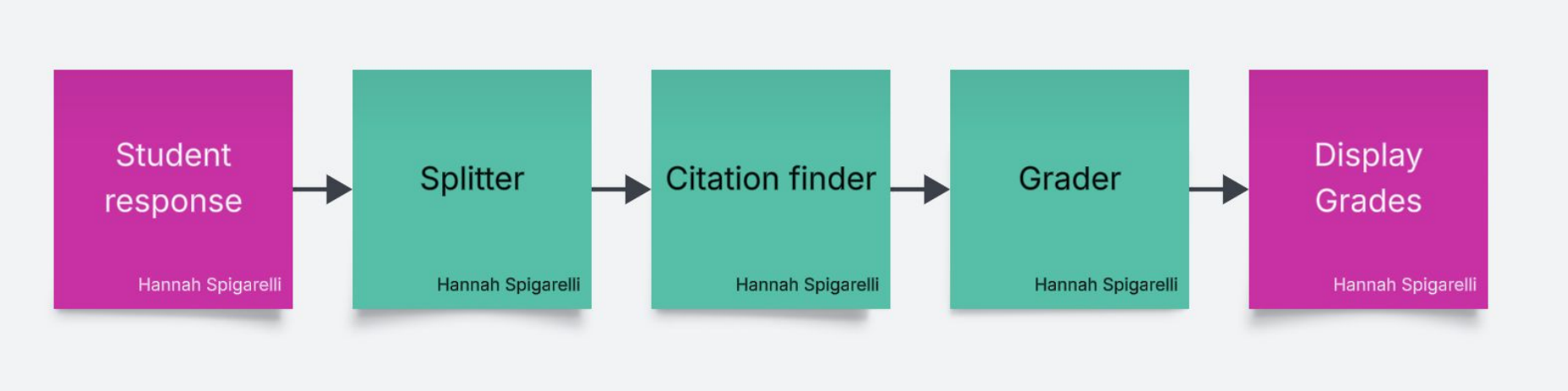
AI Grader for CS 452

In CS 452 students learn about what tends to make software projects succeed or fail. For the midterm they are given five hypothetical scenarios and they need to identify 5 risks and/or mitigations for the scenario. They are supposed to back up each part of their response with a citation to one of the course readings. Grading the midterm takes hours, and many responses are close to the same. It's easy for a human grader to begin missing details. I believe that having an AI system that can assess the quality of an answer based on the grading rubric would significantly improve the fairness and accuracy of the grading.

Project structure

For this project I take the initial student response to each question, and then run it through a series of agents that eventually leads to outputting a grade for the response.



Example question

A large enterprise is refactoring its monolithic enterprise resource planning (ERP) system into a containerized, cloud-native architecture with AI-enhanced predictive analytics for supply chain optimization. The 20-person team includes a mix of legacy experts and new hires specializing in DevOps, but they're operating in a hybrid remote/office setup with varying tool proficiencies (e.g., some prefer Kubernetes, others Docker Compose). Midway through, a cyber threat actor exploits a vulnerability in a third-party library, prompting an urgent security patch that disrupts ongoing migrations, while executives demand accelerated rollout to align with quarterly earnings. Requirements creep in as business units request custom dashboards without impact assessments, and testing is fragmented across automated CI/CD pipelines and manual reviews due to incompatible environments. Identify five risks and/or mitigations (numbered #1 through #5), with explanations and citations. Then, state AI usage.

