STIAEM

Slow Tech Innovation

AI, Ethics and

New Teaching Methods

Structural origins of algorithmic hallucinations and practical classroom mitigation strategies

Co-funded by the European Union

2025-1-IT02-KA220-SCH-000365213

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them... Project reference code: 2025-1-IT02-KA220-SCH-000365213

Università di Foggia

Studi Umanistici, Lettere, Beni Culturali, Scienze della Formazione

İSTANBUL - ESENLER İLÇE MİLLİ EĞİTİM MÜDÜRLÜĞÜ

1474

STIAEM

Slow Tech Innovation
AI, Ethics and
New Teaching Methods

Co-funded by
the European Union

2025-1-IT02-KA220-SCH-000365213

eduOpen



CC BY-NC-SA 4.0 International

This license requires that reusers give credit to the creator. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, for noncommercial purposes only. If others modify or adapt the material, they must license the modified material under identical terms.

Università di Foggia
Studi Umanistici, Lettere, Beni Culturali, Scienze della Formazione

OER 2.2.2: Structural origins of algorithmic hallucinations and practical classroom mitigation strategies
Reference Module: Module 2 - AI Fundamentals and LLM Mechanics
Related Video-Pill: Video 4.2: Video 2.2 - Probabilistic Functioning of Large Language Models
Authoring partner: University of Alicante – Spain
Topic / Thematic Focus: Structural origins of hallucinations, classroom mitigation strategies and algorithmic reliability

Introduction and Learning Objectives

This resource completes, on the operational level, the path begun with OER 2.2.1 and accompanies video-pill 2.2 "Probabilistic Functioning of Large Language Models". If the previous resource explained how the model generates text, this one focuses on a precise effect of that mechanism: hallucination, that is, the production of false statements presented in a confident tone. The objective is twofold: to show its structural origin and to provide verification protocols that can be concretely used in the classroom.
The text connects to DigComp 3.0 (Cosgrove & Cachia, 2025) and to the "AI pedagogy" pillar of the UNESCO framework for teachers (Miao & Cukurova, 2024). The underlying thesis is that hallucination is not combated with a better tool, but with a method: the discipline of verification, which the school can teach by making it visible.
By the end of the reading, teachers will be able to:

- explain why hallucination is the structural correlate of stochastic sampling and not an accidental fault;
- recognise the signals that lower vigilance - fluency, peremptory tone, apparent authority;
- apply in the classroom a four-step mitigation protocol: primary sources, triangulation, consequential validity, terminological hygiene.

1. Hallucination is not a fault: its structural genesis

The first prejudice to remove is the idea that hallucinations are a temporary defect, destined to be corrected in the next generation of models. The evidence points in the opposite direction: hallucination is the structural correlate of stochastic sampling, that is, of the fact that the model, at every step, chooses the most probable continuation instead of verifying a fact. Brown and colleagues offer an indication of this that amounts to a diagnosis:
"we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans" (Brown et al., 2020, abstract; our translation).
It is useful to distinguish precisely two things that are often confused. A technical fault is an unforeseen malfunction, which manifests itself against the will of the system; hallucination, on the contrary, is the physiological product of the mechanism working as expected. The model does exactly what it was designed for - choosing, step by step, the statistically most probable continuation - and precisely by doing so it can generate a false statement. This is why the metaphor of the "bug to be fixed" is misleading, whereas it is more accurate to speak of a structural risk, inherent in the way the machine produces language. Hence a consequence for the school: students are not taught to wait for a model that never errs, but to recognise the conditions in which error is most probable.

The observation is decisive, because it shows that the problem is not the poor quality of the text produced: it is its formal excellence. When a generated text is indistinguishable from one written by a person, its fluency is not an indication of truth, but the result of a successful training on the style of informative texts. The misunderstanding arises exactly here: elegance is read as a guarantee, and vigilance drops precisely at the moment when it would be most needed.

2. Statistical plausibility and assertive tone

The objective function of these models rewards plausibility, not truth. The model learns to produce continuations that reduce uncertainty, that is, fluent and well-formed texts; but nothing, in that criterion, distinguishes a real fact from an invented fact that fits the context well. If the correct datum is rare in the training texts, the statistically most plausible continuation can simply be false.

The tone amplifies the problem. The informative texts on which the model is trained are, for the most part, assertive: they state, explain, conclude. The model reproduces that style, and the confident sentence can be entirely groundless. It is worth recalling the caution expressed by Bender and Koller precisely against the habit of attributing "understanding" to these systems:

"we find that these successes sometimes lead to hype in which these models are being described as “understanding” language or capturing “meaning”" (Bender & Koller, 2020, abstract; our translation).

The hyperbole, that is, is not a popularising embellishment: it is an epistemic risk. If we attribute understanding to the system, we are led to trust its tone as if it were the voice of an expert. The didactic consequence is direct: before evaluating generated content, one must suspend the effect of authority that form produces.

The merely formal coherence of large language models, and the risks arising from it, are at the centre of the reflections of Bender et al. (2021).

