

A smaller AI. A clearer job.

Anthropic's 7 October Haiku 5.5 launch highlights a useful idea: match the model to the task, then test whether it does the job.



NARROW TASK



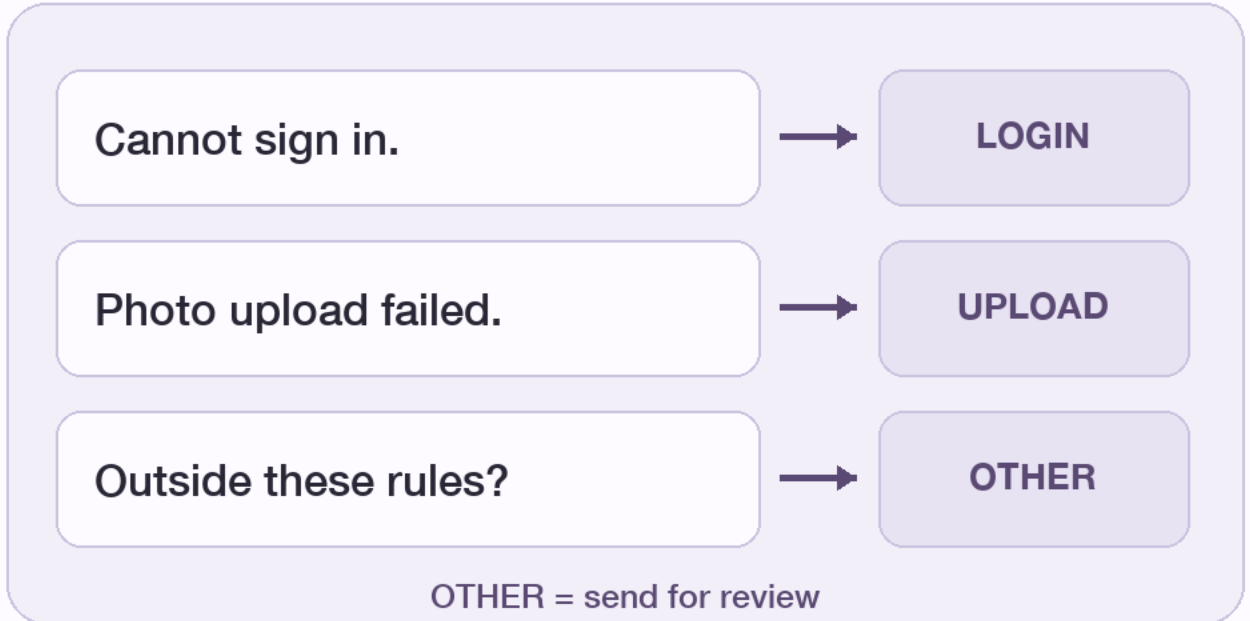
COMPLEX TASK

Match the model. Then test it.

Fictional examples. No measured performance or savings.

Give it a bounded task.

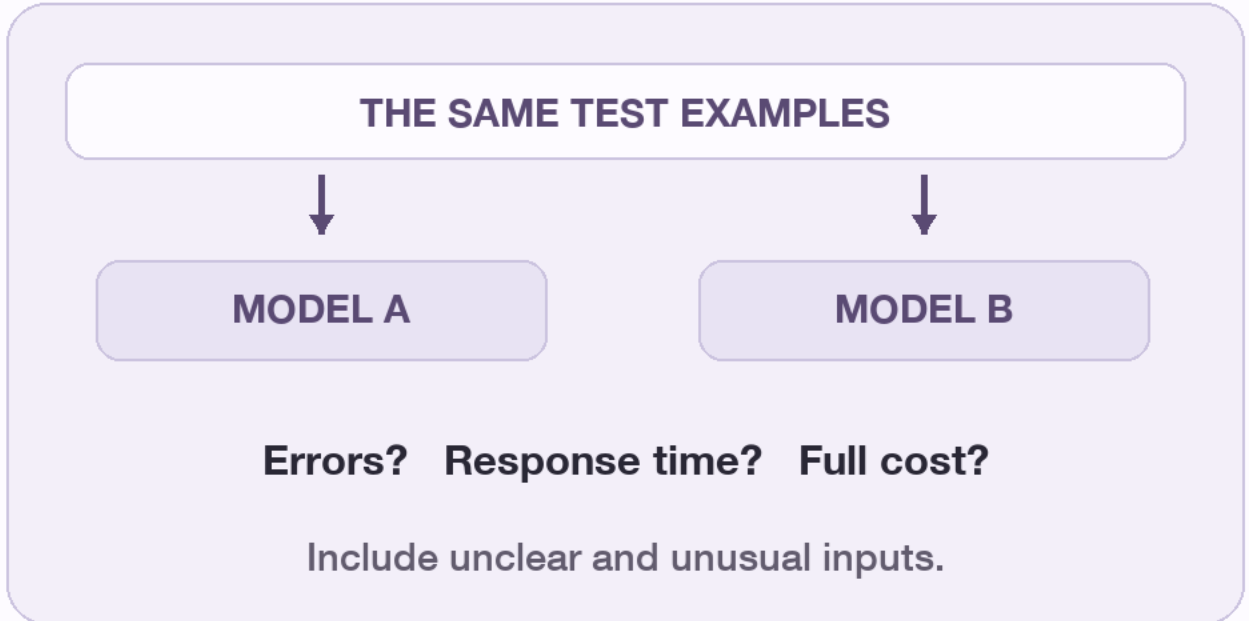
For a fictional app, label a message Login, Upload or Other. Define the labels and what should happen when the message does not fit.



