Illinois Summer CS Teaching Workshop

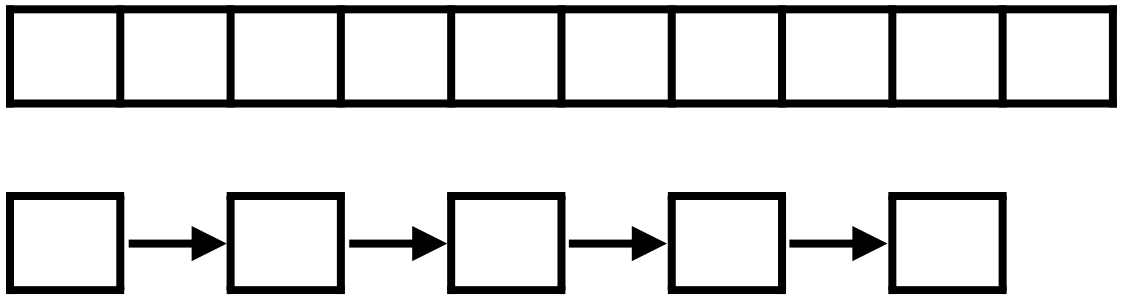
A Scalable Comparison of AI vs Manual Grading of CS Theory Content in Data Structures

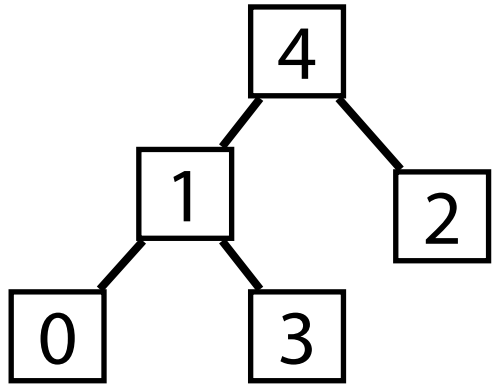
Brad Solomon

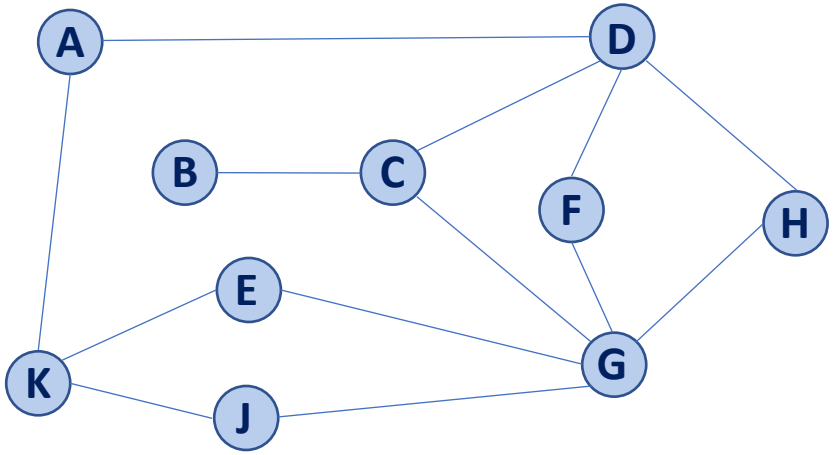
Siebel School of Computing and Data Science

CS 2(25): Algorithms and Data Structures

What theory content?







CS 2(25): Algorithms and Data Structures

Theory Goal 1: Efficiency of Data Structures

Passable: Memorize Big O for each data structure

	Singly Linked List	Array
Look up **arbitrary** location	O(n)	O(1)
Insert after **given** element	O(1)	O(n)
Remove after **given** element	O(1)	O(n)
Insert at **arbitrary** location	O(n)	O(n)
Remove at **arbitrary** location	O(n)	O(n)

Mastery: Understand why and apply in novel contexts!

CS 2(25): Algorithms and Data Structures

Theory Goal 2: Propose and justify design decisions

You are designing a **historical database** where individuals are tracked by a unique ID. As genealogical records are uncovered, previously unknown relationships between people are discovered, merging families. The system must efficiently **determine whether two people share a common relative** and update family groups as new relationships are documented.