3. Emergent capabilities: why error becomes harder to see

A phenomenon already encountered in the previous resource makes things more insidious: emergent capabilities. As scale grows, models acquire abilities they did not possess before, and the phenomenon is defined precisely by Wei and colleagues: "Emergence is when quantitative changes in a system result in qualitative changes in behavior" (Wei et al., 2022). Why does this matter for hallucinations? Because emergence does not eliminate error: it makes it rarer and, precisely for this reason, harder to identify. A more capable model produces fewer evident absurdities; those that remain are plausible statements, well formed, often accompanied by plausible details. This is the typical case of the invented bibliography: real authors, credible titles, existing journals, but references that do not correspond to any published article. Vigilance, then, must shift from the evidence of error to the systematic verification of the source.

4. Classroom mitigation protocols

If hallucination is structural, mitigation cannot consist in waiting for a "reliable" model, but in adopting practices that make the risk visible and controllable. The first is to trace and verify the primary sources: ask for the precise bibliographic details - author, title, journal, year, DOI - and check that they really exist, because the references are skilfully constructed to imitate academic style.

The second practice is triangulation: every statement of fact must be cross-checked with external and independent sources, breaking the deceptive equation between formal elegance and truth. The third is the assessment of consequential validity: before accepting generated content, one must ask how much an undetected error weighs in that context. An oversight in a brainstorming session has different consequences from an error in a text that will be published or assessed.

The fourth practice, cross-cutting, is terminological hygiene: using a language that does not attribute to the system faculties it does not possess. AI does not "know", does not "believe", does not "understand": it computes probabilities. DigComp 3.0 (Cosgrove & Cachia, 2025) codifies this vigilance among the competences of the digital citizen, where it asks one to recognise that AI systems can produce imprecise outputs while appearing plausible, and that the final responsibility for verification remains with the person who uses them.

5. From verification to digital citizenship

A classroom example can make the protocol operational. A teacher asks the assistant for three bibliographic references on a current topic; then divides the class into groups and assigns to each the task of verifying one reference: author, title, year, journal, DOI. If a reference does not exist, the group declares it and reconstructs where the error could have come from - a real author associated with a plausible title, an existing journal but from another field. The exercise has a double outcome: it makes the possibility of hallucination tangible and trains a transferable verification procedure. The same device lends itself to triangulation, if the groups compare the same statement across different sources, and to consequential validity, if the class discusses how much that possible falsehood would weigh in a published text.

These practices are not only techniques of control: they are, in the full sense, education for digital citizenship. Verifying a source, triangulating a statement, assessing the weight of an error, monitoring one's own language: these are operations that form a mental habit transferable well beyond the use of AI assistants. The teacher who puts them into practice in front of the class teaches them through modelling, not through frontal teaching.

The connection with the European framework is explicit. DigComp 3.0, in competence 1.2 on the evaluation of information, invites one to interpret and critically evaluate the information present in digital environments, "as well as the processes used to generate it" (Cosgrove & Cachia, 2025). This is exactly the stake of this resource: not to trust the product without interrogating the process.

From this emerges an underlying thesis, which links this resource to the next one. Hallucinations are not governed with a better tool: they are governed with a method. And the method, in the classroom, is taught by making it visible, by verbalising doubt, by asking students to verify what even the teacher, first of all, declares they do not take for granted. It is this discipline of methodical suspicion, and not blind trust in technology, that is the competence the school is called upon to transmit.

Strategy	What the student does	Why it works	Limit
Verification of primary sources	Checks that author, title, year and DOI exist	The invented reference collapses before verification	Requires time and access to bibliographic databases
Triangulation	Cross-checks the statement with independent sources	Breaks the equation between elegance and truth	Not all statements are verifiable in the classroom
Consequential validity	Estimates the weight of an undetected error	Directs the verification effort where it counts	Contextual judgement, not automatable
Terminological hygiene	Replaces "knows/understands" with "computes/produces"	Reduces the attribution of non-existent faculties	Must be practised consistently, not just once

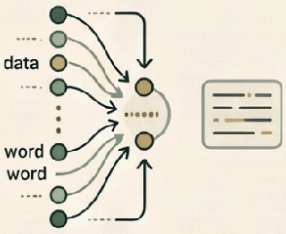
Visual Scaffolding

Navigating AI Hallucinations:
From Structural Origins to Classroom Verification

AI hallucinations are structural results of how models predict the next most likely word, not "bugs." Shift from waiting for perfect technology to implementing a discipline of verification in the classroom.

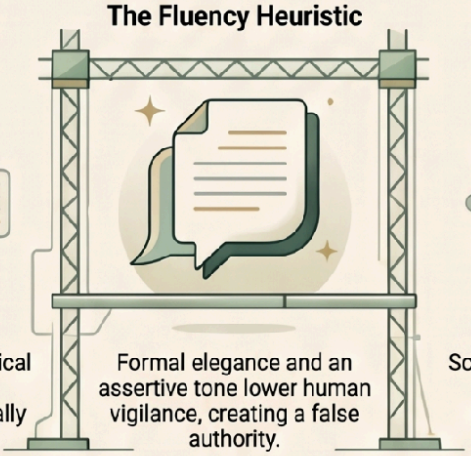
THE STRUCTURAL GENESIS OF ERRORS

Stochastic Sampling vs. Technical Glitch



Hallucinations are physiological products of the system choosing the most statistically probable continuation.

