OpenAI banned goblins.

Buried in ChatGPT's code was an explicit rule: never mention goblins, gremlins, raccoons, trolls, or pigeons. This is what it tells us about AI writing.

By the end of this, you'll have a free PDF checklist and a skill file to run on everything you produce. Links below.

AI with Kyle

SLIDE 2

WHY GOBLINS?

The Real Story

AI with Kyle

openai.com/index/where-the-goblins-came-from

OpenAI Research Products Safety Company

May 9, 2025 • Research

Where the goblins came from

How a harmless-seeming personality trait led to unexpected goblin talk — and what we learned about model behavior along the way.

Authors	Reading time	Share
OpenAI	8 min	

1. Nerdy personality reward scored creature-heavy metaphors higher in **76.2%** of audited datasets
2. Nerdy = only **2.5%** of responses, but produced **66.7%** of all goblin mentions
3. By GPT-5.4, goblin mentions in Nerdy mode were up **3,881%** vs GPT-5.2
4. The behaviour leaked into the base model via fine-tuning data
5. **Fix:** hard system-prompt ban — 'Never talk about goblins, gremlins, raccoons, trolls, ogres, or pigeons'

This isn't just a funny story. It's proof that AI models develop verbal habits — patterns they default to — without anyone intending it. Those patterns are what we call **tells.**

Source: openai.com/index/where-the-goblins-came-from

WHAT ARE AI TELLS?

An **AI tell** is a language pattern that appears so consistently in AI output that it signals — consciously or not — that a machine wrote it.

Vocabulary

Structure

Tone

Tells exist because AI models are trained on the same data, rewarded for the same signals, and optimised against the same human preferences. Predictable input → predictable output.

→ **The PDF at the end is a ready-to-paste checklist of these.**

SLIDE 4

THE HIT LIST

Famous AI Tells

AI with Kyle

WALL OF SHAME

① VOCABULARY

delve
tapestry
nuanced
navigate
foster
underscore
leverage
meticulous
pivotal
vibrant
realm
intricate
robust
showcase
ensure
multifaceted

② STRUCTURE

"It's not X, it's Y"

"You're not X — you're Y"

- bullet-everything

markdown headers for a two-sentence answer

"Key Takeaways:"

"In conclusion."

③ TONE / OPENERS

"Certainly!"

"Absolutely!"

"Great question!"

"I'd be happy to help"

"It's important to note"

"It's worth mentioning"

④ PUNCTUATION

— Em dash overuse — like this — at every clause boundary

" Curly quotes

> **Paul Graham,** *April 7 2024:*
> *"Someone sent me a cold email... I noticed it used the word delve. It's a sign that text was written by ChatGPT."*

> **Kobak et al., Science Advances 2025:**
> *at least* **13.5%** *of 2024 biomedical abstracts show LLM-assisted writing fingerprints."*

BEFORE

In today's rapidly evolving landscape, it is pivotal to leverage innovative solutions that foster growth and drive impactful outcomes. This multifaceted approach ensures stakeholders can navigate complex challenges with confidence and maximize value across the board. It's important to note that collaboration, when underscored by clear communication, showcases our commitment to excellence and paves the way for sustained success in this intricate and dynamic realm.

