MONASH University

ASSESSMENT 1 · COHORT DEBRIEF

WHAT 57 PROPOSALS TOLD US

Coverage, performance, and the one thing that separated the strongest from the rest.

ANONYMISED · NO INDIVIDUAL IS IDENTIFIED IN THIS SESSION

PERFORMANCE

AT A GLANCE

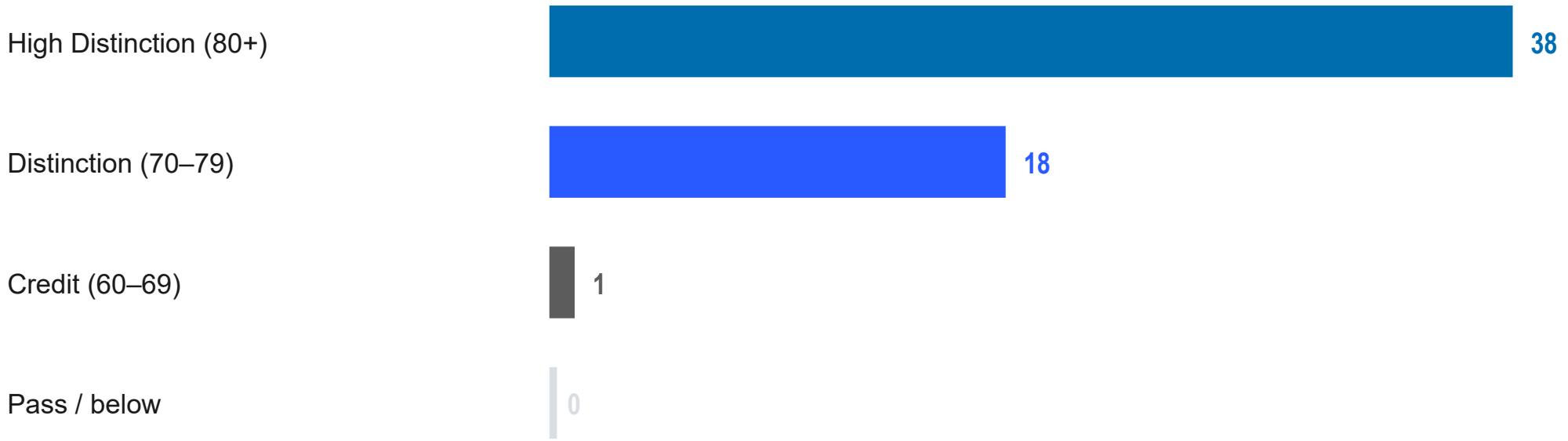
57	83.5	84.6	98%
proposals marked	mean mark	median	at Distinction or above

Range 69.8 to 96.2. Standard deviation 6.8 — a tight, strong cohort with a long tail of small differences rather than a split between those who could do it and those who could not.

What that tightness means for you: the marks were not decided by whether you understood AI. Almost everyone did. They were decided by how well your claims were evidenced — which is a skill you can move quickly, and the rest of this session is about how.

PERFORMANCE

WHERE THE COHORT LANDED



Two-thirds of the cohort reached HD. That is unusually high, and it is genuine — the proposals were, on the whole, thoughtful and well-researched.

It also means the marginal mark was expensive. Moving from 82 to 88 took something specific — see slide 10.

MONASH University

PERFORMANCE

WHERE MARKS WERE WON AND LOST

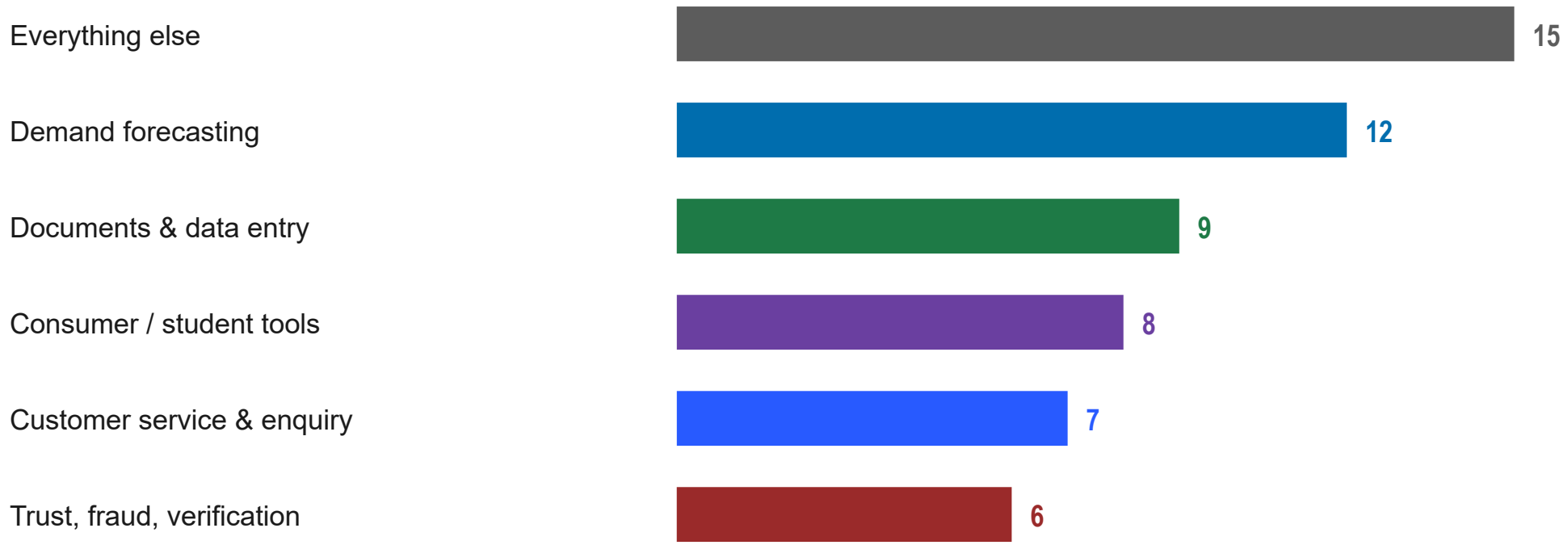
Average score on each criterion, across all 57 scripts.

