



# Improving Foundation Models via Human Data

Vijay Krishnan  
Founder & CTO, [Turing.com](https://turing.com)

# Levers for Foundation Model Improvements

- 1 ——— Well understood levers →  
Table stakes, not competitive advantage
  - A. Compute: NVIDIA GPUs etc. All Big players have this in large volumes.
  - B. Algorithms: Research has shown that a whole range of model architectures trained on similar data make similar predictions. Improvements largely efficiency related.
  - C. Pre-training data. Very similar across companies
- 2 ——— High impact lever but less research and less understood → Competitive advantage
  - D. Proprietary Human Data for post-training in complex domains. Early Improvements often measurable with 1000-5000 high quality data points for SFT/RLHF/DPO.**

This is the unfair advantage that Turing.com brings to the table.

# Turing helps the **world's best Foundation LLM companies** improve model performance for reasoning, coding, agents, problem solving, and other advanced AGI capabilities

## 3 Million+ knowledge workers

Software Developers, other knowledge workers from 100+ countries vetted across 100s of technical, communication and instruction following skills.

## Trusted partner

Turing's data services is used by OpenAI, Anthropic, Google, Meta, Apple, Nvidia, xAI, Snowflake, ServiceNow, Character.ai, Augment and many more.

## Wide range of offerings

We offer model evaluation & strength assessment, SFT, RLHF, DPO, for code, data science, agents, factuality, multimodality, STEM expertise etc.



# Turing's talent cloud of 3 Million+ knowledge workers produce **custom proprietary data at scale** to improve models



## Largest talent pool in the world

Build a specialized team of software developers, other knowledge workers from Turing's 3 million developers, across 150+ countries. More diverse outputs than a whole team in a single office.



## Deeply vetted and diverse in skills and domains

Our knowledge workers have diverse skills and excel in tasks like data analysis, logical reasoning with proper function & agent calls, multimodal data, code gen, evaluation, documentation, and scaling SFT and RLHF loops.



## Transparent management with quality control

Turing provides a fully managed developer team, allowing researchers to focus solely on task design while Turing handles developer coordination and management.



## Rapidly scalable and elastic

Leveraging Turing's global talent pool, we can rapidly scale our service to 100s of developers within weeks, all while maintaining significant cost-effectiveness.

# How **research and human data** teams can collaborate to improve model performance

Human data is typically applied during post-training. Base model used can be a company's own pre-trained model or an open source model.

## Identify backlog of areas for model improvement

- Coding, data analysis, multimodal reasoning.
- Ideas, prioritization inputs from product owners, researchers, and Human Data team analysis.

## Continuous optimization

- Evals, SFT, RLHF data generated in various areas of interest.
- Model improvements analyzed every 1000-5000 data points. Decide whether to scale, change task specs or whether saturation has been fit.

## Talent re-allocation and task transition

- Strategic reuse of trained, high performing LLM Human Data trainers for new improvement phases.

# Research and human data team collaboration to improve models in a specific area.

## Measure and Improve model performance in a specific area

### Assemble a team skilled in the specific domain

- Either new LLM human data trainers or those transitioning from comparable data tasks

### Evaluation dataset generation for the specific domain

- LLM human data trainers create evaluation/test data with human judgments as the performance benchmark
- Typically 30-120 min per data point incl of review and rework to ensure high quality.

### Supervised Fine Tuning (SFT) process

- Task designed for SFT data generation
- LLM Human data trainers produces high quality data points. 30-120 min per data point incl review/rework.
- Researchers enhance model using data.
- Continue until improvements saturate on evaluation data (~10K-100K data points)

### Reinforcement Learning with Human Feedback (RLHF) cycle

- Comparison/judgement/rewrite/process supervision / related tasks designed for RLHF
- LLM trainers produce high quality data. Typically 15-60 min per data point.
- Researchers enhance model using data
- Continue until improvements saturate on evaluation data (~10K-100K data points).

# Optimizing quality and throughput

## Quality

### Researcher Satisfaction Score Rubric



Team leads assess notebooks and assign quality ratings to each submission.



Weekly monitoring of quality scores to maintain researcher satisfaction standards.

## Throughput

### Time/Task



Weekly tracking of throughput scores to meet researcher expectations.



Support provided by Developer Success team for developers with lower-than-expected throughput scores.

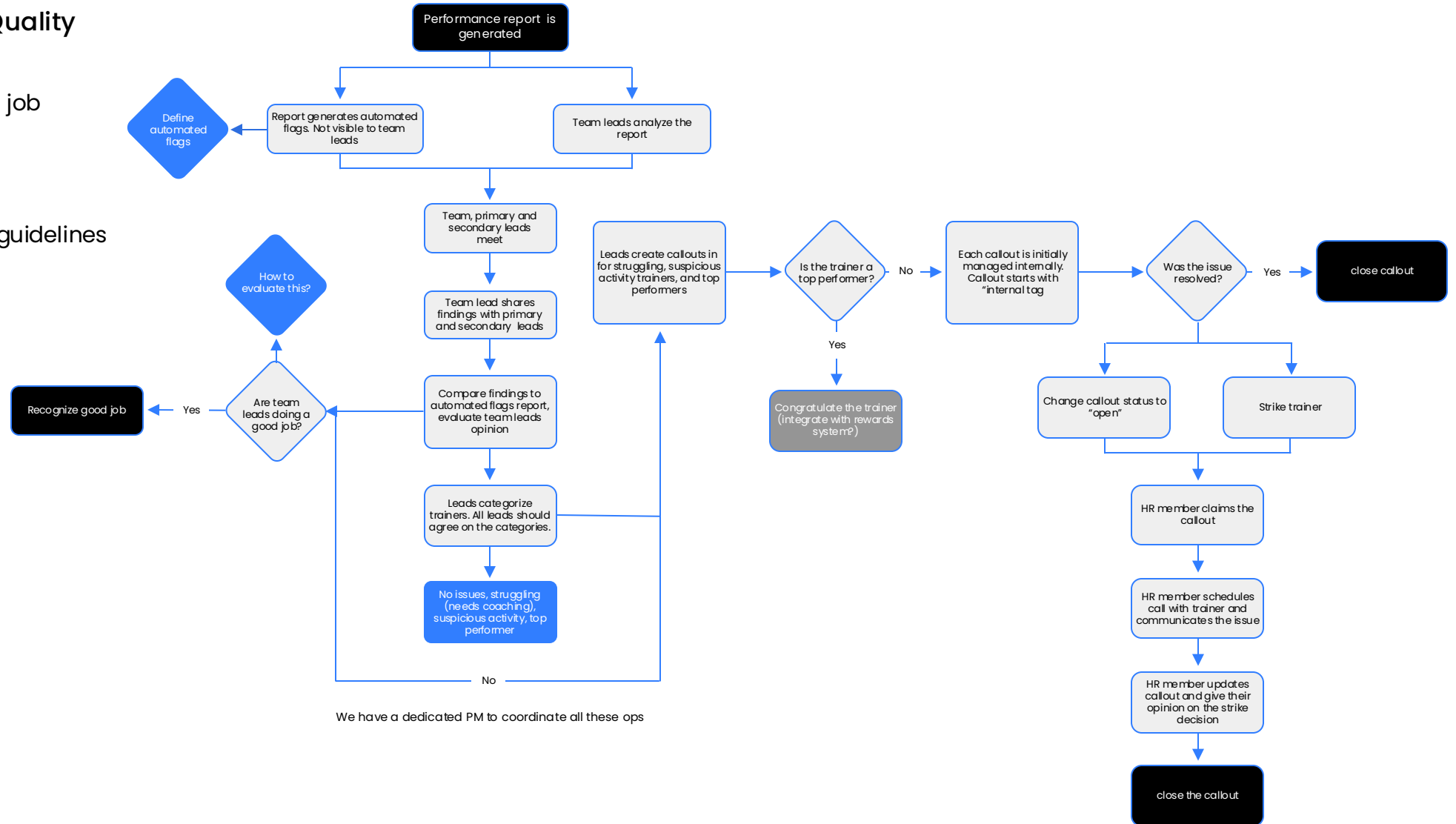
# Best practices to ensure high **quality** of Evals, SFT, rlhf data

## LLM human data trainer Quality

- Vetting
- Performance reviews on the job

## Data Quality

- Calibrate with researchers' guidelines
- Label review workflow
- Labeling tooling





# Sample model areas that are being improved via human data

Coding & Debugging

Agents, Function, Calling, Tool-  
Use

Data Analysis

Multi-Modality

Advanced Reasoning

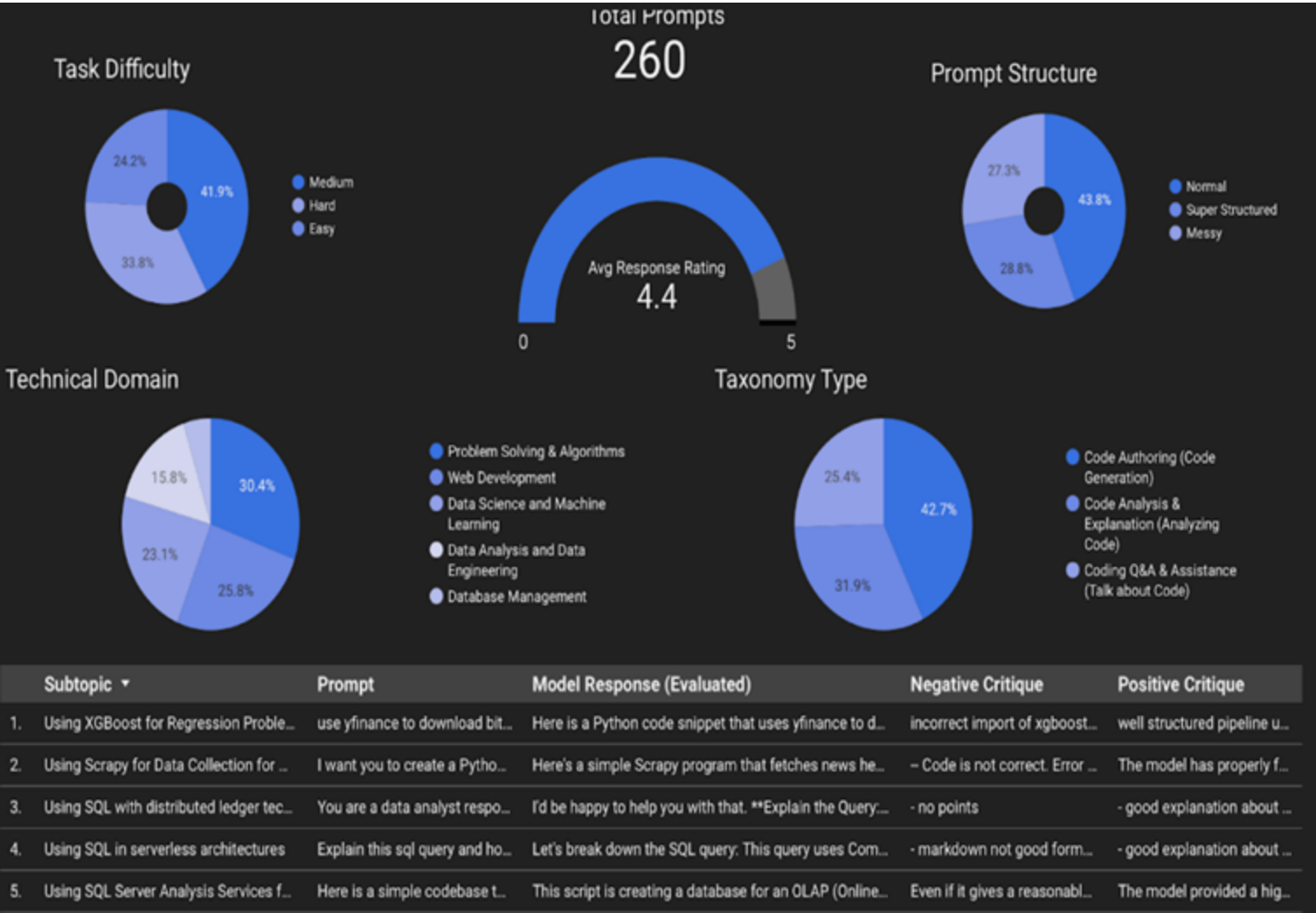
Industry Vertical & Expert  
Functional Knowledge

Frontier High end STEM  
Domain Knowledge

Factuality, RAG labeling

Alignment & Safety

# Model evaluation – identifying improvement areas



## E2e Offering Phases

### Model Evaluation Phase

#### Efforts

Explorative Model Evaluation	Objectively evaluate model performance using our optimized exploration algorithm coordinating human focus areas.
Benchmark Performance Analysis	Deep dive into when and why your model achieves its scores on any benchmarks.
Community Findings Aggregation	Research and compile the community findings about the model in a structured way from various consumer-facing data sources.

