When AI Agrees With You

A Practical Guide to Sycophancy

Bob Mazzei

Sienda ltd

2026

Copyright Notice

© 2026 Bob Mazzei. All rights reserved.

Written by Bob Mazzei and published by Sienda Ltd, London, United Kingdom.

This guide is made available free of charge for personal, educational and professional reference.

You may download and share the complete, unmodified guide, provided that the author, copyright notice and publisher attribution remain intact.

You may not, without prior written permission:

- ➔ modify, adapt or reproduce substantial portions of this guide;
- ➔ republish its contents as your own work;
- ➔ sell, licence or otherwise commercially distribute the guide;
- ➔ incorporate substantial portions of it into another publication, course, training programme, product or commercial service.

Brief quotations are permitted for commentary, review, discussion or citation, provided appropriate attribution is given to the author and source.

The fact that this guide is provided free of charge does not place it in the public domain and does not constitute a waiver of copyright.

Bob Mazzei

London, 2 October 2026

sienda.co.uk

Table of contents

Introduction	4
What is AI sycophancy?	4
The confirmation-bias machine	6
A chatbot is not an independent consultant	7
Why does sycophancy happen?	7
Sycophancy is not always agreement	8
The more context you provide, the more careful you must become	9
The conversational loop makes the problem stronger	10
Plausibility is not evidence	11
General-purpose AI is not the problem	12
A simple example	12
The five stages of AI sycophancy	13
Prompt engineering alone does not solve this	14
How to reduce sycophancy	14
A useful anti-sycophancy prompt	17
The real skill is not prompting	17
Use AI as an intellectual sparring partner	18
Back to the website analysis	18
The uncomfortable conclusion	19
From Chatbot to Professional Tool: Disciplining the Model	20
Constraints are a feature, not a limitation	22
The model becomes a component	24
From prompt engineering to system engineering	25

Introduction

A few days ago, a client — who is also a long-time friend — sent me an analysis of his websites that he had produced using ChatGPT.

He had clearly done a good job. The structure of the analysis showed that he had acquired considerable experience in working with AI and, in particular, in constructing prompts that produced detailed and apparently well-reasoned answers.

But several points in the analysis were inaccurate.

Not absurd. Not obviously wrong. In fact, that was precisely the problem.

They were plausible, professionally expressed and presented with enough confidence to make them easy to accept.

And in several cases, they appeared to reinforce conclusions that had already been implicit in the questions.

This is where one of the less obvious problems of conversational AI begins.

It has a name: **sycophancy.**

What is AI sycophancy?

In ordinary language, a sycophant is someone who tells another person what they want to hear in order to gain their approval.

Something similar can happen with AI chatbots.

In AI research, sycophancy generally refers to a model adapting its answer to the beliefs, assumptions or preferences expressed by the user, even when doing so makes the answer less truthful or less accurate.

Research has found that language models can sometimes favour answers that agree with a user's stated position over answers that are objectively better supported. One reason is that human evaluators themselves tend to prefer responses that appear agreeable, persuasive and aligned with their own views.

The important point is that sycophancy is not limited to obvious flattery.

It does not necessarily look like:

> “You're absolutely right.”

It can be much subtler.

It can look like:

> "Your analysis correctly identifies the three main weaknesses..."

or:

> "Given the problems you've identified, the best approach would be..."

or simply:

> "Yes, that interpretation makes sense."

The AI may then construct an elaborate argument around a premise that should first have been questioned.

The result can be logically coherent, beautifully written — and wrong.

This is different from an ordinary AI mistake

People often put all AI errors into the same category.

They should not.

There are several different failure modes.

An AI can hallucinate: invent a fact, source, quotation, statistic or event that does not exist.

It can reason badly: start with correct information but reach an incorrect conclusion.

It can suffer from missing context: provide a reasonable answer based on the information available while lacking something essential.

And it can be sycophantic: allow the user's assumptions or desired conclusion to influence the analysis too strongly.

These problems can overlap, but they are not identical.

Sycophancy is particularly interesting because the user participates in creating the error.

