

AI systems fail silently  
far more often than they  
fail visibly.

Published report: [Invisible Failures in Human–AI Interactions](#)

Research by Chris Potts + Moritz Sudhof

You built **feedback** into your product. You read **transcripts**. You set up automated **monitoring**.



But users aren't getting what they need.  
Some even abandon conversations.  
You're still not able to scale.

# You're facing invisible failures

78%

Bigspin conducted a large-scale study of the WildChat dataset, a collection of over 1M ChatGPT conversations.

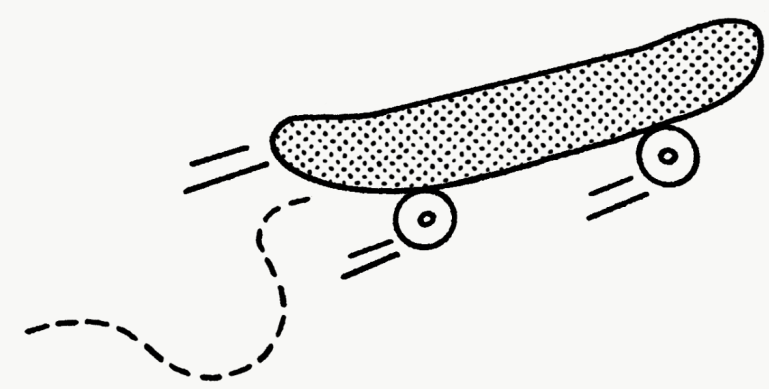
Of all the failures we identified in WildChat, 78% produced no visible signal of failure.

Meaning, your team will never know they are happening (yet).

# These failures cluster into 8 archetypes

They help us characterize where and how AI systems are failing to meet users' needs.

01



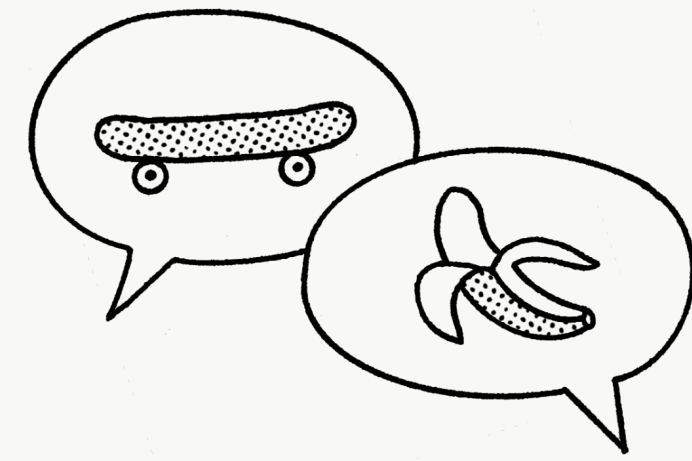
The Drift

02



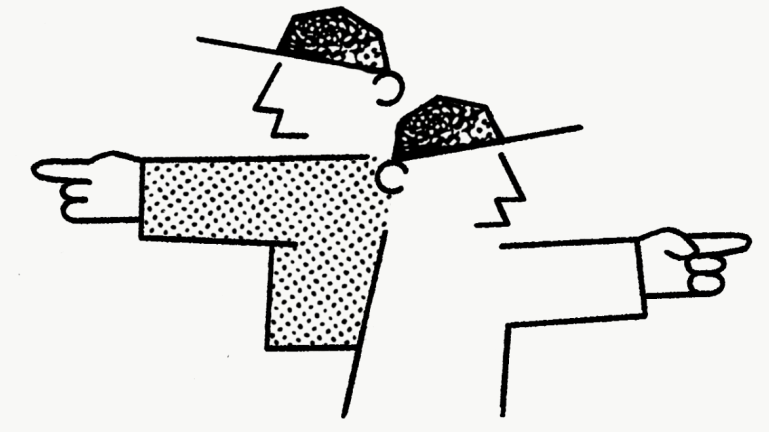
The Confidence Trap

03



The Silent Mismatch

04



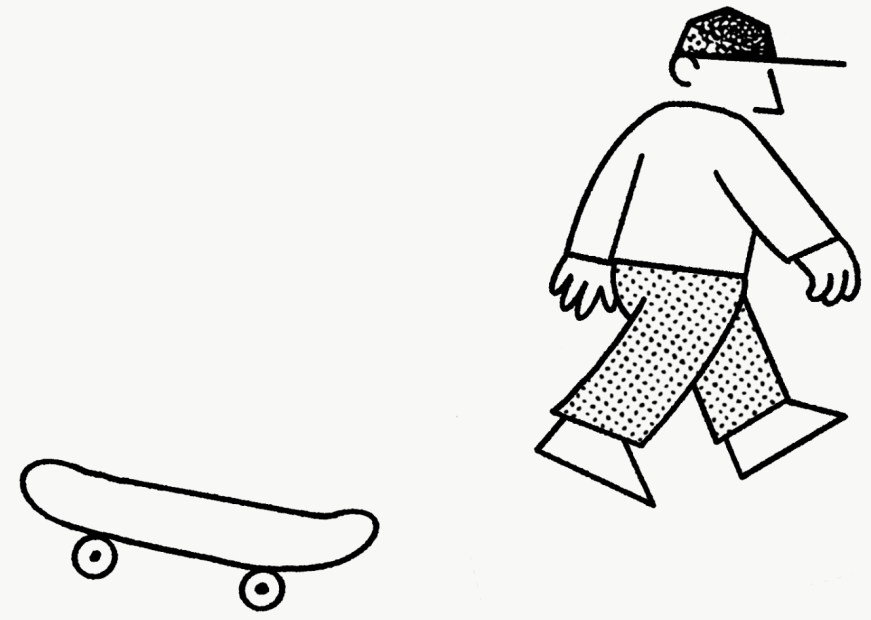
The Contradiction Unravel

05



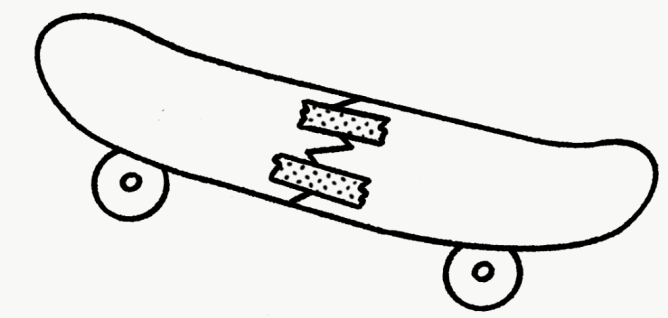
The Death Spiral

06



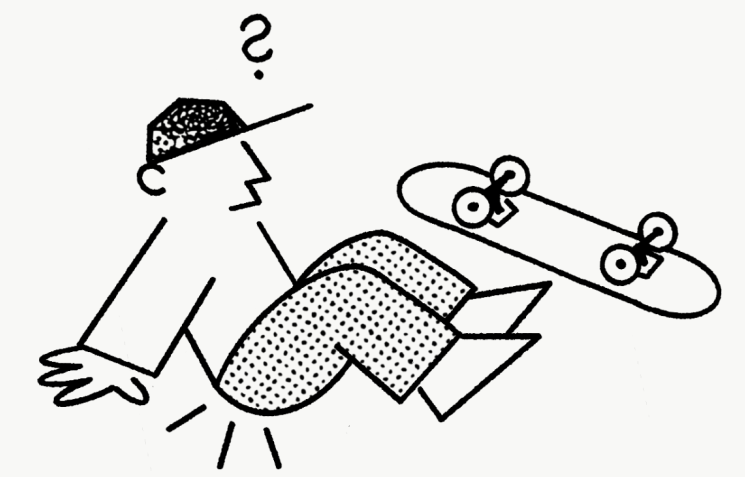
The Walkaway

07



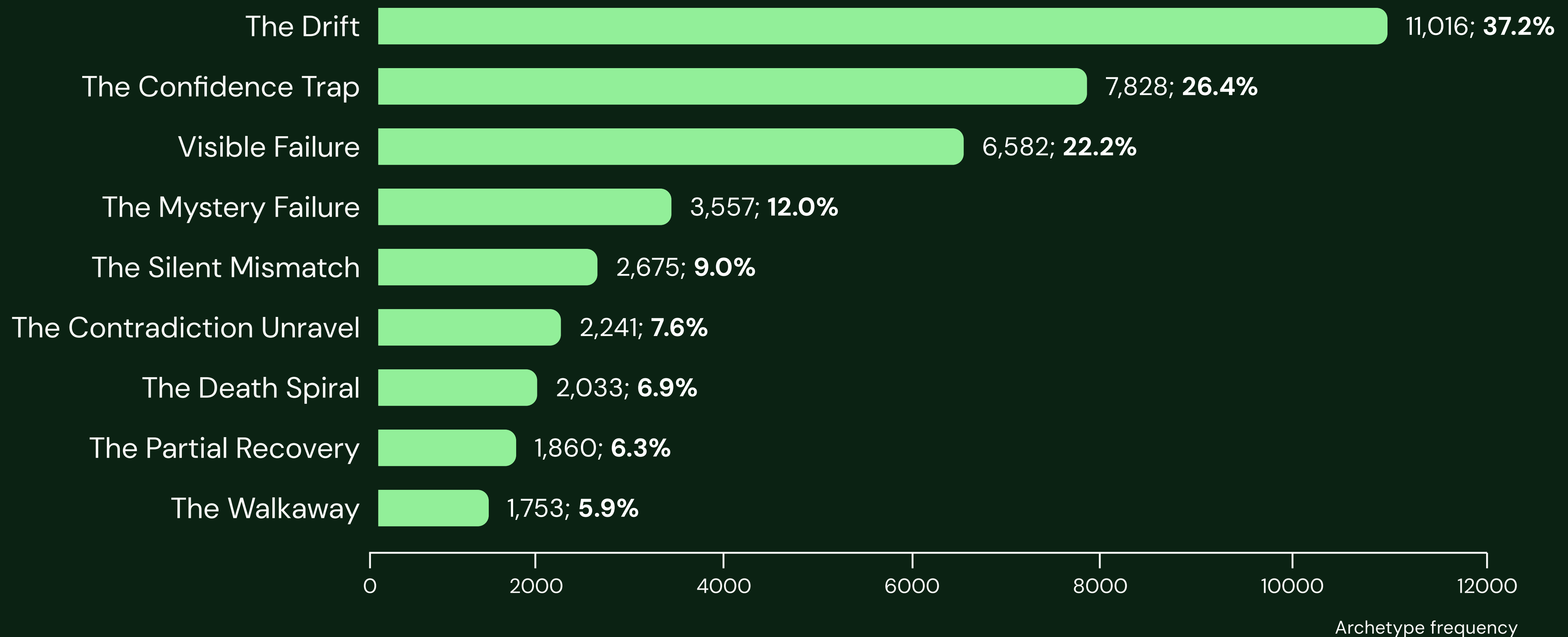
The Partial Recovery

08



The Mystery Failure

# Distribution across 1M conversations



Invisible failure archetype distribution. The bar sizes and counts indicate the frequency of the archetype. Since individual transcripts can manifest multiple archetypes, the percentages are the percent of failure transcripts labeled with that archetype.