Student input

#1: Risk: Toolchain inconsistencies in hybrid setups leading to deployment failures. With team members using varying DevOps tools like Kubernetes vs. Docker Compose, integration errors could arise during containerization, delaying the ERP refactor and amplifying downtime risks. Citation: DeMarco & Lister, Peopleware, Ch. 5 (discusses how mismatched environments erode productivity and foster "flow" interruptions) #2: Mitigation: Standardize tooling through team training sessions. Form a subgroup to enforce unified tools and conduct workshops, ensuring compatibility in CI/CD pipelines to streamline migrations and patches. Citation: Lecture on Team Dynamics (Knutson podcast #4), emphasizing sociological alignment in ambiguous tech stacks #3: Risk: Security vulnerabilities from third-party libraries post-exploit. The urgent patch could introduce regressions in AI analytics if not thoroughly tested, potentially exposing supply chain data in a high-stakes enterprise system. Citation: Brooks, The Mythical Man-Month, Ch. 6 (highlights the challenges of system integration and unforeseen bugs in complex additions like patches) #4: Mitigation: Implement incremental rollouts with monitoring. Deploy changes in stages, using feedback loops to catch issues early from the cyber incident, balancing speed with robustness. Citation: Brooks, The Mythical Man-Month, Ch. 7 (advocates for incremental planning and safeguards against complexities in large-scale projects) #5: Risk: Requirements creep from business units without assessments. Adding custom dashboards mid-refactor could inflate scope, straining resources already diverted by the security incident and executive pressures. Citation: DeMarco & Lister, Peopleware, Ch. 3 (warns about managerial pressures that lead to unchecked interruptions and scope expansion) AI usage: I used Grok with the prompt “Give me a list of risks that cloud-native refactors could have involving security incidents.” I then built off this list and combined with the preliminary ideas to ensure I covered everything we talked about in class.

Student input

Problems

- Submission software is not great quality,
  - \n characters were removed from student responses
  - No easy way to export responses
- Guidelines for answer submission are very relaxed
  - No standardized format for indicating each risk/mitigation
- All 5 components of the answer are in one text box

Splitter

Goal

- Split full question answer into 6 components. One for each risk/mitigation and one for the explanation of AI usage

Did it work? - Mostly!

- Almost all answers were split perfectly!
- A few students didn’t follow the instructions and explained their AI usage throughout the question, or submitted ‘bonus’ risks/mitigations beyond the required 6
- A few responses ended up having the AI usage show up in both risk/mitigation 5 and in the AI usage category.

Splitter

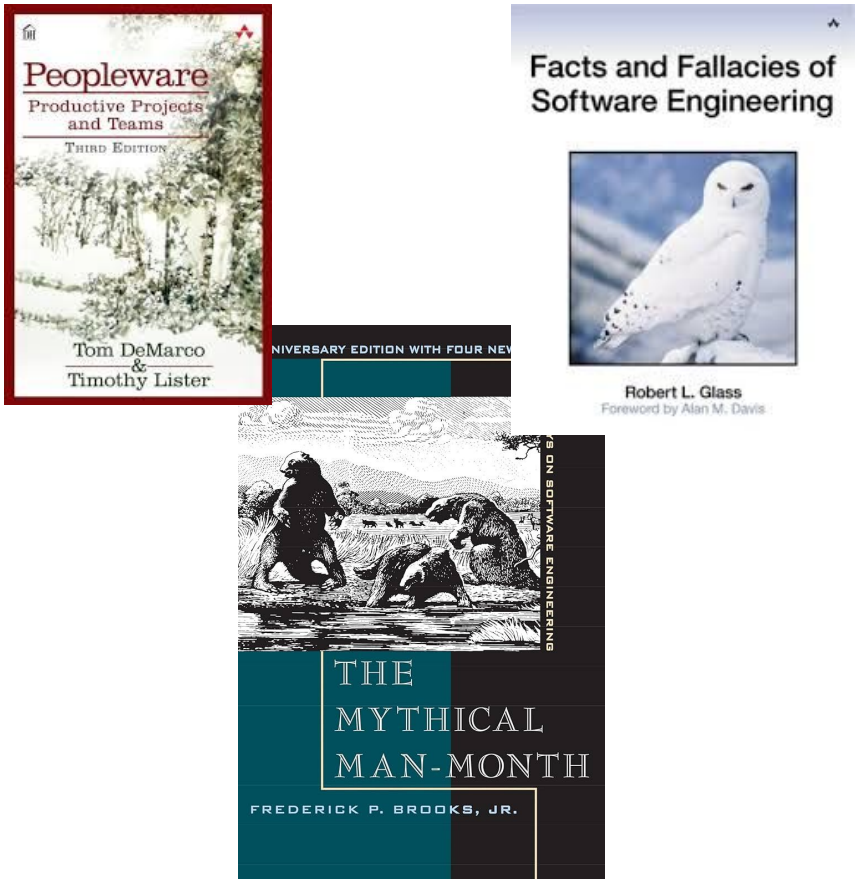

```
{
   "Students": [
      {
         "Student": "sample",
         "Question 1": {
            "Risk/mitigation 1": "#1: Risk: Toolchain inconsistencies in hybrid setups leading to deployment failures. With team members using varying DevOps tools like Kubernetes vs. Docker Compose, integration errors could arise during containerization, delaying the ERP refactor and amplifying downtime risks. Citation: DeMarco & Lister, Peopleware, Ch. 5 (discusses how mismatched environments erode productivity and foster \"flow\" interruptions)",
            "Risk/mitigation 2": "#2: Mitigation: Standardize tooling through team training sessions. Form a subgroup to enforce unified tools and conduct workshops, ensuring compatibility in CI/CD pipelines to streamline migrations and patches. Citation: Lecture on Team Dynamics (Knutson podcast #4), emphasizing sociological alignment in ambiguous tech stacks",
            "Risk/mitigation 3": "#3: Risk: Security vulnerabilities from third-party libraries post-exploit. The urgent patch could introduce regressions in AI analytics if not thoroughly tested, potentially exposing supply chain data in a high-stakes enterprise system. Citation: Brooks, The Mythical Man-Month, Ch. 6 (highlights the challenges of system integration and unforeseen bugs in complex additions like patches)",
            "Risk/mitigation 4": "#4: Mitigation: Implement incremental rollouts with monitoring. Deploy changes in stages, using feedback loops to catch issues early from the cyber incident, balancing speed with robustness. Citation: Brooks, The Mythical Man-Month, Ch. 7 (advocates for incremental planning and safeguards against complexities in large-scale projects)",
            "Risk/mitigation 5": "#5: Risk: Requirements creep from business units without assessments. Adding custom dashboards mid-refactor could inflate scope, straining resources already diverted by the security incident and executive pressures. Citation: DeMarco & Lister, Peopleware, Ch. 3 (warns about managerial pressures that lead to unchecked interruptions and scope expansion)",
            "AI usage": "I used Grok with the prompt \u201cGive me a list of risks that cloud-native refactors could have involving security incidents.\u201d I then built off this list and combined with the preliminary ideas to ensure I covered everything we talked about in class."
         }
      }
   ]
}
```

