Assignment 0

Building a Memory-Safe Red-Black Tree in C Under Kernel-Style Constraints

Custom allocators, memory-safety verification, and disciplined AI-assisted development

Version 1.0 July 21, 2026

Version history appears in Section 21.

Contents

Nota Bene: This assignment arrives before the semester begins. That is deliberate. It is a preview in two senses: it previews the *shape* every assignment in this course will take, so the form is familiar by the time the stakes are not, and it is itself the first of those assignments, meant to be worked, not skimmed. Work it the way the later ones will *demand*: through the steps as they are laid out, not around them. The effort compounds, too. The red-black tree you are about to build is not a self-contained exercise; it is the load-bearing structure of Assignment 4, where a nearly identical tree will sit at the heart of a CPU scheduler that you will build. What you understand now, you will lean on then. What you merely borrow now, you will have to explain then, to a version of yourself who has forgotten where it came from. One thing about that shape is genuinely new, and it is the reason for this note: for the first time in a CS course here, an AI coding agent (Claude Code) is not merely tolerated but *required*. Working with Claude Code is woven into both the workflow you will follow and the grade you will earn. The tool many of you already reach for, with varying degrees of permission, is now the one I am openly asking you to master. That turns out to be a very different assignment. As you can see, the assignments will also be commensurately harder.

1 Overview

Everything in this assignment orbits a single data structure. A red-black tree is a small thing to describe and a demanding thing to keep alive. Every node is a separately allocated heap object stitched to its neighbors by pointers. Every insertion may allocate several times and then rebalance by rewiring pointers far from the insertion point. Every deletion must repair global invariants while releasing memory in exactly the right order. That combination makes the red-black tree an ideal specimen for studying memory management, which is the part this course actually cares about.

You will build the tree once (Part 1) and then keep it correct while the assignment repeatedly changes the rules about *where* its bytes live and *how* they may be reclaimed (Part 2). By the end, the logical tree is unchanged: same API, same invariants, same answers. Everything underneath it has been replaced. If you finish with the sense that the balancing algorithm was the easy part, the assignment has worked. The structure matters beyond the assignment's perimeters, too. The Linux kernel ships its own red-black tree library and leans on it wherever predictable $O(\log n)$ behavior and tight memory control must coexist: most famously in the CFS scheduler's runqueue (one of the crown jewels of the CPU-scheduling module, where we will study it in depth) and, for several years, in tracking the memory regions of every process's address space.

The core theme here is *discipline under a moving target*. Writing a red-black tree that returns the right answer is fairly straightforward. Keeping that tree correct when allocation can fail at an arbitrary point, when nodes must come from your own allocator, when there is no stack to recurse on, and when two logical trees quietly share the same bytes: that is the harder skill, and the one this assignment is actually grading. A tree that looks up the right key but leaks, double-frees, or recurses unboundedly fails. Every correct answer it returns is beside the point.

Grading methodology and the Quiz that you can't Google

This assignment is graded out of **10 points**, and those points answer two different questions, calling for two different kinds of evidence.

7 points answer the question the tools already ask: *does the program work, and does it work without corrupting memory*? That question is settled for you, mechanically and without opinion, by the test batteries, the fault-injection sweep, and AddressSanitizer and valgrind.

The remaining **3 points** answer a harder question: *do you understand what you built?* **1 point** comes from an analysis of your Claude Code session transcripts: whether they show the short, specific, iterative exchanges of someone building and debugging their own program, or the single large paste of someone outsourcing it. **2 points** come from a short quiz generated from your own code, administered at a live walkthrough: your own function names, your own error-path structure, why one of your teardown loops cannot run forever, what frees a given allocation on a given failure path. Two students can submit code that behaves identically on every test case and still walk away from that quiz with very different scores, because the quiz was never testing the program. It was testing you.

There is a reason for the split. A working red-black tree is a nice thing to have. A student who can explain, unprompted, why a failed insert must leave the tree byte-for-byte as it was is a nice thing to become. Most of the hours you put into this assignment build the former. The prompt analysis and the quiz at the end are where we find out whether you also built the latter.

2 Learning Objectives

- Implement a red-black tree in C and reason precisely about node ownership, lifetimes, and the constant-time pointer surgery in rotation and fixup.
- Separate the logical data structure from its memory model, and hold the former invariant while the latter is replaced: fault-tolerant allocation, a slab pool, $O(1)$-space traversal, and copy-on-write sharing.
- Write allocation-failure-safe code that unwinds cleanly, mirroring kernel paths that must cope with `kmalloc` returning `NULL`.
- Verify memory-management correctness with sanitizers and valgrind, treating a single leaked byte or invalid access as a failure, not a warning.
- Design and defend a C module against a frozen public API: struct layout, ownership contract, and the error-code discipline the grading harness depends on.
- Direct an AI coding agent (specifically, *Claude Code*) the way a senior engineer directs a capable but overconfident junior: give it context, make it plan, keep changes small, verify everything with tools, and merge nothing you cannot explain.

Item	Detail
Language	C (C23), compiled with `gcc` and `make`, must build cleanly with `-Wall -Wextra -Werror`
Verification	AddressSanitizer/UBSan (`make asan`) and valgrind (`make memcheck`); a single leak or UB report caps that milestone's correctness at 50%
Expected effort	Approximately 12–18 hours across three milestones, including required use of Claude Code
Tooling	Claude Code is required for this assignment (see Section 11)

3 On Asking Us for Help

This specification is the support. It is deliberately long, deliberately specific, and deliberately finished: every algorithm, every interface, every threshold you need is somewhere in these pages. Where a choice is left to you, that is the assignment, not an omission. What this document is not is the opening move in a conversation. We will *not* be answering questions about this assignment. That is a considered decision: not an oversight or a shortage of goodwill.

Here is the reason. The skill this assignment builds cannot be handed to you across a desk, or over a Teams/Zoom call. You are working with a tool that will answer any question you put to it, at any hour, in unlimited quantity. Finding an answer is no longer the hard part. Deciding which answers to trust, and how to check them, is. A student who brings us "what should my node struct look like?" has skipped the exercise. A student who asks the agent, argues with what comes back, tests it, and can then tell us why the answer was wrong has done the whole of it. Every question we answer for you is one you never learn to answer for yourself, and the semester ahead assumes you have learned it here.

Treat the spec as a map, not a hotline. When you are stuck, the order of resort is: reread the relevant section, ask Claude Code, write a test that settles the question, and (failing all of that) make a defensible choice and document why in your design notes. That last option is not a consolation prize. Researchers spend most of their careers there.

One last thing, and it is the part most likely to be misread. This assignment is worth **zero** points. Zero. That is not a signal to skim it; it is the only invitation of its kind you will get. Every later assignment counts, which means this is the one place where a wrong turn costs you *nothing* but the afternoon. So *please* take the wrong turn, break the tree, argue with Claude Code, and find out what you actually understand while the answer is still free. Take it seriously precisely because it is safe to.

We start on **Tuesday, August 25th**. Lectures are Tuesdays and Thursdays, 2:00–3:15 pm, in *Natural Resources* Room 140. Show up having built something.

4 What is a Red-Black Tree?

A red-black tree is a binary search tree that keeps itself approximately balanced by attaching one bit of bookkeeping, a *color*, to every node. Five rules govern the colors:

1. Every node is red or black.
2. The root is black.
3. Every leaf is a black sentinel node, written NIL.
4. A red node has only black children, so reds never stack.
5. Every path from a given node down to a NIL leaf crosses the same number of black nodes, called the *black-height*.

Rules 4 and 5 together squeeze the tree's shape. The longest possible root-to-NIL path alternates red and black, while the shortest is all black, so no path can be more than twice as long as any other. From this it follows that the height satisfies $h \leq 2\log_2(n+1)$, and search, insert, and delete all run in $O(\log n)$ worst-case time. Figure 1 shows a valid tree. The figures here use small integer keys, which make the ordering easy to see at a glance; your implementation stores C-string keys compared with `strcmp`, but the structure and the invariants are identical either way.