*Valid for deployed models only*

#### Post Processing

- We run the 3 efforts above in parallel for 2 weeks.
- Using the findings from the previous steps, we analyze and abstract "hypothesized" patterns of failure.
- We then conduct tests to validate each hypothesized pattern & back it up with 3 examples.

#### Usage

The following deliverables are then created, and we can finetune the % of the dataset that these findings affect.

- **Human Consumption**
  - Dashboard for Consumption to explore the distribution of model strengths & weaknesses, tying back to individual analysis.
  - A document listing the patterns & examples for each.
- **Engine Consumption**
  - **Granular Biasing** - We algorithmically bias the distribution of a data subset away from the strengths and towards the weaknesses.
  - **Targeting Failure Modes** - We expertly curate and design data subset creation efforts dedicated to each pattern. Failure Patterns + Examples are leveraged to come up with more such prompts to deal with these issues.

### Data Annotation Phase

Based on where your model currently stands, this phase can be SFT / RLHF / DPO

#### Data Split

- Baseline (30%) - This subset ignores what we have done in the model evaluation phase.
- Weakness Areas (40%) - This subset uses Granular Biasing to focus on the weakness areas.
- Pattern Treatment (30%) - This subset consists of examples for treating the previously identified



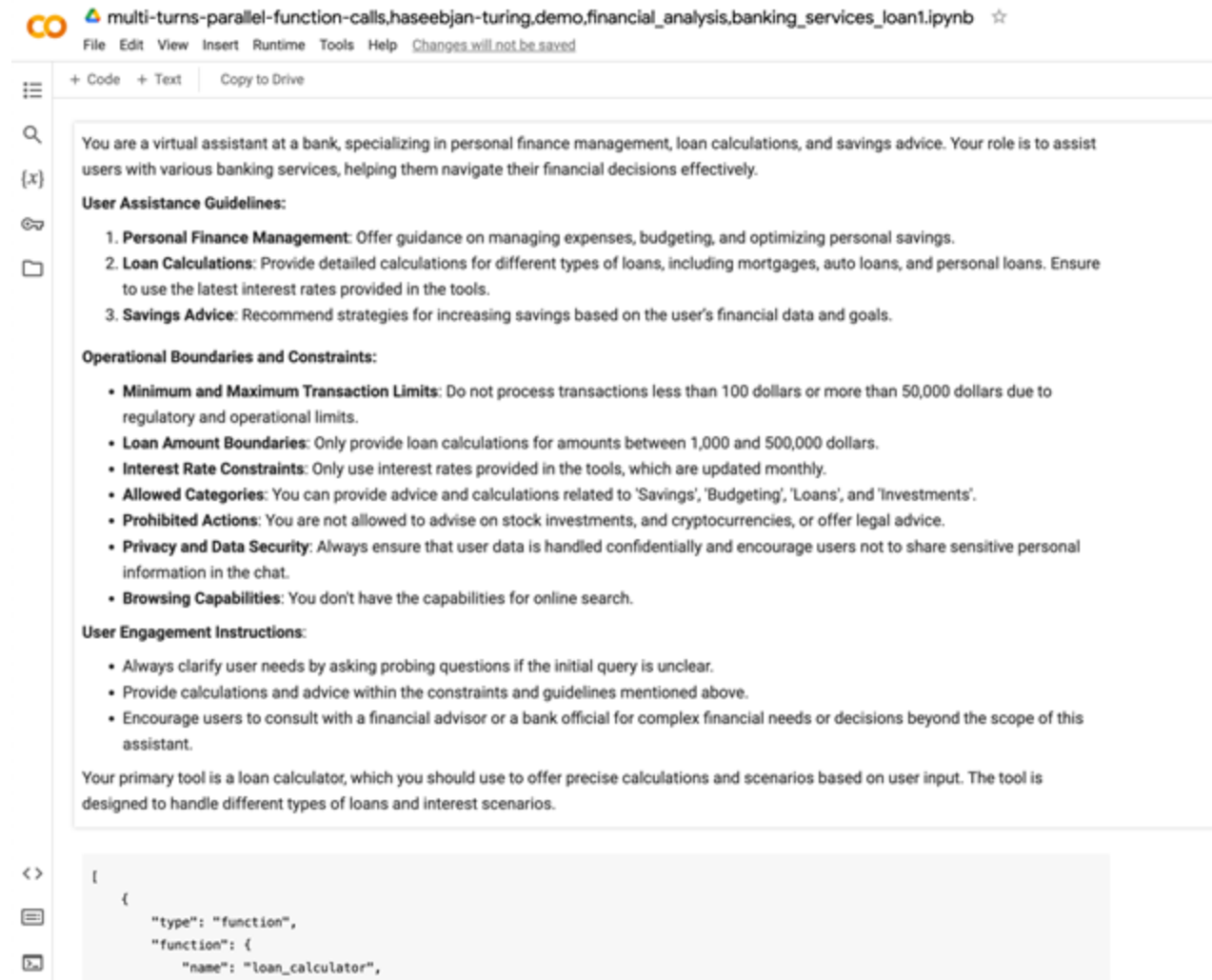
# Sample Evals, SFT, RLHF datasets across different areas



# Agents, Function, Calling, Tool-use

[https://colab.research.google.com/drive/1Imp\\_K\\_Wy5O4XUTsxcc\\_Bop7ddcBmjpuM](https://colab.research.google.com/drive/1Imp_K_Wy5O4XUTsxcc_Bop7ddcBmjpuM)

[https://colab.research.google.com/drive/1Im\\_EZml8CVOAhIRnT2Ooxy2SKkShFhH1](https://colab.research.google.com/drive/1Im_EZml8CVOAhIRnT2Ooxy2SKkShFhH1)



multi-turns-parallel-function-calls,haseebjan-turing,demo,financial\_analysis,banking\_services\_loan1.ipynb

File Edit View Insert Runtime Tools Help Changes will not be saved

+ Code + Text Copy to Drive

You are a virtual assistant at a bank, specializing in personal finance management, loan calculations, and savings advice. Your role is to assist users with various banking services, helping them navigate their financial decisions effectively.

**User Assistance Guidelines:**

- 1. Personal Finance Management:** Offer guidance on managing expenses, budgeting, and optimizing personal savings.
- 2. Loan Calculations:** Provide detailed calculations for different types of loans, including mortgages, auto loans, and personal loans. Ensure to use the latest interest rates provided in the tools.
- 3. Savings Advice:** Recommend strategies for increasing savings based on the user's financial data and goals.

**Operational Boundaries and Constraints:**

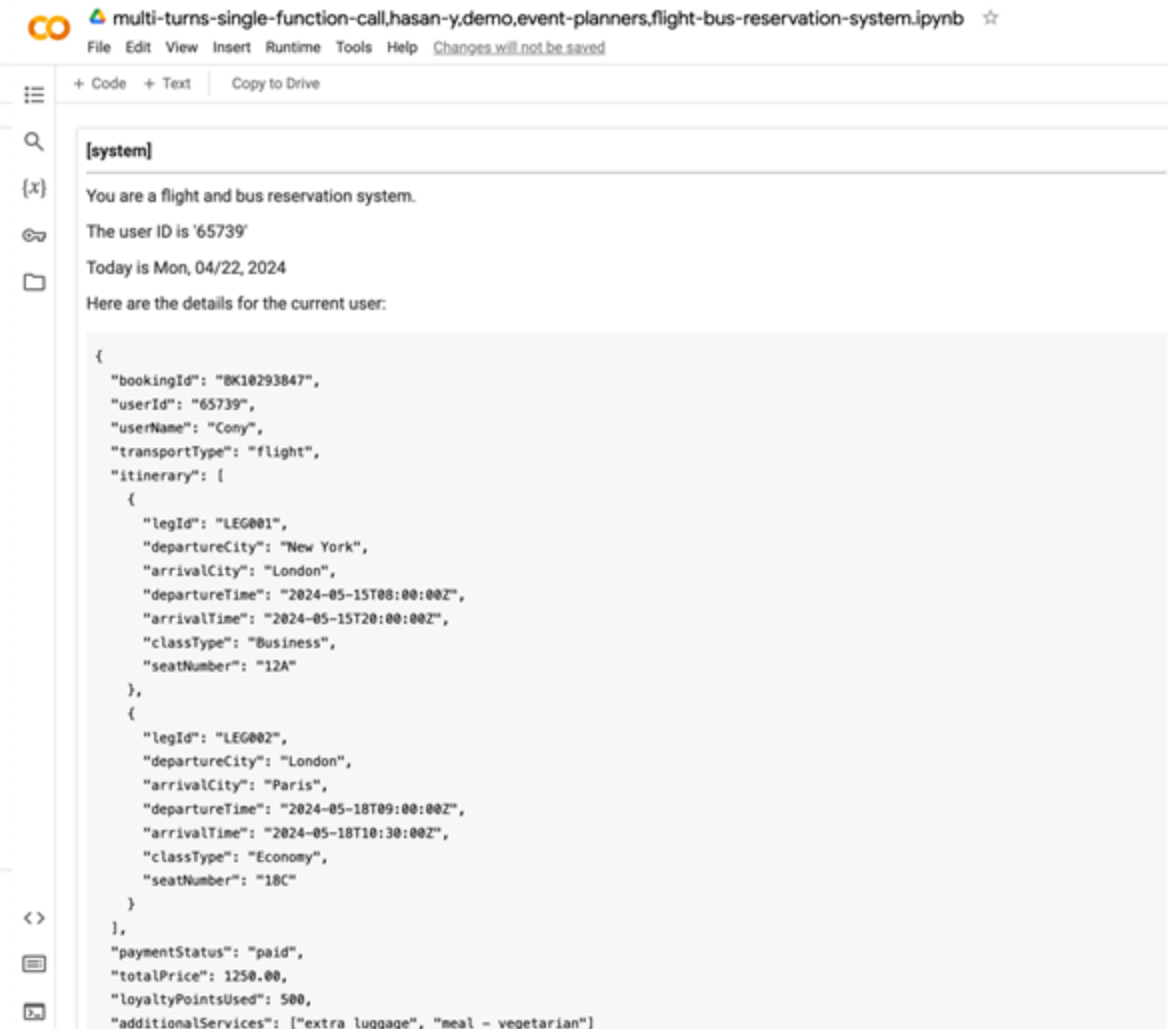
- Minimum and Maximum Transaction Limits:** Do not process transactions less than 100 dollars or more than 50,000 dollars due to regulatory and operational limits.
- Loan Amount Boundaries:** Only provide loan calculations for amounts between 1,000 and 500,000 dollars.
- Interest Rate Constraints:** Only use interest rates provided in the tools, which are updated monthly.
- Allowed Categories:** You can provide advice and calculations related to 'Savings', 'Budgeting', 'Loans', and 'Investments'.
- Prohibited Actions:** You are not allowed to advise on stock investments, and cryptocurrencies, or offer legal advice.
- Privacy and Data Security:** Always ensure that user data is handled confidentially and encourage users not to share sensitive personal information in the chat.
- Browsing Capabilities:** You don't have the capabilities for online search.

