

Does it help with your real task?

Record evidence first. Score only what you can support.

Task / decision: _____

Tool + version: _____ Test date: _____

Usable-result standard: _____

Actual cost / expected frequency: _____

TIME Score: ____ / 5 or not assessed

Total time including preparation, review and corrections.

Evidence: _____

QUALITY Score: ____ / 5 or not assessed

Accuracy, completeness and usability. Keep both outputs.

Evidence: _____

RELIABILITY Score: ____ / 5 or not assessed

Repeated trials, failures and conditions where the tool struggles.

Evidence: _____

EFFORT Score: ____ / 5 or not assessed

Setup, prompting, learning, supervision and cleanup.

Evidence: _____

VALUE Score: ____ / 5 or not assessed

Practical benefit versus actual cost, frequency and alternatives.

Evidence: _____

Verdict: USE IT / USE WITH LIMITS / SKIP IT / TOO EARLY TO TELL

Main limitation: _____

A critical error can outweigh any total. No evidence? Leave the score unassessed.

How to score without guessing

Use the same finish line for your usual and AI-assisted workflows.

1. Define the task and the usable-result standard before testing.
2. Time both workflows. Include setup, checking and corrections.
3. Keep the outputs. Record mistakes and any unsupported commitments.
4. Repeat across comparable tasks and record every failure.
5. Explain each score and select a practical verdict.

1 - Clearly worse

Unacceptable result or substantial net disadvantage.

2 - Limited benefit

Major drawbacks make routine use difficult.

3 - Acceptable

Useful in some cases, with meaningful tradeoffs.

4 - Useful

A practical benefit with minor limitations.

5 - Strong benefit

Clear benefit under the conditions you actually tested.

These anchors apply within each dimension, not across all tools.

Record raw times, errors, effort and costs beside the judgment.

Do not total an incomplete scorecard or turn a total into an accuracy percentage.

When evidence is insufficient, choose TOO EARLY TO TELL.

Completed example: Test #001

ChatGPT assistance for short, routine professional emails.

Observed timing result

One participant. 10 manual trials; 3 comparable AI-assisted trials.

Manual median: 151.5 sec. Assisted median: 16 sec.

About 89% less time in this sample; not a universal productivity claim.

Baseline score: 24/25 - PROVISIONAL

Time 5/5 | Quality 5/5 | Reliability 4/5 | Effort 5/5 | Value 5/5

Quality: non-blind evaluation; manual examples were written after timing.

The original ten timed manual emails were not saved.

Reliability: only three timed assisted trials; no general reliability estimate.

Effort: setup and supervision were not separately quantified.

Value: no paid-subscription cost-benefit calculation was performed.

What the score does not prove

It is not an independently validated product rating.

It does not establish the quality of the ten timed manual outputs.

It does not show that buying a paid subscription is worthwhile.

Times were participant-reported; exact model/version was not recorded.

VERDICT: USE IT

Use it for drafting, not blind sending.

Always verify facts, commitments and context.

Full method, results and limitations: tasklitmus.com/tests