CS 2(25): Algorithms and Data Structures

Theory Goal 2: Propose and justify design decisions

Passable: Pick your favorite data structure that solves the problem

E.g. Solve a disjoint set problem using a map implemented as AVL Tree

Functionally it can work in small scale, in practice its slow!

Mastery: Pick an optimal data structure and be able to justify it!

E.g. Using smart union by size and path compression, disjoint set is $O(\alpha(N))$

CS 2(25): Algorithms and Data Structures

Additional Challenges:

800 — 1000 students per semester

Not the only learning goals (~10 minutes of time on 50 minute exam)

Partial credit is necessary for those still developing mastery

How can we assess theory on this level in an exam?

CS 2(25): Algorithms and Data Structures

Additional Challenges:

800 — 1000 students per semester

Not the only learning goals (~10 minutes of time on 50 minute exam)

Partial credit is necessary for those still developing mastery

How can we assess theory on this level in an exam?

Manually graded free-response questions

Free Response Theory Question Overview

Two broad categories of problem, one for each 'theory goal':

1) Given <data structure> with <twist>, explain Big O to do <X>.

Encourages students to study 'why' rather than memorize Big O

2) Given <open ended problem>, propose and justify implementation

Encourages students to review material and identify optimal use cases

FRQ Assessment Strategy

To ensure accuracy, TAs are trained in rubric-based grading

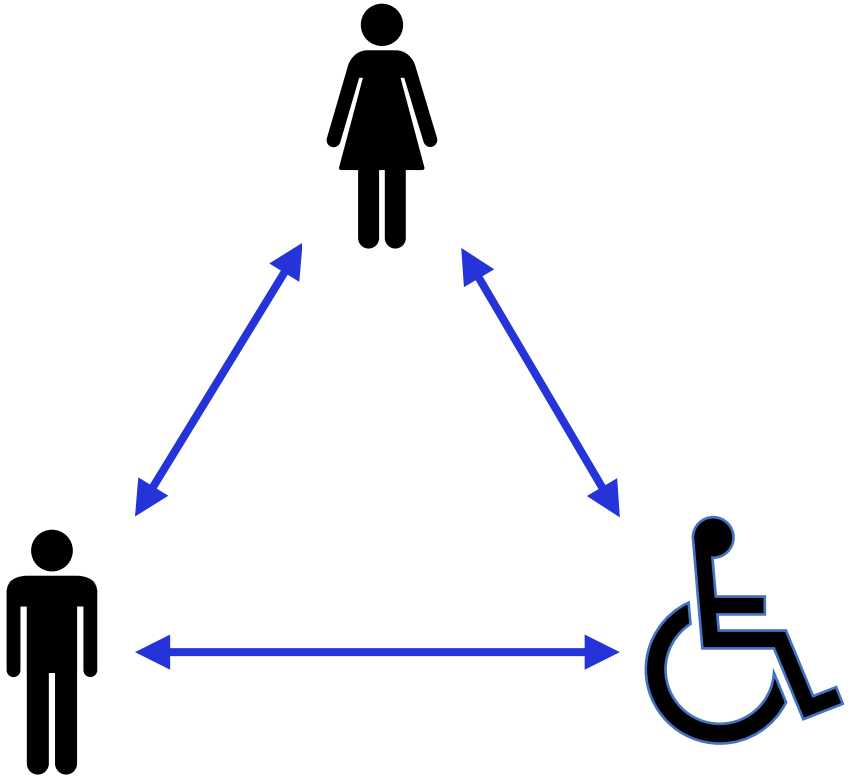
Points	Rubric Item
2	Overall Big O correct
3	Minimum necessary steps described
1	Individual step Big O correct
2	Explain how ‘twist’ affects Big O
1	Answer given in complete sentences
1	Answer has appropriate structure

A single TA has to grade between ~35 - 70 submissions per exam

FRQ Assessment Strategy

For consistency, TAs peer evaluate as assigned teams

All TAs assigned a problem grade all peers



0—40%, 50—80%, 90—100% grade entry

Please give the link to a 90–100% grade submission you are double checking. *

Your answer

Is this grade correct? *

- ◯ Yes -- I agree with the grading
- ◯ No -- I think the grading is incorrect
- ◯ Maybe -- its ambiguous and would like another opinion

If you answered no or maybe, explain why.