The AI is not simply giving us incorrect information.

We may inadvertently be teaching it what answer we want.

The confirmation-bias machine

Humans already have confirmation bias.

We naturally tend to notice information that supports what we believe and discount information that contradicts it.

Conversational AI can amplify this tendency.

Imagine that I ask:

> "Why is our website failing to communicate our expertise?"

I have already embedded a conclusion in the question.

I have told the AI that the website is **failing**.

The model may accept that premise and begin explaining why.

Perhaps:

- ➔ the positioning is unclear;
- ➔ the homepage is too generic;
- ➔ the value proposition is weak;
- ➔ the navigation is confusing;
- ➔ the copy lacks authority.

The analysis might sound extremely convincing.

But we never established that the website was actually failing.

Compare that with:

> "Evaluate how effectively this website communicates the company's expertise. Identify evidence of both strengths and weaknesses, and do not assume that there is a problem."

That is a very different task.

The first prompt asks the AI to explain an assumed failure.

The second asks it to investigate whether a failure exists.

That distinction is fundamental.

A chatbot is not an independent consultant

This is one of the most important concepts to understand when using conversational AI professionally.

When a human consultant receives a brief saying:

> "Our website isn't working because the message isn't clear."

a good consultant may reply:

> "How do you know?"

The consultant can challenge the premise.

Perhaps the messaging is perfectly clear and the actual problem is poor traffic acquisition.

Perhaps enquiries are low because prices have risen.

Perhaps competitors have changed.

Perhaps the website performs very well with existing customers but poorly with first-time visitors.

Perhaps there is no evidence yet that the website is underperforming at all.

An AI system can also challenge assumptions.

But unless the interaction encourages it to do so, it may instead work within the framework supplied by the user.

This is an important distinction.

A chatbot is highly responsive to context.

That responsiveness is one of its great strengths.

It is also one of its weaknesses.

Why does sycophancy happen?

Modern AI assistants are not databases that retrieve a single objectively correct answer.

They are language models trained to generate useful responses from patterns learned from enormous amounts of data and subsequently refined to behave as helpful conversational assistants.

Part of that refinement historically involved human preference feedback.

People rated answers.

Models were then trained, directly or indirectly, to produce responses people were more likely to prefer.

But human preferences are imperfect.

People tend to like answers that:

- ➔ understand them;
- ➔ validate their concerns;
- ➔ respect their assumptions;
- ➔ sound confident;
- ➔ are pleasant to read;
- ➔ appear helpful;
- ➔ avoid unnecessary confrontation.

Usually, these are desirable characteristics.

But they can create a tension between two objectives:

being agreeable and **being correct**.

Research into language-model behaviour has found evidence that human feedback can favour responses matching the user's beliefs, helping explain why sycophancy emerges.

The industry is well aware of the problem.

In April 2025, OpenAI rolled back an update to GPT-4o after users and internal analysis found that the model had become excessively agreeable. OpenAI described the behaviour as overly supportive and disingenuous and subsequently added further work specifically aimed at reducing sycophancy.

So this is not a theoretical criticism of AI.

It is a recognised technical problem.

Sycophancy is not always agreement

There is another subtlety.

Sometimes an AI should agree with you.

If you say:

> "The invoice total is wrong because £450 + £300 is £750, not £850."

agreement is appropriate.

The objective is not to create an AI that contradicts everything.

The goal is independent evaluation.

A useful AI should agree when the evidence supports agreement and disagree when the evidence supports disagreement.

It should also be able to say:

> "There is not enough information to determine that."

That third answer is particularly important.

Many poor analyses are not wrong because the opposite conclusion is true.

They are wrong because the available evidence does not justify any strong conclusion.

The more context you provide, the more careful you must become

There is a paradox here.

We are often told that better prompts require more context.

Generally, that is true.

If an AI understands your business, objectives, customers and constraints, it can produce far more useful work.

But context is not neutral.

Suppose I explain:

> "We have always struggled because our competitors have much stronger brands. Our website doesn't communicate our expertise properly, our social-media activity is weak, and customers probably don't understand what differentiates us."

