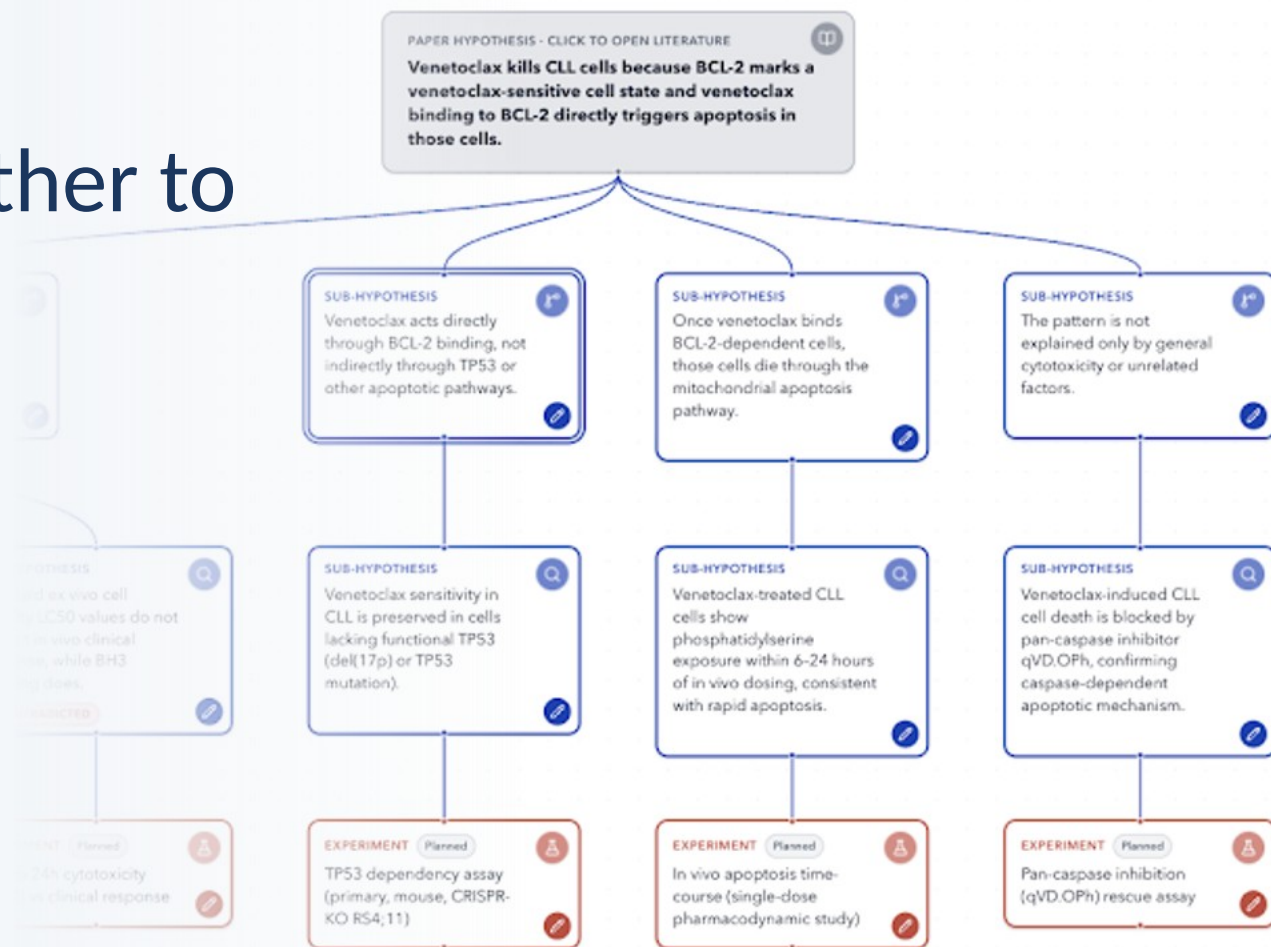


Scientists and AI Working Together to Test a Hypothesis

HypoTree | Internship Project S26

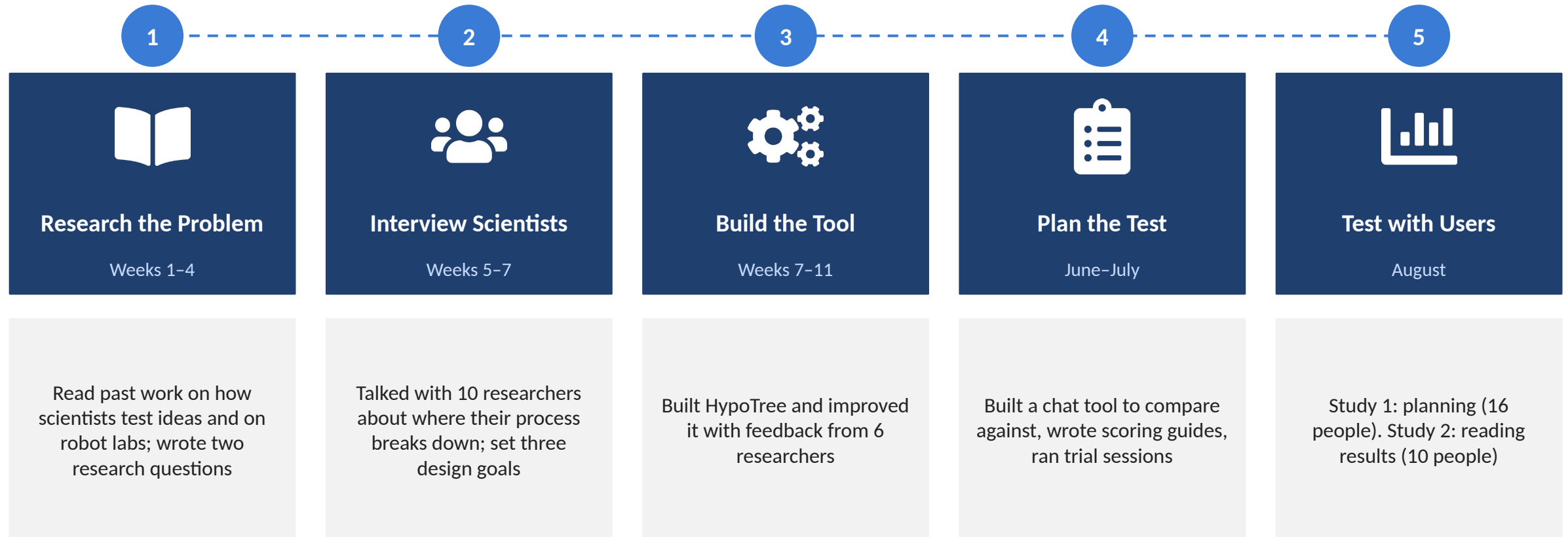
September 2026

Wonjin Kim



Project Roadmap

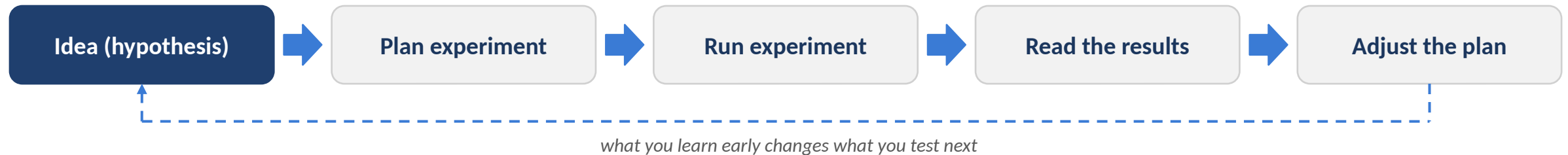
How the project went: learn the problem, build a tool, test it with scientists



Outcome: a working tool, two user studies where scientists clearly preferred it over a regular AI chat, and a research paper draft

Background: Testing an Idea Takes Many Rounds

Scientists rarely prove an idea with one experiment. They run a chain of experiments and adjust as they go



One idea, many experiments

A big idea has several parts, and each part needs its own experiment. Each result decides what to test next.



People and robot labs

People can make mistakes or bring bias. Robot labs ("self-driving labs") run routine steps the same way every time.



Help the scientist, don't replace them

Past research says scientists should stay in charge of the goals and the big decisions.

The gap: few tools help scientists and AI work together **while** an idea is being tested: talking through results, handling unclear outcomes, and deciding the next experiment.

Most AI science tools try to work on their own. Few help the scientist decide what to do next

Co-Scientist	Coscientist	CRISPR-GPT	The AI Scientist	Robin
Several AIs debate to come up with new ideas	AI plans chemistry experiments and runs lab machines	AI plans gene-editing experiments step by step	AI writes code experiments and papers on its own	AI suggests drug ideas and analyzes lab data
Limitation Only a chat; does not break ideas into parts or help discuss results	Limitation Works alone; little back-and-forth with the scientist about results	Limitation Follows a set plan; rarely changes it based on new results	Limitation Computer-only; no real lab experiments	Limitation Scientist only gives the starting goal, then is left out

Tests show AI alone is not enough: AI spots useful links but struggles with long, multi-step research. In one test, AI agents got single steps right 20–35% of the time but finished a full experiment only 0.5% of the time.

Our approach: instead of AI working alone, **HypoTree** keeps the scientist in charge of breaking down the idea, reading results, and choosing the next step.

First, find out what goes wrong today. Then, build a tool and test whether it helps



RQ1

Understand the problem

Where do scientists get stuck when they test an idea over many experiments and have to keep deciding what to do next?

- Interviewed 10 researchers
- Each walked us through a real project of their own
- Result: a list of problems and three design goals



RQ2

Build and test a tool

Does a tree-shaped AI workspace help scientists organize their experiments, understand results, and plan the next step?

