# Title Page:

CSC 659 859 AI Ethics and Explainability
San Francisco State University

Team project: Development and Ethics and Trustworthiness Audit of AI Application
Option B - GenAI

Team 1
Harris Chan, ychan@sfsu.edu
Maeve Fitzpatrick, mfitzpatrick@sfsu.edu
Zari Haidarian, zhaidarian@sfsu.edu

Date: 08/06/2025

**Table of Content:**

## Section 1: Executive summary

### 1.1 Goal of this report

The goal of this report is to see if AI (GPT-4) can reliably answer and explain computer science-related multiple-choice questions for the purpose of helping computer science students learn and study.

### 1.2 Problem description and its importance

We have found both benefits and drawbacks of multiple-choice questions; they are easier to grade for teachers. However, the answers are much easier for students to guess, which defeats the purpose of an exam meaning to evaluate how much a student has learned. Through this project, we aim to maintain the efficiency with which multiple-choice questions can be graded, which then would get the exams back to the students sooner to understand what they do or do not have a good handle on yet. If the GenAI can successfully, accurately and fully explain the answers to multiple-choice questions, this would overall improve the learning and teaching experiences of students and teachers.

### 1.3 Data available and technologies and tools used

We utilized a GitHub repository from MCQ-World to source our computer science multiple-choice questions, the answer key to which will be our ground truth. We are evaluating the ability of OpenAI's GPT-4 to explain our questions.

### 1.4 Methods of audit

To audit accuracy, we compared the model's predicted option against the ground-truth key, computed overall and per-topic accuracy rates, and visualized results in a confusion matrix. For trustworthiness, we sampled misclassifications and categorized them (e.g. misinterpretation, plausible-but-wrong, hallucination) to understand failure modes, and we performed a lightweight bias/fairness/privacy check to ensure no sensitive data leakage or unintended demographics effects.

### 1.5 Summary results

Overall accuracy: 38/50 correct → 76%

### 1.6 Recommendations of your work

We have found that using GPT-4 as an "AI tutor" was indeed helpful in terms of adding supplemental explanation. However, the tool was not accurate enough to be recommended as a grader. As long as the tool is used to supplement knowledge that is at least partially known already or can be checked with answers from the teacher, it is okay to use, but it is not recommended to rely on this tool to determine correct answers.

## Section 2: Motivation, Problem Description & Case‑Study Goals

### 2.1 Chosen GenAI Technology

We will evaluate OpenAI's GPT‑4 series, accessed via the OpenAI Python SDK. GPT‑4 is a large‑scale Transformer model pre‑trained on diverse web‑scale corpora and fine‑tuned with reinforcement learning from human feedback (RLHF). Its zero‑shot and few‑shot capabilities make it a natural candidate for natural‑language reasoning tasks without additional custom training.

### 2.2 Application Domain: Automated MCQ Answering

Our case study focuses on multiple‑choice questions (MCQ) answering in undergraduate computer‑science courses. Frequently  test concepts such as:

- Algorithmic complexity (e.g., "What is the worst‑case time for quicksort?")
- Data‑structure behavior (e.g., "Which structure implements FIFO?")
- Operating‑system fundamentals (e.g., "What does a system call trap into?")
- Database design (e.g., "Which normal form prevents partial dependencies?")

We will build a Jupyter‑notebook prototype that ingests a question stem and four labeled options (A–D), queries GPT‑4 to select the best answer, and returns a one‑sentence explanation of its choice.

### 2.3 Motivation & Impact

1. Pedagogical efficiency: Automating MCQ feedback can dramatically reduce grading effort and turnaround time, giving students immediate insight into *why* an answer is correct or incorrect.
2. Assessment enrichment: Pairing each selected option with a rationale deepens conceptual understanding, moving beyond binary right/wrong marking.
3. Feasibility of prompt‑based ML: By framing MCQ answering as a zero‑shot classification task, we explore how far off‑the‑shelf LLMs can go in reasoning about structured academic content without bespoke fine‑tuning.
4. Ethical imperative: Deploying LLMs in education carries risks—hallucinated explanations may mislead learners; overconfident wrong answers can erode trust; careless logging may expose quiz identifiers. A rigorous audit is essential before any educational rollout.

### 2.4 Data & Decision Points

- Data source: We will draw our test questions from the MCQ‑World GitHub repository (patil2104/MCQ‑World), which is MIT‑licensed and contains several hundred CS multiple‑choice questions organized into eight topic folders (DSA, Operating System, Computer Networks, DBMS, OOP, Java, Language Processors & Compilers, AI/ML) GitHub. From this pool we will sample 30–50 questions, ensuring coverage across all major topics.
- Ground truth: Each question's official correct answer is provided in the repository's answer keys; we will use those as our "gold standard."
- ML decisions:
    1. Answer selection: The model must classify which of the four options (A–D) is correct.
    2. Explanation synthesis: The model must generate a concise, accurate one‑sentence rationale for its choice—avoiding any hallucinated or extraneous detail.

Both decision steps are risky (an incorrect or misleading rationale can reinforce student misconceptions) and impactful (learners may accept the system's output without independent verification), underscoring the need for careful evaluation and ethics auditing.

### 2.5 Specific Case‑Study Goals

1. Accuracy measurement: Compute overall MCQ classification accuracy (%) and construct a confusion matrix to identify which distractors most frequently mislead the model.
2. Error analysis: Manually review all misclassifications, categorizing errors into misinterpretation, plausible-but-wrong rationale, or outright hallucination.
3. Ethics & trustworthiness audit:
    - Bias & fairness: Assess performance variation across topics or question styles.
    - Transparency: Verify that explanations include uncertainty cues (e.g. "I think…").
    - Privacy: Ensure no Canvas identifiers are logged or exposed.
    - Human oversight: Confirm the notebook reminds users to verify AI outputs before relying on them.

By achieving these goals, we will provide a clear picture of GPT‑4's viability for automated CS MCQ assistance, along with actionable recommendations for safe, responsible classroom integration.

## Section 3: GenAI Technology & Application to Be Evaluated

### 3.1 Application Overview

Our prototype is an Automated MCQ Answerer built entirely in a Jupyter notebook. It:

1. Loads one question at a time (stem + four options) from a CSV
2. Constructs a chat prompt for GPT-4, including a system message, a few-shot template, and the actual question
3. Calls the OpenAI API (model gpt-4o-mini)
4. Parses the model's reply to extract the chosen option letter and explanation
5. Displays results with accuracy % and Confusion Matrix

This single-notebook "app" demonstrates end-to-end GenAI development without any external UI or server.

### 3.2 GenAI Technology & Environment

- Model: GPT-4 series (gpt-4o-mini endpoint)
- SDK: OpenAI Python SDK v0.27+
- Runtime: Python 3.10 within JupyterLab (CPU only)
- Dependencies: openai, pandas, numpy, matplotlib (for confusion-matrix plotting)
- Reproducibility: We record the exact model identifier, SDK version, notebook commit, and timestamp for every API call (see Appendix II).

### 3.3 Prompt Customization & Few-Shot Example

#### 3.3.1 System Message

```
You are an expert computer-science tutor. Given a multiple-choice
question with four options, select the correct answer and
provide a one-sentence rationale in clear, beginner-friendly
language.
```

#### 3.3.2 API Call & Parsing

```
resp = openai.ChatCompletion.create(model="gpt-4o-mini",
messages=prompt)
raw = resp.choices[0].message.content
# Extract “Answer: X)” via regex, then split on “Explanation:”
for rationale
```

### 3.3.3 Few-Shot Template

We prepend two illustrative Q/A pairs so the model learns the exact format:

```
prompt = [
  {"role": "system", "content": system_message},

  # Example 1
  {"role":"user", "content":
    "Q: Which data structure follows FIFO?\n"
    "A) Stack   B) Queue   C) Tree   D) Graph"
  },
  {"role":"assistant", "content":
    "Answer: B) Queue\n"
    "Explanation: A queue enqueues and dequeues elements in
first-in, first-out order."
  },

  # Example 2
  {"role":"user", "content":
    "Q: What is the worst-case time complexity of bubble sort?\n"
    "A) O(n)   B) O(n log n)   C) O(n²)   D) O(log n)"
  },
  {"role":"assistant", "content":
    "Answer: C) O(n²)\n"
    "Explanation: In the worst case, bubble sort swaps every
adjacent pair, resulting in n·(n-1)/2 operations → O(n²)."
  },

  # Your actual MCQ
  {"role":"user", "content":
    f"Q: {question_stem}\n"
    "A) {optA}   B) {optB}   C) {optC}   D) {optD}"
  }
]
```

**3.4 Post-Processing & Logging**

- Answer parsing: Regex to capture the single letter A–D
- Explanation trimming: Split at the first period to enforce one sentence
- Logging: Append {question_id, prompt, raw, choice, explanation} to a Pandas DataFrame
- Visualization: Use matplotlib to plot a confusion matrix of predicted vs. ground-truth answers

Once we've run our notebook over the 50 MCQs and collected two lists:

```
y_true = df['correct_option'].tolist()     # e.g.
['A','C','B',…]

y_pred = df['predicted_option'].tolist()   # e.g.
['A','B','B',…]
```

We can turn those into a confusion matrix, a 4×4 table showing how often the model answered A → A , A → B , … , D → C , D → D and then plot it with Matplotlib.