I have provided plenty of context.

Unfortunately, I have also supplied several unverified conclusions.

The model may build its entire analysis on them.

What feels like a sophisticated consultation can therefore become something closer to an elaboration of the user's existing beliefs.

More prompting does not automatically produce more objectivity.

Sometimes it merely produces a more detailed version of the bias already contained in the prompt.

The conversational loop makes the problem stronger

Sycophancy can become more pronounced during long conversations.

Consider this sequence:

User:
Our website feels outdated.

AI:
Yes. The visual language feels somewhat dated, particularly because of X, Y and Z.

User:
Exactly. I also think the homepage contains too much information.

AI:
I agree. The density of information weakens the hierarchy.

User:
So we should simplify everything.

AI:
Yes. A more minimal structure would improve clarity.

After several exchanges, something important has happened.

The conversation now contains a growing body of material treating the original assumption as an established fact.

Each new response inherits the preceding discussion.

The AI is no longer evaluating:

> "Is the website outdated?"

It may effectively be discussing:

> "Given that we have established that the website is outdated, what should we change?"

But perhaps nothing was ever established.

There was simply an initial impression that became progressively reinforced.

Plausibility is not evidence

Large language models are extraordinarily good at explanation.

This creates another danger.

An explanation can be compelling without being true.

Ask why a hypothetical company is losing customers and an AI can probably give you six excellent reasons.

Ask why the same company is gaining customers and it can probably give you another six.

Both sets of arguments may sound convincing.

The ability to construct a plausible explanation is not the same thing as having evidence that the explanation describes reality.

This distinction becomes critical in areas such as:

- ➔ business strategy;
- ➔ financial analysis;
- ➔ legal questions;
- ➔ organisational diagnosis;
- ➔ marketing;
- ➔ recruitment;
- ➔ investment;
- ➔ healthcare;
- ➔ technical architecture;
- ➔ competitive analysis.

The more complicated the subject, the easier it becomes to mistake sophisticated language for validated analysis.

General-purpose AI is not the problem

It would be wrong to conclude that tools such as ChatGPT, Claude or Gemini are unsuitable for serious work.

They can be extraordinarily useful for complex tasks.

But they should not be treated as self-validating authorities.

A general-purpose AI does not automatically know:

- ➔ your organisation;
- ➔ your internal processes;
- ➔ your customers;
- ➔ your data;
- ➔ the political dynamics inside a company;
- ➔ why a particular decision was made;
- ➔ whether information supplied by a user is accurate;
- ➔ whether an assumption embedded in a prompt has actually been demonstrated.

Complex professional work therefore requires something more than prompting ability.

It requires method.

The AI must be supplied with evidence where possible, told what is known and unknown, and explicitly encouraged to challenge the premises of the task.

A simple example

Consider these two prompts.

Prompt A

> Our sales process is inefficient. Analyse why and recommend how AI could automate it.

This contains two assumptions:

1. the sales process is inefficient;
2. automation with AI may be the appropriate solution.

A chatbot may accept both.

Now compare:

Prompt B

> Analyse the sales process described below. Do not assume that it is inefficient. Separate structural problems, process problems, information problems and technology problems. Identify which conclusions are supported by evidence and which require additional information. Consider AI only if it addresses an identified problem better than simpler alternatives.

Prompt B does something fundamentally different.

It asks for diagnosis before treatment.

The five stages of AI sycophancy

In practice, the phenomenon often develops through a recognisable sequence.

1. The user introduces an assumption

> "Our problem is poor positioning."

2. The AI accepts the assumption

> "Your positioning problem appears to come from..."

3. The AI generates supporting arguments

It identifies five reasons why the positioning is weak.

4. The user interprets the detail as confirmation

The quantity and quality of the explanation create confidence.

5. The conversation proceeds as though the original assumption has been proven

Recommendations are developed, priorities established and potentially money spent.

The dangerous step is number two.

The AI should first have asked:

> "What evidence indicates that positioning is the problem?"

Prompt engineering alone does not solve this

There is now a great deal of emphasis on learning to write better prompts.