- Built HypoTree with feedback from 6 researchers
- Each person tried both HypoTree and a regular AI chat
- Study 1: planning (16 people). Study 2: results (10)



Find problems (RQ1)



Set 3 design goals



Build HypoTree



Test with users (RQ2)

How the project fits together

Interviews: Where Scientists Get Stuck (RQ1)

We talked with 10 researchers in biology and materials science about a real project of theirs

How we did it

Who

10 researchers (PhD students, postdocs, and staff scientists) in biology, bioengineering, biochemistry, and materials science

How

45–60 minute interviews. Each person showed us one real project using their own lab notes, protocols, and planning documents

What we asked

How they organize ideas and experiments, what tools they use, how they read results, and how they pick the next step

What we found



Steps are not connected

The idea, its parts, the experiments, the results, and the next step were rarely tracked in one place



The "why" gets lost

Reasons for choosing an experiment are rarely written down and are forgotten later



Notes are scattered

Plans and results are spread across Word files, slides, spreadsheets, and memory



AI chat only helps one step at a time

ChatGPT answers single questions but can't keep track of a long project, so trust stays low

What they wanted: one place that shows how ideas, experiments, results, and decisions connect, and that remembers why each choice was made.

From Problems to Design Goals

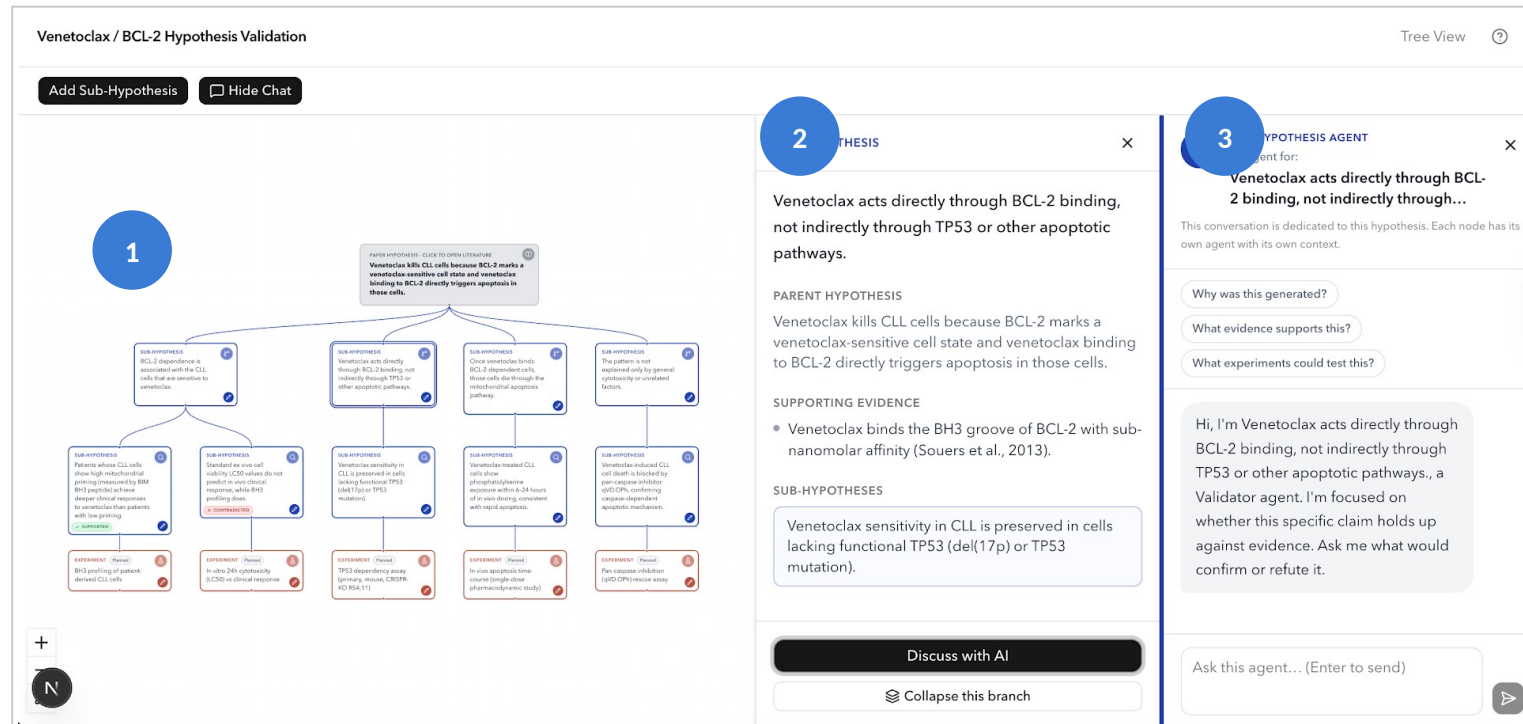
Each problem we found led to one goal for the tool