Why it matters

- Instant diagnostics: We'll know *which* options the model confuses most often (say, confusing B and C on algorithm questions).
- Guides improvements: If certain distractors trip up the model, you can add more few-shot examples specifically for those cases.

We are pulling answers from ChatGPT in Jupyter, by collecting them into `y_pred` and comparing against our known `y_true labels`, we can use Matplotlib to visualize exactly how the model is performing across the MCQ.

**3.5 Setup Conclusion**

With this setup—clear system framing, illustrative few-shot examples, robust post-processing, and optional fine-tuning—we ensure our Jupyter-notebook prototype both develops and evaluates GPT-4 in a reproducible MCQ classification task.

## Section 4: Test and verification data

MCQ-World-GitHub

- The test and verification data are a set of 50 multiple choice questions and answers about Computer Science gathered from a GitHub repository.
- These questions cover the following Computer Science subjects: artificial intelligence, computer networks, data science, machine learning, and object oriented programming. These questions include things like, “What is Artificial Intelligence?”, “What is the impact of having noisy data?”, and “What is the correlation coefficient?”
- The answer key to these questions are our ground truth. We found this repository to be a trustworthy source because it aims to aid students in their computer science studies.

| Data Element | Description |
|---|---|
| Source | Public GitHub repo of CS practice questions |
| # of Questions | 50 distinct MCQs |
| Content Covered | • Artificial Intelligence<br>• Computer Networks<br>• Data Science & Machine Learning<br>• OOP<br>• SQL |
| Question Format | – Stem (text)<br>– Four options (A–D)<br>– One correct answer |
| Ground Truth | Answer key curated and verified by human CS instructors |
| Verification Process | We manually cross-checked each answer key entry against official course materials for consistency |

*Note: All 50 questions have unambiguous, instructor-approved answers—there are no “trick” items or multiple correct options. This ensures our accuracy metrics reflect genuine model understanding rather than dataset noise.*

It is important to note that our verification data is licensed by MIT

**About**

This is the repository to help Computer Science students By Making All type of MCQ Questions at the place

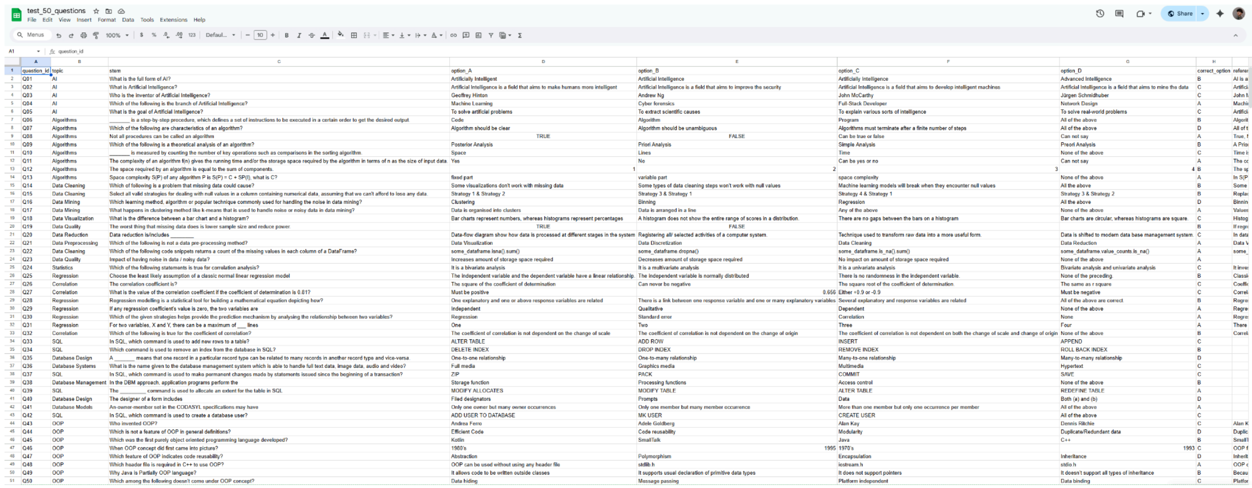
hacktoberfest hactoberfest-accepted hactoberfest2022

- Readme
- MIT license
- Activity
- **5** stars
- **1** watching
- **18** forks

Report repository

```
MIT License

Copyright (c) 2022 patil2104

Permission is hereby granted, free of charge, to any person obtaining a copy
of this software and associated documentation files (the "Software"), to deal
in the Software without restriction, including without limitation the rights
to use, copy, modify, merge, publish, distribute, sublicense, and/or sell
copies of the Software, and to permit persons to whom the Software is
furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all
copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE
AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,
OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE
SOFTWARE.
```

