

Single Day Event

Transforming from Statistical Programmer to Data Scientist for Clinical Trials in the GenAI Era

Hao(Harry) Chen,
BeiGene



DISCLAIMER

The views and opinions presented in this paper, along with the associated presentation and slides, are solely those of the presenter and do not necessarily represent those of BeiGene. This analysis is based entirely on publicly available data, and BeiGene does not ensure the accuracy or reliability of the information provided.



Agenda

Background



Data Science



GenAI Implementation



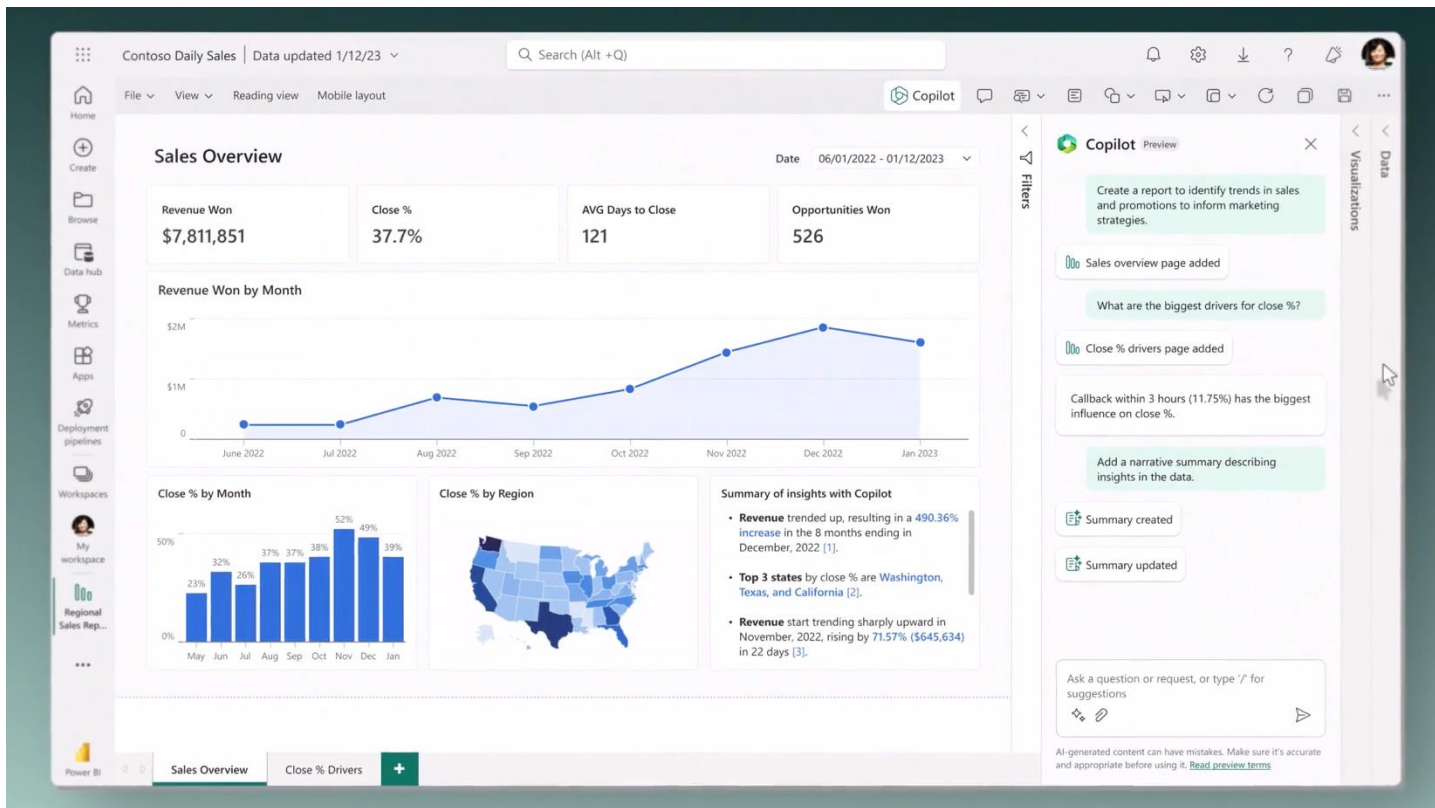
Case Demo



Summary

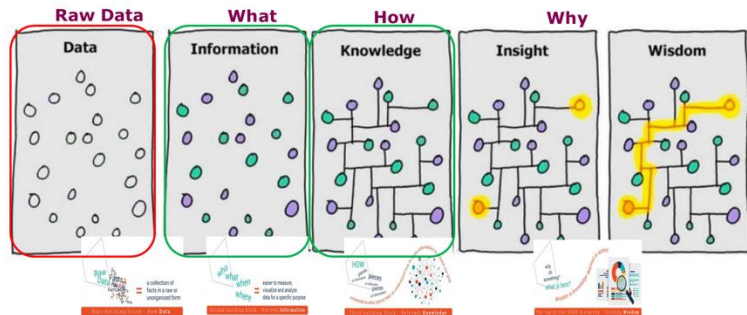


Accelerating the data to insight with AI



DIKW Model

(Data-to-Information-to-Knowledge-to-Wisdom)



Autonomous driving vehicle (ADV) example

Case Study: Going Straight at an Intersection – System Engineering Problem

Our solution: system level optimization.

Perception, prediction and planning modules need to work together.

- Perception: detect vehicles, pedestrians and cyclists in the intersection.
- Prediction: predict the future intention and trajectory of other vehicles, e.g. yield for ADV or pass in front of ADV.
- Planning: decide the right timing and trajectory to pass the intersection.



Predicted intentions

Pass in front of ADV

Yield for ADV

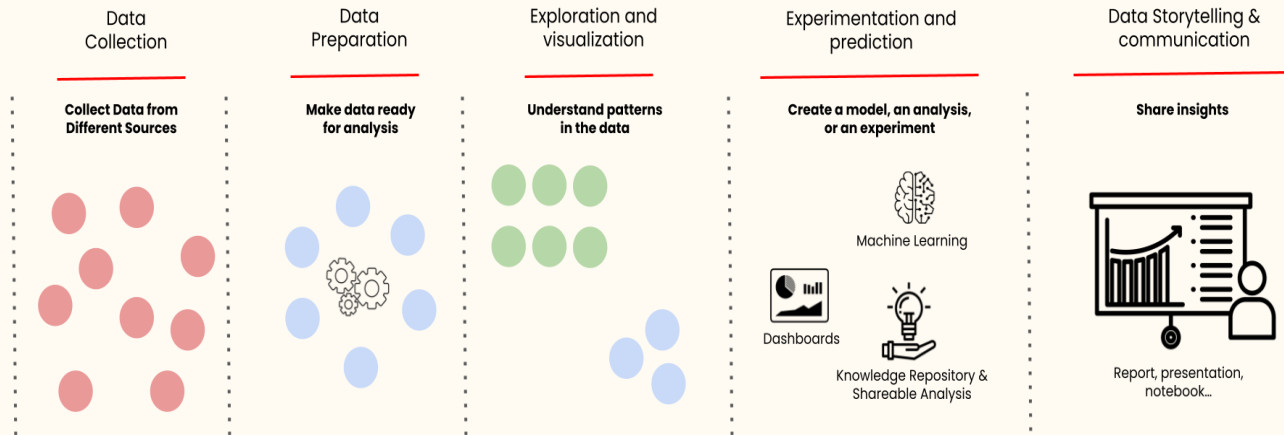
Caution to ADV

Crossing a complex multi-lane intersection in Beijing



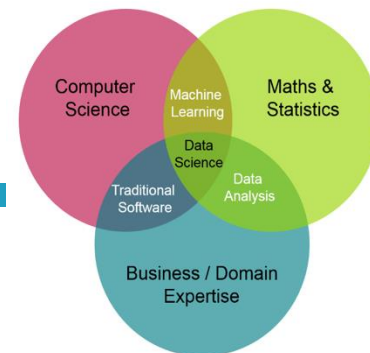
Data Science

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.





Data Scientist



Data Scientist

- Analyze data to provide actionable insights.
- Work across multiple teams in a business environment.
- Lay the foundations for robust analytics.
- Develop strategies to capture, gather, and clean data from various sources.
- Organize and explore data.
- Build solutions based on data analysis.
- Communicate findings to the wider business.

Data Analysts vs. Data Scientists

Data Analysts:

- Interpret existing data to provide actionable insights.
- Utilize simpler models and tools.

Data Scientists:

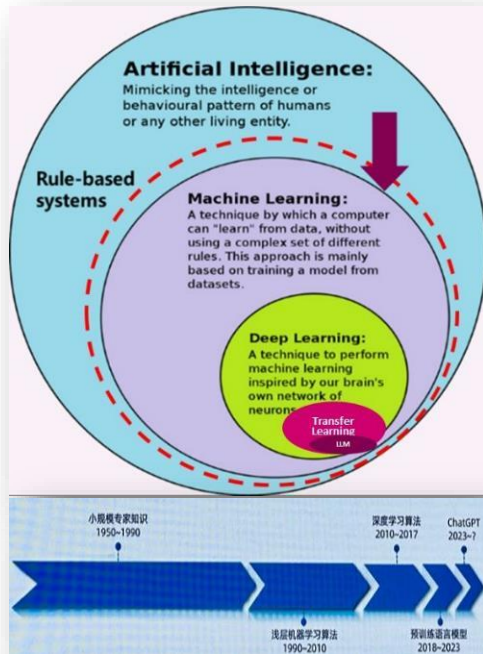
- Take on a broader role by formulating questions that need answering.
- Derive insights using advanced tools and models, including machine learning algorithms

Technical skills

- Python/R
- Statistics and math
- SQL and NoSQL
- Data visualization
- Machine learning
- Deep learning
- Natural language processing
- Big data
- Cloud computing



Democratized AI Technology + Easy Web Apps



How deep learning is a subset of machine learning and how machine learning is a subset of artificial intelligence (AI) https://om.wikiwand.com/en/deep_learning

Cloud Gallery Components Generative AI Community Docs Blog Sign in Sign up

A faster way to build and share data apps

Streamlit turns data scripts into shareable web apps in minutes.
All in pure Python. No front-end experience required.

