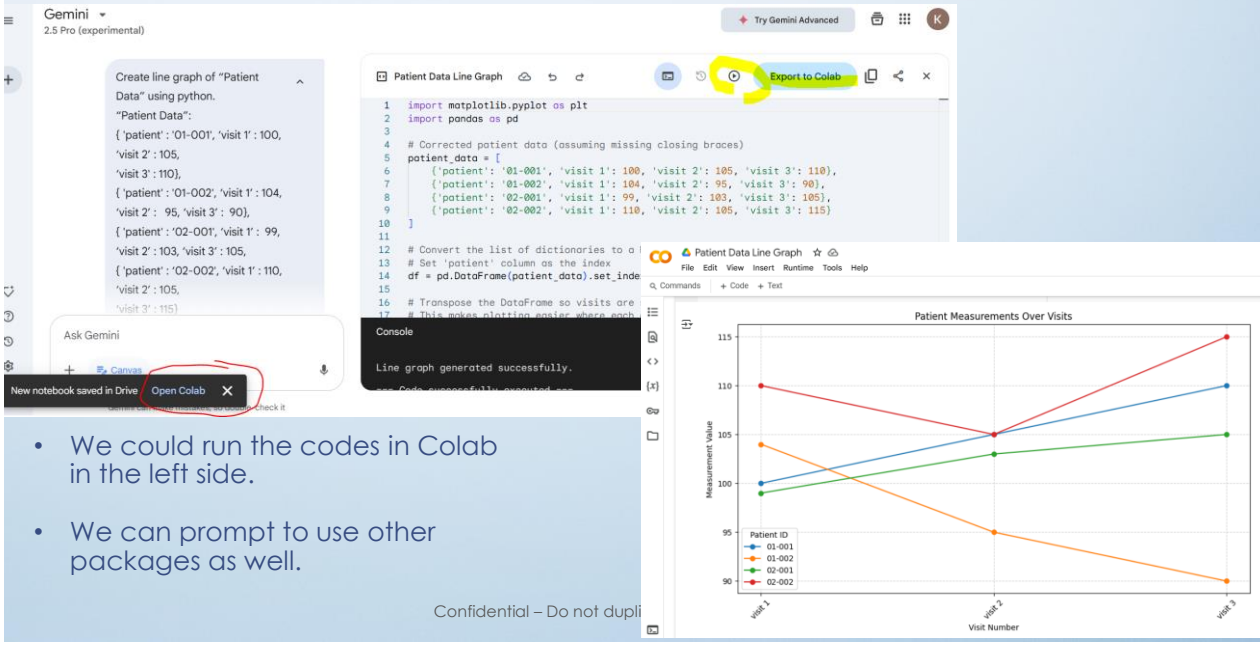
- Click "canvas" to start
- You could see the prompt in the left and answer to the right,
- You can continue to prompt in the left side while seeing the response in the right side.
- You can also highlight and prompt to modify them.

Using Canvas (coding)



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

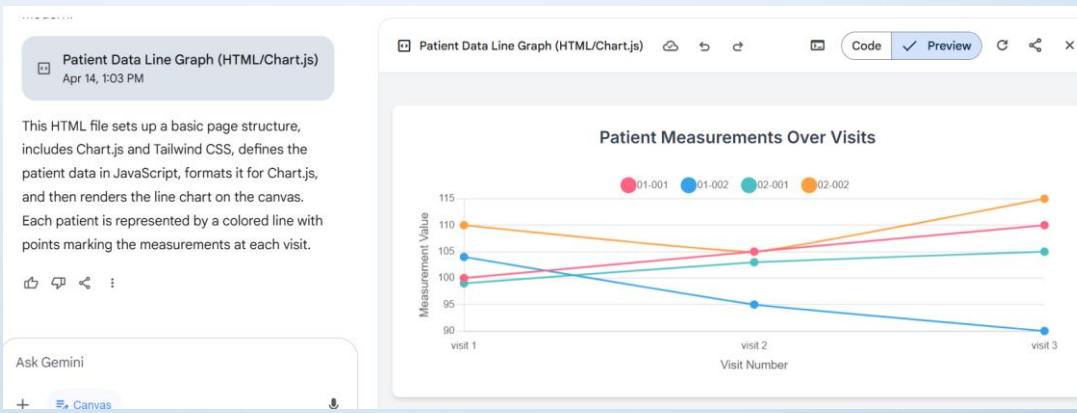
Using Canvas (coding)



- We could run the codes in Colab in the left side.
- We can prompt to use other packages as well.

Confidential – Do not dupli

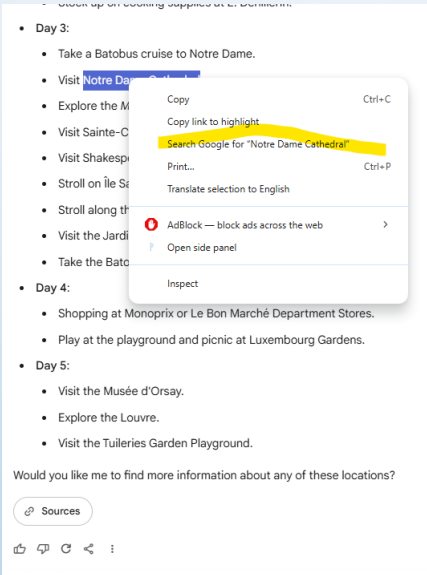
Using Canvas (coding)

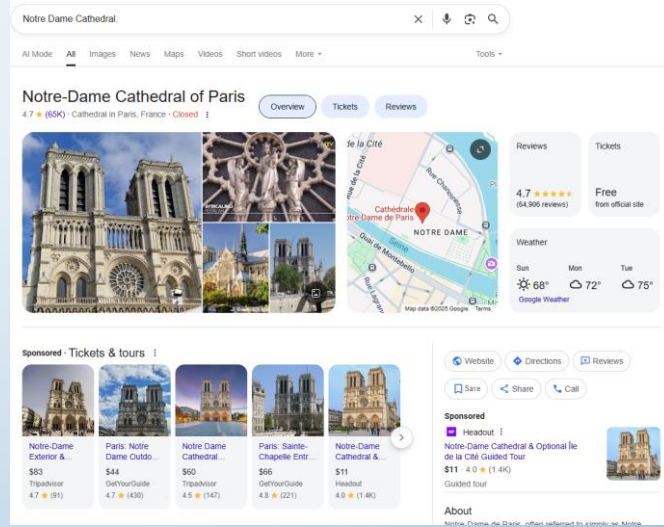


- We could also create it in html and JS.
- We could view actual coding in "code" tab.

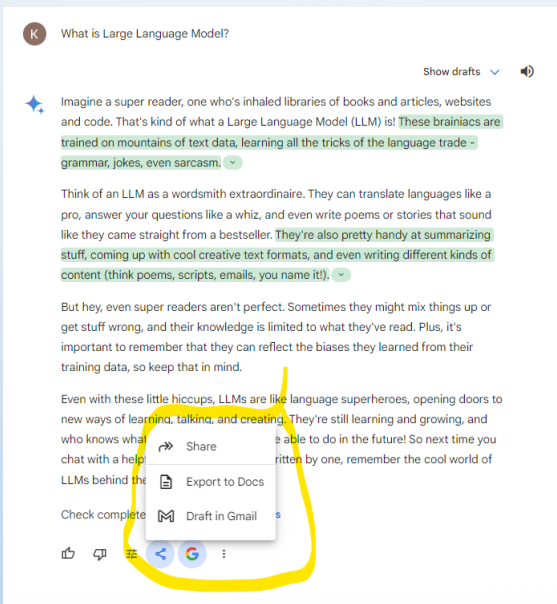
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Response Search in Google

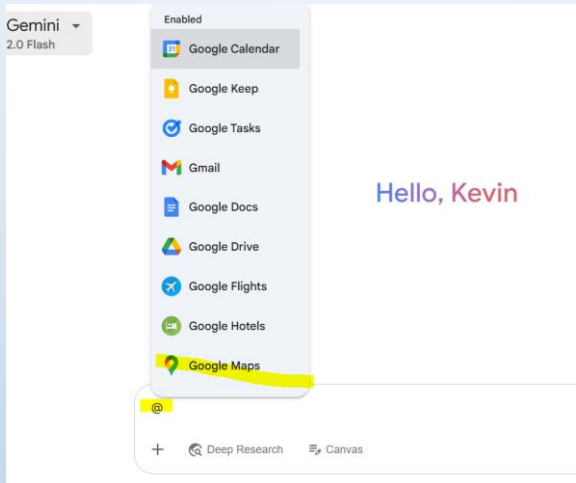


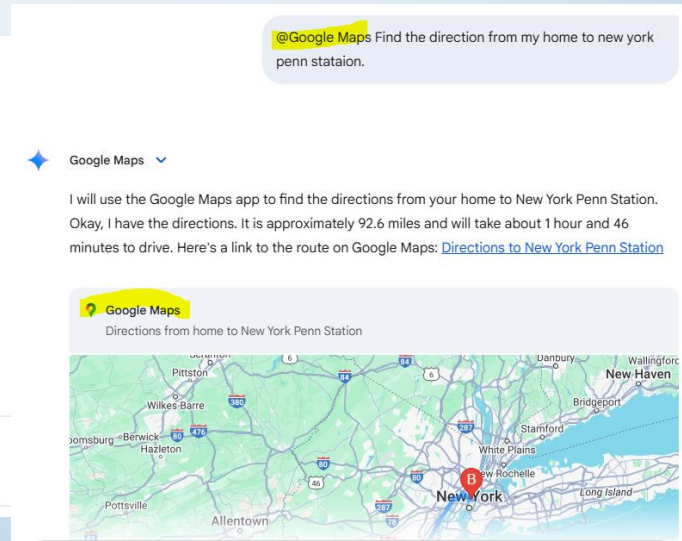


Share



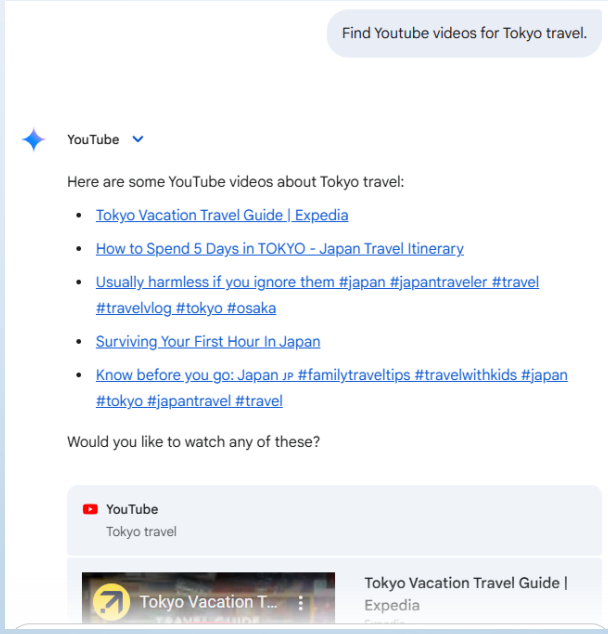
Better Search Google tools (AI Agent)

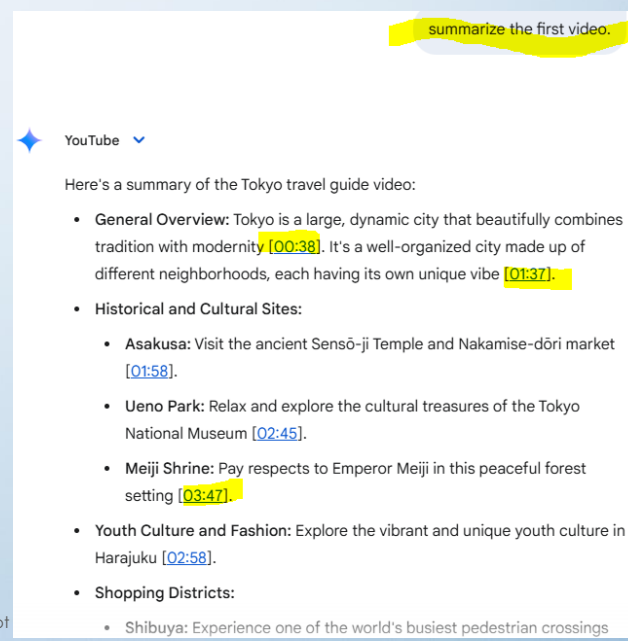




Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Better Search with YouTube extension





Use extension for the prompt – AI agent

K I will visit Boston from 2024/04/23 to 2024/04/26. I need to attend conference in Boston Convention center. Find the hotel near by.

Show drafts

Google Maps & 1 more

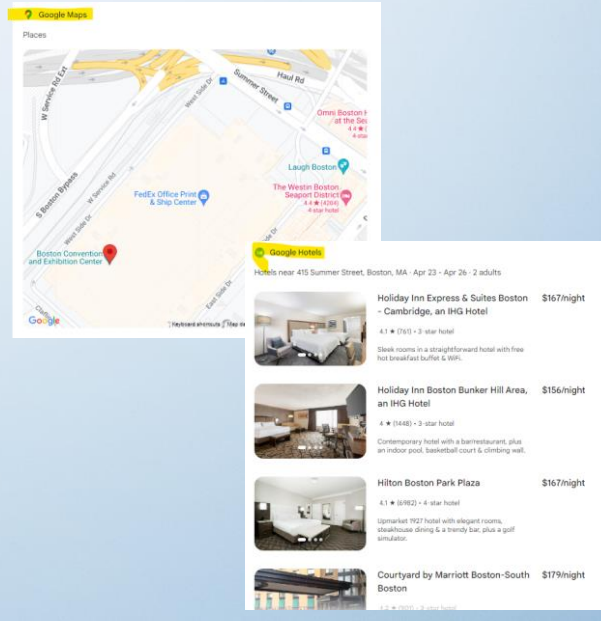
Gemini doesn't always get it right. Be sure to check the following hotel details:

Hotels near 415 Summer Street, Boston, MA · Apr 23 - Apr 26 · 2 adults

Sure, here are some hotels available near the Boston Convention Center between April 23rd and April 26th:

- Holiday Inn Express & Suites Boston - Cambridge, an IHG Hotel - Sleek rooms in a straightforward hotel with free hot breakfast buffet & WiFi. Price: $167 per night.
- Holiday Inn Boston Bunker Hill Area, an IHG Hotel - Contemporary hotel with a bar/restaurant, plus an indoor pool, basketball court & climbing wall. Price: $156 per night.
- Hilton Boston Park Plaza - Upmarket 1927 hotel with elegant rooms, steakhouse dining & a trendy bar, plus a golf simulator. Price: $167 per night.
- Courtyard by Marriott Boston-South Boston - Modern hotel with a restaurant & an outdoor courtyard, plus free Wi-Fi. Price: $179 per night.
- Courtyard by Marriott Boston Cambridge - Modern riverside hotel with 2 restaurants & a ballroom, plus free Wi-Fi. Price: $171 per night.

I hope this helps with planning your trip to Boston!



Connect to my Google Drive thru Google Workspace extension

K what is the latest documents that I uploaded in my google doc?

Show drafts

Google Workspace

Assuming by "latest" you mean the recency of creation date, then the latest documents you uploaded to your Google Drive are:

- Outline for Clinical Data Management with Python, created on July 7, 2023.
- Strategic Implementation of Open Source Software in Clinical Trials, created on May 8, 2021.
- cs329s_03_note_data_engineering, created on January 19, 2021.
- RESUME, created on May 16, 2012.
- hero.pdf, created on June 8, 2014.

I hope this helps!

Google Drive Items considered for this response. Learn more

Outline for Clinical Data Management with Python Nilesh Kowadkar

K summary "Outline for Clinical Data Management with Python" from Google Drive in one sentence.