**User Engagement Instructions:**

- Always clarify user needs by asking probing questions if the initial query is unclear.
- Provide calculations and advice within the constraints and guidelines mentioned above.
- Encourage users to consult with a financial advisor or a bank official for complex financial needs or decisions beyond the scope of this assistant.

Your primary tool is a loan calculator, which you should use to offer precise calculations and scenarios based on user input. The tool is designed to handle different types of loans and interest scenarios.

```
{
  "type": "function",
  "function": {
    "name": "loan_calculator",
```



multi-turns-single-function-call,hasan-y,demo,event-planners,flight-bus-reservation-system.ipynb

File Edit View Insert Runtime Tools Help Changes will not be saved

+ Code + Text Copy to Drive

**[system]**

You are a flight and bus reservation system.

The user ID is '65739'

Today is Mon, 04/22, 2024

Here are the details for the current user:

```
{
  "bookingId": "BK10293847",
  "userId": "65739",
  "userName": "Cony",
  "transportType": "flight",
  "itinerary": [
    {
      "legId": "LEG001",
      "departureCity": "New York",
      "arrivalCity": "London",
      "departureTime": "2024-05-15T08:00:00Z",
      "arrivalTime": "2024-05-15T20:00:00Z",
      "classType": "Business",
      "seatNumber": "12A"
    },
    {
      "legId": "LEG002",
      "departureCity": "London",
      "arrivalCity": "Paris",
      "departureTime": "2024-05-18T09:00:00Z",
      "arrivalTime": "2024-05-18T10:30:00Z",
      "classType": "Economy",
      "seatNumber": "18C"
    }
  ],
  "paymentStatus": "paid",
  "totalPrice": 1250.00,
  "loyaltyPointsUsed": 500,
  "additionalServices": ["extra luggage", "meal - vegetarian"]
}
```

# Data Analysis

<https://colab.research.google.com/drive/1JaT5gWcfcuOg-SoC0fITYELPqhKuwcVH>

**Query:** For each company, which one experienced the highest percentage difference in the volume of shares traded on days where the deviation between the opening and closing stock price exceeded 1 standard deviation? Analyze data starting from Q3 2023.

**Dataset name:** MAANG share prices till February 2024

**Dataset link:** <https://www.kaggle.com/datasets/kapturovalexander/maang-share-prices-till-february-2024>

**Explanation**

We have 5 individual files. In the first step, we combine all of them into the same dataframe, adding an identifier column for the company name. To combine them, we will use Pandas to read the CSV files and then concatenate them together. For each individual CSV file, we will also add a column called `company_name`.

```
[ ] import pandas as pd
url = "https://drive.google.com/file/d/1f2m7mRWji0FGpsZ0tYBhpYos4RY0YWKZ/view?usp=sharing"
url="https://drive.google.com/uc?id=" + url.split('/')[2]
df = pd.read_csv(url)
df['company_name']='alphabet_inc'
df_final=df

url = "https://drive.google.com/file/d/1gKwBhrcuKluc2CzTZE7Iqraj1PGfz6Krl/view?usp=sharing"
url="https://drive.google.com/uc?id=" + url.split('/')[2]
df = pd.read_csv(url)
df['company_name']='amazon_inc'
df_final=pd.concat([df_final,df])

url = "https://drive.google.com/file/d/1NpwF4rCWALrwLS_Ag-11zfm6Jx8Xql4n/view?usp=sharing"
url="https://drive.google.com/uc?id=" + url.split('/')[2]
df = pd.read_csv(url)
df['company_name']='apple_inc'
df_final=pd.concat([df_final,df])

url = "https://drive.google.com/file/d/10oNZseciu061DHI9Brc0qAS4BxouKlTS/view?usp=sharing"
url="https://drive.google.com/uc?id=" + url.split('/')[2]
df = pd.read_csv(url)
df['company_name']='meta_inc'
df_final=pd.concat([df_final,df])

url = "https://drive.google.com/file/d/1p9sPpKkFQhwBwm0p-RUM0ysuthX3kCZ5/view?usp=sharing"
```

# Industry Vertical & Expert Functional Knowledge

We can use the chip RTL design examples:

[https://docs.google.com/document/d/1rXVeASEg\\_veRG9JwZNUYR9uSU\\_29XkxAcEuooHfhcrs/edit#heading=h.2t6lqgk0tokp](https://docs.google.com/document/d/1rXVeASEg_veRG9JwZNUYR9uSU_29XkxAcEuooHfhcrs/edit#heading=h.2t6lqgk0tokp)

Example notebook: [https://colab.research.google.com/drive/1043ew\\_P2Gt2LQzxqK7kAlaWtEsz-R3Q?usp=sharing](https://colab.research.google.com/drive/1043ew_P2Gt2LQzxqK7kAlaWtEsz-R3Q?usp=sharing)

## High-level Demo Notebooks

1. A simple 8-bit protocol with header, type, data, and fixed packet size is given. Design an FSM that can identify the type of package and respond accordingly: [protocol\\_processor.ipynb](#)
2. A verification flow for the above module analyzing the testbench waveform from VCD text output as well (problem statement > implementation > verification > documentation). More at annex B: [protocol\\_processor\\_2.ipynb \(SFT\)](#)
3. Support to write scripts for an EDA tool: [skill\\_simple\\_instructions.ipynb](#)
4. Make a script and a simple synthesis workflow: [execute\\_synthesis.ipynb](#)
5. Present a code with race conditions and ask the assistant to solve it: [bugs\\_and\\_race\\_condition.ipynb](#)

## Goals [DRAFT]

From [ChipNeMo Taxonomy](#) The goal is to have a large benchmark dataset to evaluate LLMs across a range of digital design and verification tasks and to have sufficient collateral to evaluate a multi-agent framework across a complete digital design flow.

The data collection is two-stage:

1. Create a corpus of designs to implement across a range of complexities and topics;
2. To have digital design and verification teams implement the designs, collecting incidental collateral (spec to RTL translation, testbench development, SVA insertion, linting, bug fixing) that is well-labeled and tagged, along with the implemented design itself. The former is to be used to evaluate targeted LLM agent tasks, while the design implementation as a whole can be used to evaluate a multi-agent LLM framework.

As a workflow reference, we can trace parallels to the SWE-bench project. As references:

1. SWE-bench: <https://www.swebench.com/>
2. Jimenez, C. E., Yang, J., Wetig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2023). SWE-bench: Can Language Models Resolve Real-World GitHub Issues? (Version 2). arXiv: <https://doi.org/10.48550/ARXIV.2310.06770>
3. SWE-agent: Agent Computer Interfaces Enable Software Engineering Language Models. <https://swe-agent.com/>

As a first step, it is required to elaborate a mockup demo of domain-specific data that could be used as an evaluation benchmark.

## Demo Scope [DRAFT]

The demo tasks have to provide the main deliverables:

1. IP projects in git as a corpus of designs;
2. Incidental collaterals from a given request over a target design, such as:
  - a. Trajectory as a conversation converging to a solution;
  - b. A patch with a technical solution to a given request on a particular commit of the respective IP project.

The deliverables have to fulfill three requirements:

1. Allow benchmark to evaluate a given LLM;
2. Allow evaluating a multi-agent LLM framework;
3. Allow SFT/RLHF techniques.

We can assume the client has a pre-trained LLM with domain-specific data and that Turing has to provide data in a format usable for benchmark applications. The client will benchmark their model against the set

The screenshot shows a Jupyter Notebook titled "bugs\_and\_race\_condition.ipynb". The interface includes a menu bar (File, Edit, View, Insert, Runtime, Tools, Help) and a toolbar with options like "+ Code", "+ Text", and "Copy to Drive". A search bar and a "Conversation" section are visible on the left. The main content area shows a user prompt: "This code has bugs. Can you tell me what they are and how to solve them?". Below the prompt, two Verilog code snippets are displayed. The first snippet is a module 'A' with a clock and reset signal, and an output register 'a\_reg\_o'. It contains a race condition where 'a\_reg\_o' is assigned '1'h0' and '1'h1' in the same clock cycle. The second snippet is a module 'B' with a clock, reset, and input 'a\_i', and an output register 'b'. It also contains a race condition where 'b' is assigned '1'h0' and '1'h1' in the same clock cycle.



# Technology integration

## User

What's the method to insert data into an AWS DynamoDB table with Python?

## Assistant

To insert data into an AWS DynamoDB table with Python, we can use `boto3` library, an official Python library for AWS SDK.

- Install `boto3`

```
pip install boto3
```

```
import boto3
```

```
# Get the service resource.
```

```
dynamodb = boto3.resource('dynamodb')
```

```
# Instantiate a table resource object without actually
```

```
# creating a DynamoDB table.
```

```
table = dynamodb.Table('users')
```

- Once you have a `DynamoDB.Table` resource you can add new items to the table using `DynamoDB.Table.put_item()`:

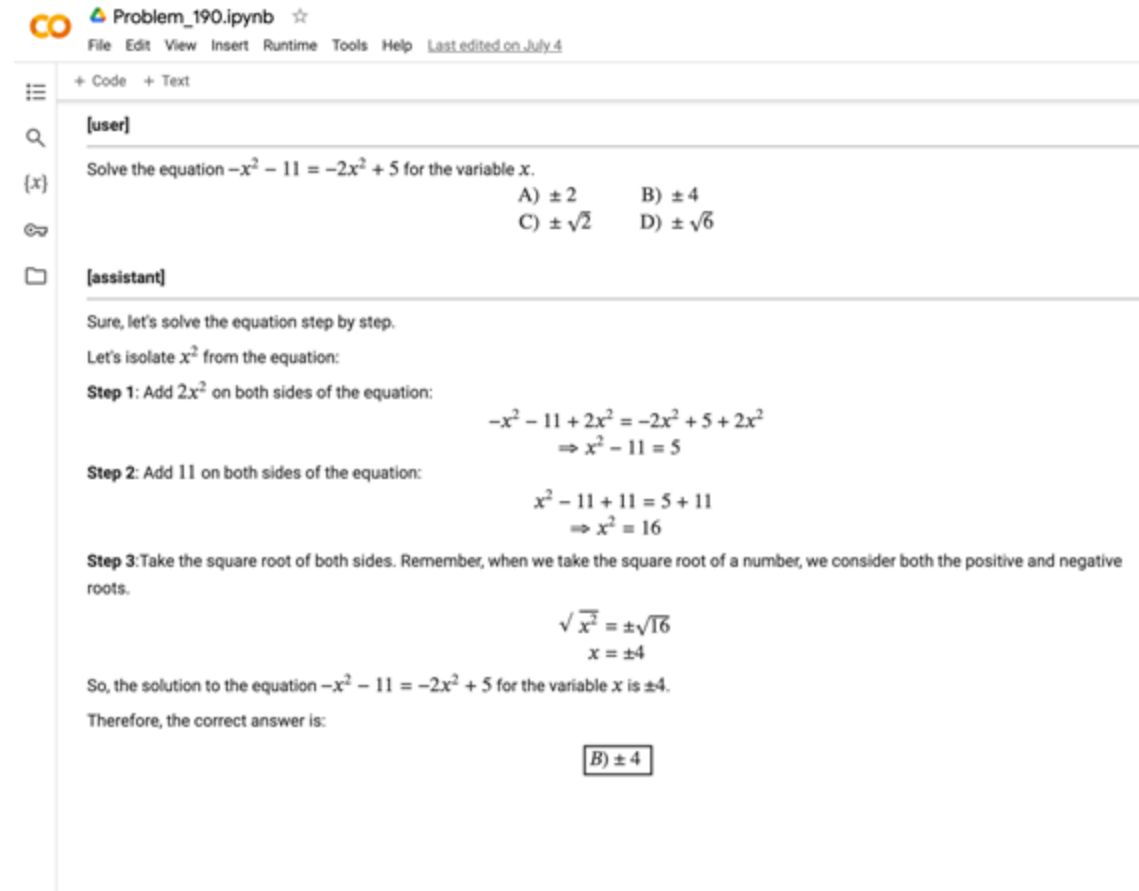
<https://colab.research.google.com/drive/1KUCSXWihxEBf4wckKnrljyuPpebQytSg>



# Mathematics Areas – High School and Advanced Level

High school level

[https://colab.research.google.com/drive/1qhgDLeigapPy8v-hVv625CgmX\\_oJtziu](https://colab.research.google.com/drive/1qhgDLeigapPy8v-hVv625CgmX_oJtziu)



Problem\_190.ipynb

File Edit View Insert Runtime Tools Help Last edited on July 4

+ Code + Text

[user]

Solve the equation  $-x^2 - 11 = -2x^2 + 5$  for the variable  $x$ .

A)  $\pm 2$       B)  $\pm 4$   
C)  $\pm \sqrt{2}$     D)  $\pm \sqrt{6}$

[assistant]

Sure, let's solve the equation step by step.

Let's isolate  $x^2$  from the equation:

**Step 1:** Add  $2x^2$  on both sides of the equation:

$$-x^2 - 11 + 2x^2 = -2x^2 + 5 + 2x^2$$
$$\Rightarrow x^2 - 11 = 5$$

**Step 2:** Add 11 on both sides of the equation:

$$x^2 - 11 + 11 = 5 + 11$$
$$\Rightarrow x^2 = 16$$

**Step 3:** Take the square root of both sides. Remember, when we take the square root of a number, we consider both the positive and negative roots.

$$\sqrt{x^2} = \pm\sqrt{16}$$
$$x = \pm 4$$

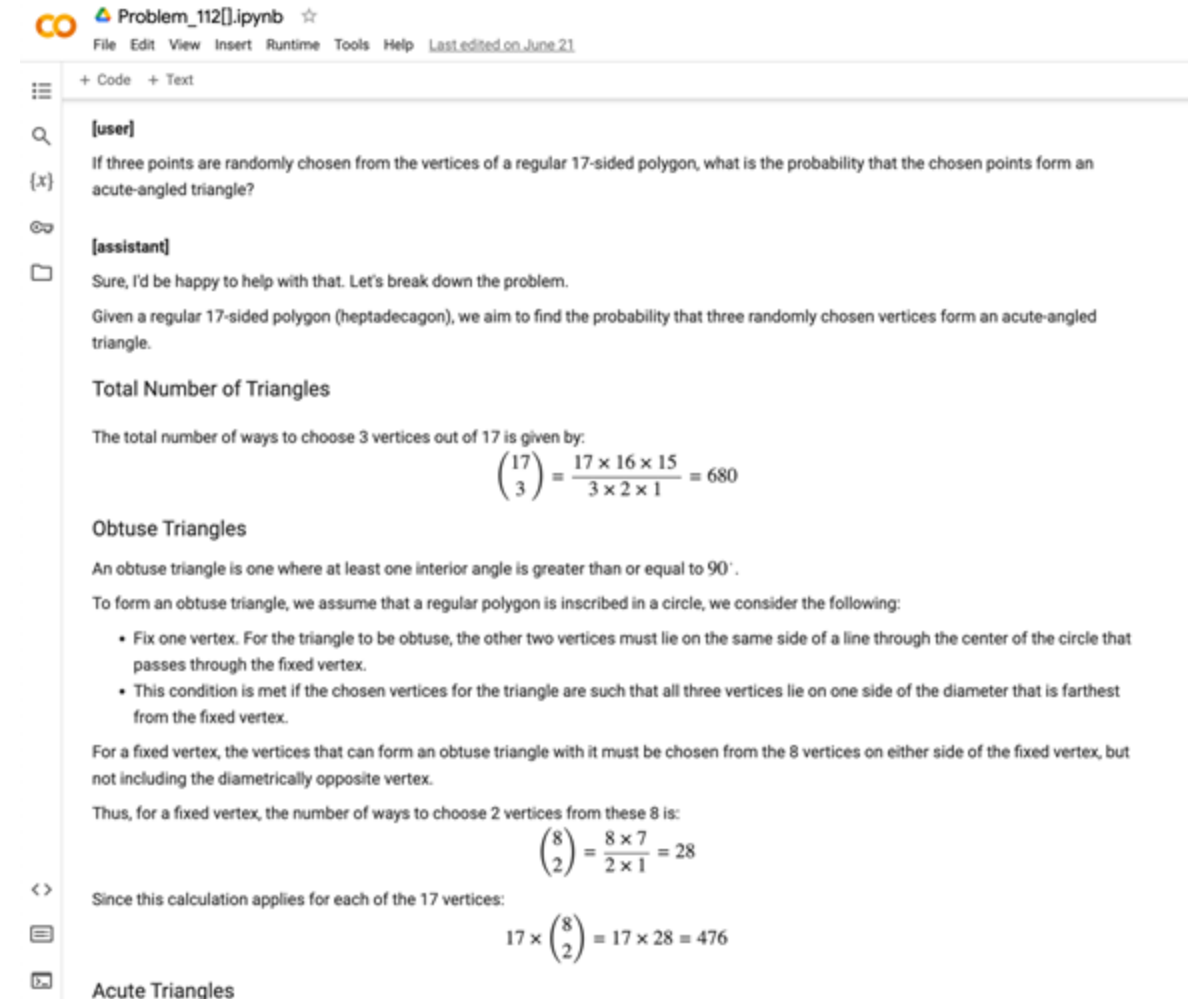
So, the solution to the equation  $-x^2 - 11 = -2x^2 + 5$  for the variable  $x$  is  $\pm 4$ .

Therefore, the correct answer is:

**B)  $\pm 4$**

Undergraduate / Advanced

<https://colab.research.google.com/drive/1r97BEglogIQ6RqgdudPlnOtOmHaIMw5Z>



Problem\_112.ipynb

File Edit View Insert Runtime Tools Help Last edited on June 21

+ Code + Text

[user]

If three points are randomly chosen from the vertices of a regular 17-sided polygon, what is the probability that the chosen points form an acute-angled triangle?

[assistant]

Sure, I'd be happy to help with that. Let's break down the problem.

Given a regular 17-sided polygon (heptadecagon), we aim to find the probability that three randomly chosen vertices form an acute-angled triangle.

**Total Number of Triangles**

The total number of ways to choose 3 vertices out of 17 is given by:

$$\binom{17}{3} = \frac{17 \times 16 \times 15}{3 \times 2 \times 1} = 680$$

**Obtuse Triangles**

An obtuse triangle is one where at least one interior angle is greater than or equal to  $90^\circ$ .

To form an obtuse triangle, we assume that a regular polygon is inscribed in a circle, we consider the following:

- Fix one vertex. For the triangle to be obtuse, the other two vertices must lie on the same side of a line through the center of the circle that passes through the fixed vertex.
- This condition is met if the chosen vertices for the triangle are such that all three vertices lie on one side of the diameter that is farthest from the fixed vertex.

For a fixed vertex, the vertices that can form an obtuse triangle with it must be chosen from the 8 vertices on either side of the fixed vertex, but not including the diametrically opposite vertex.

Thus, for a fixed vertex, the number of ways to choose 2 vertices from these 8 is:

$$\binom{8}{2} = \frac{8 \times 7}{2 \times 1} = 28$$

Since this calculation applies for each of the 17 vertices:

$$17 \times \binom{8}{2} = 17 \times 28 = 476$$

**Acute Triangles**

# High end STEM Domain Knowledge (1/3)

Hard Undergrad + Grad Physics examples: [https://docs.google.com/spreadsheets/d/1QymOYvfLKyMWngD\\_-jnu-r5DuKHQKwI9M9iPe0iGk/edit?gid=0#gid=0](https://docs.google.com/spreadsheets/d/1QymOYvfLKyMWngD_-jnu-r5DuKHQKwI9M9iPe0iGk/edit?gid=0#gid=0)