[Try Streamlit now](#) [Deploy on Community Cloud \(it's free!\)](#)

```
1 import streamlit as st
2 import mymodel as m
3
4 st.write("Hi!")
5 # Sales model
6 Below are our sales predictions
7 for this customer.
8
9
10 window = st.slider("Forecast window (days)",
11 st.write(m.run(window=window)))
```

Sales model

Below are our sales predictions for this customer:

Forecast window (days) 42

0:25 / 6:42

Learn more: Why Generative AI + Streamlit are a perfect match

[Streamlit • A faster way to build and share data apps](#)

Easy web apps for data science without the compromises

No web development skills required

[Get started in R](#) [Get started in Python](#)

Here is a Shiny app

Shiny apps are easy to write. Let users interact with your data and your analysis, all with R or Python.

Python

```
1 # Import libraries
2 library(shiny)
3 library(ggplot2)
4 library(dplyr)
5 library(reshape2)
6
7 # Define UI elements
8 ui <- fluidPage(
9   titlePanel("Shiny App"),
10   textInput("x", "Enter a value for x"),
11   textInput("y", "Enter a value for y"),
12   textInput("z", "Enter a value for z"),
13   textInput("w", "Enter a value for w"),
14   textInput("v", "Enter a value for v"),
15   textInput("t", "Enter a value for t"),
16   textInput("f", "Enter a value for f"),
17   textInput("d", "Enter a value for d"),
18   textInput("s", "Enter a value for s"),
19   textInput("a", "Enter a value for a"),
20   textInput("n", "Enter a value for n"),
21   textInput("h", "Enter a value for h"),
22   textInput("e", "Enter a value for e"),
23   textInput("l", "Enter a value for l"),
24   textInput("o", "Enter a value for o"),
25   textInput("g", "Enter a value for g"),
26   textInput("i", "Enter a value for i"),
27   textInput("v", "Enter a value for v"),
28   textInput("w", "Enter a value for w"),
29   textInput("x", "Enter a value for x"),
30   textInput("y", "Enter a value for y"),
31   textInput("z", "Enter a value for z"),
32   textInput("w", "Enter a value for w"),
33   textInput("x", "Enter a value for x"),
34   textInput("y", "Enter a value for y"),
35   textInput("z", "Enter a value for z"),
36   textInput("w", "Enter a value for w"),
37   textInput("x", "Enter a value for x"),
38   textInput("y", "Enter a value for y"),
39   textInput("z", "Enter a value for z"),
40   textInput("w", "Enter a value for w"),
41   textInput("x", "Enter a value for x"),
42   textInput("y", "Enter a value for y"),
43   textInput("z", "Enter a value for z"),
44   textInput("w", "Enter a value for w"),
45   textInput("x", "Enter a value for x"),
46   textInput("y", "Enter a value for y"),
47   textInput("z", "Enter a value for z"),
48   textInput("w", "Enter a value for w"),
49   textInput("x", "Enter a value for x"),
50   textInput("y", "Enter a value for y"),
51   textInput("z", "Enter a value for z"),
52   textInput("w", "Enter a value for w"),
53   textInput("x", "Enter a value for x"),
54   textInput("y", "Enter a value for y"),
55   textInput("z", "Enter a value for z"),
56   textInput("w", "Enter a value for w"),
57   textInput("x", "Enter a value for x"),
58   textInput("y", "Enter a value for y"),
59   textInput("z", "Enter a value for z"),
60   textInput("w", "Enter a value for w"),
61   textInput("x", "Enter a value for x"),
62   textInput("y", "Enter a value for y"),
63   textInput("z", "Enter a value for z"),
64   textInput("w", "Enter a value for w"),
65   textInput("x", "Enter a value for x"),
66   textInput("y", "Enter a value for y"),
67   textInput("z", "Enter a value for z"),
68   textInput("w", "Enter a value for w"),
69   textInput("x", "Enter a value for x"),
70   textInput("y", "Enter a value for y"),
71   textInput("z", "Enter a value for z"),
72   textInput("w", "Enter a value for w"),
73   textInput("x", "Enter a value for x"),
74   textInput("y", "Enter a value for y"),
75   textInput("z", "Enter a value for z"),
76   textInput("w", "Enter a value for w"),
77   textInput("x", "Enter a value for x"),
78   textInput("y", "Enter a value for y"),
79   textInput("z", "Enter a value for z"),
80   textInput("w", "Enter a value for w"),
81   textInput("x", "Enter a value for x"),
82   textInput("y", "Enter a value for y"),
83   textInput("z", "Enter a value for z"),
84   textInput("w", "Enter a value for w"),
85   textInput("x", "Enter a value for x"),
86   textInput("y", "Enter a value for y"),
87   textInput("z", "Enter a value for z"),
88   textInput("w", "Enter a value for w"),
89   textInput("x", "Enter a value for x"),
90   textInput("y", "Enter a value for y"),
91   textInput("z", "Enter a value for z"),
92   textInput("w", "Enter a value for w"),
93   textInput("x", "Enter a value for x"),
94   textInput("y", "Enter a value for y"),
95   textInput("z", "Enter a value for z"),
96   textInput("w", "Enter a value for w"),
97   textInput("x", "Enter a value for x"),
98   textInput("y", "Enter a value for y"),
99   textInput("z", "Enter a value for z"),
100  textInput("w", "Enter a value for w"),
101  textInput("x", "Enter a value for x"),
102  textInput("y", "Enter a value for y"),
103  textInput("z", "Enter a value for z"),
104  textInput("w", "Enter a value for w"),
105  textInput("x", "Enter a value for x"),
106  textInput("y", "Enter a value for y"),
107  textInput("z", "Enter a value for z"),
108  textInput("w", "Enter a value for w"),
109  textInput("x", "Enter a value for x"),
110  textInput("y", "Enter a value for y"),
111  textInput("z", "Enter a value for z"),
112  textInput("w", "Enter a value for w"),
113  textInput("x", "Enter a value for x"),
114  textInput("y", "Enter a value for y"),
115  textInput("z", "Enter a value for z"),
116  textInput("w", "Enter a value for w"),
117  textInput("x", "Enter a value for x"),
118  textInput("y", "Enter a value for y"),
119  textInput("z", "Enter a value for z"),
120  textInput("w", "Enter a value for w"),
121  textInput("x", "Enter a value for x"),
122  textInput("y", "Enter a value for y"),
123  textInput("z", "Enter a value for z"),
124  textInput("w", "Enter a value for w"),
125  textInput("x", "Enter a value for x"),
126  textInput("y", "Enter a value for y"),
127  textInput("z", "Enter a value for z"),
128  textInput("w", "Enter a value for w"),
129  textInput("x", "Enter a value for x"),
130  textInput("y", "Enter a value for y"),
131  textInput("z", "Enter a value for z"),
132  textInput("w", "Enter a value for w"),
133  textInput("x", "Enter a value for x"),
134  textInput("y", "Enter a value for y"),
135  textInput("z", "Enter a value for z"),
136  textInput("w", "Enter a value for w"),
137  textInput("x", "Enter a value for x"),
138  textInput("y", "Enter a value for y"),
139  textInput("z", "Enter a value for z"),
140  textInput("w", "Enter a value for w"),
141  textInput("x", "Enter a value for x"),
142  textInput("y", "Enter a value for y"),
143  textInput("z", "Enter a value for z"),
144  textInput("w", "Enter a value for w"),
145  textInput("x", "Enter a value for x"),
146  textInput("y", "Enter a value for y"),
147  textInput("z", "Enter a value for z"),
148  textInput("w", "Enter a value for w"),
149  textInput("x", "Enter a value for x"),
150  textInput("y", "Enter a value for y"),
151  textInput("z", "Enter a value for z"),
152  textInput("w", "Enter a value for w"),
153  textInput("x", "Enter a value for x"),
154  textInput("y", "Enter a value for y"),
155  textInput("z", "Enter a value for z"),
156  textInput("w", "Enter a value for w"),
157  textInput("x", "Enter a value for x"),
158  textInput("y", "Enter a value for y"),
159  textInput("z", "Enter a value for z"),
160  textInput("w", "Enter a value for w"),
161  textInput("x", "Enter a value for x"),
162  textInput("y", "Enter a value for y"),
163  textInput("z", "Enter a value for z"),
164  textInput("w", "Enter a value for w"),
165  textInput("x", "Enter a value for x"),
166  textInput("y", "Enter a value for y"),
167  textInput("z", "Enter a value for z"),
168  textInput("w", "Enter a value for w"),
169  textInput("x", "Enter a value for x"),
170  textInput("y", "Enter a value for y"),
171  textInput("z", "Enter a value for z"),
172  textInput("w", "Enter a value for w"),
173  textInput("x", "Enter a value for x"),
174  textInput("y", "Enter a value for y"),
175  textInput("z", "Enter a value for z"),
176  textInput("w", "Enter a value for w"),
177  textInput("x", "Enter a value for x"),
178  textInput("y", "Enter a value for y"),
179  textInput("z", "Enter a value for z"),
180  textInput("w", "Enter a value for w"),
181  textInput("x", "Enter a value for x"),
182  textInput("y", "Enter a value for y"),
183  textInput("z", "Enter a value for z"),
184  textInput("w", "Enter a value for w"),
185  textInput("x", "Enter a value for x"),
186  textInput("y", "Enter a value for y"),
187  textInput("z", "Enter a value for z"),
188  textInput("w", "Enter a value for w"),
189  textInput("x", "Enter a value for x"),
190  textInput("y", "Enter a value for y"),
191  textInput("z", "Enter a value for z"),
192  textInput("w", "Enter a value for w"),
193  textInput("x", "Enter a value for x"),
194  textInput("y", "Enter a value for y"),
195  textInput("z", "Enter a value for z"),
196  textInput("w", "Enter a value for w"),
197  textInput("x", "Enter a value for x"),
198  textInput("y", "Enter a value for y"),
199  textInput("z", "Enter a value for z"),
200  textInput("w", "Enter a value for w"),
201  textInput("x", "Enter
```

GenAI Implementation





Using LLMs for Programing

Don't write code if the model can do it; the model will get better, but the code won't.

Trade leverage for precision; use interaction to mitigate.

Code is for syntax and process; models are for semantics and intent.

The system will be as brittle as its most brittle part.

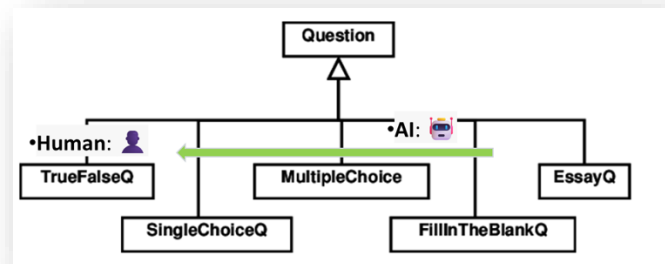
Ask Smart to Get Smart.