Book Embeddings and Files

The course readings include:

- Mythical Man month
- Facts and Fallacies
- Peopleware
- Assorted articles, mostly from Bruce F. Webster

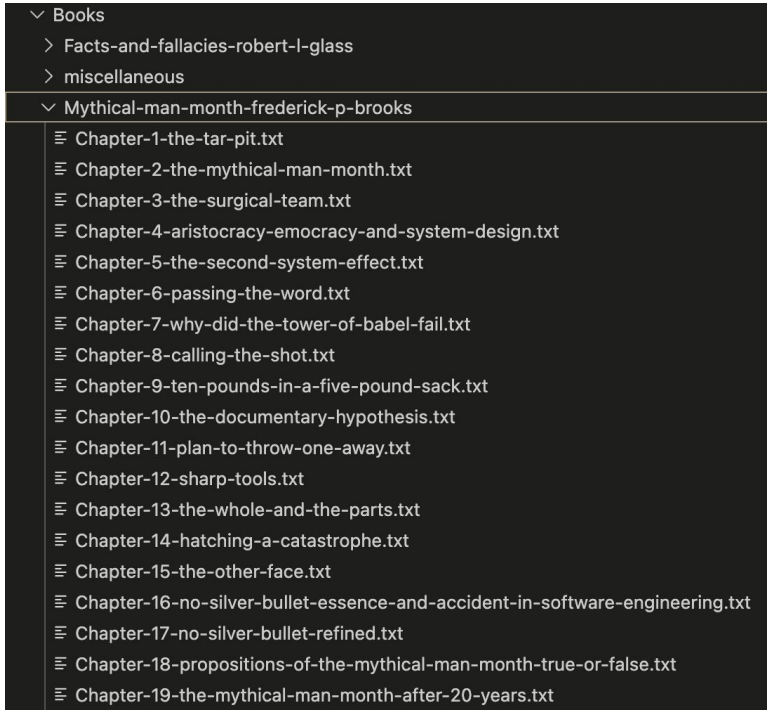
In order to grade the response I needed to check the accuracy of each citation based on the course readings.



