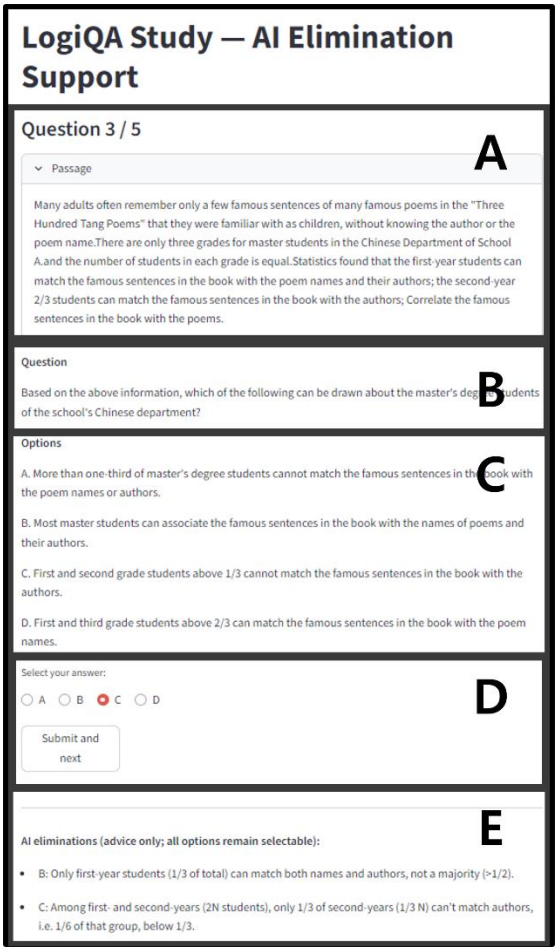
# CSE 594
# Assignment #2

Minseo Park

2025.10.12

University of Michigan

In this task, each performer will be presented with a context, a question and four options. The performer's goal is to select the correct answer with AI's assistance. Each performer will solve 5 questions per round, attempting to complete them as quickly and accurately as possible. The experiment will consist of 4 rounds, so each performer will answer a total of 20 questions.




<Fig 1. Example Interface>

Fig 1. demonstrates the interface that performers will see consisting of five sections. (A) displays the reading passage providing the context. (B) presents the question that performers must solve. (C) lists the multiple-choice options. (D) is the area where performers select their answer. (E) contains the two options eliminated by the AI, along with the corresponding reasons.

Several datasets were considered as potential candidates for this assignment, including LogiQA, ReClor, CommonsenseQA, RACE, and OpenBookQA.

- ReClor was excluded due to its source, GMAT and LSAT, which include dense logical reasoning passages that go far beyond the scope of this assignment
- RACE was excluded because of the long reading passages, making it impractical for the test.

- OpenBookQA was excluded because many questions rely on specialized science knowledge, which may disadvantage performers from diverse backgrounds.

After this screening, the two strongest candidates were CommonsenseQA and LogiQA. CommonsenseQA primarily depends on general common sense which humans usually handle easily, while AI often struggles. In contrast, LogiQA requires systematic logical reasoning which humans can also perform but where AI's assistance can be helpful. AI can eliminate implausible options quickly but may fail to reach the correct final conclusion.

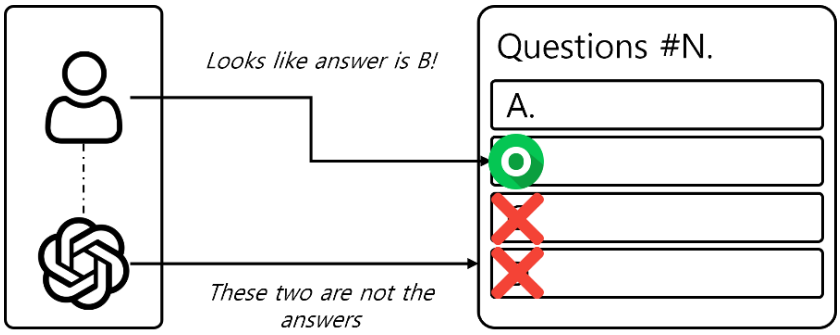
Human-AI Collaboration tends to be most effective on tasks of intermediate difficulty, where humans and AI make different types of errors and can complement each other's weaknesses. It works best when AI provides elimination cues while humans can focus on actual candidates and retain final decision authority. With this reasoning, LogiQA was chosen as the final dataset.