Uncertainty is an exception throw.

Text is the universal wire protocol.

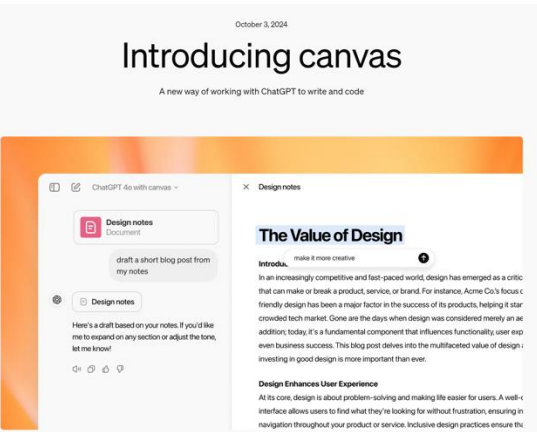
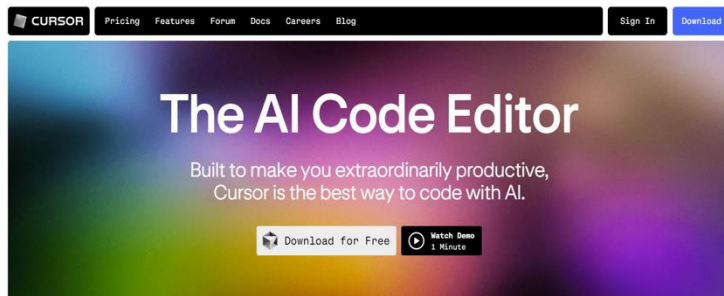
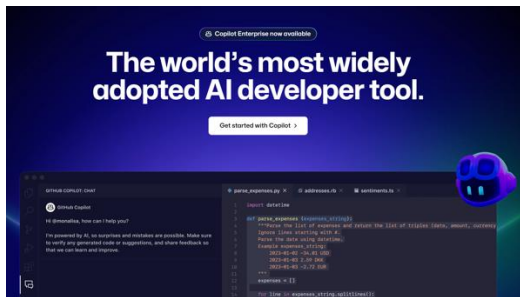
Hard for you is hard for the model.

Beware pareidolia of consciousness; the model can be used against itself.





Using LLMs for Programing



Shiny Assistant



Hello, I'm Shiny Assistant! I'm here to help you with [Shiny](#), a web framework for data driven apps. You can ask me questions about how to use Shiny, to explain how certain things work in Shiny, or even ask me to build a Shiny app for you.

Here are some examples:

- "How do I add a plot to an application?"
- "Create an app that shows a normal distribution."
- "Show me how make it so a table will update only after a button is clicked."
- Ask me, "Open the editor", then copy and paste your existing Shiny code into the editor, and then ask me to make changes to it.

Let's get started! 🚀

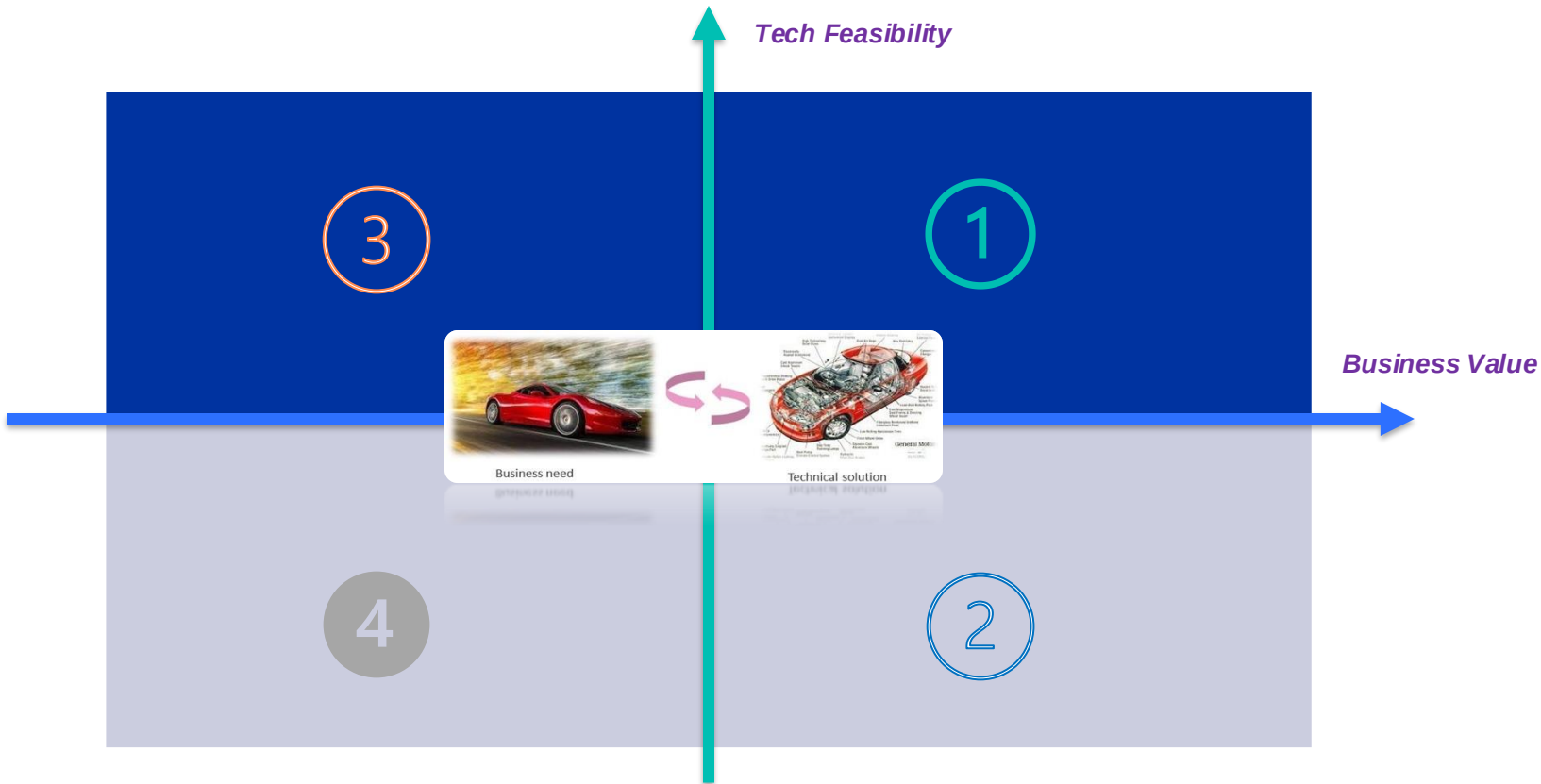
Who can see my activity? ⓘ

Enter a message...

[Shiny - Shiny Assistant \(posit.co\)](#)