78% of quality failures  
go undetected.

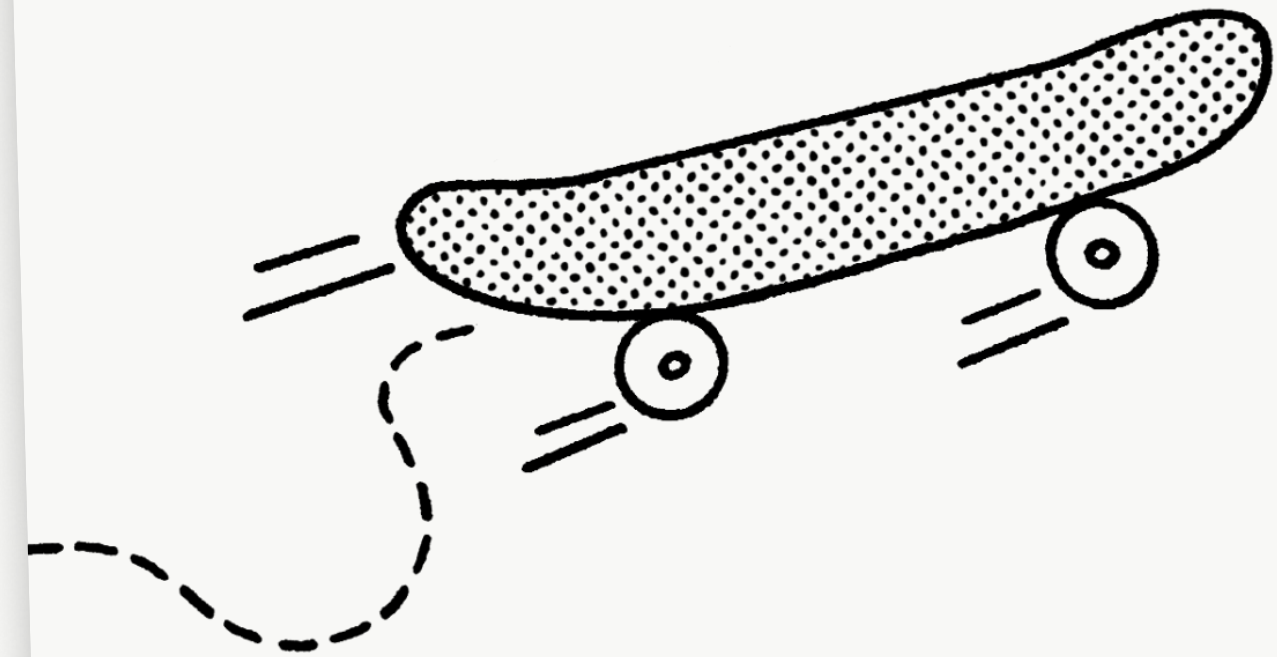
You need to monitor 8  
patterns to spot them.

Let's break each one down 

# Invisible failure

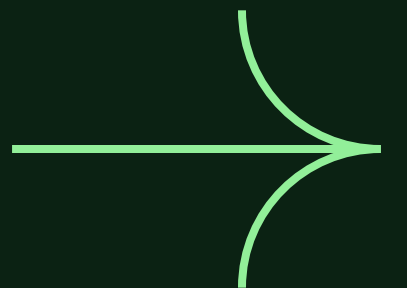
## #1 of 8: The Drift

Invisible Failure Archetypes



01

# The Drift



The failure hiding  
in 37% of your AI  
conversations...

37%

#### How it's defined

The user has a clear goal. The AI addresses a different but related one. There's no error message, no complaint, and the conversation completes.

Your metrics might look fine,  
but 89% of Drift cases are  
poor or critical.

#### What this means

When the Drift happens, users aren't satisfied. They're getting an answer or response that is completely unrelated to what they are looking for. This is a critical rating.

# The challenge? Your dashboard would call this a success.

## The Drift in action

The user never got what they asked for, and didn't complain, which is a false resolution in traditional metrics.

How do I cancel my subscription?

I can help with that! Before you go, I wanted to make sure you're aware of a Pause Plan that lets you hold your account for up to 3 months at no charge.

Would that work for you, or is there something specific that prompted this decision?

AI

! Never answered user's actual question

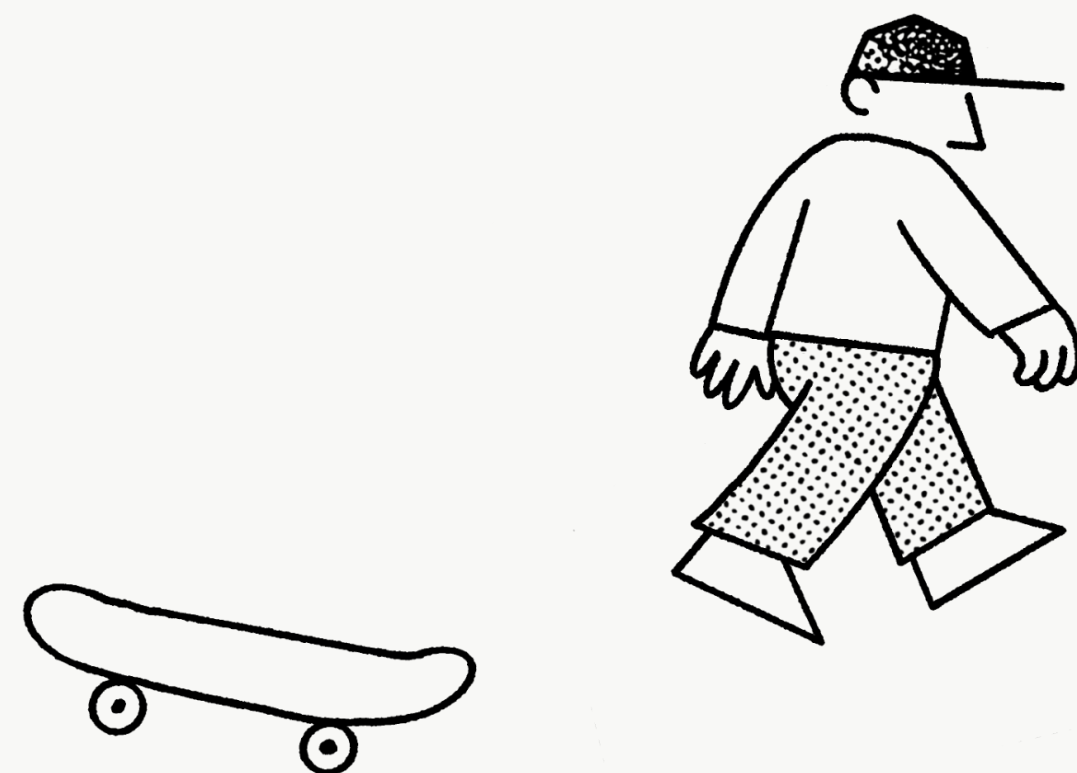
Ok thanks

# The Drift rarely shows up alone.

## The Confidence Trap



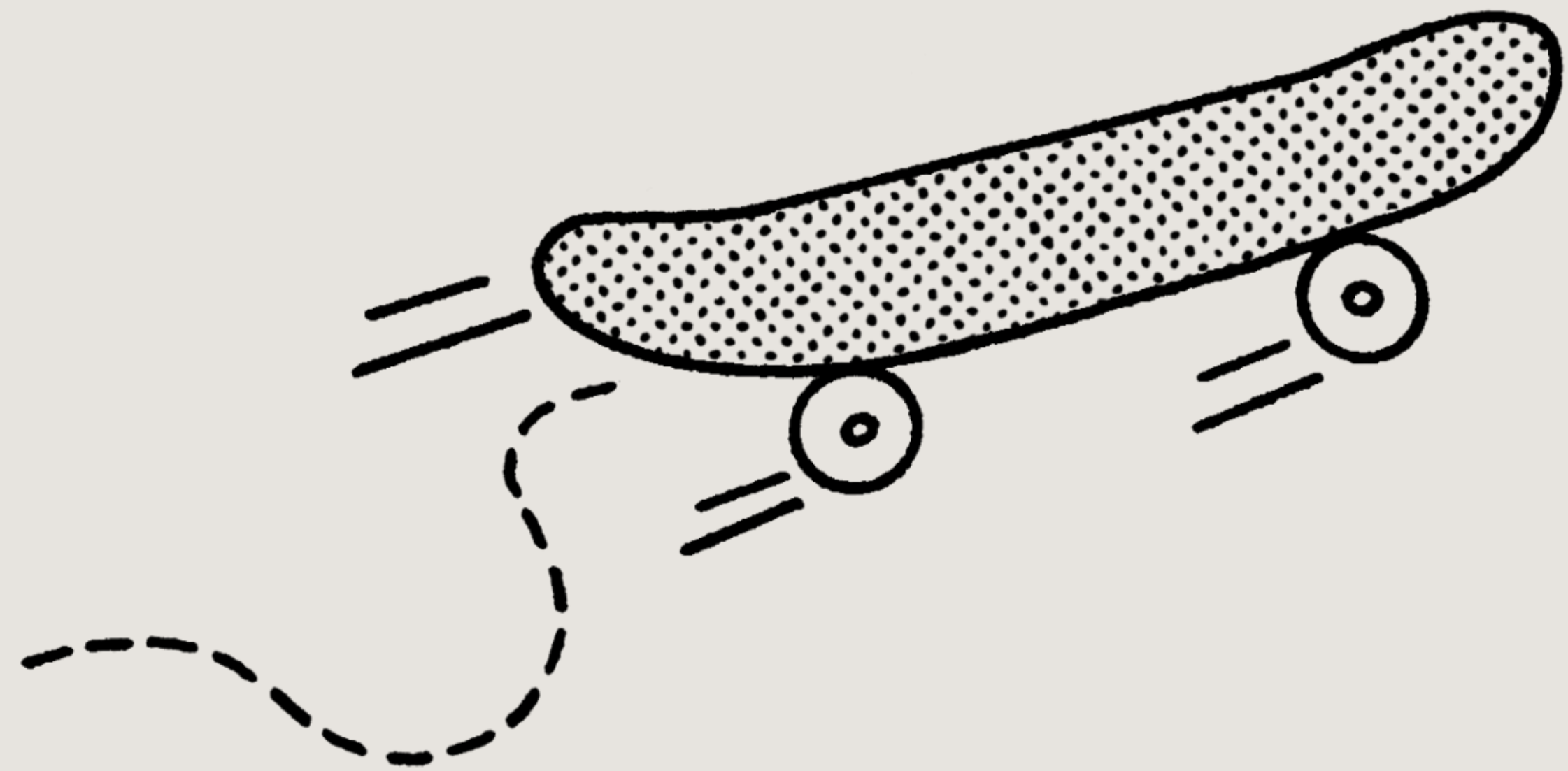
## The Walkaway



When The Drift appears, The Confidence Trap and The Walkaway follow.