Problem	Why it hurts	Design goal (DG)
Thinking in branches, writing in lists Scientists think in branches (parts, causes, alternatives) but write in top-to-bottom notes	Hard to keep track They have to hold the whole picture in their head	1 DG1 Show the whole testing process as a tree that grows over time
Reasons are not written down Why an idea was split a certain way, or why an experiment was chosen, lives in papers and memory	Hard to explain later Reasons are lost, so work is hard to share or reuse	2 DG2 Attach the reason, backed by papers, to every link in the tree
Results don't lead to next steps Surprising results are seen as "it failed" instead of a hint about what to change	Hard to decide People stick to one path and miss good alternatives	3 DG3 Help read each result and suggest next steps that fit the project

Core idea: put the plan, the reasons, and the options on screen so scientists don't have to keep them all in their head

HypoTree: What the Tool Looks Like

A tree-shaped workspace where the scientist and AI assistants test an idea together



1 Tree view (DG1)

The main idea at the top, its smaller parts below, and experiments at the bottom. Each box shows if it is planned, supported, or disproved.

2 Detail panel (DG2)

Click any box to see where it came from, the research papers that support it, and what sits below it.

3 AI assistant per box (DG3)

Each box has its own AI assistant that remembers its topic. Quick buttons: "Why was this suggested?", "What supports this?", "How could we test this?"

Add or edit any idea

Fold away ideas that failed

Upload results (data, graphs, notes)

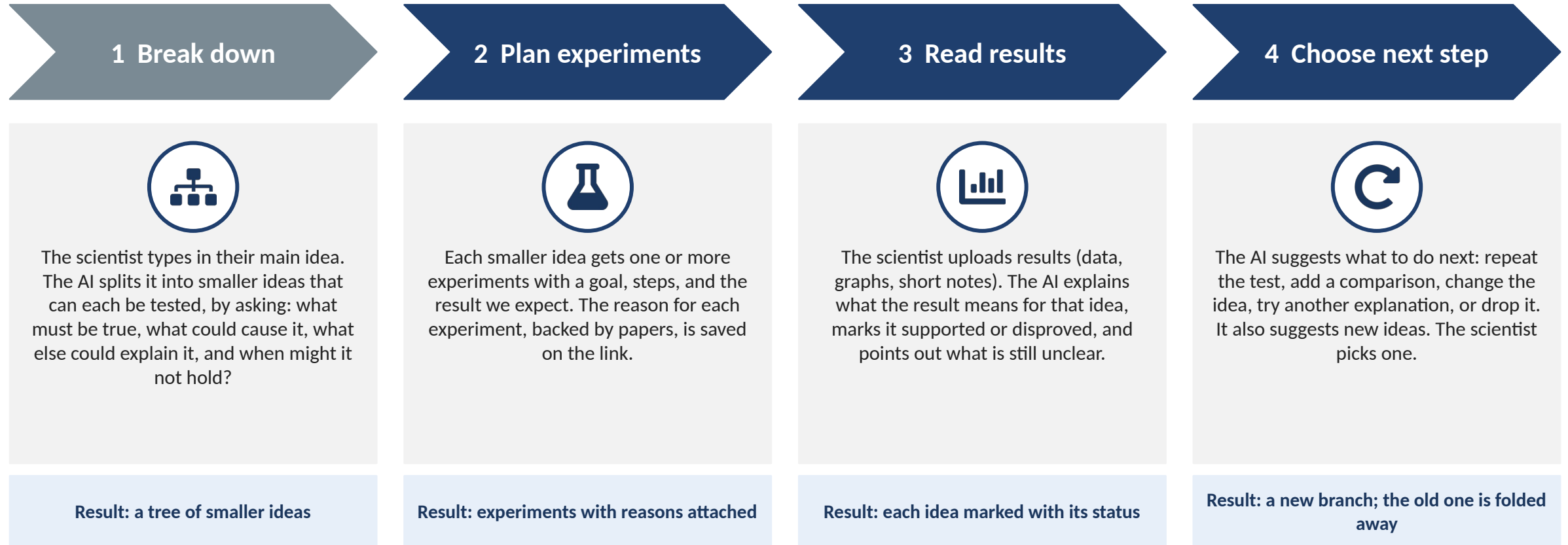
Approve the next step

The scientist stays in control: nothing changes in the tree without their approval