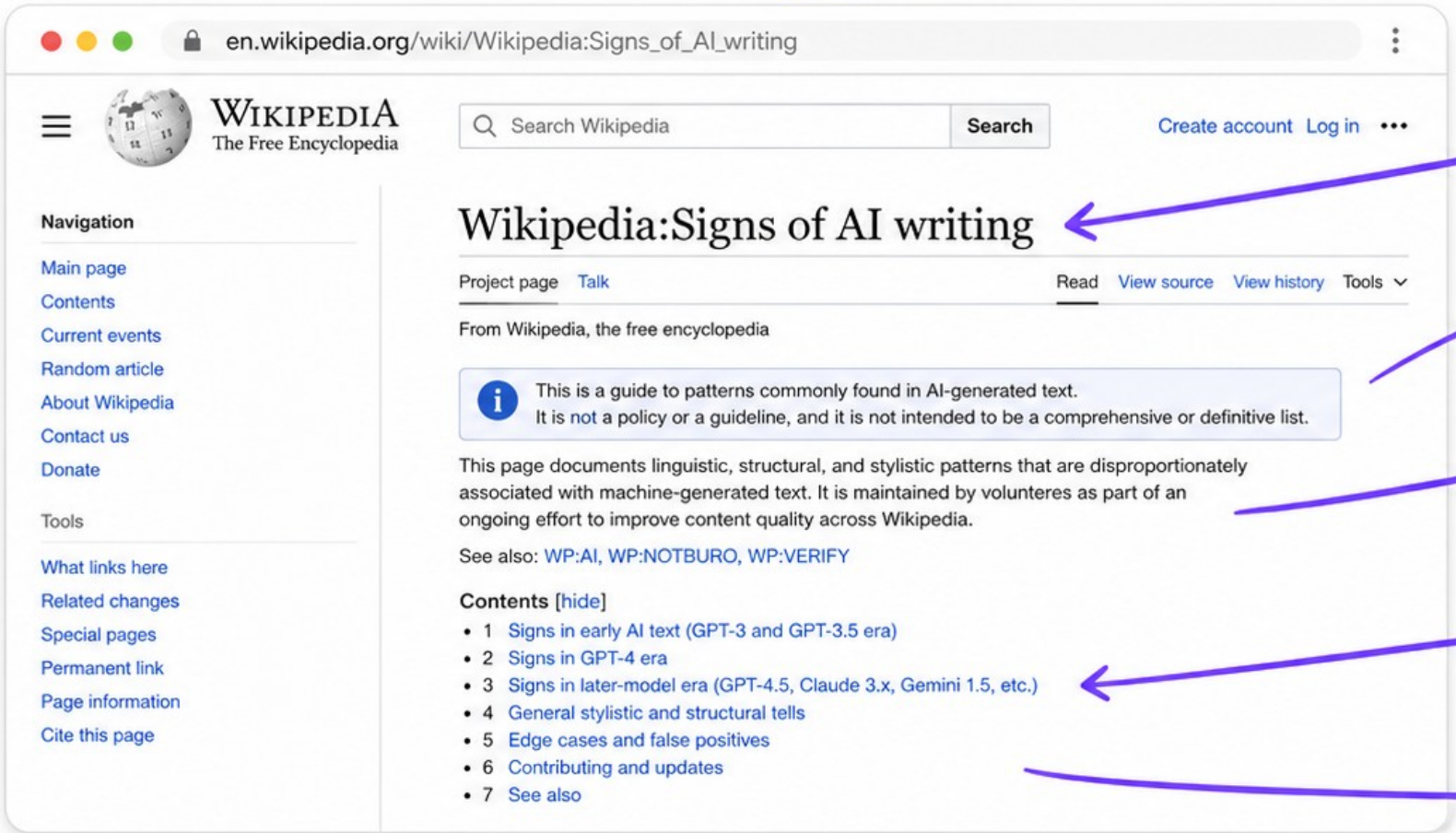
AFTER

The world is changing fast. To keep up, we need practical solutions that drive real results. When we communicate clearly and work together, we can solve hard problems, create value, and build long-term success.

SLIDE 5

WIKIPEDIA'S LIVING LIST

AI with Kyle



Wikipedia:Signs of AI writing

~15,000 words, maintained by a live editor community

Organised by era:
GPT-4 tells vs later-model tells

Wikipedia banned AI-generated article content in March 2026 (WP:LLM)

Related: WikiProject AI Cleanup

This page is updated by humans whose job is detecting AI.

