How Large Language Models Support Statistical Analysis

Peter H. Gruber
Università della Svizzera italiana, Lugano
peter.gruber@usi.ch

Workshop on Empirical Research in the AI Era

BPMRL Workshop, Porto, Dec 16-17 2024

Starting point: Large Language Models (LLMs)

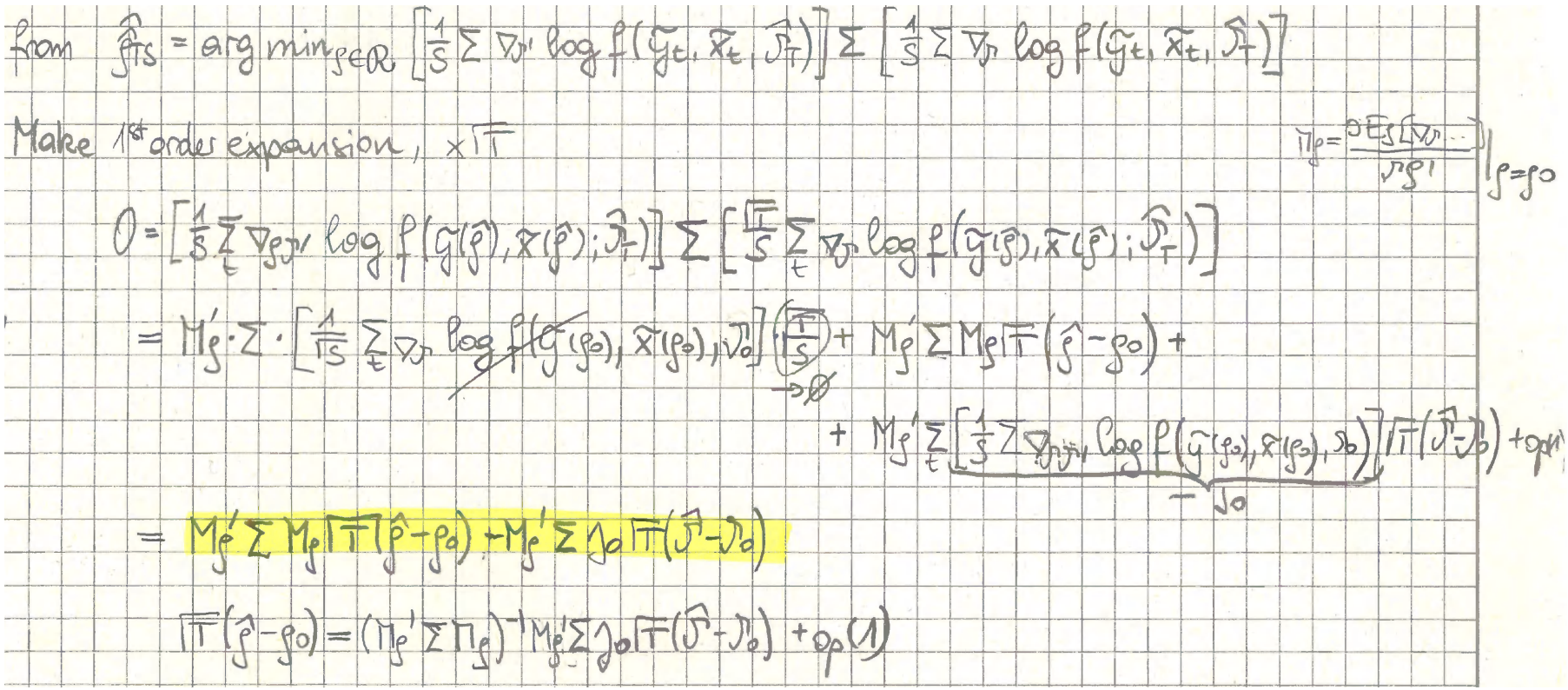
- **LLMs are ...**
  - Statistical models to generate text/code
  - Method: sequentially predict a likely next word
- **LLMs are Language models.**
  They can ...
  - Transform and translate information
  - Learn provided information
- **LLMs are not Knowledge/Computation models.**
  They cannot ...
  - Know everything
  - Analyze a problem formally
  - Understand *your* background

Hypothesis:

