STIAEM

Slow Tech Innovation

AI, Ethics and New Teaching Methods

The Mechanics of Large Language Models: Deconstructing the 'Statistical Mirror'

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 2.2.1: The Mechanics of Large Language Models: Deconstructing the 'Statistical Mirror'
Reference Module: Module 2 - AI Fundamentals and LLM Mechanics
Related Video-Pill: Video 2.2 - Probabilistic Functioning of Large Language Models
Authoring partner: University of Alicante – Spain
Topic / Thematic Focus: LLM mechanics, statistical mirror deconstruction and language modeling

Introduction and Learning Objectives

This resource accompanies, together with OER 2.2.2, video-pill 2.2 "Probabilistic Functioning of Large Language Models" and constitutes its conceptual in-depth study. Its purpose is to make comprehensible, even to those without a background in computer science, the mechanism by which a large language model generates its own text; and to draw from that mechanism a precise teaching consequence: the fluency of an answer is not proof of its truth.

The text connects to DigComp 3.0 (Cosgrove & Cachia, 2025), the European framework of digital competences in which AI-related competence runs, explicitly or implicitly, through all competences, and to the "AI pedagogy" pillar of the UNESCO framework for teachers (Miao & Cukurova, 2024). The metaphor that organises the resource is that of the statistical mirror: the model reflects the patterns sedimented in the training data, without any anchoring to the reality it describes.

By the end of the reading, teachers will be able to:

- describe the autoregressive generation of text and explain why the model chooses each word on the basis of probability;
- distinguish formal coherence from factual accuracy, recognising in the "statistical mirror" a simulation and not an understanding;
- recognise, on the basis of few-shot learning and emergent capabilities, that scale produces statistical jumps and not mental faculties.

1. The statistical mirror: formal coherence, not factual accuracy

A teacher asks an artificial intelligence assistant for a list of peer-reviewed scientific articles on working memory. The answer arrives within seconds and is impeccable: real authors, plausible titles, authoritative journals, consistent years. And yet those articles, in whole or in part, may not exist. This is not a temporary defect destined to disappear with the next version of the model: it is a structural property of the way in which the text is generated.

Brown and colleagues, in presenting the GPT-3 model, describe its functioning as follows:

"we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting" (Brown et al., 2020, abstract; our translation).

The decisive point is not the size, but the absence of a consultable base of facts. The model does not compare its sentences with the state of the world: it computes probability distributions over sequences of words. If a plausible sequence does not correspond to any article actually published, the system has no way of noticing, because nothing, in its architecture, distinguishes the true from the plausible. It is here that the metaphor of the statistical mirror becomes precise: the system reflects the patterns of the texts on which it

was trained, and in that reflection an artificially constructed reference and a real one are indistinguishable.
For the teacher there follows an immediate criterion of deontological hygiene: an elegant bibliographic list does not prove that the sources exist, and a fluent synthesis does not prove that the facts it contains are true. Form, in these systems, is a clue about the training, not about reality.

2. Autoregression: generating one word at a time

The underlying mechanism is the same as that of the smartphone's autocorrect, but on an incomparably larger scale. Given a text, the model computes which is the most probable next element and chooses it; then that fragment re-enters the input, and the computation starts again. This is autoregressive generation: word after word, the text grows by accumulation of local choices. The term "autoregressive", used by Brown and colleagues to characterise GPT-3, indicates exactly this: every output depends on the previous outputs, which in turn become input.
From this derives a property with direct consequences in the classroom. Every decision of the model is a statistical bet on the next fragment, informed by the context but not verified: fluency is not a sign of understanding, it is the effect of an objective function that rewards statistically plausible continuations. When an assistant is used to write a synthesis, the fluency of the result says that the model has learned the style of similar texts well; it says nothing, however, about the truth of what it asserts.
Distinguishing form and accuracy is therefore the first critical gesture that this resource intends to make available to the teacher. It is not an abstraction: it is the competence that allows one to decide when an output can be accepted as it is and when, instead, it must be verified before entering teaching material or an assessment.

3. Few-shot learning: the simulation of adaptation

One of the most surprising properties of large models is the ability to perform new tasks starting from very few examples provided in the text of the request. The description of GPT-3 compares this ability with the human one:
"humans can generally perform a new language task from only a few examples or from simple instructions... GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model" (Brown et al., 2020, abstract; our translation).
The didactically relevant point is the misunderstanding that this similarity produces. It comes naturally to say that the model "learns on the spot", as a pupil would. But few-shot learning works differently: the pupil abstracts a rule and applies it to new cases; the model, simply, prolongs the context by maximising statistical plausibility. It seems to adapt, it does not adapt. And it is precisely when the similarity is most convincing that critical vigilance is most necessary: the sensation of being in front of an interlocutor who understands is the effect of a simulation, not the proof of an understanding.

4. Form and meaning: why the model does not "understand"

This distinction has a precise theoretical formulation. Bender and Koller, in a contribution presented at the ACL conference of 2020, state it clearly:

"we argue that a system trained only on form has a priori no way to learn meaning" (Bender & Koller, 2020, abstract; our translation).

