M A V E R I C K A I

THE TRUTH PROMPT

STOP AI FROM CONFIDENTLY MAKING THINGS UP

One set of instructions that forces Claude (or ChatGPT, or Gemini) to separate fact from guess, score its own confidence, and tell you when it does not know.

The full Truth Prompt, ready to copy and paste

Setup for Claude, ChatGPT, Gemini, and Claude Code

Why AI makes things up in the first place

How to read the confidence score without being fooled

8 follow-up prompts that force it to stay honest

The mistakes that quietly turn the prompt off

A one-page cheat sheet you can keep open

2026 Edition

The Truth Prompt

Here it is. Copy it, paste it into your AI's custom instructions, and read the rest of the guide after. The whole thing is one block. Do not shorten it.

THE TRUTH PROMPT (COPY ALL OF IT)

```
Adopt the role of a rigorous, truth-seeking
reasoning assistant.

For every complex or factual request, follow
this process:

# 1. DECOMPOSE
Break the request into smaller questions or
claims that can be evaluated independently.

# 2. DISTINGUISH
Clearly separate:
- Verified facts
- Reasonable inferences
- Assumptions
- Opinions
- Unknown or missing information

# 3. SOLVE
Address each part carefully. Do not invent
facts, sources, quotes, statistics, links, or
details to fill gaps.

# 4. VERIFY
Before answering, check:
- Logical consistency
- Factual accuracy
- Whether the answer fully addresses the
request
- Whether important context is missing
- Whether bias or unsupported assumptions may
be affecting the answer

# 5. CALIBRATE CONFIDENCE
Assign a confidence score from 0.0 to 1.0
based on the quality of the available
evidence, not on how persuasive the answer
sounds.
```

```
# 6. RETRY WHEN NEEDED
If confidence is below 0.8:
- Identify the weakest claims
- Reconsider the reasoning
- Revise the answer
- Ask for clarification or state what
information is needed if the uncertainty
cannot be resolved

# 7. BE HONEST ABOUT UNCERTAINTY
Never present an assumption, prediction, or
uncertain claim as a confirmed fact. If
something cannot be verified, say so directly.

# OUTPUT FORMAT
For every final response, use this format:

Clear Answer:
Provide the most accurate and useful answer
possible.

Confidence Level:
Give an overall confidence score from 0.0 to
1.0 and briefly explain it.

Key Caveats:
List any assumptions, uncertainties, missing
information, conflicting evidence, or facts
that require verification.
```

DO THIS NOW

Open Claude. Click your name in the bottom left, then Settings, then find "Instructions for Claude." Paste the prompt in. Hit save. That is it, it now applies to every new chat.

Why AI Confidently Makes Things Up

AI does not lie the way a person lies. There is no intent. What actually happens is simpler and, in some ways, worse.

A language model is a prediction engine. It looks at everything in front of it and predicts what text should come next. That is the entire job. It is not looking things up in a database. It is not checking a source. It is producing the most plausible continuation of a sentence.

Here is the problem. A sentence that is confidently wrong is just as plausible as a sentence that is confidently right. "The study was published in 2019 by researchers at Stanford" reads perfectly whether that study exists or not. The model has no internal alarm that goes off when it crosses from recall into invention. Both feel the same to it.

THE ACTUAL TERM: HALLUCINATION

When an AI produces content that sounds right but is not grounded in anything real, researchers call it a hallucination. It shows up most often as fake citations, invented statistics, made-up URLs, quotes nobody said, and function names that do not exist in a library. The output is fluent. That is exactly what makes it dangerous.

The three failure modes you actually run into

1. Gap filling

You ask a question, the model knows 80 percent of the answer, and it smooths over the missing 20 percent instead of flagging it. You never see the seam.

2. Confidence without evidence

The model's tone is the same whether it is certain or guessing. There is no verbal hedge, no wobble. You read confidence and assume accuracy. They are unrelated.

3. Agreeing with you

Models are trained on human feedback, and humans rate agreeable answers higher. So the model drifts toward telling you what you want to hear. Push back on a correct answer and watch it fold. That is not the model changing its mind, that is it optimizing for your approval.

HEADS UP

You will see viral claims that AI "lies X percent of the time." Treat any single number with suspicion. Hallucination rates swing wildly depending on the model, the topic, whether web search is on, and how the question is asked. The rate is not zero, and that is the only part you need to act on.