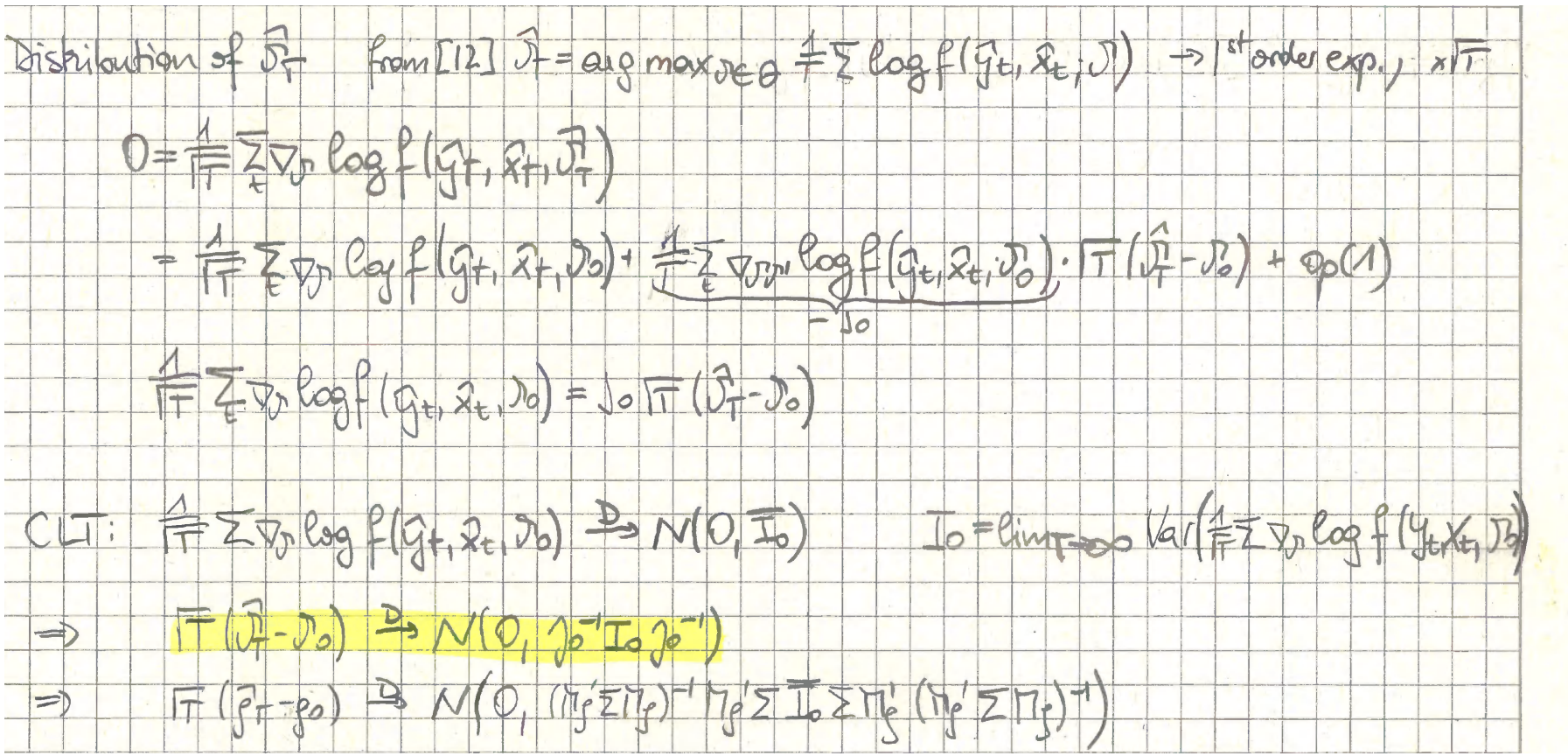
Statistics problems are
(in large parts) language problems

What are relevant questions in statistics?

Questions like these?



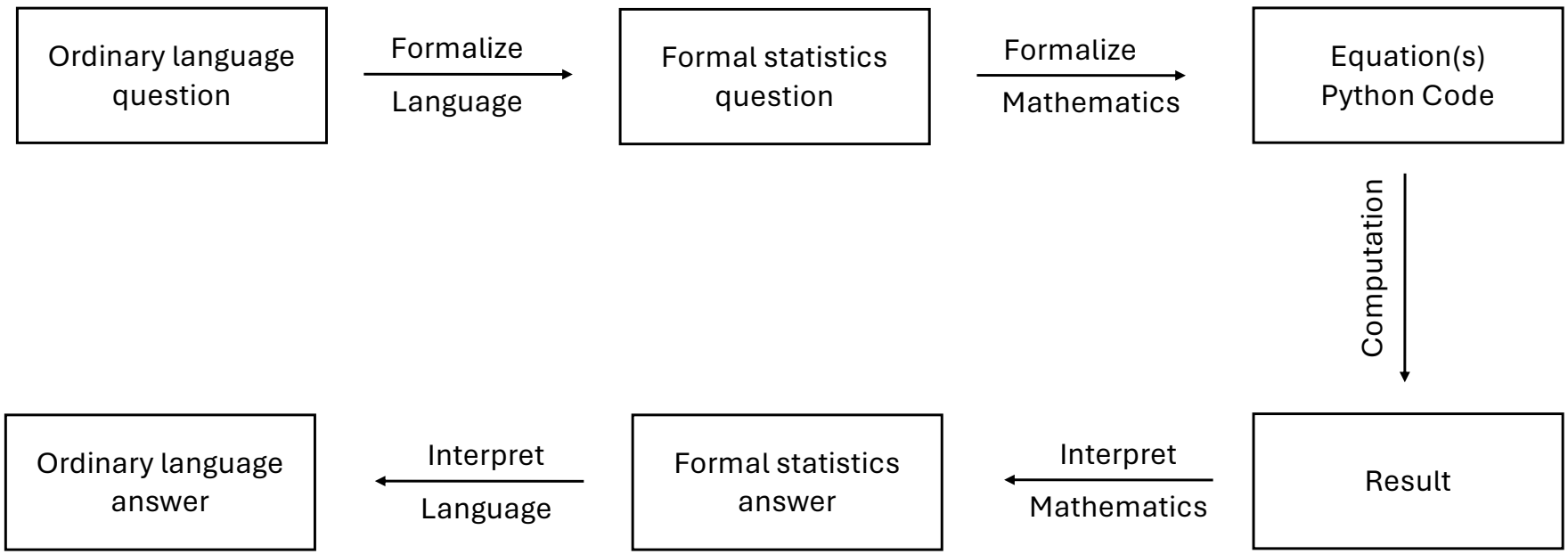
Questions like these? (part 2)



Or questions like these?