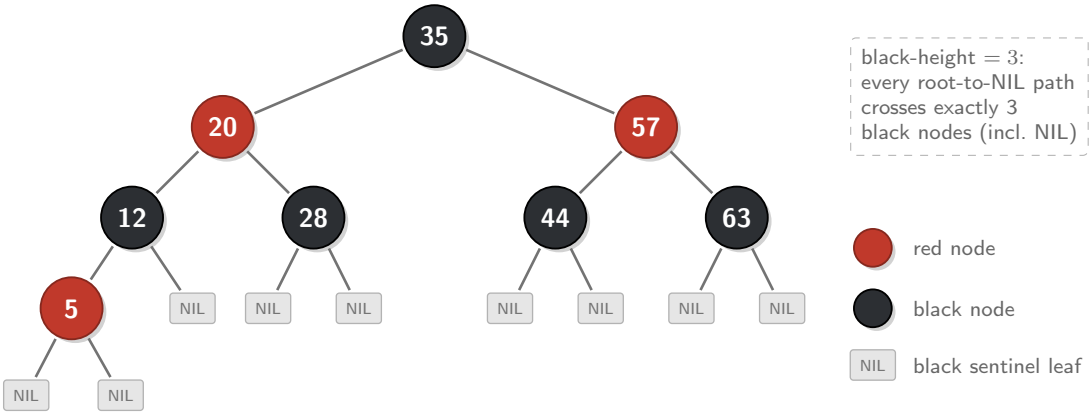


Figure 1: A valid red-black tree over integer keys. No red node has a red child, every root-to-NIL path crosses the same number of black nodes, and in-order traversal yields the keys in sorted order. The path through 5 is the longest and the path to 28's children is among the shortest, but the two differ in length by less than a factor of two.

Balance is maintained, not assumed. An insertion places a new red node at a leaf position, which may create a red-red violation; a deletion may remove a black node, which breaks the black-height rule. Both are repaired by walking toward the root with two local tools. *Recoloring* flips colors and moves no pointers. *Rotation* (Figure 2) is a constant-time pointer surgery that lifts a child over its parent while preserving the in-order sequence. Insert fixup needs at most two rotations; delete fixup needs at most three. Deletion is famously the fiddly one, and its case analysis is where both

students and AI agents most often go subtly wrong, which is precisely why this assignment requires all of it.

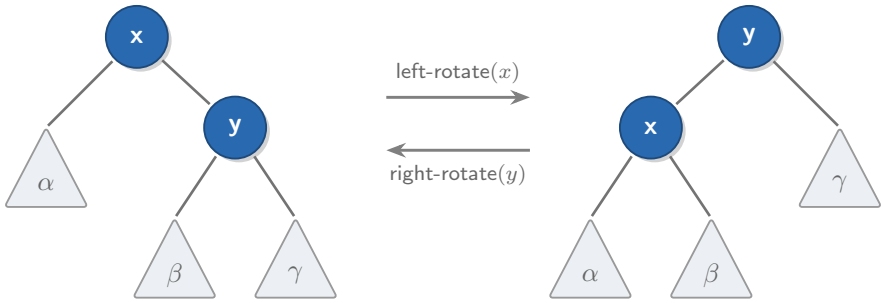


Figure 2: Rotation, the balancing primitive. Three pointers are rewritten in $O(1)$: y rises, x sinks, and subtree β changes parents. The in-order sequence $\alpha, x, \beta, y, \gamma$ is identical on both sides, so search order is preserved. Rotations move no colors; recoloring is a separate step in the fixup procedures.

That is the whole idea. What the definition hides, and what this assignment is built to expose, is that each of those circles is a heap allocation with an owner, a lifetime, and neighbors holding pointers to it. The remainder of this specification is about getting that part right.

5 Assignment Structure

The assignment proceeds through three milestones plus a setup step and an optional extra-credit mutation. Do them in order; each mutation builds on the last. The hour estimates for each element are included in the table below.

Milestone	Title	Est. hours
M0	Setup: repository, `CLAUDE.md`, toolchain	1
M1	Baseline red-black tree + fuzzer	4–6
M2	Mutations 1 & 2 (fault sweep, then slab pool)	4–6
M3	Mutation 3 ($O(1)$-space teardown), hardening, process artifacts	3–4
The Reach	Mutation 4 (copy-on-write snapshots); ungraded, see Section 10.4	2–4
–	Process artifacts (`CLAUDE.md`, `PROMPTLOG.md`, `REFLECTION.md`), written throughout	1–2

Each mutation targets a property the tree already has:

Property of the tree	Mutation	OS analog
`rb_insert` performs several dependent allocations (node, key copy), any of which can fail	M1: fault-injection sweep	unwinding cleanly when `kmalloc` returns `NULL`
nodes are small, uniform, and allocated in bulk	M2: slab pool	the kernel slab allocator (`kmem_cache`)
height is $O(\log n)$, but teardown still visits all n nodes	M3: $O(1)$-space walks	8–16 KB kernel stacks
an update touches only the $O(\log n)$ path back to the root	M4: COW snapshots	copy-on-write pages after `fork()`

6 Why This Assignment Looks Different

You are *required* to use Claude Code for this assignment, and that requirement shaped its entire design. Modern AI coding agents can produce a working red-black tree in one shot. If that were the whole assignment, you would learn nothing. This assignment has two deliberately intertwined goals:

1. **Systems goal:** Build a red-black tree, then survive a series of *mutations* that stress memory management the way an OS kernel stresses it: custom allocators, tiny stacks, allocation failure at arbitrary points, and copy-on-write structural sharing. A correct tree that leaks, double-frees, or recurses unboundedly **fails**, even if every lookup returns the right answer.
2. **Process-Engineering-with-AI goal:** Learn to direct an AI agent the way a senior engineer directs a capable but overconfident junior: give it context, make it plan, keep changes small, verify everything with tools, and never merge code you can't explain.

The anti-goal: pasting this assignment into Claude and typing "solve this." That approach will actually *hurt* you here, for three reasons: (a) the fault-injection and grading harness punishes code nobody reviewed, (b) a large fraction of your grade comes from your process artifacts and a live walkthrough, and (c) you will be asked to explain arbitrary lines of your submission (*live*), and to extend it on parameters you have never seen.

7 Rules of Engagement for AI Use

- **Allowed and expected:** using Claude Code to explore, plan, implement, test, debug, and review; asking it to explain any concept; having it write test scaffolding and Makefiles.
- **Required:** the process deliverables in Section 15 (`CLAUDE.md`, `PROMPTLOG.md`, `REFLECTION.md`), the raw session transcripts, and a git history that shows incremental work (≥ 10 meaningful commits; a single "final submission" commit is an automatic process-grade of zero).

- **Forbidden:** submitting code you cannot explain. The live walkthrough (Section 14) exists to check exactly this. "Claude wrote it" is not an explanation; "this loop restructures the tree into a right-spine so teardown needs $O(1)$ space, and here's why that can't loop forever" is.
- **You own the bugs:** If the agent introduces a use-after-free and you commit it, it is *your* use-after-free.

8 Part-0: Setup (Milestone 0)

8.1 Repository layout

Create (or clone the provided) skeleton:

```
rbtree-lab/
|-- CLAUDE.md              # you will write this in the workflow section
|-- Makefile               # Appendix B
|-- include/
|   `-- rbtree.h           # fixed public API (Appendix A). DO NOT MODIFY
|-- src/
|   |-- rbtree.c           # your implementation
|   `-- pool.c             # Mutation 2
|-- tests/
|   |-- test_rbtree.c      # your unit tests
|   |-- fault_alloc.c      # fault-injecting allocator (Mutation 1)
|   |-- fault_alloc.h
|   `-- fuzz.c             # randomized stress driver
`-- PROMPTLOG.md           # annotated AI-interaction log (see Deliverables)
```

8.2 Install Claude Code

macOS / Linux / WSL:

```
curl -fsSL https://claude.ai/install.sh | bash
```

Windows PowerShell:

```
irm https://claude.ai/install.ps1 | iex
```

(Homebrew users: `brew install --cask claude-code`.) Then, from inside `rbtree-lab/`, run `claude` and follow the login prompt. Full instructions: `https://code.claude.com/docs/en/quickstart`.