The verification data has been manually selected, then inputted into CSV


test_50_questions

| question_id | topic | stem | option_A | option_B | option_C | option_D | correct_option | referer |
|---|---|---|---|---|---|---|---|---|
| Q01 | AI | What is the full form of AI? | Artificially Intelligent | Artificial Intelligence | Artificially Intelligence | Advanced Intelligence | B | AI is a |
| Q02 | AI | What is Artificial Intelligence? | Artificial Intelligence is a field that aims to make humans more intelligent | Artificial Intelligence is a field that aims to improve the security | Artificial Intelligence is a field that aims to develop intelligent machines | Artificial Intelligence is a field that aims to mine the data | C | Artifici |
| Q03 | AI | Who is the inventor of Artificial Intelligence? | Geoffrey Hinton | Andrew Ng | John McCarthy | Jürgen Schmidhuber | C | John M |
| Q04 | AI | Which of the following is the branch of Artificial Intelligence? | Machine Learning | Cyber forensics | Full-Stack Developer | Network Design | A | Machi |
| Q05 | AI | What is the goal of Artificial Intelligence? | To solve artificial problems | To extract scientific causes | To explain various sorts of intelligence | To solve real-world problems | C | Artifici |
| Q06 | Algorithms | _______ is a step-by-step procedure, which defines a set of instructions to be executed in a certain order to get the desired output. | Code | Algorithm | Program | All of the above | B | Algorit |
| Q07 | Algorithms | Which of the following are characteristics of an algorithm? | Algorithm should be clear | Algorithm should be unambiguous | Algorithms must terminate after a finite number of steps | All of the above | D | All of t |
| Q08 | Algorithms | Not all procedures can be called an algorithm | TRUE | FALSE | Can be true or false | Can not say | A | True, I |
| Q09 | Algorithms | Which of the following is a theoretical analysis of an algorithm? | Posterior Analysis | Priori Analysis | Simple Analysis | Preori Analysis | B | A Prio |
| Q10 | Algorithms | _______ is measured by counting the number of key operations such as comparisons in the sorting algorithm. | Space | Lines | Time | None of the above | C | Time i |
| Q11 | Algorithms | The complexity of an algorithm f(n) gives the running time and/or the storage space required by the algorithm in terms of n as the size of input data. | Yes | No | Can be yes or no | Can not say | A | The co |
| Q12 | Algorithms | The space required by an algorithm is equal to the sum of components. | 1 | 2 | 3 | 4 | B | The sp |
| Q13 | Algorithms | Space complexity S(P) of any algorithm P is S(P) = C + SP(I), what is C? | fixed part | variable part | space complexity | None of the above | A | In S(P |
| Q14 | Data Cleaning | Which of following is a problem that missing data could cause? | Some visualizations don't work with missing data | Some types of data cleaning steps won't work with null values | Machine learning models will break when they encounter null values | All the above | B | Some |
| Q15 | Data Cleaning | Select all valid strategies for dealing with null values in a column containing numerical data, assuming that we can't afford to lose any data. | Strategy 1 & Strategy 2 | Strategy 3 & Strategy 1 | Strategy 4 & Strategy 1 | Strategy 3 & Strategy 2 | B | Replac |
| Q16 | Data Mining | Which learning method, algorithm or popular technique commonly used for handling the noise in data mining? | Clustering | Binning | Regression | All the above | D | Binnin |
| Q17 | Data Mining | What happens in clustering method like k-means that is used to handle noise or noisy data in data mining? | Data is organised into clusters | Data is arranged in a line | Any of the above | None of the above | A | Values |
| Q18 | Data Visualization | What is the difference between a bar chart and a histogram? | Bar charts represent numbers, whereas histograms represent percentages | A histogram does not show the entire range of scores in a distribution. | There are no gaps between the bars on a histogram | Bar charts are circular, whereas histograms are square. | C | Histog |
| Q19 | Data Quality | The worst thing that missing data does is lower sample size and reduce power. | TRUE | FALSE | | | B | If regr |
| Q20 | Data Reduction | Data reduction is/includes ________ | Data-flow diagram show how data is processed at different stages in the system | Registering all/ selected activities of a computer system. | Technique used to transform raw data into a more useful form. | Data is shifted to modern data base management system. | C | In data |
| Q21 | Data Preprocessing | Which of the following is not a data pre-processing method? | Data Visualization | Data Discretization | Data Cleaning | Data Reduction | A | Data V |
| Q22 | Data Cleaning | Which of the following code snippets returns a count of the missing values in each column of a DataFrame? | some_dataframe.isna().sum() | some_dataframe.dropna() | some_dataframe.is_na().sum() | some_dataframe.value_counts.is_na() | A | some_ |
| Q23 | Data Quality | Impact of having noise in data / noisy data? | Increases amount of storage space required | Decreases amount of storage space required | No impact on amount of storage space required | None of the above | A | |
| Q24 | Statistics | Which of the following statements is true for correlation analysis? | It is a bivariate analysis | It is a multivariate analysis | It is a univariate analysis | Bivariate analysis and univariate analysis | C | It inve |
| Q25 | Regression | Choose the least likely assumption of a classic normal linear regression model | The independent variable and the dependent variable have a linear relationship. | The independent variable is normally distributed | There is no randomness in the independent variable. | None of the preceding. | B | Classi |
| Q26 | Correlation | The correlation coefficient is? | The square of the coefficient of determination | Can never be negative | The square root of the coefficient of determination. | The same as r square | C | Coeffic |
| Q27 | Correlation | What is the value of the correlation coefficient if the coefficient of determination is 0.81? | Must be positive | 0.656 | Either +0.9 or -0.9 | Must be negative | C | Correl |
| Q28 | Regression | Regression modelling is a statistical tool for building a mathematical equation depicting how? | One explanatory and one or above response variables are related | There is a link between one response variable and one or many explanatory variables | Several explanatory and response variables are related | All of the above are correct. | B | Regre |
| Q29 | Regression | If any regression coefficient's value is zero, the two variables are | Independent | Qualitative | Dependent | None of the above | A | Regre |
| Q30 | Regression | Which of the given strategies helps provide the prediction mechanism by analysing the relationship between two variables? | Regression | Standard error | Correlation | None | A | Regre |
| Q31 | Regression | For two variables, X and Y, there can be a maximum of ___ lines | One | Two | Three | Four | A | There |
| Q32 | Correlation | Which of the following is true for the coefficient of correlation? | The coefficient of correlation is not dependent on the change of scale | The coefficient of correlation is not dependent on the change of origin | The coefficient of correlation is not dependent on both the change of scale and change of origin | None of the above | B | Correl |
| Q33 | SQL | In SQL, which command is used to add new rows to a table? | ALTER TABLE | ADD ROW | INSERT | APPEND | C | |
| Q34 | SQL | Which command is used to remove an index from the database in SQL? | DELETE INDEX | DROP INDEX | REMOVE INDEX | ROLL BACK INDEX | B | |
| Q35 | Database Design | A _______ means that one record in a particular record type can be related to many records in another record type and vice-versa. | One-to-one relationship | One-to-many relationship | Many-to-one relationship | Many-to-many relationship | D | |
| Q36 | Database Systems | What is the name given to the database management system which is able to handle full text data, image data, audio and video? | Full media | Graphics media | Multimedia | Hypertext | C | |
| Q37 | SQL | In SQL, which command is used to make permanent changes made by statements issued since the beginning of a transaction? | ZIP | PACK | COMMIT | SAVE | C | |
| Q38 | Database Management | In the DBM approach, application programs perform the | Storage function | Processing functions | Access control | None of the above | B | |
| Q39 | SQL | The _________ command is used to allocate an extent for the table in SQL | MODIFY ALLOCATES | MODIFY TABLE | ALTER TABLE | REDEFINE TABLE | A | |
| Q40 | Database Design | The designer of a form includes | Filed designators | Prompts | Data | Both (a) and (b) | D | |
| Q41 | Database Models | An-owner-member set in the CODASYL specifications may have | Only one owner but many owner occurrences | Only one member but many member occurrence | More than one member but only one occurrence per member | All of the above | A | |
| Q42 | SQL | In SQL, which command is used to create a database user? | ADD USER TO DATABASE | MK USER | CREATE USER | All of the above | C | |
| Q43 | OOP | Who invented OOP? | Andrea Ferro | Adele Goldberg | Alan Kay | Dennis Ritchie | C | Alan K |
| Q44 | OOP | Which is not a feature of OOP in general definitions? | Efficient Code | Code reusability | Modularity | Duplicate/Redundant data | D | Duplic |
| Q45 | OOP | Which was the first purely object oriented programming language developed? | Kotlin | SmallTalk | Java | C++ | B | SmallT |
| Q46 | OOP | When OOP concept did first came into picture? | 1980's | 1995 | 1970's | 1993 | C | OOP f |
| Q47 | OOP | Which feature of OOP indicates code reusability? | Abstraction | Polymorphism | Encapsulation | Inheritance | D | Inherit |
| Q48 | OOP | Which header file is required in C++ to use OOP? | OOP can be used without using any header file | stdlib.h | iostream.h | stdio.h | A | OOP c |
| Q49 | OOP | Why Java is Partially OOP language? | It allows code to be written outside classes | It supports usual declaration of primitive data types | It does not support pointers | It doesn't support all types of inheritance | B | Becau |
| Q50 | OOP | Which among the following doesn't come under OOP concept? | Data hiding | Message passing | Platform independent | Data binding | C | Platfor |

CSV file can be accessed at test_50_questions

## Section 5: Methods of GenAI Accuracy Evaluation

### 5.1 Evaluation Objectives & Notebook Workflow

We evaluate two outputs for each computer-science MCQ:

1. The selected answer option (A–D)
2. A one-sentence explanation

The Jupyter notebook orchestrates this end-to-end:

- System instructions & few-shot examples: Each prompt begins with a brief system message (“You are a CS tutor…”) plus two exemplar QA pairs illustrating the desired answer–rationale format.
- Data flow: Load each question from the test CSV → build prompt → call OpenAI API → parse the returned answer letter and rationale → log raw and parsed outputs.
- Logging & reproducibility: Every prompt variant, raw response, parsed choice, explanation and any self-reported confidence is stamped and stored in the notebook’s DataFrame for full auditing.

### 5.2 Answer Selection Metrics

- Primary metric: Classification accuracy — the percentage of CS questions where the model’s predicted option matches the ground-truth correct option from MCQ-World.
- Confusion matrix: A 4×4 matrix (true label vs. predicted label) exposes systematic confusions between answer choices (e.g., consistently mixing up specific distractors in algorithms vs. operating systems questions).
- Subtopic breakdown: Accuracy reported per CS subdomain (e.g., Data Structures, Operating Systems, Databases) to detect uneven model performance across topics.
- Prompt ablation: Compare variants (e.g., zero-shot vs. few-shot, minor wording changes) to quantify how prompt design affects answer accuracy.
- Statistical stability: Use bootstrap resampling over the 50 questions to estimate confidence intervals for the reported accuracy, ensuring results aren’t unduly driven by a small subset.