Behind the scenes: AI language models + a research paper search

How HypoTree Works

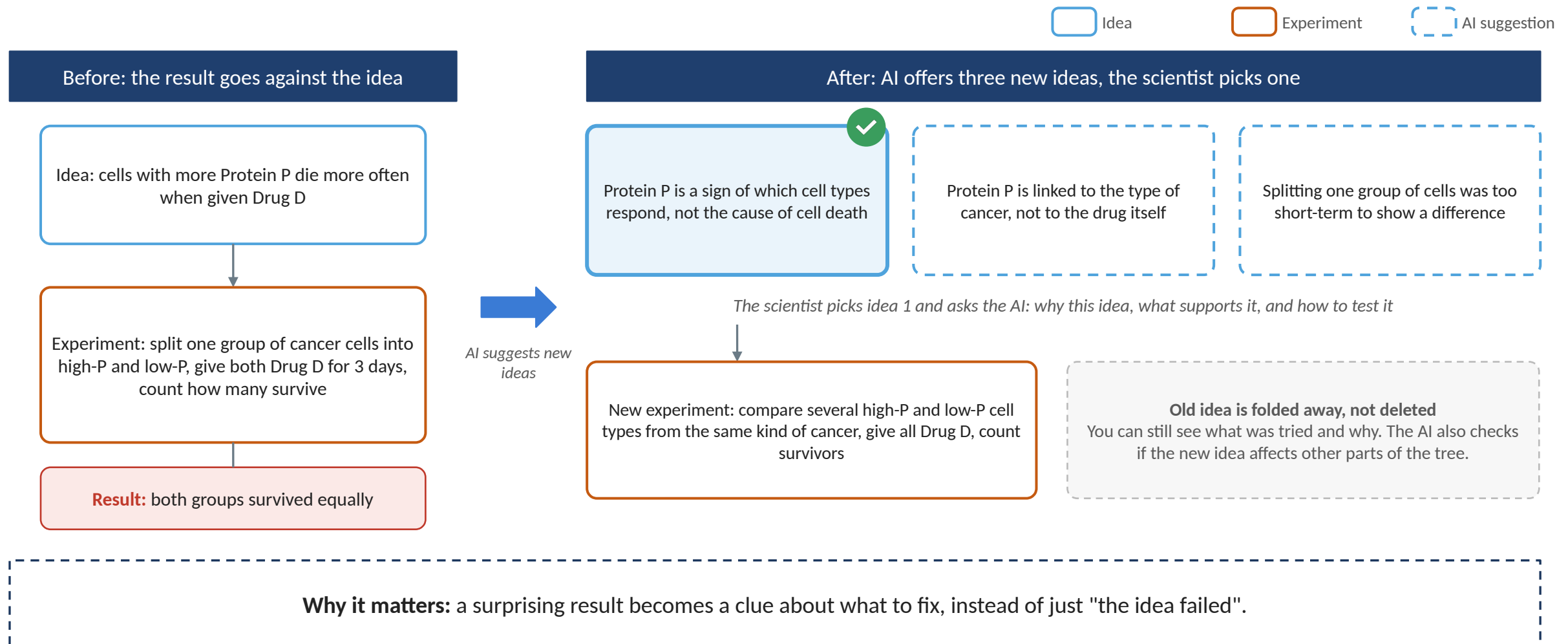
Four steps. At each one, the scientist and AI talk it through instead of the AI giving one answer



The scientist decides at every step: they can edit anything, choose between AI suggestions, and approve each change.

Example: When a Result Is Unexpected

The AI suggests new explanations, and the scientist picks which one to test next



What We Compared Against

A regular AI chat next to a document, using the same AI and the same research papers

Comparison tool: document + AI chat

The screenshot shows a web application titled "Hypothesis Validation Project". On the left is a document with sections: "1. Main hypothesis", "2. Sub-hypotheses / hypothesis aspects", "3. Experiment plan", "4. Result interpretation", and "5. Next-step decision". Each section has a "fill-in" area for notes. On the right is an "AI Assistant" chat window. The chat shows a conversation where the user asks for help generating sub-hypotheses, and the AI provides a list of suggestions. The user then asks for experiments to test the main hypothesis, and the AI provides a list of experimental approaches. The user then asks for a summary of a paper, and the AI provides a summary. The user then asks for interpretation of a result, and the AI provides an interpretation. The user then asks for the next step, and the AI provides a list of next steps.

- A document with fill-in sections: main idea, smaller ideas, experiment plan, results, and next step
- The AI chat can suggest ideas, experiments, paper summaries, and next steps as text
- The scientist has to copy and organize the answers into the document by hand