Show drafts

Google Workspace

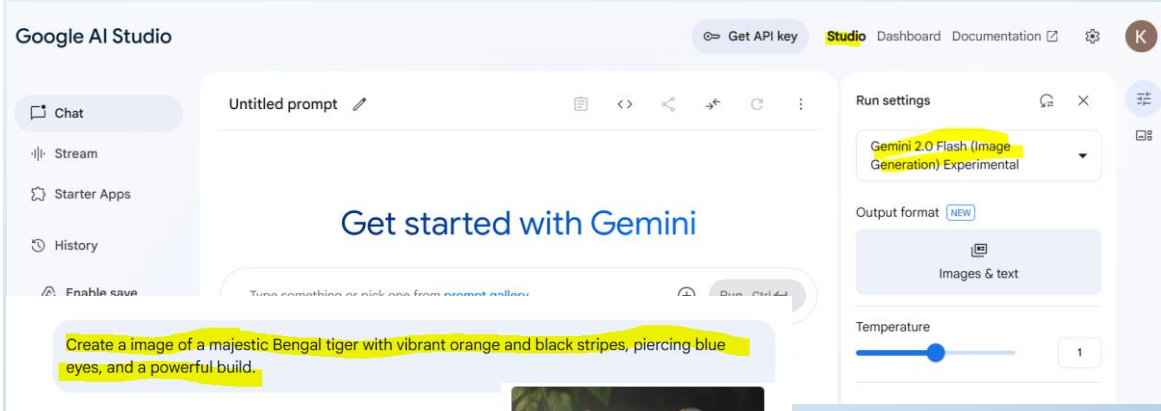
"Outline for Clinical Data Management with Python" provides a comprehensive guide to using Python for clinical trial data management and analysis, covering topics from fundamental concepts to advanced analytics and regulatory considerations.

Google Drive Items considered for this response. Learn more

Outline for Clinical Data Management with Python Nilesh Kowadkar

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Google AI Studio – image

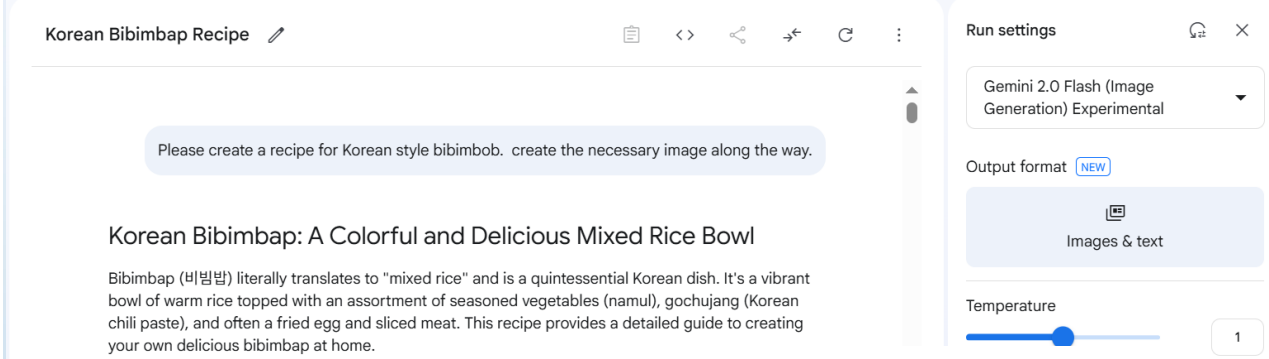


Create a image of a majestic Bengal tiger with vibrant orange and black stripes, piercing blue eyes, and a powerful build.

- https://aistudio.google.com/prompts/new_chat
- Much faster than ChatGPT in creating the images

thout the permission of Kevin Lee

Google AI Studio – contents with image



Please create a recipe for Korean style bibimbob. create the necessary image along the way.

Korean Bibimbap: A Colorful and Delicious Mixed Rice Bowl

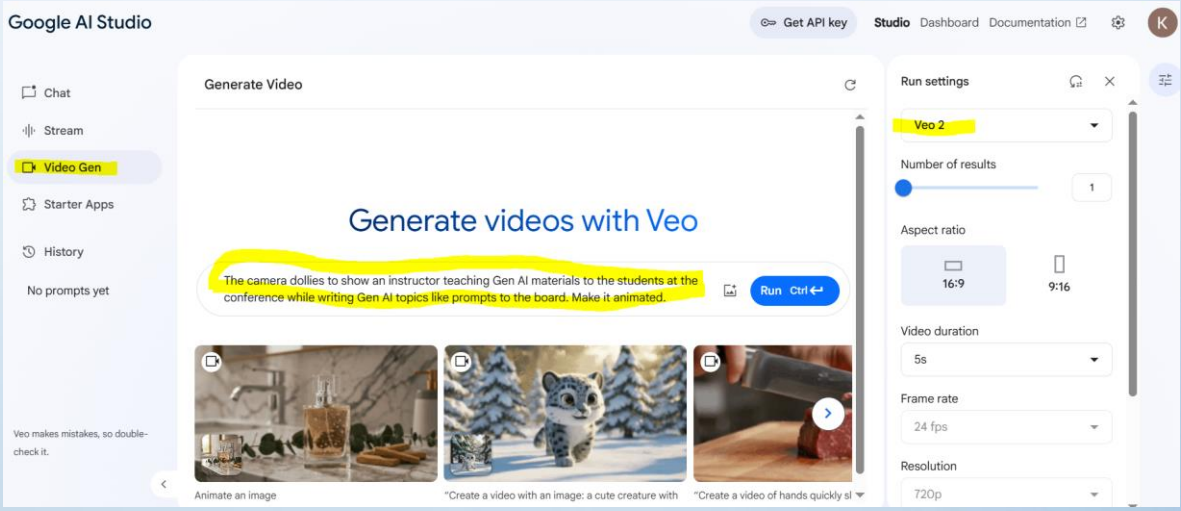
Bibimbap (비빔밥) literally translates to "mixed rice" and is a quintessential Korean dish. It's a vibrant bowl of warm rice topped with an assortment of seasoned vegetables (namul), gochujang (Korean chili paste), and often a fried egg and sliced meat. This recipe provides a detailed guide to creating your own delicious bibimbap at home.

Korean Bibimbap Recipe

Image 1: Raw short-grain white rice in a measuring cup.

Image 13: A person using chopsticks to mix all the ingredients of a bibimbap bowl together.

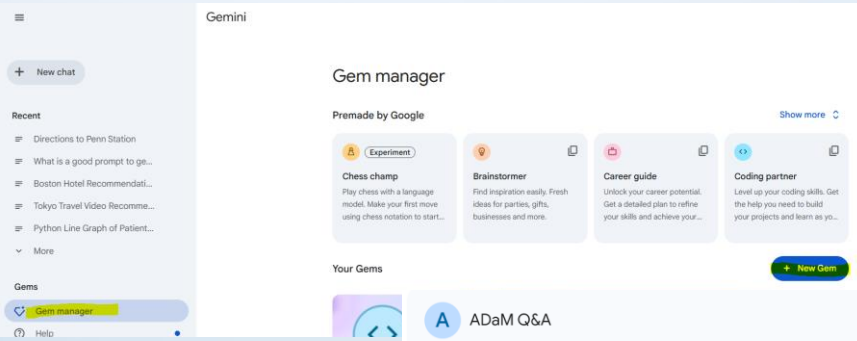
Google AI Studio – Creating video



- There is a limit (5 times a day)

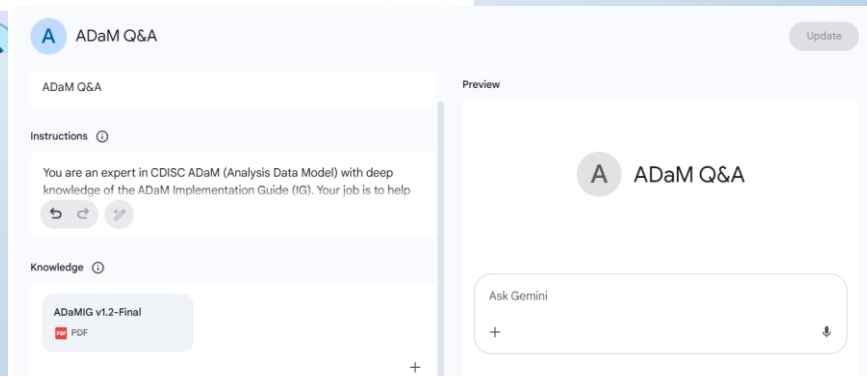
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Gems (GPTs)



components of Gems

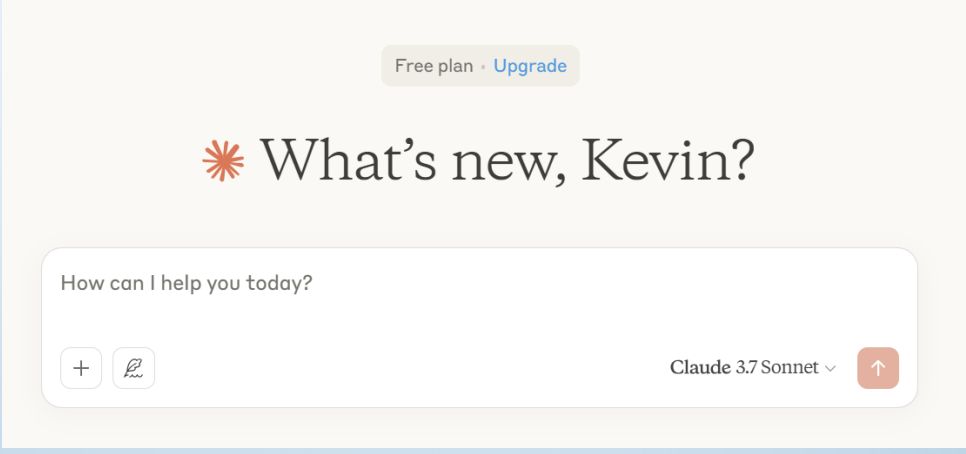
- Title
- Instructions
- reference files



Claude Sonnet Functionality

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

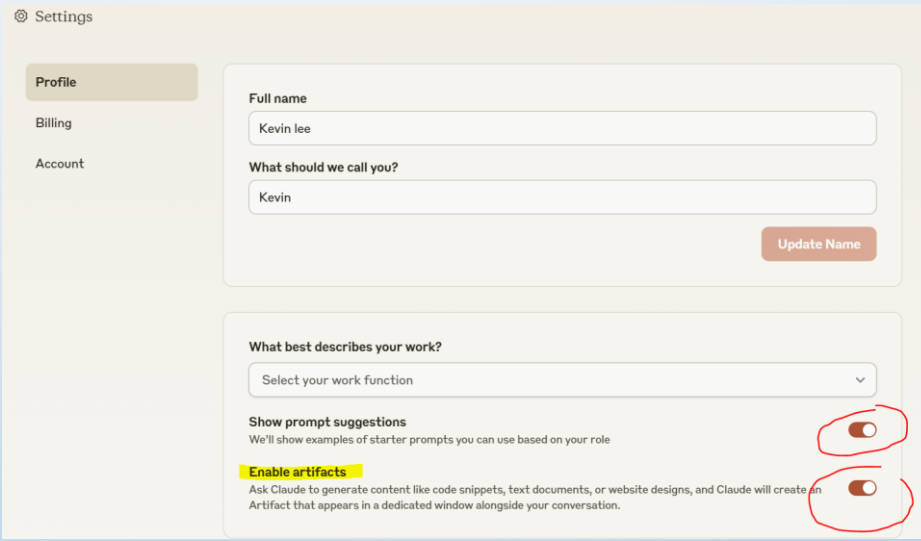
Claude Sonnet



- https://claude.ai/new
- You could sign in with current google account / new userid and pw.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

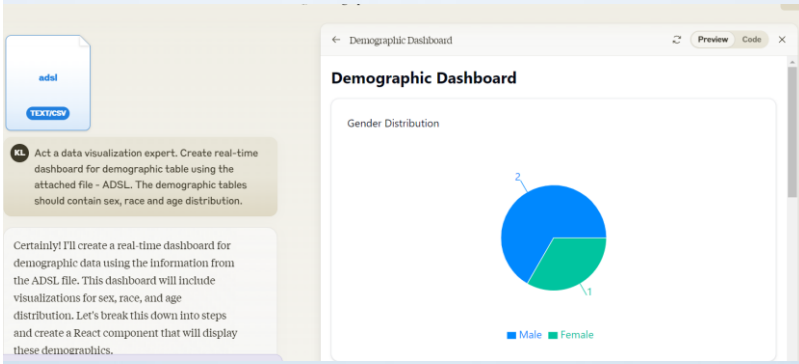
Set up artifact



- Click the setting
- Enable artifacts
- Claude will automatically generate Artifacts for specific content types like code or graphics, displaying them in a dedicated window next to the chat.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Data Analysis



1. Attach the file - adsl
2. Prompt : Act a data visualization expert. Create real-time dashboard for demographic table using the attached file - ADSL. The demographic tables should contain sex, race and age distribution.
3. It will create
   1. Answer.
   2. Artifacts – Preview and Codes

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Data Analysis - Answer

Certainly! I'll create a real-time dashboard for demographic data using the information from the ADSL file. This dashboard will include visualizations for sex, race, and age distribution. Let's break this down into steps and create a React component that will display these demographics.

I've created a React component that generates a real-time dashboard for demographic data based on the ADSL file. Here's a breakdown of the dashboard:

1. Gender Distribution: A pie chart showing the distribution of males and females in the study.
2. Race Distribution: A pie chart displaying the racial breakdown of participants.
3. Age Group Distribution: A bar chart illustrating the distribution of participants across different age groups.

The dashboard is designed to update every 5 seconds, simulating real-time data updates. In a real-world scenario, you would replace the simulated data fetch with an actual API call to get the latest data.

Key features of this dashboard:

1. Responsive design: The charts will resize based on the screen size.
2. Real-time updates: The data refreshes every 5 seconds.
3. Interactive charts: Users can hover over chart elements to see detailed information.
4. Clear visualization: Each demographic aspect (gender, race, age) is presented in its own card for easy comprehension.

To use this dashboard:

1. Make sure you have the necessary dependencies installed (react, recharts, and the shadcn/ui components).
2. Replace the simulated data in the `fetchData` function with your actual data fetching logic.
3. Adjust the styling as needed to fit your application's design.

This dashboard provides a clear and interactive way to visualize the demographic data from your ADSL file. Users can quickly understand the distribution of participants across different demographic categories.