Fictional examples. No measured performance or savings.

Test before you trust it.

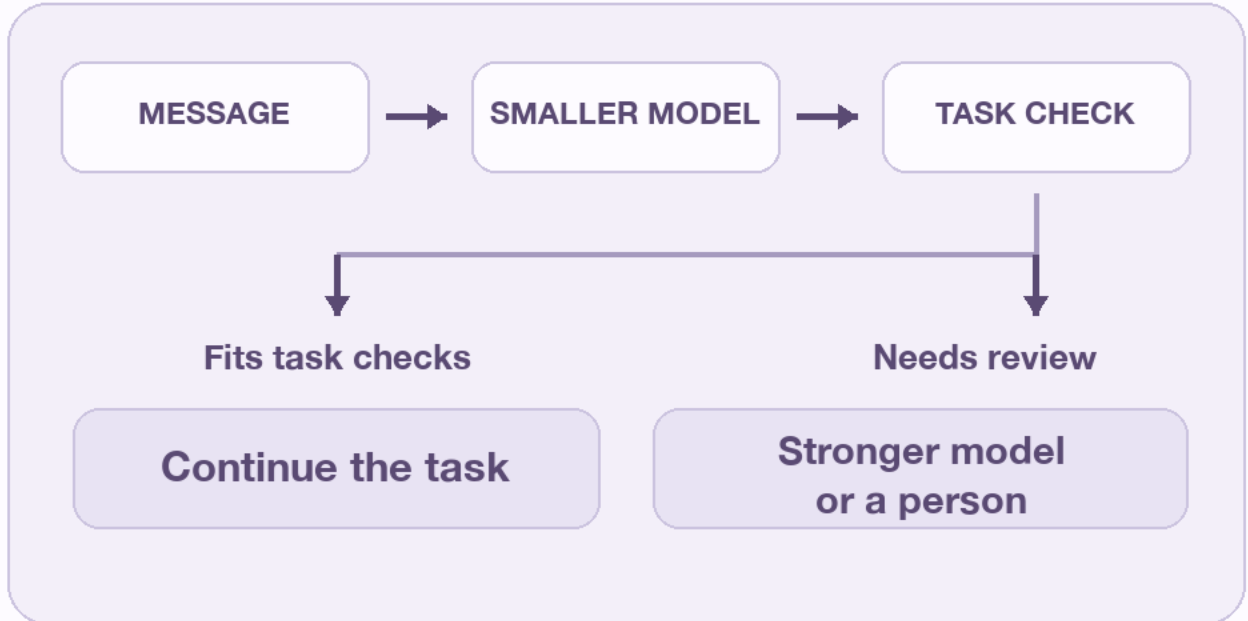
Use examples with known answers, including ambiguous cases. Check errors, response time and the full cost of completing the task.



Fictional examples. No measured performance or savings.

Keep an escape route.

A case outside the rules goes to a stronger model or a person. A model sounding sure is not proof that its answer is right.



Fictional examples. No measured performance or savings.

Count every model call.

Assume 100 first passes at a fictional \$0.02 each, plus 20 extra calls at \$0.20. Those model-call fees add to \$6.

FICTIONAL USD FEES	
100 FIRST CALLS	$100 \times \$0.02 = \2
20 EXTRA CALLS	$20 \times \$0.20 = \4
Model-call total	\$6

Fictional examples. No measured performance or savings.

Compare the whole job.

All 100 calls at that fictional \$0.20 would cost \$20. This is price arithmetic, not a quality comparison. Add routing, retries and review.

SMALLER FIRST

\$6

LARGER ONLY

\$20

Do both meet the quality bar?

+ routing + retries + review

Fictional fees. No measured saving.

Fictional examples. No measured performance or savings.

Check access.

Check real rates.

Model access and prices depend on the product, provider and account. Real API bills depend on tokens, prompt length, caching and tools.

ACCESS

Product / provider / account

USAGE

Input + output + prompt length

EXTRAS

Caching + effort + tools

Fictional examples. No measured performance or savings.

Match. Test. Then route.

Choose a model that meets the task's quality bar. Keep a review path and compare total cost per accepted result.

01 What must be correct?

02 What if the case is unclear?

03 Cost per accepted result?

A proposed workflow. Test it first.

Fictional examples. No measured performance or savings.