8.3 Toolchain rules

Everything must compile with `-std=c23 -Wall -Wextra -Werror`. Two verification builds are required and graded, and both lean on tools you may not have met before.

AddressSanitizer (ASan, switched on with `-fsanitize=address`) is instrumentation the compiler weaves directly into your binary. It watches every memory access as the program runs and halts the instant one goes wrong: a read past the end of a heap block, a use of memory after it was freed, a double free. It then prints the offending source line together with a backtrace of where that memory was originally allocated. Its companion *UBSan* (UndefinedBehaviorSanitizer, `-fsanitize=undefined`) catches the C standard's undefined behavior in the act: signed overflow, out-of-range shifts, misaligned or null pointer dereferences. Because both are compiled into the program, they slow it down. In exchange, they pin each bug to the exact instruction that commits it.

valgrind takes the opposite approach. It is a separate program that runs your *unmodified* binary on a synthetic CPU, interposing on every allocation and memory access from the outside, so it needs no special compile flags and no recompilation. Its `memcheck` tool finds the same families of fault (leaks, invalid reads and writes, and reads of uninitialized memory) at the cost of running many times slower than native. The two tools overlap but are not redundant: ASan is faster and sharper about the access that goes wrong, while valgrind needs no rebuild and catches some uninitialized-value reads that ASan lets through. Requiring both to come back clean is deliberate belt-and-suspenders.

- `make asan`: AddressSanitizer + UBSan build, runs the test suite.
- `make memcheck`: valgrind `--leak-check=full --error-exitcode=1` over the test suite.

A single leaked byte, invalid read/write, or UB report in either target caps the correctness portion of that milestone at 50%. This is the assignment's central discipline: memory bugs are not warnings. They are failures.

9 Part-1: The Baseline Tree (Milestone 1)

Implement a classic red-black tree behind the fixed API in `include/rbtree.h` (see Appendix A). Keys are C strings; the tree **copies** keys on insert and **owns** the copies. Values are opaque `void *` pointers; the tree takes ownership of a value on successful insert and releases it (via the destructor supplied to `rb_create`) on delete/destroy/overwrite.

Required operations: `rb_create`, `rb_insert` (with rebalancing), `rb_find`, `rb_delete` (with the full deletion-fixup cases; yes, all of them), `rb_size`, `rb_foreach` (in-order), `rb_validate`, `rb_destroy`.

`rb_validate` must check, and your tests must exercise, all invariants (Section 4, Figure 1):

1. The root is black.
2. No red node has a red child.
3. Every root-to-NIL path contains the same number b of black nodes (the tree's *black-height*); so the tree height is $O(\log n)$.
4. In-order traversal yields strictly increasing keys: $k_1 < k_2 < \cdots < k_n$ under `strcmp`.
5. `rb_size` matches the actual node count n.

Testing bar: your fuzzer (`tests/fuzz.c`) performs $\geq 10^5$ random insert/find/delete operations against a reference model (a sorted array or a simple linked list is fine), calling `rb_validate` at least every 100 operations, under both `asan` and `memcheck`.

10 Part-2: The Mutations (Milestones 2–3)

Up to this point the tree has lived in the most forgiving world imaginable: every node it asked for, it received, and every node it finished with, it was free to forget. That forgetting is the luxury this part takes away. Three mutations follow, and not one of them changes what the tree computes; each changes only the ground it stands on. To keep your footing you will need three ideas the baseline never forced you to meet.

The first is the *memory leak*. When you ask the allocator for a block, it hands you the bytes and remembers nothing else about them; keeping track of that block is now entirely your problem. Free it and it returns to circulation. Lose the last pointer to it without freeing, and the block is stranded: still allocated, forever unreachable, dead weight until the process exits. In a program that runs for a second, such a leak is a tree falling in an empty forest. In one that runs for months, which is to say a kernel, a database, or a web server, leaks accumulate, and accumulation is fatal. The cruelty is that they gather on precisely the paths you are least likely to test: the error paths, where one allocation failed halfway through a chain of them, you returned early, and you quietly forgot to release what you had already grabbed. Mutation 1 (Section 10.1) exists to walk down every one of those paths and check.

The second idea is that asking for memory one node at a time is not free. Every call to `malloc` must find a suitable block, record its size, align it, and stitch its own bookkeeping around it. Do that a million times for a million identical little nodes and the bookkeeping starts to rival the payload! While the nodes themselves scatter across the heap and your cache spends the day chasing pointers. When every object is the same size, you can decline to play that game. Grab memory wholesale in big page-sized chunks, slice each chunk into node-sized slots, and hand them out yourself; when a node dies, thread its corpse onto a free list so the next request can reuse it without troubling `malloc` at all. That is a *pool allocator*, and it turns both allocation and freeing into an $O(1)$ pointer shuffle.

The third idea is just a name for the wholesale chunk: a *slab*. The pattern has a distinguished pedigree, because the Linux kernel allocates staggering numbers of identical small objects, one structure per open file, per running process, per network packet, and it manages them with exactly this machinery, the slab allocator that lives behind `kmem_cache`. Mutation 2 (Section 10.2) asks you to build a small, honest version of the same thing. Do all three in order; each builds on the last. And notice, at the end, that the tree never noticed: same keys, same invariants, same answers, and an entirely different story underneath about where its nodes come from and where they go to die. That story is, not coincidentally, the whole job description of a memory manager.

10.1 Mutation 1: "Every path leaks nothing" (allocation-failure injection)

Route **all** allocation in `src/rbtree.c` through `rb_malloc`/`rb_free` wrappers (declared in `tests/fault_alloc.h`, trivially forwarding to `malloc`/`free` in production). The test build

interposes a fault-injecting version with this contract:

```
#include <stddef.h>

/* Allocation wrappers. In production these forward to malloc/free;
 * the test build swaps in the fault injector below. ALL allocation
 * in src/rbtree.c (and src/pool.c) must go through these. */
void *rb_malloc(size_t n);
void  rb_free(void *p);

/* the n-th allocation from now returns NULL */
void  fault_alloc_arm(long n);
void  fault_alloc_disarm(void);
/* allocations observed so far */
long  fault_alloc_total(void);
```

Your harness must run the **fault sweep**: for $n = 1, 2, 3, \ldots$ re-run a fixed scenario (e.g., insert 200 keys, delete 100, overwrite 50) with the n-th allocation failing, until a run completes without hitting the fault. After *every* run, whether it failed midway or not:

- `rb_validate` must still pass: a failed insert must leave the tree exactly as it was;
- `rb_destroy` must free everything; ASan/valgrind must report **zero** leaks across the entire sweep;
- every API call that hit the failure must have returned its documented error code.

Why this is hard: `rb_insert` allocates up to three times (node, key copy, and possibly pool metadata later). Failure *between* allocations is where real systems leak. This is the same discipline as kernel code paths that must unwind cleanly when `kmalloc` returns `NULL`.

A guarantee about the sweep. A fixed scenario performs a bounded number of allocations, so the fault sweep is guaranteed to terminate: some finite n is larger than every allocation the scenario makes, and the run at that n completes without ever hitting the injected fault. The injected failures are deterministic for a given scenario and n. If your sweep appears not to terminate, or if the same n gives different results on different runs, treat that as a bug in your own code, most often a hidden dependence on uninitialized memory or on allocation addresses, not as a sign that the harness is unfair.

10.2 Mutation 2: "Roll your own slab" (pool allocator)

Nodes may no longer come from per-node `malloc`. Implement a slab-style pool in `src/pool.c`:

```
typedef struct rb_pool rb_pool_t;
/* slabs are 4096 bytes */
rb_pool_t *pool_create(size_t obj_size);
/* O(1), from a free list */
void      *pool_alloc(rb_pool_t *p);
```

```
/* O(1), returns to the free list */
void       pool_free(rb_pool_t *p, void *obj);
void       pool_stats(const rb_pool_t *p,
                      size_t *slabs, size_t *live, size_t *free_objs);
/* releases all slabs */
void       pool_destroy(rb_pool_t *p);
```

Constraints: slabs are page-sized ($4096 = 2^{12}$ bytes) chunks obtained with `rb_malloc`; freed objects are chained into an intrusive free list *inside the dead objects themselves* (no side table); `pool_alloc`/`pool_free` are $O(1)$; `pool_destroy` frees every slab regardless of how many objects are live. The tree gains `rb_create_pooled(...)` and must pass the *entire* Milestone 1 + Mutation 1 test battery (including the fault sweep, since slab allocation can fail too) on the pooled build.