The dataset is originally available at : https://github.com/csitfun/LogiQA2.0

For this assignment, only the validation part was used. Since importing from Hugging Face did not work properly in my local environment, the raw data file was extracted directly and reformatted into a CSV file. The name of the file is logiqa_eval.csv. This file contains a total 651 problems, from which each performer will be given a random set of 5 problems to solve.

In this assignment, Open AI's gpt-o4-mini was chosen and accessed through the API.

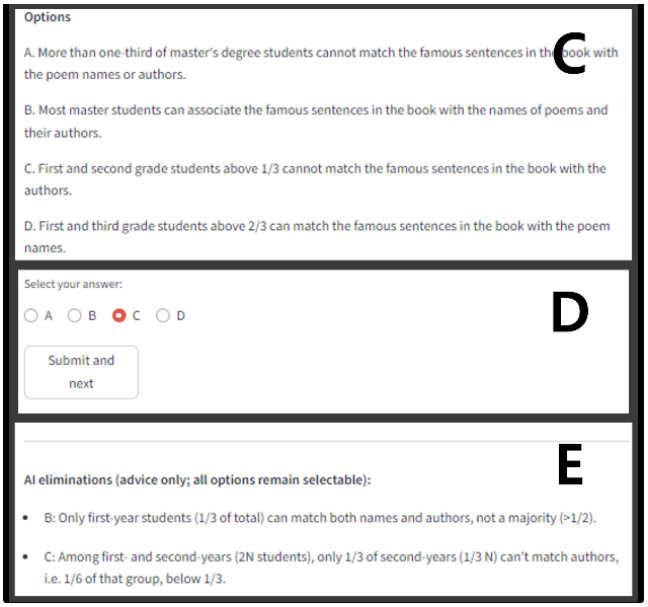
The AI provides assistance before the performer submits an answer. Specifically, the AI evaluates the questions and options, eliminates two options that are most likely incorrect and presents the remaining two options with brief rationales.




<Fig 2. AI Assistance to Human>

Fig. 2 shows the simplified problem solving process. This process allows the performer to focus on distinguishing between the plausible options instead of spending time discarding obvious distractors.

Fig. 3 shows the partial part of expected interface. The E section is written by AI assistance with two wrong options. As it is written in E section and can be seen at D section, human can still ignore the AI's advice and select among the four original options.




<Fig 3. AI Assistance to Human with rationale>

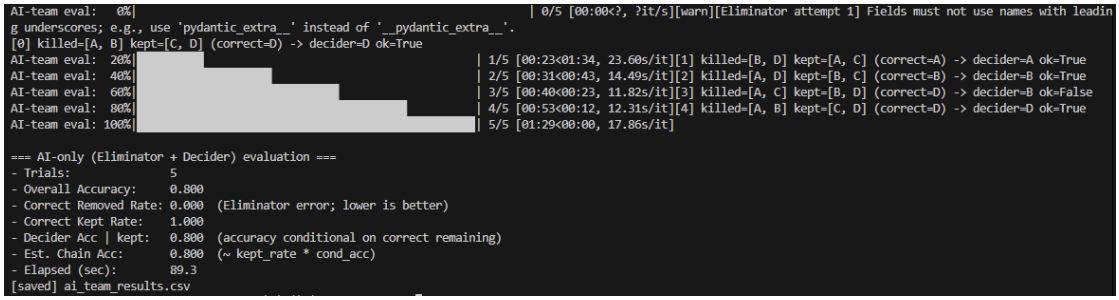
Ultimately, this is about comparing the performance of AI-only, Human-AI Collaboration and Human-only. As it is mentioned earlier, human can ignore AI Assistances' recommendations. However, in the AI-only scenario, I set that recommendations are always accepted. There will be two AI agents in AI-only scenario, one that gives advice by removing wrong options(assistant agent) and one that submits the answer(decider agent).

Solving the quiz alone does not necessarily justify the use of AI, since humans can also achieve correct answers given enough time. However, under time pressure, AI can provide valuable support by quickly eliminating implausible options, reducing cognitive load and enabling humans to focus on more complex reasoning. At the same time, relying on AI-only is dangerous either because AI often generate flawed reasoning and can't guarantee the correctness of their outputs. In particular, if an AI mistakenly eliminates the correct answer, the task becomes unsolvable for humans which highlights the risk of full delegation.