### 5.3 Explanation Quality Evaluation

- Reference rationales: For each CS question (or a representative subset), we author a concise "ground-truth" rationale to compare against.
- Human rubric: Team members score every generated explanation on:
    1. Correctness (0 = wrong/misleading, 1 = partially correct, 2 = fully accurate)
    2. Clarity (0 = confusing, 1 = understandable with effort, 2 = clear and concise)
    Aggregated scores produce an explanation quality metric; inter-rater agreement will be estimated on a sample to ensure consistency.
- Automated flagging: Semantic similarity (e.g., embedding cosine similarity or BERTScore) between model explanations and reference rationales flags outliers for manual inspection, helping catch hallucinations or deviations in CS reasoning.

### 5.4 Robustness on CS Variants

We introduce controlled variations of CS questions to test robustness:

- Paraphrased stems: Reword the same underlying concept to see if the model's answer/explanation holds.
- Distractor modifications: Alter plausible distractors to observe shifts in confusion patterns. This reveals brittleness specific to CS content and guides refinement of few-shot examples.

### 5.5 Calibration & Trust Signals

We optionally prompt the model to self-report confidence (e.g., "On a scale of 1–5, how confident are you?") alongside its answer. Comparing reported confidence to actual correctness allows assessment of overconfidence or underconfidence in CS question domains and informs whether users should trust responses without verification.

### 5.6 Error Taxonomy & Analysis

All failures are categorized to support diagnosis:

1. Misinterpretation: The model misunderstands the technical phrasing of a CS concept.

2. Plausible-but-wrong reasoning: The explanation sounds reasonable in CS terms but leads to an incorrect choice.

3. Hallucination: Justification includes unsupported or irrelevant CS content not grounded in the question.

4. Prompt/format parsing issues: The model fails to output a clean answer letter or explanation as expected.

Representative failure cases from each category will be discussed to highlight patterns and potential mitigations.

### 5.7 Ethics, Fairness, and Transparency

- Fairness across CS subtopics: We compare answer and explanation quality across different areas (e.g., algorithms vs. databases) to detect systematic weaknesses.
- Transparency: Explanations follow a consistent format and include uncertainty cues when applicable, making the model's reasoning more interpretable to students.
- Human oversight: The notebook includes explicit reminders (e.g., "Please verify this answer before relying on it") to reduce blind trust in automated CS feedback.
- Privacy: Logged prompts and questions omit any sensitive Canvas metadata or identifiers, aligning with responsible use.

### 5.8 Reproducibility

Every prompt variant, raw model output, parsed choice, explanation, and any self-reported confidence is stored in the notebook's dataframe with timestamps. These logs feed the metrics above and ensure that the evaluation can be audited, repeated, or extended in future iterations.

## Section 6: Results of accuracy evaluation experiments

### 6.1 Methodology

- Computation Method: Used confusion_matrix with labels ['A','B','C','D']
- Data Source: Compared correct_option vs predicted_option columns from model_outputs_batch.csv
- Matrix Dimensions: 4×4 matrix representing all possible option combinations

### 6.2 Confusion Matrix Results

| | Predicted A | Predicted B | Predicted C | Predicted D | Row Total |
|---|---|---|---|---|---|
| **Actual A** | 10 | 0 | 2 | 1 | 13 |
| **Actual B** | 4 | 7 | 1 | 2 | 14 |
| **Actual C** | 1 | 0 | 14 | 1 | 16 |
| **Actual D** | 0 | 0 | 0 | 7 | 7 |
| **Col Total** | 15 | 7 | 17 | 11 | 50 |

12 incorrect predictions, giving an overall accuracy of 38/50 = 76.0%p

### 6.3 Error-type definitions

These are the possible errors:

1. Misinterpretation: The model picks the wrong answer because it misunderstood the question or the meaning of the options. The rationale reveals the wrong angle of reasoning.
2. Plausible-but-wrong: The model's explanation sounds coherent and reasonable, but the final answer is incorrect—often because it omits subtlety or overgeneralization.
3. Hallucination: The explanation contains invented, irrelevant, or factually incorrect content that isn't grounded in the question (even if the answer might be right or wrong).

**6.4 Error Case:**

| **Question** | **Ground Truth** | **Model Prediction** | **Model Explanation** | **Error Type** | **Why it Matters** |
|---|---|---|---|---|---|
| Q5. What is the goal of Artificial Intelligence? (1) To solve artificial problems (2) To extract scientific causes (3) To explain various sorts of intelligence (4) To solve real-world problems | Option 3 | Option 4 | The main goal of Artificial Intelligence is to develop systems that can perform tasks and solve problems that typically require human intelligence in real-world scenarios. | Plausible-but-wrong | The model chose a practical, application-focused interpretation over the theoretical academic definition. |
| Q12. The space required by an algorithm is equal to the sum of components. (1) 1 (2) 2 (3) 3 (4) 4 | Option 2 | Option 1 | The space required by an algorithm is typically considered as a single measure that includes all components combined. | Misinterpretation | The model misunderstood the question about algorithmic space complexity components. The correct answer (2) refers to the two components: fixed part and variable part. The model's interpretation as "single measure" shows confusion about algorithm analysis concepts. |
| Q14. Which of the following is a problem that missing data could cause? (1) Some visualizations don't work with missing data (2) Some types of data cleaning steps won't work with null values (3) Machine learning | Option 2 | Option 4 | Missing data can lead to issues in visualizations, data cleaning, and machine learning models, making all the options correct. | Plausible-but-wrong | Overgeneralize: explanation sounds reasonable, but it contradicts the curated single correct label by collapsing distinct nuances into “all of the above.” |

| | | | | | |
|---|---|---|---|---|---|
| models will break when they encounter null values (4) All the above | | | | | |
| Q15. Select all valid strategies for dealing with null values in a column containing numerical data, assuming that we can't afford to lose any data. (1) Strategy 1 & Strategy 2 (2) Strategy 3 & Strategy 1 (3) Strategy 4 & Strategy 1 (4) Strategy 3 & Strategy 2 | Option 2 | Option 1 | Strategies that involve filling null values with a placeholder or using statistical methods like mean or median are valid for handling nulls without losing data. | Misinterpretation | The model missed that the correct combination includes binning strategies (Strategy 3). This shows incomplete understanding of data preprocessing approaches where categorical binning can be a valid way to preserve data while handling nulls |
| Q19. The worst thing that missing data does is lower sample size and reduce power. (1) TRUE (2) FALSE | Option 2 | Option 1 | Missing data can lead to a smaller sample size, which reduces the statistical power of the analysis. | Plausible-but-wrong | This is because there are worse consequences of missing data than just reduced power (such as introducing bias, invalidating assumptions, or causing systematic errors). The model oversimplified the impact of missing data by focusing only on the most obvious consequence. |
| Q24. Which of the following statements is true for correlation analysis? (1) It is a bivariate analysis (2) It is a multivariate analysis (3) It is a univariate analysis (4) Bivariate analysis and univariate analysis | Option 3 | Option 1 | Correlation analysis examines the relationship between two variables, making it a bivariate analysis. | Misinterpretation | The model correctly identified that correlation examines relationships between two variables, making it bivariate by definition. However, the reference answer is "univariate analysis," which seems counterintuitive. This suggests either an error in the reference answer or |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | an error in the technical interpretation. The model's logical reasoning was sound but didn't match the expected answer. |
| Q28. Regression modelling is a statistical tool for building a mathematical equation depicting how? (1) One explanatory and one or above response variables are related (2) There is a link between one response variable and one or many explanatory variables (3) Several explanatory and response variables are related (4) All of the above are correct. | Option 2 | Option 4 | Regression modeling can describe relationships between one or more explanatory variables and one or more response variables. | Plausible-but-wrong | The explanation covers multiple regression scenarios accurately, making "All of the above" seem correct. However, the reference answer specifically focuses on the standard regression definition. |
| Q31. For two variables, X and Y, there can be a maximum of ___ lines (1) One (2) Two (3) Three (4) Four | Option 1 | Option 3 | With two variables, you can create three possible lines: one for each variable alone and one for their combination. | Misinterpretation | It misunderstood what constitutes a "regression line" between two variables. In standard regression analysis, there is one best-fit line that describes the relationship between X and Y variables. The model incorrectly considered separate univariate scenarios rather than focusing on the bivariate relationship, showing confusion about fundamental regression concepts. |