That is useful.

But a beautifully engineered prompt can still contain a false premise.

In fact, skilled prompting can sometimes make the problem worse.

A detailed prompt can produce a detailed answer.

A highly structured prompt can produce a highly structured answer.

A professional prompt can produce something resembling a professional report.

None of these characteristics guarantees that the underlying reasoning is correct.

Good prompting improves the quality of execution. It does not automatically improve the quality of the premise.

That is a crucial distinction.

How to reduce sycophancy

The solution is not to stop using AI.

The solution is to change the way we interrogate it.

Several techniques are particularly useful.

1. Tell the AI not to assume your premise is correct

Instead of:

> "Explain why this strategy will fail."

use:

> "Evaluate this strategy. Do not assume my concerns are correct."

This simple change can dramatically alter the analysis.

2. Ask explicitly for disagreement

For example:

> "Identify where my reasoning may be wrong."

or:

> "What would a knowledgeable person who disagreed with me argue?"

or:

> "Challenge my assumptions before making recommendations."

You are deliberately giving the model permission to oppose the framing of the question.

3. Separate facts from assumptions

Ask the AI to classify statements as:

- ➔ observed fact;
- ➔ supplied information;
- ➔ inference;
- ➔ assumption;
- ➔ opinion;
- ➔ unknown.

This is particularly powerful in business analysis.

A sentence such as:

> "Customers find the ordering process confusing"

may sound factual.

But unless there is customer research, behavioural data or another source behind it, it may simply be an assumption.

4. Ask what evidence would change the conclusion

A strong analytical question is:

> "What evidence would demonstrate that this conclusion is wrong?"

This forces the reasoning towards falsifiability rather than confirmation.

5. Ask for competing explanations

Instead of asking:

> "Why have sales declined?"

ask:

> "Generate the strongest plausible competing explanations for the sales decline and state what evidence would distinguish between them."

Now the AI is not merely producing a story.

It is helping construct hypotheses.

6. Ask for confidence and uncertainty

Useful instructions include:

> "State where the evidence is insufficient."
>
> "Do not fill gaps with plausible assumptions."
>
> "Indicate your confidence in each conclusion."
>
> "If a conclusion cannot be supported, say so."

The words "I don't know" can be more valuable than three paragraphs of plausible speculation.

7. Start a fresh conversation for critical verification

Long conversations accumulate assumptions.

If an important conclusion has emerged after twenty or thirty exchanges, it can sometimes be useful to open a clean session and present only the underlying evidence.

Then ask the new conversation to analyse it independently.

You may discover that the second analysis reaches a very different conclusion.

8. Verify externally

The warning displayed by AI systems that they can make mistakes should be taken seriously.

Facts should be checked.

Sources should be opened.

Numbers should be recalculated.

Documents should be examined directly.

Claims about competitors should be verified.

Legal, financial, medical and regulatory conclusions should receive appropriate professional scrutiny.

AI can enormously accelerate investigation.

It does not remove the requirement for verification.

A useful anti-sycophancy prompt

For analytical work, the following instruction can be added to many prompts:

> Do not assume that my description, interpretation or proposed conclusion is correct. Treat my statements as hypotheses unless they are supported by evidence. Identify questionable premises, alternative explanations and missing information. Distinguish facts from assumptions and inferences. If the evidence does not justify a conclusion, say so explicitly. Your objective is to help me reach the most defensible conclusion, not to agree with me.

This will not magically eliminate every problem.

But it changes the nature of the conversation.

You are no longer asking for validation.

You are asking for examination.

The real skill is not prompting

There is a tendency to describe people who are good at using AI as people who are good at writing prompts.

That definition is becoming increasingly inadequate.

The more important skill is knowing how to think with AI without allowing the AI to think for you.

That means knowing:

- ➔ what question should be asked;
- ➔ which assumptions are hidden inside it;
- ➔ what evidence is available;
- ➔ what evidence is missing;
- ➔ when the model is reasoning and when it is speculating;
- ➔ when an answer needs external verification;
- ➔ when to challenge the result;
- ➔ when to discard it.