- Which of two cancer therapies should I choose?
- What policy for combatting poverty works best?
- Are girls better students than boys?

So how do we answer empirical questions?

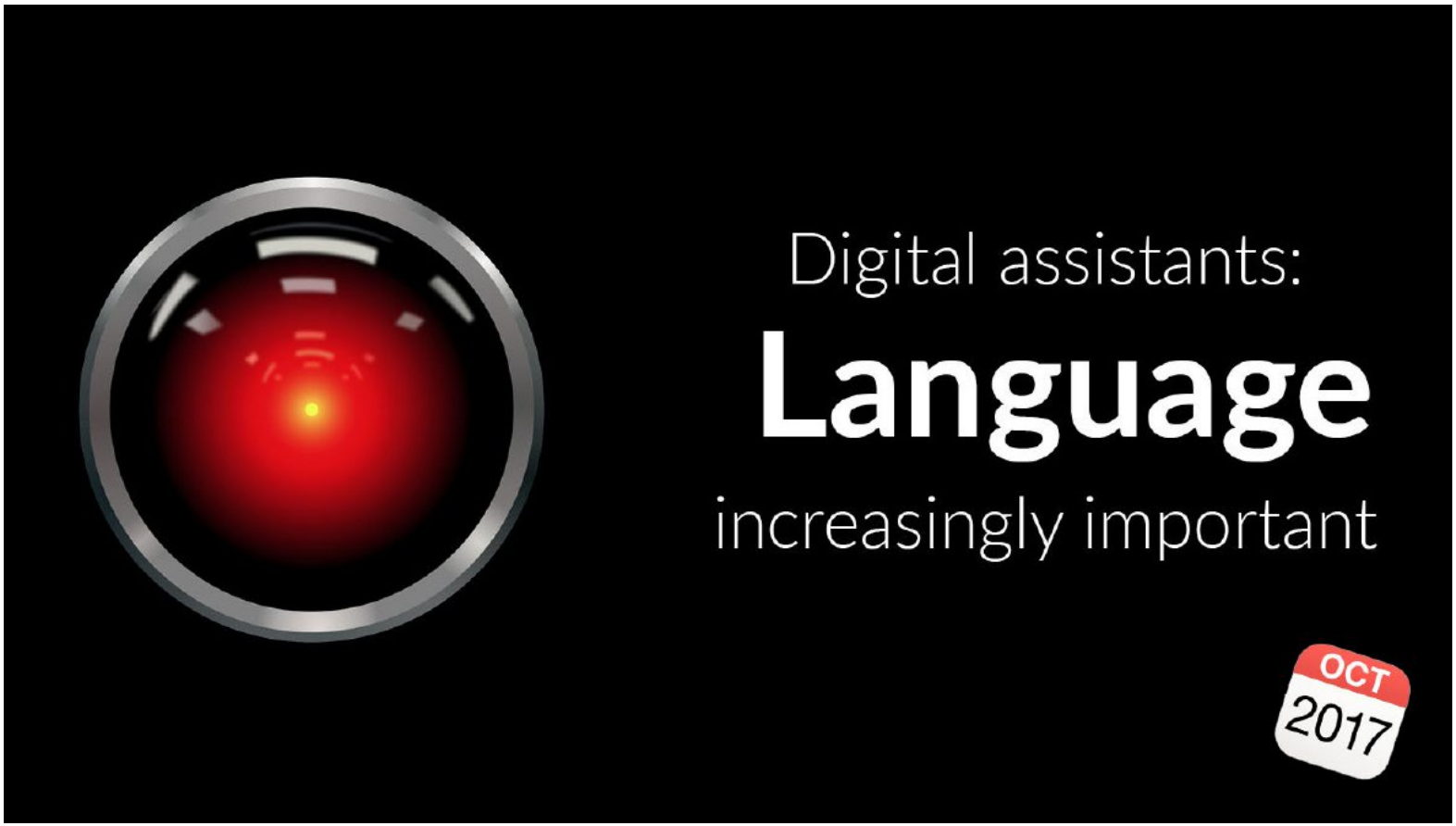


Demo

– ChatGPT

– Microsoft Copilot

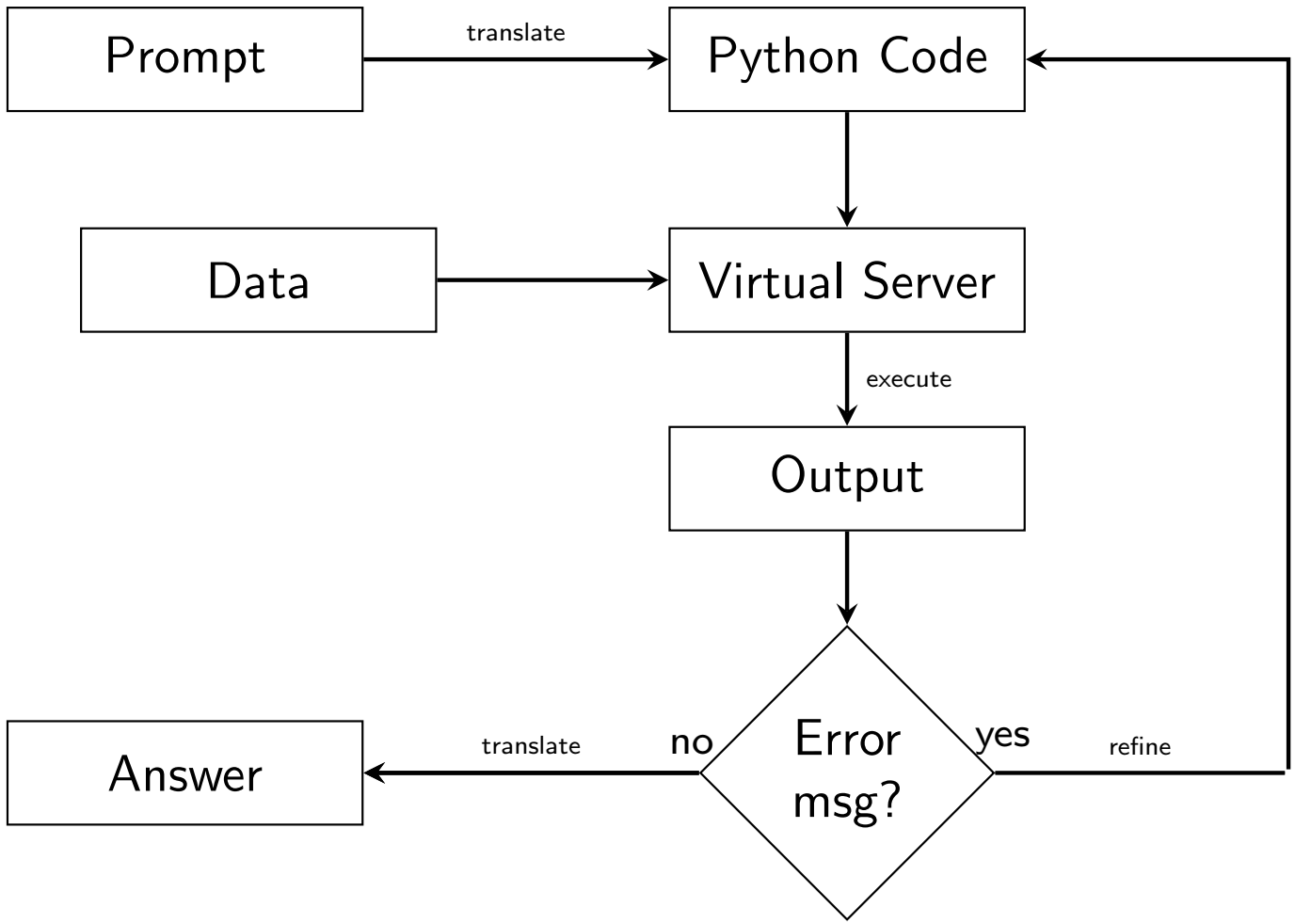
So language is important



Recap: Large Language Models (LLMs)

- **LLMs are ...**
  - Statistical models to generate text/code.
  - Method: sequentially predict (most) likely next word.
- **LLMs are Language models.**
  They can ...
  - Transform and translate information
  - Incorporate provided information
- **LLMs are not Knowledge/Computation models.**
  They cannot ...
  - Know everything
  - Analyze a problem formally
  - Understand *your* background

A closer look at the Advanced Data Analytics



4th wave of democratisation . . .