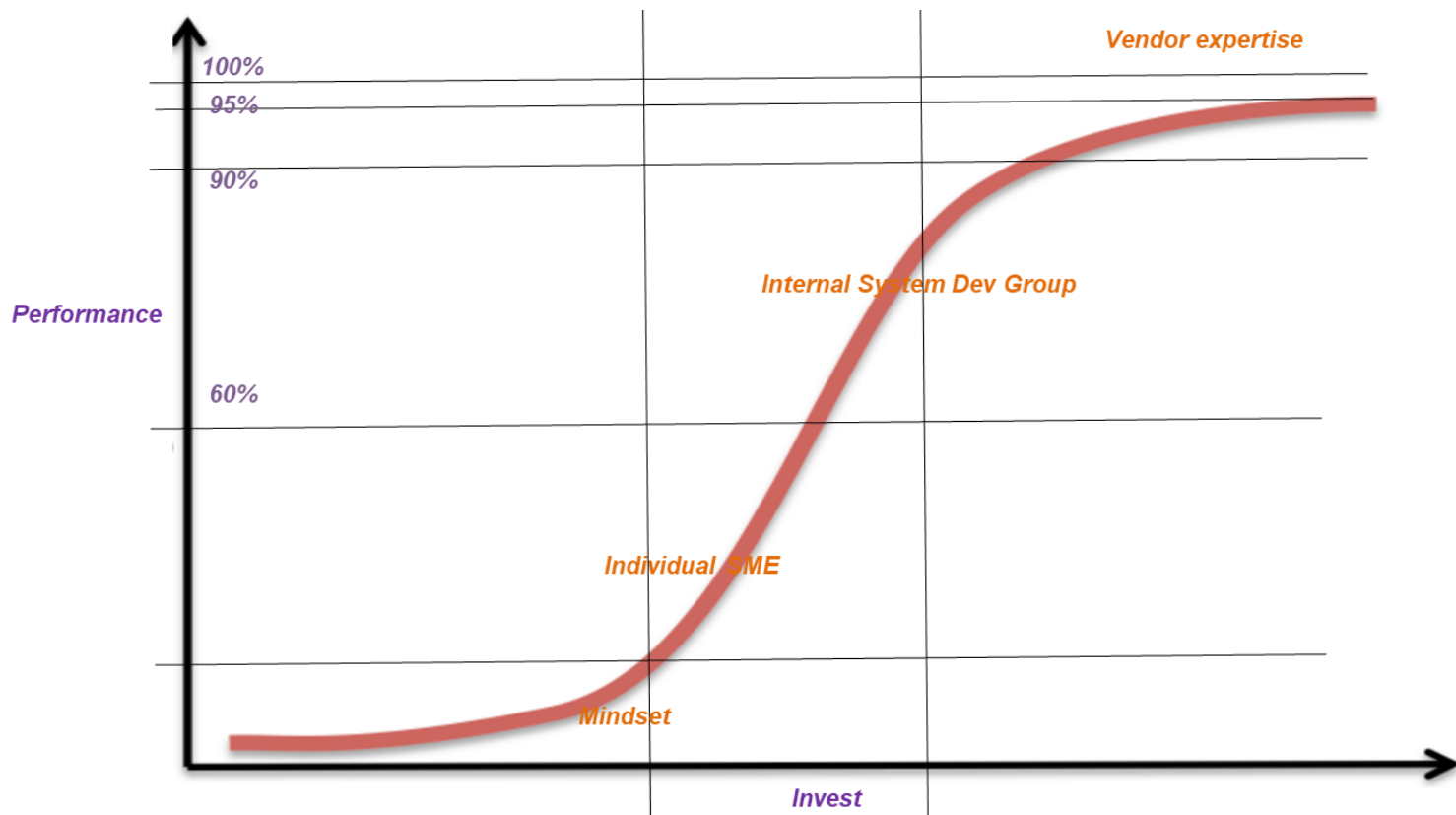
AI Implementation Scenarios





GenAI Implementation Stages

Clear Goal
Good Training Data



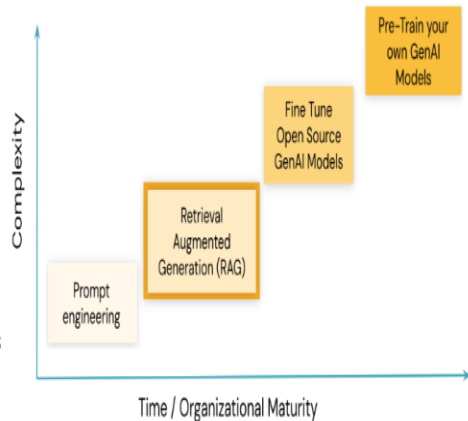


GenAI Deployment Approaches

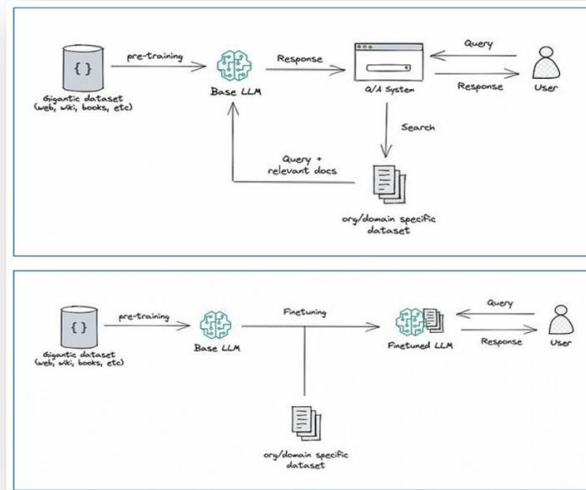
GenAI & Maturity curve

Deploying GenAI can be done in multiple ways:

- **Prompt engineering on public APIs** (e.g. Llama 2, openAI): answer from public information, retail (think ChatGPT)
- **Retrieval Augmented Generation (RAG)**: specialize your model with additional content. *This is what we'll focus on in this demo*
- **OSS model Fine tuning**: when you have a large corpus of custom data and need specific model behavior (execute a task)
- **Train your own LLM**: for full control on the underlying data sources of the model (biomedical, Code, Finance...)

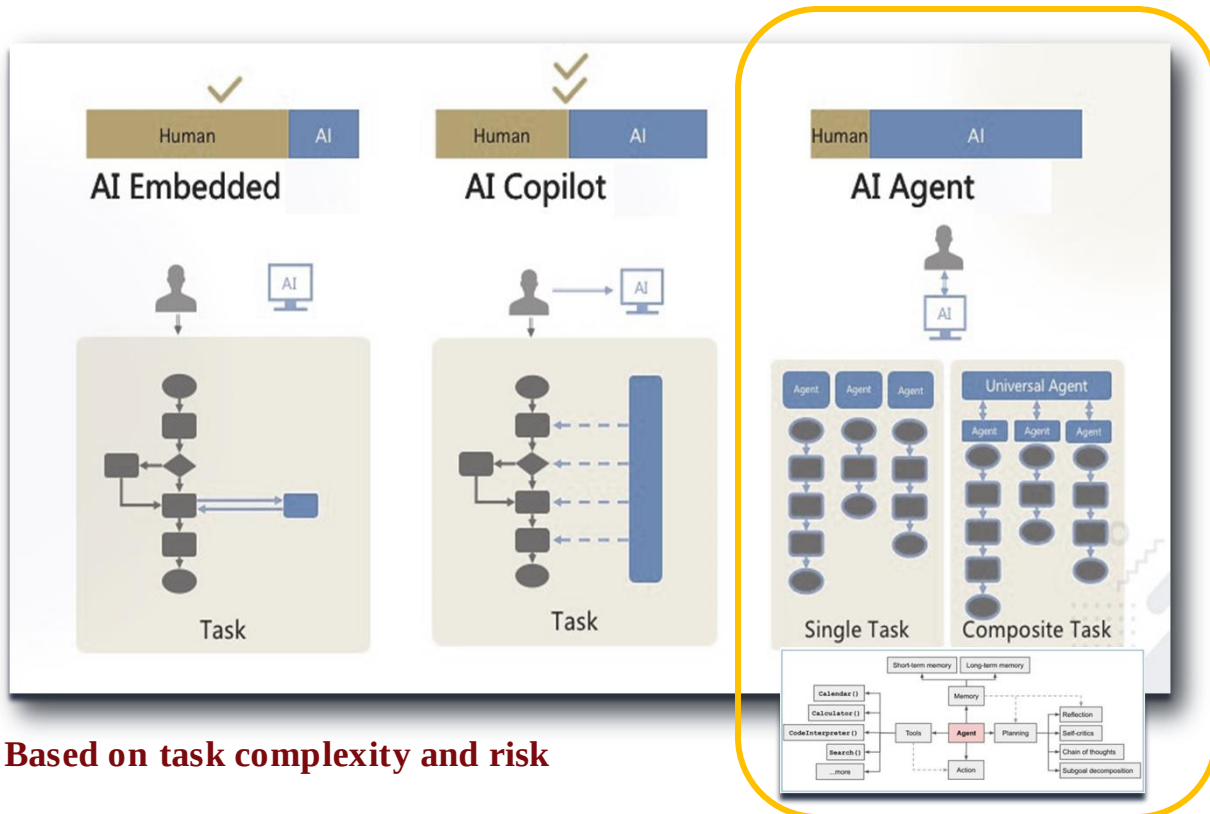


RAG (Retrieval-Augmented Generation) vs. FT (Fine-Tuning)





GenAI Implementation Mode



Based on task complexity and risk

AI Agent

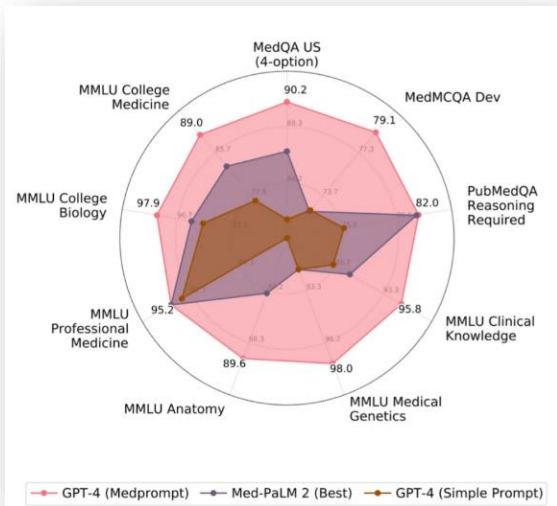
Types of Tasks

- *Single Task:**
 - Each agent is designed to handle a specific task independently.
- *Composite Task:**
 - Universal agents manage multiple agents to coordinate complex, multi-step tasks.

Functional Components

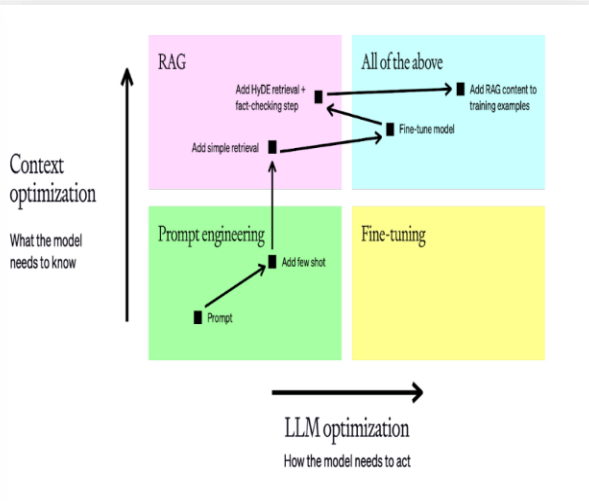
- *Tools:**
 - Agents utilize various tools such as Calendar, Calculator, Code Interpreter, and Search functions.
- *Memory:**
 - Short-term and long-term memory storage for efficient task execution.
- *Planning and Action:**
 - Agents use planning mechanisms to determine actions based on memory and tools.
- *Advanced Cognitive Functions:**
 - Reflection:** Allows agents to evaluate their actions.
 - Self-criticism:** Enables self-assessment and improvement.
 - Chain of Thoughts:** Supports logical reasoning and decision-making.
 - Subgoal Decomposition:** Breaks down complex tasks into manageable subgoals.

LLM Optimization Process



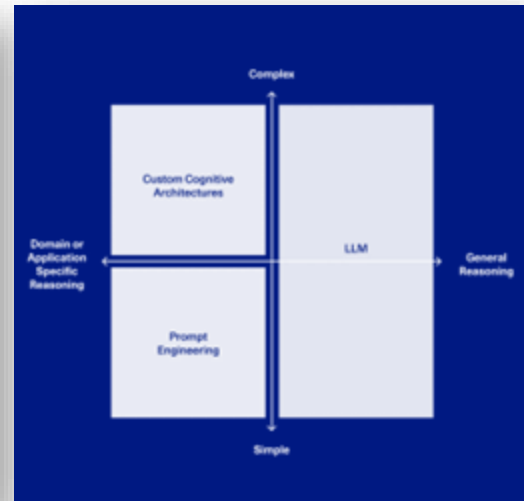
[The Power of Prompting](#)

No free lunch



[Optimizing LLM Accuracy](#)

Customization is needed



[Planning for Agents](#)

Task-specific agent needs custom cognitive architecture

#	Project	Language	Code	Domain Knowledge	Reasoning
1	Medical Coding	X		X	X
2	ChatPaper	X		X	X
3	ChatDataCopilot	X	X	X	X
4	SASCodeGenius	X	X	X	X
5	Translation	X		X	X