Since these tests have ground truth, human rating was excluded from the evaluation process. Human-AI collaboration will be tested by classmates in Assignment #3, so this assignment focuses only on evaluating the AI-only scenario.

The only evaluation metric chosen is accuracy. There are two types of accuracy for the AI-only scenario: (1) whether the decider agent picks the correct answer, and (2) whether the assistant agent excludes the wrong options properly.

Among all 651 problems, 200 questions were picked and then solved by the AI team. The output was recorded in **ai_team_results.csv**, and the rationales explaining why the assistant agent eliminated certain options are also included in that file.




<Fig 4. AI-Scenario sample results>

As shown in Fig. 4, each line displays the process and summarizes the decisions made by the AI agents.

**killed=[A, C] kept=[D, D] (correct=D) -> decider=D ok=True – (1)**

Sentence (1) indicates that the assistant agent eliminated options A and C, leaving B and D. Among the remaining options, the decider agent chose D, which was the correct answer.

After all the tasks are done, several parameters will be printed out including overall accuracy and correct kept rate which are accuracy of decider and assistant respectively.

In addition, the rationale for each elimination is provided in the file **ai_team_results.csv**. An example is shown below:

**{"A": "States Daejeon failed, contradicting the only derivable conclusion that Daejeon passed.", "C": "Introduces non-taking of the exam, which is unsupported by the given conditional."} – (2)**

Sentence (2) illustrates how the assistant agent's rationales are recorded and presented in the output file.

| | Count | Percentage |
|---|---|---|
| Decider - Accuracy (Right Answer / Total Questions) | 141 / 200 | 70.5% |
| Assistant – Accuracy (Eliminated Properly / Total Questions) | 163 / 200 | 81.5% |

<Table 1. Accuracies>

Table 1 summarizes the two types of accuracy measured in the AI-only setting. The decider achieved an accuracy of 70.5%, while the assistant reached 81.5%. Among the 200 trials, the decider failed in 59 cases. In 37 of these, the correct answer had already been eliminated by the assistant, leaving the decider with no chance of success. This highlights how assistant errors critically constrain the overall system.

We conducted an error analysis of the assistant. When the assistant provides misleading eliminations, it significantly affects the performers' ability to reach the correct answer. We also examined whether such errors could potentially be mitigated in a Human–AI collaboration setting.

To further investigate, we divided the dataset into cases where the assistant mistakenly eliminated the correct answer (37/200) and those where it did not (163/200). We then examined

linguistic cues associated with errors. As shown in Table 2, negation markers (e.g., except, least, cannot, neither, only) were present in 43.2% of error cases but only 9.2% of correct cases, indicating a roughly 4.7× higher risk. Conditional markers (e.g., if, unless, only if) also increased error likelihood (48.7% vs. 38.0%).

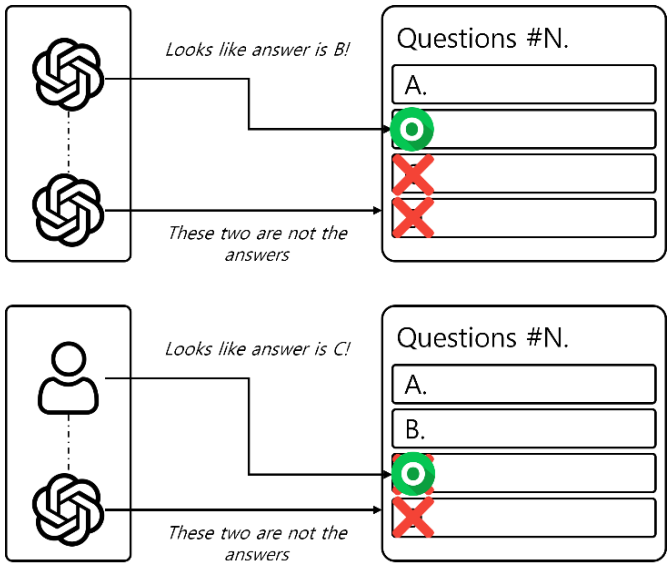
| Type | Keywords | Eliminated Truth(error) | Eliminated Wrong(correct) |
|---|---|---|---|
| Negation | No, not, without, neither, only, cannot, can't never, none, nor | 16/37 = 43.24% | 15/163 = 9.2% |
| Conditional | If, unless, only if | 18/37 = 48.65% | 62/163 = 38.04% |