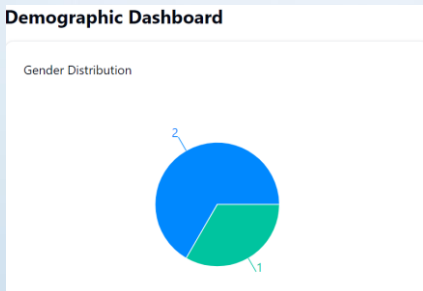
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

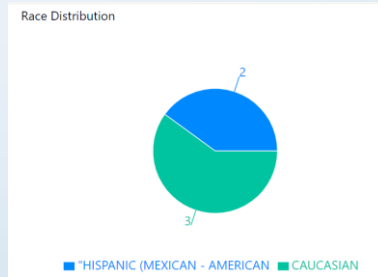
Data Analysis – Artifacts (code)

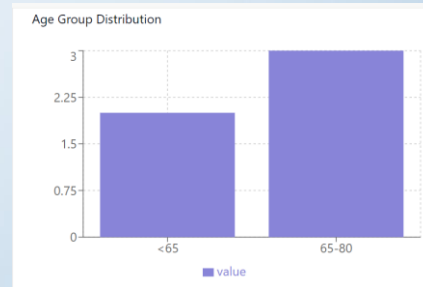
```
import React, { useState, useEffect } from 'react';
import { PieChart, Pie, Cell, BarChart, Bar, XAxis, YAxis, CartesianGrid, Tooltip, Legend, ResponsiveContainer } from 'recharts';
import { Card, CardHeader, CardContent } from '@/components/ui/card';

const COLORS = ['#0088FE', '#00C49F', '#FFBB28', '#FF8042', '#8884D8'];

const DemographicDashboard = () => {
  const [data, setData] = useState([]);

  useEffect(() => {
    // Simulating data fetch
    const fetchData = () => {
      // In a real scenario, you'd fetch this data from an API
      const rawData = `USUBJID,DISCONT,DSREASCD,DEATH,TRTP,TRTPCD,AGE,AGEGRP,RACE,SEX
01-701-1015,,Completed,,Placebo,Pbo,63,<65,"HISPANIC (MEXICAN - AMERICAN, MEXICO, CENTRAL AND SOUTH AMERICA)",F
01-701-1023,Y,Adverse Event,,Placebo,Pbo,64,<65,"HISPANIC (MEXICAN - AMERICAN, MEXICO, CENTRAL AND SOUTH AMERICA)",M
01-701-1028,,Completed,,Xanomeline High Dose,Xan_Hi,71,65-80,CAUCASIAN,M
01-701-1033,Y,Sponsor Decision,,Xanomeline Low Dose,Xan_Lo,74,65-80,CAUCASIAN,M
01-701-1034,,Completed,,Xanomeline High Dose,Xan_Hi,77,65-80,CAUCASIAN,F`;

      const parsedData = rawData.split('\n').slice(1).map(row => {
        const [,,,,,,AGE,AGEGRP,RACE,SEX] = row.split(',');
        return { AGE: parseInt(AGE), AGEGRP, RACE, SEX };
      });

      setData(parsedData);
    };
```

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Data Analysis - Artifacts

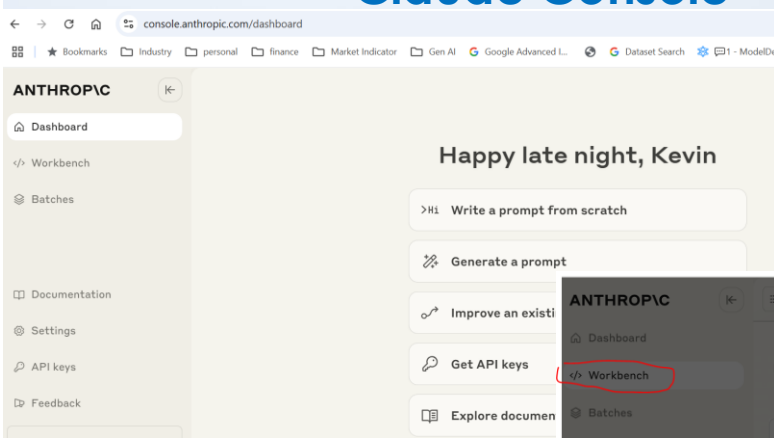






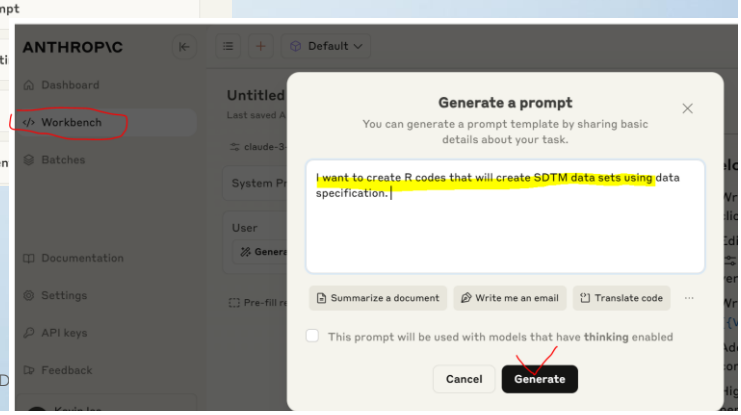
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Claude Console



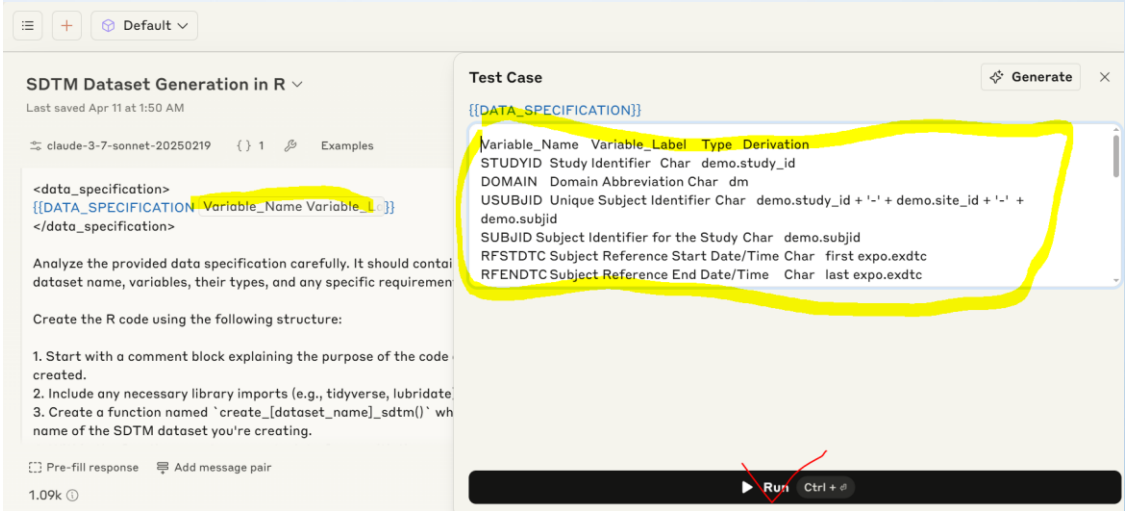
- http://console.anthropic.com
- You could sign in with current google account / new userid and pw.

- Click on "Workbench"
- Click on "Generate a Prompt"



Confidential – D

Claude Console



- We can test and update prompts very easily in the console.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Claude Console - response

You are tasked with creating R code that will generate SDTM (Study Data Tabulation Model) datasets using only a provided data specification. This task is crucial for standardizing clinical trial data. Follow these instructions carefully to create accurate and efficient R code.

You will be provided with a data specification in the following format:
<data_specification>
{{DATA_SPECIFICATION}}
</data_specification>

Analyze the provided data specification carefully. It should contain information about the dataset name, variables, their types, and any specific requirements or derivations.

Create the R code using the following structure:
1. Start with a comment block explaining the purpose of the code and the dataset being created.
2. Include any necessary library imports (e.g., tidyverse, lubridate).
3. Create a function named `create_[dataset_name]_sdtm()` where [dataset_name] is the name of the SDTM dataset you're creating.
4. Within the function, create an empty data frame with the specified variables.
5. Add code to populate the data frame based on the specifications for each variable.
6. Include error handling and input validation where appropriate.
7. Return the completed data frame at the end of the function.
When handling different variable types, use appropriate R functions and data types:
- For character variables, use `as.character()`
- For numeric variables, use `as.numeric()`
- For date variables, use appropriate date conversion functions from the lubridate package
- For derived variables, include the derivation logic as specified
Include necessary R libraries at the beginning of your code, such as:

```r
library(tidyverse)
library(lubridate)
```

Add error handling and input validation where appropriate, using tryCatch() or if-else statements. Include comments throughout your code to explain complex operations or derivations.

Present your final R code within <r_code> tags. Ensure your code is properly formatted and follows R best practices.

<r_code>
Your R code here
</r_code>

After the code, provide a brief explanation of how the code works and any assumptions made during its creation. Present this explanation within <explanation> tags.

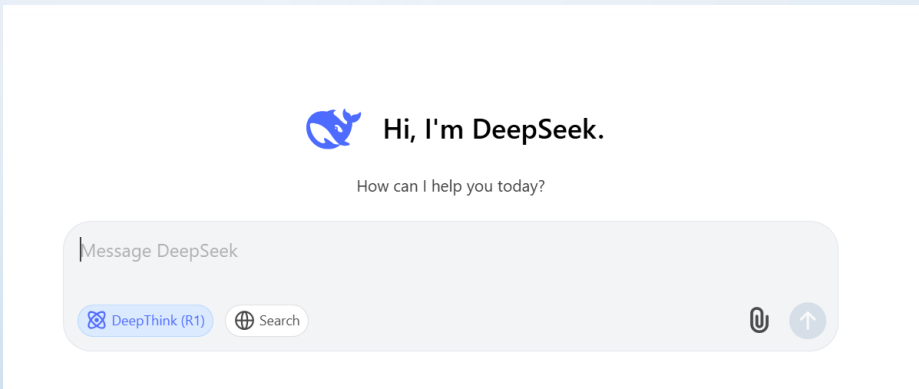
<explanation>
Your explanation here
</explanation>

Remember to adhere strictly to the provided data specification and SDTM standards when creating the R code.

DeepSeek Functionality

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

DeepSeek R1



- https://chat.deepseek.com/
- You could sign in with current google account / new userid and pw.

- Only for texts
- Free
- Long Text (128K tokens)
- Strong Performance in Coding & Math
- File Support – PDF, Word, Excel, PPT
- Good in Deep Research – Thinking long

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

DeepSeek R1

Based on the attached file, tell me what are required variables in ADSL?

DeepThink (R1) Search

Based on the attached file, tell me what are required variables in ADSL?

Edit

Thought for 26 seconds

Okay, let me try to figure out what the required variables in ADSL are based on the attached file. The user is asking specifically about ADSL, which is the Subject-Level Analysis Dataset. From what I remember from the document, ADSL has to have certain variables that are mandatory.

First, I'll start by recalling that in Section 3.2 of the document, there are tables that list ADSL variables. Table 3.2.1 talks about identifier variables. The first few variables listed there are STUDYID, USUBJID, SUBJID, and SITEID. The "Core" column for these says "Req" for required. So those must be required.

marked "Req" in the Core column and the conditional ones that are necessary based on study design.

So the final list should include the identifiers, demographics with their units, treatment variables for the first period, ARM, and at least one population flag. The date variables are required only if applicable.

Based on the **ADaMIG v1.2** document, the **required variables in ADSL** (Subject-Level Analysis Dataset) are as follows:

1. Identifier Variables (Table 3.2.1)

- `STUDYID` : Study Identifier (from SDTM DM).
- `USUBJID` : Unique Subject Identifier (from SDTM DM).
- `SUBJID` : Subject Identifier for the Study (from SDTM DM).
- `SITEID` : Study Site Identifier (from SDTM DM).

2. Demographic Variables (Table 3.2.2)

- `AGE` : Subject's Age (from SDTM DM).
- `AGEU` : Age Units (from SDTM DM).
- `SEX` : Sex (from SDTM DM).
- `RACE` : Race (from SDTM DM).

- Once clicking on "DeepThink(R1), it will start "Deep Research" functionality and thinking about 26 secondes.
- We will see its thinking process, then eventually it provide the response.

Confidential – Do not duplica

Gen AI capabilities

- Reasoning
- Web Access and Research
- Image Generation
- Read documents, images, videos
-

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Reasoning

Recent Breakthrough:

- Letting an AI "think" before answering leads to better results
- The thinking process often feels surprisingly human-like

How to Use Reasoning Models

- Designed for deep thinking – like scholars, not chatty assistants
- Expect some delay – they take time to generate thoughtful answers
- Best for:
  - Academic research
  - Complex math and computer science questions
- Tip: Always provide clear, detailed context for better answers

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Model Comparison

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Data Analysis - Artifacts

	OpenAI	Gemini	Anthropic	DeepSeek
Reasoning	o1, GPT-4o	2.0 Flash Thinking	Sonnet 3.7	r1
Web Access	GPT	2.0 Flash		r1
image	sora	Imagen 3		
Code Execution	Canvas		artifact	
Live mode				

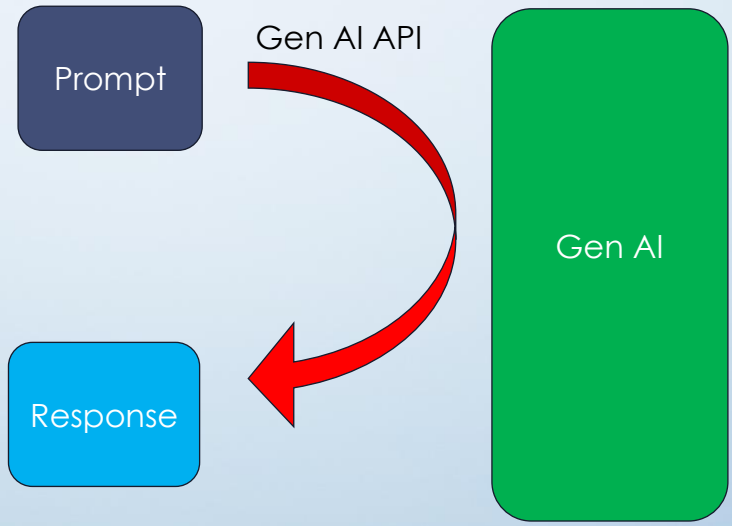
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Application Development using Gen AI API

- SAS
- Python
- R

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

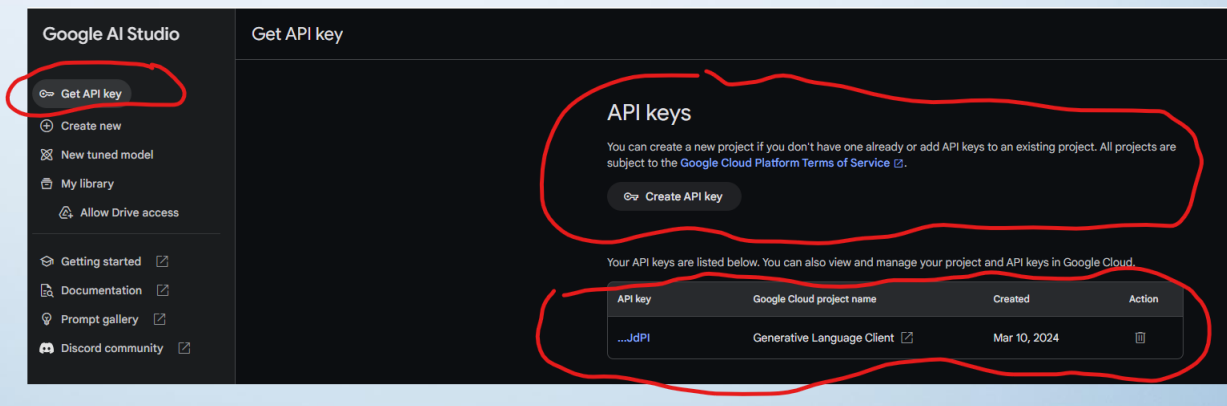
Application Developing using Gen AI API



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

How to create Gemini API

- API Keys in https://aistudio.google.com/app/apikey



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

SAS Application using Gemini API (free)

```
{ "contents" :
    [{ "parts" :
        [{"text" : <prompt> }]
    }]
}
```

```
{ "candidates" :
    [ "content" :
        { "parts" :
            [{ "text" : <response>
            }]
        }
    ],

  "usageMetadata" :
    { "promptTokenCount": 12,,},

  "modelVersion":"gemini-1.5-flash"
}
```

```
%let api_key = xxxxxx;
%let url =
https://generativelanguage.googleapis.com/v1/models/gemini-1.5-flash:generateContent;
%let url_query = &url.?key=&api_key. ;

filename prompt "F:\biometrics\project\GenAI\Gemini\prompt.txt";
filename resp "F:\biometrics\project\GenAI\ Gemini\resp.json";

Proc http
    method = "POST"
    url="&url_query"
    ct="application/json"
    in=promt
    out=resp;
Run;

libname response JSON fileref=resp;

data r_sas;
    set response.content_parts;
run;
```

