

Evaluating AI Output

Computing

How to check an answer, and the failure modes to check for specifically.

The five failure modes that account for most bad output.

Watch for

Hallucination — confident and wrong

Stale knowledge — a training cutoff

Sycophancy — agreeing with your framing

Lost detail — the middle of long context

Plausible citations that do not exist

Check by

Asking for sources, then opening them

Asking the same thing two ways

Giving it the document rather than trusting recall

Verifying every number independently

Lower the risk

Supply the source material yourself

Ask it to say when it does not know

Keep the task narrow

Ask for reasoning, then check the reasoning

A model has no reliable access to its own certainty. Hedged wording and confident wording are both just style.