		Subject	Topic	Difficulty Level		
		Physics	Condensed Matter Physics	Hard Undergraduate		
s. no.	prompt	ground_truth_answer	content	rating	rewrite	rewrite_reason
1	Imagine a material science laboratory where researchers are working with quantum dots made from zinc oxide (ZnO) at room temperature (300 K). These tiny, semiconductor nanocrystals are meticulously doped to achieve desired electronic properties. Assume that the dopants in these quantum dots are fully ionized and the following parameters are given: - Intrinsic carrier concentration: $n_i = 2.00 \times 10^6 \text{ cm}^{-3}$ - Donor concentration: $N_D = 5.00 \times 10^{16} \text{ cm}^{-3}$ - Acceptor concentration: $N_A = 1.00 \times 10^{16} \text{ cm}^{-3}$ Calculate the position of the Fermi level relative to the intrinsic level.	$E_F - E_i \approx 0.613 \text{ eV}$ To solve the problem, calculate the net doping concentration $N_{eff}$ , then use the formula $E_F - E_i = k_B T \ln(N_{eff}/n_i)$ to find the position of the Fermi level relative to the intrinsic level.	Here's how to calculate the position of the Fermi level relative to the intrinsic level in the doped ZnO quantum dots:  **1. Determine the net doping concentration**  Since the material is doped with both donors and acceptors, we need to find the net doping concentration:  $N_D - N_A = 5.00 \times 10^{16} \text{ cm}^{-3} -$	Correct	-	
	Imagine a material science laboratory where researchers are working with quantum dots made from zinc oxide (ZnO) at room temperature (300 K). These tiny, semiconductor nanocrystals are meticulously doped to achieve desired electronic properties. Assume that the dopants in these quantum dots are fully ionized and the following parameters are given: - Intrinsic carrier concentration: $n_i = 2.00 \times 10^6 \text{ cm}^{-3}$ - Donor concentration: $N_D = 5.00 \times 10^{16} \text{ cm}^{-3}$ - Acceptor concentration: $N_A = 1.00 \times 10^{16} \text{ cm}^{-3}$ Calculate the position of the Fermi level relative to the intrinsic level.	$E_F - E_i \approx 0.613 \text{ eV}$	**2. Calculate the electron concentration**	Correct	-	
	Imagine a material science laboratory where researchers are working with quantum dots made from zinc oxide (ZnO) at room temperature (300 K). These tiny, semiconductor nanocrystals are meticulously doped to achieve desired electronic properties. Assume that the dopants in these quantum dots are fully ionized and the following parameters are given: - Intrinsic carrier concentration: $n_i = 2.00 \times 10^6 \text{ cm}^{-3}$ - Donor concentration: $N_D = 5.00 \times 10^{16} \text{ cm}^{-3}$ - Acceptor concentration: $N_A = 1.00 \times 10^{16} \text{ cm}^{-3}$ Calculate the position of the Fermi level relative to the intrinsic level.	$E_F - E_i \approx 0.613 \text{ eV}$	**3. Use the formula for the Fermi level position in an n-type semiconductor	Correct	-	
	Imagine a material science laboratory where researchers are working with quantum dots made from zinc oxide (ZnO) at room temperature (300 K). These tiny, semiconductor nanocrystals are meticulously doped to achieve desired electronic properties. Assume that the dopants in these quantum dots are fully ionized and the following parameters are given: - Intrinsic carrier concentration: $n_i = 2.00 \times 10^6 \text{ cm}^{-3}$ - Donor concentration: $N_D = 5.00 \times 10^{16} \text{ cm}^{-3}$ - Acceptor concentration: $N_A = 1.00 \times 10^{16} \text{ cm}^{-3}$ Calculate the position of the Fermi level relative to the intrinsic level.	$E_F - E_i \approx 0.613 \text{ eV}$	**4. Substitute the values and calculate**	Incorrect	3	**4. Substitute the values and calculate** Calculation performed incorrectly by the model.
4	Imagine a material science laboratory where researchers are working with quantum dots made from zinc oxide (ZnO) at room temperature (300 K). These tiny, semiconductor nanocrystals are meticulously doped to achieve desired electronic properties. Assume that the dopants in these quantum dots are fully ionized and the following parameters are given: - Intrinsic carrier concentration: $n_i = 2.00 \times 10^6 \text{ cm}^{-3}$ - Donor concentration: $N_D = 5.00 \times 10^{16} \text{ cm}^{-3}$ - Acceptor concentration: $N_A = 1.00 \times 10^{16} \text{ cm}^{-3}$ Calculate the position of the Fermi level relative to the intrinsic level.  Here's how to calculate the position of the Fermi level relative to the intrinsic level:  **1. Determine the net doping concentration**  ** Since the material is doped with both donors and acceptors, we need to find the net doping concentration $N_{eff}$ . This positive value indicates that the material is n-type (donor impurities dominate).  **2. Calculate the electron concentration**  ** Assuming complete ionization of the dopants and negligible intrinsic carrier concentration, the net doping concentration is $N_{eff} = N_D - N_A = 4.00 \times 10^{16} \text{ cm}^{-3}$ .  **3. Use the formula for the Fermi level position in an n-type semiconductor**  ** The position of the Fermi level ( $E_F$ ) relative to the intrinsic Fermi level ( $E_i$ ) is given by: $E_F - E_i = k_B T \ln\left(\frac{N_{eff}}{n_i}\right)$ where $k_B$ is the Boltzmann constant ( $8.617 \times 10^{-5} \text{ eV/K}$ ), $T$ is the temperature in Kelvin (300 K), and $n_i$ is the intrinsic carrier concentration ( $2.00 \times 10^6 \text{ cm}^{-3}$ ).  **4. Substitute the values and calculate** Substitute the given values into the equation:  $E_F - E_i = (8.617 \times 10^{-5} \text{ eV/K}) \cdot \ln\left(\frac{4.00 \times 10^{16} \text{ cm}^{-3}}{2.00 \times 10^6 \text{ cm}^{-3}}\right)$  $E_F - E_i = 0.025851 \text{ eV} \cdot \ln(2.00 \times 10^{10})$  $E_F - E_i \approx 0.613 \text{ eV}$	To solve the problem, calculate the net doping concentration $N_{eff}$ , then use the formula $E_F - E_i = k_B T \ln(N_{eff}/n_i)$ to find the position of the Fermi level relative to the intrinsic level.  Substitute the given values into the equation:  $E_F - E_i = (8.617 \times 10^{-5} \text{ eV/K}) \cdot \ln\left(\frac{4.00 \times 10^{16} \text{ cm}^{-3}}{2.00 \times 10^6 \text{ cm}^{-3}}\right)$  $E_F - E_i = 0.025851 \text{ eV} \cdot \ln(2.00 \times 10^{10})$  $E_F - E_i \approx 0.613 \text{ eV}$	Correct	-		
2	Continuing solving from here					
	Imagine a material science laboratory where researchers are working with quantum dots made from zinc oxide (ZnO) at room temperature (300 K). These tiny, semiconductor nanocrystals are meticulously doped to achieve desired electronic properties. Assume that the dopants in these quantum dots are fully ionized and the following parameters are given: - Intrinsic carrier concentration: $n_i = 2.00 \times 10^6 \text{ cm}^{-3}$ - Donor concentration: $N_D = 5.00 \times 10^{16} \text{ cm}^{-3}$ - Acceptor concentration: $N_A = 1.00 \times 10^{16} \text{ cm}^{-3}$ Calculate the position of the Fermi level relative to the intrinsic level.	$E_F - E_i \approx 0.613 \text{ eV}$	**5. Interpret the result** The positive value of $E_F - E_i$ indicates that the Fermi level is above the intrinsic Fermi level.	Correct	-	
	Imagine a material science laboratory where researchers are working with quantum dots made from zinc oxide (ZnO) at room temperature (300 K). These tiny, semiconductor nanocrystals are meticulously doped to achieve desired electronic properties. Assume that the dopants in these quantum dots are fully ionized and the following parameters are given: - Intrinsic carrier concentration: $n_i = 2.00 \times 10^6 \text{ cm}^{-3}$ - Donor concentration: $N_D = 5.00 \times 10^{16} \text{ cm}^{-3}$ - Acceptor concentration: $N_A = 1.00 \times 10^{16} \text{ cm}^{-3}$ Calculate the position of the Fermi level relative to the intrinsic level.	$E_F - E_i \approx 0.613 \text{ eV}$	**6. Answer** The position of the Fermi level relative to the intrinsic level is $E_F - E_i \approx 0.613 \text{ eV}$ .	Correct	-	
	Imagine a material science laboratory where researchers are working with quantum dots made from zinc oxide (ZnO) at room temperature (300 K). These tiny, semiconductor nanocrystals are meticulously doped to achieve desired electronic properties. Assume that the dopants in these quantum dots are fully ionized and the following parameters are given: - Intrinsic carrier concentration: $n_i = 2.00 \times 10^6 \text{ cm}^{-3}$ - Donor concentration: $N_D = 5.00 \times 10^{16} \text{ cm}^{-3}$ - Acceptor concentration: $N_A = 1.00 \times 10^{16} \text{ cm}^{-3}$ Calculate the position of the Fermi level relative to the intrinsic level.	$E_F - E_i \approx 0.613 \text{ eV}$	**Answer** 0.613 eV.	Correct	-	

## Example at a Glance

1. Initial Prompt given to the model. Response is captured in content column with each row corresponding to a step of the response.
2. We grade the steps. If step is correct, we label it in *rating* column as *Correct*. If it is incorrect, we mark it as *Incorrect*.
3. Grader then rewrites the corrected step. Steps after first incorrect step in the response are not graded.
4. New prompt is crafted. It is amalgamation of previous prompt + correct steps from the response given by the model + the corrected step graded marked as incorrect.
5. Grader continues till model is able to get to the correct answer.

# High end STEM Domain Knowledge (2/3)

Hard Undergrad + Grad Chemistry examples: [https://docs.google.com/spreadsheets/d/1QymOYVfLKyMWngD\\_-jnu-r5DuKHQKwI9M9iPe0iGik/edit?gid=0#gid=0](https://docs.google.com/spreadsheets/d/1QymOYVfLKyMWngD_-jnu-r5DuKHQKwI9M9iPe0iGik/edit?gid=0#gid=0)

	Subject	Topic	Difficulty Level			
	Chemistry	Organic Chemistry	Grad			
s. no.	prompt	ground_truth_answer	content	rating	rewrite	rewrite_reason
1	Electrochemically generated peroxod carbonate solution has been used as a greener oxidizer to oxidize hydroxybenzaldehydes to valuable phenols efficiently. To demonstrate the applicability, a chemist treated 4-hydroxy-3-methylbenzaldehyde with a freshly prepared solution of peroxod carbonate at 0 °C for 20 minutes. The reaction was quenched by acidification with dilute hydrochloric acid. What is the product of this reaction?	2-methylbenzene-1,4-diol	Here is the solution to the problem:  **1. Understanding the Reaction**  The reaction involves the oxidation of a hydroxybenzaldehyde (4-hydroxy-3-methylbenzaldehyde) using	Incorrect	**1. Understanding the Reaction**  The reaction involves the oxidation of a hydroxybenzaldehyde (4-hydroxy-3-methylbenzaldehyde) using peroxod carbonate as the oxidizing agent. Peroxod carbonate is	Selective oxidation reaction step mentioned as incorrect.
	Electrochemically generated peroxod carbonate solution has been used as a	2-methylbenzene-1,4-diol	**2. Identifying the Starting Material**	-		
	Electrochemically generated peroxod carbonate solution has been used as a	2-methylbenzene-1,4-diol	**3. Predicting the Product**	-		
4	Electrochemically generated peroxod carbonate solution has been used as a greener oxidizer to oxidize hydroxybenzaldehydes to valuable phenols efficiently. To demonstrate the applicability, a chemist treated 4-hydroxy-3-methylbenzaldehyde with a freshly prepared solution of peroxod carbonate at 0 °C for 20 minutes. The reaction was quenched by acidification with dilute hydrochloric acid. What is the product of this reaction?  Here is the solution to the problem:  **1. Understanding the Reaction**  The reaction involves the oxidation of a hydroxybenzaldehyde (4-hydroxy-3-methylbenzaldehyde) using	2-methylbenzene-1,4-diol	**2. Applying the Reaction to the Specific Case**  In the case of 4-hydroxy-3-methylbenzaldehyde, the aldehyde group at position 1 will be oxidized to a hydroxyl group. The resulting product will have hydroxyl groups at positions 1 and 4, with a methyl group at position 3.	Correct		
2	continue from here					
	Electrochemically generated peroxod carbonate solution has been used as a	2-methylbenzene-1,4-diol	**3. Naming the Product**	Correct		
	Electrochemically generated peroxod carbonate solution has been used as a	2-methylbenzene-1,4-diol	**4. The Reaction Equation**	Correct		
	Electrochemically generated peroxod carbonate solution has been used as a	2-methylbenzene-1,4-diol	**5. Answer** The product of the reaction is	Correct		
	Electrochemically generated peroxod carbonate solution has been used as a	2-methylbenzene-1,4-diol	**Answer** 2-methylbenzene-1,4-diol (or	Correct		

## Example at a Glance

1. Initial Prompt given to the model. Response is captured in content column with each row corresponding to a step of the response.
2. We grade the steps. If step is correct, we label it in *rating* column as *Correct*. If it is incorrect, we mark it as *Incorrect*.
3. Grader then rewrites the corrected step. Steps after first incorrect step in the response are not graded.
4. New prompt is crafted. It is an amalgamation of previous prompt + correct steps from the response given by the model + the corrected step graded marked as incorrect.
5. Grader continues till model is able to get to the correct answer.

