Founders AI
founders-ai.dev

PRACTICAL AI FIELD GUIDE

AI Quality System

How to Check AI Work Before You Trust It

A practical review guide for founders, business owners, creators, consultants, and professionals who want to check AI output before trusting, sending, publishing, or automating it.

- Define quality
- Extract claims
- Verify sources
- Check logic
- Score with rubrics
- Save failures

Most people try to get better results from AI by asking better questions. That helps, but it is only half the system. If you want to use AI for real work, you need a way to check the work before you trust it.

The rule is simple: **AI can draft, but proof needs a process.** AI can help you think, write, plan, research, summarize, sell, code, automate, and organize. But if the output is going to affect a customer, a decision, a payment, a public post, a legal claim, a medical claim, a business process, or an automation, you need a review layer before it moves forward.

This guide gives you that review layer. It is built for founders, business owners, creators, consultants, professionals, and everyday people who want to use AI seriously without pretending AI is always right.

> **The goal is not to distrust AI. The goal is to stop trusting AI in places where you have not checked it yet.**

Why AI output fails

AI output usually feels cleaner than it is. It can sound confident, structured, and useful while still being wrong, incomplete, outdated, overconfident, or unsafe. NIST calls this risk "confabulation," where generative AI systems confidently present false or erroneous content, including fabricated citations or reasoning (NIST AI 600-1).

For normal users, the danger is not that AI makes mistakes. The danger is that AI makes mistakes in a format that looks finished. That is why the review process has to separate presentation quality from actual quality.

The five common failure types:

- **Fabricated facts**: The answer includes claims, names, numbers, studies, quotes, links, or citations that are wrong or nonexistent.
- **Misgrounded sources**: The source exists, but it does not actually support the claim.
- **Bad logic**: The answer has steps that sound reasonable but do not follow from the facts.
- **Missing context**: The answer ignores constraints, edge cases, local rules, customer reality, or timing.
- **Unsafe action-readiness**: The answer is written as if it is ready to send, publish, automate, or execute when it still needs approval.

The Founders AI quality rule

Use AI in two separate modes:

1. **Creation mode**: Ask AI to draft, explore, summarize, generate options, or build a first version.
2. **Review mode**: Ask AI, another tool, or a human to check claims, assumptions, logic, tone, risk, and readiness.

Do not let the same answer pass directly from creation to action without review. OpenAI's evaluation guidance makes the same broader point for AI systems: because generative AI is variable, teams need structured evaluations rather than ordinary one-time testing (OpenAI evaluation best practices).

The everyday version is this:

> **Do not ask, "Does this look good?" Ask, "What would make this fail if I used it?"**

The quality ladder

Not every AI output deserves the same level of review. A brainstorm does not need the same process as a customer email, a sales proposal, a pricing page, or an automation that touches your CRM.

Risk level	Example output	Review required
Level 1: Low-risk thinking	Ideas, outlines, rough drafts	Quick read for usefulness
Level 2: Personal productivity	Notes, summaries, task lists	Check for missing context
Level 3: Business-facing	Emails, posts, sales copy, SOPs	Claim check, tone check, human approval
Level 4: Decision support	Market research, pricing, hiring, vendors	Source check, logic check, assumptions check
Level 5: Automation/action	CRM updates, invoices, code, messages, deletes	Validation, permission limits, logs, human approval
Level 6: High-stakes	Legal, medical, financial, safety, regulated work	Expert review before use

If the output can embarrass you, cost you money, mislead a customer, break a system, or create legal risk, move it up the ladder.

The AI Quality Loop

Use this loop every time the AI output matters:

1. **Define the job**: State what the output is supposed to do.
2. **Define quality before generating**: Write the standard before you judge the answer.
3. **Extract the claims**: Pull out facts, assumptions, recommendations, and actions.
4. **Verify facts and sources**: Check claims against trusted sources.
5. **Check logic and math**: Look for unsupported jumps, wrong calculations, and hidden assumptions.
6. **Check fit**: Review tone, audience, brand, timing, and practical usefulness.
7. **Check risk**: Decide what happens if this is wrong.
8. **Score with a rubric**: Use pass/fail or a 1 to 5 score.
9. **Revise or reject**: Do not polish a broken answer.
10. **Save the failure**: Turn mistakes into future tests.