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

SAS Application using Gen AI (Gemini) API

prompt.txt

```
{ "contents" :
    [{ "parts" :
        [{"text" : <prompt> }]
    }]
}
```

resp.json

```
{ "candidates" :
    [ "content" :
        { "parts" :
            [{ "text" : <resp>
            }]
        }
    ],
}
```

```
%let api_key = xxxxxx;
%let url =
https://generativelanguage.googleapis.com/v1/models/gemini-1.5-flash:generateContent;

%let url_query = &url.?key=&api_key. ;

filename prompt
"F:\biometrics\project\GenAI\Gemini\prompt.txt";
filename resp "F:\biometrics\project\GenAI\
Gemini\resp.json";

Proc http
    method = "POST"
    url="&url_query"
    ct="application/json"
    in=prompt
    out=resp;
Run;

libname response JSON fileref=resp;
data r_sas;
    set response.content_parts;
run;
```

Python Application using Gemini API

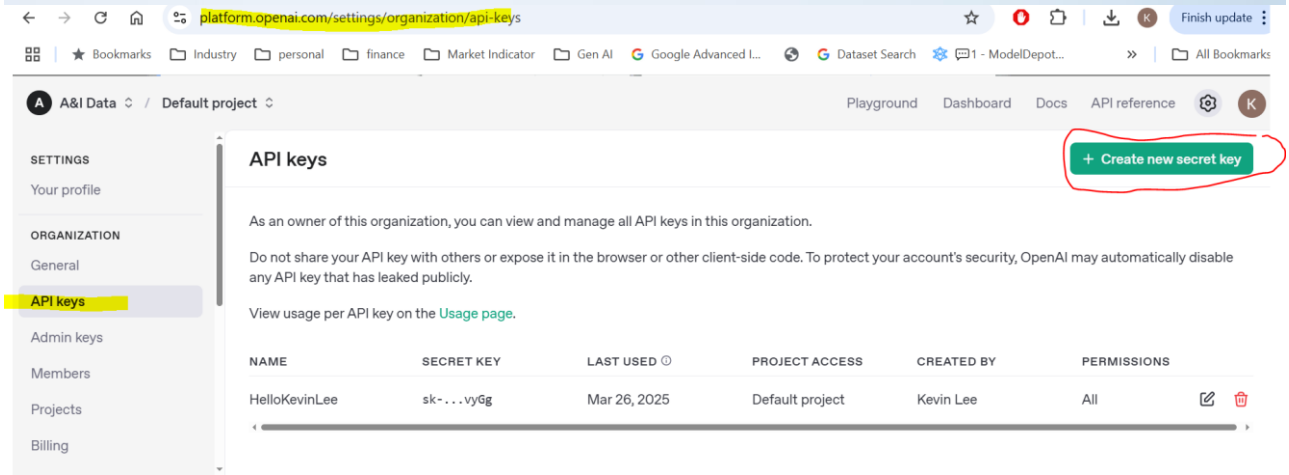
```
## import Google Gemini library
import google.generativeai as genai

## Create Prompt message
prompt = " Tell me about CDISC in 200 words" }]

## Set up Google Gemini  API
genai.configure(api_key="xxxx")

## Send Prompt message to Gemini and get the response
google_model = genai.GenerativeModel('gemini-pro')

response = google_model.generate_content(prompt)
```

```
response:
GenerateContentResponse(
    done=True,
    iterator=None,
    result=glm.GenerateContentResponse({'candidates': [{'content': {'parts': [{'text': 'CDISC (Clinical Data Interchange Standards Consortium) is a non-profit organiz
ation dedicated to developing and maintaining data standards for clinical research. Its mission is to enhance data quality, facilitate data sharing, and streamline th
e clinical research process. CDISC standards provide a common language for researchers, data managers, and regulatory agencies to exchange and interpret clinical dat
a. They include standards for data models, terminology, and data exchange formats, ensuring consistency and interoperability of clinical research data.'}], 'role': 'm
odel'}, 'finish_reason': 1, 'index': 0, 'safety_ratings': [{'category': 9, 'probability': 1, 'blocked': False}, {'category': 8, 'probability': 1, 'blocked': False},
{'category': 7, 'probability': 1, 'blocked': False}, {'category': 10, 'probability': 1, 'blocked': False}], 'token_count': 0, 'grounding_attributions': []}]}),
)
```

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

How to create ChatGPT API (paid)

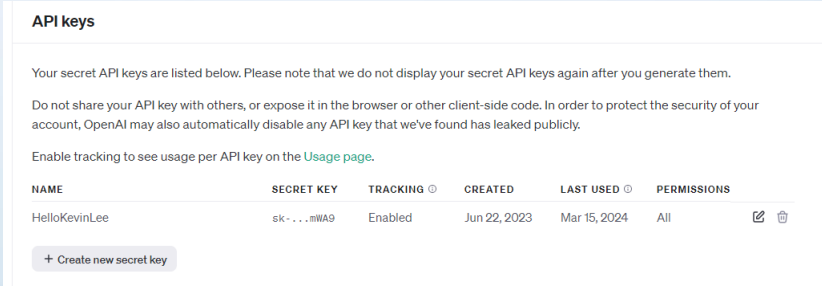
- API Keys in https://platform.openai.com/api-keys



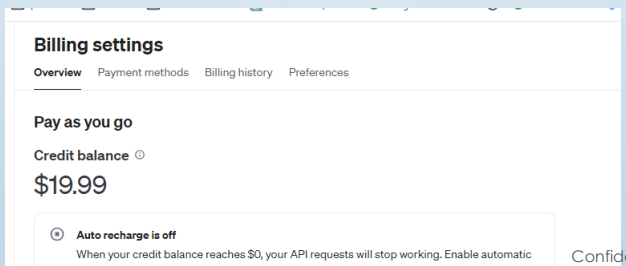
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

How to create ChatGPT API (paid)

- API Keys in https://platform.openai.com/api-keys



- add credit for API in https://platform.openai.com/account/billing/overview



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

SAS Application using ChatGPT API

```
{ "model" : "gpt-3.5-turbo",
  "message" :
    [{ "role" : "user",
       "content" : <prompt>
    }]
}
```

```
{ "id" : "xxxx",
  "object" : "chat.completion",
  "created" :  1710533285,
  "model" : "gpt-3.5-turbo-0125",
  "choices" :
    [{  "index" : 0,
        "message" :
          { "role" : "assistant",
            "content" : <resp>
          },
        "logprobs" : null,
        "finish_reason" : "stop"
    }],
  "usage": { },
  "system_fingerprint": "xxx"
}
```

```
%let api_key = xxxxxxx ;

filename prompt
"F:\biometrics\project\GenAI\OpenAI\prompt.txt";
filename resp
"F:\biometrics\project\GenAI\OpenAI\resp.json";

Proc http
      method = "POST"
      url="https://api.openai.com/v1/chat/completions"
      ct="application/json"
      in=promt
      out=resp;
      Headers "Authorization" = "Bearer &api_key.";
Run;

libname response JSON fileref=resp;
data r_sas;
      set r_json.choices_message;
run;
```

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

SAS Application using Gen AI (ChatGPT) API

prompt.txt

```
{ "model" : "gpt-3.5-turbo",
  "message" :
    [{ "role" : "user",
       "content" : <prompt>
    }]
}
```

resp.json

```
{ "id" : "xxxx",
  "object" : "chat.completion",
  "created" :  1710533285,
  "model" : "gpt-3.5-turbo-0125",
  "choices" :
    [{  "index" : 0,
        "message" :
          { "role" : "assistant",
            "content" : <resp>
          },
        "logprobs" : null,
        "finish_reason" : "stop"
    }],
  "usage": { },
  "system_fingerprint": "xxx"
}
```

```
%let api_key = xxxxxxx ;

filename prompt
"F:\biometrics\project\GenAI\prompt.txt";
filename resp
"F:\biometrics\project\GenAI\resp.json";

Proc http
      method = "POST"
      url="https://api.openai.com/v1/chat/completions"
      ct="application/json"
      in=prompt
      out=resp;
      Headers "Authorization" = "Bearer &api_key.";
Run;

libname response JSON fileref=resp;
data r_sas;
      set response.choices_message;
run;
```

Python application using ChatGPT API

```
## import OpenAI ChatGPT library
import openai

## Create Prompt message
message = [{"role" : "user", "content " : " Tell me about CDISC in 200 words" }]

## Set up OpenAI ChatGPT API
openai.api_key = XXXXXXXX

## Send Prompt message to OpenAI
## Get the response
response = openai.chat.completion.create(
      model="gpt-3.5-turbo",
      messages=message,
      temperature=0
)
```


```
{
  "id": "chatcmpl-98x12kqDpn0K1YUFMjddpoKY5JwRY",
  "object": "chat.completion",
  "created": 1711920508,
  "model": "gpt-3.5-turbo-0125",
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": "CDISC, or the Clinical Data Interchange Standards Consortium, is a global non-profit organization that develops an
d promotes data standards for clinical research. Founded in 1997, CDISC aims to streamline the process of collecting, analyzing, and s
haring clinical trial data to improve efficiency and facilitate data sharing among researchers, regulators, and industry stakeholder
s.\n\nCDISC standards cover a wide range of areas, including data collection, data submission, and data analysis. These standards help
ensure that data is collected consistently across different studies, making it easier to compare and combine data from multiple source
s. By using CDISC standards, researchers can save time and resources, reduce errors, and improve the quality and reliability of their
data.\n\nIn 2002, CDISC released its first set of standards, known as the Study Data Tabulation Model (SDTM). Since then, CDISC has de
veloped several other standards, including the Analysis Data Model (ADaM) and the Define-XML standard for defining metadata in clinica
l trials.\n\nOverall, CDISC plays a crucial role in advancing the field of clinical research by promoting data standardization and int
eroperability, ultimately leading to better and more efficient healthcare outcomes for patients."
      },
      "logprobs": null,
      "finish_reason": "stop"
    }
  ],
  "usage": {
    "prompt_tokens": 18,
    "completion_tokens": 235,
    "total_tokens": 253
  },
  "system_fingerprint": "fp_b28b39ffa8"
}
```


Confidential – Do not duplicate or distribute without the permission of Kevin Lee

R application using ChatGPT API

```
## import OpenAI ChatGPT library
library(httr)
library(jsonlite)

# Set your OpenAI API key from .env file
api_key <- "xxxx"

# Define the request URL
api_url <- "https://api.openai.com/v1/chat/completions"

prompt <- "Tell me about CDISC in 200 words"

# Define the request body
body <- list(
  model = "gpt-3.5-turbo",
  messages = list(
    list(role = "system", content = "You are a helpful assistant."),
    list(role = "user", content = prompt )
  )
)
```

```
# Make the API request
response <- POST(
  api_url,
  add_headers('Authorization' = paste("Bearer", api_key),
              'Content-Type' = "application/json"),
  body = jsonlite::toJSON(body, auto_unbox = TRUE),
  encode = "json"
)

# Parse the response
result <- content(response, as = "parsed")

# Print the response
print(result$choices[[1]]$message$content)
```

```
result                  List of  8
    $ id                  : chr "chatcmpl-B3rIkdVKWMXRhLCdJkiPSrCyLSG1W"
    $ object              : chr "chat.completion"
    $ created             : int 1740259574
    $ model               : chr "gpt-3.5-turbo-0125"
    $ choices             :List of 1
     ..$ :List of 4
     .. ..$ index         : int 0
     .. ..$ message       :List of 3
     .. .. ..$ role    : chr "assistant"
     .. .. ..$ content: chr "CDISC, or the Clinical Data Interchange Standar…
     .. .. ..$ refusal: NULL
     .. ..$ logprobs      : NULL
     .. ..$ finish_reason: chr "stop"
```

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

LangChain

- Definition
- Functions
- Components

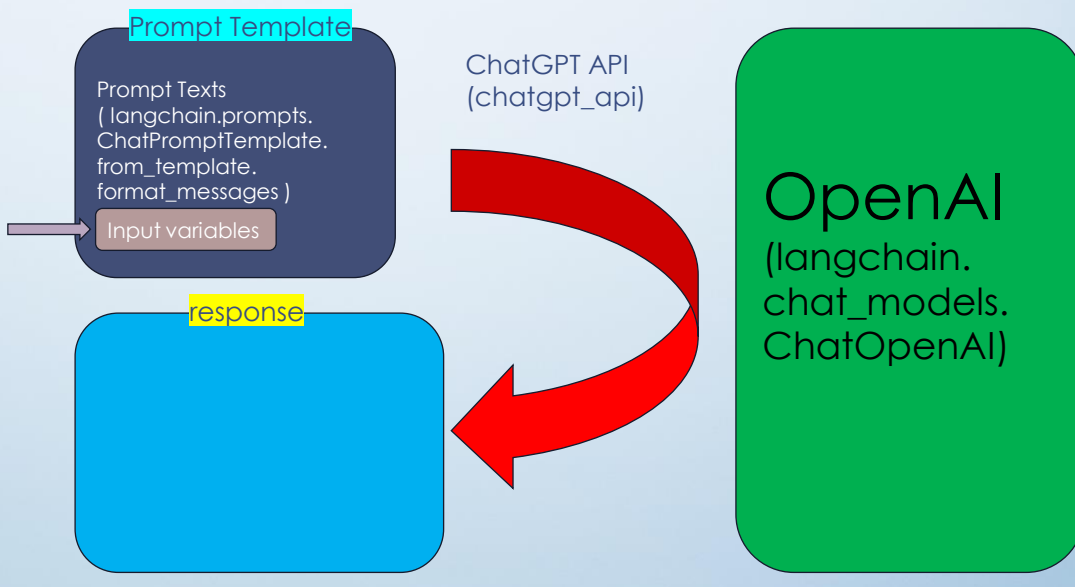
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

LangChain

- Definition : Open-source development framework for developing LLM application.
- Functions
  - Modular components
  - Use cases : Common ways to combine components
- Components
  - Models
    - LLM : 20+ integrations
    - Chat models
    - Text Embedding models
  - Prompts
    - Prompt Templates
    - Output Parsers
    - Example Selections
- Indexes
  - Document Loading
  - Text Splitting
  - Vector Stores
  - Retrievers
- Chains
  - Prompt + LLM + Output Parsing
  - Can be used for building longer chains
- Agents
  - Agent Types :
  - Algorithms for getting LLMs to use tools
  - Agent Toolkits : sort, sum

Confidential – Do not duplicate or distribute without the permission of Kevin

LangChain Framework



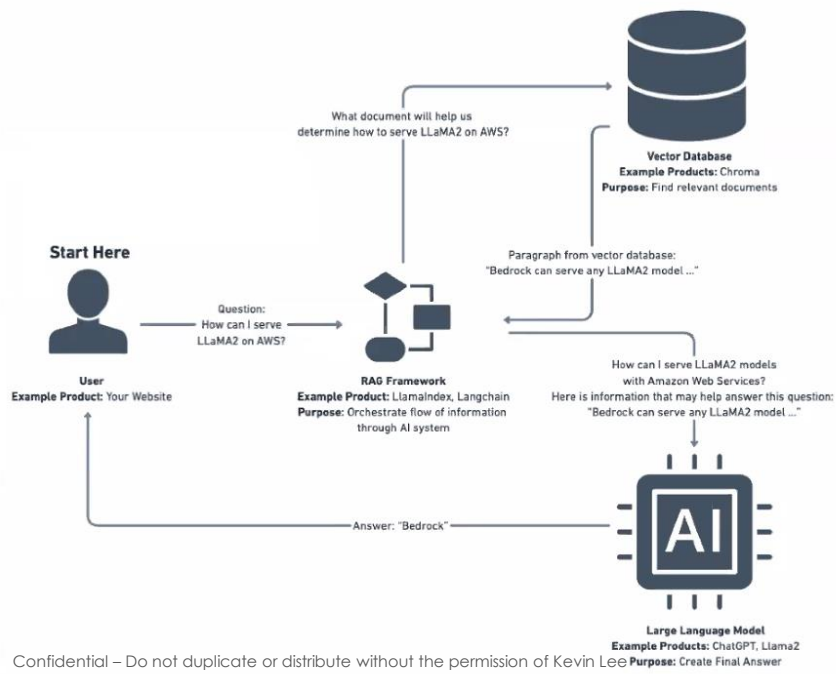
Retrieval Augmented Generation (RAG)