The weakest criterion was the FIRST one. Problem framing scored lowest and varied most. Everything downstream — which AI, which risks, which model — is only as good as the problem you named. Narrow it until it can be acted on.

MONASH University

COVERAGE

WHAT YOU CHOSE TO SOLVE



57 distinct businesses with almost no duplication of concept — but heavy clustering of problem TYPE. Roughly a third are some form of demand forecasting; another quarter automate a document or an enquiry a small business currently handles by hand.

MONASH University

COVERAGE

THE FORECASTING CLUSTER — TWELVE OF YOU

THE SAME ANSWER, VIA THE SAME ARGUMENT

Bakery production, cafe prep, restaurant demand, convenience-store movers, grocery ordering, markdown timing, event staffing, campus queues, hospital registration.

Nearly all chose classical machine learning, and nearly all reasoned identically: structured point-of-sale data, tabular prediction, tree-based model, tolerable cost of error.

That reasoning is correct — which is precisely the problem. Twelve people arrived at the same defensible answer, so the criterion stopped telling us anything about you.

TWO ESCAPED — HERE IS HOW

One reframed from "food waste" to daily SKU-level production planning under uncertain demand — then proposed benchmarking against a rule-based baseline to prove the complexity earned its place.

The other moved from forecasting to markdown TIMING — when to discount, and by how much. That is a pricing decision, not a quantity one. Then they spotted that customers would learn the pattern and wait for it.

Both changed the question, not the technology.

If you are in a crowded lane, the differentiation has to happen at the framing stage — not at the solution stage.

MONASH University

COVERAGE

WHICH LAYER YOU CHOSE — AND WHETHER IT FITTED

21	**Classical ML**	structured data, interpretable, cheap, error recoverable
17	**Foundation model (usually + RAG)**	input is unstructured language; the knowledge already exists to retrieve
9	**Explicit hybrid**	the problem has two halves needing different layers
7	**Deep learning**	the input is an image and no lower layer can see it
3	**Rule-based**	answers must be consistent; no training data exists

The split is close to even between the two ends of the stack — and almost nobody proposed an agentic system as the deliverable. That is the right instinct, and mostly for the right reason.

MONASH University

WHAT THE STRONGEST DID

THE HYBRIDS WERE THE BEST THINKING IN THE COHORT

Nine of you recognised your problem had two parts, and assigned a different layer to each.

A generative model for the image, a separate model for the size prediction

two different jobs, two different tools

Vision to read the receipt, deterministic code to do the arithmetic

a model that reads text probabilistically should not be trusted with sums

A cheap keyword filter to clear the obvious cases before the expensive multimodal model runs

cost control as a design decision, not an afterthought

Notice what these have in common: the student decomposed the problem before choosing anything.

MONASH University

WHAT THE STRONGEST DID

CHOOSING THE SIMPLEST THING THAT WORKS IS A RESULT

Three of you chose rules outright, and defended it: police report intake, where the force must own every question and the answers must be identical every time. Rental repair intake, where the letting agents were interviewed and found not to hold enough labelled history to train anything. Construction timesheets, where the exceptions are already defined in an award.

Higher is not better. A lower layer, chosen deliberately and justified out loud, scored better than a foundation model chosen because it sounded impressive.

THE SINGLE BIGGEST SEPARATOR

WHERE YOUR EVIDENCE CAME FROM

The strongest and the weakest proposals often described the SAME business problem. What separated them was almost never the idea — it was whether the evidence could bear weight.

THE SELLER SOLD BOTH

Several proposals established that the manual method is inadequate, and that AI fixes it — with both claims sourced to companies selling AI for that task. A vendor page comparing its product to the status quo is not a measurement of the status quo.

