Cheat sheet: the seven ways AI review fails

A one-page summary of the seven dimensions above, one control per dimension, no claim beyond what is documented in this article.

Before every AI review of a document, a codebase or an analysis, run this list.

1. Attention decay (lost in the middle). Long input recalled well at the start and the end, poorly in the middle. That is where the requirement usually sits. Control: split the input; check each requirement against the result on its own.

2. Effort avoidance. A completeness claim states everything is captured. It quietly covers only part of the list. Control: build the coverage list outside the model first, then reconcile the answer against it.

3. Sycophancy. A correct finding retracted the moment a human pushes back. No counter-evidence required. Control: never take agreement as a signal; require a passing test, a real reference target and a matching coverage number.

4. Mode collapse. Fewer findings and more apologies with every round of self-correction. Control: bound the number of correction rounds; judge the result against an external standard.

5. Hallucinated references. A fluent, specific reference to a function, clause or source that does not exist. Control: void any finding whose named entity has no real target in the source.

6. Security illusion. A confident, well-argued approval with a real gap the argument never names. Control: judge safety against executable tests and known attack patterns, never the model's stated confidence.

7. Operational fragility. The system around the model quietly served a different model or a truncated input and every check still passed. Control: log which model actually served the request and whether the full input reached it.

Full mechanism, source and example: gates-blog-06b-seven-dimensions-technical