OS connection: this is a miniature of the kernel slab allocator (`kmem_cache_alloc`). Write one paragraph in `REFLECTION.md` comparing your design to it: what does the kernel's version add, and why?

Watch out: reusing a freed node's memory hides use-after-free bugs from ASan (the memory is still "valid" to the sanitizer). Add **poisoning** (`memset` freed objects to `0xDD` in debug builds) and explain in your reflection what class of bug this catches that ASan now cannot.

10.3 Mutation 3: "No stack to spare" ($O(1)$-space teardown and traversal)

Before the requirement, a word about the thing you are about to run out of.

Every running function needs somewhere to keep its private business: its parameters, its local variables, the address it should jump back to when it's done, and whatever registers its caller is trusting it not to clobber. That bundle is a *stack frame*. It lives on the *stack*, a region of memory handed to your thread when the thread is created, which grows as calls nest and shrinks as they return.

The nesting is the whole point. Say `x()` calls `y()`, and `y()` calls `z()`. At the instant `z` begins executing, all three frames are on the stack at once, piled in call order: `x` at the bottom, `y` on top of it, `z` on top of that. Nothing of `x` has gone anywhere. Its local variables (or locals, for short) are just sitting there, buried but alive, waiting patiently for the frames above them to clear out. When `z` returns, its frame pops and `y` resumes with everything exactly where it left it; when `y` returns, same again for `x`. Frames come off in exactly the reverse of the order they went on, which is the entire reason this thing is called a stack and not something more imaginative.

Now here's the catch, and it's a big one. That region is finite. Its size gets fixed the moment the thread is born, and nobody grows it for you on demand, not even a little. Each nested call eats another few dozen bytes of it, and a function that calls *itself* is really just an especially enthusiastic version of `x` calling `y` calling `z`: recursion of depth d means d frames stacked up and alive at once, all of them leaning on the innermost one to finish first. A recursive tree teardown recurses once per level, so its peak stack usage tracks the height of the tree, a number you don't control and, mid-teardown, can't easily bound.

Do the arithmetic and it gets uncomfortable fast. Give a frame a modest 48 bytes, which is optimistic for anything doing real work, and a 64 KB stack buys you somewhere north of a thousand

nested calls before you are out of room. A kernel thread, which is what you should be imagining throughout this course, gets roughly 8–16 KB in total. That is a couple of hundred frames for *everything*, your recursion included, and you are sharing it with whatever called you.

What happens when you run off the end is worth knowing, because the two answers are very different. In userspace on Linux, a guard page sits below the stack and the write into it kills your process with a **segfault**. That is the *good* outcome. You get a core dump, a backtrace, and an afternoon. Inside a kernel, and on older kernels without guard pages especially, you get no such courtesy: the stack simply keeps growing into whatever memory was unlucky enough to be adjacent, silently corrupting it, and the resulting bug surfaces somewhere else entirely, in a different subsystem, an hour later, wearing a disguise. It's a genuinely miserable class of bug to chase and a large part of why kernel code avoids unbounded recursion as a matter of policy rather than taste.

The obvious dodge: allocating your own explicit stack on the heap and pushing node pointers onto it, is *not* available to you here either. It relocates the problem without solving it. It makes teardown cost $O(n)$ extra memory at exactly the moment you are trying to give memory back, and (per Mutation 1) that allocation can fail, which leaves you needing to free a tree with no way to walk it. Destruction paths that can fail are their own special kind of unpleasant.

So: no recursion, no heap, no borrowing. Reimplement:

- `rb_destroy`: no recursion, no heap-allocated stack/queue, $O(1)$ auxiliary space. (Classic technique: rotate the tree into a right spine as you free it, or use pointer reversal.)
- `rb_foreach`: same constraints. Morris traversal or an explicit-parent-pointer walk both work; if you add parent pointers, account for the memory cost in `pool_stats` and your reflection.

Both techniques lean on one idea worth appreciating before you write a line of code: the tree already has plenty of pointers lying around, and most of them are pointing at nodes you're finished with. Instead of recording where you've been in some separate structure, you borrow the tree's own pointer fields to remember it, then put them back (or free them) on the way out. Your bookkeeping lives inside the data itself, which is exactly why the auxiliary space is a couple of local variables and nothing more. It's a lovely trick. It's also precisely the sort of code whose loop invariant you'll be asked to justify at the walkthrough, out loud, without hedging, without reaching for your notes. Know what it maintains. Know why it can't spin forever.

Your test must build a tree of $\geq 10^6$ nodes (the pooled build makes this fast) and then destroy it under a deliberately small stack. Run the teardown inside a `pthread` created with a 64 KB stack (`pthread_attr_setstacksize`). Confirm no crash. Then confirm zero leaks.

10.4 Mutation 4: "`fork()` your tree" (The Reach: COW snapshots)

This mutation carries no points. Do it anyway. It is the part of the assignment that most resembles real kernel work, and the one you will be proudest to have gotten right. Nobody is checking whether you finished it, which is exactly what makes finishing it mean something. If the first three mutations taught you to keep a tree alive under pressure, this one asks whether you can make two trees share a life without either one corrupting the other.

Copying is expensive, and most copies are never even scribbled on. That single observation is the whole idea behind *copy-on-write*, one of the great lazy triumphs of systems design. When you

ask for a snapshot of the tree, the honest but foolish reply is to duplicate every node; the clever reply is to hand back a second doorway into the very same nodes and to promise the real copying for later, if and only if someone actually tries to change something. As long as the original and the snapshot only ever read, that promise never comes due, and one set of nodes quietly serves two masters.

The bill arrives on the first write. You cannot scribble on a shared node without the other handle seeing it, so before changing a node you copy it, and because its parent pointed at the old version you must copy the parent too, and its parent, all the way up to the root. A write clones exactly the $O(\log n)$ nodes on the path back to the root, which is why the technique is called *path copying*; every subtree hanging off that path stays shared, now claimed by two parents at once (Figure 3). That shared claiming is what makes teardown interesting, because a node no longer belongs to one tree and cannot simply be freed when one handle goes away. Enter the per-node *reference count*: raise it when a new handle comes to share a node, lower it when a handle lets go, and free the node only at the instant its count reaches zero. Get the order of those decrements wrong and you have produced either a leak or a use-after-free, which is precisely the bug class that has kept kernel security researchers gainfully employed for decades.

If all of this is starting to sound like `fork()`, that is because it is `fork()`. When a process forks (as we will see in the second week of classes), the kernel does not copy its address space; it marks the pages read-only, shares them between parent and child, and waits. The first write to a shared page faults, the kernel copies that one page, and both processes carry on as though the copy had been there all along, with the reference count living in `struct page`. Mutation 4 asks you to rebuild that mechanism in miniature, on a tree instead of an address space, which is exactly why the reflection below asks you what plays the role of the page fault.

Add:

```
/* O(1): new handle sharing all nodes */
rbtree_t *rb_snapshot(rbtree_t *t);
```

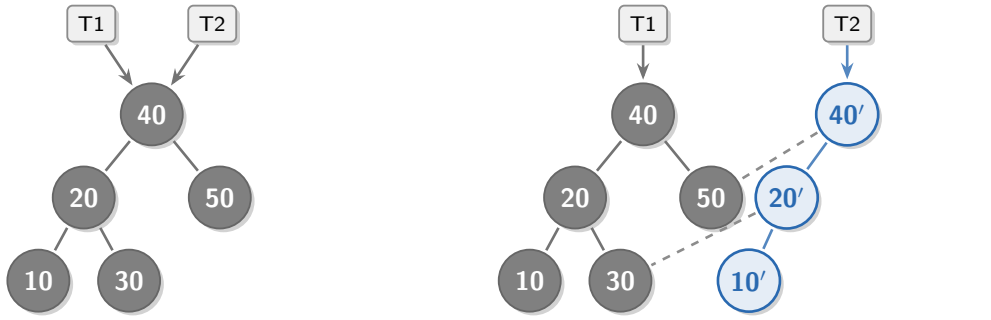
After a snapshot, both handles are fully usable. Writes through either handle must not be visible through the other: copy the modified node's path to the root ("path copying", Figure 3), leaving shared subtrees intact. Manage node lifetime with per-node reference counts; `rb_destroy` on one handle must free exactly the nodes no longer reachable from any live handle.