Google's Gen AI evaluation guidance describes rubrics as pass/fail tests similar to unit tests, which is the right mental model for everyday users too (Google Gen AI evaluation service). If an output cannot pass the test, it is not ready.

Step one: write the quality brief

Before asking AI to create the output, tell it what "good" means. This prevents vague outputs and makes the review easier.

Use this prompt:

```
Before creating the output, restate the quality standard.

The output should:
- Serve this audience: [audience]
- Help them accomplish: [job]
- Be useful because: [value]
- Avoid: [things that would make this bad]
- Include: [required sections, facts, examples, or format]
- Exclude: [claims, tone, fluff, unsupported advice, risky content]
- Be checked for: [facts, sources, logic, math, tone, risk]

After you create it, include a short self-review against those criteria.
```

This turns "make it good" into a visible standard. The standard becomes your first quality control.

Step two: extract claims before checking

You cannot fact-check a wall of text. First, separate the output into pieces.

Use this prompt:

```
Review the output below.

Extract it into four lists:
1. Factual claims that need verification
2. Assumptions that may or may not be true
3. Recommendations or decisions being suggested
4. Actions someone might take based on this

Do not verify yet. Only extract.

Output:
[paste output]
```

Once the claims are visible, the review becomes manageable. You can check the five claims that matter instead of rereading the whole answer and hoping you notice the problem.

Step three: use the claim/source/action table

This table is the fastest way to check AI output before you use it.

Claim or action	Why it matters	Evidence needed	Status
“This tool is best for small teams”	Could influence a purchase	Current pricing, reviews, feature fit	Unchecked
“Send this offer to cold leads”	Could affect reputation	Compliance, audience fit, tone	Needs review
“The market is growing quickly”	Could affect strategy	Current market data from credible source	Unchecked
“Use this API route”	Could affect code/security	Docs, tests, permissions	Needs technical check

Do not mark something approved just because the AI sounded confident. Mark it approved only when the evidence is good enough for the risk level.

Step four: fact-check the highest-risk claims

Not every claim needs a research project. Start with claims that are specific, recent, numerical, legal, medical, financial, technical, or decision-changing.

Use this prompt:

```
Fact-check the claims below.

For each claim:
- Mark it Supported, Unsupported, Misleading, Outdated, or Needs More Context
- Explain the issue in plain English
- Identify what source would be needed to verify it
- Rewrite the claim so it is safer and more accurate

Claims:
[paste extracted claims]
```

For important claims, do not rely only on the model's answer. Cross-check with trusted external sources. OWASP's misinformation guidance specifically recommends cross-checking LLM outputs with trusted external sources and using human oversight for critical or sensitive information (OWASP LLM09:2025 Misinformation).

Step five: check whether sources actually support the claim

The source existing is not enough. A citation can be real and still not prove the sentence it is attached to. Stanford HAI's analysis of legal AI tools found that some answers were "misgrounded," meaning the tool described the law correctly or plausibly but cited a source that did not support the claim (Stanford HAI).

Use this source-check prompt:

```
Check source grounding.

For each cited source:
- What exact claim is the source supposed to support?
- Does the source directly support that claim?
- Is the source current enough?
- Is the source primary, authoritative, or just commentary?
- What caveat should be added?

Return a table with: Claim | Source | Supports claim? | Problem | Safer wording.
```

This is especially important for research, pricing, medical claims, legal claims, financial claims, tool comparisons, statistics, and anything with a date.

Step six: check logic, not just facts

An answer can have true facts and still reach the wrong conclusion. This happens when AI jumps from "this could work" to "you should do this," or from "some people do this" to "this is the best strategy."

Use this prompt:

```
Act as a skeptical operator.

Review this output for logic problems:
- Unsupported conclusions
- Hidden assumptions
- Missing constraints
- Overgeneralization
- Wrong cause-and-effect
- Advice that ignores cost, time, risk, or skill level

For each issue, explain the problem and rewrite the recommendation more carefully.
```

The goal is not to make AI timid. The goal is to make it honest about what it knows, what it assumes, and what still needs judgment.

Step seven: score the output with a rubric

Rubrics make quality visible. They also keep you from approving work just because it is well-written.