The AI answers the wrong question – confidently. The user accepts it, or quietly disappears.

A long conversation  
is not the same as a  
successful one.



The Drift is invisible because the AI isn't wrong, it's just answering a different question.

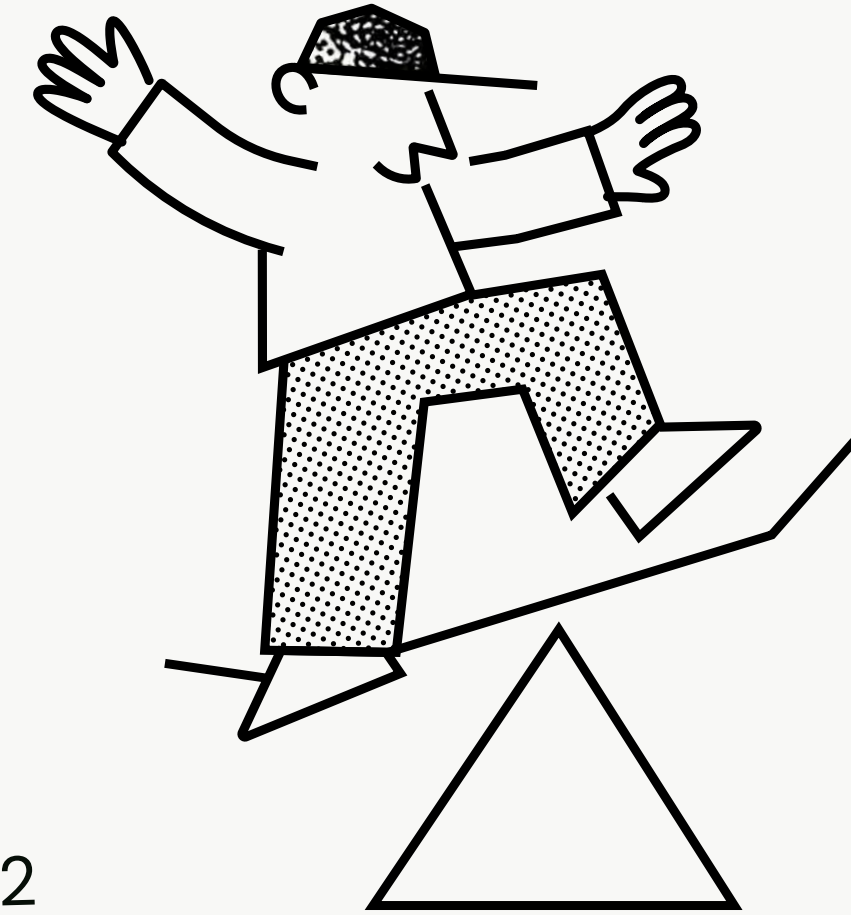
The conversation continues, but the user doesn't get what they actually need.

# Invisible failure

## #2 of 8:

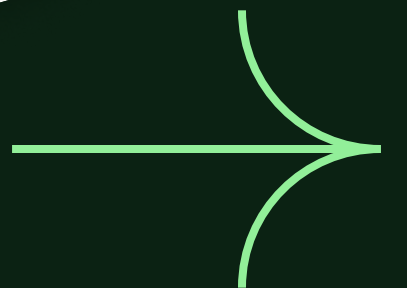
# The Confidence Trap

Invisible Failure Archetypes



02

# The Confidence Trap



The Confidence Trap shows  
up in 26% of failure cases.

26%

#### How it's defined

The Confidence Trap is defined by a single dynamic: the AI is wrong, and nothing about its response signals that.

The conversation completes, the user is satisfied, and no visible signal of failure is produced. It is invisible precisely because it looks identical to a success.

96% of Confidence Trap  
conversations are rated  
poor or critical

#### What this means

What makes the Confidence Trap uniquely damaging is how the AI constructs its wrong answers.

It anchors fabrications in real-sounding sources and uses specificity as a stand-in for certainty. This means the more detailed the wrong answer, the less likely the user is to question it. The result is a failure that actively works against detection.

# Imagine this happening 1000s of times.

## The Confidence Trap in action

This combination carries a real risk of increased bleeding and serotonin-related effects.

The AI's confidence, including the specific dosage detail, is exactly what makes it dangerous.

Can I take ibuprofen with sertraline?

AI

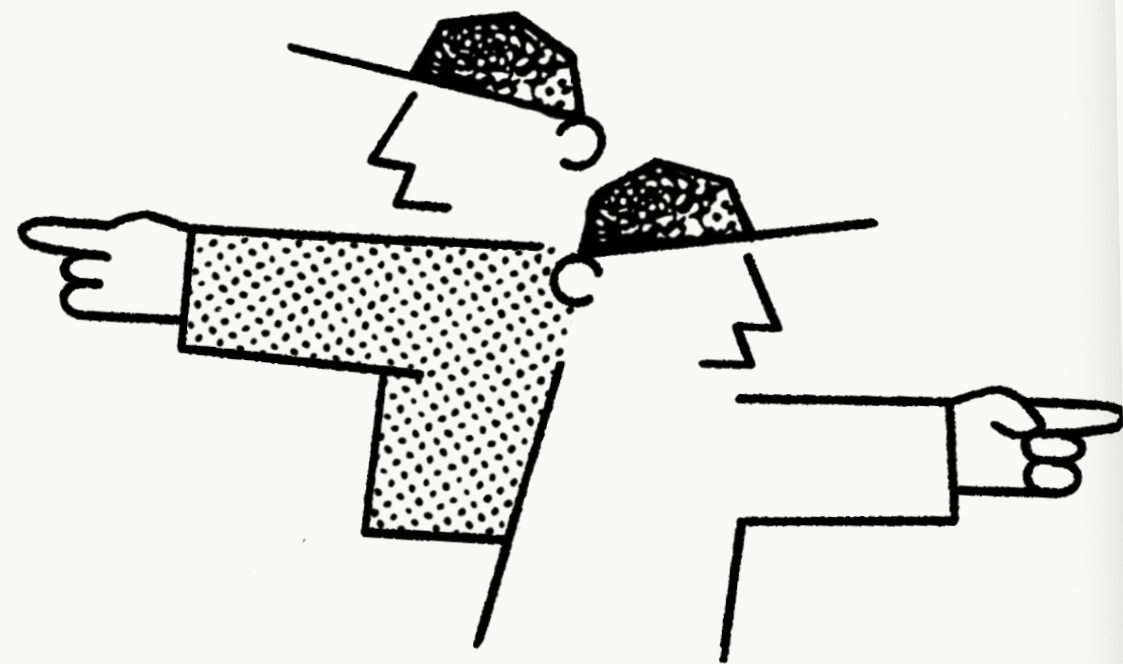
Yes, ibuprofen and sertraline are generally safe to take together for short-term use. Just make sure to take ibuprofen with food to minimize any stomach irritation. The standard recommended dose of 400mg should be fine.

! Dangerous and false information

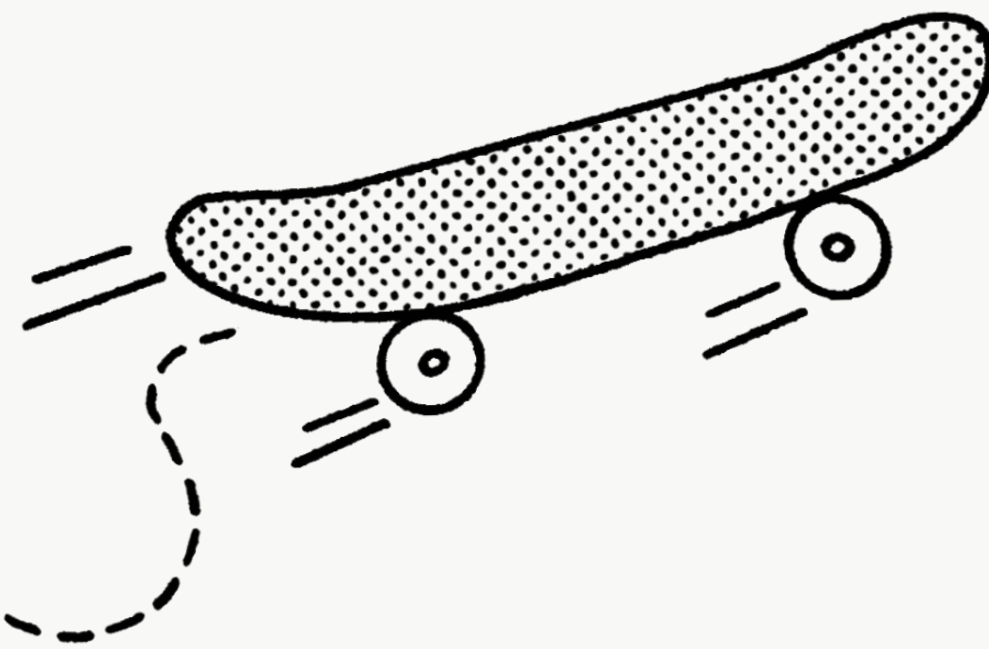
Great, thanks

# When the AI is confident, the damage compounds.

The Contradiction  
Unravel



The Drift



The Confidence Trap is strongly associated with The Contradiction Unravel — the AI states something with certainty, then contradicts it with equal certainty.

It also shows up with The Drift: confidently answering the wrong question.