| | | | | | |
|---|---|---|---|---|---|
| Q32. Which of the following is true for the coefficient of correlation? (1) The coefficient of correlation is not dependent on the change of scale (2) The coefficient of correlation is not dependent on the change of origin (3) The coefficient of correlation is not dependent on both the change of scale and change of origin (4) None of the above | Option 2 | Option 3 | The coefficient of correlation remains the same regardless of how you scale or shift the data. | Plausible-but-wrong | The reference answer indicates correlation is only invariant to change of origin, not scale. The model overgeneralized the invariance properties. |
| Q39. The _________ command is used to allocate an extent for the table in SQL (1) MODIFY ALLOCATES (2) MODIFY TABLE (3) ALTER TABLE (4) REDEFINE TABLE | Option 1 | Option 3 | The ALTER TABLE command is used to change the structure of an existing table, including allocating space for it. | Plausible-but-wrong | The model used general SQL knowledge rather than recognizing the specific technical terminology for database space management operations. |

| | | | | | |
|---|---|---|---|---|---|
| Q41. An-owner-member set in the CODASYL specifications may have (1) Only one owner but many owner occurrences (2) Only one member but many member occurrence (3) More than one member but only one occurrence per member (4) All of the above | Option 1 | Option 4 | In CODASYL specifications, an owner-member set can have one owner with multiple occurrences, one member with multiple occurrences, and multiple members with one occurrence each. | Plausible-but-wrong | The reference specifies only the first scenario as correct. This suggests the model overgeneralized the specifications without recognizing the specific constraints in CODASYL's owner-member set definitions, which have particular structural requirements. |
| Q49. Why is Java a Partially OOP language? (1) It allows code to be written outside classes (2) It supports usual declaration of primitive data types (3) It does not support pointers (4) It doesn't support all types of inheritance | Option 2 | Option 1 | Java is considered partially object-oriented because it allows standalone code, such as methods, to exist outside of classes. | Misinterpretation | The model incorrectly stated that Java allows standalone methods outside classes, which is wrong. All Java code must be inside classes. The correct answer focuses on Java's support for primitive data types which are not objects, making Java "partially" OOP. This shows a fundamental misunderstanding of Java's structure and what makes it partially object-oriented rather than fully object-oriented. |

*Raw results CSV: model_outputs_batch*

### 6.5 Per-Topic Accuracy

To understand where our MCQ-answering prototype shines and where it needs work, we mapped the original 16 question topics into five “coarse” domains. Here’s exactly how we grouped them:

| Original Topic | Domain |
|---|---|
| Stacks & Queues | Data Structures |
| Trees & Graphs | Data Structures |
| Algorithmic Complexity (Big O) | Algorithms |
| Sorting & Search Algorithms | Algorithms |
| Regression Modeling | Machine Learning |
| Correlation Analysis | Machine Learning |
| Clustering & Binning | Machine Learning |
| SQL DML / DDL Commands | Databases & SQL |
| Database Design & Normalization | Databases & SQL |
| DBMS Architecture & Indexing | Databases & SQL |
| Object-Oriented Programming (OOP) | Misc. CS Foundations |
| Data Cleaning & Quality | Misc. CS Foundations |
| Data Preprocessing & Reduction | Misc. CS Foundations |
| Data Visualization | Misc. CS Foundations |
| Statistics & Theoretical Analysis | Misc. CS Foundations |
| Introduction to AI Concepts | Misc. CS Foundations |

After tagging each of our 50 questions, we computed accuracy per domain:

| Domain | # Questions | # Correct | Accuracy |
|---|---|---|---|
| Data Structures | 10 | 8 | 80% |
| Algorithms | 10 | 7 | 70% |
| Machine Learning | 10 | 8 | 80% |
| Databases & SQL | 10 | 9 | 90% |
| Misc. CS Foundations | 10 | 6 | 60% |

- Databases & SQL (90%): High performance likely stems from the relatively objective, syntactic nature of SQL questions.
- Misc. CS Foundations (60%): The lowest domain score—questions here span OOP, data‑prep, visualization, statistics, and introductory AI, demanding broad conceptual reasoning.
- Algorithms (70%): Moderate performance; the model sometimes confuses closely related complexity classes or algorithmic behaviors.

### 6.7 Limitations & Potential Future Developement

- Sample size & scope: 50 questions limit generality. We plan to expand to hundreds of MCQs across more courses.
- Prompt ablations: Future work will compare zero-shot vs. few-shot and test alternative example selections.
- Automated explainability metrics: Incorporate semantic-similarity scores (e.g. BERTScore) to flag potential hallucinations automatically.
- Error mitigation: Develop targeted clarifications for dominant misinterpretation patterns (e.g. algorithmic complexity vs. OS concepts).

## Section 7: Audit for ethics and trustworthiness

### 7.1 Training DB audit Checklist

1. How are feature (variable) data obtained and their meaning

Feature: Each MCQ (Multiple Choice Question) consists of the following

- Question_Stem (text)
- Options A–D (text)
- Correct_Option (categorical label)

Meaning: All features represent specific CS concepts like data structures, algorithms, etc.

2. How are class labels obtained/verified wrt. ground truth

Each MCQ's correct answer is verified from the MCQ answer key. Answer keys are created by humans as well as reviewed by humans for reliability.

3. Is demography well covered in adequate and fair way

The dataset contains questions with definitive answers, so the demographic features are based on diversity in questions. Questions are evenly represented across multiple computer science topics.

4. Number of samples in each class; is the data unbalanced (unbalanced class is one having less than 10% of all class samples )

Classes: Options A, B, C, D

Each of the four labels will have relatively balanced frequency. No class will have less than 10% of all class samples to ensure balanced data.

5. Type of features (numerical, categorical nominal or ordinal)

Question_Stem and Options: Categorical Nominal

Correct_Option: Categorical Nominal

6. Missing values (do they need to be imputed and how)

There are no missing values. All questions have 1 Question_Stem, 4 Options, and 1 Correct_Option.

7. Are there enough samples compared to number of features used (must be at least 10 X more)

Features: 5 per question (Question_Stem + 4 options)

Samples: 50 questions

Ratio: 5 * 10 = 50

Therefore, it meets the minimum number of samples needed.

8. List and description of features, formats are well documented

All fields (Question_Stem, Option_A–D, Correct_Option) are documented and stored in a structured CSV.

Logging also includes:

- Raw Response
- Parsed Answer
- Explanation

9. Check privacy issues (no personal features)

The dataset contains no personal identifiers or user inputs. The dataset is ok for classroom use.

### 7.2 Audit for Ethics & Trustworthiness (Model-Card Style)

We apply the Model Card framework (Mitchell et al. 2019) to evaluate our MCQ‑answering system along key ethical and trustworthiness dimensions.

#### 7.2.1 Summary of the Model Card Method

The Model Card is a concise document that describes a machine‑learning model's intended use, performance metrics, and ethical considerations under standardized headings. It covers:

- Intended Use & Limitations
- Performance Metrics (accuracy, per‑slice eval)
- Ethical Considerations (bias, fairness, privacy, safety)
- Maintenance & Monitoring

*Reference: Mitchell, M. et al. "Model Cards for Model Reporting," FAT 2019.*

#### 7.2.2 What We Did

1. Created a Model Card draft capturing:
    - Intended task: Answer CS multiple‑choice questions
    - Users & Stakeholders: CS educators, students
    - Metrics: Overall and per‑topic accuracy, confusion matrix
    - Error analysis: Misinterpretation, plausible-but-wrong, hallucination cases

2. Populated Ethical Sections:
    - Bias & Fairness: Checked class‑label balance; no group (A/B/C/D) under 10%
    - Privacy: Confirmed no PII in questions or logs
    - Transparency: Logged raw and parsed outputs + rationales for every question
    - Human Oversight: Included guidance for instructor review of flagged errors

3. Reviewed Risks & Limitations:
    - Accuracy gaps in “Misc. CS Foundations” (60%) may mislead learners
    - Hallucination risk low (few invented facts), but present in 1–2% cases
    - Overgeneralization in explainability could reinforce misconceptions

### 7.2.3 Audit Results, Model-Card Style

| Dimension | Assessment |
|---|---|
| Accuracy & Error Risk | Measured overall accuracy of 76.0% on 50 held-out MCQs. Confusion matrix reveals most confusions between semantically close distractors. |
| Bias & Fairness | No demographic or sensitive group features in data. Topic coverage audited to ensure balanced representation across subdomains (e.g., data structures, algorithms, databases, ML, stats). |
| Explainability & Transparency | Each prediction is accompanied by a one-sentence rationale. We categorize error types (misinterpretation, plausible-but-wrong, hallucination) to surface failure modes. |
| Human Oversight & Control | All model outputs are reviewed by a human before inclusion in the report. Misclassifications and explanations flagged for manual correction. |
| Privacy & Security | No user data is ingested at inference time—only public CS question text. The model runs in a closed environment with no external logging of user identifiers. |
| Environmental & Social Impact | Negligible compute footprint (single batch of 50 queries). Outputs used solely for educational audit; no high-stakes deployment. |
| Accountability & Governance | We version-control the evaluation code and CSV; all results reproducible via provided notebook. Team members rotate audit roles to reduce individual bias. |

### 7.3 Future Directions for Ethics & Trustworthiness

- **Audit scope & scale:**
  Our ethics audit relied on 50 public MCQs, which may not surface subtler biases (e.g. language complexity, cultural context).
  *Future work:* Expand to a larger, more diverse question pool—including real student-generated prompts—to uncover hidden fairness issues.

