

5–20

rows per sub-task

5

Criticals, minimum

1 in 5

Criticals a strong attempt should fail

90%

inter-judge agreement, three runs

4–5

numbers per row, maximum

THE SIX COLUMNS

#	Integers. No 4a, no 4.1
Section / Sheet	The slide title or Excel tab word for word to match the golden, or Universal. Never “Slide 1”
Scope	Deck / Model / Both
Rubric	Full sentences that stand alone. Two to four points at most
Category	Structural Numeric Logic Methodology Formatting Consistency & Coherence
Tier	Critical / Non-Critical

WHICH CATEGORY

Structural	Is the element there?
Numeric	Is the number right?
Logic	Does the claim hold?
Methodology	Was it built the right way?
Formatting	Does it follow the stated presentation rules?
Consistency & Coherence	Do the same figures agree wherever they appear?

THE TOLERANCE LADDER

Tolerance follows where the number came from, not how confident you feel.

Extracted from the VDR or source documents	Tight — rounding only
Calculated from figures the prompt fixed or the VDR provided	About ±1%
Worked out by the analyst: ratios, valuation, returns	About ±5%
Handed to the agent by the prompt or the input model	Never tested
Choices the prompt left open	Fix in the prompt, or widen downstream

Write the range out in full: 18.6% (acceptable range 17.7% to 19.5%). Never “within ±5%”.

CHECKS THAT NEED NO NUMBER

Impossible to fake, seconds to grade, and most first-time writers never think of them.

- Sources equal uses
- The case selector works
- The model moves when a driver changes
- Forecast tabs pull from the historical columns by formula

CRITICAL — THREE TESTS, ALL THREE REQUIRED

Mandatory	Every correct answer has it. Work that fails it is wrong, not just different
Objective	Visible in the deck or model, with the expected value written into the row
Stable	Separate judges would reach the same verdict

Miss any one and it is Non-Critical. When in doubt, leave it out.

A Critical gates. Fail one and the whole sub-task fails, whatever else passed. Nineteen of twenty rows passing is not 95%. It is a fail.

THREE QUESTIONS THAT DECIDE MOST ROWS

Wrong, or different?	If a different sensible build would fail it, it is not Critical
One row, or two?	Would you ever pass one part and fail the other? If the report should tell them apart, two rows
Does it flip?	Break the build on purpose. If the verdict does not flip to FAIL, the row checks nothing

FIVE-MINUTE READ, BEFORE HANDOVER

1. Read every criterion aloud. Cannot say in one sentence what a passing slide contains? Not finished
2. Where did each number come from? Handed to the agent, delete the row. Otherwise check the ladder
3. What can the judge see? Never the prompt, the input file or the golden
4. Vague words. “That”, “it”, “they” — name the thing in the same row
5. Count the numbers. More than four or five, split. Unrelated checks, split
6. How did you use the golden? For the facts. Never for its layout, wording or order

THE FIVE GATES

1. Grade the golden against your own set. It must pass every Critical. If not, say whether the golden or the rubric is wrong
2. Two-reader test. If two reviewers could read a row two ways, reword it
3. Try to break it. What would an empty file or a confident fake score? Both should fail the number and method rows
4. Check three runs agree. 90% of criteria or better
5. Check it is hard enough. A strong attempt should still fail about one Critical in five

NEVER

- Test a number the agent was handed
- Write a check the judge cannot see in the submitted work
- Make a Critical out of the golden's layout, wording or ordering
- Let one row lean on another — every row stands alone
- Count the same mistake twice
- Add a second sentence, a contrastive tail, or a restatement of the verdict