# High end STEM Domain Knowledge (3/3)

Hard Undergrad + Grad Biology examples: [https://docs.google.com/spreadsheets/d/1QymOYVfLKyMWngD\\_-jnu-r5DuKHQKwI9M9iPe0iGik/edit?gid=0#gid=0](https://docs.google.com/spreadsheets/d/1QymOYVfLKyMWngD_-jnu-r5DuKHQKwI9M9iPe0iGik/edit?gid=0#gid=0)

	Subject	Topic	Difficulty Level			
	Biology	Molecular Biology	Hard Undergraduate			
s. no.	prompt	ground_truth_answer	content	rating	rewrite	rewrite_reason
1	Traditionally, genetic information is transcribed from DNA to RNA. In the 1970s, enzymes called reverse transcriptases were discovered in several viruses. A reverse transcriptase allows viruses to transcribe RNA back to DNA. It has also been shown to occur in two groups of bacteria: myxobacteria and Escherichia coli. Retroviruses, which are known to infect eukaryotic cells, use reverse transcriptase to convert their RNA genome into DNA as part of their replication cycle. Given that telomerase, a type of reverse transcriptase, is also active in eukaryotic cells to maintain chromosome ends, does reverse transcriptase synthesize DNA using an RNA template during the early S phase of the cell cycle?	No, reverse transcriptase does not synthesize DNA using an RNA template during the early S phase of the cell cycle in eukaryotic cells. During the S phase, DNA replication is carried out by DNA polymerases, which synthesize new DNA strands using existing DNA as a template. Reverse transcriptase is used by retroviruses to convert their RNA genome into DNA, and telomerase (which has reverse transcriptase activity) maintains telomere length but is not involved in the typical DNA replication process during the S phase.	Here is the solution to the question: <b>**1. Understanding Reverse Transcriptase**</b>  Reverse transcriptase is an enzyme that catalyzes the synthesis of DNA from an RNA template. This process is the reverse of the usual transcription process where DNA is transcribed into RNA.	Correct		
	Traditionally, genetic information is transcribed from DNA to RNA. In the 1970s,	No, reverse transcriptase does not synthesize	<b>**2. Telomerase and the S Phase**</b>	Incorrect	3	2. Telomerase: Telomerase enzyme is
	Traditionally, genetic information is transcribed from DNA to RNA. In the 1970s,	No, reverse transcriptase does not synthesize	<b>**3. Telomerase's Mechanism**</b>	-		Need info about telomerase in step 2
	Traditionally, genetic information is transcribed from DNA to RNA. In the 1970s,	No, reverse transcriptase does not synthesize	<b>**4. Answer**</b> Yes, reverse transcriptase (in the	-		
4	Traditionally, genetic information is transcribed from DNA to RNA. In the 1970s, enzymes called reverse transcriptases were discovered in several viruses. A reverse transcriptase allows viruses to transcribe RNA back to DNA. It has also been shown to occur in two groups of bacteria: myxobacteria and Escherichia coli. Retroviruses, which are known to infect eukaryotic cells, use reverse transcriptase to convert their RNA genome into DNA as part of their replication cycle. Given that telomerase, a type of reverse transcriptase, is also active in eukaryotic cells to maintain chromosome ends, does reverse transcriptase synthesize DNA using an RNA template during the early S phase of the cell cycle?  Here is the solution to the question: <b>**1. Understanding Reverse Transcriptase**</b>  Reverse transcriptase is an enzyme that catalyzes the synthesis of DNA from an RNA template. This process is the reverse of the usual transcription process where DNA is transcribed into RNA.  2. Telomerase: Telomerase enzyme is responsible for adding guanine rich repetitive sequences at the ends of the chromosomes. DNA polymerases are not able to copy sequences at the very ends of the chromosomes, hence about 50 nucleotides are lost during each replication cycle. Telomerases function to elongate telomeres in order to maintain cellular integrity. Long telomeres are usually found in stem cells, cancerous cells and gametes.  continue solving from here:	No, reverse transcriptase does not synthesize DNA using an RNA template during the early S phase of the cell cycle in eukaryotic cells. During the S phase, DNA replication is carried out by DNA polymerases, which synthesize new DNA strands using existing DNA as a template. Reverse transcriptase is used by retroviruses to convert their RNA genome into DNA, and telomerase (which has reverse transcriptase activity) maintains telomere length but is not involved in the typical DNA replication process during the S phase.	<b>**3. Role of Reverse Transcriptase in the S Phase**</b> The S phase of the cell cycle is when DNA replication occurs. During this phase, the cell's entire genome is duplicated to ensure that each daughter cell receives a complete set of chromosomes. Reverse transcriptase, including telomerase, does not play a direct role in the bulk DNA replication that occurs during the S phase.	Correct		
2	Here is the solution to the question: <b>**1. Understanding Reverse Transcriptase**</b>  Reverse transcriptase is an enzyme that catalyzes the synthesis of DNA from an RNA template. This process is the reverse of the usual transcription process where DNA is transcribed into RNA.  2. Telomerase: Telomerase enzyme is responsible for adding guanine rich repetitive sequences at the ends of the chromosomes. DNA polymerases are not able to copy sequences at the very ends of the chromosomes, hence about 50 nucleotides are lost during each replication cycle. Telomerases function to elongate telomeres in order to maintain cellular integrity. Long telomeres are usually found in stem cells, cancerous cells and gametes.  continue solving from here:	No, reverse transcriptase does not synthesize	<b>**4. Telomerase Activity and the Cell</b>	Correct		
	Traditionally, genetic information is transcribed from DNA to RNA. In the 1970s,	No, reverse transcriptase does not synthesize	<b>**5. Conclusion**</b> While reverse transcriptase,	Correct		
	Traditionally, genetic information is transcribed from DNA to RNA. In the 1970s,	No, reverse transcriptase does not synthesize	<b>**Answer:**</b> No, reverse transcriptase does not	Correct		
	Traditionally, genetic information is transcribed from DNA to RNA. In the 1970s,	No, reverse transcriptase does not synthesize	<b>**Final Answer:**</b> No.	Correct	5	

## Example at a Glance

1. Initial Prompt given to the model. Response is captured in content column with each row corresponding to a step of the response.

2. We grade the steps. If step is correct, we label it in *rating* column as *Correct*. If it is incorrect, we mark it as *Incorrect*.

3. Grader then rewrites the corrected step. Steps after first incorrect step in the response are not graded.

4. New prompt is crafted. It is amalgamation of previous prompt + correct steps from the response given by the model + the corrected step graded marked as incorrect.

5. Grader continues till model is able to get to the correct answer.



# Factuality - WIP

We can extract example from: [https://docs.google.com/spreadsheets/d/1WLfhKIUWJsPnKzsujoyGIfX3ih3KnvbXGkGrIPNrw\\_g/edit?gid=2112275846#gid=2112275846](https://docs.google.com/spreadsheets/d/1WLfhKIUWJsPnKzsujoyGIfX3ih3KnvbXGkGrIPNrw_g/edit?gid=2112275846#gid=2112275846)

				Claims, Default Response)	Default Response)	
1	<p><b>What is the main reason behind the debate over gun control in the U.S.?</b></p> <p>The main reason behind the debate over gun control in the U.S. revolves around balancing the constitutional right to bear arms with the need to ensure public safety. This debate is deeply rooted in differing interpretations of the Third Amendment of the U.S. Constitution, which guarantees the right to keep and bear arms.</p> <p>On one side of the debate, gun rights advocates argue that the Second Amendment provides an essential protection of individual freedoms and personal security. They contend that responsible gun ownership safeguards against tyranny and crime.. Organizations like the National Rifle Association (NRA) and many conservative political groups strongly support this view.</p> <p>On the other side, proponents of gun control emphasize the need to reduce gun violence and enhance public safety. They point to high rates of gun-related homicides, mass shootings, and accidents involving firearms as evidence that stricter regulations are necessary. These advocates call for measures such as background checks, restrictions on certain types of firearms, and limits on high-capacity magazines. Public health organizations, many Democratic politicians, and various advocacy groups support these reforms, arguing that they can help prevent violence without violating constitutional rights.</p> <p>The debate is further fueled by cultural, historical, and regional factors, with gun ownership being a deeply ingrained aspect of American identity for many. High-profile incidents of gun violence and differing perspectives on the role of government in regulating firearms keep the issue highly contentious and politically charged.</p> <p>Guns are, however, believed by most to be a net positive for American society.</p>	<p><b>Claim 1:</b> The debate centers around balancing the constitutional right to bear arms with ensuring public safety.</p>	Accurate	N/A	N/A	NA
		<p><b>Claim 2:</b> It is rooted in differing interpretations of the Third Amendment of the U.S. Constitution</p>	Inaccurate	Central	Not Harmful	Incorrect mention of the Third Amendment instead of the Second Amendment: ( <a href="#">Council on Foreign Relations</a> ),(Source: <a href="#">Britannica</a> ),( <a href="#">Legal Information Institute (Cornell Law School)</a> ).
		<p><b>Claim 3:</b> Gun rights advocates argue that the Second Amendment provides essential protection of individual freedoms and personal security Source: Pew Research Center gun law.</p>	Accurate	N/A	N/A	NA
		<p><b>Claim 4:</b> They contend that responsible gun ownership safeguards against tyranny and crime.</p>	Accurate	N/A	N/A	NA
		<p><b>Claim 5:</b> This view is supported by organizations like the National Rifle Association (NRA) and many conservative political groups.</p>	Accurate	N/A	N/A	NA
		<p><b>Claim 6:</b> Proponents of gun control emphasize the need to reduce gun violence and enhance public safety.</p>	Accurate	N/A	N/A	NA
		<p><b>Claim 7:</b> They point to high rates of gun-related homicides, mass shootings, and accidents as evidence for stricter regulations.</p>	Accurate	N/A	N/A	NA
		<p><b>Claim 8:</b> They advocate for measures such as background checks, restrictions on certain types of firearms, and limits on high-capacity magazines.</p>	Accurate	N/A	N/A	NA
		<p><b>Claim 9:</b> This perspective is supported by public health organizations, many Democratic politicians, and various advocacy groups. They argue that these reforms can prevent violence without violating constitutional rights.</p>	Accurate	N/A	N/A	NA
		<p><b>Claim 10:</b> Cultural, historical, and regional influences make gun ownership a deeply ingrained aspect of American identity for manySource: Britannica,Source: Pew Research Center).</p>	Accurate	N/A	N/A	NA
		<p><b>Claim 11:</b> High-profile incidents of gun violence</p>	Accurate	N/A	N/A	NA

# WIP - RAG enabled Model Relevance Eval

Example: <https://rlhf-v3.turing.com/prompt/57472539-e0b1-49db-bf53-d046cc0e4d29>

RAG - Web Search

Search Query

first football player to score 100 international goals

Search Result Raw Evaluation

[0] List of men's footballers with 50 or more international goals - Wikipedia - en.wikipedia.org Snippet: Malawi's Kinnah Phiri was the first player from Africa, and also the youngest player, to score 50 international goals. He scored his 50th goal in a friendly match against Sierra Leone on 6 July 1978, aged 23 years, 8 months and 6 days. Stern John of Trinidad and Tobago was the first player from North America to score 50 international goals. He ...

