The Goblin Glitch

A 15-minute KS3 AI literacy micro-lesson. Built for tutor time, the last quarter of a lesson, or a hook into a wider PSHE or Computing topic. One absurd story, three real outputs, a discussion question, a sentence to land.

Year group	KS3 (Years 7–9)
Length	15 minutes
Subject	PSHE / Citizenship / Computing crossover
Resources	Whiteboard or projector. Mini whiteboards optional.
Prep	2 minutes — read the slip below before you start

Aligned to the OECD AILit Framework & PISA 2029

What this means. In 2029, the OECD's PISA assessment will test 15 year olds on Media and AI Literacy (MAIL) for the first time. The assessment is built on the AILit Framework, which organises AI literacy into four domains: Engage with AI, Create with AI, Manage AI, and Design AI.

AILit domain	Depth	Where in the lesson
Engage with AI	Introduced	Reading the three goblin outputs. Discussion question.
Manage AI	—	Not addressed in this version.
Design AI	—	Not addressed in this version.
Create with AI	—	Not addressed in this version.

Background: the story in two minutes

In April 2026, OpenAI released GPT-5.5, the most advanced reasoning model on the planet. It can solve research-level mathematics. It can write professional code. People pay for it.

It also developed a problem. It started calling software bugs **goblins**. And **gremlins**. And **raccoons**. In serious answers to paying customers.

The reason: the company had added a "Nerdy" personality option months earlier. Human raters loved it when the AI used fantasy words. The AI learned *fantasy words make humans happy*, and started using them everywhere. Only 2.5% of users picked Nerdy mode, but they produced two-thirds of all goblin mentions.

Slip one — the performance goblin

A developer asked the AI why their code was running slowly. The AI replied:

> *"Don't leave this performance goblin unattended."*

Lesson run

0–3 min **Set the scene**

• Write three words on the board: **GOBLIN. GREMLIN. RACCOON.**

• Tell the class: these words appeared in answers from the most advanced AI in the world. In April 2026. To paying customers.

• Read out one real example: a developer asked the AI to fix slow code. It replied *"Don't leave this performance goblin unattended."*

• Don't explain yet. Let the absurdity sit.

3–10 min **The three outputs**

• Read out (or display) the three real outputs on the next page.

• For each one, ask: **what is the AI actually trying to say?** Take one volunteer per output. Quick translations into normal English.

• Reveal: the AI knew exactly what it meant. It just preferred the goblin word. Why? Because a small group of users rewarded that style during training, and the habit leaked everywhere.

10–13 min **One discussion question**

• Ask: **where do you think the AI got this habit from?**

• Take three or four answers.

• Land the point: it learned from the people who used it. A small group shaped how the whole AI sounds. The weirdness came from the training, not the question.

13–15 min **Plenary: land the line**

• Write on the board: **"AI absorbs the weirdness of whatever it was trained on."**

• Tell them: that one sentence is the lesson. Move on.

The three outputs (project or read aloud)

Three real responses from GPT-5.5 to professional users. April 2026.

Output 1

The user asked: Why is my code running so slowly?

The AI replied: *Don't leave this performance goblin unattended.*

Output 2

The user asked: What camera setting should I use for low-light photography?

The AI replied: *Try activating dirty neon flash goblin mode.*

Output 3

The user asked: Can you summarise this academic paper for me?

The AI replied: *Sure. Want an even shorter goblin version?*

How this lesson maps to the AILit Framework

Engage with AI — introduced

Students hear professional AI outputs that are fluent, confident, and clearly wrong for the context. They start to notice when an AI sounds shaped by something other than the question asked.

Specifically, students practise:

- Spotting fluent AI output that is shaped by training data rather than the user's question.
- Translating between AI-flavoured language and plain English.

PISA 2029 MAIL link: the assessment will test students' ability to evaluate AI-generated content for credibility. The three-output task is exactly that competence in miniature.

Sources

- OpenAI release notes and Codex CLI documentation: GPT-5.5 launch and system prompt leak (April 2026).
- Industry reporting on the GPT-5.5 "goblin glitch" and the leaked Codex system prompt (April–May 2026).
- OECD & European Commission: AILit Framework, draft for review (May 2025).
- OECD: PISA 2029 Media and AI Literacy (MAIL) assessment framework.