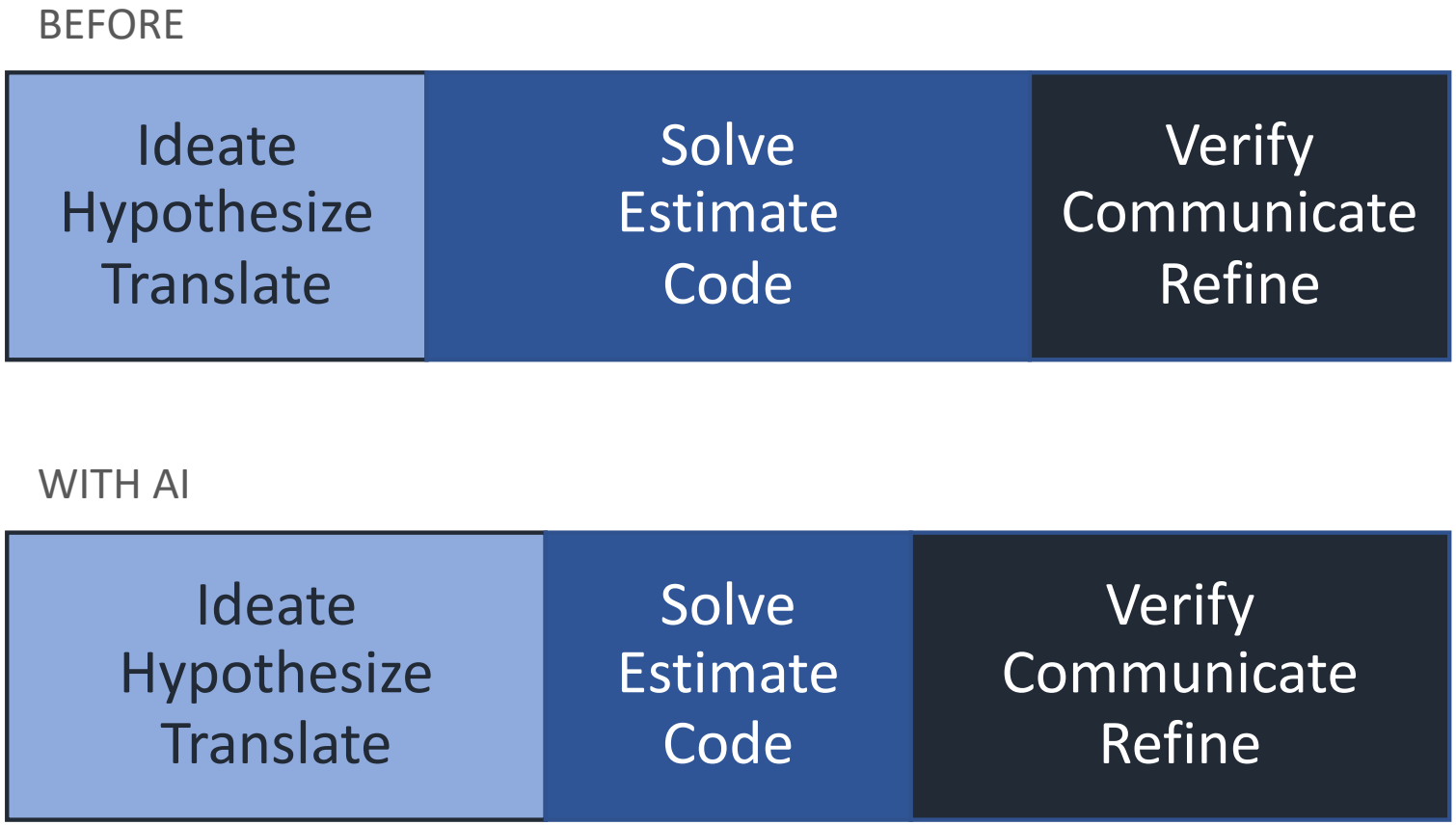
Communication

Knowledge

Computation

Analytics

Accordion model



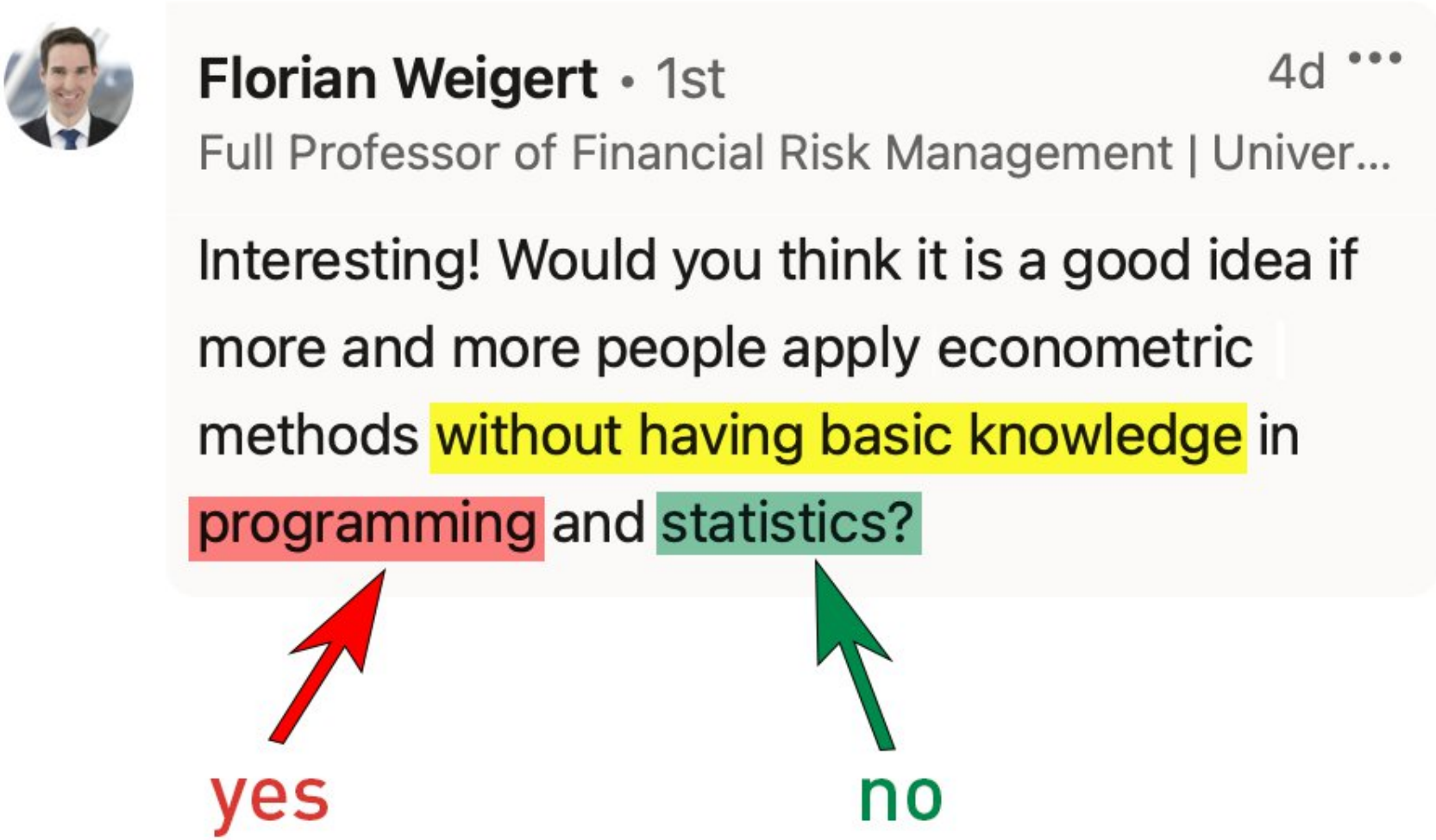
Controversy

Florian Weigert • 1st 4d •••

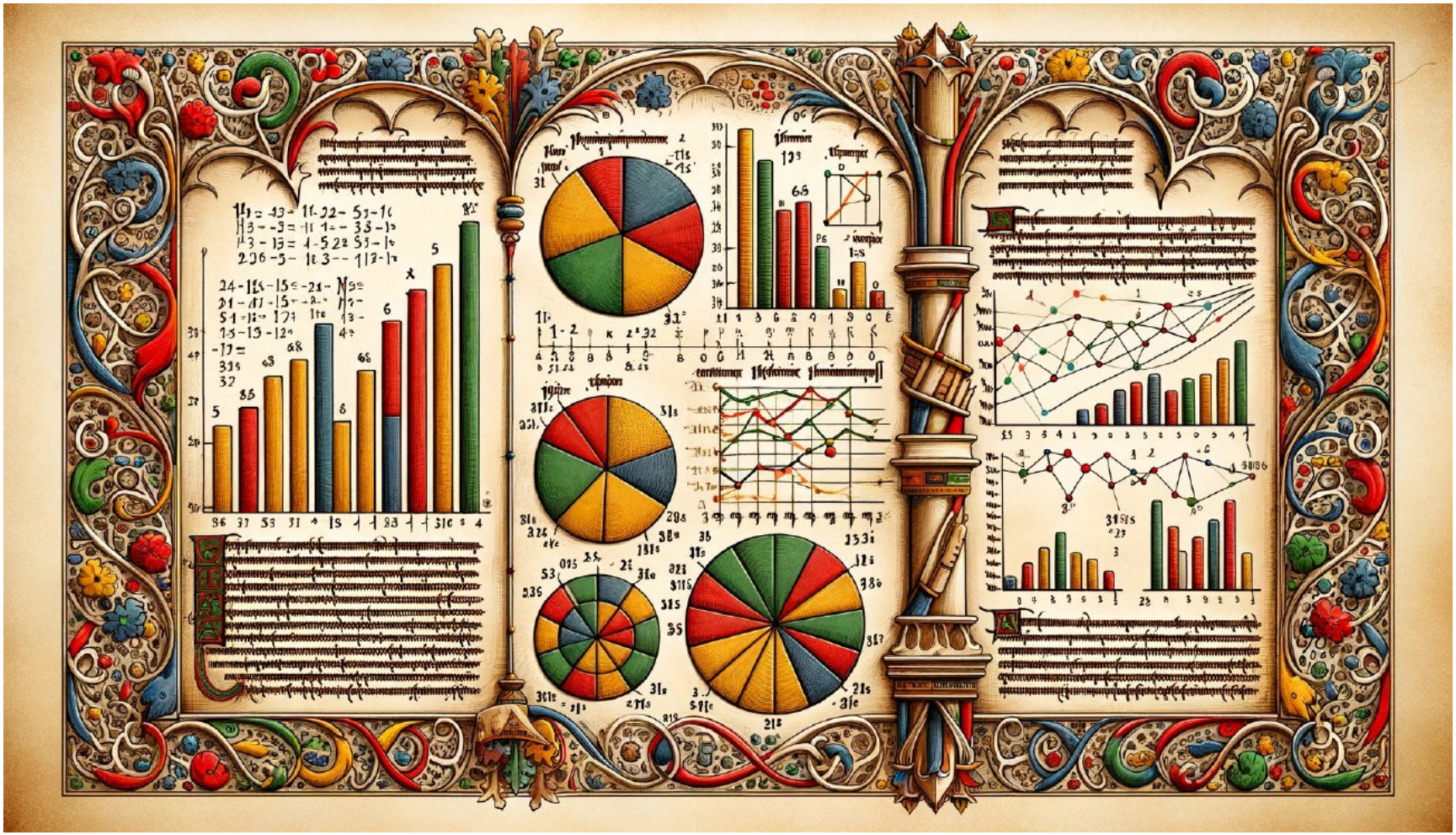