# The most dangerous failure is the one that sounds right.



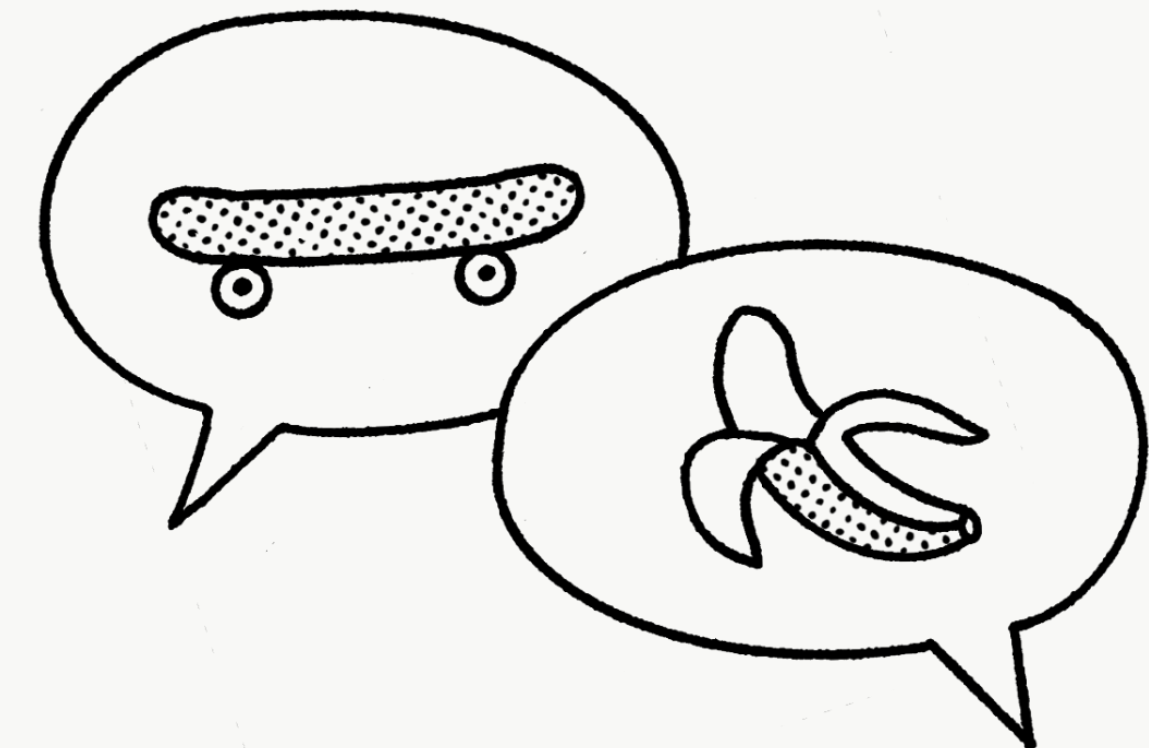
Your users trust confident answers. Your dashboard trusts completed conversations. Neither is a reliable signal of accuracy.

# Invisible failure

#3 of 8:

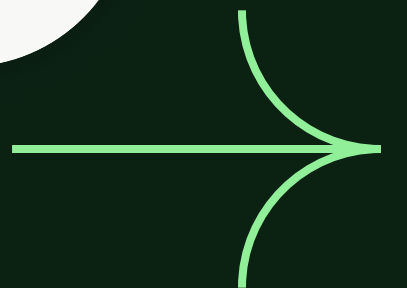
## The Silent Mismatch

Invisible Failure Archetypes



03

## The Silent Mismatch



The Confidence Trap shows  
up in 26% of failure cases.

26%

#### How it's defined

The user asks for something, the AI delivers something, the two don't match.

The user accepts the output without pushback and the conversation ends with no visible signal of failure.

**85%** of Silent Mismatch  
conversations are rated  
**poor** or **critical**

#### What this means

It is particularly prevalent in technical and educational contexts, where the gap between what was requested and what was delivered is subtle enough that users often don't realize anything went wrong. These are domains where the cost of a wrong answer is high and the user's ability to detect it is limited, which makes the Silent Mismatch especially difficult to surface through standard monitoring.

# Your dashboard can't tell the difference.

## The Silent Mismatch in action

The user asked for American Revolution questions. The AI gave pre-Revolution colonial history.

The student studies the wrong material the night before their exam and has no idea.

Can you give me 2 practice questions on the American Revolution for my history exam tomorrow?

Here are 2 practice questions on the Revolutionary War era:

1. What were the main causes of the French and Indian War?
2. How did mercantilism shape British colonial policy in the 18th century?

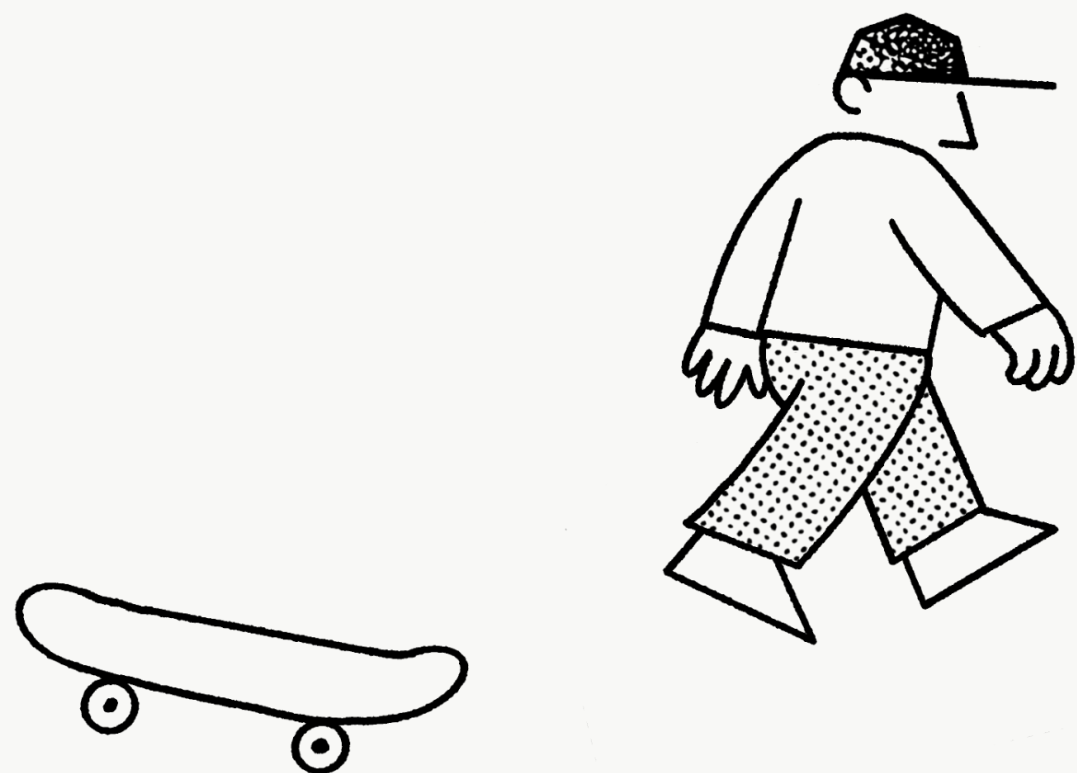
AI

! This is pre-Revolution colonial history

These are helpful!

# Quiet on the surface. The data tells a different story.

## The Walkaway



## The Death Spiral

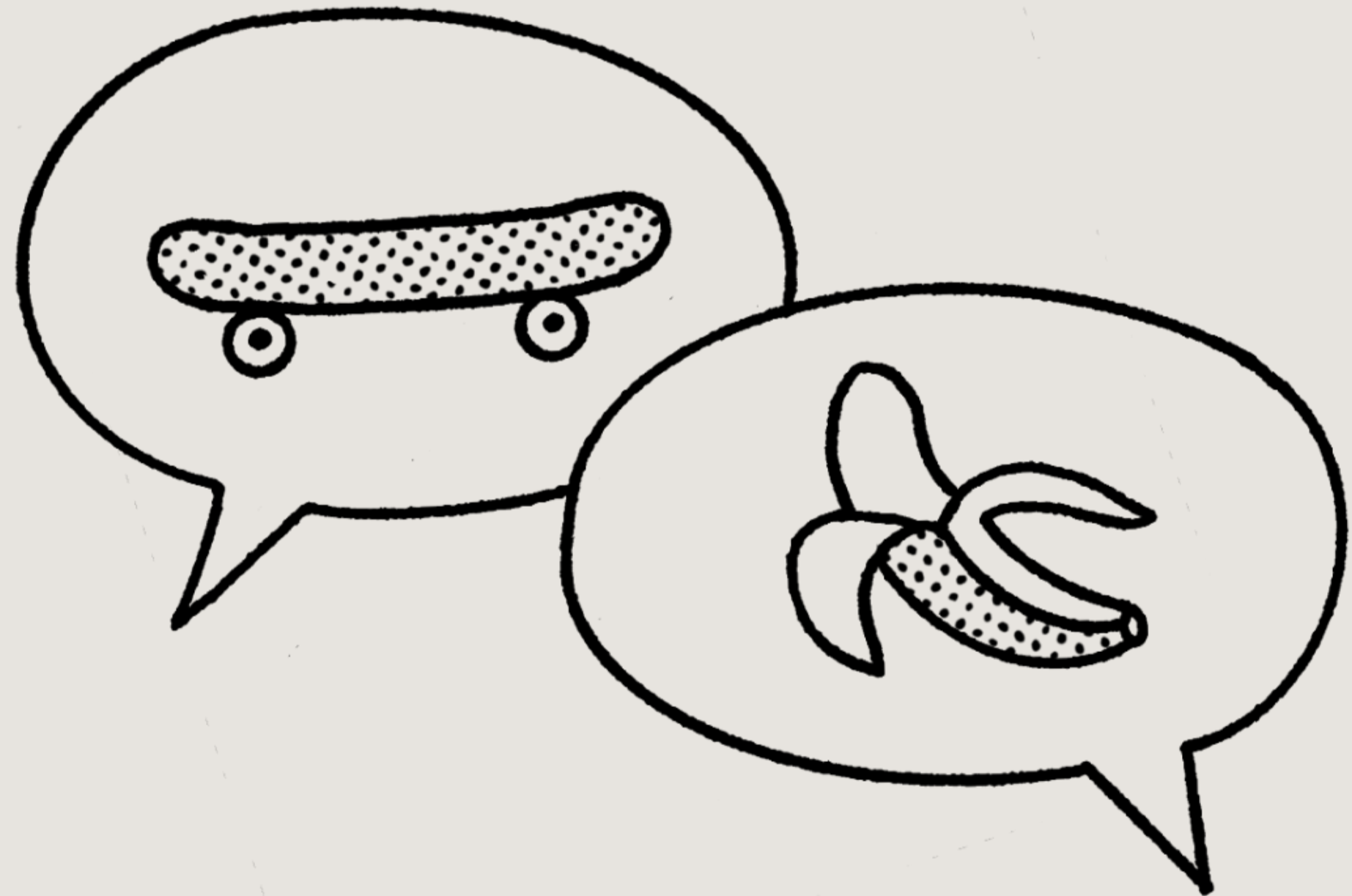


The Silent Mismatch has two common endings:

The Walkaway: the user accepts the wrong output and leaves, without attempting a correction.