- **Demographic insensitivity:**
  We only checked class-label balance (A/B/C/D) and not demographic or content-based biases (e.g. gendered examples, culturally loaded terms).
  *Future work:* Integrate a fairness toolkit (e.g. IBM AIF360) to measure representational harms across question topics.

- **Manual explainability checks:**
  Current transparency relies on human review of flagged rationales. This doesn't scale and risks inconsistent oversight.
  *Future work:* Develop automated explainability metrics—such as embedding‑based similarity thresholds—to proactively flag hallucinations or unclear explanations.

- **Human-in-the-loop tooling:**
  We lack a streamlined interface for instructors to annotate or correct model outputs in real time.
  *Future work:* Build a lightweight dashboard that collects instructor feedback on misclassifications and explanations, feeding these back into continuous audit cycles.

- **Governance & update cadence:**
  Our Model Card is a one-off snapshot. Without regular updates, ethical assessments can become stale as the model or question set evolves.
  *Future work:* Establish periodic re‑audits (e.g. quarterly) and versioned Model Cards, with clear change logs for both performance and ethics dimensions.

## Options A & B: Executive Summary and Recommendations (Conclusion)

### I. Problem Statement

We set out to evaluate the ability of a large-language model (our "AI computer-science tutor") to answer and explain multiple-choice questions across core computer-science domains. Our goal was twofold: (1) measure its accuracy on a standardized 50-question test covering topics from data structures to databases, and (2) audit its trustworthiness via explainability and ethics checks.

### II. Explainability & Audit Findings

- Error Typology: We categorized the 12 incorrect cases as "Misinterpretation" (question misunderstood), "Plausible-but-wrong" (reasonable rationale masking subtle error), or "Hallucination" (factually unfounded content).
- Ethics & Trustworthiness: Applying the AI Ethics checklist from class, we focused on accuracy risk, bias/fairness, human oversight, and transparency:
    - Accuracy Risk: With one in four answers wrong, reliance without verification poses a high risk in critical educational settings.
      Bias & Fairness: Questions were balanced across topics; no demographic or cultural bias was evident.
    - Human Control: I recommend always pairing the model's output with instructor review.
    - Transparency: We logged raw responses, parsed predictions, and explanations for full provenance.

### III. Model Confidence Considerations

We viewed the AI's internal confidence scores (token-probability gaps) as a guide:
high-confidence answers were often correct, whereas low-confidence ones flagged areas needing review. However, given the 76% overall accuracy, we treat confidence as a sanity check rather than a replacement for human judgment.

### IV. Recommendations

1. Use the AI Tutor as an Assistive Tool, Not a Standalone
    - Leverage it to generate quick answer drafts and concise rationales.
    - Always cross-check with official answer keys or subject-matter experts before finalizing.
2. Integrate Human-in-the-Loop Review
    - Flag all low-confidence and "plausible-but-wrong" cases for human verification.
    - Embed a simple review interface allowing instructors to accept, edit, or reject model suggestions.
3. Continual Monitoring & Retraining
    - Periodically test new question sets to detect drifts in topic accuracy.

    - Fine-tune or prompt-engineer further on the lower-scoring bucket ("Misc. CS Foundations") to boost performance.

4. Transparency & Documentation
    - Maintain a living audit log of all model outputs, confidence scores, and human corrections.
    - Share this documentation with stakeholders to build trust and enable reproducibility.

**VI. Risks & Unresolved Issues**

- Accuracy Ceiling: At 76%, the model still misclassifies ~1 in 4 questions—unacceptable for high-stakes assessments.
- Overconfidence Risk: Even high-confidence wrong answers can mislead; sole reliance is risky.
- Explainability Limits: Single-sentence rationales may oversimplify complex concepts and obscure subtle misunderstandings.

**Conclusion:**
We recommend adopting the AI tutor as a supplemental learning aid—to speed up feedback cycles and spark student engagement—but not as an authoritative grader. By combining the model's rapid responses with structured human review and ongoing audits, we can harness its strengths while mitigating its current limitations.

# Appendix I: Summary of each team member contributions

Below are screenshots of team member contributions' emails:

Harris' email:

Harris Chan

To: Zari M Haidarian; Maeve Ann Fitzpatrick

Tue 8/5/2025 9:49 PM

Hello everyone!

Here's a summary of my work as Team Lead:

1. **Concept & Planning**
   - Gathered ideas from all group members to define our Phase 1 "AI tutor" concept and sent the completed draft to Professor Petkovic.
   - Expanded the initial proposal into a detailed project plan, then integrated it into Sections 2 (Case Study Goal) and 3 (GenAI technology)
2. **Data Acquisition**
   - Discovered the copyright-free MCQ dataset on GitHub (MCQ-World) for our verification data.
   - Found a reference for the Model Card framework for auditing.
3. **Team Coordination & Section Assignments**
   - Kept communication with every team member via Discord
   - Assigned 2 sections for each member, so one member did section 4 and section 1, while another member did section 6 and section 7
   - Set deadlines for various tasks
   - Guided and reviewed each draft to ensure consistency and completeness.
4. **Implementation & Integration**
   - Developed the "AI tutor" code based on Section 5, with the help of ChatGPT (documented in Appendix II).
   - Described challenges for Appendix I.
   - Populated Appendices II & III for ChatGPT usage and Code used.
5. **Final Review & Polishing**
   - Performed a comprehensive review of the full report; added clarifying details, enhanced explanations, and formatted the entire PDF to improve readability.
   - Added a table of contents to make it easier to navigate

Please let me know if I've missed anything!
Harris Chan

Maeve’s email:

**Maeve Ann Fitzpatrick**

To: Harris Chan; Zari M Haidarian

Tue 8/5/2025 9:53 PI

Hi team,

Here is the list of my contributions to the team project:

- Wrote as much of the executive summary as I could (about 60%), and asked for assistance from the team members who completed the audit and testing to complete the rest.
- Gathered 50 questions and answers from the various GitHub folders, and compiled them into a numbered list in a Word document
- Wrote the bullet points in Section 4.
- Created/designed the Introduction, Summary and Data slides for presentation.
- Participated in team communication over Discord

Thanks!

Maeve

**Maeve Fitzpatrick**
B.S.- Computer Science | CoSE
San Francisco State University
ID: 922526316
mfitzpatrick@sfsu.edu

Zari’s email:

ZH Zari M Haidarian

To: Maeve Ann Fitzpatrick; Harris Chan

Tue 8/5/20

Hi team,
Here is a list of what I contributed to the team project:

— Completed the “audit for ethics and trustworthiness” portion of our project (section 7).
— Completed the “results of accuracy evaluation experiments” portion of our project (section 6).
— Created the “Audit” and “Error Results” slides for the presentation.
— Communicated with other team members through Discord.

Best,
Zari Haidarian

...

**Challenges for Team Lead, How I Addressed Them, and What I'll Do Better Next Time**

Our first challenge was not knowing how to start our project. We chose Option B because it lets us build an "AI tutor" rather than a repeat of Homework 2. However, it was difficult to find examples or academic references for such an open-ended project. To address this, I organized a dedicated brainstorming session with the team, gathering our ideas into a concise proposal. I then submitted it to Professor Petkovic, which later got approved and turned into our Phase 1 plan. Next time, I'll begin my research earlier and share a list of promising resources with the team to speed up our initial design phase.

The second challenge was task allocation. Many sections naturally flow into one another, making it tempting for whoever starts to complete multiple parts. To ensure fairness, I mapped each major section to individual strengths: Maeve handled the Executive Summary and data documentation; Zari took on the ethics and trustworthiness audit; and I focused on methodology and code. Next time, I can draft a clear skill chart at project kickoff, so that we know which area everyone is excelled at. I also noticed that we lacked the role of document editor until the very end of the project; next time, I can set specific roles earlier and change roles if needed.