Full Professor of Financial Risk Management | Univer...

Interesting! Would you think it is a good idea if more and more people apply econometric methods without having basic knowledge in programming and statistics?

Controversy



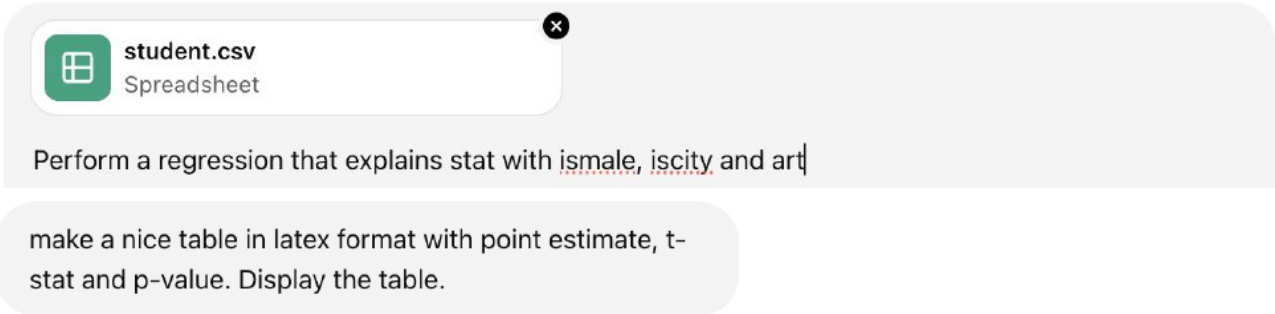
Controversy (2)