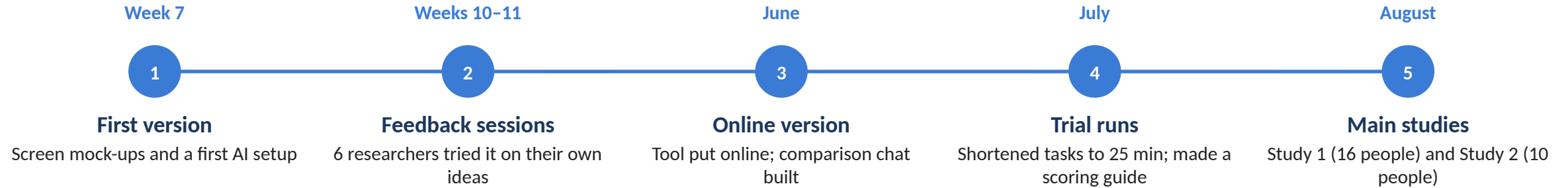
HypoTree: tree workspace

The screenshot shows a web application titled "Venetoclax / BCL-2 Hypothesis Validation". It features a hierarchical tree of hypotheses. The root node is "Venetoclax kills CLL cells because BCL-2 marks a venetoclax-sensitive cell state and venetoclax binding to BCL-2 directly triggers apoptosis in these cells." This node branches into four sub-hypotheses: "BCL-2 expression is upregulated in CLL cells that are venetoclax-sensitive", "Venetoclax acts directly through BCL-2 binding, not indirectly through TP53 or other apoptotic pathways", "Once venetoclax binds BCL-2 in venetoclax-sensitive cells, these cells die through the mitochondrial apoptosis pathway", and "The pattern is not explained by a general sensitivity or unrelated factors". Each sub-hypothesis node further branches into supporting evidence and sub-hypotheses. On the right is an "AI Assistant" chat window. The chat shows a conversation where the user asks for a summary of a paper, and the AI provides a summary. The user then asks for interpretation of a result, and the AI provides an interpretation. The user then asks for the next step, and the AI provides a list of next steps.

- The same AI suggestions appear as boxes in a tree that grows (DG1)
- Each link shows its reason and papers; each box has its own AI that remembers (DG2)
- New results update the tree and bring up suggested next steps (DG3)

Improving the Tool With Researchers

We built a first version, showed it to scientists, and improved it based on what they said



What researchers asked for, and what we changed

"Show how sure the AI is"

Links drawn thicker when the AI is more confident

"Tell me what result to expect"

Every experiment now lists its expected result

"Experiment steps are too long and vague"

Shorter steps inside each box, fitted to the idea being tested

"Let me add and change ideas myself"

Any box can be added, edited, or folded away

"Remember what we discussed"

Each box has its own AI that keeps its chat history

"After a result, suggest new ideas and new tests"

The AI now offers both; the scientist chooses

How We Tested HypoTree

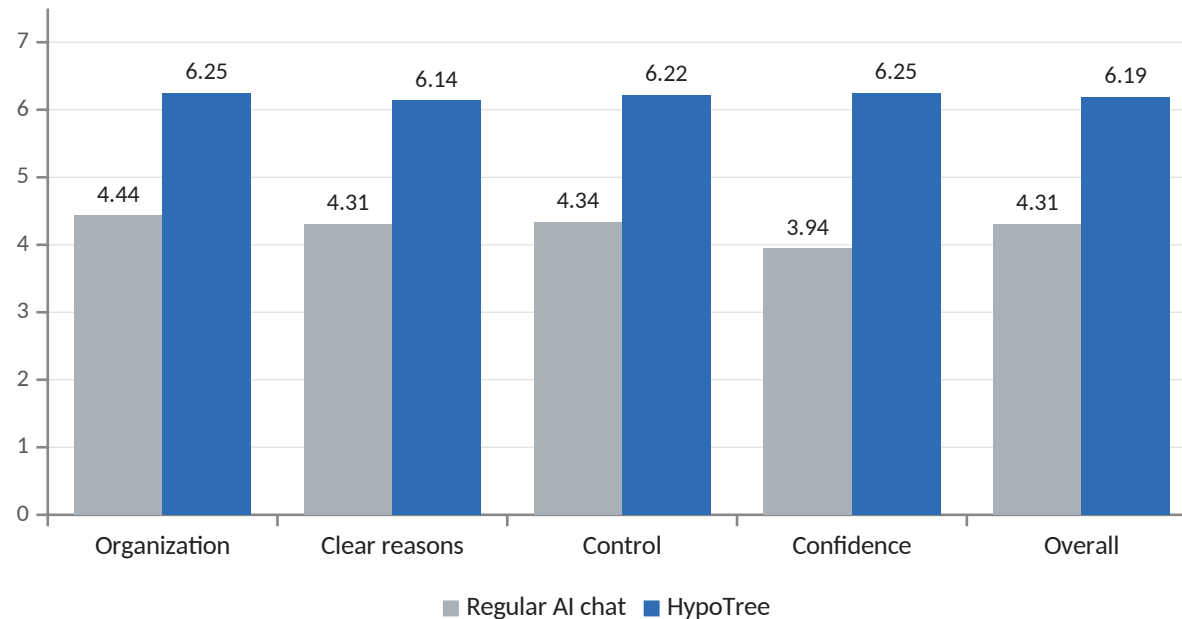
Two studies. Each person used both HypoTree and the regular AI chat on their own research