What The Prompt Actually Does

The Truth Prompt does not give the model new knowledge. It cannot. What it does is change the model's process before it answers, and force it to show you its work afterward.

Think of it as the difference between asking someone a question in the hallway versus asking them to write you a memo. Same person, same knowledge. Wildly different rigor.

The seven steps, and why each one is there

Step	What it forces	Why it matters
1. Decompose	Split the request into separate claims	A wrong sub-claim hides inside a right-sounding answer. Splitting it exposes the weak link.
2. Distinguish	Label each piece: fact, inference, assumption, opinion, or unknown	This is the whole game. Most bad AI answers blend all five into one confident paragraph.
3. Solve	Answer without inventing sources, stats, quotes, or links	An explicit ban on gap filling. Naming the exact things it fabricates works better than "be accurate."
4. Verify	Self-check logic, facts, completeness, and bias	A second pass catches contradictions the first pass wrote in.
5. Calibrate	Score confidence 0.0 to 1.0 based on evidence, not on how good the answer sounds	Separates "I am sure" from "this reads well." Those are different things.
6. Retry	If confidence is under 0.8, find the weakest claim and redo it	The model gets a chance to catch itself before you ever see the answer.
7. Be honest	Say "I do not know" out loud	Gives the model explicit permission to not have an answer. Without it, silence feels like failure.

THE OUTPUT FORMAT IS HALF THE VALUE

Clear Answer, Confidence Level, Key Caveats. The caveats section is the one people skip and it is the most useful part of the whole thing. It is where the model tells you what it was unsure about, what it assumed, and what you should go verify yourself.

How To Install It Everywhere

Do this once per tool and it applies to every conversation from then on. You never have to paste it again.

STEP 1: Claude (web and desktop app)

Click your name in the bottom left corner. Go to Settings. Find the section called "Instructions for Claude" (also called personal preferences). Paste the full prompt in and save.

If you want it on one project only rather than everywhere, open a Project instead, click "Set project instructions," and paste it there. That keeps your casual chats casual and your research chats rigorous.

STEP 2: ChatGPT

Click your profile in the bottom left. Go to Settings, then Personalization, then Custom Instructions. Paste the prompt into the box labeled "How would you like ChatGPT to respond?" and save.

ChatGPT's custom instruction field has a character limit. If the full prompt does not fit, use the compressed version in the next chapter.

STEP 3: Gemini

Open Settings, then Saved Info (or Personalization, depending on your version). Add the prompt as a saved instruction. Gemini's memory is less strict than Claude's instructions field, so re-paste the prompt at the top of any high-stakes chat as a backup.

STEP 4: Claude Code (or any coding agent)

Put it in your CLAUDE.md file at the root of your project. Every session reads that file automatically. For coding specifically, add one extra line, because the failure mode is different:

CODING ADD-ON (PASTE UNDER THE MAIN PROMPT)

```
Additional rule for code:
Never invent library functions, API endpoints,
flags, or config keys. If you are not certain a
function exists, say so and tell me to check the
docs. A wrong function name that looks right
costs more time than saying "I do not know."
```

STEP 5: API or system prompt

Drop it in as the system message. Same text, no changes needed. If you are building anything that touches facts, numbers, or citations, this belongs in your system prompt permanently.

TIP

Instructions apply to NEW chats. If you have a conversation already open, close it and start a fresh one, otherwise you will think the prompt did not work.

Short Versions For When You Need Them

The full prompt is the right one for custom instructions. But sometimes you need something that fits in a character limit, or you just want to spot-check one answer without changing your settings. Here are three sizes.

The compressed version (fits any character limit)

SHORT TRUTH PROMPT

```
Be a rigorous, truth-seeking assistant.
For factual requests: break the question into
separate claims. Label each as verified fact,
inference, assumption, or unknown. Never invent
sources, stats, quotes, or links to fill gaps.
Check your answer for logic, accuracy, and
missing context before replying. If your
confidence is under 0.8, revise it or tell me
what you would need to know.

End every factual answer with:
Confidence: 0.0-1.0 (based on evidence, not on
how good the answer sounds)
Caveats: assumptions, unknowns, things to verify
```

The one-liner (paste into any single chat)

ONE-LINE TRUTH CHECK

```
Answer this, then give me a confidence score
from 0.0 to 1.0 and list every assumption or
unverified claim you made. Do not invent
sources or numbers. If you do not know, say so.
```

The audit (run it on an answer you already got)

AUDIT AN EXISTING ANSWER

```
Re-read your last answer as a skeptical
fact-checker who wants to find errors.

List every factual claim you made.
For each one, mark it: VERIFIED, LIKELY,
```

```
UNCERTAIN, or FABRICATED.

Then rewrite the answer keeping only what

survives.
```

TIP

The audit prompt is the highest-leverage one in this guide. Run it on any AI answer you are about to act on. You will be surprised how often something gets downgraded.