Required tests: interleaved writes to 3+ snapshots with model checking; the fault sweep over a snapshot-heavy scenario; a leak-free teardown in *every* snapshot destruction order. In `REFLECTION.md`, connect this to copy-on-write pages after `fork()`: what plays the role of the page fault? Of the reference count in `struct page`?

This mutation is where refcount bugs (the classic kernel CVE generator) live. Expect Claude to get it *almost* right on the first try. "Almost" is the lesson.

11 Part-3: Required Use of Claude Code (the workflow)

You are required to use Claude Code for this assignment and to log that usage, but the goal is for you to leave the assignment understanding every line you submit, not just to produce working



(a) after `T2 = rb_snapshot(T1)`: every node shared, refcounts bumped

(b) after a write to key 10 through T2: only the path is copied (blue); 50, 30 stay shared (dashed)

Figure 3: Copy-on-write snapshots by path copying (Mutation 4). The snapshot itself is $O(1)$: the new handle shares every node, and per-node reference counts record the sharing. A later write through T2 copies only the $O(\log n)$ path it touches ($40'$, $20'$, $10'$); subtrees rooted at 50 and 30 remain shared and are now reachable from two parents each. Destroying either handle must free exactly the nodes whose reference counts drop to zero, in any destruction order. Node colors are omitted here; the copied path carries its red-black colors with it.

code by any means available. This section is graded via your transcripts and your `PROMPTLOG.md`. Work through it step by step; each step names the failure it prevents.

Step 1: Give the agent persistent context in `CLAUDE.md`

Claude Code reads a file named `CLAUDE.md` at the start of every session; it is the one place where your rules survive between conversations. In your project root, run `claude`, then type `/init` to have it generate a starter file from your repo, then **edit it** to contain your project law. A working example is in Appendix C; yours must at minimum pin down:

- build/test commands (`make test`, `make asan`, `make memcheck`) and the rule that *nothing is "done" until all three pass*;
- `include/rbtree.h` is frozen and the agent must never edit it;
- all allocation goes through `rb_malloc`/`rb_free`;
- style rules (C23, no VLAs, no `goto` except cleanup-label unwinding, every allocation checked);
- "make the smallest change that passes; do not refactor unrelated code."

Failure this prevents: re-explaining your constraints every session, and the agent "helpfully" rewriting your header.

Step 2: Explore before anything is written

Start a fresh session and ask questions with no edits: "Walk me through the deletion fixup cases and why each rotation preserves black-height." "Where in this codebase would an allocation failure between the node alloc and the key copy leak memory?" You are calibrating two things: your own understanding, and whether the agent's picture of the problem matches yours.

Failure this prevents: discovering at the walkthrough that you never understood deletion fixup.

Step 3: Plan first, in plan mode

For every milestone, press **Shift+Tab** to cycle permission modes until the status line shows **plan mode** (or use `/plan`). In plan mode the agent can read and analyze but not modify anything. Prompt like this:

> *Plan Mutation 1. Read src/rbtree.c and tests/fault_alloc.h first. I want: (1) every allocation site in rb_insert identified, (2) for each, what must be unwound if it fails, (3) the exact error-path code structure you propose (I prefer goto-cleanup), (4) the fault-sweep harness design, (5) which tests you'd write first. List risks. Do not write code yet.*

Then **argue with the plan**. Ask "what breaks if the key allocation succeeds but the node allocation fails?" If the plan hand-waves, say so and make it revise. Only approve when *you* could defend every step. Record at least one plan revision in `PROMPTLOG.md`; a plan you accepted unmodified on the first try is a red flag we will look for.

Failure this prevents: 400 lines of confident code solving the wrong problem.

Step 4: Tests first, then the smallest slice

Out of plan mode, direct implementation in slices, tests leading:

> *Write the failing tests for the fault sweep first: arm fault_alloc at $n = 1 \ldots N$ over the standard scenario, asserting rb_validate passes and error codes are correct after each run. Don't touch rbtree.c yet. Run make test and show me the failures.*

Then: "Now make only the insert path survive the sweep. Smallest diff. Run `make asan` and show me the output." One slice per prompt. If a diff comes back touching five files, reject it: "Too big. Insert path only."

Failure this prevents: the un-reviewable mega-diff, and the agent quietly weakening tests to make them pass (read every test diff, because this is a real failure mode).

Step 5: Close the loop with tools, demand evidence

Never accept "the tests should pass now." Make the agent run the commands and show output. Feed tool output back verbatim; the agent reads a valgrind report better than a vague description of one:

> *make memcheck reports "40 bytes definitely lost, allocated at rbtree.c:212 in rb_insert". Trace the ownership of that allocation through every path out of rb_insert and fix the leak. Then re-run make memcheck and show me a clean report.*

Failure this prevents: the agent asserting success; you graduating without ever reading a valgrind report.

Step 6: Manage the context window

The context is the agent's working memory; a session full of dead ends makes it dumber. Use `/clear` between milestones and after finished sub-tasks; use `/compact` mid-task if a long debugging session gets noisy but you need the thread. Rule of thumb: new milestone, new session. `CLAUDE.md` carries the constraints forward, which is why Step 1 matters.

Failure this prevents: the "it keeps forgetting my instructions" spiral late in a long session.

Step 7: Adversarial review in a fresh context

Before each milestone commit, get a review from a context that did not write the code, because a reviewer that shares the author's assumptions inherits its blind spots. Either run the bundled `/code-review` skill, or `/clear` and prompt:

> *Review the diff for Mutation 2 as a hostile kernel reviewer. Hunt specifically for: use-after-free enabled by pool reuse, off-by-one in slab carving, alignment bugs for the intrusive free-list pointer, paths where pool_ destroy is called with live objects. Report findings with file:line; do not fix anything.*

Triage the findings yourself and decide which are real. Log one real finding and one false positive in `PROMPTLOG.md` with your reasoning.

Failure this prevents: the author grading its own homework.

Step 8: Git discipline, conversationally

Commit at every green state: "run the full battery; if clean, commit as 'M2: pool allocator survives fault sweep'." Small commits are your undo button when a later "fix" regresses the sweep, and they are also your process grade.

Step 9 (optional): Automate your loop

Create a project slash command, a file at `.claude/commands/battery.md`, so the whole battery is one keystroke:

```
Run make test, make asan, and make memcheck in sequence.
If any fail, stop and diagnose the first failure: $ARGUMENTS
Show all output. Do not modify tests to make them pass.
```

Then `/battery` runs it in any session. Docs for going further (hooks that block edits to `rbtree.h`, subagents): `https://code.claude.com/docs/en/common-workflows`.

11.1 Prompt patterns: good vs. useless

Useless	Effective
"solve the assignment"	"Plan Mutation 1 per CLAUDE.md; identify every allocation site in `rb_insert` and what must unwind on failure. No code yet."
"fix the bug"	"The fault sweep fails at $n = 7$: `rb_validate` reports a black-height violation after a failed insert of key 'm'. Reproduce with the seed in `tests/fuzz.c` line 14, find the state the failed insert leaves behind, and propose a fix."
"make it not leak"	"Here is the full valgrind output: [paste]. Explain the ownership of the 40 bytes at `rbtree.c:212` on the early-return path at line 230, then fix only that path."
"write tests"	"Write table-driven tests for `rb_delete` covering: red leaf, black leaf with red sibling, root deletion, both children present. Each asserts `rb_validate` and `rb_size` afterward."
"is this right?"	"State the invariant this loop maintains at the top of each iteration, and show why termination follows. If you can't, the code is wrong; say so."