- Definition
- Framework with LangChain

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

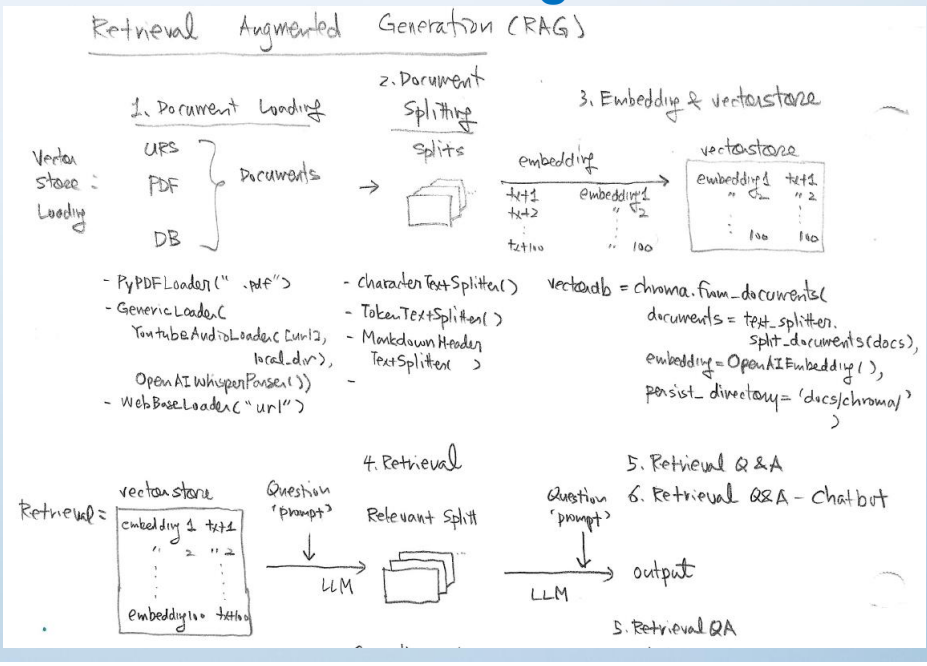
RAG Definition

- Retrieval-Augmented Generation
- Process used to improve the accuracy and reliability of LLMs
- Process
  1. Create vector DB using LLM embedding.
  2. Retrieve the information relavant to questions from vector DB
  3. Feed the retrieved info to LLMs
  4. LLMs generates the contents from retrieved info.

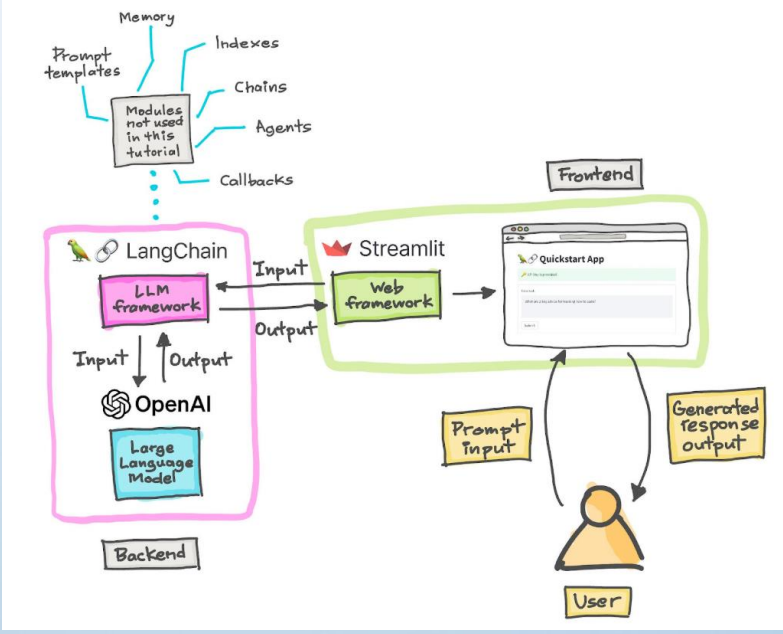


Confidential – Do not duplicate or distribute without the permission of Kevin Lee

RAG framework in LangChain



Gen AI application architecture



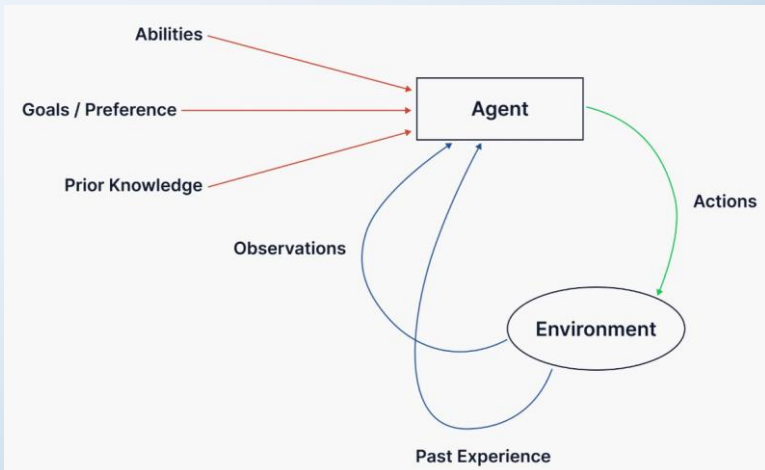
AI Agent

- Definition
- Framework with LangChain

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

AI Agent

- System that can interact with its environment, collect data, and autonomously peform tasks and make decisions to achieve the goals.
- Multiple prompts using different tools
- Main Components
  - LLM
  - Prompts
  - Tools
  - Memory (short-term /long-term memory, RAG)



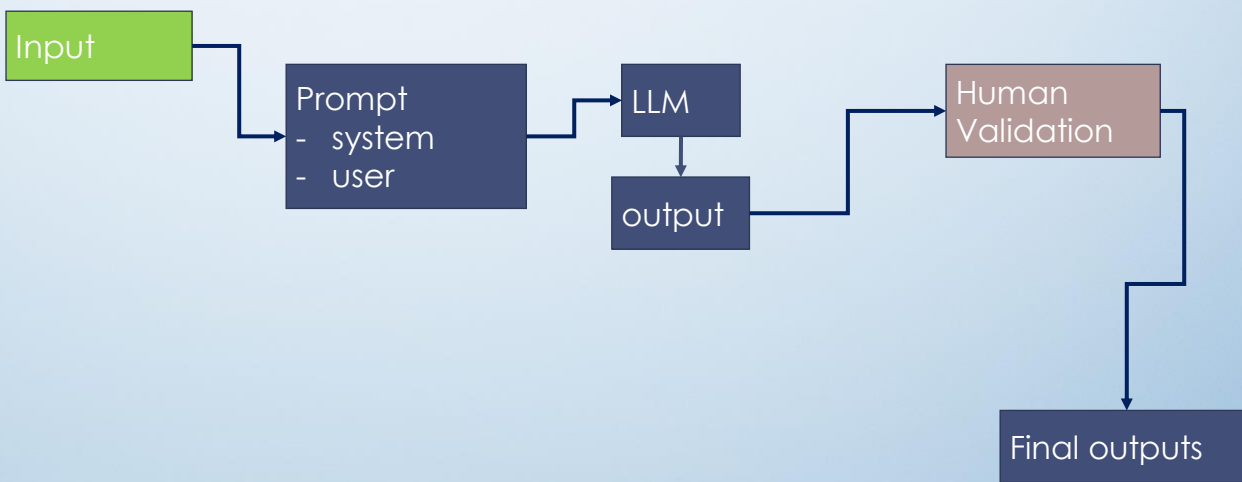
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Traditional GenAI vs AI Agent

Traditional GenAI	AI Agent
One time direct answer generation – one prompt at a time	Multi-step problem processing – using Agentic workflow.
Dependent – require human prompt continuously	Autonomous – can start next actions without human prompt
Can answer	Can reason and decide
Stateless – No memory or very short-term memory	Stateful – can keep a context as long as the session is running
No tool use	Tool use – databases, RAG, APIs, MCP

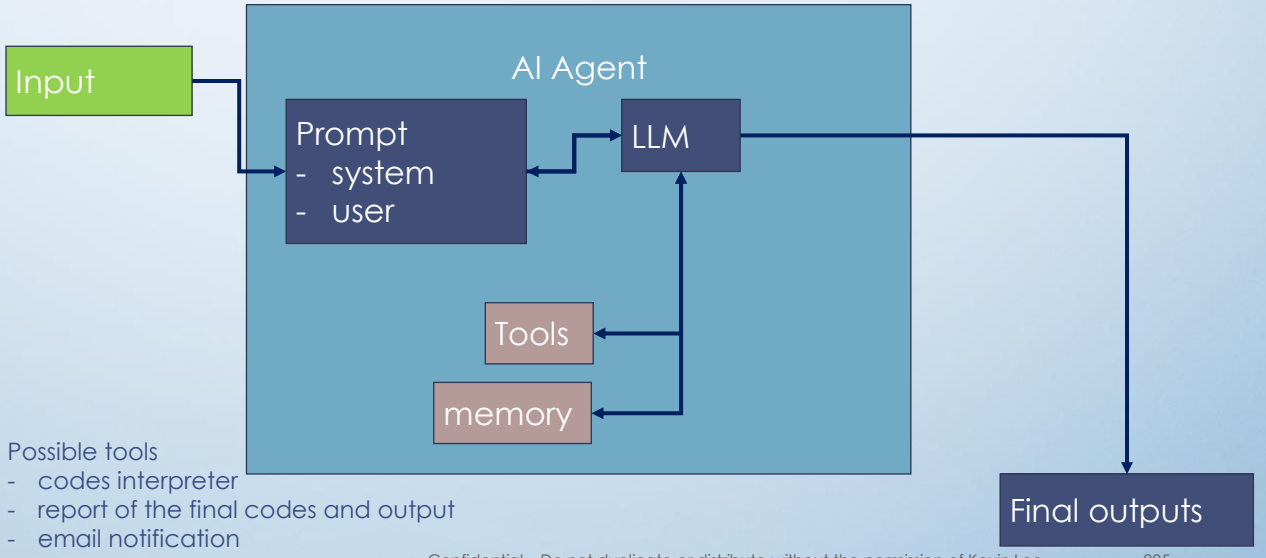
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Artifacts Development using Traditional Gen AI



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Artifacts Development using AI Agent

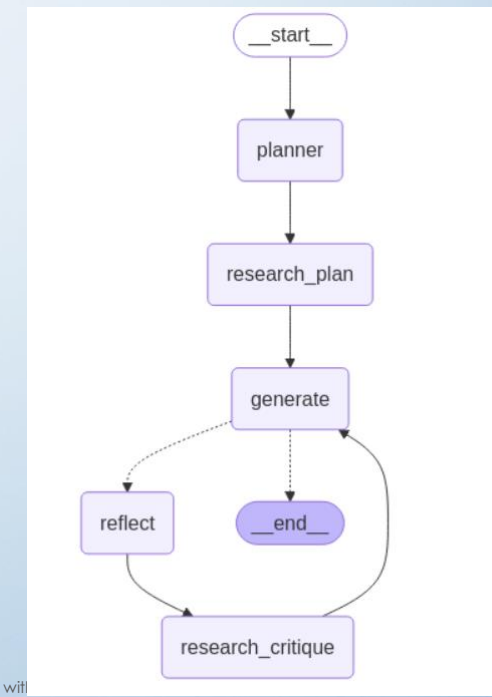


Possible tools
- codes interpreter
- report of the final codes and output
- email notification
- human intervention

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

How to build AI Agent workflow

1. Define Agent State
2. Define LLM model
3. Define Prompts
4. Define tools for AI Agents
5. Define functions for nodes
6. Define Memory
7. Define workflow
   a) Start workflow with Agent State
   b) Assign nodes to functions
   c) Assign starting node
   d) Assign conditional edges
   e) Assign edges
   f) Compile memory
8. Compile AI Agent
9. Prepare input (user prompts)
10. Run AI Agent



Confidential – Do not duplicate or distribute wit

Sample Codes for AI Agent

```
class AgentState(TypedDict):
    task: str
    plan: str
    draft: str
    ,,,,

llm = ChatOpenAI(model="gpt-3.5-turbo")

prompt_system = """  ,,,,  """
prompt_user = """  ,,,,  """

tool_tavily = TavillyClient( )

def planner(state : AgentState ) :
    response = llm.invoke(prompt_user)
    return {"plan" : response.content }
def research_plan(state : AgentState ) :
    result = tool_tavily.search( )
    return {"draft" : result.content )
,,,,,
mem_store = InMemoryStore( index={"embed":
"openai:text-embedding-3-small"})
```

```
workflow = StateGraph(AgentState)
workflow.add_node("planner", planner)
workflow.ddd_node("research plan",
research_plan)
workflow.add_node("generate", generate)

workflow.set_entry_point("planner")

workflow.add_conditional_edges("generator"
, continue, {END: END, "reflect" : "reflect"}  )

workflow.add_edge("planner",
"research_plan")
workflow.add_edge("research_plan",
"generate")

ai_agent =
workflow.compile(store=mem_store)

input = {  }

final_response = ai_agent.invoke( input )
```

icate or di

6 different AI Agentic Design

1. **ReAct Agent**
➜Thinks, takes action, looks at the result, repeats.
➜Classic loop: "Should I Google this?" → Does it → Adjusts.
➜ Used in most AI products today (like basic chat assistants).

2. **CodeAct Agent**:
➜ Runs real code, not just JSON.
➜ So instead of saying "call API X," it writes and runs a Python script.
➜ More powerful. Used in research agents and dev assistants.

3. **Modern Tool Use**:
➜ Sends tasks to smart APIs (search, cloud, data), and lets them do the heavy lifting.
➜ The agent mostly routes and formats info.
➜ Think: a smart middleman between you and powerful services.

4. **Self-Reflection**:
➜ Agent checks its own work.
➜ Did it make a mistake? It catches it, critiques it, and tries again.
➜ Most AI errors happen **because this step is missing.**

5. **Multi-Agent Workflow**:
➜ One agent isn't doing everything.
➜ You have a planner, a researcher, and a writer — all working together.
➜ Like a mini team of AIs. More accurate. Less chaos.