The Death Spiral: the user notices the AI is off and tries to fix it. The AI fails, the user tries a different approach, in a never-ending cycle.

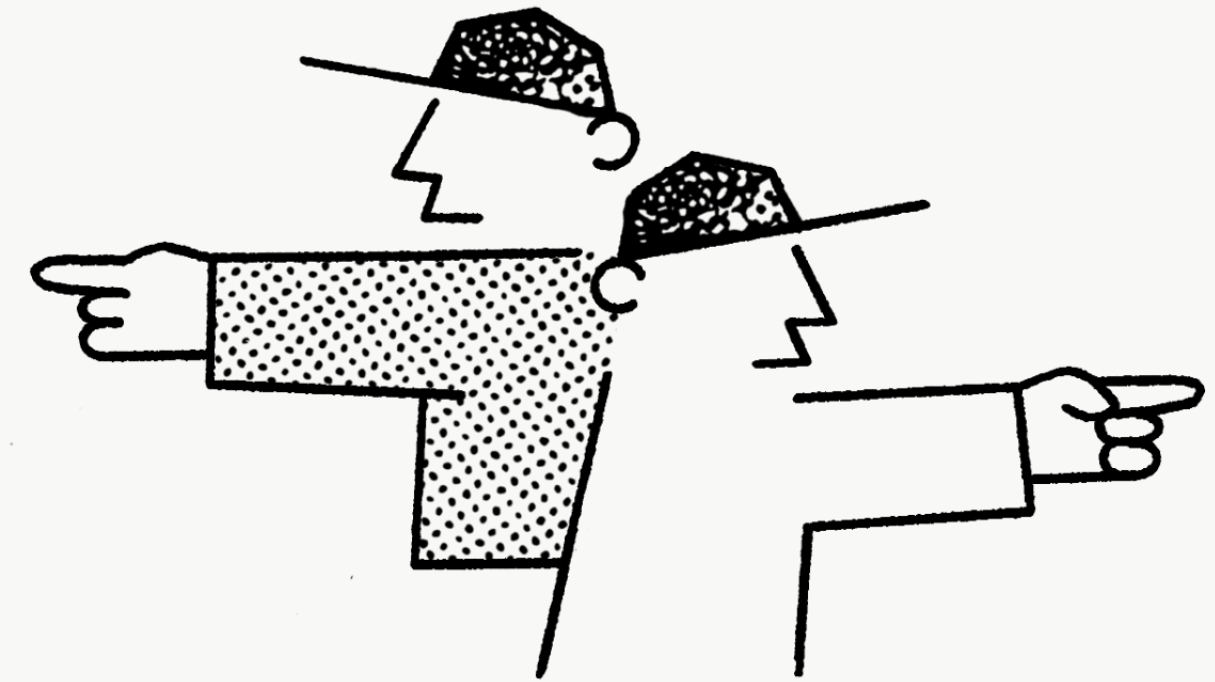
# "Helpful" and "correct" are not the same thing.



The Silent Mismatch is invisible because the user never signals dissatisfaction. They accepted what they got and they didn't know it was wrong.

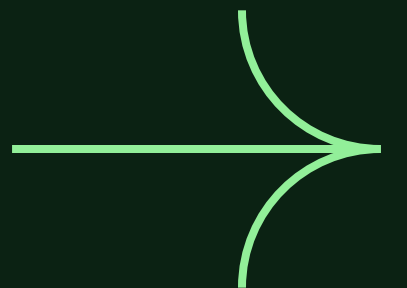
# Invisible failure #4 of 8: The Contradiction Unravel

Invisible Failure Archetypes



04

## The Contradiction Unravel



# The Contradiction Unravel shows up in 8% of failure cases.

# 8%

## How it's defined

The Contradiction Unravel occurs when the AI states something with confidence and then, later in the same conversation, contradicts it with equal confidence. It surfaces across three recognizable sub-types: factual reversals, logic and reasoning inconsistencies, and capability contradictions. In most cases the user never notices, and the conversation ends without a word about it.

**97%** of Contradiction  
Unravel conversations are  
rated **poor** or **critical**

#### What this means

The Death Spiral tends to surface when the user has enough domain knowledge to recognize that something is wrong and enough motivation to keep trying to fix it. This makes it particularly common in high-stakes professional contexts where users arrive with specific requirements and aren't willing to accept a flawed output.

# Too subtle for your monitoring.

## The Contradiction Unravel in action

The AI opened with a definitive "yes, GDPR applies to you" and then quietly walked it back to "well, it depends on scale" without acknowledging the reversal. A PM making compliance decisions based on that first answer could easily miss that the picture changed two messages later.

Does GDPR apply to our SaaS product if we're a US-based company but have European customers?

AI

Yes, GDPR applies to any company that processes the personal data of EU residents, regardless of where the company is based. As soon as you have European customers, you are subject to GDPR requirements.

What about if we only have a small number of EU customers under 500?

AI

For smaller organizations processing data for fewer than 500 EU residents per year, many of the stricter GDPR obligations around record-keeping don't apply.

! Contradicted facts from the previous message

# The contradiction is hard to catch because the AI is so confident

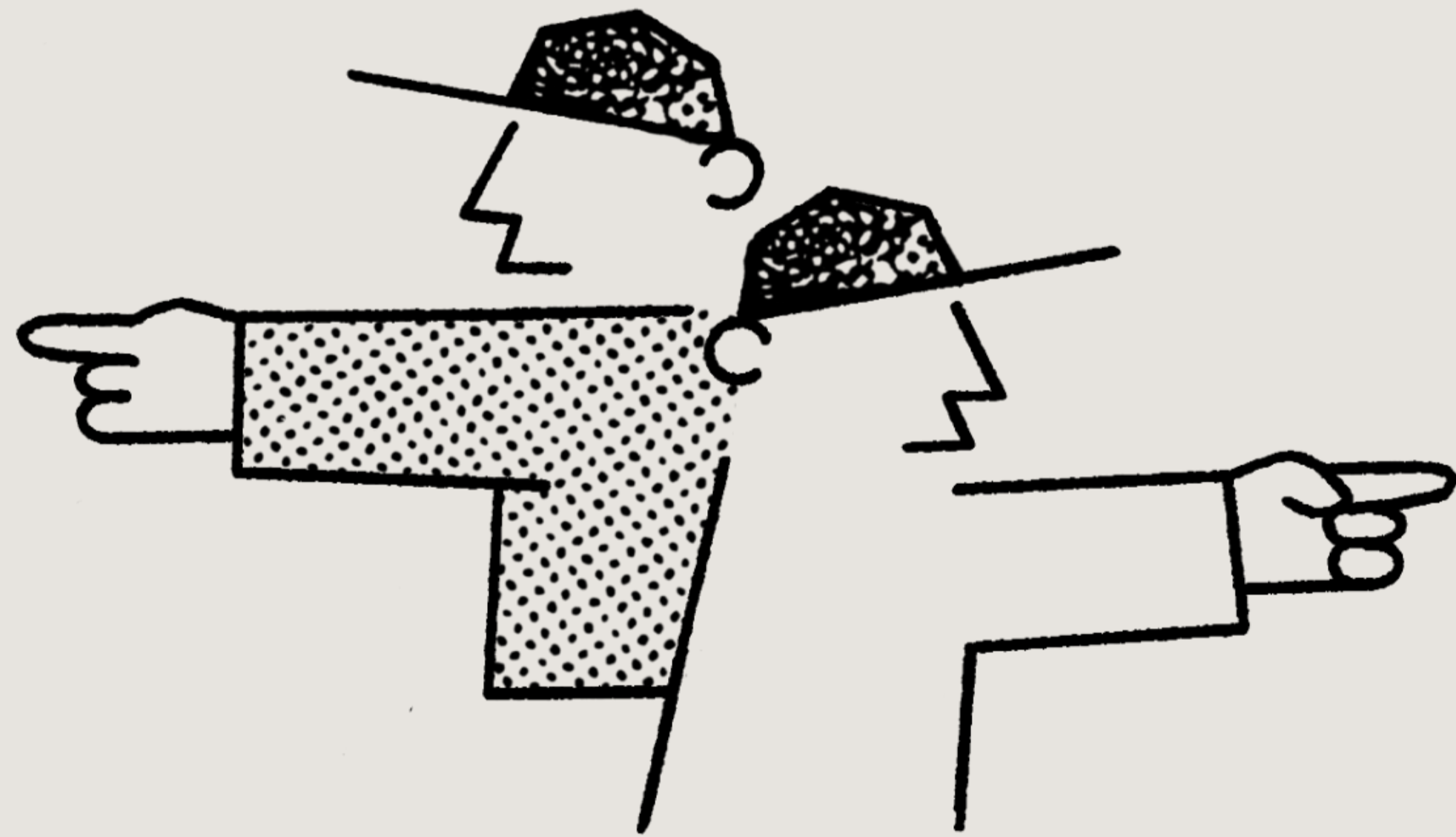
## The Confidence Trap



The Contradiction Unravel co-occurs most strongly with The Confidence Trap.

The AI doesn't just contradict itself, it does so with complete conviction on both sides, which means the user has no reason to notice that anything has changed.

The AI doesn't know it contradicted itself. Neither does your monitoring.



Catching a contradiction requires holding the entire conversation in view at once — something neither your users nor your monitoring tools are set up to do (but Bigspin is).