Transcript submission. You must submit the raw session transcript(s) Claude Code already saves for you, not a hand-written or `/export`'d summary. By default these live locally as JSONL files at `~/.claude/projects/<project> /<session-id>.jsonl`, one file per session, timestamped and recorded tool call by tool call. Copy the `.jsonl` file(s) for this assignment's project directory into your submission as-is; do not edit, reformat, retype, or hand-transcribe them. Work on this assignment in its own dedicated project directory, both so this copy step is a single `cp`, and so the file doesn't contain unrelated conversations from other work. Claude Code purges local session transcripts automatically after 30 days by default; if you start early, copy them out periodically or raise `cleanupPeriodDays` in your settings. Losing early transcripts to this default is not a valid excuse for an incomplete submission, so plan around it.

12 Helpful Claude Code Resources

You will not need most of Claude Code's documentation for this assignment, and chasing all of it is its own way of not writing the tree. What follows is the short path, ordered the way you will actually meet each idea. Read the first two before you write a line; keep the middle three open while you work; reach for the last two when the basics have gone automatic. Each maps to a step in the workflow of Section 11, so if a step there feels underspecified, the matching link is where it is spelled out in full.

- **Quickstart**: install Claude Code, log in, and run your first session. Start here, once.
  `https://code.claude.com/docs/en/quickstart`

- **How Claude remembers your project**: what `CLAUDE.md` is and how it carries your project's rules across sessions. This is Step 1, and it is the difference between correcting the agent once and correcting it every morning.
  `https://code.claude.com/docs/en/memory`
- **Choose a permission mode**: what **plan mode** is, and how **Shift+Tab** cycles between reading, planning, and editing. This is the single habit (Step 3) that separates directing the agent from being surprised by it.
  `https://code.claude.com/docs/en/permission-modes`
- **Best practices**: the house rules for getting good work out of the agent: specific prompts, small diffs, evidence over assertion. Read once early, skim again when a session starts going in circles.
  `https://code.claude.com/docs/en/best-practices`
- **Common workflows**: worked recipes for exploring a codebase, fixing a bug, and writing tests, which is most of what Steps 4 and 5 ask of you. Also your pointer to going further with hooks and subagents.
  `https://code.claude.com/docs/en/common-workflows`
- **Commands**: the reference for the session commands this spec uses: `/clear` and `/compact` for managing context (Step 6), `/plan`, and defining your own slash commands like the `/battery` loop in Step 9.
  `https://code.claude.com/docs/en/commands`
- **Code Review**: how the `/code-review` skill runs an adversarial pass over your diff from a fresh context, which is exactly the hostile-reviewer move in Step 7.
  `https://code.claude.com/docs/en/code-review`

Two standing tips that outlast any link: inside a session, `/help` lists everything available right now, and asking Claude Code "how do I . . ." in plain language is often faster than finding the page. The documentation moves quickly, so if a link has drifted, the index at `https://code.claude.com/docs` will have its new home.

13 Using Your Claude Code Allowance Well

This course requires a paid Claude subscription, currently $20/month for the Pro plan; the free tier will not carry you through this assignment or the ones that follow it this semester. A subscription is not bottomless, though. Your allowance is metered on two clocks at once: (1) a session window that resets 5 hours after your first message, and (2) a weekly cap that resets at a fixed day and time assigned to your account. Both are shared across everything you do with Claude, so the evening you spend having it adjudicate an argument about tabs versus spaces is drawn from the same purse as the fault sweep. Neither clock is generous enough to be ignored. Neither is tight enough to be a problem if you work deliberately. The difference between those two outcomes is almost entirely a matter of habit, which is the good news, because the habits that conserve your allowance are the same ones this assignment already demands of you for entirely unrelated reasons.

Here is the mechanism behind that happy coincidence. Every message you send carries your whole conversation with it, because the agent has no memory between turns and must be handed the *context* afresh each time. So cost scales with context, not with the number of questions you ask, and a long session grows more expensive per message the longer it runs, whether or not the accumulated history is still relevant. A conversation that has wandered through three dead ends is paying rent on all three ... indefinitely, like a storage unit full of furniture you no longer own.

Consider the anti-goal named earlier in this assignment: pasting the whole specification into a session and typing "solve this." This document is not short. You pay for it once when you paste it, and then again on every subsequent message for the remainder of that session, which is a remarkable sum to spend on an answer that Sections 14 and 17 exist specifically to catch. It is the only prompt in this course that manages to be expensive twice.

That is also why Section 11 tells you to `/clear` between milestones and `/compact` when a debugging thread gets noisy: those instructions were written to keep the agent sharp, and they happen to be the single most effective thing you can do to keep your allowance intact. Nearly every other rule there pays the same double dividend. Plan mode is cheap precisely because it prevents the expensive outcome: four hundred lines of confident, beautifully formatted code implementing the deletion fixup for a tree whose keys are integers rather than the C strings in the frozen header. You then pay a second time to have it undone. Specific prompts are cheap because "add the goto-cleanup unwind to `rb_insert`" reads one function, whereas "improve this code" reads `src/`, then `tests/`, then the Makefile, and eventually `PROMPTLOG.md`, where it will encounter your earlier remarks about it. Small diffs are cheap because you review them once instead of three times. You were going to do all of this anyway; you may as well know that it pays twice.

Habits that matter most here

- **One milestone, one session:** Start fresh with `/clear` when you move from the baseline tree to the fault sweep, or from the pool to the teardown. Stale context is charged on every subsequent message, and the pool allocator gains nothing from remembering the afternoon you spent certain that `rb_validate` was miscounting black-height when in fact you were counting the NIL sentinel twice. Use `/rename` before clearing and `/resume` to return to a session you may want again.
- **Compact rather than continue:** When a long debugging thread is finished but the task is not, `/compact` summarizes what came before. You can steer what survives: `/compact keep the failing test output and the current diff`, not `/compact keep everything`, which is the command for people who have made peace with their weekly cap.
- **Match the model to the job:** Sonnet handles ordinary implementation work well and costs less. Save the heavier model for the genuinely hard reasoning in this assignment, which is roughly two things: arguing through the deletion-fixup cases, and the refcount decrement order in Mutation 4. It is not required for a `pool_stats` function that returns three numbers through three pointers. Switch with `/model`.
- **Filter tool output before the agent sees it:** This assignment generates spectacular quantities of log. The fault sweep prints a result per run. A 10^6-node teardown will print one million cheerful

lines if you let it, and it will let you. A valgrind report can run to hundreds of lines of which four matter. Step 5 asks you to paste tool output verbatim, and you should, but paste the twenty lines containing the error, not the two thousand containing the word `PASS`. `grep`, `head`, and `tail` are free; context is not. `make memcheck 2>&1 | grep -A6 'definitely lost'` costs you nothing and tells the agent everything.

- **Stop early when it goes wrong:** Press Escape the moment the agent heads somewhere you did not intend, rather than letting a bad edit finish and paying a second time to unwind it. Escape is the cheapest key on your keyboard, and it gets cheaper the sooner you press it.
- **Keep `CLAUDE.md` lean:** It loads at the start of every session, so every line is a standing charge. The example in Appendix C is about the right size. If yours has grown a subsection on your preferred brace style, you are paying a small fee to relitigate brace style every single morning for three weeks.

Watch the meter

You can't manage what you can't see, and both of the commands that show you are free.

Command	What it tells you
`/usage`	How much of your plan's session and weekly allowance you have spent, with a breakdown of what consumed it. Press `d` or `w` to switch between the last day and the last week.
`/context`	What is occupying your context window right now, which is the thing you are paying for on every message.

Check `/usage` once at the start of a work session and once when something feels slow. If you would rather not think about it at all, configure the status line to display context usage continuously; the goal is to notice a ballooning context before it forces a compaction, not to perform an autopsy afterwards.