Book Embeddings and Files

What I tried:

- Embedding each book and chapter
  - Successfully created the embeddings
  - Didn’t have enough expertise to take next steps
- Adding citation text to LLM as direct context for answer
  - Associated each risk/mitigation with content file
  - Loaded content as context when calling grader



Citation Finder

```
Processing student: [redacted]
  Question 1
    Processing Risk/mitigation 1... ✓ Mythical-man-month/Chapter-4-aristocracy-emocracy-and-system-design.txt
    Processing Risk/mitigation 2... ✓ miscellaneous/the-dead-sea-effect.txt
    Processing Risk/mitigation 3... ✓ miscellaneous/the-five-orders-of-ignorance.txt
    Processing Risk/mitigation 4... ✓ Facts-and-fallacies/chapter-5-about-management.txt
    Processing Risk/mitigation 5... ✓ Peopleware/chapter21.txt
  Question 2
    Processing Risk/mitigation 1... ✓ Mythical-man-month/Chapter-7-why-did-the-tower-of-babel-fail.txt
    Processing Risk/mitigation 2... ✓ Peopleware/chapter1.txt
    Processing Risk/mitigation 3... ✓ miscellaneous/do-not-defer-the-difficult-in-it-projects.txt
    Processing Risk/mitigation 4... ✓ Facts-and-fallacies/chapter-6-about-the-life-cycle.txt
    Processing Risk/mitigation 5... ✓ miscellaneous/remember-conways-law.txt
  Question 3
    Processing Risk/mitigation 1... ✓ Mythical-man-month/Chapter-2-the-mythical-man-month.txt
    Processing Risk/mitigation 2... ✓ miscellaneous/the-five-orders-of-ignorance.txt
    Processing Risk/mitigation 3... ✓ miscellaneous/anatomy-of-a-runaway-it-project.txt
    Processing Risk/mitigation 4... ✓ Peopleware/chapter3.txt
    Processing Risk/mitigation 5... ✗ NOT_FOUND
  Question 4
    Processing Risk/mitigation 1... ✓ Peopleware/chapter1.txt
    Processing Risk/mitigation 2... ✓ Peopleware/chapter10.txt
    Processing Risk/mitigation 3... ✓ Facts-and-fallacies/chapter-6-about-the-life-cycle.txt
    Processing Risk/mitigation 4... ✓ miscellaneous/remember-conways-law.txt
    Processing Risk/mitigation 5... ✓ Mythical-man-month/Chapter-2-the-mythical-man-month.txt
  Question 5
    Processing Risk/mitigation 1... ✓ Mythical-man-month/Chapter-4-aristocracy-emocracy-and-system-design.txt
    Processing Risk/mitigation 2... ✓ miscellaneous/getting-technology-lifecycles-in-sync.txt
    Processing Risk/mitigation 3... ✓ miscellaneous/the-dead-sea-effect.txt
    Processing Risk/mitigation 4... ✓ Mythical-man-month/Chapter-11-plan-to-throw-one-away.txt
    Processing Risk/mitigation 5... ✓ miscellaneous/do-not-defer-the-difficult-in-it-projects.txt

Processing student: [redacted]
```

```
============================================================
Processing complete!
Total citations processed: 780
Citations found: 758
Citations not found: 22
Output saved to: responses_with_citations.json
============================================================
```

Citation Finder

Outcome

- Added to JSON objects a new value for each citation

Did it work? - Mostly!

- Ran it 5 times with slightly different prompts and additional context
- Depending on the run failed to find 22-60 citations out of 780. Only 5 responses had citations missing
- Found that changing chat model and prompt didn’t seem to make a significant difference
- Each run would vary slightly in what it identified, occasionally changing citation (ie if student said “Chapters 5-7” one run would identify chapter 5 as the citation, and another run would identify chapter 6 or 7 as the citation)

Grader

Passed the question prompt, risk/mitigation text and citation, along with the grading rubric to the Open AI API asking it to grade the question.

