Ethical Prompting

The AI Governance Playbook

> **One-liner:** Design a practical AI governance framework for a team or project — covering when to use AI, how to verify outputs, and what requires human judgment.

? Jump in (Tinkerers start here)

Pick a real team, project, or organization you work with. You're going to design an AI usage framework they could actually adopt.

Step 1 — Map the AI touchpoints. Send this prompt:

> I'm designing an AI governance framework for a **[team type/project type]** that does **[describe the work]**. Map out all the places where team members might use AI in their workflow. For each touchpoint, classify the risk level:
>
> - **Low risk:** AI errors are easily caught and consequences are minor (e.g., drafting internal emails, brainstorming)
> - **Medium risk:** AI errors could waste significant time or create confusion (e.g., research summaries, data analysis, first drafts of client deliverables)
> - **High risk:** AI errors could cause reputational, legal, or financial harm (e.g., published content, financial recommendations, legal language, customer-facing decisions)
>
> Present this as a table with: Touchpoint | Description | Risk Level | Why

Step 2 — Design the verification tiers. Based on the risk map, create a tiered verification system:

> Based on the risk map above, design a 3-tier verification system:

> **Tier 1 (Low risk):** What's the minimum verification needed? What can proceed without review? **Tier 2 (Medium risk):** What checks are required? Who reviews? What's the turnaround expectation? **Tier 3 (High risk):** What's the full review process? Who signs off? What documentation is needed?
>
> For each tier, specify:
>
> - Verification steps (checklist)
> - Who is responsible
> - What "approved" looks like
> - What happens when issues are found

Step 3 — Write the team guidelines. Now produce the actual document:

> Write a 1-page "AI Usage Guidelines" document for this team. It should be practical, not corporate. Include:
>
> 1. **When to use AI** — Green light scenarios
> 2. **When to be careful** — Yellow light scenarios with required verification
> 3. **When NOT to use AI** — Red light scenarios or scenarios requiring explicit approval
> 4. **Verification standards** — The tier system from above, simplified
> 5. **Attribution** — When and how to disclose AI usage
> 6. **Escalation** — What to do when you're unsure whether AI use is appropriate
>
> Write it in the tone of a senior colleague giving practical advice, not a legal department issuing mandates.

Step 4 — Red-team the framework. Test it:

> Now role-play as a team member who wants to use AI in a gray area. Come up with 3 realistic scenarios where the guidelines are ambiguous or where a reasonable person might interpret them differently. For each scenario, suggest how to clarify the guideline.

Revise the guidelines based on the edge cases.

? Plan first (Planners start here)

Here's what you're about to do:

1. **Choose your context** — Pick a real team or project. The framework should be one you could actually share or implement.
2. **Map AI touchpoints and risk levels** — Identify every place AI could be used in the workflow and classify each by potential harm from errors.
3. **Design tiered verification** — Create different verification processes for different risk levels. Not everything needs the same scrutiny.
4. **Write the guidelines** — Produce a practical 1-page document that a team member could reference in their daily work.
5. **Red-team with edge cases** — Test the framework against ambiguous scenarios. Fix any gaps before sharing.

"Done" looks like: A complete, practical AI governance framework (risk map + tiered verification + 1-page guidelines) that you could present to your team, plus documentation of edge cases you tested against.

? Why this matters (Strategists start here)

Individual verification habits (from EP-Intermediate-01) don't scale to teams. When five people use AI differently with different standards, the team's AI output quality is only as good as the weakest link. A governance framework creates **shared standards without bureaucracy** — it tells people what's safe to do quickly and what requires care, without making every AI interaction feel like a compliance exercise. This is also the document organizations will pay for: a practical, calibrated AI usage policy that actually gets followed.

Reflection

- Which risk classification was hardest to assign? What does that ambiguity tell you?
- Would your team actually follow these guidelines? What would make them ignore it?
- Did the red-teaming step reveal fundamental gaps, or just edge cases?
- □□ *Present your framework to a colleague and ask: "Would you follow this?" Their honest reaction is more useful than any AI review.* (Social Learners)

?? Level up

You've reached the advanced level for Ethical Prompting & Judgment. From here, consider:

- Presenting this framework to your actual team and iterating based on feedback
- Combining this with AC-Advanced-01 to add governance to multi-agent workflows
- Building a case study of how the framework changed AI usage behavior in your team

Back to Ethical Prompting & Judgment

The Fact-Check Habit

> **One-liner:** Catch an AI making a confident mistake — and build a simple verification process you'll use every time.

? Jump in (Tinkerers start here)

Pick a topic you know well — your industry, your hobby, your area of expertise. Something where you can spot errors.

Step 1 — Get a confident answer. Send this prompt:

> Give me a detailed overview of **[topic you know well]**. Include specific facts, statistics, and examples. Be thorough and authoritative.

Read the output carefully. **Find at least one claim that feels off.** It might be a statistic that seems too round, a date that feels wrong, a name that's slightly off, or a causal claim that oversimplifies reality.

Step 2 — Make the AI check itself. Send this:

> Look at your previous response. I want you to fact-check yourself. For each specific claim, statistic, or example you cited:
>
> 1. Rate your confidence (high / medium / low)
> 2. Flag anything you might have fabricated or estimated
> 3. Identify which claims are most likely to be wrong and why
>
> Be ruthlessly honest. I'd rather know what you're uncertain about than have you defend everything.

Step 3 — Verify. Pick the 2-3 claims the AI flagged as lowest confidence. Google them. Were they accurate, close but wrong, or completely fabricated?

Step 4 — Build your check. Based on what you just learned, write a 3-line "verification prompt" you can append to any AI output:

> “ Before I use this, tell me:
>
> 1. Which specific claims are you least confident about?
> 2. What did you estimate or approximate vs. know with certainty?
> 3. What should I verify independently before sharing this?