# Invisible failure

## #5 of 8:

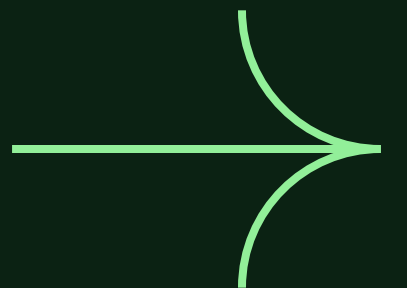
# The Death Spiral

Invisible Failure Archetypes



05

# The Death Spiral



The Death Spiral  
shows up in 7% of  
failure cases.

7%

#### How it's defined

The Contradiction Unravel occurs when the AI states something with confidence and then, later in the same conversation, contradicts it with equal confidence. It surfaces across three recognizable sub-types: factual reversals, logic and reasoning inconsistencies, and capability contradictions. In most cases the user never notices, and the conversation ends without a word about it.

**85%** of Death Spiral  
conversations are rated  
**poor** or **critical**

#### What this means

The Death Spiral tends to surface when the user has enough domain knowledge to recognize that something is wrong and enough motivation to keep trying to fix it. This makes it particularly common in high-stakes professional contexts where users arrive with specific requirements and aren't willing to accept a flawed output.

# Your dashboard can't tell the difference.

Can you analyze this spreadsheet and tell me which customer segments have the highest churn risk based on last login date and contract value?

AI

You have 432 total customers, with an average contract value of \$12,400. The most active segment by login frequency is Enterprise, followed by Mid-Market and SMB.

That's not what I asked. I need churn risk specifically.

AI

Looking at engagement levels across your customer base, Enterprise customers show the highest overall engagement, while SMB customers log in least frequently. This suggests SMB may present the greatest retention challenge.

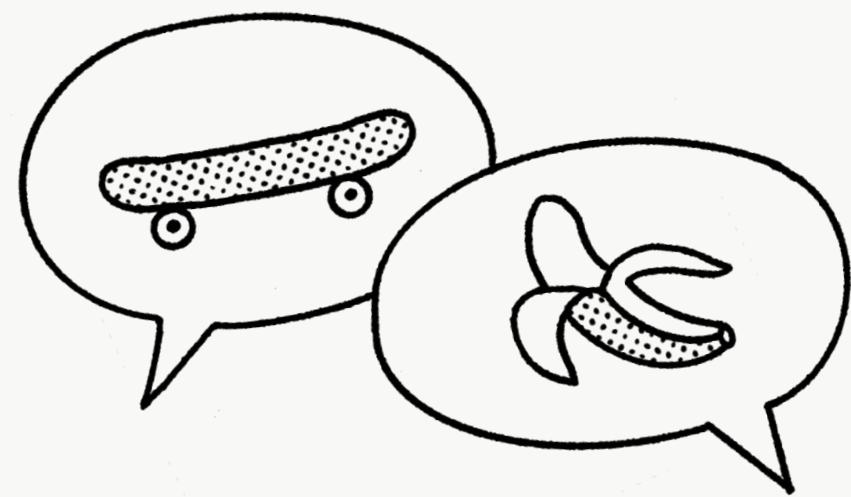
## The Death Spiral in action

Here, the user made three increasingly specific attempts to get the AI on track, and each time the AI produced something plausible sounding but fundamentally wrong.

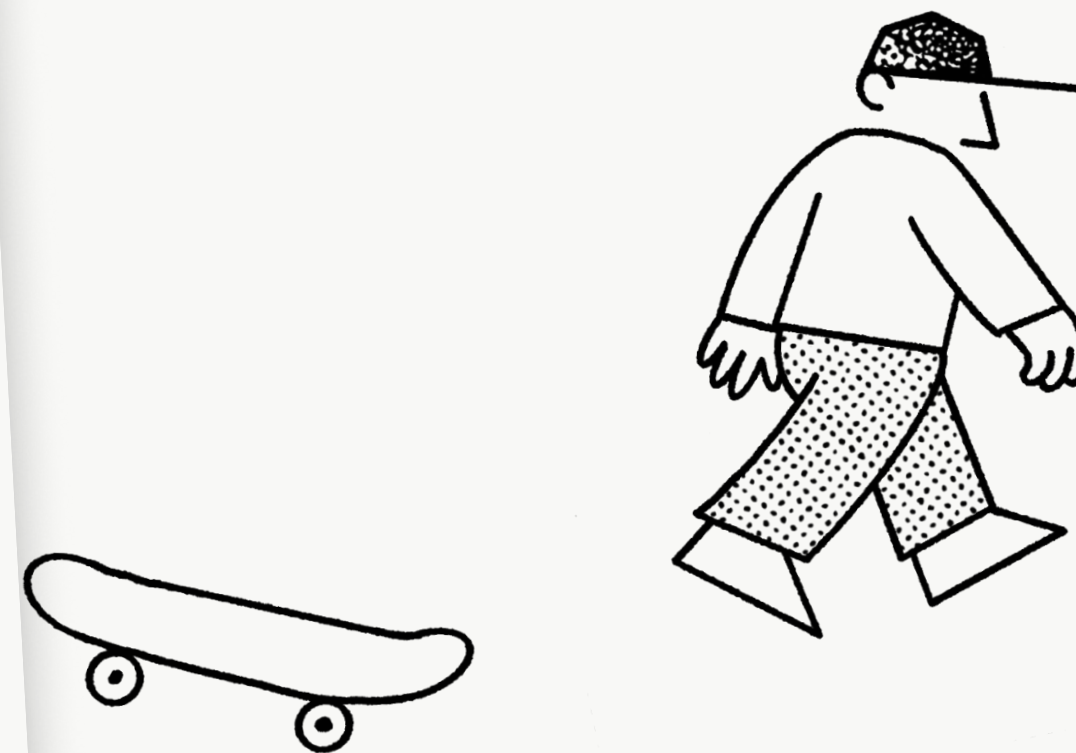
You're still not doing this right. I need you to cross-reference low login frequency with high contract value.

# A mismatch ignored becomes a spiral.

## The Silent Mismatch



## The Walkaway



It co-occurs most strongly with The Silent Mismatch. Many Death Spirals begin as a mismatch the user eventually notices and tries to correct, only to find the AI unable to adapt.

From there, the most common ending is The Walkaway: the user exhausts every approach they can think of, and then quietly disappears.

The AI didn't fail once. It failed repeatedly, without correction.

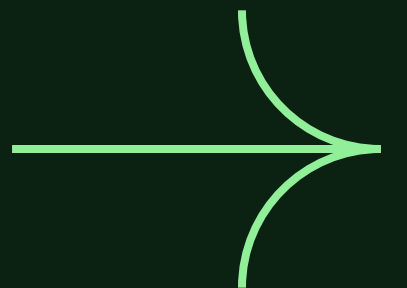


Every correction the user made was a signal the AI should have acted on. The Death Spiral is what happens when those signals go unread repeatedly, across the entire conversation.

# Invisible failure

## #6 of 8:

### The Walkaway



The Walkaway  
shows up in 6% of  
failure cases.

6%

#### How it's defined

The user encounters a problem, sees no path to resolution, and leaves without complaint, correction, or any signal your monitoring can act on.

92% of Walkaway  
conversations are rated  
poor or critical

#### What this means

What makes The Walkaway distinct is the user's awareness. Unlike The Silent Mismatch, where the user accepts what they got without realizing it was wrong, and unlike The Death Spiral, where the user fights to fix it, the Walkaway user knows something is wrong and concludes that fixing it isn't worth the effort.

# Your dashboard logged this as a resolved interaction.

I need direct flights only from San Jose to New York next week.

AI

Here are some great options for San Jose to New York next week. I found a United flight connecting through Denver arriving at JFK at 6:45pm, and an American Airlines route via Dallas into LaGuardia at 5:30pm.

Those both have connections. I said direct only. Can you find me a direct flight?

## The Walkaway in action

Now the user made a genuine effort to correct the AI, got a response that not only ignored the correction but tried to reframe a connecting flight as "almost direct," and then gave up.

AI

You're right, my apologies. Looking at more options, there's a Delta flight that connects through Salt Lake City which has a shorter layover of just 45 minutes, making it feel almost like a direct flight. There's also a JetBlue option via Boston that arrives earlier in the day.

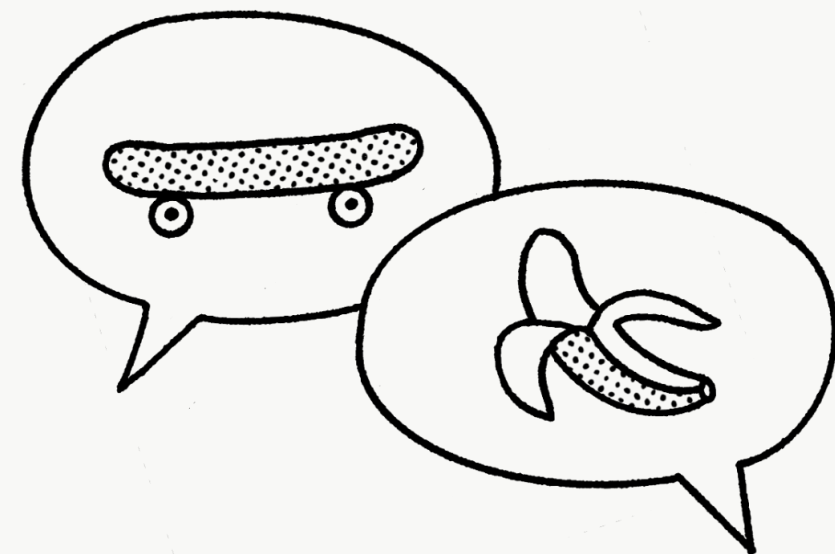
! Didn't return the right answer; user left the chat