How To Read The Confidence Score

This part matters and almost nobody gets it right. The confidence score is useful. It is not a measurement.

The model is not computing a probability. It is estimating one, using the same prediction machinery that produces everything else it says. That means the score is a signal, not a guarantee. A 0.9 is not a promise. It is the model saying "this feels well-supported to me."

So why bother? Because the relative movement is informative. When a model that normally hands you 0.9 suddenly hands you 0.55, something in that answer is shaky, and it just told you. That is worth a lot.

What the numbers should mean to you

Score	What it usually means	What you should do
0.9 - 1.0	Well-established, widely documented, unlikely to be contested	Proceed. Still verify anything with money, health, or legal consequences.
0.75 - 0.9	Solid, but with assumptions baked in or some detail the model is fuzzy on	Read the caveats section carefully. That is where the soft spot is.
0.5 - 0.75	The model is reasoning, not recalling. Real uncertainty	Do not act on this without a second source. Ask it what would change its mind.
Below 0.5	It is guessing and admitting it	Treat as a starting point for your own research, nothing more.

HEADS UP

A high confidence score on a hallucinated fact is still possible. The prompt reduces this, it does not eliminate it. If an answer will cost you money, time, or credibility when wrong, verify it against a real source. The prompt is a filter, not a guarantee.

The follow-up that breaks a fake high score

PRESSURE-TEST A CONFIDENT ANSWER

```
You gave that a high confidence score. Now
argue against yourself.

What is the strongest case that your answer is
wrong? What would I find if I checked your
sources and they did not exist? Where is the
weakest link in your reasoning?

Then re-score your confidence.
```

8 Follow-Ups That Force It To Stay Honest

Custom instructions set the baseline. These are what you type in the moment, when an answer feels a little too smooth.

1. The source check

```
For every source, statistic, and quote you just
used: do you actually know this exists, or are
you reconstructing what a plausible source would
say? Be specific about which is which.
```

2. The anti-agreement check

```
I think you changed your answer because I
pushed, not because I was right. Were you
correct the first time? Tell me honestly.
```

3. The missing context check

```
What did you leave out of that answer that a
real expert would have included? What would they
say you got dangerously wrong?
```

4. The disconfirming evidence check

```
Give me the strongest evidence AGAINST what you
just told me. Do not soften it.
```

5. The blank-slate check

```
Forget everything I told you I wanted. If you
had no idea what answer I was hoping for, what
would you say?
```

6. The verification path

```
Do not tell me the answer. Tell me exactly how I
would verify it myself, and what I should search
for. Then give the answer.
```

7. The knowledge boundary check

```
Is this something you actually have reliable
training data on, or is it recent, niche, or
obscure enough that you are extrapolating? Be
specific about where your knowledge thins out.
```

8. The stakes check

```
I am going to act on this. Money and reputation
are on the line. Given that, what do you want to
change or caveat about your answer?
```

TIP

Number 2 is the one you will use most. Models fold under pressure. If you push back and the AI immediately agrees with you, that is a red flag, not a win.

The Mistakes That Turn The Prompt Off

The prompt works. But there are specific things people do that quietly cancel it out, and then they conclude it did not work.

Pasting it into an old chat

Custom instructions only load at the start of a conversation. If your chat was already open when you saved the prompt, it is not running. Start a new one.

Trimming it down to "be accurate"

"Be accurate and do not hallucinate" does almost nothing. The model already thinks it is being accurate. What works is the specific list of things it is banned from inventing (sources, quotes, statistics, links) and the required output format. Vague instructions get vague compliance.

Ignoring the caveats section

People read the Clear Answer, glance at the confidence number, and skip the caveats. The caveats are the payload. That is where the model tells you what it is unsure about. Skipping them means you installed a smoke detector and unplugged it.

Treating 0.9 as verified

It is not verified. It is the model's self-assessment. For anything that matters, you still check.

Leaving it on for creative work

If you are brainstorming, writing fiction, or riffing on ideas, this prompt makes the AI stiff and hedgy. It will caveat a poem. Turn it off or use a Project so it only runs where you need it.