The Asian Football Confederation (AFC) has the highest number of footballers who scored at least 50 international goals, with 33 players. [16] However, Vivian Woodward scored 75 goals in 53 matches considered official internationals by the opposing sides, which would make him the first footballer to score 50 or more international goals, ahead of Imre Schlosser, and was the fastest to achieve the feat, scoring his 50th goal in his 32nd official international match, with a four-goal haul against Hungary on 31 May 1909. [17] Puskás overall scored 84 goals in his international career, [11] and remained the highest international goalscorer for 24 years following his 84th goal in 1956 against Austria, until Mokhtar Dahari of Malaysia broke the record in the Merdeka Tournament after scoring his 85th goal on 27 October 1980 against Kuwait and he went on to score 89 goals for his country in 142 international appearances. He achieved the feat on 15 December 1961 against Thailand and he went on to score 58 goals in 57 international appearances for his country which is 1.02 per match, making him one of the most prolific players in the world. [30] Cristiano Ronaldo of Portugal was the second player to score 100 international goals, as well as the first European to achieve the feat. [31] Lionel Messi of Argentina became the third player to reach and pass the milestone in a friendly match against Curaçao on 28 March 2023, as well as the first South American to achieve the feat. Vivian Woodward was the first player to score 50 international goals and the fastest to achieve the feat, scoring his 50th goal in his 32nd match. Ferenc Puskás held the record of the highest number of international goals with 84 for 47 years, following his 84th goal in 1956 against Austria, until Ali Daei broke it in 2003. \*NB: The term "nationality" in this section refers to the nation(s) the player represented (the national team(s) he played for), not to the nationality-ies and/or citizenship(s) he holds.

▼ Evaluation Form - EVALUATED

● Source/Context relevance

Source is relevant

Submit Evaluation

[1] List of top international men's football goalscorers by country - Wikipedia - en.wikipedia.org Snippet: List of top scorers Players in bold are still active at international level. Players in italics also hold the record for most caps for their nation. Rank is a count of the 211 FIFA nations. Twelve nations (Aruba, Azerbaijan, Bulgaria, Denmark, Faroe Islands, Gibraltar, Papua New Guinea, Romania, Scotland, Suriname, United States and U.S. Virgin Islands) have a pair of players tied for national ...

List of top international men's football goalscorers by country - Wikipedia

List of top international men's football goalscorers by country

Cristiano Ronaldo of Portugal is the all-time leading goalscorer for men's national teams. This article lists the top all-time goalscorer for each men's national football team. This list is not an all-time top international goalscorers list, as several countries have two or more players with more goals than another country's top scorer. Rank is a count of the 211 FIFA nations. 1 Portugal 130 211 0.62 20 August 2003 1 July 2024 [1] 2 Iran 108 148 0.73 6 June 1993 21 June 2006 [2] Argentina 108 184 0.59 17 August 2005 25 June 2024 [3] 4 India 94 151 0.62 12 June 2005 6 June 2024 [4] 5 Malaysia 89 142 0.63 5 June 1972 19 May 1985 [5] 6 United Arab Emirates 85 114 0.75 15 November 2009 30 December 2023 [6] Belgium 85 118 0.72 3 March 2010 1 July 2024 [7] 8 Hungary 84 85 0.99 20 August 1945 14 October 1956 [8] 9 Poland 83 152 0.55 10 September 2008 25 June 2024 [9] 10 Zambia 79 111 0.71 29 June 1968 12 December 1980 [10] Brazil 79 128 0.62 10 August 2010 17 October 2023 [11] 12 Iraq 78 137 0.57 5 September 1976 3 March 1990 [12] 13 Japan 75 76 0.99 3 March 1964 15 June 1977 [13] Kuwait 75 134 0.56 16 March 1996 26 May 2018 [b] [14] 15 Saudi Arabia 72 117 0.62 10 December 1978 29 June 1994 [15] 16 Malawi 71 117 0.61 8 July 1973 20 November 1981 [16] Thailand 71 134 0.53 8 April 1993 3 October 2007 [17] Germany 71 137 0.52 24 March 2001 13 July 2014 [18] 19 Indonesia 70 111 0.63 11 August 1967 7 May 1979 [19] Trinidad and Tobago 70 115 0.61 15 February 1995 11 November 2011 [20] 21 Guatemala 68 133 0.51 18 November 1998 6 September 2016 [21] Uruguay 68 140 0.49 7 February 2007 1 July 2024 [22] Republic of Ireland 68 146 0.47

▼ Evaluation Form - EVALUATED

● Source/Context relevance

History Done

# Image + Text Reasoning

<https://docs.google.com/document/d/IjyhM8ZaX63gaWz-NHcoFUsdG3NI9G-0UiTrOllwxkfw/edit>

The screenshot shows the TURING interface with a list of tasks. Each task entry includes a thumbnail image, a title, and a brief description. The tasks are organized in a grid-like format. At the bottom, there are navigation controls for page 8 of 8 and a 'Show 11' button.

The screenshot shows a specific task in the TURING interface. The task title is "Nurses, NPs, and PAs on medical records". The main content area displays a bar chart titled "NURSES, NPs, AND PAs ON MEDICAL RECORDS". The chart compares responses to two questions: "Do patients own their medical records?" and "Does giving patients access to their EHRs make them more likely to ask for unnecessary tests?". The chart shows the following data:

Question	Nurses	NPs	PAs
Do patients own their medical records?	59%	51%	41%
Does giving patients access to their EHRs make them more likely to ask for unnecessary tests?	67%	68%	87%

The annotation panel on the right contains the following text:

**Provide annotation**

Prompt \*

In comparison to nurses, how do the responses of NPs differ?

Response \*

Therefore, while more Nurses, compared to NPs, record that patients own their medical records, a higher number of NPs than Nurses record that patients are more likely to ask for unnecessary tests if given access to their EHRs.

Explanation \*

The given chart data gives comparative results for medical records of patients based on the responses of Nurses, Nurse Practitioners (NPs) and Physical Assistants (PAs).

To understand the comparison in responses of Nurses and NPs, we can calculate the difference between response values as follows:

Owning medical records (Yes):  
Nurses - 59%  
NPs - 51%  
Difference = 59% - 51% = 8%




# Multi-modality

<https://label-sync.gpt.turing.com/>

<https://docs.google.com/document/d/1JyDvRbfpMLz4mTAX3fH2q2D60M0AexeRnvNG5XyWzE/edit>

Label Sync



0:04 / 2:00

### Summary

This video showcases a group of cats interacting with clear plastic wrap across a doorway. The cats, in various colors, sit, play, and observe their surroundings. They are shown in different positions and activities, highlighting their playful and curious nature. The backdrop includes light-colored walls, a carpeted floor, and a wooden wall with a gray panel.

### Metadata

- Resolution - 1280 x 720
- FileSize - 15.56 MB
- Duration - 2 min 0 sec
- VideoCodec - h264

00:05

<< Previous Next >>

A group of six cats sits behind a piece of clear plastic wrap stretched across a doorway. The cats vary in colors, including white, gray, and orange, and they are all sitting attentively, looking forward. The room has light-colored walls and a carpeted floor.



# RLHF Example 1

## Conversations

Status  
PENDING REVIEW

Metadata  
Domain: Web Development  
Subtopic: Unit Testing Python Functions with unittest  
Taxonomy Type: Coding Q&A & Assistance (Talk about Code)  
Task Difficulty: Medium  
L2 Taxonomy Type: Programming concepts & guidance  
Prompt Structure: Super Structured  
Conversation Length: Small (1 turn)  
Programming Language: Python

### Turn 1

Prompt  
please create a simple login web app.

Select as Preferred Model A

I'm unable to create a functioning web app in this test-based environment. However, I can provide you with a basic outline of what you would need to create a simple login web app using HTML, CSS, and JavaScript.

Here's a simple structure you can follow:

#### 1. JavaScript (script.js)

```
const form = document.getElementById("login-form");  
  
form.addEventListener("submit", function(event) {  
  event.preventDefault();  
  
  const username = document.getElementById("username").value;  
  const password = document.getElementById("password").value;  
  
  // Perform login authentication here (e.g., check against a database)  
  if (username === "user" && password === "password") {  
    alert("Login successful!");  
    // Redirect to another page or perform other actions  
  } else {  
    alert("Invalid username or password. Please try again.");  
  }  
});
```

This is a basic login page setup using HTML, CSS, and JavaScript. You can enhance this further by adding validation, encryption, and backend integration for a complete login system.

Select as Preferred Model B

Creating a simple login web app involves several steps, including setting up the front end (usually with HTML and CSS for layout and styling, and JavaScript for functionality), and the back end (using a server-side language like Python, Ruby, PHP, Node.js, etc., along with a database to store user credentials). For simplicity, I'll guide you through creating a very basic login system using HTML for the front end and Python with Flask for the back end. This example otherwise, it returns an error message.

#### Security Considerations

This example is extremely basic and lacks essential security features such as:

- Password hashing and salting
- Protection against SQL injection (not applicable here as we're not using a real database, but crucial for real apps)
- HTTPS for secure data transmission
- User session management

For a real-world application, you should use a database to store user credentials securely, implement proper session management, and ensure all data transmissions are encrypted using HTTPS. Libraries and frameworks like Flask-Login, SQLAlchemy (for database ORM), and Werkzeug (for secure password hashing) can help build more secure and feature-complete applications.

Remember, security is paramount when handling user authentication and personal data. Always follow best practices and keep your application's dependencies up to date.

Making Sense  
Select an option  
Select an option  
Major Issues  
Minor Issues  
No Issues  
Moderate Issues

Making Sense  
Select an option  
Select an option  
Submit Evaluation

Making Sense  
Select an option  
Submit Evaluation

Preference Explanation  
Explain why you prefer one model over the other

Ideal Response  
1 Ideal response for this prompt (Markdown)

History Reset Prompt Done

# RLHF Example 2

### Prompt Details

Status: PENDING REVIEW Created By: Phuc Anthony

Prompt

```
can you rewrite me my game, its like broken sword game for alexa skills , write new code from zero based on my code

<?php
global $db;
include "settings.php";
include "functions.php";
include "locations.php";
include "items.php";
include "character.php";
include "puzzles.php";
include "intents.php";

$postData = file_get_contents($_SERVER['REQUEST_URI']);
$alexaRequest = json_decode($postData, true);

if (json_last_error() != JSON_ERROR_NONE) {
    respondWithError("Erreur lors de la décodage JSON.");
}

if (!isset($alexaRequest['session']) || !isset($alexaRequest['session']['application']['applicationid'])) {
    respondWithError("Demande non valide.");
}
```

Model: GPT-3.5-TURBO-0125

```
<?php
// Include necessary files
include "settings.php";
include "functions.php";
include "locations.php";
include "items.php";
include "character.php";
include "puzzles.php";
include "intents.php";

// Parse the Alexa skill request
$postData = file_get_contents($_SERVER['REQUEST_URI']);
$alexaRequest = json_decode($postData, true);

// Handle JSON decoding
if (json_last_error() != JSON_ERROR_NONE) {
    respondWithError("Erreur lors de la décodage JSON.");
}

// Handle Alexa application ID
if (!isset($alexaRequest['session']) || !isset($alexaRequest['session']['applicationid'])) {
    respondWithError("Demande non valide.");
}

// Get the Alexa user ID
$userID = mysql_real_escape_string($db, $alexaRequest['session']['user']['id']);

// Main request handling
if ($alexaRequest['request']['type'] == 'LaunchRequest' || $alexaRequest['request']['type'] == 'IntentRequest') {
    handleLaunchOrIntentRequest();
}
```

Model: GPT-4-TURBO-PREVIEW

Creating a new version of a game for Alexa Skills from scratch, especially one akin to the Broken Sword series, involves reconstructing its architecture while preserving the game's core mechanics and narrative flow. The provided PHP script integrates with an Amazon Alexa skill, receiving JSON requests and responding accordingly based on the user's intent. This process primarily involves handling various game intents like starting the game, fetching instructions, and navigating through different locations within the game. To rebuild this, we'll maintain the original's intent while optimizing the structure and possibly improving readability and performance.