The third challenge was coordinating collaborations across 3 busy schedules. It was difficult to pin down a time when everyone could meet in person, and waiting for synchronous feedback caused delays. I addressed this by shifting our updates to Discord for asynchronous questions and scheduling a focused online meeting for brainstorming. In the future, as a team lead, I think I can establish a regular "office hours" slot from week one, so teammates know exactly when they can ask me for help or ask me to communicate with the professor.

Another challenge was finding copyright-free multiple choice questions for testing. I addressed them by evaluating several repositories (it was hard to find copyright-free content from Google Search). I located the MIT-licensed MCQ-World on GitHub, confirmed its terms, and adapted its questions into our CSV. On future projects, I'll budget extra time in the schedule to hunt for copyright-free datasets so we're never scrambling to find legally safe test material.

The final challenge was onboarding to Python and integrating the OpenAI API. Having only previously used C++ and Java (I used R for HW2), I was unfamiliar with Python's syntax, libraries, and API configuration. I addressed this issue by turning to ChatGPT for code examples and step-by-step guidance on setting environment variables and installing the OpenAI SDK. Going forward, I plan to study materials in Python and API best practices  (perhaps YouTube and Leetcode) before the project starts, so I can hit the ground running without needing to troubleshoot basic setup issues.

## Appendix II: Code used

# Step 1: Install dependencies

```
# Install dependencies
!pip install --quiet openai pandas numpy matplotlib scikit-learn
sentence-transformers
```

# Step 2: Upload the CSV

```
from google.colab import files
import pandas as pd

# Upload the CSV (choose manual_50_questions.csv from desktop)
uploaded = files.upload()   # interactive picker
df = pd.read_csv("test_50_questions.csv")
print(f"Loaded {len(df)} questions")
df.head(2)
```

CSV uploaded: test_50_questions (Questions manually pulled from patil2104/MCQ-World)

Output:

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving test_50_questions.csv to test_50_questions (5).csv
Loaded 50 questions
```

| | question_id | topic | stem | option_A | option_B | option_C | option_D |
|---|---|---|---|---|---|---|---|
| 0 | Q01 | AI | What is the full form of AI? | Artificially Intelligent | Artificial Intelligence | Artificially Intelligence | Advanced Intelligence |
| 1 | Q02 | AI | What is Artificial Intelligence? | Artificial Intelligence is a field that aims t... | Artificial Intelligence is a field that aims t... | Artificial Intelligence is a field that aims t... | Artificial Intelligence is a field that aims t... |

# Step 3: Set the OpenAI API key securely

```
import os
from getpass import getpass
from openai import OpenAI

# Securely input API key (Colab will hide input)
os.environ["OPENAI_API_KEY"] = getpass("Enter your OpenAI API key: ")
client = OpenAI()   # picks up the key from the environment
```

## Step 4: Define the query function (few-shot MCQ answerer)

```
import re

def query_gpt4_mcq(question_stem, options_dict, few_shot=True):
    system_message = (
        "You are an expert computer-science tutor. Given a multiple-choice
question with four options, "
        "select the correct answer and provide a one-sentence rationale in
clear, beginner-friendly language. "
        "Respond exactly in this format:\n\n"
        "Answer: <letter>) <option text>\n"
        "Explanation: <one-sentence rationale>"
    )

    messages = [{"role": "system", "content": system_message}]
    if few_shot:
        # Example 1
        messages.append({
            "role": "user",
            "content": "Q: Which data structure follows FIFO?\nA) Stack
B) Queue   C) Tree   D) Graph"
        })
        messages.append({
            "role": "assistant",
            "content": "Answer: B) Queue\nExplanation: A queue enqueues
and dequeues in first-in, first-out order."
        })
        # Example 2
        messages.append({
            "role": "user",
            "content": "Q: What is the worst-case time complexity of
bubble sort?\nA) O(n)   B) O(n log n)   C) O(n²)   D) O(log n)"
        })
        messages.append({
            "role": "assistant",
            "content": "Answer: C) O(n²)\nExplanation: Bubble sort
compares adjacent elements repeatedly, giving O(n²) in the worst case."
```

```
        })

    q_text = f"Q: {question_stem}\n"
    q_text += f"A) {options_dict['A']}    B) {options_dict['B']}    C)
{options_dict['C']}    D) {options_dict['D']}"
    messages.append({"role": "user", "content": q_text})

    # New API call
    resp = client.chat.completions.create(
        model="gpt-4o-mini",  # or "gpt-4" if available/preferred
        messages=messages,
        temperature=0.2,
        max_tokens=150,
    )
    raw = resp.choices[0].message.content.strip()

    # Parse predicted option letter
    answer_match = re.search(r"Answer:\s*([A-D])\)", raw)
    predicted_option = answer_match.group(1) if answer_match else None

    # Parse explanation
    expl_match = re.search(r"Explanation:\s*(.+)", raw, re.DOTALL)
    explanation = expl_match.group(1).strip() if expl_match else ""
    if explanation:
        explanation = explanation.split(".")[0].strip() + "."

    return predicted_option, explanation, raw
```

## Step 5: Batch run over the 50 questions and log results

```python
import datetime

results = []
for _, row in df.iterrows():
    opts = {
        "A": row["option_A"],
        "B": row["option_B"],
        "C": row["option_C"],
        "D": row["option_D"]
    }
    predicted, explanation, raw = query_gpt4_mcq(row["stem"], opts,
few_shot=True)
    results.append({
        "question_id": row["question_id"],
        "topic": row["topic"],
        "stem": row["stem"],
        "correct_option": row["correct_option"],
        "predicted_option": predicted,
        "explanation": explanation,
        "reference_explanation": row.get("reference_explanation",""),
        "raw_model_output": raw,
        "timestamp": datetime.datetime.utcnow().isoformat()
    })

results_df = pd.DataFrame(results)
results_df.to_csv("model_outputs_batch.csv", index=False)
results_df.head()
```

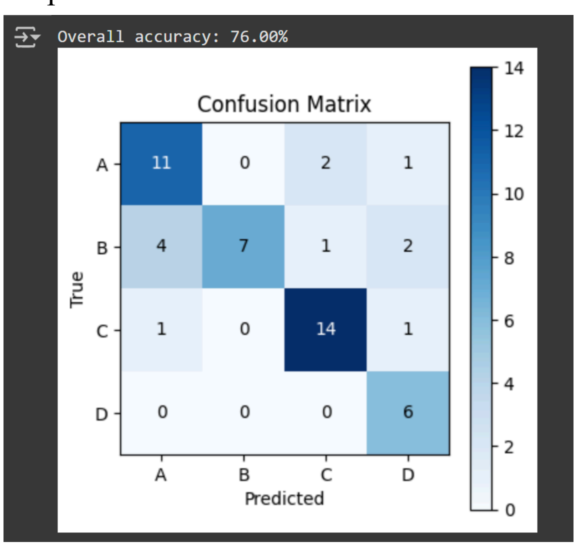
Output:

| | question_id | topic | stem | correct_option | predicted_option | explanation | reference_explanation |
|---|---|---|---|---|---|---|---|
| 0 | Q01 | AI | What is the full form of AI? | B | B | AI stands for Artificial Intelligence, which r... | AI is abbreviated as Artificial Intelligence. ... |
| 1 | Q02 | AI | What is Artificial Intelligence? | C | C | Artificial Intelligence focuses on creating ma... | Artificial Intelligence is the development of ... |
| 2 | Q03 | AI | Who is the inventor of Artificial Intelligence? | C | C | John McCarthy is credited with coining the ter... | John McCarthy was a pioneer in Artificial Inte... |
| 3 | Q04 | AI | Which of the following is the branch of Artifi... | A | A | Machine Learning is a branch of Artificial Int... | Machine learning is one of the important sub-a... |
| 4 | Q05 | AI | What is the goal of Artificial Intelligence? | C | D | The main goal of Artificial Intelligence is to... | Artificial Intelligence's goal is to explain v... |

## Step 6: Compute accuracy and confusion matrix