Assuming it stops all hallucination

It does not. Nothing does, right now. It makes hallucination less frequent and, more importantly, more visible. Visible is the win. A wrong answer that says "confidence 0.6, I am extrapolating here" is a wrong answer you will not act on.

HEADS UP

The single biggest gain from this prompt is not fewer errors. It is that the errors now announce themselves. Do not skip the confidence score and caveats, they are the entire point.

When To Use It, When Not To

Rigor has a cost. The answers get longer, more hedged, and slower. That is a great trade for research and a terrible one for a quick idea. Know which mode you are in.

Turn it ON for	Turn it OFF for
Research and fact-finding	Brainstorming and idea generation
Anything with statistics or numbers	Creative writing, fiction, poetry
Medical, legal, or financial questions	Casual conversation
Writing you will publish under your name	First drafts you will heavily rewrite
Code you will ship	Rough scratch code and experiments
Decisions with real money attached	Rewording, summarizing, formatting
Anything you will cite to someone else	Tasks where you already know the answer

THE PROJECT TRICK

In Claude, create one Project called "Research" with the Truth Prompt as its project instructions, and leave your global settings clean. Now you get rigor when you want it and a normal, fast Claude the rest of the time. Same idea works in ChatGPT with a custom GPT.

Pair it with search

The prompt makes the model honest about what it does not know. Web search gives it a way to actually find out. Together they are far stronger than either alone. If your tool has search or browsing, turn it on for anything factual and recent, and the confidence scores will get noticeably more useful.

Test It Yourself In Two Minutes

Do not take my word for it. Run this. It is the fastest way to see the difference.

STEP 1: Ask an impossible question WITHOUT the prompt

Open a normal chat and ask something the model cannot possibly know. Something niche, recent, or fully invented.

THE BAIT

```
Summarize the key findings of the 2024 Hartwell
study on remote work and team cohesion.
```

There is no Hartwell study. It does not exist. Watch what a lot of models do anyway: they produce a clean, confident, three-bullet summary with plausible numbers attached.

STEP 2: Install the Truth Prompt

Custom instructions. Save. Start a fresh chat.

STEP 3: Ask the exact same question

Now you should get something closer to: "I have no record of a 2024 Hartwell study on this topic. I may be missing it, or the study may not exist. Confidence: 0.2. Caveat: I am not able to verify this citation and will not summarize a paper I cannot confirm."

THAT DIFFERENCE IS THE WHOLE PRODUCT

Same model. Same knowledge. Same question. One version made something up. The other told you it did not know. The prompt did not make the AI smarter, it made it honest about the edge of what it knows.

TIP

Try the bait question again with a topic in your own field, where you will instantly spot a fake answer. It is a lot more convincing when you catch it yourself.

The One-Page Cheat Sheet

Everything in this guide, compressed. Screenshot this page.

Setup

Tool	Where the prompt goes
Claude	Your name (bottom left) > Settings > Instructions for Claude
Claude (one project only)	Open a Project > Set project instructions
ChatGPT	Profile > Settings > Personalization > Custom Instructions
Gemini	Settings > Saved Info
Claude Code	CLAUDE.md at the project root
API	System prompt

The seven steps

Decompose the question. **Distinguish** fact from inference from assumption from unknown. **Solve** without inventing anything. **Verify** logic, facts, completeness, bias. **Calibrate** confidence 0.0 to 1.0. **Retry** if under 0.8. **Be honest** about what it does not know.

The output you should now see every time

```
EXPECTED FORMAT

  Clear Answer:

  ...

  Confidence Level:

  0.0 to 1.0, with a reason

  Key Caveats:

  Assumptions, unknowns, things to verify
```

The three follow-ups worth memorizing

1. "Which of those sources do you actually know exist, and which are you reconstructing?"
2. "Did you change your answer because I pushed, or because I was right?"
3. "Argue against yourself. Where is the weakest link?"

YOUR 5-MINUTE ACTION PLAN

1. Paste the full prompt into Claude's Instructions for Claude and save. 2. Start a NEW chat. 3. Ask it the fake-study bait question and watch it refuse. 4. Save the audit prompt somewhere you can grab it. 5. From now on, read the Key Caveats section before you act on anything.

Want More AI Tips Like This?

I post daily AI content, tool breakdowns, and prompts that actually help you get more done. Follow me for the latest.

@MavGPT

TikTok | Instagram | Facebook

An AI that says "I do not know" is worth ten that sound sure.