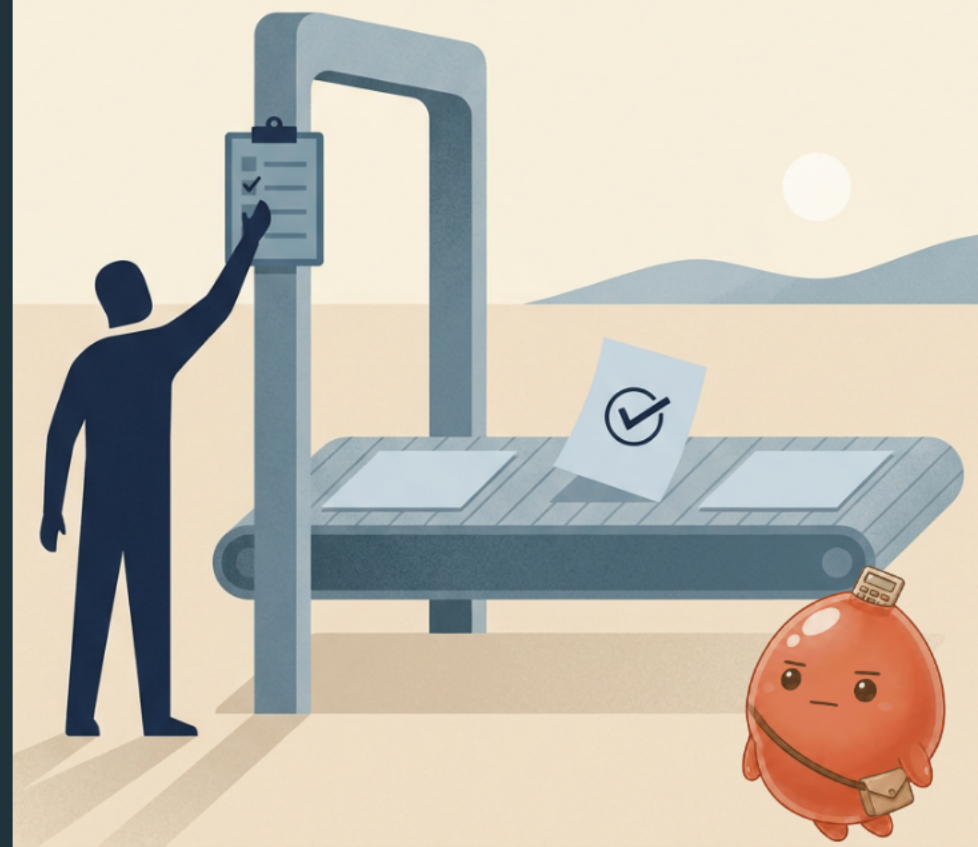


検証を仕組み化する

「確認して」と毎回頼まなくていい状態へ



今回は、「確認」そのものを 仕組みとして育てる

前回は、委任を広げるほど、確認をどこに置くかの設計が重要になる、という話だった。
その「確認」を、毎回自分の目でやるのではなく、決まった形で勝手に動くようにしていく。
手順書に「確認すること」と書いてあるだけでは、思い出さなければ動かない。
思い出さないと動かないものは、仕組みではなく、まだ心がけにすぎない。



検証の理想形 — 3つの条件

基準

判定基準を、先に渡す

なんとなく良さそう、で見ない。何をもって合格とするかが、先に決まっている。

自動

発動が自動である

「確認して」と毎回頼まなくても、決まったタイミングで動く。

反復

直るまでループする

1回指摘して終わりではなく、基準を満たすまで直させる。



「確認すること」と書いてあるだけなら、 それは仕組みではない

手順書に一行あるだけの状態は、思い出したときにしか動かない。
3つの条件がそろって、はじめて「仕組み」と呼べる。

心がけを、仕組みに変える。



なぜ「判定基準」が絶対に要るのか

AIの推敲能力を測った、ある調査がある。

自分が出したものを自分だけで見直させても、最先端のモデルで正答率は2～3割ほど。

何度繰り返しても、大きくは変わらない。

でも、どの基準を満たせていないかを外から毎回指摘してやると、数回のやりとりでほぼ完璧まで直せる。

出典：RefineBench (arXiv:2511.22173／ICLR 2026採択) — AIの推敲能力を測るベンチマーク

自己推敲 正答率 約21～31% → 外部からの不合格項目指摘つきの見直しで ほぼ完璧に



実習：検証専用スキルを1つ作る — まず材料あつめ

選ぶ

繰り返す確認作業を1つ

自分の目で必ず確認している作業を1つ。何度も繰り返すものを選ぶ。

見比べ

良い例・ダメな例を3つずつ

良かったもの・ダメだったものを、3つずつ思い出す。実物があれば見比べる。



Step3：違いを「感想」ではなく「基準」にする

多すぎると運用しにくい。まずは3～5個のチェックリストにして、あとで足していく。

感想のまま

- ・なんとなく読みにくい
- ・雰囲気合っていない
- ・人によって判断がぶれる

基準に言い換える

- ・一文が60字を超えている
- ・主語が抜けている
- ・誰が見ても同じ判定になる



実習のつづき — スキルにして、試す

保存

その基準でAIに確認させ、スキルに

チェックリストに沿って確認させ、全項目クリアまで直させる。
うまくいったら、その指示をスキルとして保存する。

試す

過去の成果物で走らせて調整

ダメだった例に、ちゃんと引っかかるか。
引っかからなければ、基準がまだ曖昧。具体的に直す。



今日つくれたのは、 「呼べば動く」ところまで

スキルは、近い作業を始めれば自動で立ち上がることもある。
でも、AIがその場で判断して呼ぶだけ。毎回かならず、ではない。
下書きができるたび確実に検証を走らせるには、フックが要る
(置くだけでは動かず、設定に登録して動く。詳しくは第6回)。

感想でチェックする、から、基準でチェックする、へ。それだけで精度は別物になる。



今日のチェックリスト

- ・ 繰り返し確認している作業を1つ選んだ
- ・ 良い例・ダメな例を見比べ、基準を3~5個の言葉にした
- ・ その基準でAIに確認させ、スキルとして保存した
- ・ 過去の成果物で試し、正しく引っかかるか確認した

次回は、設計。

そのループ、いつまで続けるのか。

止めどきを、先に決める。