Your answer

Enter the era of AI…

The system works — but could it be better?

TA time is valuable — is grading the best use of their time?

TA grading can be inconsistent (though peer evaluation helps)!

I want more FRQs but don’t have the staff to run them!

An alternative AI approach

1) Staff training proceeds unchanged

2) AI grader acts as a **first pass filter**

3) TA grading becomes **team oversight**

4) Final regrade requests proceed unchanged

Study goals motivated by this pipeline

1) Staff training proceeds unchanged

2) AI grader acts as a **first pass filter**

What population or question type does the LLM struggle with?

3) TA grading becomes **team oversight**

How good are LLMs at rubric-based grading of CS theory content?

4) Final regrade requests proceed unchanged

The data used

Motivation was years of data: ~4000 unique submissions per exam

… due to data privacy we can't use them except on secure machines

Narrowed preliminary study, focusing on Spring 2026:

We have ~300 student's who have provided **informed consent**!

The Study Objectives

How good are LLMs at rubric-based grading of CS theory content?

Summer 2026:

1. Develop and validate an automated pipeline for LLM FRQ grading

2. Evaluate consistency and accuracy of **numeric grading values**

Fall 2026:

3. Evaluate accuracy and clarity of LLM-generated **feedback**

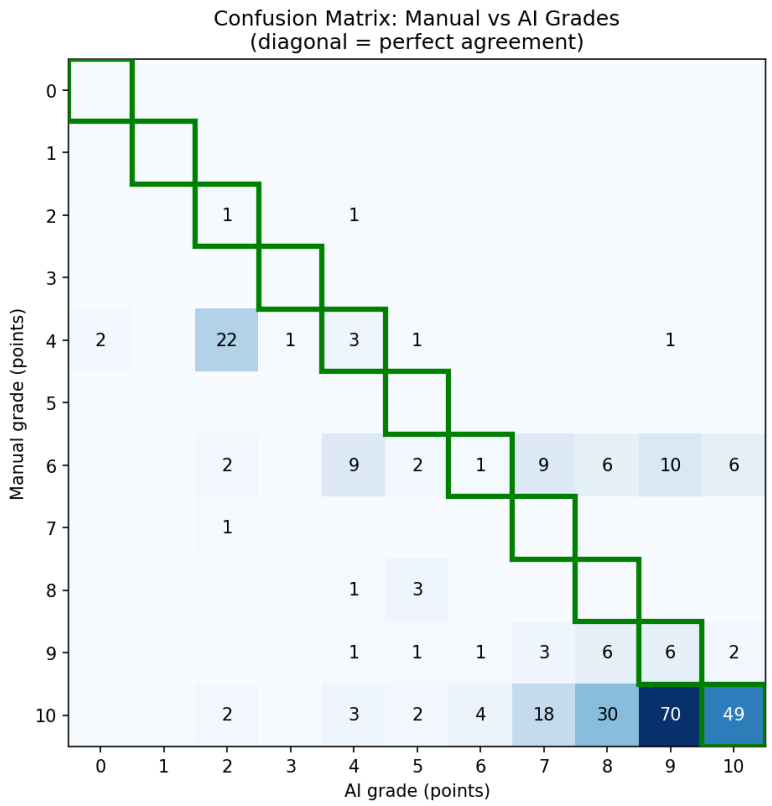
Models Tested

(Per 1M Tokens)