Study 1: Planning experiments (16 people)		Study 2: Understanding results (10 people)	
Setup Each person brought two similar ideas from their own research: one for HypoTree, one for the chat. Half started with each tool		Setup We gave each person sample experiment results for their own idea, in both tools	
Task Break the idea into smaller parts, choose experiments, and write a short plan: the main idea, its parts, the reasons, the experiments, and expected results		Task Add the results, figure out what they mean for the idea, decide the next step, and update the plan	
What we measured 1–7 ratings of organization, reasons, control, and confidence; how much effort it took; 3 experts scored each plan without knowing which tool made it		What we measured 1–7 ratings of help with reading results and choosing next steps, confidence, fit with their work, and control; follow-up interview	
Same-person comparison each person's two scores compared directly	Two statistical tests checked results two ways to be sure	Stricter threshold adjusted for testing many things at once	Expert scoring guide clear, organized, traceable, testable, sound, doable

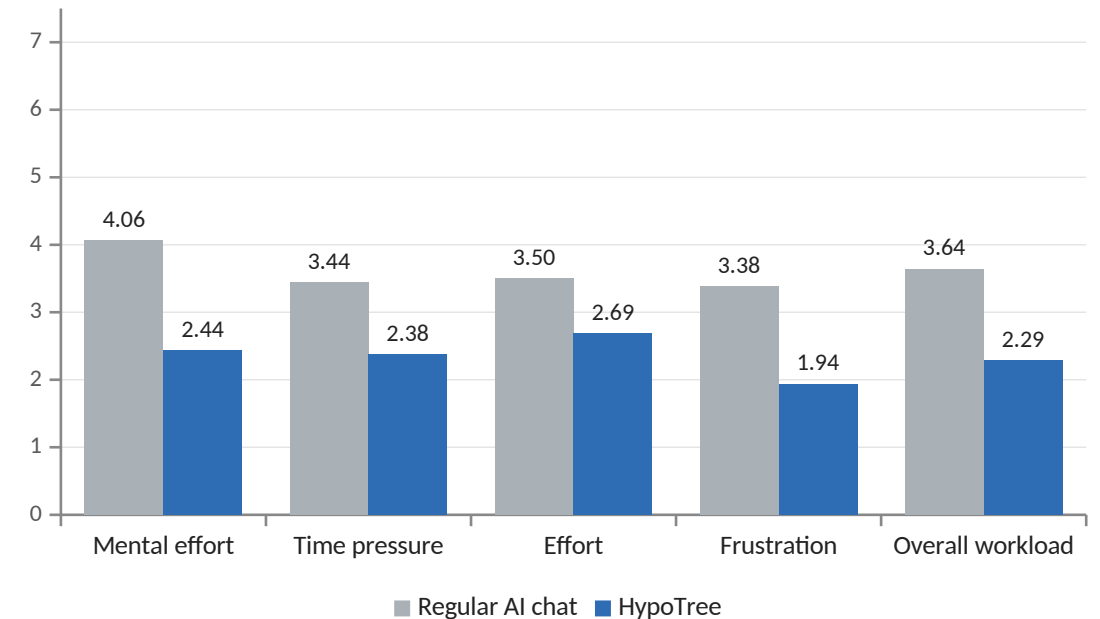
Study 1 Results: Planning

16 people rated HypoTree clearly higher on every measure, and found it less tiring to use

User ratings, 1-7 (higher is better)



How demanding it felt, 1-7 (lower is better)



+1.9

points higher overall rating (4.3 → 6.2 out of 7)