Save this somewhere you'll see it. Use it as a default follow-up to any AI output you plan to rely on.

? Plan first (Planners start here)

Here's what you're about to do:

1. **Choose a familiar topic** — You need to be able to spot errors, so pick something in your area of knowledge. Don't use an unfamiliar topic — you won't know what to verify.
2. **Generate an authoritative-sounding response** — Ask AI for a detailed, factual overview. The more specific and confident the output, the more likely it contains subtle errors.
3. **Ask AI to fact-check itself** — Use the self-audit prompt to force the AI to rate its own confidence and flag potential fabrications.
4. **Independently verify** — Pick the lowest-confidence claims and check them against reliable sources. Track what was right, close, and wrong.
5. **Create your verification template** — Build a reusable 3-question follow-up that you'll use after any AI output you plan to act on.

"Done" looks like: You've caught at least one AI error, you understand *why* the AI got it wrong, and you have a saved verification prompt you can use going forward.

? Why this matters (Strategists start here)

The community's Ethical Prompting score is 75% — the highest of all five pillars. Most people *know* they should verify AI output, but few have a systematic process for doing so. This exercise closes the gap between awareness and practice by giving you a concrete, reusable tool. The verification prompt you build here becomes a habit — a 30-second step that catches errors before they become problems. At the intermediate level, you'll build a more comprehensive verification

checklist; this exercise establishes the baseline behavior.

Reflection

- What type of error did the AI make — a fabricated fact, a wrong date, or a subtle logical leap? Does the category matter for how you'd catch it?
- Did the AI's self-assessment match what you found when you verified manually? Was it too confident, too cautious, or well-calibrated?
- Will you actually use your verification prompt going forward? What would make it stick as a habit vs. something you forget about?
- □□ *Share a specific AI error you caught with a colleague. Ask them how they currently verify AI output — you may discover they don't.* (Social Learners)

?? Level up

Ready for more? Try EP-Intermediate-01 — where you'll build a comprehensive verification checklist and stress-test it against real AI outputs.

Back to Ethical Prompting & Judgment

The Verification Checklist

> **One-liner:** Build a personal AI verification system — a checklist you'll actually use — and stress-test it against real AI outputs to find its limits.

? Jump in (Tinkerers start here)

You're going to build a verification checklist, then immediately try to break it.

Step 1 — Generate something to verify. Ask AI to produce a piece of content you might actually use in your work:

> Write a **[deliverable type — e.g., client email, project proposal, market analysis, technical recommendation]** about **[topic relevant to your work]** . Make it detailed and specific. Include data points, recommendations, and reasoning.

Step 2 — Build your checklist. Before reading the output carefully, write your own verification checklist. Start with these categories and add your own:

Check	Question	Pass/Fail
Factual claims	Are specific numbers, dates, or statistics verifiable?	
Sources	Could I find the original source for any cited information?	
Reasoning	Does the logic hold? Are there hidden assumptions?	
Completeness	What important perspective or consideration is missing?	
Tone/audience	Is the tone appropriate? Would the intended audience trust this?	
Actionability	Are the recommendations specific enough to actually follow?	
Your domain check	[Add a check specific to your field]	

Check	Question	Pass/Fail
Your domain check	[Add another check specific to your field]	

Step 3 — Apply the checklist. Go through the AI output line by line using your checklist. Mark each check as pass or fail. For every fail, note what the issue is.

Step 4 — Stress-test the checklist. Now deliberately ask AI to produce something harder to verify:

> Write the same type of **[deliverable]** but on a topic I'm less familiar with: **[topic outside your expertise]**. Make it equally detailed and authoritative.

Apply your checklist again. Where does it fail to catch problems? What check do you need to add?

Step 5 — Finalize. Update your checklist based on what you learned. Save it where you'll actually use it — bookmark it, pin it, print it, whatever works.

? Plan first (Planners start here)

Here's what you're about to do:

1. **Generate test content** — Ask AI to produce a work-relevant deliverable. This gives you realistic material to verify.
2. **Draft your checklist** — Build a structured verification list covering factual accuracy, reasoning quality, completeness, tone, and domain-specific concerns.
3. **Apply to familiar territory** — Use the checklist on AI output about a topic you know. This lets you calibrate how well the checklist catches real errors.
4. **Apply to unfamiliar territory** — Use the checklist on AI output about a topic you *don't* know well. This exposes gaps in your process — the errors you can only catch with domain knowledge.
5. **Iterate and save** — Update the checklist based on what it missed. Save it in a format you'll actually reach for.

"Done" looks like: A tested, refined verification checklist (8-12 items) saved in a usable format, with evidence of at least one error it caught and one gap you identified and fixed.

? Why this matters (Strategists start here)

In EP-Basic-01, you built a simple 3-question verification prompt. Here, you're building a **systematic process** — a checklist that works regardless of topic, catches both factual and reasoning errors, and is tuned to your specific work context. The community's 75% Ethical Prompting score means most people *intend* to verify AI output but lack a consistent method. A checklist turns good intentions into reliable behavior. At the advanced level, you'll scale this into a governance framework for a team; this exercise builds the individual practice first.

Reflection

- Which check caught the most problems? Which was least useful?
- How did your verification experience change between the familiar topic and the unfamiliar one?
- Is your checklist something you'd actually pull up before sending an AI-generated deliverable? What format makes it most likely you'll use it?
- □□ *Trade checklists with a colleague. Have them apply yours to an AI output from their work — their feedback will reveal blind spots specific to your domain.* (Social Learners)

?? Level up

Ready for more? Try EP-Advanced-01 — where you'll design an AI governance framework for a team or project.

Back to Ethical Prompting & Judgment