6. **Agentic RAG**:
➜ This is what powers apps like Perplexity.
➜ The agent looks stuff up (retrieval), thinks about it, uses tools, and gives you a smarter answer.
➜ Works with real-time data, not just model memory.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

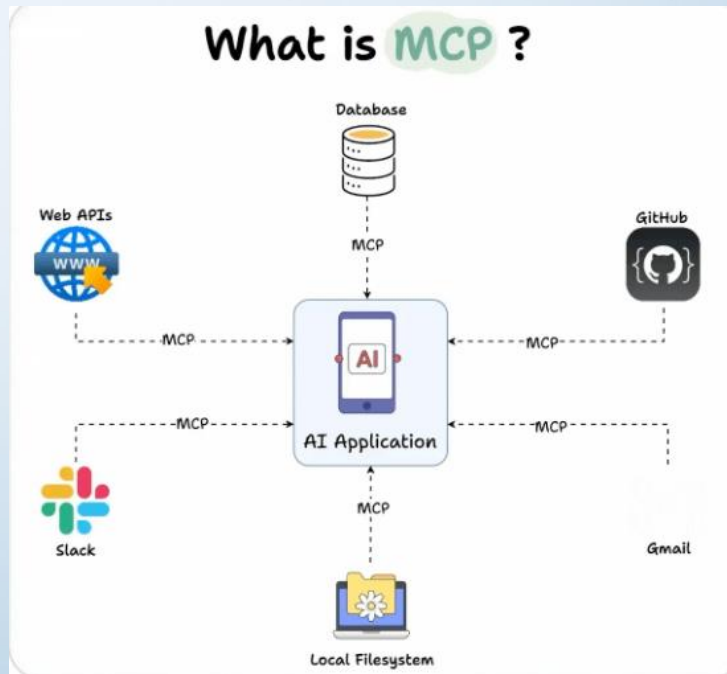
MCP

- Definition
- Structure

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

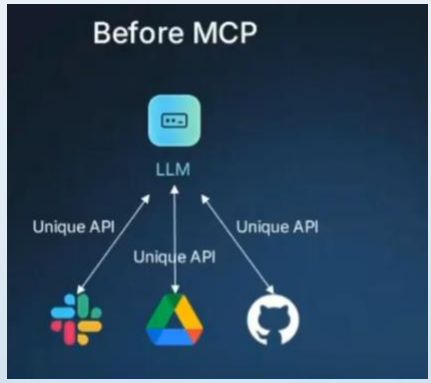
MCP

- Model Context Protocol (MCP) :
  - Open standard on how AI apps connect to different data sources and tools.
  - To simplify how AI models interact with external data and tools.
  - Initially developed by Anthropic in https://modelcontextprotocol.io/introduction



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

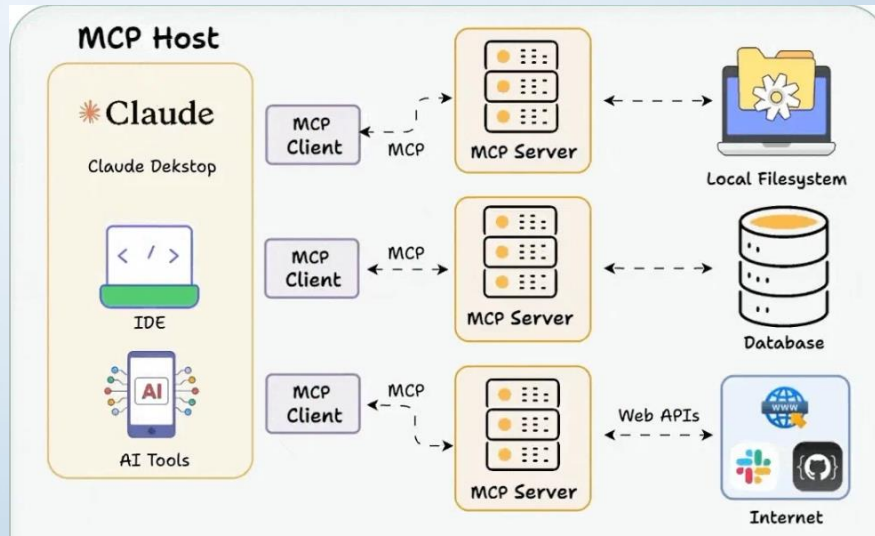
MCP



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

MCP- high lever structure

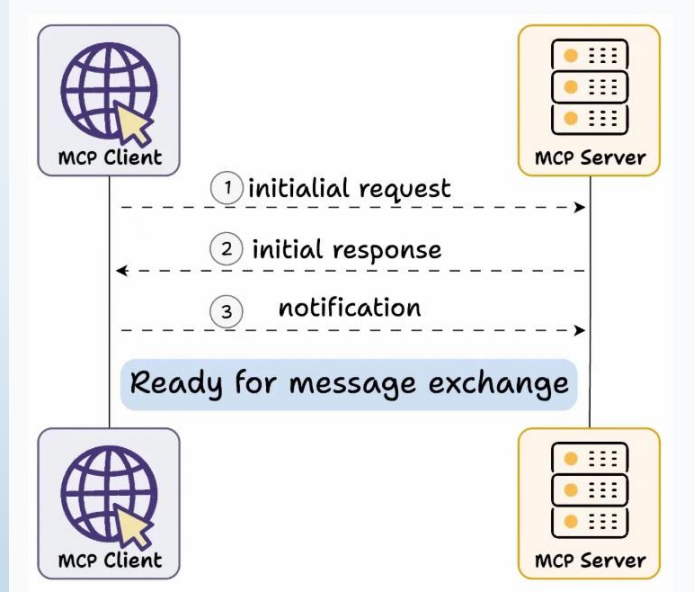
- MCP host : AI app with LLM
- MCP client : operates within the host to enable communication with MCP servers
- MCP Server : expose specific capabilities and provide access to
  - Tools : enable LLM to perform actions thru server
  - Resources : Expose data and content from servers to LLMs
  - Prompts : Create reusable prompt template and workflows



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

MCP- how MCP client interact with MCP Servers

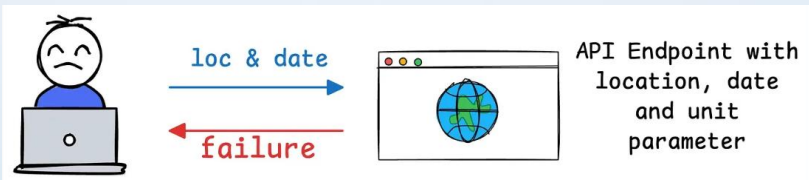
- MCP client :
- MCP Server



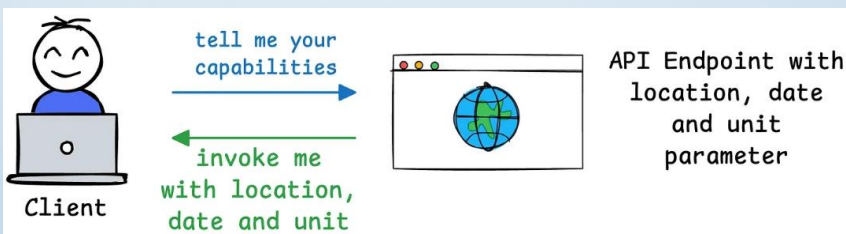
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

MCP vs traditional APIs

- Traditional APIs



- MCP



Advantages of MCP

- Use LLMs to refine the inputs much easier
- Minimize the failures, but invoke the clients to adjust.
- Clients can adjust their request much easier.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

AI Code Assistant

- Vibe-coding
- Github Copilot
- Gemini Code Assistant
- Tabnine
- Codeium

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Vibe Coding

- Andrej Karpathy (co-founder of openAI) – “There is a new kind of coding I call ‘vibe coding’, where you fully give in to the vibes, embrace exponentials, and forget that code even exists.”
- Coding with the help of AI
- MS CEO Satya Nadeli – 30% of the company’s code is now written by AI.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Vibe Coding Rules

1. Start from templates: No need to reinvent
2. Use agent mode: Don’t code, compose
3. Chat in tasks: New Composer chat = new job
4. Voice-to-text is speed: Tools like Whispr Flow unlock flow
5. Run + test early: Don't wait for the deploy
6. Deploy fast: Use Vercel, catch errors upstream
7. Fix bugs with AI: Copy/paste errors into Composer
8. Save good prompts: They're reusable gold
9. Commit often: Cursor can push for you
10. Secure your secrets: Always .env
11. Iterate without fear: Polish comes later
12. Keep Composer lean: Short, scoped, clean
13. Use Perplexity to ideate: Let it find APIs & designs
14. Fork wisely: Use GitHub like a buffet
15. Just vibe: Build fast. Think less. Have fun.

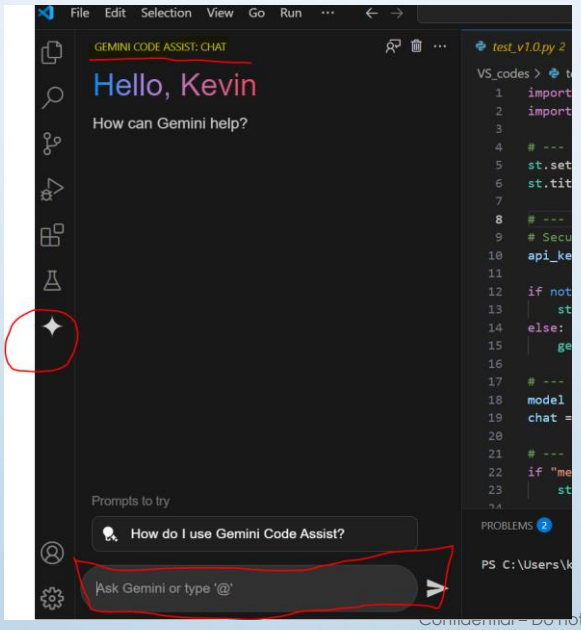
Confidential – Do not duplicate or distribute without the permission of Kevin Lee

GitHub Copilot

- Code suggestions,
- chat functionality,
- easy auto-complete navigation,
- multiple language and IDE support

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Gemini Code Assistant



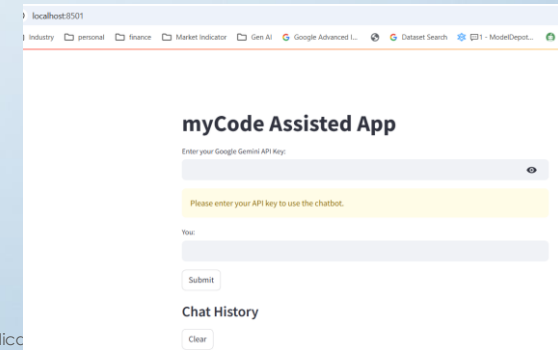
Prompt in Chat:

Create a streamlit script for a chatbot for google gemini.

Have a API kye placeholder, a input text, a submit button, and a output text area.

Have a clear button on the page.

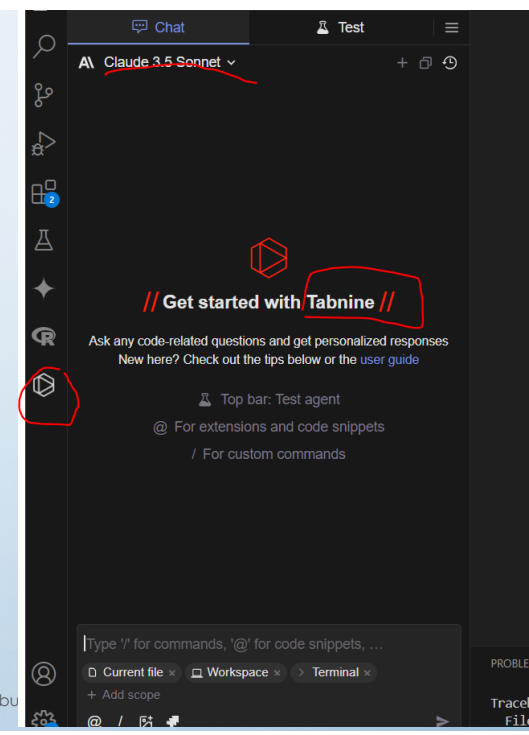
Title of page is my code Assisted App.



Confidential – Do not duplicc

Tabnine

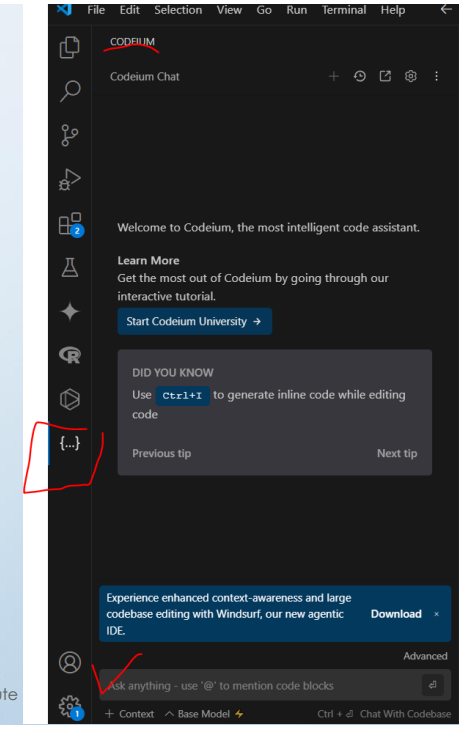
- Code suggestions,
- chat functionality,
- easy auto-complete navigation,
- multiple language and IDE support



Confidential – Do not duplicate or distribu

Codeium

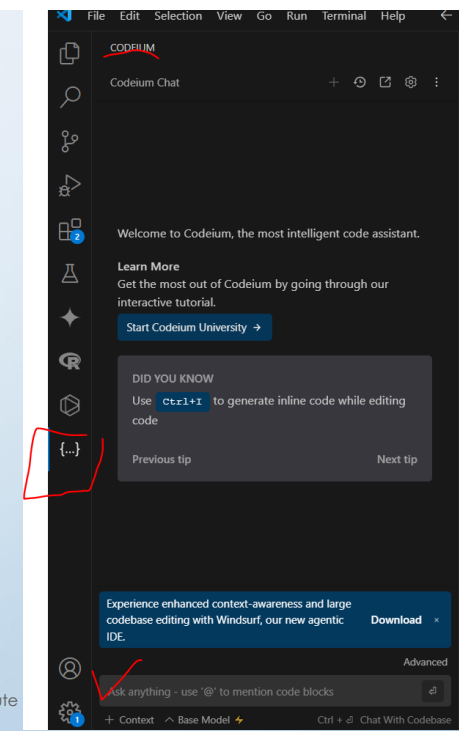
- AI autocomplete,
- prompt-driven development,
- test generation,
- multiple programming languages support and IDE integration



Confidential – Do not duplicate or distribute

Intellicode

- Real-world usage examples from GitHub,
- improved privacy



Confidential – Do not duplicate or distribute

Gen AI powered Application Development

- Pros
- Cons

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Gen AI Application Development - Pros

1. Rapid Prototyping and Development
- Gen AI APIs enable programmers to quickly build sophisticated applications by leveraging pre-trained Gen AI models.
- Significantly reduces development time compared to building AI capabilities from scratch

2. Advanced Functionality with Minimal Expertise
- Provides access to cutting-edge Gen AI capabilities like natural language processing, image generation, and complex reasoning
- Democratizes access to sophisticated Gen AI technologies across department

3. Scalability and Flexibility
- Cloud-based AI APIs can easily scale to handle varying levels of computational demand
- Flexible integration across different programming languages and frameworks
- Can be quickly adapted to different use cases and business requirements

4. Cost-Effectiveness
- Reduces infrastructure and computational costs associated with training and maintaining Gen AI models
- Pay-as-you-go pricing models allow for more predictable budgeting
- Eliminates the need for massive upfront investments in AI infrastructure

5. Continuous Improvement
- Gen AI providers continuously update and improve their models
- Access to the latest Gen AI advancements with minimal effort

Gen AI Application Development - Cons

1. Dependency on Gen AI company
- Heavy reliance on third-party Gen AI API providers
- Potential challenges in switching between different Gen AI service providers
- Risk of service disruptions or significant pricing changes

2. Privacy and Data Security Concerns
- Sending sensitive data through external APIs raises potential security risks
- Compliance challenges with data protection regulations
- Limited control over how data is processed and stored by API providers

3. Performance and Latency Issues
- Network latency can impact real-time application responsiveness
- Potential performance bottlenecks during high-traffic periods
- Dependence on external service availability and response times

Limited Customization
- Pre-trained Gen AI models might not perfectly align with specific use cases
- Restricted ability to fine-tune models compared to custom-built solutions
- Generic responses that may lack domain-specific nuance

5. Cost Unpredictability
- Usage-based pricing can lead to unexpected expenses (e.g., Gen AI application not suitable for record-based programming)
- Complex pricing structures might make budgeting challenging
- Potential for high costs with increased API usage

6. Ethical and Bias Considerations
- Inherit potential biases from pre-trained models
- Limited transparency in model decision-making processes
- Potential for generating inappropriate or controversial content

Strategic Gen AI / LLM / ChatGPT Implementation at Enterprise

- Risks and Concerns
- Benefits and Use Cases
- Cross Functional Team
- Policy and Guidelines
- Training and Education
- PoC & Evaluation
- Deployment

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Does your company allow ChatGPT at Work?