FIGURES YOU COULD NOT FOLLOW

One proposal carried 27 figures and 2 sources. Another rested its headline statistic on a page whose name it spelled two different ways in one sentence. A number you cannot trace back to who measured it is decoration.

CITATIONS REACHED VIA BLOGS

The consultancy names everyone recognises appeared repeatedly — but reached through posts summarising the reports rather than the reports. In one case the URL led to unrelated vendor documentation.

The contrast: a handful of you cited peer-reviewed work in the exact application area — restaurant demand forecasting, delivery attempt prediction, fairness in credit scoring. Those proposals read completely differently, because the claims stopped being assertions.

MONASH University

THE GAP BETWEEN GOOD AND EXCELLENT

FIVE THINGS THE BEST ANSWERS DID

1	**Went and asked someone**	Interviews with tenants and letting agents. A student survey with the instrument reported. Two accounts of doing the work for named clients. Those problem statements were unarguable, because they were not arguments.
2	**Narrowed until it could be acted on**	Not food waste — next-day item-level production planning. Not counterfeit goods — the monitoring capacity a small firm lacks. The narrowing is what lets you name a layer at all.
3	**Refused the fanciest option, out loud**	Chose a lower layer and said why: the data is structured; the answers must be consistent; a regulator requires a human signature.
4	**Costed being wrong in BOTH directions**	Overpredicting costs this, underpredicting costs that, and they are not the same size — then used that asymmetry to decide how much oversight to build.
5	**Found the risk their own design created**	Brand-voice drift because everything runs through one system. A queue predictor that empties the queue it predicted. A markdown pattern customers learn to wait for.

NAME THE CATEGORY

SECOND-ORDER RISK

Bias, privacy and hallucination are the risks everyone lists. They are real, and listing them earns you the middle of the band.

The risk your own design creates is harder to see — and it is the one that would actually bite.

A queue predictor	**A markdown pattern**	**A recommender**	**Brand-voice drift**
that empties the queue it predicted	customers learn to wait for	funded by the restaurants it recommends	because every client runs through one system

Students who found one of these wrote visibly better proposals than students who listed the standard three. Ask it of your Assessment 3 build: what does my design make possible that was not possible before?

THE MOST COMMON WEAKNESS

PART 5 — IDEATION WAS NOT IDEATION

✗ THE SAME COMPARISON TWICE

In five scripts the three “alternative ideas” were the three technology layers already compared in Part 2 — rule-based, then machine learning, then a foundation model.

That is the Part 2 comparison restated, not ideation.

An alternative concept is a different PRODUCT, a different CUSTOMER, or a different FRAMING of the problem.

✓ REASONS THAT ACTUALLY KILL AN IDEA

“Set aside — an incumbent already ships it in this market.”

“Set aside — there would be no data to sell before there were users.”

“Set aside — it would be a thin wrapper anyone could copy.”

Those kill an idea. “It is less flexible” does not.

And almost nobody said how they would find out whether they were right. The few who did — five to eight interviews, these four measures, and if corrections stay high we fix the data before adding features — stood out immediately.

MONASH University

ON THE AI-USE DECLARATIONS

THE DECLARATION IS EVIDENCE OF JUDGEMENT

Disclosures ranged from twenty-three passage-level comments to none at all. The best were not the longest — they were the ones that distinguished between FINDING something with AI and WRITING something with AI, and recorded what was checked.

Deleted AI-generated percentage claims because nothing supported them

Other proposals published exactly that kind of figure.

Checked that three named products actually existed before citing them

Two minutes of work; the difference between a citation and a guess.

Chose hedged wording — “estimated”, “could” — because the study covered one setting

Knowing the limit of your evidence is the skill.

A comment saying only “I used AI here” discloses that a tool was involved and tells your reader nothing about whether you checked it.

CARRYING IT FORWARD

WHAT THIS MEANS FOR ASSESSMENTS 2 AND 3

1

Name an incumbent

If someone already ships your idea, say so and say why yours differs. It forces differentiation at the framing stage.

2

Go and ask someone

Five interviews beats fifty citations. It was the single strongest move in this cohort and almost nobody made it.

3

Price being wrong

Both directions, with numbers. Then use the asymmetry to decide something real about your build.

4

Find your second-order risk

The one your design creates. Not bias, privacy and hallucination — the one specific to what you are making.

All four are cheap. None of them requires a better idea — only better evidence for the one you have.

MONASH University

MONASH University

THE ONE SENTENCE

THE IDEA WAS RARELY THE PROBLEM. THE EVIDENCE WAS.

Almost everyone in this room understood the technology well enough. What separated 82 from 90 was whether a claim could be followed back to someone who measured it — and whether you had gone and found out anything yourself.

Week 6 is where this becomes a habit: name a baseline, build a small eval set, get one number, say it honestly.

YOUR INDIVIDUAL FEEDBACK IS ON MOODLE.