Use this simple five-part rubric:

Category	Pass standard
Accuracy	Important facts are verified or clearly caveated
Usefulness	The output helps the intended person take the next step
Completeness	It includes the key context, constraints, and edge cases
Fit	Tone, format, and depth match the audience
Risk control	Sensitive claims, actions, and automations are reviewed before use

Use this prompt:

```
Score this output using the rubric below.

For each category, give:
- Pass or Fail
- 1 to 5 score
- Specific reason
- What must change to pass

Rubric:
Accuracy, Usefulness, Completeness, Fit, Risk Control

Output:
[paste output]
```

Anthropic's eval guidance recommends combining grader types and using human judgment for subjective or expert tasks, while automated checks can help with reproducibility and speed (Anthropic agent evals). For everyday use, that means AI can help grade, but you still approve the final risk.

Step eight: add zero-trust before downstream action

The moment AI output moves from text into action, the risk changes. If AI output is passed into a website, database, email system, CRM, shell command, invoice, codebase, or automation, treat it like untrusted input.

OWASP's improper output handling guidance says LLM output should be validated, sanitized, and handled before it is passed downstream, and recommends treating the model with a zero-trust approach (OWASP LLM05:2025 Improper Output Handling).

Practical zero-trust rules:

- **Do not execute AI-generated code without review and testing.**
- **Do not let AI send, delete, purchase, publish, or modify records without approval.**
- **Do not paste AI-generated SQL, shell commands, or API calls into production blindly.**
- **Do not let AI output become customer-facing copy without tone and claim review.**
- **Do not let AI summarize private data into public spaces without privacy review.**

For automations, the approval rule is simple:

Action type	Default approval rule
Draft only	AI can create without approval
Internal note	Human reviews if factual or sensitive
Customer message	Human approves before send
Public content	Human approves before publish
Money movement	Human approves every time
Data deletion	Human approves every time
System/code change	Tests plus human approval

Step nine: build a failure log

Every AI mistake is training data for your future process. Do not just fix the output and move on. Save the failure.

Use this failure log:

Date	Task	What failed	Why it failed	New rule
2026-01-08	Pricing research	Used outdated price	Did not require current source	Add “source date required”
2026-01-12	Sales email	Too aggressive	No tone check	Add audience/tone rubric
2026-01-15	Automation	Wrong field updated	No test record	Require sandbox test first

OpenAI recommends logging during development so teams can mine logs for eval cases, and Anthropic recommends starting evals with 20 to 50 tasks drawn from real failures (OpenAI evaluation best practices, Anthropic agent evals). The everyday version is a simple mistake log that becomes your checklist.

Step ten: create your approval language

Before you trust AI output, force a clear decision:

- **Approved**: Ready to use.
- **Approved with caveats**: Safe if the caveats are included.
- **Needs revision**: Useful direction, but not ready.
- **Rejected**: Too wrong, risky, vague, or unsupported.

Use this prompt:

```
Give me an approval memo for this AI output.

Return:
- Decision: Approved, Approved with caveats, Needs revision, or Rejected
- Top three risks
- Claims that still need proof
- Assumptions the user must confirm
- Final changes required before use
- Whether a human expert should review this before action
```

This is the difference between “AI gave me an answer” and “I made a controlled decision.”

The 30-minute setup

If you only have 30 minutes, build the lightweight version:

1. Create a note called **AI Quality System**.
2. Add the five-part rubric: Accuracy, Usefulness, Completeness, Fit, Risk Control.
3. Add the claim/source/action table.
4. Add the failure log.
5. Save the claim extractor prompt.
6. Save the approval memo prompt.
7. Use it on the next important AI output before you send, publish, buy, code, or automate anything.

That is enough to create a real quality loop. You can improve it later.

The one-page checklist

Before trusting AI output, ask:

- What is this output supposed to do?
- Who will rely on it?
- What happens if it is wrong?
- Which claims need proof?
- Are the sources real and relevant?
- Does the logic actually follow?
- Is any advice missing context?
- Is the tone right for the audience?
- Is this going into a customer, public, financial, legal, technical, or automated workflow?
- Does a human need to approve it first?
- What failure should be saved for next time?

If you cannot answer those questions, the output is not ready. It may still be useful, but it is not ready to trust.

The Founders AI takeaway

The people who win with AI will not be the people who believe every output. They will be the people who build systems that let AI move fast without letting unchecked work move too far.

Use AI to create more. Use a quality system to decide what deserves to ship.