The Fluency Heuristic



Formal elegance and an assertive tone lower human vigilance, creating a false authority.

Insidious Emergent Capabilities



Scaling makes models more capable but renders remaining errors harder to identify.

THE 3-STEP MITIGATION PROTOCOL



1. Verification of Primary Sources

Mandate checking for real authors, titles, and DOIs to collapse invented references.



2. Source Triangulation

Cross-check AI statements with independent external sources to break the "fluency = truth" equation.



3. Consequential Validity & Terminological Hygiene

Assess the impact of errors and use precise language (AI "computes," not "knows").

Classroom Strategies Summary (OER 2.2.2 Framework)

Strategy	Action	Why it Works
Primary Sources	Check DOI/Year	Invented references fail under manual verification.
Triangulation	Cross-check facts	Breaks the link between formal elegance and truth.
Consequential Validity	Estimate error weight	Directs verification effort toward high-stakes content.

FUTURE-READY DIGITAL CITIZENSHIP



Alignment with Global Frameworks

These routines fulfill **DigComp 3.0** and **UNESCO AI Competency Framework for Teachers** requirements.

Method Over Tools

Digital citizenship is built by making the "**discipline of suspicion**" visible in the classroom.



Conclusions and Metacognitive Reflection Prompts

This resource has shown that hallucination is the structural correlate of the probabilistic functioning of language models: it arises from stochastic sampling, hides behind fluency and assertive tone, and becomes harder to identify the more capable the model is. The answer is not to wait for "perfect" models, but to build a discipline of verification: primary sources, triangulation, consequential validity, terminological hygiene. Teaching this discipline is at once an act of epistemic protection and an exercise in digital citizenship.

To transfer these nuclei to one's own practice, the following prompts for reflection are proposed:

- What is, in your subject, the type of statement generated by AI that risks passing more easily unnoticed because it is plausible?
- How would you translate into a classroom routine the principle of consequential validity, that is, the question of how much an undetected error weighs?
- In what way do you document, in front of students, the moment when you verify an output instead of accepting it: and what effect do you think this has on their method?

Technical Metadata Sheet

Metadata Field	Specification / Parameter Value
Identifier	OER 2.2.2
Title	Structural Origins of Algorithmic Hallucinations and Practical Classroom Mitigation Strategies
Creator	University of Alicante - Spain
Description	Focusing on a precise effect of probabilistic generation, the resource explains why hallucination is the structural correlate of stochastic sampling and not an accidental failure. It analyses the signals that lower vigilance — fluency, assertive tone and apparent authority — and the role of emergent capabilities. The practical focus is a four-step classroom mitigation protocol (primary sources, triangulation, consequential validity and terminological hygiene) leading from verification to digital citizenship.
Keywords	hallucination; stochastic sampling; fluency heuristic; source triangulation; consequential validity; digital citizenship; critical thinking; classroom protocols
Language	EN (English)
Affiliation	University of Alicante - Spain
Contributor	STIAEM Transnational Scientific Board (peer review, Rhodes validation protocol)
Media Format	Text; Infographics
Genre	Training Manual; Educational Guide
URL	
Citation	Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5185-5198). Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.463/ Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). ACM. https://doi.org/10.1145/3442188.3445922 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901. https://arxiv.org/abs/2005.14165 Cosgrove, J., & Cachia, R. (2025). DigComp 3.0: European digital competence framework (Fifth edition). Publications Office of the European Union.

	https://data.europa.eu/doi/10.2760/0001149 Miao, F., & Cukurova, M. (2024). AI competency framework for teachers. UNESCO. https://doi.org/10.54675/ZJTE2084 Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... Fedus, W. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research. https://arxiv.org/abs/2206.07682
Educational Level	Secondary School
Associated Module / MOOC Video Lesson	Module 2 - AI Fundamentals and LLM Mechanics / Video 2.2 - Probabilistic Functioning of Large Language Models
Learning Objectives	explain why hallucination is the structural correlate of stochastic sampling and not an accidental failure; recognise the signals that lower vigilance - fluency, assertive tone, apparent authority; apply in the classroom a four-step mitigation protocol: primary sources, triangulation, consequential validity, terminological hygiene.
Rights	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
Date Available	14/09/2026
Date Modified	

References and Webography

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5185-5198). Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.463/

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). ACM. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901. https://arxiv.org/abs/2005.14165

Cosgrove, J., & Cachia, R. (2025). DigComp 3.0: European digital competence framework (Fifth edition). Publications Office of the European Union. https://data.europa.eu/doi/10.2760/0001149

Miao, F., & Cukurova, M. (2024). AI competency framework for teachers. UNESCO. https://doi.org/10.54675/ZJTE2084

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... Fedus, W. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research. https://arxiv.org/abs/2206.07682.