Three ways students will burn a week's allowance in an afternoon

1. **The session that never closes:** Opened Friday for Milestone 1, still open Sunday for Milestone 2. Every question you ask on Sunday arrives escorted by Friday's confident and entirely wrong theory that the leak was inside `strcmp`, and you are billed for the escort.
2. **Pointing the agent at everything:** The off-by-one is in your slab-carving loop, 40 lines of `pool.c`. There is no reason for the agent to read the fuzzer, the Makefile, and eleven hundred lines of tests on its way there. Name the file. Name the function.
3. **Convening a committee:** Agent teams spawn multiple instances, each with a context window of its own, and can consume several times the tokens of an ordinary session. Four agents hunting one missing `rb_free` is not a productivity strategy; it is a way of buying the same wrong answer four times, concurrently!! They are off by default. For this assignment, leave them off.

A last word, and it is the least technical of all: start early. The 5-hour window is a coffee break, but the weekly cap is a week, and it does not care that your deadline is Thursday. A student who works three unhurried evenings will finish comfortably inside the allowance. A student who begins the night before will meet both clocks at once, and only one of them can be waited out. The weekly cap is also the only entity associated with this course that has no discretion, no sympathy, and no email address.

Figures and mechanics here were current when this assignment was written and do change; `/usage` *in your own session is always the final authority. For details, see* `https://code.claude.com/docs/en/costs` *and* `https://support.claude.com/en/articles/8325606-what-is-the-pro-plan`.

14 Personalized Verification

Passing every test battery does not end your obligation on this assignment. Every student, not a flagged subset, sits for a short mandatory live walkthrough after the deadline. It is not a bonus check triggered by a mismatch between your log and your code; it happens for everyone.

At the walkthrough you will be given a small set of parameters specific to you and one small, previously unseen modification to make to your own already-submitted code, for example adding a new command-line flag to the test driver, handling a key-length range you did not originally test, or tightening a `pool_stats` figure. You will make this change live, starting from your own submitted source and nothing else. We will also pick lines of *your* code ("why is this node recolored?", "what frees this on the error path?", "why can't this teardown loop forever?") and one episode from your `PROMPTLOG.md` and ask you to defend your judgment call. This is the concrete form of the 2-point quiz described in the grading breakdown.

Why this exists. Every algorithm in here is specified precisely enough to autograde by test battery, sanitizer, and fault sweep, which is exactly what makes rigorous grading possible. It is also, unavoidably, what makes this document alone reproducible from a single request to an AI coding tool. Personalized Verification exists because a static submission, however complete, can only ever get you to a plausible-looking result. Only your ability to navigate, defend, and extend your own code after the fact, on parameters and modifications neither you nor any AI assistant could have seen in advance, can confirm the work behind it was actually yours.

15 Deliverables

Submit, via Canvas, a tar file that contains the following.

- **Source code:** a `Makefile` (Appendix B) that builds the test and fuzz binaries with `gcc -Wall -Wextra -Werror -std=c23` and no warnings; `src/rbtree.c`, `src/pool.c`, and your tests; the *unmodified* frozen `include/rbtree.h` (Appendix A).
- `CLAUDE.md`, checked in and evolving across the milestones.

- `PROMPTLOG.md`: 6–10 annotated episodes, each with the prompt, what came back, and *your* judgment (accepted, rejected, or pushed back on, and why). Must include one plan you revised (Step 3), one rejected/oversized diff (Step 4), one tool-output debugging loop (Step 5), and one review finding you triaged (Step 7). Raw transcript dumps here score nothing; annotation is the deliverable.
- `REFLECTION.md` (1–2 pages): the slab and (if attempted) COW comparisons from Section 10; where the agent was most and least reliable; one bug it introduced that you caught, and how; your poisoning rationale.
- The raw Claude Code session transcript(s) (`.jsonl` files) for this project's directory, copied in as-is. Not a summary, not a `/export`, not retyped.
- A git history with ≥ 10 meaningful commits tracking the milestones. A single "final submission" commit is an automatic process-grade of zero.

`PROMPTLOG.md` and `REFLECTION.md` carry no separate point weight, but they are reviewed by the TA alongside your transcripts and are used to seed quiz questions; a missing or perfunctory one will cost you where it hurts, on the quiz.

16 Grading

Note that this assignment is worth 0 points. Zero. The description below is to give you a sense of the shape and structure of how assignments in CS370 will be graded.

This assignment is worth **10 points**: 7 points for program correctness, 1 point for an analysis of your Claude Code transcripts, and 2 points for the quiz on your own code described in Sections 11 and 14.

Within the 7 correctness points, the governing discipline is not a grading technicality; it reflects what this course is actually about. A red-black tree that returns the right answer but leaks memory, double-frees, reads freed memory, or recurses without bound has not solved the problem this assignment poses; it has solved an easier one. So correctness is graded in two layers, both settled by tools rather than by opinion.

- **Layer 1: functional correctness.** Your fuzzer agrees with a reference model across $\geq 10^5$ operations; failed allocations unwind to the prior tree state with the documented error code; the pooled build passes the entire Milestone 1 and Mutation 1 battery; a 10^6-node tree tears down under a 64 KB stack. This is checked by running your suite, and by running our scenarios against your tree.
- **Layer 2: the memory-bug cap.** A single leaked byte, invalid read/write, or UB report under `make asan` or `make memcheck` caps the correctness portion of that milestone at 50%. The measurement is mechanical and reproducible; there is no room for disagreement, which is exactly why we grade the hardest part of the assignment this way.

Component	Method	Points
Part 1: baseline tree	Fuzzer ($\geq 10^5$ ops) agrees with reference model under asan + memcheck; `rb_validate` exercises all invariants	2
Mutation 1: fault sweep	Full sweep leak-free; every failed operation unwinds to the prior state with the correct error code	1.5
Mutation 2: pool allocator	$O(1)$ slab pool with intrusive free list passes the entire M1 + Mutation-1 battery on the pooled build; debug poisoning present	1.5
Mutation 3: $O(1)$-space teardown	$\geq 10^6$ nodes destroyed under a 64 KB stack, zero leaks, no recursion or heap auxiliary space	1
Code quality	Compiles clean under `-Wall -Wextra -Werror`; zero ASan/UBSan/valgrind findings; sane structure	1
Program correctness subtotal		γ
Prompt analysis	Manual TA review of your raw Claude Code transcripts (Section 17) for genuine, iterative use rather than wholesale generation, corroborated against your repository	1
Quiz	Short quiz generated from your own submitted code, administered at the live walkthrough (Sections 14 and 11)	2
Total		**10**
The Reach	Mutation 4 (COW snapshots) complete with tests and reflection: attempted for the craft, not for points	–

Autograder preview. Before you submit, conformance looks like: your fuzzer runs green against the reference model under both sanitizer targets; the fault sweep completes leak-free with correct error codes after every run; the pooled build passes the same sweep; a 10^6-node tree tears down under a 64 KB `pthread` stack with no crash and no leaks; and, if attempted, interleaved writes across three or more snapshots match the model with a leak-free teardown in every destruction order. Full public test cases will be posted and made available later.

17 How We Grade Your Claude Code Transcripts

One of the 10 points comes from your Claude Code transcripts. It is worth being precise about what that point measures, because it is easy to misread as a participation check that any submitted log passes. It is not. The point measures whether your transcripts show the fingerprints of someone building and debugging their own program, and that is a claim we can check against your repository rather than take on faith.

The governing principle is simple, and everything below follows from it: **the grade is based on what can be corroborated against the repository, never on the log's own narrative**. A transcript can say anything. It can claim you caught a subtle refcount bug, demanded a test, or rejected a bad design. What earns credit is not the claim but the evidence for it sitting in your git history, your source, and your tool output. A confident story with nothing behind it scores below a modest one that checks out.

17.1 Mechanism

The grader runs headless (`claude -p`) or as a slash command, with your repository checked out alongside the log, not the log alone. It reads your transcripts as a set of claims and then verifies those claims by inspection against the repo. It asks questions of this shape: does the commit a prompt refers to actually exist; does the fix for a diff you rejected in conversation show up later in history; does the table-driven `rb_delete` test you told Claude you wanted actually appear in your test files; does the `valgrind` leak you debugged at, say, `rbtree.c:212` correspond to code that in fact changed around that line? Episodes that cross-check against a commit, a test, or a diff count. Episodes that float free of any repository artifact do not, no matter how well written they are.