Very large

difference, far beyond chance

-1.3

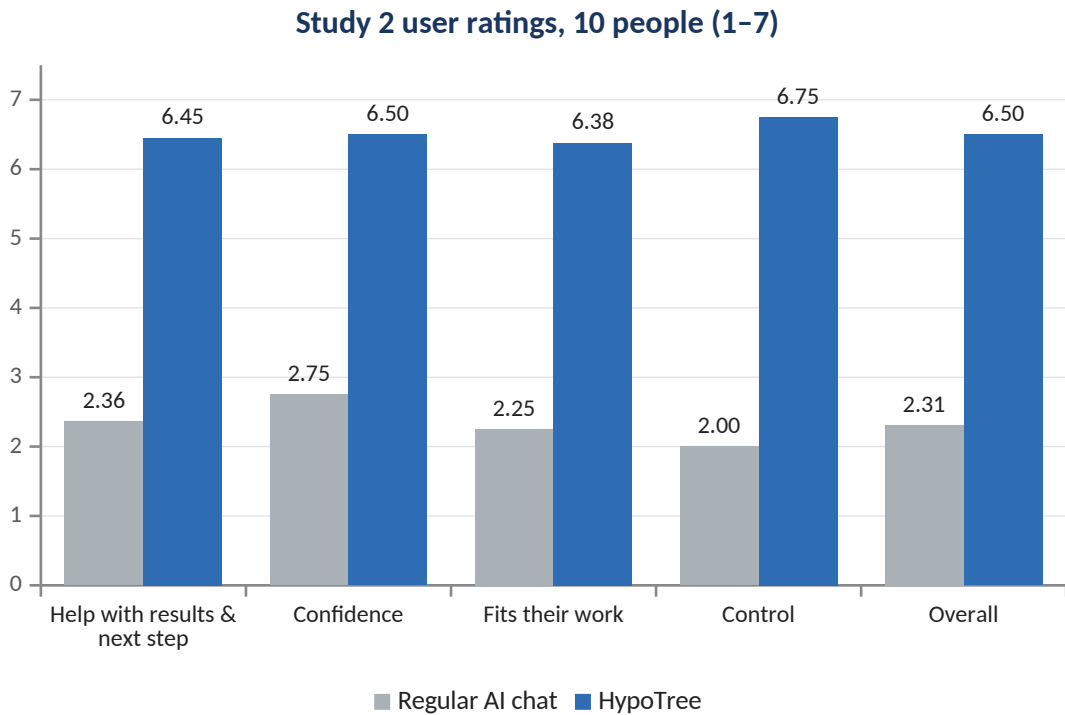
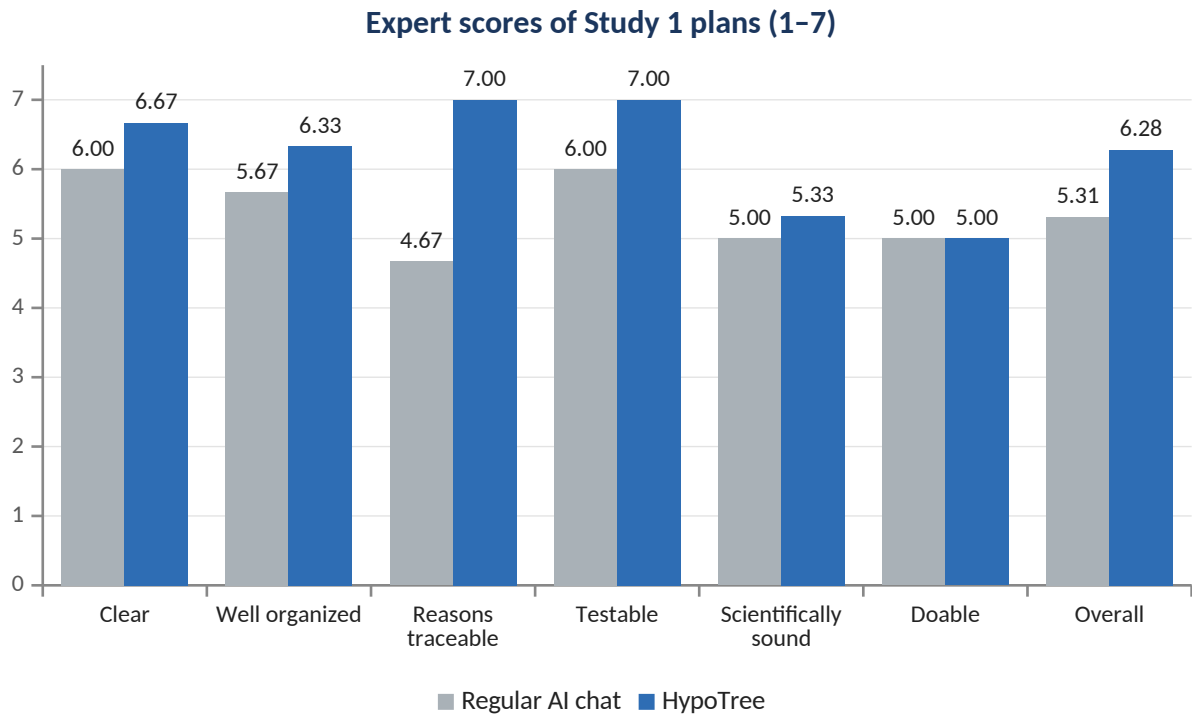
points lower workload (3.6 → 2.3)

5.8 vs 4.2

how successful people felt, HypoTree vs chat

Expert Scores and Study 2 Results

Experts scored HypoTree plans higher, and HypoTree helped most with understanding results



Biggest gain: clear reasons. Experts could trace why each experiment was chosen (4.7 → 7.0) and found plans easier to test. "Doable" stayed the same, which the tool was not designed to change.

Study 2: ratings rose by about 4 points out of 7 on every measure, and every single person preferred HypoTree.

Key Takeaways

What we learned, and what comes next



A tree works across fields

Researchers from biology to materials science put their own ideas into the tree and found it far clearer than regular notes.



AI answers you can check

Saved reasons, one AI per box, and suggested ideas turn AI text into steps a scientist can follow and verify.



The scientist stays in charge

The AI breaks down, explains, and suggests. The scientist approves every step, and it felt like less work, not more.

Next steps

Paper

Finish the research paper for CHI, a top conference on people and computers

Robot lab

Connect HypoTree to a robot lab so experiments can be split between the scientist and the robot

Real-world use

Let scientists use it for months on real projects and see how their trees grow