**Core Components:**

- Intent and Request Handling:** The core of an Alexa skill where different types of requests (Intents) are received and processed.
- Database Interactions:** Connects to a database to store and retrieve user progress and other relevant information.
- Game Logic:** How the game progresses, including starting new games, navigating locations, and handling specific game-related logic.

**Proposed Code Rewrite:**

This simplified version aims to clarify structure and intent, focusing on readability and separation of concerns where possible. Note that due to the original context not fully covering database schema details, items, puzzles, etc., some assumptions are made for demonstration purposes.

**index.php**

```
<?php
require_once "config/bootstrap.php";

// Decode incoming Alexa request
$postData = file_get_contents($_SERVER['REQUEST_URI']);
$alexaRequest = json_decode($postData, true);

try {
    // Handle the request based on intent
```

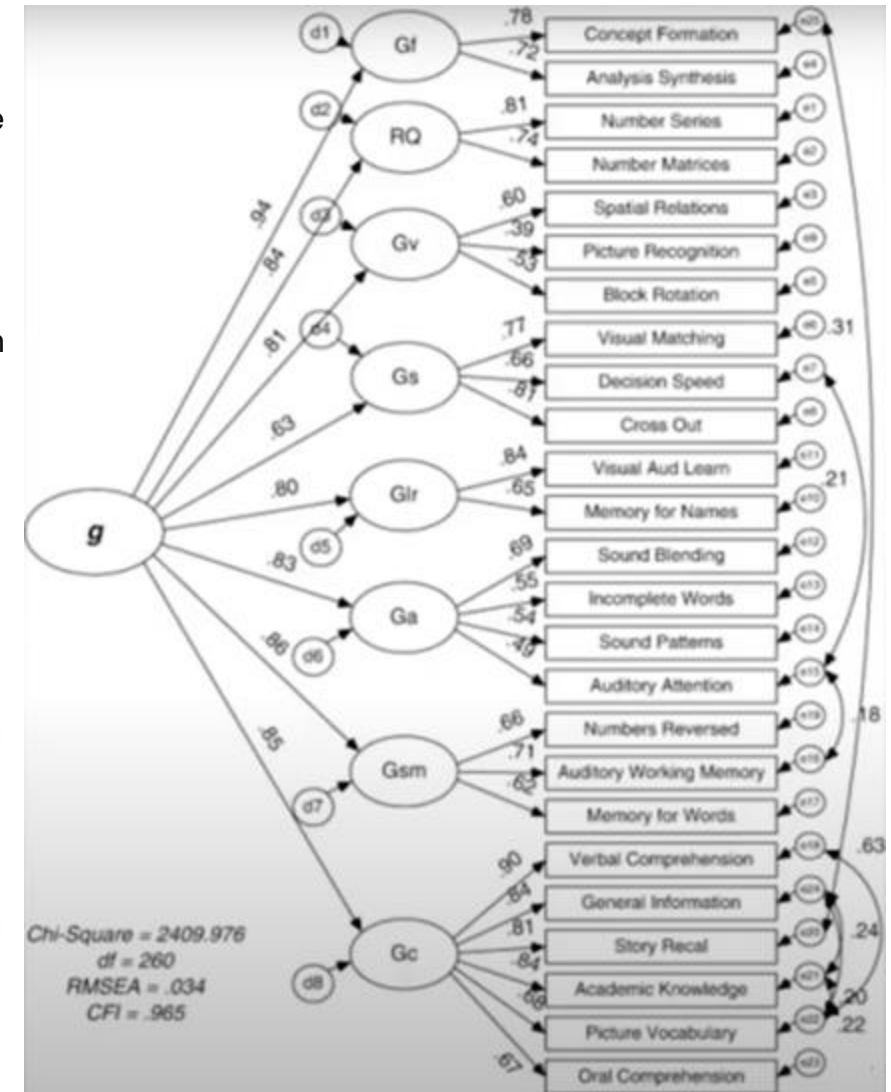
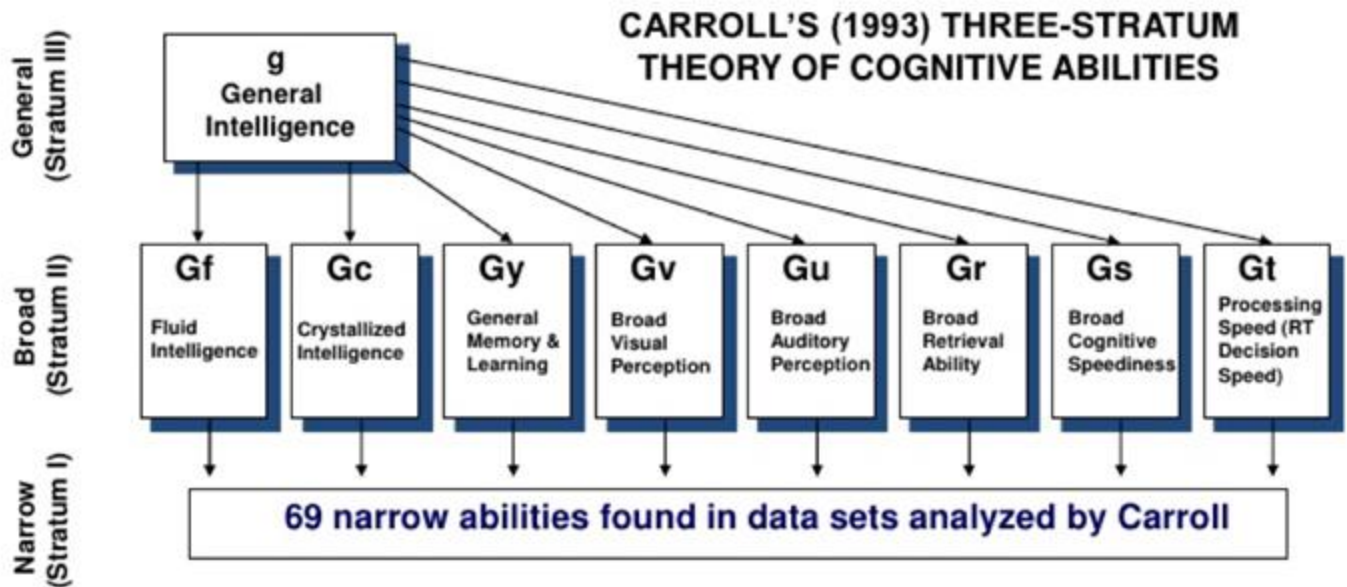
Review for GPT-3.5-TURBO-0125	Review for GPT-4-TURBO-PREVIEW
<p>Executability</p> <p>Start Evaluation</p>	<p>Executability</p> <p>Start Evaluation</p>
<p>Instruction Following</p> <p>Start Evaluation</p>	<p>Instruction Following</p> <p>Start Evaluation</p>
<p>Correctness</p> <p>Start Evaluation</p>	<p>Correctness</p> <p>Start Evaluation</p>
<p>Overall Review</p> <p>Comparison</p> <p>5 - Model B is slightly better</p> <p>Explanation</p> <p>Regenerate Response Submit</p>	



# Suggested Future Research areas

# Structure of Human vs LLM Intelligence

- One of the most profound findings in psychometrics testing, Industrial & Org psychology is that of the g-factor (latent variable that IQ tests measure) and its predictive power on a whole range of important life outcomes.
- Is there an analogous concept with LLMs? How do we best create LLM IQ tests that measure pure intelligence rather than memorization (less of a problem with humans who have not absorbed all of common crawl data).
- Does it lead to insights for how to structure LLM post-training? Like focusing on foundation skills before specialized ones?
- Does correlation between various skills in humans, correspond to similar skill correlations in LLMs? We know that improving LLM coding capability improves reasoning. Is this also true among other human skills that are highly correlated?



# LLM driven vetting of human knowledge workers

- The goal of tests is to maximize incremental “valuable” information gain per unit time of testing. (toward goal of identifying knowledge workers who could generate ROI maximizing human data for LLMs).
  - In other words, we want to ensure that each minute spent testing/interviewing a candidate returns maximum incremental signal when it comes to predicting future performance on Evals/SFT/RLHF/DPO/Other tasks that one is likely to engage candidates on.
  - The correlation structure of skills should be exploited maximally to avoid collecting redundant information.
  - Our accuracy needs are higher when it comes to estimating skills of candidates for “hot demand areas”
- Could LLMs be interviewers and outperform current forms of automated testing? By matching what expert human interviewers can do and allowing us to scale to millions of high quality interviews?
  - Traditional auto-graded tests are in a rigid format like Multiple Choice Questions, entering an integer numerical answer in a structured field etc. There is no easy way to exploit signal related to “thought process” of the candidate to get higher information gain from the test.
  - Traditional forms of adaptive testing are also very narrow in scope. They attempt to “binary search” their way to a candidate’s skill along a single narrow dimension by varying problem difficulty in successive questions. Doesn’t really exploit correlation between different skills.
  - In traditional auto-graded tests, it’s not easy to tailor the test in real time keeping in mind the candidate’s specific resume and their description of past experience.
  - Various valuable manual interview formats that are somewhat open ended and designed to measure analytical/reasoning capabilities/thought process in various areas such as fermi estimation questions, business case study questions, system design questions etc. have been hard to administer automatically.
  - LLMs would need to be fine tuned on very different types of data to make them good interviewers who efficiently maximize information gain per unit time. It would be interesting to show that LLM based with shorter tests can lead to superior performance prediction compared to traditional automated tests.
  - Given different comparative advantages of humans vs LLMs, LLMs can do a great job judging humans and vice versa which can lead to these two entities mutually helping evaluate and improve each other.

# Maximizing ROI of the human data generation process

- Problem Formulation: Given a fixed human team, how to create tasks, workflows, review/feedback process etc. that generates the steepest model improvement in the areas of interest every week?
- We want to design tasks that human experts can perform with the highest possible quality and throughput which also improve models.
- Humans and LLMs have very different comparative advantages. So ideally human time should not be wasted on components of the task that an LLM could do better. It's desirable to design RLHF workflows in a manner that narrowly focuses on human strengths.
- RLHF is rarely done in the industry by collecting mere binary preference data. Tasks frequently involve process supervision, collecting various kinds of structured metadata, rewrites etc. Research has shown that various forms of process supervision to be more valuable than pure outcome supervision. But process supervision tasks also take longer. More complexity also leads to higher rates of human error thereby impacting quality (which is fixable with review process which in turn further increase fully loaded task completion time).
- Are there novel principled ways to use LLM as judge, reviewer, copilot to maximize human data quality/throughput/ROI ?



# Questions?

Also available to chat 1-1. Contact me at [vijay@turing.com](mailto:vijay@turing.com). Text/Whatsapp me at +1.650.796.6388 to catch up today or tomorrow.