This is the same philosophy as the memory-bug cap in Section 8.3 and the live walkthrough in Section 14, applied to your process instead of your output. A plausible-looking artifact that does not actually connect to real work is easy to produce and, deliberately, easy to catch.

17.2 The scale

Your transcripts are scored on the five-level scale below for technical substantiveness. The level determines the point value directly, in even steps of 0.2: level 5 earns the full point, and each level down subtracts two tenths.

Level	Name	What it looks like	Points
5	**Verified engineering dialogue**	Prompts reference concrete artifacts (file and line, invariants, tool output); your pushback demonstrably changed outcomes; every required episode cross-checks against a commit, test, or diff.	1.0
4	**Substantive, partially corroborated**	The same quality of reasoning, but one or two episodes cannot be tied back to repository evidence.	0.8
3	**Concrete but shallow**	Real artifacts are cited, but the judgments are acceptance without technical reasoning, of the "looked good, kept it" variety.	0.6
2	**Generic**	Plausible transcripts with no verifiable anchors; the log could describe any student's project as easily as yours.	0.4
1	**Performative or contradicted**	The narrative conflicts with git history, or episodes appear fabricated.	0.2

17.3 What this means for how you work

You do not earn a high score by narrating impressively. You earn it by working in a way that leaves a verifiable trail: small commits, tests you actually asked for and actually wrote, real error messages pasted from real runs, and design decisions you can point to in the code. If you use Claude Code the way Section 11 describes, as a collaborator you argue with one function and one bug at a time, the corroboration falls out of your normal workflow and you will land at level 4 or 5 without effort. The only way to score low here is to submit a log that has little to do with the code beside it, and that is precisely the case this point exists to catch.

18 Suggested Schedule

- **Week 1:** Milestone 0 + 1: setup, `CLAUDE.md`, baseline tree, fuzzer green under asan/memcheck.
- **Week 2:** Milestone 2: Mutations 1 and 2 (fault sweep, then pool; re-run the sweep on the pooled build).
- **Week 3:** Milestone 3: Mutation 3, hardening, `PROMPTLOG.md`/`REFLECTION.md`, and the Reach (Mutation 4) if you have time. Walkthroughs in section.

19 Late Submission

The late submission policy for CS370 is available from the course website at `https://www.cs.colostate.edu/~cs370`.

20 The Spirit of This Assignment

You are required to submit your own work, and you are required to use Claude Code to help you produce it. Those two requirements are not in conflict. Your own work in the context of this assignment has never meant work built with no help; it means work that you understand well enough to defend and extend, regardless of how much help you had along the way.

There is no reason to submit someone else's work on this assignment. Not as a shortcut. Not under deadline pressure. Whatever a classmate's tree might save you in Part 2, it costs you at the walkthrough, where a quiz built from code you did not write, and a live modification to code you do not understand, is not a shortcut but a trap you built for yourself. If the agent introduces a use-after-free and you commit it without understanding it, it is still your use-after-free to explain.

Do the assignment the way it is designed to be done. Use Claude Code as a real collaborator, not a copy-paste machine: plan first, keep the diffs small, verify everything with tools, and merge nothing you cannot explain. There is nothing left to gain by doing it any other way.

21 Version History

This section will reflect the change history for the assignment. It will list the version number, the date it was released, and the changes that were made to the preceding version. Changes to the first public release are made to clarify the assignment; the spirit or the crux of the assignment will not change.

Version	Date	Comments
1.0	7/21/2026	First public release of the assignment.

A `include/rbtree.h` (frozen)

```
#ifndef RBTREE_H
#define RBTREE_H
#include <stddef.h>

typedef struct rbtree rbtree_t;
typedef void (*rb_value_free_fn)(void *value);

/* Returns NULL on allocation failure.
 * value_free may be NULL (values not owned). */
rbtree_t *rb_create(rb_value_free_fn value_free);

/* Mutation 2: node storage drawn from an internal slab pool. */
rbtree_t *rb_create_pooled(rb_value_free_fn value_free);

/* Copies key (tree owns the copy).
 * Takes ownership of value ONLY on success.
 * Overwriting an existing key frees the old value via value_free.
 * Returns 0 on success, -1 on allocation failure (tree unchanged, value NOT
 * consumed; caller still owns it). */
int rb_insert(rbtree_t *t, const char *key, void *value);

/* Returns the value, or NULL if absent. Tree retains ownership. */
void *rb_find(const rbtree_t *t, const char *key);

/* Removes key; frees the key copy and the value.
 * Returns 0, or -1 if absent. */
int rb_delete(rbtree_t *t, const char *key);

size_t rb_size(const rbtree_t *t);

/* In-order. Mutation 3: O(1) auxiliary space, no recursion. */
void rb_foreach(const rbtree_t *t,
                void (*fn)(const char *key, void *value, void *ctx),
                void *ctx);

/* Returns 0 iff all red-black invariants hold
 * (see the baseline section). */
int rb_validate(const rbtree_t *t);

/* Frees all nodes, key copies, and (if owned) values.
 * Mutation 3: O(1) auxiliary space, no recursion. NULL-safe. */
void rb_destroy(rbtree_t *t);

/* Mutation 4 (The Reach): O(1) copy-on-write snapshot.
 * NULL on alloc failure. */
rbtree_t *rb_snapshot(rbtree_t *t);
```

```
#endif
```

B Starter `Makefile`

```
CC      := gcc
CFLAGS  := -std=c23 -Wall -Wextra -Werror -g -O1 -Iinclude
SRC     := src/rbtree.c src/pool.c
TSRC    := tests/test_rbtree.c tests/fault_alloc.c
BIN     := build/test_rbtree
FUZZBIN := build/fuzz

all: $(BIN) $(FUZZBIN)

$(BIN): $(SRC) $(TSRC) include/rbtree.h
    @mkdir -p build
    $(CC) $(CFLAGS) $(SRC) $(TSRC) -o $@ -lpthread

$(FUZZBIN): $(SRC) tests/fuzz.c tests/fault_alloc.c include/rbtree.h
    @mkdir -p build
    $(CC) $(CFLAGS) $(SRC) tests/fuzz.c tests/fault_alloc.c -o $@ -lpthread

test: $(BIN) $(FUZZBIN)
    ./$(BIN) && ./$(FUZZBIN) 100000

asan: CFLAGS += -fsanitize=address,undefined -fno-omit-frame-pointer
asan: clean test

memcheck: all
    valgrind --leak-check=full --show-leak-kinds=all \
        --error-exitcode=1 ./$(BIN)
    valgrind --leak-check=full --show-leak-kinds=all \
        --error-exitcode=1 ./$(FUZZBIN) 20000

clean:
    rm -rf build

.PHONY: all test asan memcheck clean
```

(Note: `asan` and `memcheck` are separate targets on purpose, because valgrind and ASan do not mix in one binary.)

C Sample **CLAUDE.md**

```
# rbtree-lab: project rules

## Commands
- Build & unit tests: `make test`
- Sanitizers: `make asan`   Valgrind: `make memcheck`
- A change is DONE only when all three pass. Always run them; show output.

## Hard constraints
- NEVER modify include/rbtree.h. It is the graded contract.
- All heap allocation in src/ goes through rb_malloc/rb_free
  (tests/fault_alloc.h). Direct malloc/free in src/ is a defect.
- Any allocation may fail. Every failure path must unwind completely:
  no leaks, tree left exactly as before the call, documented error code.
- NEVER weaken, skip, or delete a test to make the suite pass. If a test
  looks wrong, stop and explain why instead.

## Style
- C23. -Wall -Wextra -Werror must stay clean. No VLAs.
- Error handling: goto-cleanup pattern for multi-allocation functions.
- Prefer the smallest diff that passes. Do not refactor unrelated code.
- Every non-obvious loop gets a one-line invariant comment.

## Workflow
- For any multi-file or algorithmic change: propose a plan and wait for
  approval before editing.
- Commit only from a green state; message format "M<n>: <what>".
```