<Table 2. Error cases>

On the answer side, Table 3 shows that quantifiers and boundary terms (e.g., all, most, at least, more than) occurred more frequently in error cases (29.7%) than in correct ones (17.2%), corresponding to a 1.7× increase. These findings suggest that the assistant is particularly vulnerable to linguistic complexity involving negation, conditionals, and quantification.

| Type | Keywords | Eliminated Truth(error) | Eliminated Wrong(correct) |
|---|---|---|---|
| Quantifier | all, everyone, every, most, some, at least, at most, more than, less than, no more than, no less than | 11/37 = 29.73% | 28/163 = 17.18% |

<Table 3. Error cases>




<Fig 5. Difference between AI-only and AI-Human Collaboration>

Fig. 5 illustrates the key differences between AI-only and Human–AI Collaboration scenarios. In the collaboration setting, humans are not bound to follow the AI's eliminations blindly; they can override incorrect suggestions and still recover the correct answer if the assistant agent

makes a mistake. This flexibility marks the critical distinction between AI-only and Human–AI Collaboration.

When evaluated in terms of accuracy, the predicted ranking is: (1) Human–AI collaboration, (2) AI-only, and (3) Human-only. In terms of completion time, AI-only is likely to be the fastest, followed by Human–AI collaboration and then Human-only.

The rationale is that humans can mitigate some of the AI's errors while still benefiting from the speed of automated elimination, enabling faster and more accurate performance compared to working alone.

Importantly, the earlier error analysis showed that the assistant struggled most with questions involving negation, conditional markers, or quantifiers. In a Human–AI collaboration scenario, performers are not consciously aware of these systematic weaknesses, but their natural reasoning process still provides a safeguard. Even when the AI discards a potentially correct option, humans may recall its relevance from the passage or question wording and hesitate to accept the elimination at face value. This independent judgment reduces the likelihood that such systematic AI errors directly lead to failure, showing how collaboration can attenuate the risks inherent in AI-only decision making.

# Appendix : Code Description

**1. extract.py**

This script is responsible for **preparing the dataset**. It downloads the LogiQA evaluation set from Hugging Face (lgw863/LogiQA-dataset/Eval.txt), reformats the data, and saves it as logiqa_eval.csv. Each row in the resulting CSV includes the question ID, context, question, four multiple-choice options, and the correct answer.

- **Purpose:** To convert the original LogiQA data into a structured CSV format for later use in both AI-only and Human-AI collaboration scenarios.
- **Usage:** `python extract.py`

This generates logiqa_eval.csv in the current directory.

---

**2. ai_only.py**

This script implements the **AI-only evaluation scenario**. It runs two distinct AI agents:

- **Assistant agent**: Eliminates two options deemed least plausible and provides rationales for their elimination.
- **Decider agent**: Selects a final answer from the two remaining options.

The script processes a random subset of the LogiQA evaluation data and produces a detailed log of decisions. Each entry records the eliminated options, the kept options, the decider's final choice, correctness against ground truth, and assistant rationales. Results are saved in ai_team_results.csv.

- **Key Features:**
  - Supports different run modes (--save-raw, --dry-run).
  - Provides verbose logs of AI reasoning.
  - Tracks accuracy separately for the assistant and decider agents.
- **Usage Example:** `python ai_only.py --data logiqa_eval.csv --verbose`

---

**3. app.py**

This script provides a **Streamlit web interface** for the Human-AI collaboration scenario. It randomly selects five questions per session and presents them to the human performer.

- **Interface Sections:**
  1. Reading passage (context)
  2. Question
  3. Multiple-choice options
  4. Human answer input
  5. AI elimination support: the assistant agent eliminates two options and provides rationales, leaving the performer to choose between the remaining two.
- **Additional Features:**
  - A progress bar to track completion.
  - Automatic accuracy calculation across questions.

  - Option to download a session log (including context, question, options, AI eliminations, human answers, correctness, and time spent).

- **Usage:** pyhton -m streamlit run app.py

**4. compute_metrics.py**

This script provides decider and assistant's accuracy based on the result of app.py. The result is written on ai_team_results.csv.

- **Key Features:**
  - Calculate accuracy of decider and assistant
  - Output will be shown like below

    ==== Recomputed Metrics ====

    Trials: 200

    Decider correct: 141 / 200 (70.5%)

    Assistant kept-correct: 163 / 200 (81.5%)

- **Usage Example:** python .\compute_metrics.py