It's probably the most useful free resource in existence on this topic. **Bookmark it.**

In the PDF I've compressed the key terms from this page into a **paste-ready format.**

Source:
en.wikipedia.org/wiki/Wikipedia:Signs_of_AI_writing

SLIDE 6

WHAT TO DO WITH THIS

AI with Kyle

Belt and Braces

1 Layer 1 — Before You Write

- Compress the Wikipedia list into a short markdown file
- Paste it into your Claude Project instructions or system prompt
- Format: ‘Avoid the following words and structures: [list]’ + ‘Write in [your voice description]’
- Works for any recurring content type: newsletters, scripts, posts

2 Layer 2 — After You Write

- Create a dedicated ‘AI Signs’ skill / custom prompt
- After drafting anything, run: ‘Check this for AI tells using the following list. Flag anything and suggest a human alternative.’
- **Belt:** the system prompt stops them going in
- **Braces:** the skill catches any that slipped through

AI DETECTORS: BASICALLY USELESS

OpenAI launched its **AI Text Classifier**	**Jan 31, 2023**
Killed it on	**July 20, 2023** — *"due to its low rate of accuracy"*
Current tools:	GPTZero, Turnitin, Copyleaks
Accuracy claims often range from	**65–90%** depending on test conditions
Stanford 2023:	seven AI detectors flagged **61.3%** of TOEFL essays as AI-generated; **97%** flagged by at least one tool
These tools **disproportionately penalise** non-native English speakers and anyone who writes simply and clearly	
No major academic institution recommends using detector output as sole evidence	

> **It is a race that can't be won. The only winning move in the AI-Writing vs. AI-Detection war is not to play.**

SLIDE 8

A MOVING TARGET

2022 — 'delve' takes off

Apr 2024 — Paul Graham tweet goes viral

Feb 2025 — em dash discourse peaks

Apr 2025 — GPT-4o sycophancy crisis; OpenAI rolls back an update

Nov 2025 — OpenAI ships em dash fix

Mar 2026 — Wikipedia bans AI article content

Apr 2026 — Goblins

→ **Tells shift as models are updated.** Delve has dropped significantly since 2025.

What's flagged today may be clean in a year.
What's clean today may be the next delve.

Two implications

1. If you're trying to spot AI: read widely, keep your eyes open, and check the Wikipedia page regularly.
2. If you're trying to write without tells: the skill file isn't a one-time fix. Date it, update it, re-run it.

→ **The skill file I'm giving you is designed to be updated.** When the list changes, swap it out.

DOES ANYONE ACTUALLY CARE?

The disclosure paradox

94%

of news consumers say journalists should disclose AI use

Trusting News, 2024

42%

trust a story less after seeing an AI disclosure

Trusting News, 2025 A/B follow-up

They want to know. And when they know, they trust less.

CONTEXT

- In one study, about 50% preferred not to read AI-generated content — but about 50% didn't care or couldn't tell
- Pew, Sept 2025: many people say it matters to tell, but don't trust their own ability to spot it
- Readers tolerate AI for grammar and formatting more than for opinion, storytelling, or anything meant to be 'you'

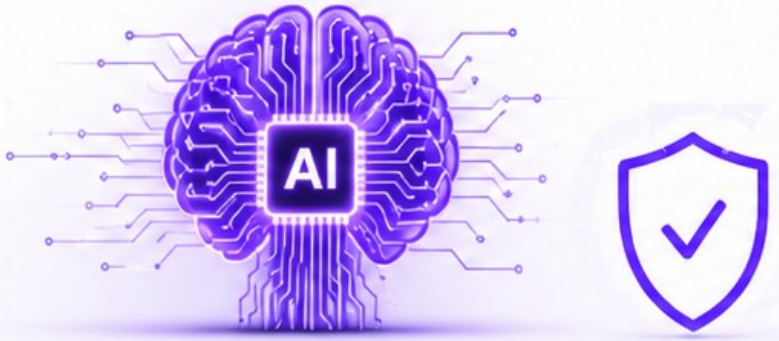
Is the problem the AI writing — or the expectation that it was human?

SLIDE 10

AI with Kyle

THE GOBLINS ARE GONE.

Something else is already taking their place.



2 FREE RESOURCES

AI Tells Cheat Sheet (PDF)
the compressed Wikipedia list, paste-ready, print-ready

AI Signs Skill File (.md)
drop into Claude Projects or your system prompt as a second-pass checker

This list will date.
The skill file is **designed to be updated**.
Treat it as a **living document**.

Links below video, in newsletter, pinned comment.