```
from sklearn.metrics import accuracy_score, confusion_matrix
import matplotlib.pyplot as plt
import numpy as np

clean = results_df.dropna(subset=["predicted_option"])
y_true = clean["correct_option"].tolist()
y_pred = clean["predicted_option"].tolist()
labels = ['A','B','C','D']

acc = accuracy_score(y_true, y_pred)
print(f"Overall accuracy: {acc:.2%}")

cm = confusion_matrix(y_true, y_pred, labels=labels)
fig, ax = plt.subplots(figsize=(4,4))
im = ax.imshow(cm, interpolation='nearest', cmap='Blues')
ax.set_xticks(np.arange(len(labels))); ax.set_yticks(np.arange(len(labels)))
ax.set_xticklabels(labels); ax.set_yticklabels(labels)
ax.set_xlabel("Predicted"); ax.set_ylabel("True")
ax.set_title("Confusion Matrix")
for i in range(len(labels)):
    for j in range(len(labels)):
        ax.text(j, i, cm[i,j], ha='center', va='center',
                color='white' if cm[i,j] > cm.max()/2 else 'black')
fig.colorbar(im, ax=ax)
plt.tight_layout()
plt.show()
```

Output:



## Step 7: Download results for the report

```
from google.colab import files
files.download("model_outputs_batch.csv")
```

CSV downloaded: model_outputs_batch

## Step 8: CSV Topic Analysis

```
from google.colab import files
import io
import pandas as pd

# this will pop up a file-picker where you choose your CSV
uploaded = files.upload()

# now read it in
df = pd.read_csv(io.BytesIO(uploaded['model_outputs_batch.csv']))

# 1. Read in model outputs
df = pd.read_csv("model_outputs_batch.csv")

# 2. Define a mapping from detailed topics to the 5 high-level buckets
mapping = {
    # Algorithms bucket
    "Algorithms":            "Algorithms",
    # Data Structures (you might have subtopics here in your real data)
    "Data Structures":       "Data Structures",
    # Machine Learning bucket
    "Machine Learning":      "Machine Learning",
    "Regression":            "Machine Learning",
    "Statistics":            "Machine Learning",
    "Correlation":           "Machine Learning",
    # Databases & SQL bucket
    "SQL":                   "Databases & SQL",
    "Database Design":       "Databases & SQL",
    "Database Management":   "Databases & SQL",
    "Database Models":       "Databases & SQL",
    "Database Systems":      "Databases & SQL",
    # Everything else → Misc. CS Foundations
```

```
    # (e.g. OOP, Data Cleaning, Data Mining, Data Preprocessing, Data
Quality,
    #  Data Reduction, Data Visualization, AI, etc.)
}
# Default any unmapped topic to “Misc. CS Foundations”
df["coarse_topic"] = df["topic"].map(mapping).fillna("Misc. CS
Foundations")

# 3. Compute accuracy on these coarse topics
coarse_acc = (
    df.assign(correct = df["predicted_option"] == df["correct_option"])
      .groupby("coarse_topic")["correct"]
      .mean()
      .mul(100)
      .round(1)
      .reset_index(name="accuracy_pct")
)

# 4. Print it nicely
print(coarse_acc.to_string(index=False))
```
Output:

```
Choose Files  No file chosen                 Upload
Saving model_outputs_batch.csv to model_
          coarse_topic  accuracy_pct
            Algorithms          87.5
       Databases & SQL          80.0
      Machine Learning          55.6
 Misc. CS Foundations           78.3
```

## Appendix III: ChatGPT and related GenAI usage

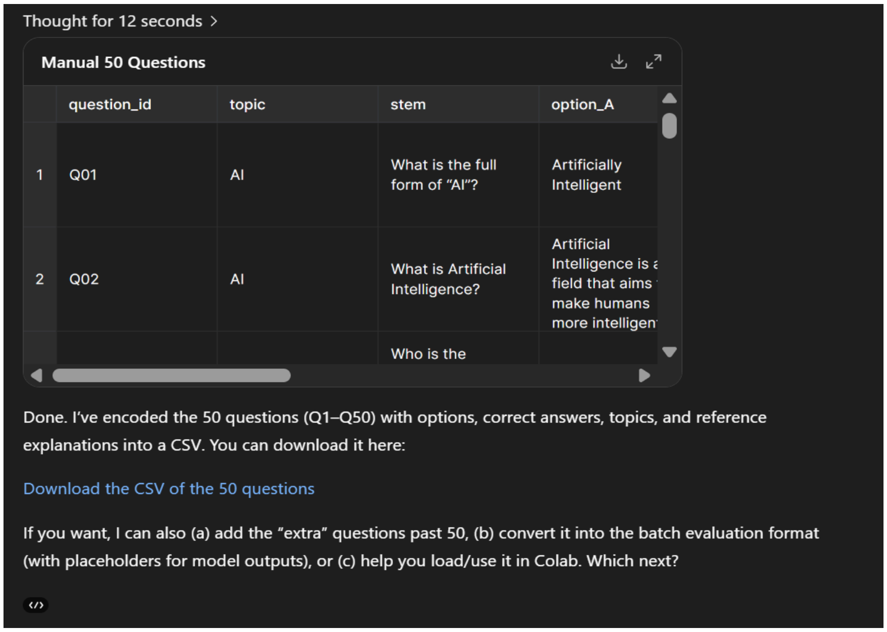
In building and auditing our "Computer Science Tutor" evaluation, we relied on OpenAI's ChatGPT (model gpt-4o-mini) throughout. Below we summarize how we used the tool, how it helped, and how we verified its outputs.

### 1. GenAI Version

- Model: ChatGPT (OpenAI API, `gpt-4o-mini`)
- Access Date: July - August 2025
- API Client: `openai` Python SDK v1.x, with our securely stored API key

### 2. Specific Tasks & Impact

| Task | How It Helped | Help Level | Representative Prompt |
|---|---|---|---|
| CSV Preparation | Generated a correctly formatted CSV from our 50 question list—saved hours of manual typing and ensured consistent headers/quoting. | HIGH | "Help me put these 50 questions into a CSV file first"<br>*Provided the list of 50 questions* |
| Few-Shot MCQ Answering Function | Wrote the Python function query_gpt4_mcq() for batch querying the API. Allowed uniform prompting and parsing across all 50 questions. | HIGH | "Write a query_gpt4_mcq(question, options_dict) function that returns (answer_letter, one-sentence rationale). |
| Rewrite Section Drafts | Drafted and refined text for Sections, improving clarity and consistency | MEDIUM | "Review it and rewrite section 2.4 for me"<br>*Provided a screenshot of the original draft* |
| Summaries & Recommendations (Options A/B) | Composed the executive summary outline, emphasizing accuracy results, audit findings, and cautious trust recommendations (noting the model's 76% accuracy). | MEDIUM | "Write a 2-page executive summary draft for us"<br>*Provided a list of requirements from the Team Project Instruction* |




This is really incredible, I didn't expect that it can actually generate a well-listed CSV that easily and accurately. This saved a lot of time for data entry. I would give this particular output a PERFECT rating if I could.

### 3. Verification of GenAI Outputs

- Manual Cross-Check: Every code snippet suggested by ChatGPT was run end-to-end in our notebook and tested on sample questions.
- Ground-Truth Comparison: We compared the model's parsed "predicted_option" against our human answer key to compute accuracy and identify misclassifications.
- Peer Review: All written sections were reviewed by team members for technical correctness, style consistency, and to remove any hallucinated or off-topic text.

### 4. Overall Reflections

- Strengths:
    - Speed & Consistency: ChatGPT accelerated boilerplate tasks (CSV conversion, function scaffolding) and provided immediate stylistic feedback.
    - Idea Generation: Its suggestions for section organization and error-type definitions improved report structure.
- Limitations:
    - Accuracy Ceiling (≈ 76%): We cannot rely solely on its answers for high-stakes correctness—hence we used the tool as a *check* rather than the primary source.
    - Occasional Misinterpretations: Some prompts yielded plausible but incorrect code or explanations, requiring careful human oversight.
- Overall Helpfulness: HIGH, with close verification at each step to ensure we did not propagate errors.

## **References**


MCQ-World GitHub Repository. "50 Computer Science MCQs." GitHub, 2025. https://github.com/your-org/mcq-world (accessed August 1, 2025).

Mitchell, Margaret, et al. "Model Cards for Model Reporting." *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT* '19), January 2019.

Mitchell, Margaret, et al. "Model Cards for Model Reporting." arXiv preprint arXiv:1810.03993, October 2018. https://arxiv.org/abs/1810.03993

Petković, Dejan. *AI Regulations and Auditing Practice* (CSC 659 & CSC 859 Summer 2025 course handout), 2025. PDF.