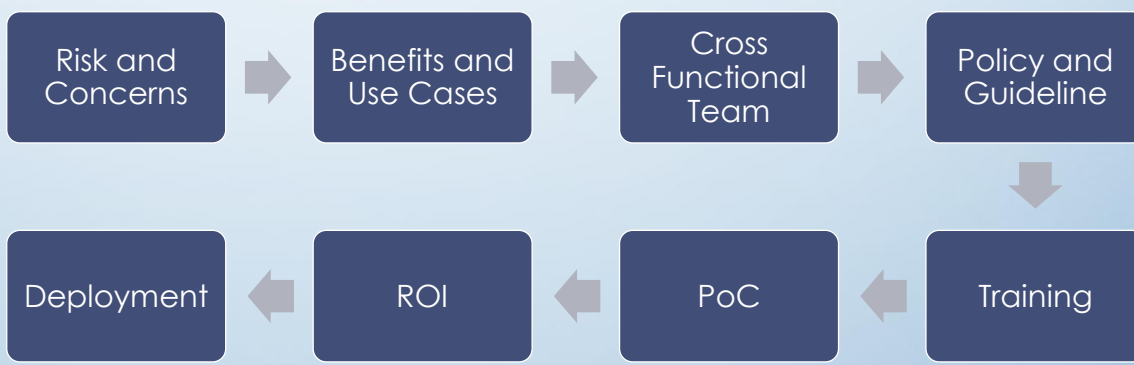
227

228

LLM (ChatGPT) Implementation Roadmap

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

LLM (ChatGPT) Implementation Roadmap



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Concerns using AI (ChatGPT)

- Data Privacy & Security
- Bias
- Accuracy
- Regulatory Compliance
- Ethical Consideration
- Black Box Feature
- Resource Requirements
- Lack of Expertise
- Cost vs ROI

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Data Privacy

- Data Privacy – the right of a person to have control over how their personal information is collected and used. (e.g., clinical trial data)
- Since ML is built using data, ML algorithms can contain sensitive information even though it is one of a Big Data.
- The growth of ML has increased the possibility of using sensitive data in ways that may violate data privacy. (e.g., ChatGPT)

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Bias of Machine Learning

- Phenomenon that occurs when ML algorithms produces results that are prejudiced due to wrong assumptions
- Examples :
  - Amazon ML recruiting engine show a bias on women.
    - ML was trained to vet applicants using 10 years of resume
    - Most came from men, a reflection of male dominance across tech industry.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Regulatory Compliance

- Regulatory Compliance – the process of adhering to laws, regulations, standards, and other rules set by governments and other regulatory bodies.
- For pre-market approval, AI-based software needs to follow regulatory guidelines (e.g., Good Machine Learning for Medical Device Development Guiding Principles)

U.S. FOOD & DRUG ADMINISTRATION | Health Canada Santé Canada | Medicines & Healthcare products Regulatory Agency

Good Machine Learning Practice for Medical Device Development: Guiding Principles

October 2021

The U.S. Food and Drug Administration (FDA), Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) have jointly identified 10 guiding principles that can inform the development of Good Machine Learning Practice (GMLP). These guiding principles will help promote safe, effective, and high-quality medical devices that use artificial intelligence and machine learning (AI/ML).

Artificial intelligence and machine learning technologies have the potential to transform health care by deriving new and important insights from the vast amount of data generated during the delivery of health care every day. They use software algorithms to learn from real-world use and in some situations may use this information to improve the product's performance. But they also present unique considerations due to their complexity and the iterative and data-driven nature of their development.

These 10 guiding principles are intended to lay the foundation for developing Good Machine Learning Practice that addresses the unique nature of these products. They will also help cultivate future growth in this rapidly progressing field.

Good Machine Learning Practice for Medical Device Development: Guiding Principles	
Multi-Disciplinary Expertise Is Leveraged Throughout the Total Product Life Cycle	Good Software Engineering and Security Practices Are Implemented
Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population	Training Data Sets Are Independent of Test Sets
Selected Reference Datasets Are Based Upon Best Available Methods	Model Design Is Tailored to the Available Data and Reflects the Intended Use of the Device
Focus Is Placed on the Performance of the Human-AI Team	Testing Demonstrates Device Performance During Clinically Relevant Conditions
Users Are Provided Clear, Essential Information	Deployed Models Are Monitored for Performance and Re-training Risks are Managed

The 10 guiding principles identify areas where the International Medical Device Regulators Forum (IMDRF), international standards organizations, and other collaborative bodies could work to advance GMLP. Areas of collaboration include research, creating educational tools and resources, international harmonization, and consensus standards, which may help inform regulatory policies and regulatory guidelines.

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Ethical Consideration

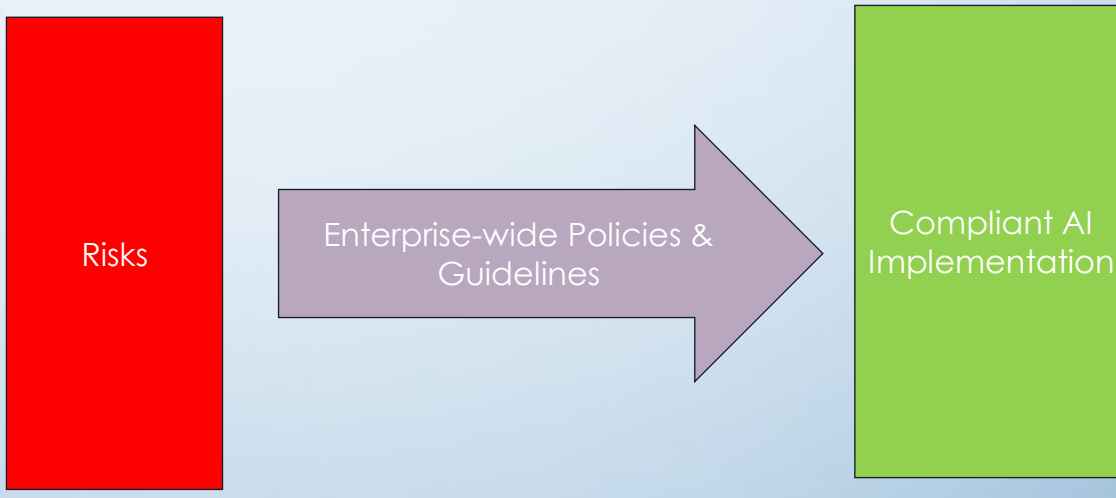
- Who will make decision? AI or human?
- Can AI be responsible for their decision?
- Is AI transparent?

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

AI Implementation through Enterprise Policy and Guidelines

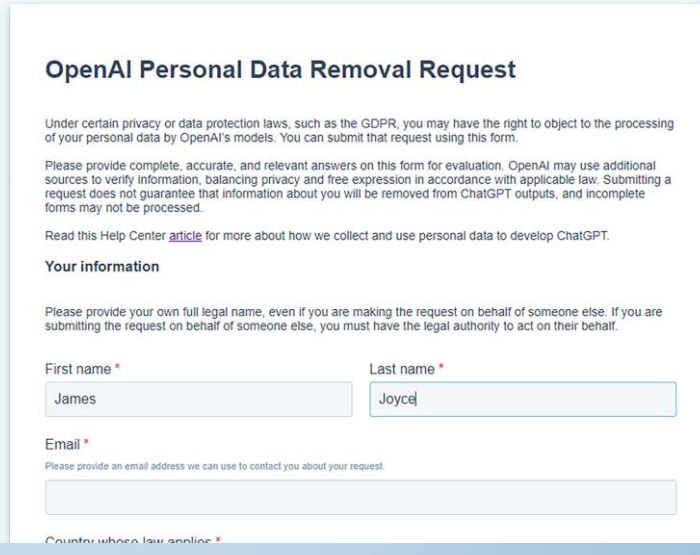
- Data Privacy
- Regulatory Compliance
- Secure Data Storage
- Secure Access in Gen AI

Risk mitigation through Policies and Guidelines



Data Privacy

- The protection and control of confidential company information collected, processed, and utilized by AI systems.
- Examples to protect data privacy in AI
  - Don't share sensitive data with AI
  - Be cautious of what you share online
  - Stay anonymous
  - Data Encryption
    - Secure Socket layer (SSL)
    - Transport Layer Security (TLS)



OpenAI Personal Data Removal Request

Under certain privacy or data protection laws, such as the GDPR, you may have the right to object to the processing of your personal data by OpenAI's models. You can submit that request using this form.

Please provide complete, accurate, and relevant answers on this form for evaluation. OpenAI may use additional sources to verify information, balancing privacy and free expression in accordance with applicable law. Submitting a request does not guarantee that information about you will be removed from ChatGPT outputs, and incomplete forms may not be processed.

Read this Help Center article for more about how we collect and use personal data to develop ChatGPT.

Your information

Please provide your own full legal name, even if you are making the request on behalf of someone else. If you are submitting the request on behalf of someone else, you must have the legal authority to act on their behalf.

First name * James

Last name * Joyce

Email *

Please provide an email address we can use to contact you about your request.

Country whose law applies *

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Regulatory Compliance

- Understand Regulatory Policy
  - FDA Comments: AI/ML will undoubtedly play a critical role in drug development, and FDA plans to develop and adopt a flexible risk-based regulatory framework that promotes innovation and protects patient safety.
- Policy Implementation
  - General Data Protection Regulation (GDPR)
  - Good Practice (GxP)
  - Code of Federal Regulation (CFR)
  - Health Insurance Portability and Accountability Act (HIPAA)

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Secure Data Storage

- Security of Data Storage and Data Processing
- Access Control of Data and ChatGPT
- Data Anonymization and Pseudonymization

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Secure Access to ChatGPT

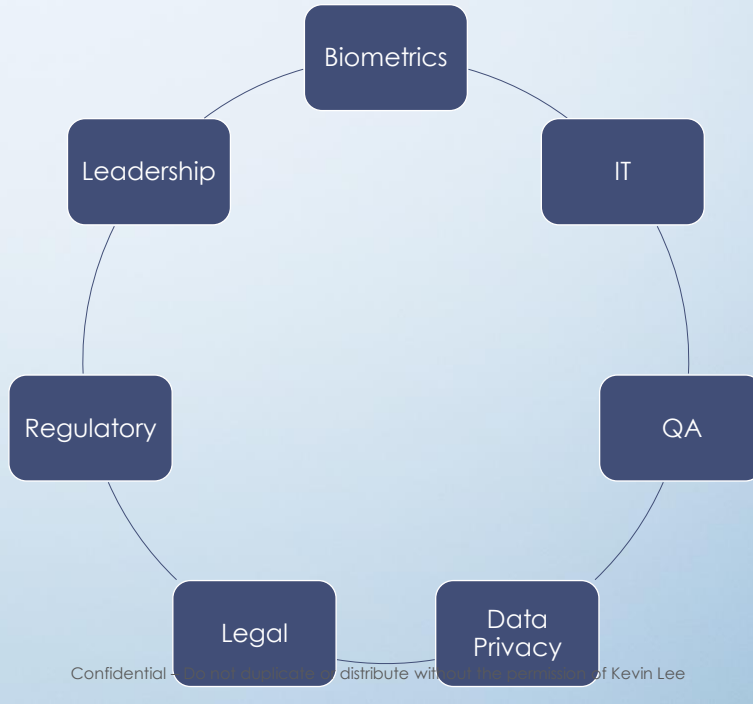
- Access Control of Data and ChatGPT
- Strong User Authentication Mechanism (e.g., Multi-factor Authentication)
- Audit Trails of user interactions with ChatGPT
- Regular Security Audits on ChatGPT
- Data Privacy Impact Assessment by ChatGPT

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

ChatGPT Implementation and Governance Cross-Functional Team

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Cross-Functional Team



Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Employee Training

- Employees who will be using and managing ChatGPT
- Security best practices
- Data Privacy
- Data Handling Procedures with ChatGPT
- Ethics and Compliance

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

PoC & Evaluation

- Security and Compliance
- ROI
- User Experience
- Training and Support
- Ethical Considerations

45

Gen AI Deployment

- Enterprise License (CoPilot, ChatGPT Enterprise)
- Home-Grown Gen AI
- Customized Gen AI implementation at function

ChatGPT Enterprise

- Enterprise-grade security & privacy
- OpenAI does not train on enterprise business data.
- All conversations are encrypted.
- Compliance : CCPA, GDPR, SOC2, SOC3
- Enterprise Security Trust Portal

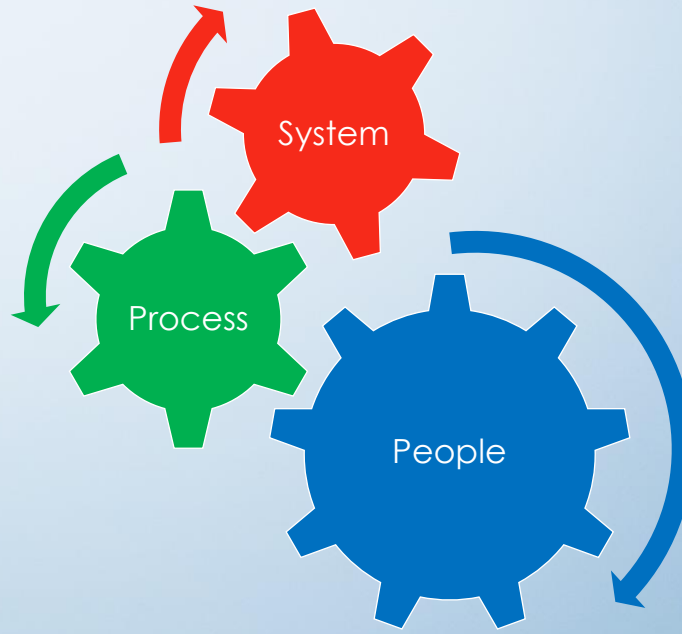
ChatGPT Enterprise

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

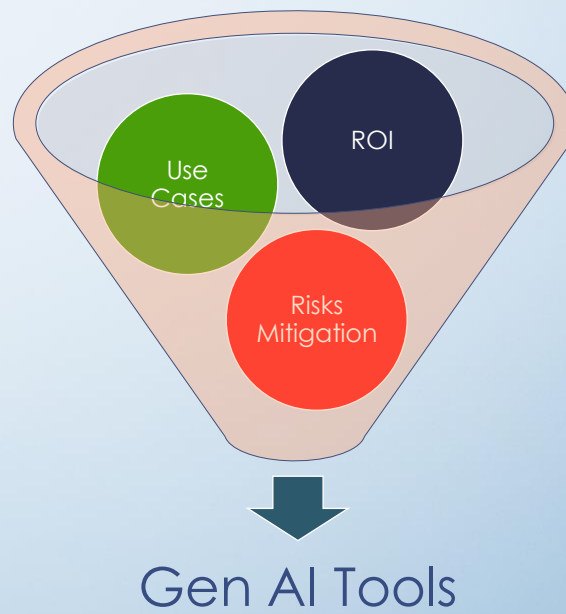
Gen AI implementation in Biometrics

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Three pillars to maximize Gen AI Capability in Biometrics



Use Cases driven Gen AI Tools



251

Gen AI Process

- Validation?
- Human intervention
- Final output or source of input?

ChatGPT is like a very smart, capable, recently graduated intern/new hire.

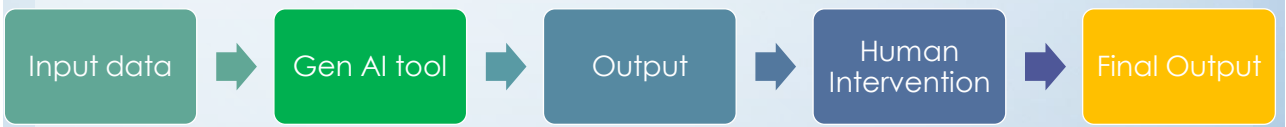
=

What will we do with the results from an intern?