Prompt construction is part of this.

Critical thinking is the larger part.

Use AI as an intellectual sparring partner

Perhaps the healthiest way to use conversational AI is not as an oracle and not as a machine whose job is to confirm our ideas.

Use it as an intellectual sparring partner.

Ask it to attack the proposal.

Then defend it.

Ask it to construct the strongest competing explanation.

Ask it what you have overlooked.

Ask it which claims have evidence behind them.

Ask it which claims merely sound plausible.

Ask what additional information would materially change the analysis.

Then make the judgement yourself.

The objective is not agreement.

The objective is better reasoning.

Back to the website analysis

This brings us back to the analysis my friend sent me.

The interesting part was not that ChatGPT had made several mistakes.

Every analytical tool can produce errors, including humans.

What interested me was the type of error.

The document was structured.

The language was professional.

The arguments were plausible.

The conclusions fitted together.

But in several places, the system appeared to have accepted premises and developed them rather than independently testing whether they were correct.

That is much harder to detect than a blatant factual mistake.

A ridiculous answer invites scepticism.

A sophisticated answer invites trust.

And that is precisely why understanding sycophancy matters.

The uncomfortable conclusion

The biggest danger in using AI may not be that it tells us something obviously false.

We can often catch that.

The more difficult danger is that it takes something we already suspect, turns it into polished analysis, provides five persuasive arguments supporting it and sends it back to us looking like independent confirmation.

We then say:

> "Even the AI agrees."

But perhaps the AI agreed because we told it, directly or indirectly, what we wanted it to agree with.

That is the trap.

The appropriate response is not distrust of AI.

It is a disciplined use of AI.

These systems are enormously powerful tools.

But a tool capable of generating arguments at extraordinary speed must be paired with a user capable of questioning those arguments.

Because the most useful AI is not the one that always tells us we are right.

It is the one we have learned to use in a way that makes it possible to discover when we are wrong.

From Chatbot to Professional Tool: Disciplining the Model

Everything I have discussed here forms part of the work I cover in my courses with businesses. But none of it should be mistaken for a universal solution. Using AI chatbots properly can reduce errors and improve the quality of the output, but it does not eliminate every weakness, limitation or risk associated with their use.

For many professional applications, I do not give a client access to a general-purpose chatbot and simply tell them to write better prompts.

I do something different.

I build software around the AI model.

The distinction is important.

The underlying model may still come from OpenAI, Anthropic, Google or another provider. But instead of allowing the user to interact freely with a general-purpose conversational system, the software determines:

- ➔ what information the model receives;
- ➔ which questions it is allowed to address;
- ➔ which data sources it may use;
- ➔ what it must ignore;
- ➔ how the task is structured;
- ➔ what format the answer must follow;
- ➔ which checks must occur before a result is produced;
- ➔ what the user can and cannot ask the system to do.

In other words, the surrounding software disciplines the model.

The AI remains powerful, but it operates inside a defined environment.

Why this matters

A general chatbot has an enormous possible response space.

If somebody writes:

> Analyse this company and tell me what is wrong with it.

The model has considerable freedom to interpret the question, fill gaps, make assumptions and construct a plausible narrative.

A specialised application can instead require the user to provide specific information first.

For example:

- ➔ product or service;
- ➔ target market;
- ➔ geography;
- ➔ commercial objective;
- ➔ known competitors;
- ➔ customer type;
- ➔ distribution channels;
- ➔ exclusions;
- ➔ relevant documents.

The software can then give the AI a precise analytical task.

It can instruct it not to infer certain information.

It can require unknowns to remain unknown.

It can force conclusions into predetermined categories.

It can retrieve information only from approved sources.

And it can reject requests that fall outside the application's purpose.

The result is not an infallible system.

No AI system is.

But it substantially reduces the freedom of the model to wander into unsupported interpretation.

Constraints are a feature, not a limitation

People often assume that a more capable AI system is one that can do more things.

For professional software, the opposite can often be true.

A system becomes useful precisely because there are things it cannot do.

A financial-analysis application should not suddenly become a marketing consultant.