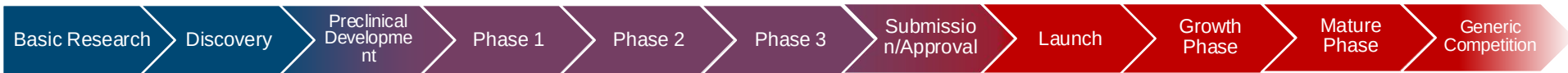


BeiGene GSDS SPIS Team

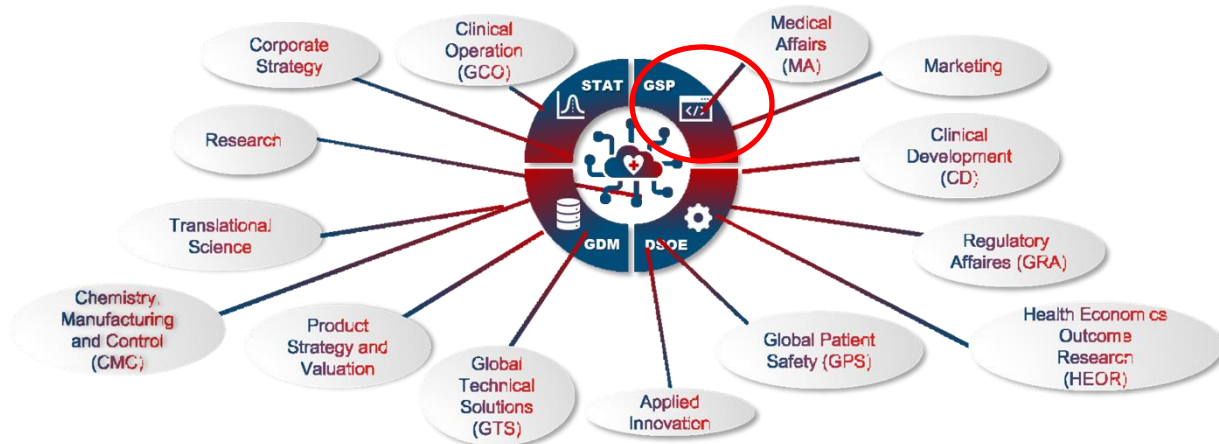
Research

Development

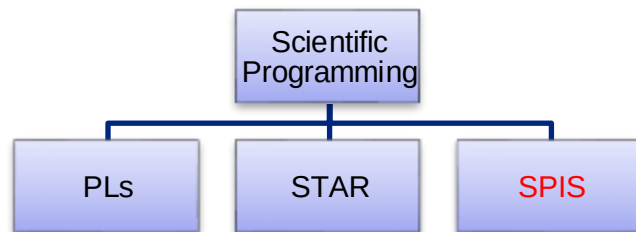
Commercialization



Global Statistics and Data Science (GSDS)



Scientific Programming Intelligent Solution (SPIS)





1. MedicalCoding - Chat Mode

Go back to Homepage

MedDRA Coding version 27.0

Chat Mode

Enter AE Verbatim Term (one line)

12TH NERVE PALSY

Lookup LLT from MedDRA

Top N Setting

1 200

Lookup LLT from Historical Data

Top N Setting

1 200

Chat with GPT

GPT Top N Setting

1 20

Model Selection

gpt4-turbo

Show prompt

GPT system knowledge

AI-Assisted Medical Coding with Large Language Models

Top Match LLT from MedDRA

	AE Verbatim Term	Matched LLT	Match Score	LLT Chinese	Match Score Rank
0	12TH NERVE PALSY	XIth nerve injury	0.8838	第XII颅神经损伤	1
1	12TH NERVE PALSY	XIth nerve paralysis	0.8745	第XI神经性瘫痪	2
2	12TH NERVE PALSY	VIth nerve palsy	0.8733	第六脑神经麻痹	3
3	12TH NERVE PALSY	VIth nerve palsy	0.873	第VII神经麻痹	4
4	12TH NERVE PALSY	VIth nerve paresis	0.8729	第六神经轻瘫	5
5	12TH NERVE PALSY	IVth nerve palsy	0.8728	第IV颅神经麻痹	6
6	12TH NERVE PALSY	Congenital paralysis c	0.8688	先天性第6和第12对脑神经麻痹	7
7	12TH NERVE PALSY	IVth nerve paresis	0.8672	第IV神经轻瘫	8
8	12TH NERVE PALSY	VIth nerve paralysis	0.8654	第VII神经性瘫痪	9
9	12TH NERVE PALSY	IXth nerve paralysis	0.8653	第IX颅神经性瘫痪	10

Chat Mode is enabled, please chat with GPT model to further discuss about the coding. Feel free to ask any questions.

S

You are the MedDRA coding expert. User will provide the verbatim term for a symptom, sign, disease, diagnosis, therapeutic indication, investigation, surgical, or medical procedure, and medical, social, or family history characteristic. Think step by step. Your work is very important to patients and doctors. According to the medical concept, please select the most precise matched term from below candidate Lowest Level Term following the MedDRA coding principle.

S

The verbatim adverse event term I need to code is: 12TH NERVE PALSY. The 100 candidate MedDRA LLT terms are as follows:



To accurately code the verbatim term "12TH NERVE PALSY," we need to identify the corresponding cranial nerve by its Roman numeral and the correct medical term for paralysis or palsy. The 12th cranial nerve is known as the hypoglossal nerve, which is responsible for controlling the muscles of the tongue.

Here are the top 5 most relevant coding terms from the provided list, ranked in order of relevance:

- [56] **Hypoglossal nerve paralysis** - This term directly corresponds to paralysis of the 12th cranial nerve, which is synonymous with "12TH NERVE PALSY." The term "paralysis" is often used interchangeably with "palsy" in medical terminology, making this the most accurate match.
- [43] **Hypoglossal nerve paresis** - This term refers to a partial paralysis or weakness of the hypoglossal nerve. While "paresis" indicates a lesser degree of severity compared to "paralysis," it is still highly relevant as it directly involves the 12th nerve.
- [32] **Cranial nerve palsy** - This term is less specific than the first two but still relevant as it refers to palsy of any cranial nerve. Since the 12th nerve is a cranial nerve, this term is applicable but lacks the specificity of directly mentioning the hypoglossal nerve.
- [51] **Cranial nerve paralysis** - Similar to the previous term, this is a general term that covers paralysis of any cranial nerve. It is relevant but less specific than terms that directly mention the hypoglossal nerve.
- [31] **Multiple cranial nerve palsies** - This term is included as a relevant option because it indicates palsy affecting more than one cranial nerve. Although it is less specific and does not exclusively refer to the 12th nerve, it remains pertinent in a broader sense.

Summary and Ranking Results:

- [56] Hypoglossal nerve paralysis
- [43] Hypoglossal nerve paresis
- [32] Cranial nerve palsy
- [51] Cranial nerve paralysis
- [31] Multiple cranial nerve palsies



Chat Mode Demo

[Go back to Homepage](#)

MedDRA Coding version
27.0

☒ Chat Mode

Enter AE Verbatim Term (one line)

☐ Lookup LLT from MedDRA

Top N Setting

1 100 200

☐ Lookup LLT from Historical Data

Top N Setting

1 100 200

Choose an option

Chat with GPT

GPT Top N Setting

1 5 20

Model Selection

gpt4o

☐ Show prompt

AI-Assisted Medical Coding with Large Language Models



1. MedicalCoding - Batch Mode

Go back to Homepage

MedDRA Coding version 27.0

☒ Batch Mode

Upload file

Drag and drop file here
Limit 200MB per file • XLSX

Browse files

test_3obs.xlsx
8.5KB

☒ Lookup LLT from MedDRA

Top N Setting
50
1 200

☐ Lookup LLT from Historical Data

Top N Setting
50
1 200

☒ Chat with GPT

GPT Top N Setting
5
1 20

Model Selection
llama-3-70b-instruct

AI-Assisted Medical Coding with Large Language Models

Top Match LLT from MedDRA

	AE Verbatim Term	Matched LLT	Match Score	LLT Chinese	Match Score Rank
0	bruising easily	Tendency to bruise easily	0.9411	容易出现瘀青	1
1	bruising easily	Bruising	0.9116	瘀伤	2
2	bruising easily	Increased tendency to bruise	0.8989	瘀青趋向增加	3
3	bruising easily	Bruise	0.8924	瘀青	4
4	bruising easily	Bruising of face	0.8821	面部瘀伤	5
5	bruising easily	Sweating easily	0.8803	容易出汗	6
6	bruising easily	Bruising of leg	0.8769	腿部瘀伤	7
7	bruising easily	Soft tissue bruising	0.8742	软组织瘀伤	
8	bruising easily	Bruising of arm	0.8709	手臂瘀伤	
9	bruising easily	Bruising of foot	0.8693	脚部瘀伤	

Detailed coding logic

After carefully reviewing the 50 MedDRA LLT terms, I have selected the top 5 most relevant coding terms for the adverse event term "bruising easily". Here are my choices, along with their original index numbers and explanations:

****Rank 1: [1] Tendency to bruise easily****

This term is an exact match for the verbatim term "bruising easily". It accurately captures the concept of having a propensity to bruise easily, which is the core meaning of the original term.

****Rank 2: [13] Spontaneous bruising****

Although not an exact match, "spontaneous bruising" implies that bruising occurs easily or without a specific cause, which is closely related to the concept of "bruising easily". This term is a strong contender for coding the adverse event.

****Rank 3: [3] Increased tendency to bruise****

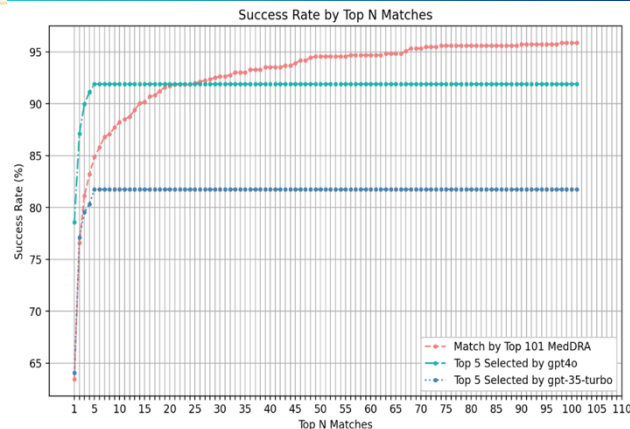
This term is similar to the top-ranked term, but with a slightly different wording. It still conveys the idea of having a higher likelihood of bruising, which aligns with the original term.

****Rank 4: [22] Bleeding tendency****

While not specifically focused on bruising, "bleeding tendency" is a related concept that implies an increased likelihood of bleeding or bruising. This term is a reasonable coding option, although it's not as precise as the top three choices.