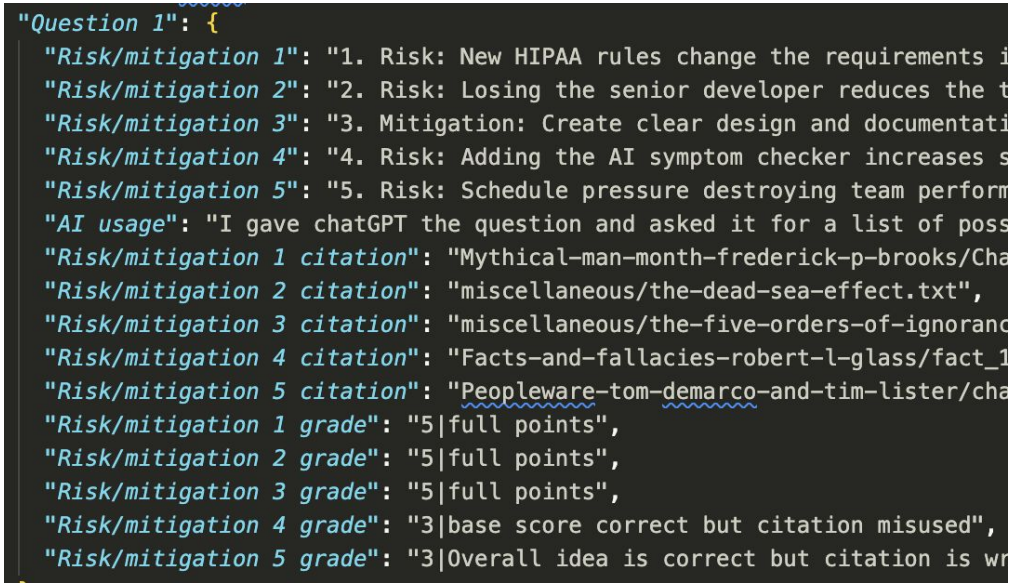
I batched the responses, feeding in responses that all cited the same course content at the same time so that I didn't have to load the context files as much and could therefore be more efficient.

```
rubric = """GRADING RUBRIC:
- 0 points: No text entered for answer (answer couldn't be found)
- 2 points: Factually incorrect
- 3 points: Overall idea is correct, but citation is wrong or misused
- 4 points: Answer is only 1-2 sentences
- 5 points: Answer is 3+ sentences
This is for each part of the question. Then the question as a whole can have these deductions:
- -1 point: Explained AI usage but didn't provide exact prompt explanation
- -2 points: Only stated they used AI without explaining how, OR didn't state whether they used AI

IMPORTANT: Start with the base points (0-25), then apply AI usage penalties at the end for the overall question (becomes 23-24 instead of 25)."""
```

Grader

The JSON object shapes after running Grader:



```
"Question 1": {
  "Risk/mitigation 1": "1. Risk: New HIPAA rules change the requirements i
  "Risk/mitigation 2": "2. Risk: Losing the senior developer reduces the t
  "Risk/mitigation 3": "3. Mitigation: Create clear design and documentati
  "Risk/mitigation 4": "4. Risk: Adding the AI symptom checker increases s
  "Risk/mitigation 5": "5. Risk: Schedule pressure destroying team perform
  "AI usage": "I gave chatGPT the question and asked it for a list of poss
  "Risk/mitigation 1 citation": "Mythical-man-month-frederick-p-brooks/Cha
  "Risk/mitigation 2 citation": "miscellaneous/the-dead-sea-effect.txt",
  "Risk/mitigation 3 citation": "miscellaneous/the-five-orders-of-ignoranc
  "Risk/mitigation 4 citation": "Facts-and-fallacies-robert-l-glass/fact_1
  "Risk/mitigation 5 citation": "Peopleware-tom-demarco-and-tim-lister/cha
  "Risk/mitigation 1 grade": "5|full points",
  "Risk/mitigation 2 grade": "5|full points",
  "Risk/mitigation 3 grade": "5|full points",
  "Risk/mitigation 4 grade": "3|base score correct but citation misused",
  "Risk/mitigation 5 grade": "3|Overall idea is correct but citation is wr
```

Grader

Outcome

- Added to JSON objects a new value for each citation

Did it work? - Not really

- Incorrectly reported on many student's responses for not having understood the cited reading or for not having written enough content
- My prompt generation skills are significantly lacking
  - Tried multiple different prompt iterations without significant improvement
- Overall didn't follow instructions well - consistently said responses with 3-4 sentences should be worth less points because they were only 1-2 sentences

Displaying Grades

- Displayed information outputted from the Grader step
- Gave a good starting point but ultimately inaccurate enough that it wasn't very helpful