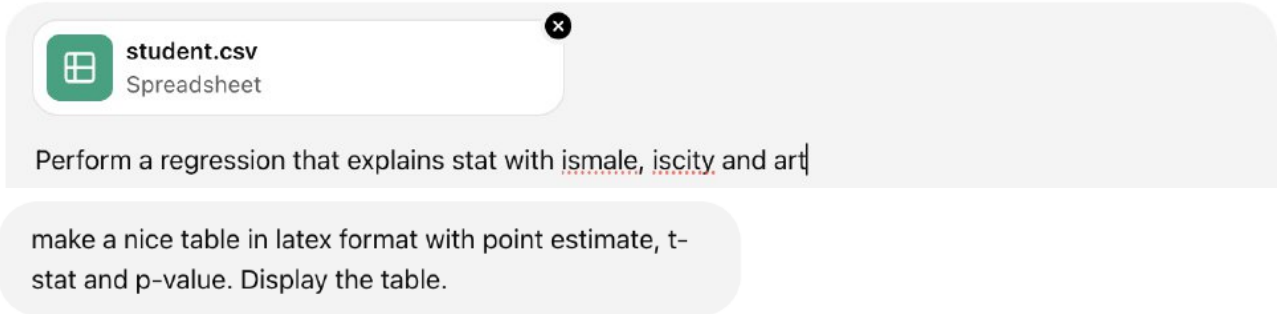
Latin is no more the language of science.

But this is not the real problem.
What can possibly go wrong?

Anatomy of a ChatGPT 4.0 fail

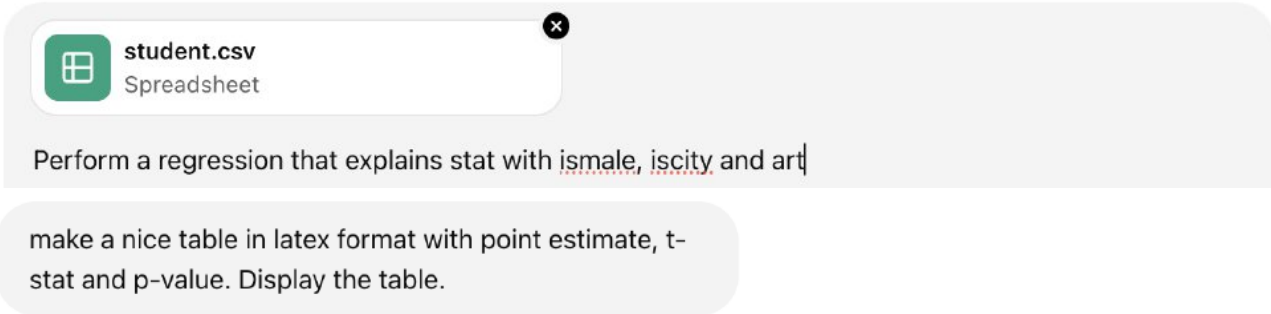


Anatomy of a ChatGPT 4.0 fail



Variable	Coefficient	Std. Err.	t-Statistic	p-Value
Constant	46.0108	4.839	9.509	0.000
isMale	10.8955	2.537	4.295	0.000
isCity	4.2464	2.387	1.779	0.077
Art	0.3016	0.069	4.388	0.000

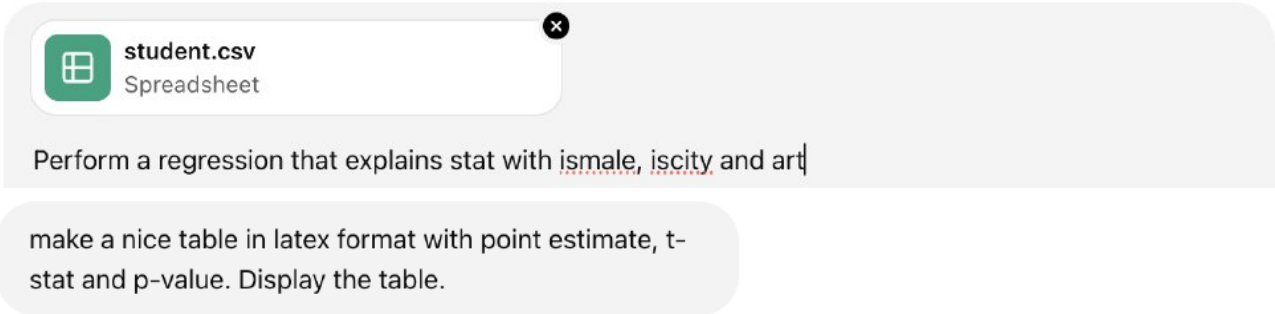
Anatomy of a ChatGPT 4.0 fail



Variable	Coefficient	Std. Err.	t-Statistic	p-Value
Constant	46.0108	4.839	9.509	0.000
isMale	10.8955	2.537	4.295	0.000
isCity	4.2464	2.387	1.779	0.077
Art	0.3016	0.069	4.388	0.000

```
> fit <- lm(stat ~ isMale + isCity + art, data=grade)
> summary(fit)
```