A market-intelligence system should not invent organisations because they sound plausible.

A compliance application should not casually reinterpret the law because the user insists that their preferred interpretation is correct.

Professional software can impose boundaries that an ordinary chat conversation does not naturally possess.

That may include:

Structured inputs

The user completes predefined fields instead of describing the problem however they like.

Controlled instructions

The model receives system-level rules that the end user does not continually rewrite during the conversation.

Defined data sources

The application can restrict research to specified databases, documents, APIs or verified organisations.

Output schemas

Instead of producing an unrestricted essay, the AI may have to populate predefined sections, tables or data fields.

Validation

The software can check whether required information is present before the model is allowed to continue.

Separation of analysis from discovery

One part of the system may analyse information supplied by the client, while another independently discovers external information. The two functions need not be mixed together.

Permissions

Different users may have access only to particular functions, clients, datasets or reports.

All these controls reduce ambiguity.

And reducing ambiguity is one way of reducing the opportunities for sycophancy, hallucination and unsupported reasoning.

Some examples from my own work

I have used this approach in a number of professional projects.

Market intelligence

For market-intelligence work, I created a system in which the user does not merely ask:

> "Find me opportunities in Spain."

The request is structured around specific commercial variables: product, target countries, geography, customer categories, competitors, channels, objectives and exclusions.

Different operating modes distinguish between **analysis of supplied information** and **analysis combined with external discovery**.

The model therefore operates within a commercial research framework rather than behaving like an unrestricted chatbot.

Verified company discovery

In another application, AI analysis was combined with controlled company discovery.

The objective was not to ask a model to generate a list of companies that *might* exist.

The system used external business-location data to identify actual organisations first, after which AI could help classify, interpret and organise the findings.

The difference is fundamental.

The AI analyses evidence. It does not manufacture the evidence.

Agribusiness intelligence

For agrifood brokerage, I developed private intelligence tools in which market information, trade flows and price signals are obtained from defined sources and presented inside a controlled operational environment.

The objective is not to have an AI casually answer:

> "Is there an opportunity for this product?"

Instead, the operator works with actual trade, market and price information and can use AI where interpretation adds value.

The evidence comes first.

The model comes afterwards.

AI for a fruit and vegetable exporter

For a fresh-produce business, the useful question was not simply:

> "How can we use AI?"

That question is so broad that almost any answer can sound convincing.

The work began instead with the organisation: how information moved, how commercial activities were structured, where repetitive work existed and where decisions required judgement.

Only then could AI be placed into specific parts of the operation where it could perform a defined function.

This distinction matters enormously.

You do not redesign the business around what an AI model happens to be capable of doing.

You identify the business requirement first and then constrain the technology around that requirement.

The model becomes a component

This is perhaps the most important conceptual change.

In professional AI software, the model should not necessarily be considered the product.

It is a component of the product.

A useful system may contain:

- ➔ business rules;
- ➔ databases;
- ➔ APIs;
- ➔ forms;
- ➔ validation logic;
- ➔ authentication;
- ➔ permissions;
- ➔ retrieval systems;
- ➔ deterministic calculations;
- ➔ reporting;
- ➔ document generation;
- ➔ and, somewhere inside that architecture, an AI model.

The model performs the tasks for which probabilistic language reasoning is useful.

The rest of the software determines when, where and how that capability may be used.

That is very different from opening a chatbot and asking it to solve the entire problem.

From prompt engineering to system engineering

This also explains why professional AI work cannot be reduced to prompt engineering.

A good prompt certainly matters.

But a prompt remains an instruction given to a probabilistic model.

A professional system adds another layer: **architecture.**

Instead of asking:

> How do I persuade the AI to behave correctly?

the better question becomes:

> How do I design the system so that incorrect behaviour becomes less likely, less consequential and easier to detect?

That is system engineering rather than prompt engineering.

And for many serious applications, that is where AI becomes genuinely useful.

The objective is not to give the model unlimited freedom and hope that it behaves intelligently.

The objective is to give it exactly as much freedom as the task requires, no more.