The output of Gen AI is

not the final result,
but the beginning of new process.

Gen AI driven Process



People : Biometrics team powered with Gen AI

-Super Biometrics Team

Gen AI - Electricity

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

ChatGPT Users vs Non-Users Comparison in Boston Consulting

- Results of ChatGPT Users
  - finished 12.2% more tasks on average
  - completed tasks 25.1% more quickly
  - produced 40% higher quality results

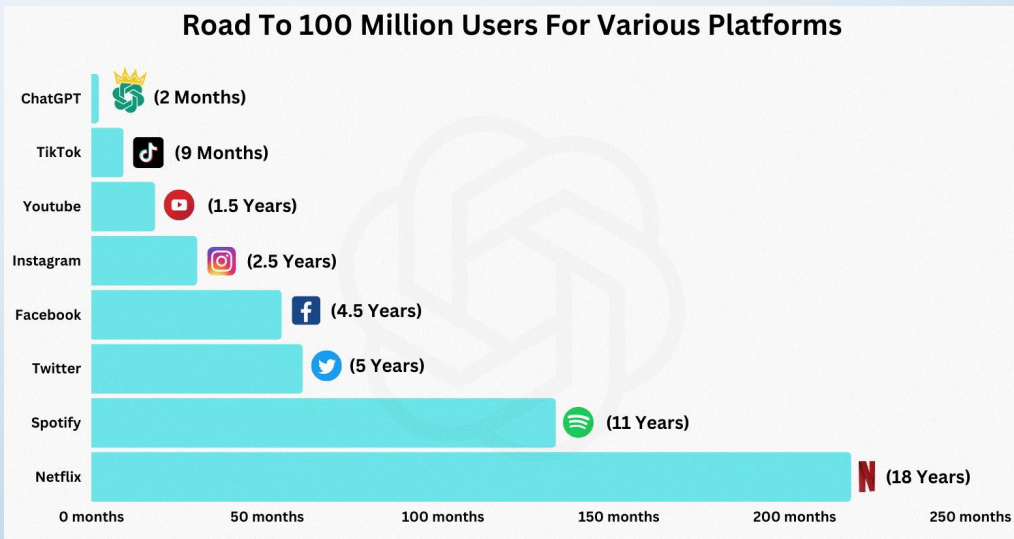
259

Gen AI gives us a Head Start. We don't need to start from the blank page.

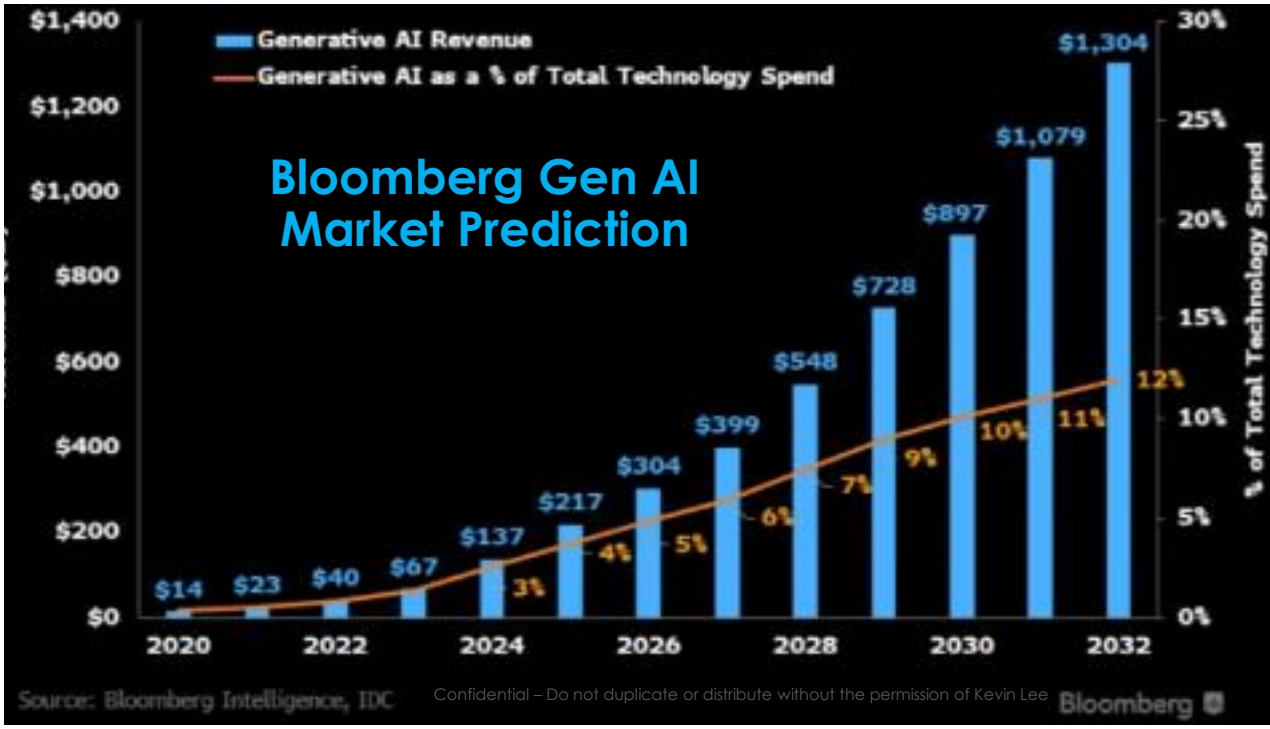
260

Future of Gen AI and US

ChatGPT Popularity



ChatGPT gained 100M users within 2 months since its release



Gen AI has arrived

264

Attacking vs Understanding

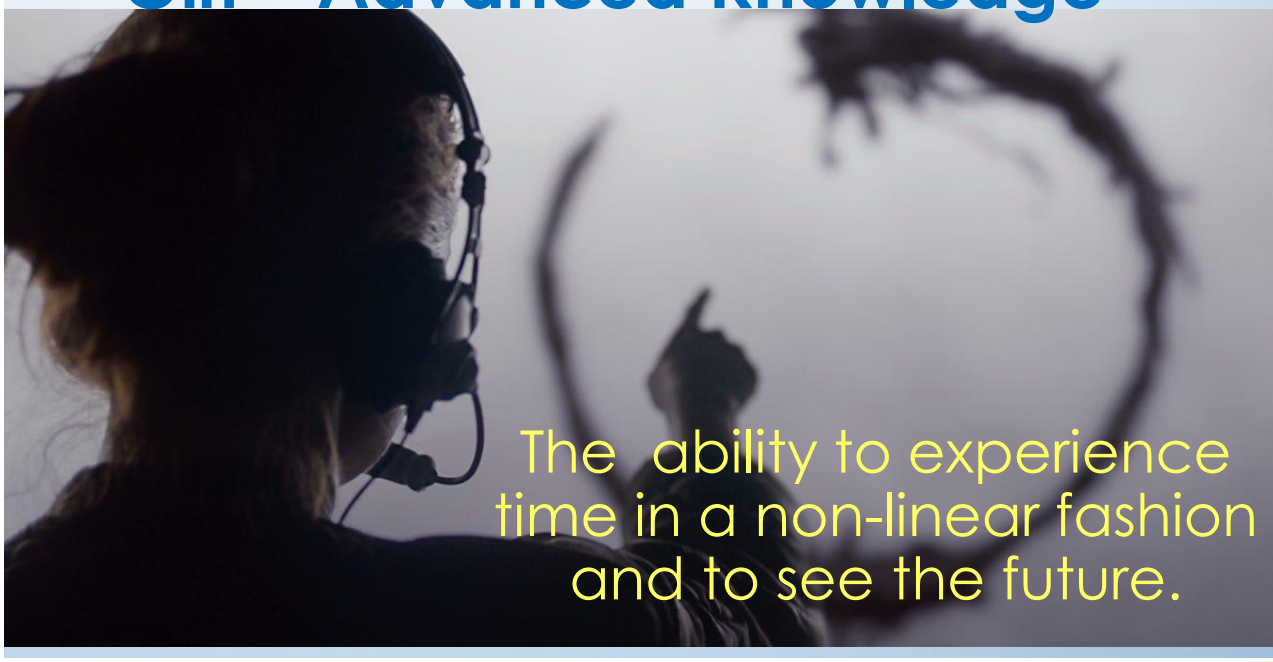
What is Gen AI for us?

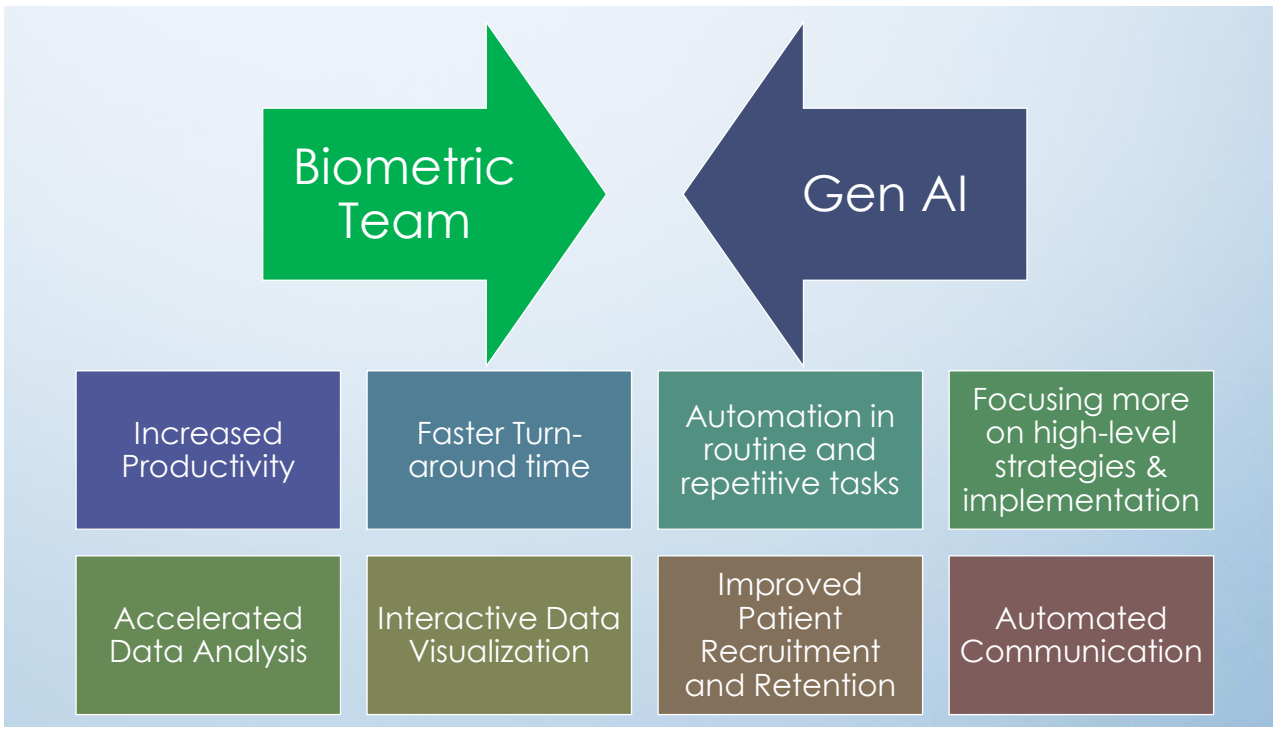
Replacing us?

Augmenting us?

266

Gift – Advanced Knowledge





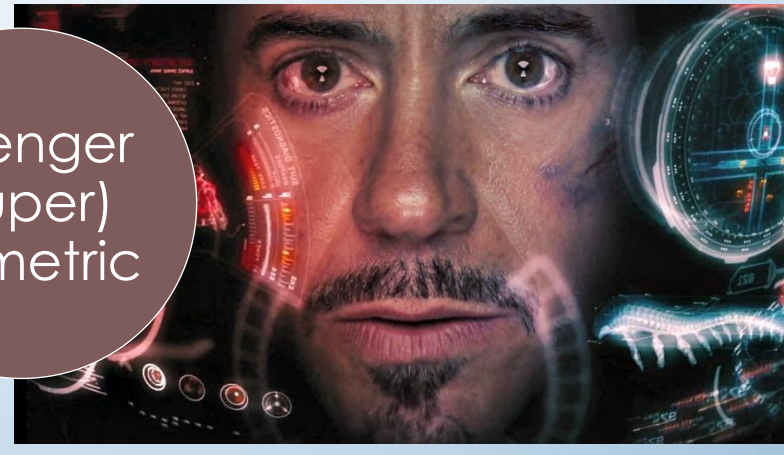
Future of Gen AI & Biometric team?

Gen AI

+

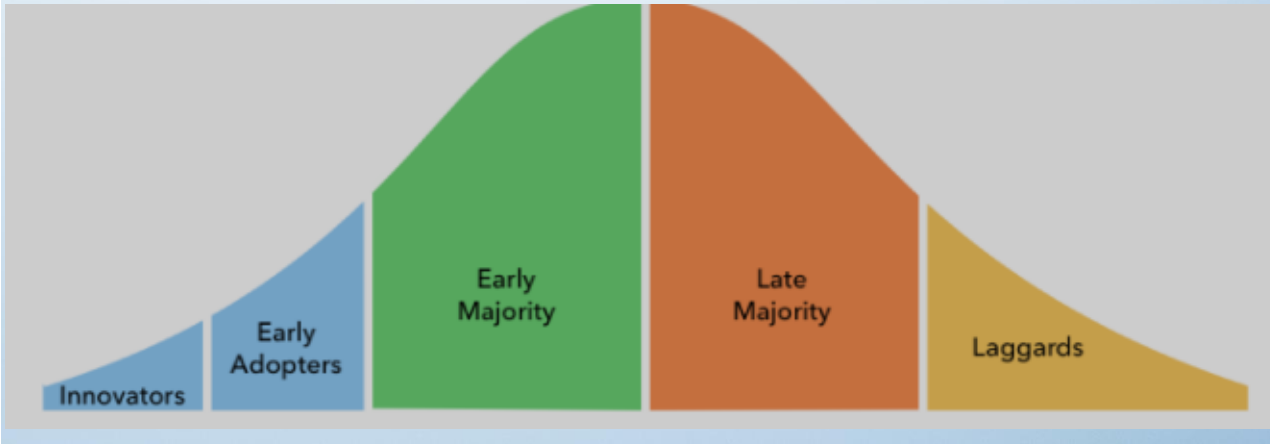
Biom etric

Avenger (Super) Biometric



Discussion :

Where do we want our Biometric team to be in Gen AI Adopter?

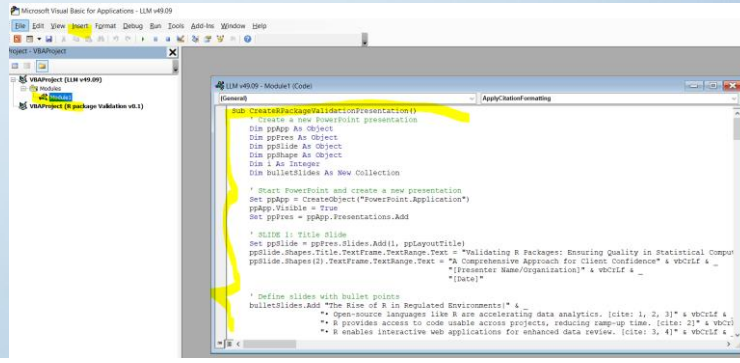


Q & A (Discussion)

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

Appendix

Confidential – Do not duplicate or distribute without the permission of Kevin Lee

How to create power-point using ChatGPT and VBA

- Gen AI
  - Prompt for the contents in ChatGPT, Gemini, Anthropic
  - Then, prompt for VBA codes for the contents (e.g., create VBA code to create the slides in power point using above contents). Copilot works best. Others gave a lot of wrong VBA.
- Open Power point application
  - Go to "Developer", "Visual Basic"
  - Select "Module" and "Insert"