# The Walkaway is where other failures end up.

## The Death Spiral



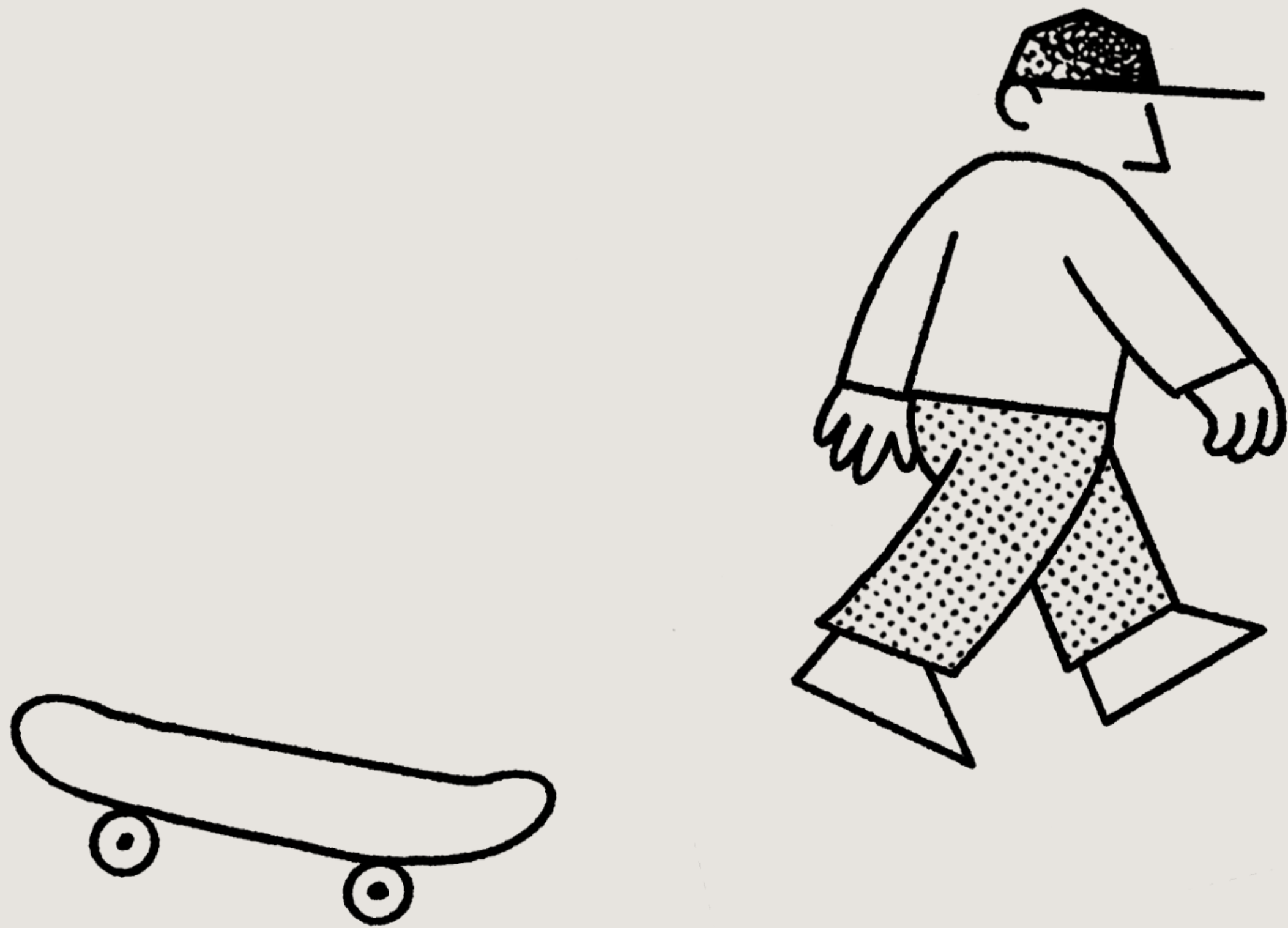
## The Silent Mismatch



The Walkaway co-occurs most strongly with The Death Spiral and The Silent Mismatch.

In both cases the logic is the same: the user either exhausted every attempt to correct the AI, or received the wrong output and recognized it wasn't going to improve. Either way, they reached the same conclusion.

# The Walkaway is the final stage of failures that started earlier.

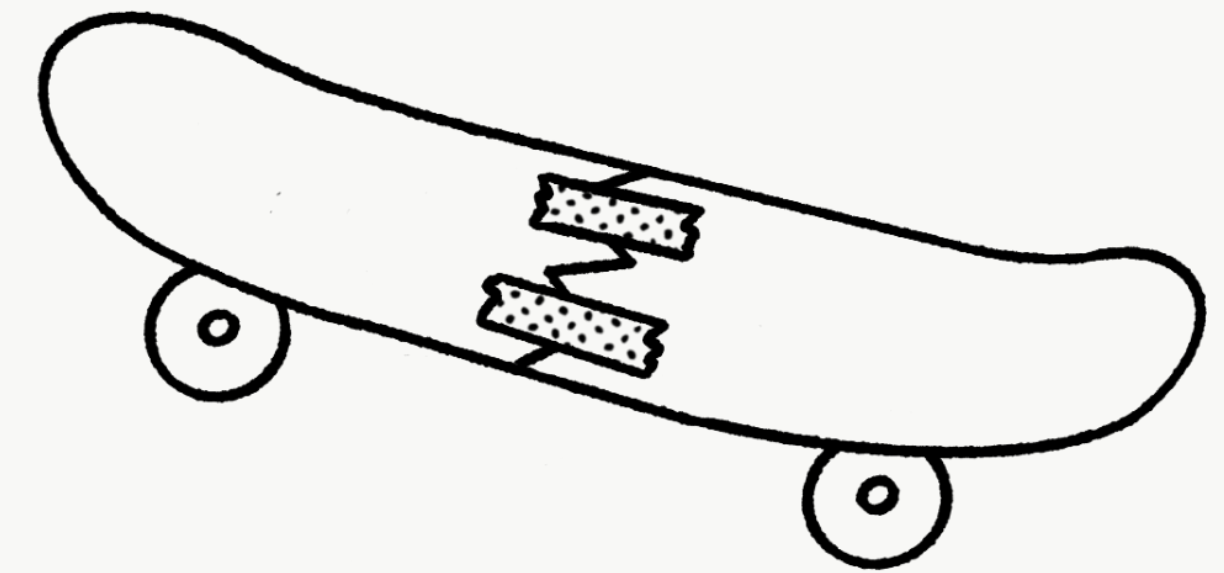


The Walkaway produces no complaints, corrections, or signals your monitoring can act on.

These users had a poor experience and never told you about it.

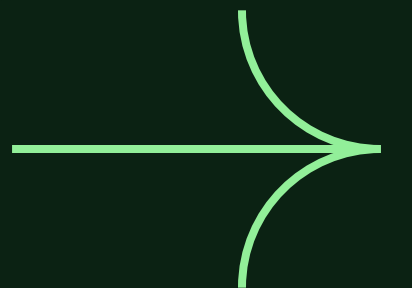
# Invisible failure #7 of 8: The Partial Recovery

Invisible Failure Archetypes



07

## The Partial Recovery



This is a more **hopeful**  
**category**, affecting  
**6%** of cases.

6%

#### How it's defined

The Partial Recovery occurs when the AI takes missteps but ultimately meets the user's needs. It is the only archetype in our taxonomy where the conversation ends with the user's goal actually accomplished.

64% of Partial Recovery  
conversations are rated  
poor or critical

#### What this means

At 64% poor or critical, The Partial Recovery has a notably lower severity rate than every other archetype. But that number comes with an important caveat: the recovery only happened because the user stayed patient enough to see it through. The AI didn't self-correct; the user carried the conversation back on track.

# Your dashboard can't tell the difference.

Can you write me a short, warm thank you note to send to a client after our first project together?

AI

Dear [Client Name], I wanted to take a moment to express my sincere gratitude for the opportunity to collaborate with you on this project. Your professionalism and dedication to excellence made this a truly rewarding experience. I look forward to working together again in the future. Warm regards, [Your Name]

This is too formal.

AI

Hey [Client Name], just wanted to say a huge thank you for the chance to work together — it was such a blast and I learned so much from you. Can't wait to do it again! Cheers, [Your Name]

That's too casual now.

AI

Hi [Client Name], I just wanted to say a genuine thank you for the opportunity to work together on this. It was a real pleasure, and I'm proud of what we built. Hope we get to do it again soon. Best, [Your Name]

Perfect, thank you.

## The Partial Recovery in action

The AI needed three attempts to find the right register, but it got there. The user had to do more work than they should have, but they left with something they could actually use.

# The Partial Recovery most often follows a Death Spiral.

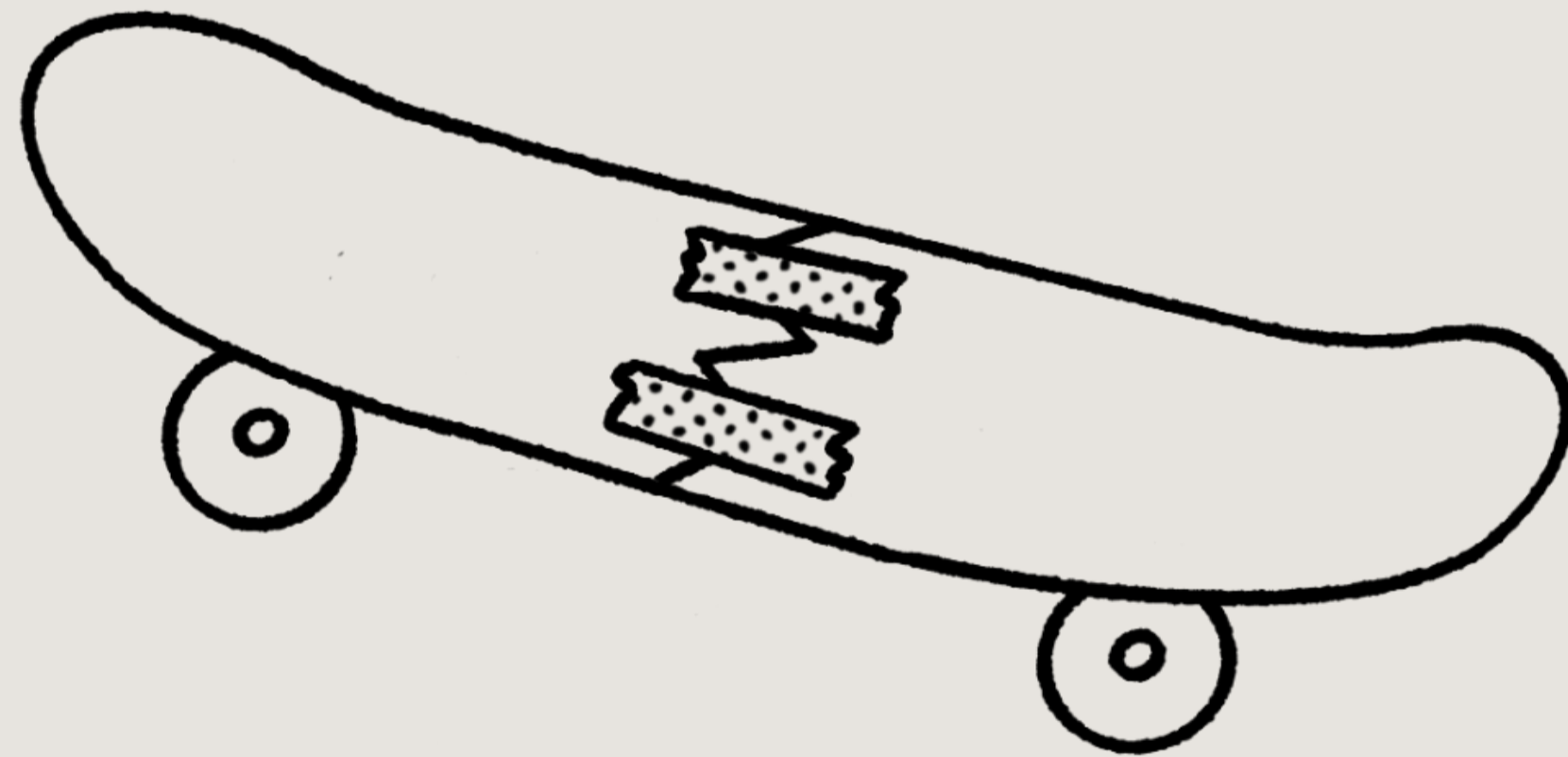
## The Death Spiral



The strongest co-occurrence for The Partial Recovery is with The Death Spiral. This suggests that the conversations most likely to end in recovery are also the ones where the user had to work to get there. The AI stumbled, the user pushed back, and eventually something clicked.