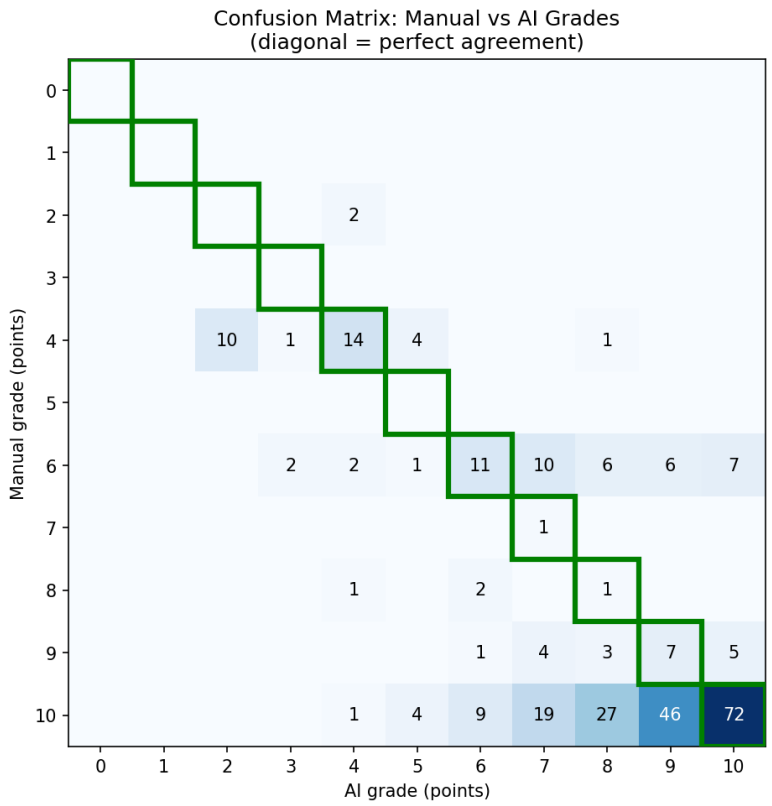
Model	Expected Input Price	Expected Output Price	Cache Capable (in context of study)	Approximate Cost
Claude Haiku 4.5	$1	$5	No	$0.68
Claude Sonnet 4.6	$5	$15	TQ2 only	$2.61
Claude Opus 4.8	$5	$25	TQ1 and TQ2	$3.67**

Theory Question 1 (Big O with a twist)