****Rank 5: [2] Bruising****

Although "bruising" is a more general term than the others, it's still a relevant coding option. It doesn't specifically convey the



2000 AE codings

- AI: ~30 mins
- Human: ~10 days



Batch Mode Demo

MedDRA Coding version
27.0

☒ Batch Mode

Upload file

Drag and drop file here
Limit: 200MB per file • XLSX

[Browse files](#)

☐ Lookup LLT from MedDRA

Top N Setting
1 50 200

☐ Lookup LLT from Historical Data

Top N Setting
1 50 200

Choose an option
Chat with GPT

GPT Top N Setting
1 5 20

Model Selection
gpt-35-turbo

AI-Assisted Medical Coding with Large Language Models

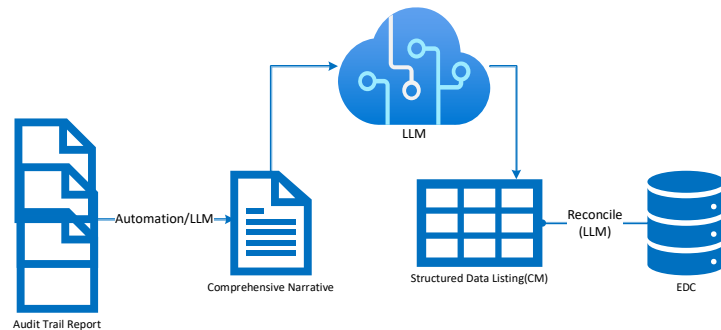
2. ChatPaper

Data Mining from Literatures

The screenshot shows a document viewer interface. On the left, there is a 'Page Thumbnails' panel displaying a grid of 9 thumbnails from a document. The main area on the right shows the full text of a scientific article. The article is titled 'Impaired histone inheritance promotes tumor progression' and is published in 'nature communications'. The article's abstract discusses the role of histone inheritance in tumor progression, mentioning that impaired histone inheritance leads to genomic instability and tumor progression. The article is dated 20 October 2022 and has been accepted for publication.

	Animal information				Cell/culture information				
	Strain	Gender	Age	Supplier	Cell	Origin	Implanted Cell Number	Implanted Volume (μl)	Estrogen label
8485.2	NOD	Female	6	JSUC	Barcoded HT-1080	N/A	5-100	50	IS-12 implanted 100-day release, Research of 1
	athymic nude mice	Female	5-6	University animal Center of the General Hospital of Shijiazhuang, Hebei, China	HT-1080	N/A	5-100	100	N/A
	nude	Female	7-9	SHALSCO Service Co. Ltd	MCF-7	ATCC	1-100	N/A	N/A
8133.03	nude	Female	6	Kunming Medical University	MCF-7	N/A	1-100	100	N/A
8422.8	NOD-GEM	Female	7	N/A	MCF-7	N/A	5-100	50	N/A
276	nude	Female	6-8	N/A	MCF-7	ATCC	1-100	N/A	E20
	nude	Female	4-5	N/A	MCF-7	N/A	1 × 10 ⁵	100	N/A
8868.13	nude	Female	10-20g	Shanghai Laboratory Animal Co. Ltd (Shanghai, China)	MCF-7	Shanghai Biotechnology Co. Ltd	1-100-1 × 10 ⁵	200	N/A
8757.2	athymic balb/c	Female	8-10day old	N/A	MCF-7	N/A	3-100	100	UPB and manged (PUB) = 10%

Smart Narrative Review



2. ChatPaper - Data Mining from Literatures

chatDoc

checkDiff

PDF parser

Show Advanced Options

File Upload Section

Choose file category:

Text Files

Data Files

Image Files

Choose text file type:

PDF

Upload your PDF Files

Drag and drop files here

Limit 200MB per file - PDF

Browse files

s41598-022...

1.9MB

s10549-021...

1.0MB

ddd1-1819...

5.9MB

Scientific Paper Q&A Assistant

Design the Q&A Form

Hierarchical Category-Name-Value Form

I have a template file to populate the form

Animal Info

Category Name

Animal Info

Name

Animal Strain

Sample Value

athymic nude mice

Remove Field

Age

Sample Value

3-6 weeks old

Remove Field

Animal Vendor

Sample Value

VR

Remove Field

Add Field

Remove Category

Cell&Isolation Information

Category Name

Cell&Isolation Information

Name

Cell

Sample Value

MCF7

Remove Field

Cell Origin or vendor

Sample Value

ATCC(HTB-22)

Remove Field

Implanted Volume (ul)

Sample Value

100

Remove Field

Add Field

Remove Category

Select a question

show info specified in the following YAML, but only include the name/value pairs in the J...

Ask a Question from the uploaded files...

show info specified in the following YAML, but only include the name/value pairs in the JSON output

Submit & Process

Processing your question...

Step 1: Read Files and Show the Answer

Step 2: Get the Intermediate Summary Data

Final Structured Summary Data

File Name	Animal Strain	Age	Animal Vendor
ddd1-181939-baicalein-induces-apoptosis-and-autophagy-of-breast-cancer-c-111518.pdf	athymic nude mice	3-6 weeks old	Experimental Animal Center of Xian Jiaotong University
s10549-021-06345-x.pdf	athymic nude mice	6-8 weeks	VR
s41598-022-26767-x.pdf	athymic balb/c nude mice	5-6 week old	Beijing Vital River Laboratory Animal Technology Co., Ltd.

Download the final report data

Total processing time: 1.10 minutes for 3 files/rows.

File 1 of 3: ddd1-181939-baicalein-induces-apoptosis-and-autophagy-of-breast-cancer-c-111518.pdf

Detailed Explanation

To address the question, we need to extract specific information from the provided context and match it to the requested YAML structure. Here's a step-by-step breakdown:

- Animal Info:**
 - Animal Strain: The context mentions "Balb/c nude mice", which are a type of athymic nude mice.
 - Age: The age range provided is "3-6 weeks old".
 - Animal Vendor: The vendor is "Experimental Animal Center of Xian Jiaotong University".
- Cell & Isolation Information:**
 - Cell: The context specifies "MCF-7" and "MDA-MB-231" breast cancer cells. For this summary, we will use "MCF-7".
 - Cell Origin or Vendor: The context does not provide specific information about the origin or vendor of the MCF-7 cells.
 - Implanted Volume (ul): The context states that MCF-7 cells were injected at a volume of "100 μ L".

Summary in Markdown Table

Animal Strain	Age	Animal Vendor	Cell	Cell Origin or Vendor	Implanted Volume (ul)
athymic nude mice	3-6 weeks old	Experimental Animal Center of Xian Jiaotong University	MCF-7	N/A	100 μ L

Source Texts

- Animal Strain and Age:**
 - "Balb/c nude mice (3-6 weeks old, body weight 18-20g) were obtained from the Experimental Animal Center of Xian Jiaotong University."
- Animal Vendor:**
 - "Balb/c nude mice (3-6 weeks old) were purchased from Experimental Animal Center of Xian Jiaotong University."
- Cell:**
 - "MCF-7 and MDA-MB-231 breast cancer cells were cultured in DMEM supplemented with 10% FBS, or 10% fetal bovine serum, and 40 ng/ml epidermal growth factor in humidified 5% CO₂ at 37°C."
- Implanted Volume:**
 - "Tumors were established by subcutaneous injection of 10⁶ cells cultured MCF-7 (100 μ L) and MDA-MB-231 (100 μ L) into nude mice."

Answer

Detailed Explanation To address the question, we need to extract specific information from the provided context and match it to the requested YAML structure. Here's a step-by-step breakdown:

Detailed Explanation ### Animal Information 1. **Animal Strain**: The study used "athymic balb/c nude mice."

Detailed Explanation

Animal Information
 1. **Animal Strain**: The study used "athymic balb/c nude mice."
 2. **Age**: The mice were "3-6 weeks old."
 3. **Animal Vendor**: The mice were provided by "Beijing Vital River Laboratory Animal Technology Co., Ltd."

Cell & Isolation Information
 1. **Cell**: The cell lines used were "MCF-7 and MCF-7/ADR."
 2. **Cell Origin or Vendor**: The specific origin or vendor for the MCF-7 cells is not explicitly mentioned in the provided context.
 3. **Implanted Volume (ul)**: The implanted volume is not specified in the provided context.

The Power of R

- Great Language for Data Analysis & Reporting

- Mature platform for data analysis and reporting developed by Posit@.
- User-friendly syntax, closely resembling natural language (e.g., Tidyverse/ggplot2)
- Widely used in the pharmaceutical industry with a thriving ecosystem.

- Open-Source

- Extensive codebase allows AI to learn and enhance capabilities.
- Enables automation of R code comprehension and generation through AI.
- Potential for developing sophisticated AI-powered R applications.

- R Shiny UI

- Superior UI framework for improved human-computer interaction.
- Integrates AI with human intelligence (AI+HI) seamlessly.
- Facilitates real-time data analysis and reporting using R in the AI+HI context.





3. ChatDataCopilot

Data Query and Listing/Narrative

ARM	USUBJID	AGE	SEX	COUNTRY	RACE	oldest_text	
1	Placebo	01-710-1083	89	F	USA	WHITE	For Placebo treatment group: subject 01-710-1083, age 89, is the oldest patient. She resides in USA and is of WHITE race
2	Screen Failure	01-705-1058	89	M	USA	WHITE	For Screen Failure treatment group: subject 01-705-1058, age 89, is the oldest patient. He resides in USA and is of WHITE race
3	Screen Failure	01-710-1376	89	F	USA	WHITE	For Screen Failure treatment group: subject 01-710-1376, age 89, is the oldest patient. She resides in USA and is of WHITE race
4	Xanomeline High Dose	01-703-1295	88	F	USA	WHITE	For Xanomeline High Dose treatment group: subject 01-703-1295, age 88, is the oldest patient. She resides in USA and is of WHITE race
5	Xanomeline Low Dose	01-710-1002	88	M	USA	WHITE	For Xanomeline Low Dose treatment group: subject 01-710-1002, age 88, is the oldest patient. He resides in USA and is of WHITE race

Showing 1 to 5 of 5 entries

```
3 oldest_patients <- dm %>%
4   group_by(ARM) %>%
5   filter(AGE == max(AGE)) %>%
6   arrange(ARM, by.group = TRUE) %>%
7   select(ARM, USUBJID, AGE, SEX, COUNTRY, RACE)
8   mutate(GENDER = ifelse(SEX == "M", "He", "She"))
9   mutate(oldest_text = paste("For", ARM, "treat",
10                             "subject", USUBJID,
11                             "is the oldest pat",
12                             "GENDER, "resides i",
13                             "subject", USUBJID,
14                             "is the oldest pat",
15                             "GENDER, "resides i"))
16 select(-GENDER)
```

Token Consumption:
Prompt: 800 Completion: 171 Total: 971

Type your request in the box

write a narrative text for oldest patient in each treatment group:

For xxx treatment group: subject xxx, age xxx, is the oldest patient, He/She resides in xxx country, and is of xxx race

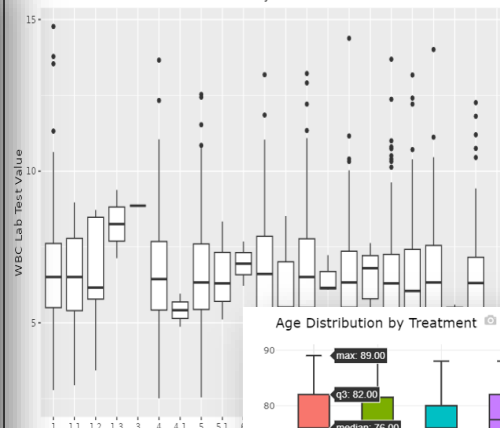
note:
1) first find the right patient (pay attention to ties), then create the narrative text, ordered by each group, and show the output table.
2) decide He/She based on gender F/M
3) no hard code please

☐ Continue

Show Code
Run
Download Output Table

Data Visualization

Distribution of WBC Lab Test Values by Visit Number



Generated Code

```
1 library(tidyverse)
2 library(ggplot2)
3
4 lbhe %>%
5   filter(LBTESTCD == 'WBC') %>%
6   ggplot(aes(x = factor(VISITNUM), y =
7     geom_boxplot() +
8     labs(x = "Visit Number", y = "WBC Le
9
10
```

Token Consumption:

Prompt: 1415 Completion: 92 Total: 1507

Type your request in the box

Generated Code

```
14 y = "Age") +
15   theme_minimal()
16
17 # Convert the ggplot2 object to a plotly
18 plotly_object <- ggplotly(box_plot, tool
19 # Print the plotly object
20 print(plotly_object)
21
22
23
```

Token Consumption:

Prompt: 942 Completion: 244 Total: 1186

Type your request in the box

age distribution by treatment, box plot

☒ Continue

Add more request/feedback

using plotly



3. ChatDataCopilot

The interface shows a terminal window with the following R code output:

```
##           Placebo   Screen Failure   Xanomeline High Dose   Xanomeline Low Dose
## (N=86)         (N=52)         (N=84)         (N=84)
## Age (yr)
## n           86           52           84           84
## Mean (SD)    75.2 (8.6)    75.1 (9.7)    74.4 (7.9)    75.7 (8.3)
## Median       76.0         76.0         76.0         77.5
## Min - Max    52.0 - 89.0  50.0 - 89.0  56.0 - 88.0  51.0 - 88.0
## Age groups
## n           86           52           84           84
## <18          0           0           0           0
## 18-64        14 (16.3%)  9 (17.3%)  11 (13.1%)  11 (13.1%)
## >=65         72 (83.7%)  43 (82.7%)  73 (86.9%)  73 (86.9%)
## Sex
## n           86           52           84           84
## Female       53 (61.6%)  36 (69.2%)  40 (47.6%)  40 (47.6%)
## Male         33 (38.4%)  16 (30.8%)  44 (52.4%)  44 (52.4%)
## Race
## n           86           52           84           84
## AMERICAN INDIAN OR ALASKA NATIVE  0           1 (1.9%)  1 (1.2%)  1 (1.2%)
## ASIAN          0           2 (3.8%)  0           0
## BLACK OR AFRICAN AMERICAN  8 (9.3%)  6 (11.5%)  9 (10.7%)  9 (10.7%)
## WHITE         78 (90.7%)  43 (82.7%)  74 (88.1%)  74 (88.1%)
```

The sidebar shows the following settings:

- Model: gpt-35-turbo
- Temperature: 0
- maxToken: 400
- Choose your option: ☒ R, ☐ SAS, ☐ VBA, ☐ PYTHON, ☐ Free Chat
- ☐ Translate to SAS
- Upload EXCEL/XPT/CSV/SAS7BDAT Data: dm.csv
- ☐ Show value of 1st row (Note: Do not do this for patient data!)
- Select Data: ☒ dm
- Upload an image as mock shell: DMshell.jpg

Table Understanding and Generation – from Shell Image

Characteristic	Placebo, N = 86	Screen Failure, N = 52	Xanomeline High Dose, N = 84	Xanomeline Low Dose, N = 84
AGE	75 ± 9	75 ± 10	74 ± 8	76 ± 8
age_group				
>64	72 (84%)	43 (83%)	73 (87%)	76 (90%)
18-64	14 (16%)	9 (17%)	11 (13%)	8 (9.5%)
SEX				
F	53 (62%)	36 (69%)	40 (48%)	50 (60%)
M	33 (38%)	16 (31%)	44 (52%)	34 (40%)
RACE				
AMERICAN INDIAN OR ALASKA NATIVE	0 (0%)	1 (1.9%)	1 (1.2%)	0 (0%)
ASIAN	0 (0%)	2 (3.8%)	0 (0%)	0 (0%)
BLACK OR AFRICAN AMERICAN	8 (9.3%)	6 (12%)	9 (11%)	6 (7.1%)
WHITE	78 (91%)	43 (83%)	74 (88%)	78 (93%)
* Mean ± SD; n (%)				

Generated Code

```
5 dm %>%
6 filter(arm %in% c("Placebo", "Screen Failure", "Xanomeline High Dose", "Xanomeline Low Dose"))
7 mutate(age_group = case_when(
8   AGE < 18 ~ "<18",
9   AGE >= 18 & AGE <= 64 ~ "18-64",
10  AGE > 64 ~ ">64") %>%
11  summarise(n = n(), p = 100 * n() / sum(n)))
12 select(arm, age, age_group, SEX, RACE) %>%
13 tbl_summary(
14   by = age_group,
15   ~>%
16     summarise(Mean = mean(age), SD = sd(age), n = n())
17 )
```

Token Consumption:
Prompt: 924 Completion: 170 Total: 1094

Type your request in the box

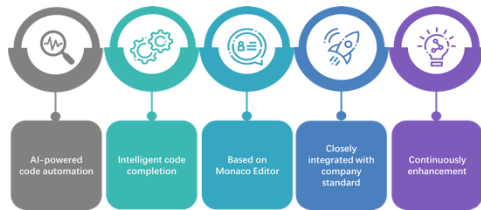
Create a statistical table comparing different groups in a clinical study, including "Placebo," "Screen Failure," "Xanomeline High Dose," and "Xanomeline Low Dose." The table should include data on age (mean, median, minimum, maximum, and age groups), sex, and race. Age groups should be <18, 18-64, and >65. Sex should be categorized into female and male, and race should include American Indian or Alaska Native, Asian, Black or African American, and White. Each category should display the number of participants (n) and the percentage in parentheses. Mean age should be presented with standard deviation (SD), and percentages should be used for the categorical variables.

Here is the template code for your reference:
Load the gtssummary package for data summarization and reporting
library(gtssummary)

Summarize the clinical trial data using tbl_summary function
Include the variables age, grade, and response in the summary table
tbl1 <-
tbl_summary(include = c(age, grade, response))



4. SASCodeGenius



The screenshot displays the Integrated SAS Programming Platform (Beta) interface. The top navigation bar includes the ChatBot tab, AI tab, Editor setting, AI Menu, and Search, Compare, Save options. The main content area shows the SAS code editor with the following code:

```
109 *** PROCESS
110 *****
111 *****
112 %m_arc_summ(data = adsl_tot
113             ,out = sum1
114             ,vars = age [ label = "Age (years)" /*digit = _auto_*/ ]
115             ,agegrln[cat label = "Age Group, n (%)" format = agegrln. preloadfmt=Y ]
116             ,sexn [cat label = "Sex, n (%)" format = sexn. preloadfmt=Y ]
117             ,racen [cat label = "Race, n (%)" format = race.
118             ,preloadfmt=Y order_data_ ]
119             ,wtgtbl [ label = "Weight (kg)" digit = 1 ]
120             ,bmibl [ label = "BMI (kg/m^2)" digit = 1 ]
121             ,ecogbl [cat label = "ECOG Performance Status, n (%)" informat =
122             ,ecog. preloadfmt=Y ]
123             ,byvar =
124             ,popds = adsl_tot
125             ,trtvvar = trtn
126             ,stat = n mean std median q1 q3 min max sum
127             ,STATLINE = %str(n | mean (std)| median
128             ,STATLBL = %str(n | Mean (SD)| Median |
129             ,totcol = N
130             ,prefix = colval,
131             ,option = %str(blabel=Y rptcol=Y)
132             );
133 data report;
134 set sum1;
```




4. SASCodeGenius

t-ae-age.sas x

```
Chat AI Result

Debug Code

data adae2;
  length SOCPT $200; /* Adjust Length as needed */
  set adae;

  /* PT Level */
  SOCPT = catx("|", AESOC, AEDECOD);
  output;

  /* SOC Level */
  SOCPT = AESOC;
  output;

  /* Patient Level */
  SOCPT = '1stLine|Patients with TEAE';
  output;
run;
```

```
35 if AESOC = '' then AESOC = "Uncoded";
36 run;
37
38
39 /*3.3 create 3 levels of records for numerator
40 data adae2;
41   length SOCPT $2;
42   set adae;
43   /*PT level*/
44   SOCPT=catx("|" AESOC, AEDECOD);
45   output;
46   /*SOC level*/
47   SOCPT=AESOC; output;
48   /*patient level*/
49   SOCPT='1stLine|Patients with TEAE';
50 run;
```

Debug Generate Com

Chat AI Result

>

Here is the input: map 1-12 to each month by proc format.

Are you ready to proceed, or do you have further guidance?

ok

Great! It sounds like you want to map 1-12 to each month. Here's an example:

```
proc format;
  value month
    1 = 'January'
    2 = 'February'
    3 = 'March'
    4 = 'April'
    5 = 'May'
    6 = 'June'
    7 = 'July'
    8 = 'August'
    9 = 'September'
    10 = 'October'
    11 = 'November'
    12 = 'December';
run;
```

Let me know if you need any further assistance with this!

make it short, 3 chars in uppercase

Certainly! You can create a SAS format to map the numbers 1-12 to the abbreviated month names in uppercase like this:

```
proc format;
  value month
    1 = 'JAN'
    2 = 'FEB'
    3 = 'MAR'
    4 = 'APR'
    5 = 'MAY'
    6 = 'JUN'
    7 = 'JUL'
    8 = 'AUG'
    9 = 'SEP'
    10 = 'OCT'
    11 = 'NOV'
    12 = 'DEC';
run;
```

Your input:

proc format;

```
value
  1 = 'JANUARY'
  2 = 'FEBRUARY'
  3 = 'MARCH'
  4 = 'APRIL'
  5 = 'MAY'
  6 = 'JUNE'
  7 = 'JULY'
  8 = 'AUGUST'
  9 = 'SEPTEMBER'
  10 = 'OCTOBER'
  11 = 'NOVEMBER'
  12 = 'DECEMBER';
run;
```

This format will map the numbers 1-12 to their respective abbreviated month names in uppercase. Let me know if you need any further assistance!