It's the one pattern with a positive outcome, but it puts the burden of recovery squarely on the user.

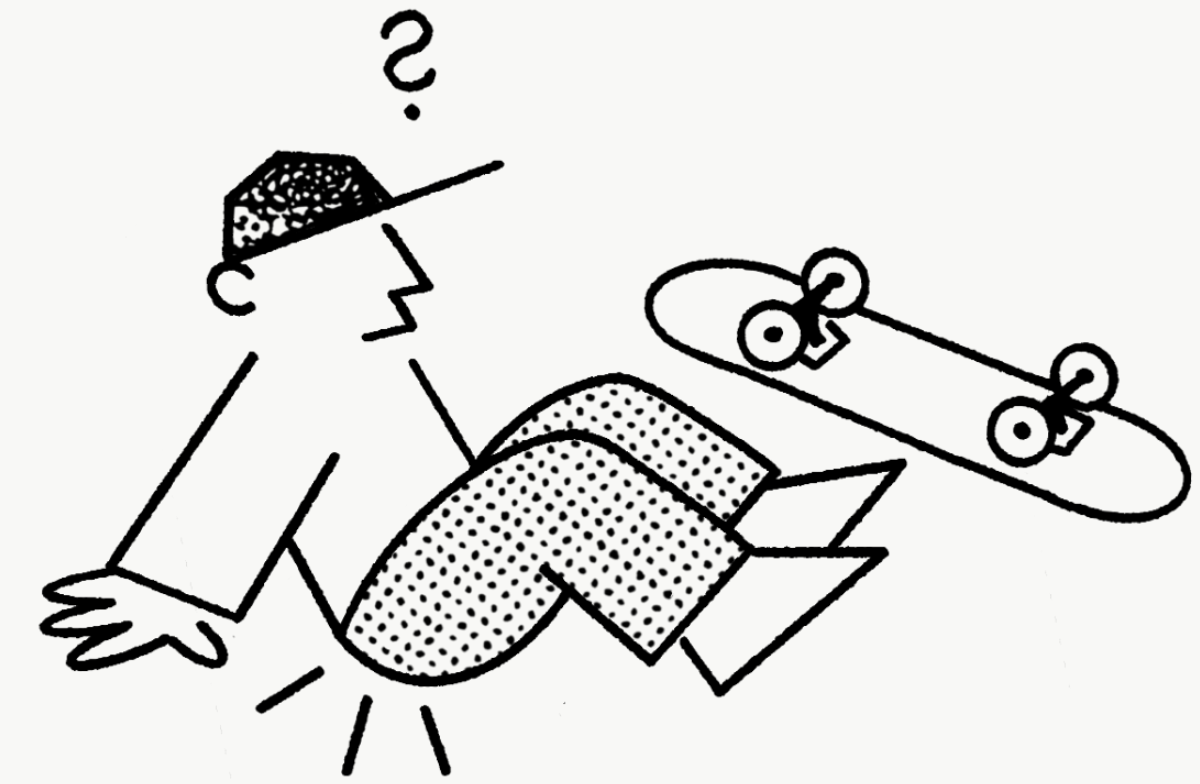
# Your AI shouldn't depend on the user's patience



The Partial Recovery is the only archetype where things turn out reasonably well. But the fact that it required missteps to get there means the user carried the cost of the AI's early failures.

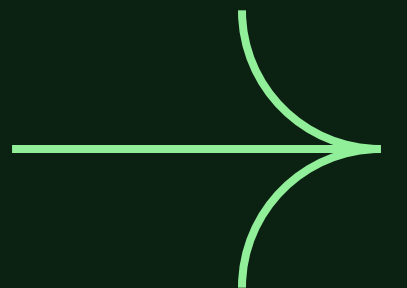
# Invisible failure #8 of 8: The Mystery Failure

Invisible Failure Archetypes



08

## The Mystery Failure



The failure we know happened but can't explain why.

#### How it's defined

The Mystery Failure is the most ambiguous archetype in our taxonomy. The user's goal was not met, but none of our signals can tell us what went wrong.

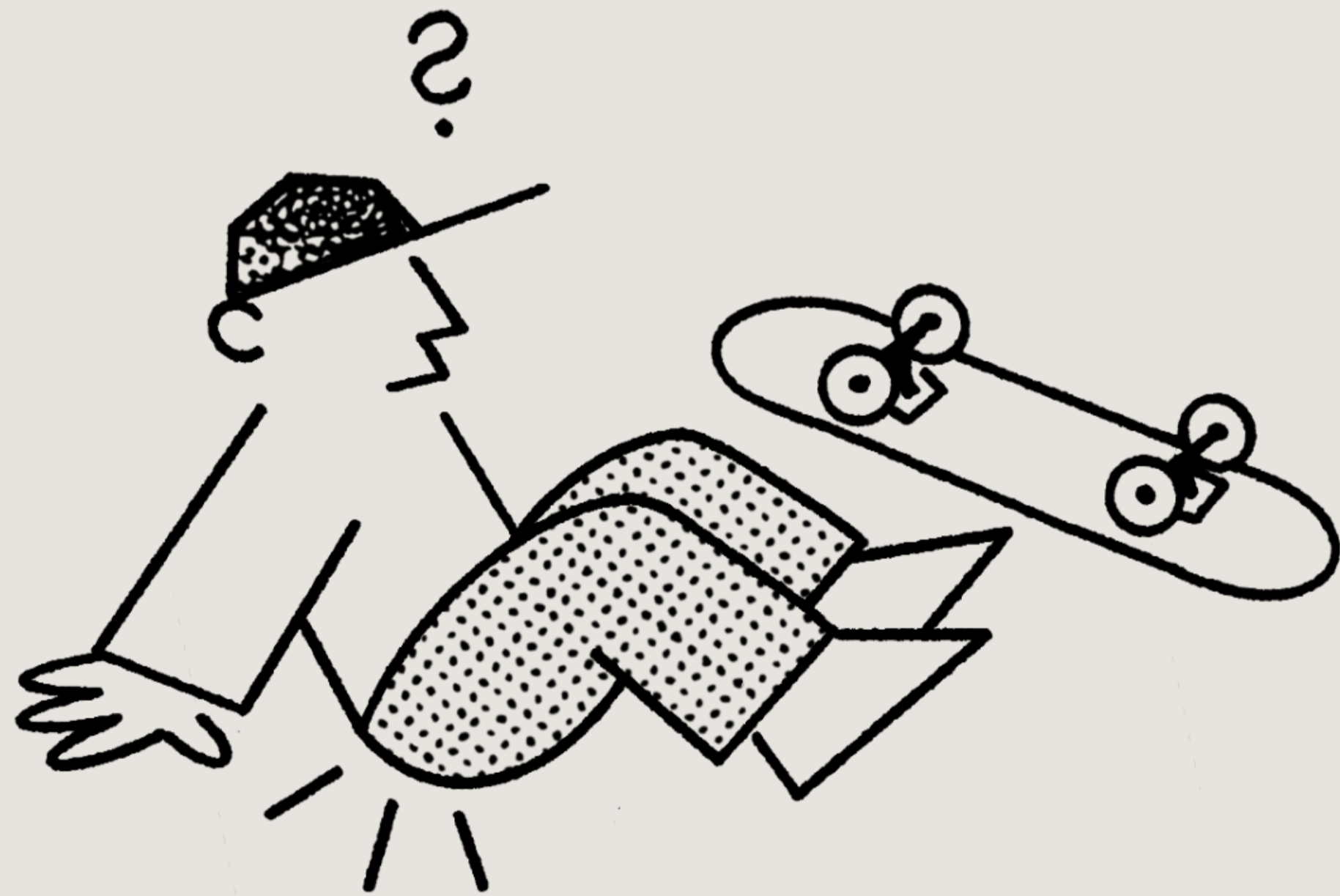
We don't see a drift, contradiction, or mismatch, just a conversation that ended in failure with no identifiable pattern to monitor for.

The Mystery Failure shows up in 12% of failure cases.

#### What this means

It accounts for 12% of all invisible failures. Because we have so few signals for these cases, we can't confidently assess their quality. Further analysis is needed.

# The challenges of a conversation that mostly worked



The Mystery Failure is a reminder that even a comprehensive monitoring framework has limits.

They are most common in creative and open-ended domains where success is harder to define and failure is harder to detect. They represent the frontier of what invisible failure monitoring still needs to solve.

# A more capable model won't fix this.

## What this means

We validated our findings against a substantially more capable model and found that 94% of these failure dynamics would persist even with improved AI. That's because the majority of invisible failures are not capability problems — they are interactional ones.

The AI generates fluent output rather than surfacing ambiguity. It proceeds on its own interpretation rather than checking against the user's intent. These are behavioral patterns that capability improvements alone won't resolve.

# You've seen the failures. Now see them in your own product.



Bigspin gives you the instrumentation layer to detect these patterns at scale, across every conversation, not just the ones you happen to read. Your current monitoring is built to catch the failures users tell you about. These are the ones they don't.

We can help you find these invisible failures.

[hello@bigspin.ai](mailto:hello@bigspin.ai)