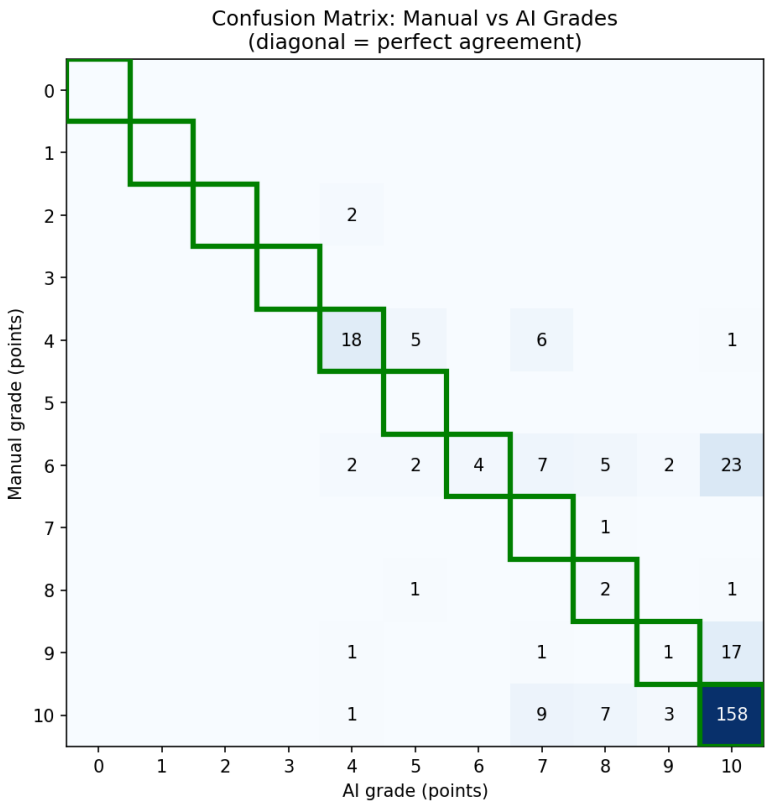
Haiku 4.5



Sonnet 4.6

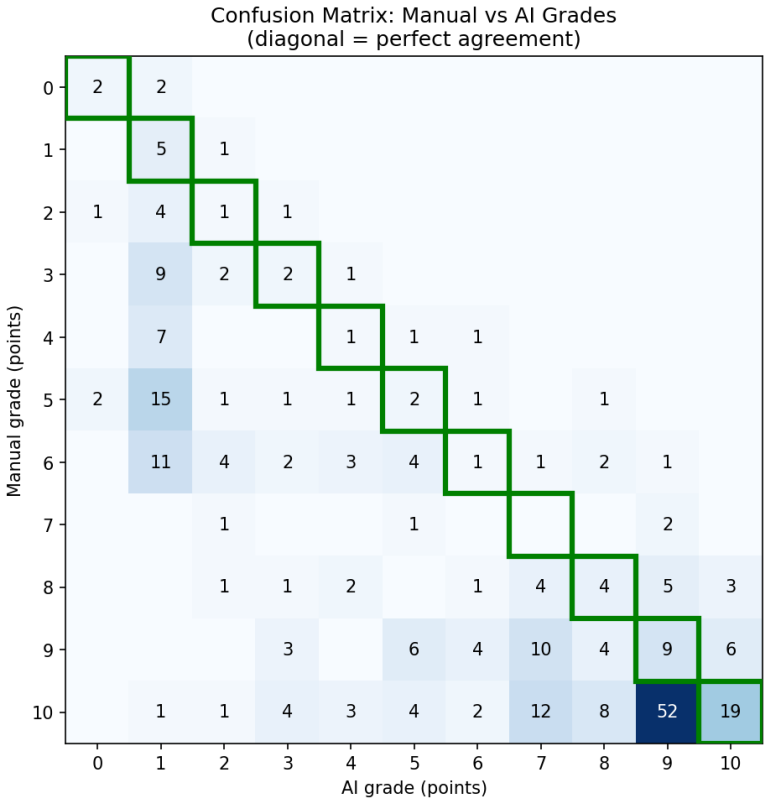


Opus 4.8

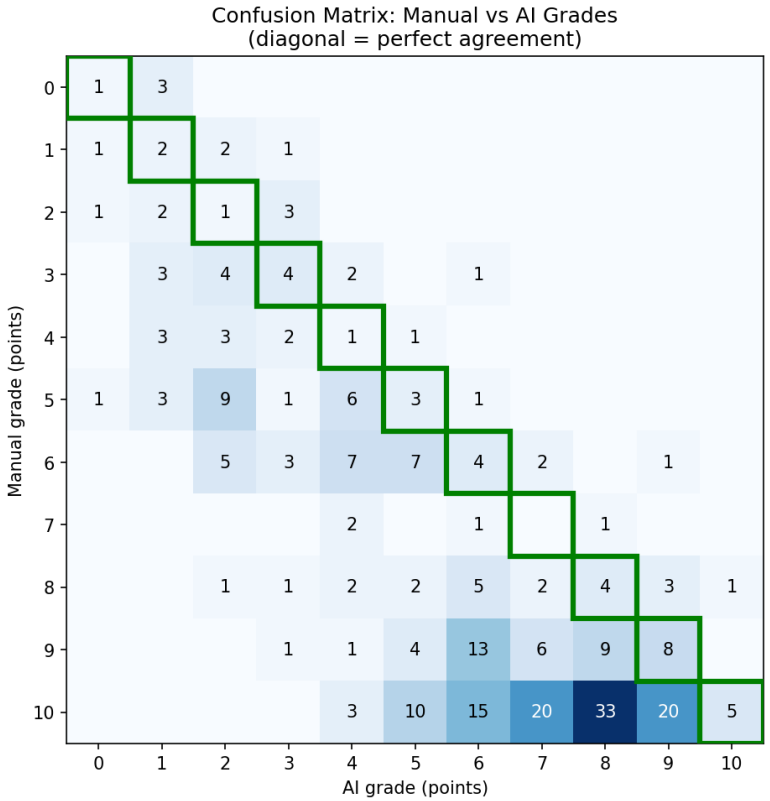


Theory Question 2 (Justify design decision)