Anatomy of a ChatGPT 4.0 fail



Variable	Coefficient	Std. Err.	t-Statistic	p-Value
Constant	46.0108	4.839	9.509	0.000
isMale	10.8955	2.537	4.295	0.000
isCity	4.2464	2.387	1.779	0.077
Art	0.3016	0.069	4.388	0.000

```
> fit <- lm(stat ~ isMale + isCity + art, data=grade)
> summary(fit)

     Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
     (Intercept) 44.87133    4.18649  10.718  < 2e-16 ***
     isMale      11.14198    2.20838   5.045 9.92e-07 ***
     isCity       5.49515    2.05437   2.675  0.00808 **
     art          0.30424    0.05972   5.094 7.90e-07 ***
```

What happened?

```
# Selecting the features and target variable
X = data[['isMale', 'isCity', 'art']]
y = data['stat']

# Splitting the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initializing and training the linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Predicting on the test set
y_pred = model.predict(X_test)

# Calculating metrics
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

model.coef_, model.intercept_, mse, r2
```

Does this always happen? No!

student.csv
Spreadsheet

Perform a regression that explains stat with ismale, iscity and art

make a nice table in latex format with point estimate, t-stat and p-value. Display the table.

Answer 1:

Variable	Coefficient (Point Estimate)	Std. Error	t-Statistic	p-Value
Intercept	44.871	4.186	10.718	1.38×10^{-21}
isMale	11.142	2.208	5.045	9.92×10^{-7}
isCity	5.495	2.054	2.675	0.0081
art	0.304	0.060	5.094	7.90×10^{-7}

Answer 2:

Variable	**Coefficient**	**t-Statistic**	**p-Value**
Constant	46.0108	9.509	0.000
isMale	10.8955	4.295	0.000
isCity	4.2464	1.779	0.077
art	0.3016	4.388	0.000

Answer 3:

Variable	**Point Estimate**	**t-Statistic**	**p-Value**
Intercept	44.87	10.72	0.000
isMale	11.14	5.05	0.000
isCity	5.50	2.67	0.008
art	0.30	5.09	0.000

ChatGPT errors

A computer will do
what you tell it to do,

. . . but that may be much different
from what you had in mind.

Joseph Weizenbaum

`https://psych.fullerton.edu/mbirnbaum/psych101/eliza.htm`

Which is your statistics *tribe*?

- What does "perform a regression of ..." Mean for *you*?
- What is *your* goal?

– Use a model
– Employ all data

– Assess its performance
– Split into training/test data

Why are we so angry about ChatGPT fails?

Why are we so angry about ChatGPT fails?

Most people have little *assessment* experience.

Adapting our profession (and teaching)

More	Less
Assessing	Doing
Natural language	Formalism
Reading (code)	Writing (code)
Why	How
Understanding principles	Practise

Challenge
How to train future executives, if we eliminate junior positions.

ChatGPT strategy

Three unusual competences

- Communication
  - Precise use of ordinary and subject-specific language
- Intercultural skills
  - Not assuming any background
  - Expressing the obvious
- Management skills
  - Setting and communicating goals
  - Structuring large problems into smaller tasks
  - Assessing results and giving useful feedback

ChatGPT strategy (2)

- Learn from computer science
  Test and verify in an organized, modular way (make a test plan)
- Use clear prompts, based on a prompting framework
- Declare your statistical tribe in the custom instructions
- Perform visual checks ($\leftarrow$ very reliable in ChatGPT)
- Use ChatGPT to check its own results
  - Ask ChatGPT how it found the solution
  - Ask ChatGPT in a new conversation to assess/verify solution
- Learn a bit of Python and run code in Google Colab

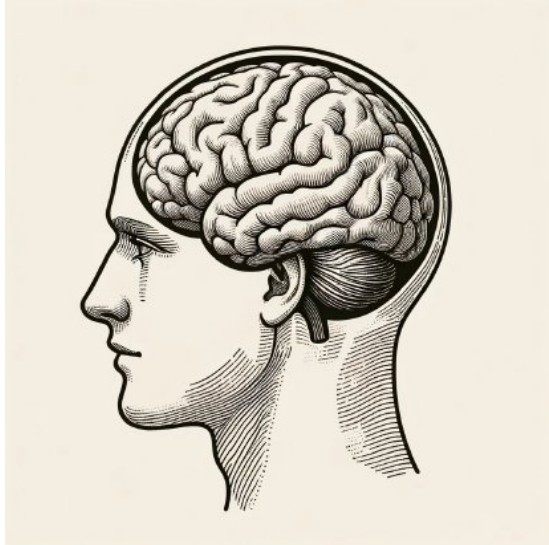
Hang up, call again (HUCA)

- Start a new conversation in order to ...
  - Reset context
  - Reset randomizer

Choose appropriate *mode* of Chat GPT

Data Analyst GPT

– Force writing code
– Regular problem
– Replicability
– Limited by virt. machine



ChatGPT classic GPT

– Only use text model
– Very messy dataset [demo?]
– Creativity
– Limited by $\frac{1}{2}\times$ context

- ChatGPT 4-o chooses mode for you (mostly correct)
  - GPTs for special tasks: Wolfram, AskYourPDF, Consensus / Scholar AI
- Cascade modes for better results
  - Fix data in ChatGPT classic $\rightarrow$ analyse with Data Analyst
  - ChatGPT classic GPT to explains code created by Data Analyst

Prompt framework

Use the **P.R.O.P.E.R.** framework!

Persona – which role should it take?
E.g. professor, helpful assistant, critic, ...

Request – what task should it should fulfill?

Operation – in which way / using which method?

Presentation – which tone/style/format for the result?
E.g. informal, short, table, "for an 8-year old", ...

Examples – provide a template for the output.

Refinement – give feedback, iterate and improve.

A meme with a grain of truth