Your input:

proc format;

It looks like you are trying to create three levels of records for the numerator in your dataset `adae2` from the original dataset `adae`. However, there are a few issues in your code that need to be addressed:

- 1.The variable `socp` is used in the patient level but is not defined in the length statement.
- 2.The variable `socp` is used inconsistently in the patient level.
- 3.The patient level should have a consistent variable name with the other levels.

Here is the corrected version of your code:

```
data adae2; length SOCPT $200; /* Adjust length as needed */ set adae; /* PT level */ SOCPT = catx("|", AESOC, AEDECOD); output; /* SOC level */ SOCPT = AESOC; output; /* Patient level */ SOCPT = '1stLine|Patients with TEAE'; output; run;
```

Explanation of changes:

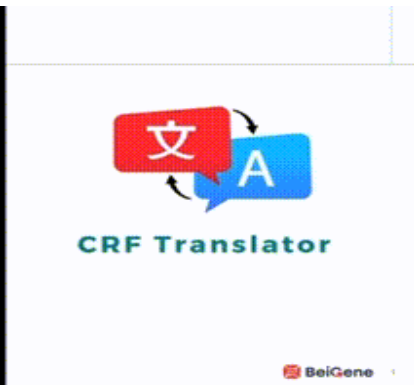
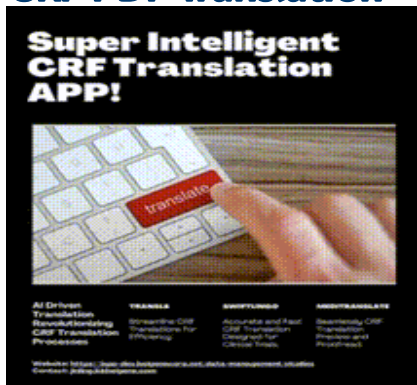
1. Defined the length of `socp` to ensure it can hold the concatenated values.
2. Used `socp` consistently for all levels.
3. Ensured the patient level record is output with the same variable name `socp`.

This should create the three levels of records as intended

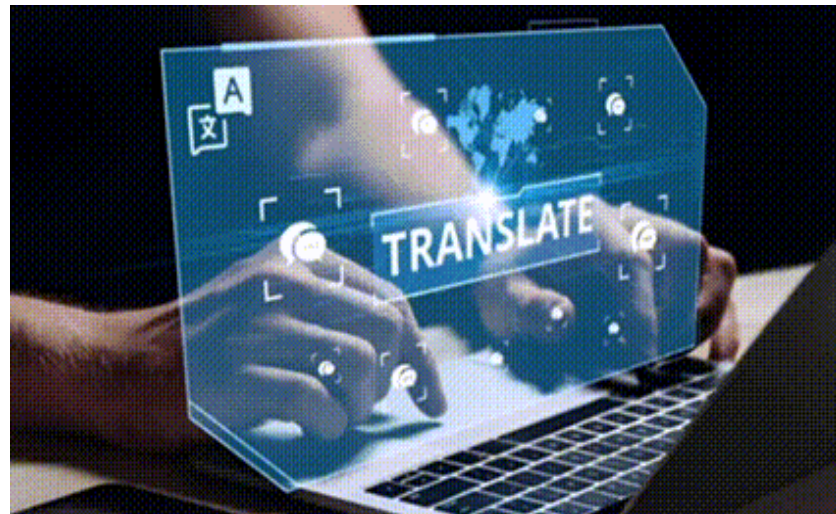


5. Translation

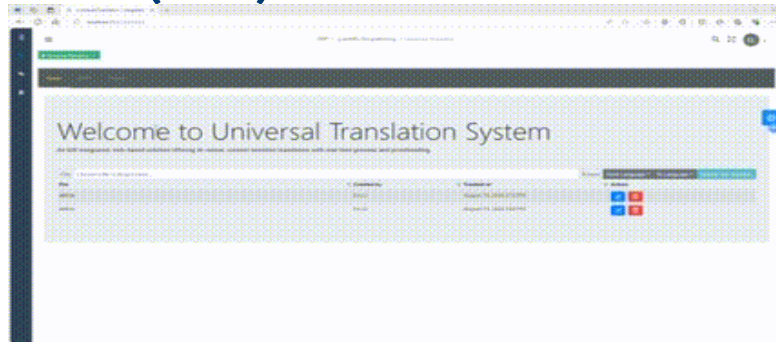
CRF PDF Translation



SDRG / ADRG (WORD) Translation



DEFINE (Excel) Translation



- AI: < 10 mins
- Human: several weeks

[illegible]

新签合同-主要 合同名称	合同	行业	完工性	基础设 施建设 投资占比	现代 服务业 占比	主要 影响	备注
A136、房屋建 筑工程总承包	房屋建筑工程			X		每个合同签一单	
A134F、平潭岛 景区基础设施	景区基础设施		X			每个合同签一单	
A234A、市政 基础设施工程	市政基础设施			X		每个合同签一单	
A232C、印刷 品及办公用品	文具办公用品			X		每个合同签每样	
A234A、市政 基础设施工程	市政基础设施		X			每个合同签一单	
A232C、印刷 品及办公用品	文具办公用品			X		每个合同签每样	
A232D、印刷 品及办公用品	文具办公用品			X		每个合同签每样	
A232D、印刷 品及办公用品	文具办公用品			X		每个合同签每样	
A232D、印刷 品及办公用品	文具办公用品			X		每个合同签每样	
A232D、印刷 品及办公用品	文具办公用品			X		每个合同签每样	



Typical Tasks

Solution : Open Code -> Macro -> Metadata Driven -> NLP -> GenAI

SDTM

Short sentence and simple logic;
Limited number of patterns;
Facilitate Mapping spec recommendation
based on text similarity

ADaM

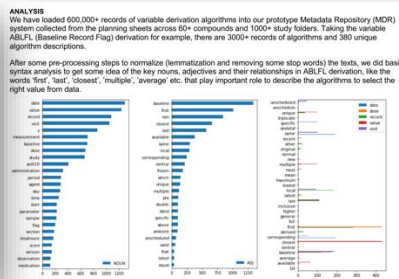
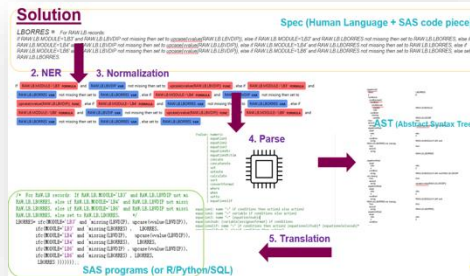
Long sentence and complex logic;
Pretty difficult for non-standard Specs;
Need more support from standard
macro/functions and knowledge database;

TLG

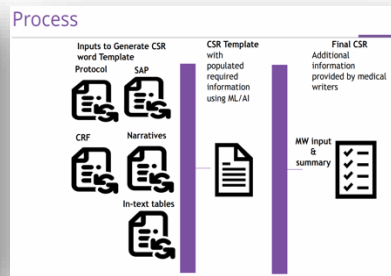
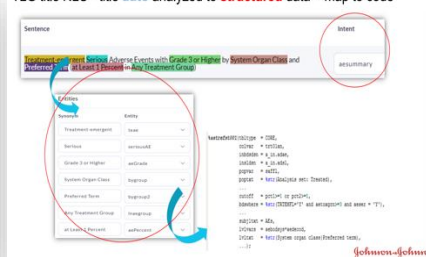
Short and relatively Standard sentence;
Need support from standards macro/functions;
Need Table layout analysis from mockup
table;

CSR

Good at key information extraction, similarity
calculation, and text summary;
Predict and identify key information in CSRs
and link to different documents and sections



TLG title NLU - title **auto** analyzed to **structured** data – map to code



Balance Between Humans and Machines – How NLP Helps Translate Text Specification to Code and Vice Versa

Metadata-based Auto-Programming Process – Part 2: Map human language to SAS® code by Natural Language Processing (NLP)

IDAR virtual AI assistant to automate statistical report generation

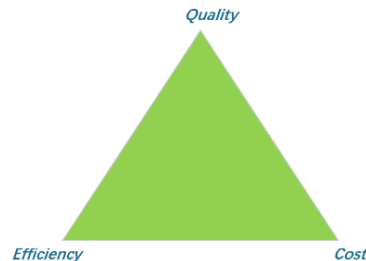
Automate Scientific Document such as CSR Generation using ML/NLP





Summary

- The rapid advancement of Generative AI (GenAI) is revolutionizing many industries including clinical trials
- This acceleration from data to insights, democratized AI technology and easy web apps present statistical programmers with a unique opportunity to evolve into data scientists.
- Expanding the role of statistical programmers can accelerate the clinical research process, create greater value and impact, and leverage our advantage in domain-specific expertise in both business and technical areas.



Moore's Law provides exponential growth in...

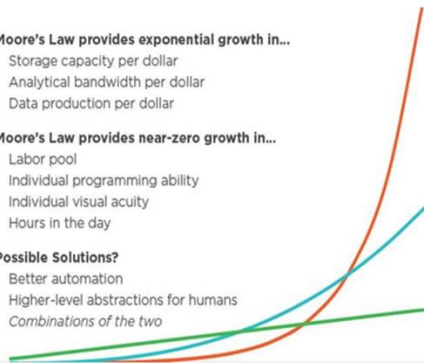
Storage capacity per dollar
Analytical bandwidth per dollar
Data production per dollar

Moore's Law provides near-zero growth in...

Labor pool
Individual programming ability
Individual visual acuity
Hours in the day

Possible Solutions?

Better automation
Higher-level abstractions for humans
Combinations of the two



Acknowledgement

Cindy Song

You Li

Jinling Li

Xuejie Zhou

Mingfeng Zhou

Wei Wang

Yaohui Zhu

Su Qiu

YunSheng Xie

Yankun Zhang

Haijing Yu

Zeqin Lian



**The Global Healthcare
Data Science Community**

Contact Channels

UK +44 1843 609600
office@phuse.global
phuse.global

Social Media

@phusetwit
/phusebook
/phusetube
/company/phuse