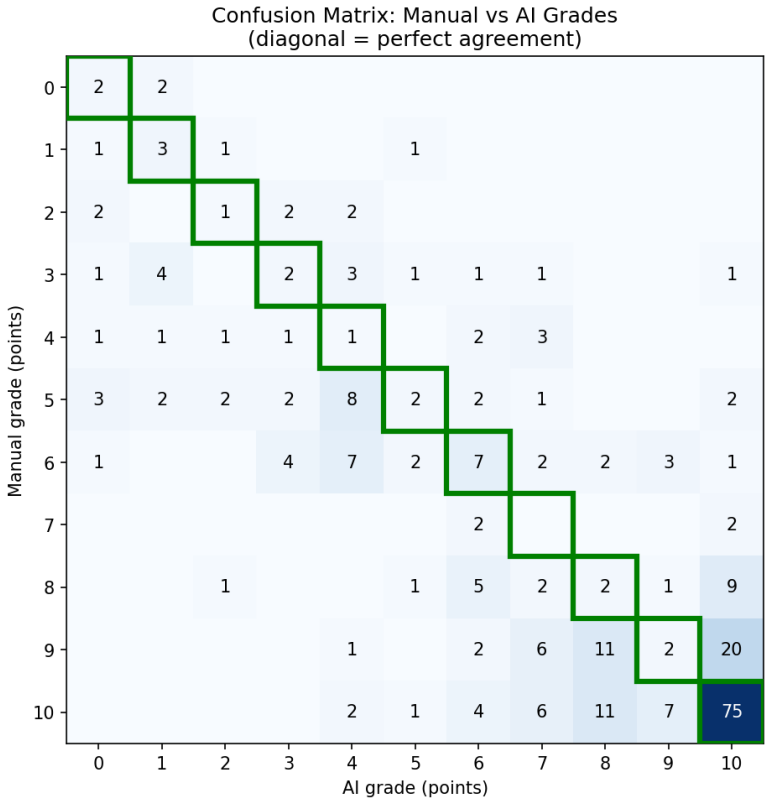
Haiku 4.5



Sonnet 4.6



Opus 4.8



Preliminary Findings

(Quadratic Weighted)

Model	Mean Abs/Signed Difference (TQ1)	Mean Abs/Signed Difference (TQ2)	Cohens Kappa (TQ1)	Cohens Kappa (TQ2)
Claude Haiku 4.5	1.61 / -1	2.08 / -1.78	0.64	0.66
Claude Sonnet 4.6	1.28 / -0.67	2.12 / -1.91	0.67	0.63
Claude Opus 4.8	**0.84 / 0.36**	**1.35 / -0.43**	**0.7**	**0.78**

Preliminary Findings

Batch / Cached queries extremely cheap to run at scale

Mid-range (and below) models are not sufficiently accurate

TA grading still required for all problems (no clear filtering step)

Will TAs prefer oversight on LLM grades versus hand-grading?

Unexpected: Sonnet did worse than Haiku on complex grade case

What's next…?

Adjusting rubrics to clarify common AI pitfalls

Pipeline improvements including:

Post processing of results to correct inconsistencies

Adjustments for isomorphic variation batching

Expanding space of tested models / focusing on what is locally hosted

Not yet tested: Goodness of feedback

Questions?

Questions for Audience

1. Would you consider using LLM graders as part of your own class?

2. What sort of work would you give to a TA who now has ~5-10 hours of extra time per week?