Their thesis is founded on an explicit definition: "We take (linguistic) meaning to be the relation between a linguistic form and communicative intent" (Bender & Koller, 2020, §1; our translation). Meaning, that is, is not a property internal to the string of text, but the relation between that form and the communicative intention of whoever produces it. A model trained only on form - on the co-occurrence of words - has no access to that relation: it never encounters the world to which the words refer. It follows, as the authors specify, that increasing the training data does not bridge the gap, because it does not introduce the missing element, namely the link with intention and with the context of use.

On this basis the formula, which has become famous, of the language model as a stochastic parrot has been consolidated: a system that reproduces with statistical competence the regularities of human language without drawing on meaning (Bender et al., 2021). For the teacher, the operational consequence is a terminological hygiene that holds in both directions: avoid saying that AI "knows", "believes" or "understands", and ask students to do the same.

5. Few-Shot Prompting Through the STIAEM Pedagogical Framework

One last property deserves attention, because it is at the origin of many recent misunderstandings. As the dimensions of the model grow, capabilities appear that are absent in the smaller versions. Wei and colleagues give a technical definition of it:

"We consider an ability to be emergent if it is not present in smaller models but is present in larger models" (Wei et al., 2022, abstract; our translation).

And they describe its form: "performance is near-random until a certain critical threshold of scale is reached, after which performance increases to substantially above random" (Wei et al., 2022, §2; our translation). The jump occurs when a purely quantitative increase - more parameters, more data - produces an observable qualitative change, for example in multi-step arithmetic tasks. The authors speak, not by chance, of a genuine phase transition of behaviour.

The correct reading of this phenomenon excludes the most suggestive and most mistaken explanation, that of a supposed algorithmic self-awareness. Emergence indicates only that, once a certain threshold is passed, the system encodes finer statistical regularities; it remains governed by sampling and acquires no understanding of the world. For the school it is a precious didactic opportunity: it shows that "emerges" does not mean "a mind is born", and that surprising performances can coexist with the absence of the ability to distinguish the true from the plausible. It is this conviction that prepares the ground for the next resource, dedicated to the structural origin of hallucinations.

Concept	What the model does	What it does not do	Source
Autoregressive generation	It computes and chooses the most probable next word	It does not plan the text or verify its truth	Brown et al., 2020

Few-shot learning	It prolongs the context with the examples provided	It does not abstract a rule like a pupil	Brown et al., 2020
Statistical mirror	It reflects the patterns of the training data	It does not anchor itself to the reality it describes	Bender & Koller, 2020; Bender et al., 2021
Form and meaning	It processes statistical relations between forms	It has no access to communicative intention	Bender & Koller, 2020
Emergent capabilities	It encodes finer regularities beyond a threshold of scale	It acquires neither understanding nor consciousness	Wei et al., 2022

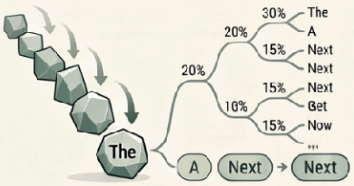
Visual Scaffolding

THE STATISTICAL MIRROR:
DECONSTRUCTING LARGE LANGUAGE MODELS

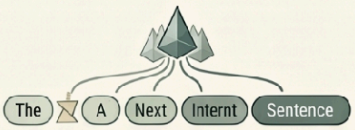
Large Language Models (LLMs) operate as a 'statistical mirror,' reflecting patterns from training data rather than accessing truth. This infographic explains the core mechanics—from autoregression to emergent capabilities—that allow AI to simulate human-like fluency without genuine understanding.

THE MECHANICS OF COHERENCE

AUTOREGRESSIVE GENERATION



Models grow text by choosing the next most probable word based on local distributions.



Models grow text by data-at word - trainting senens and enconinicated in a real persoable accor is low context to trutoamo: next precious roads of word in unorstantions.



FORM VS. COMMUNICATIVE MEANING



Models process statistical word relations but lack access to real-world intent or grounding.

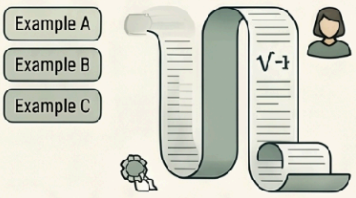
FLUENCY IS NOT FACTUALITY



Elegant syntax reflects training patterns, not a consultable base of verified facts.

SIMULATED INTELLIGENCE & SCALE

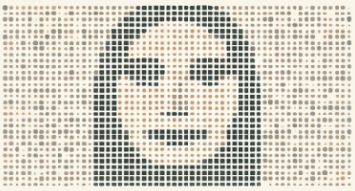
FEW-SHOT LEARNING VS. RULE ABSTRACTION



Few-Shot Learning **Rule Abstraction**

Models prolong context through examples instead of abstracting rules like human pupils.

EMERGENT CAPABILITIES



Qualitative jumps in performance are statistical phase transitions, not algorithmic self awareness.



DATA TABLE

STATISTICAL MIRROR

What the Model Does	What it Does Not Do
Reflects training data patterns	Anchor itself to described reality

AUTOREGRESSION

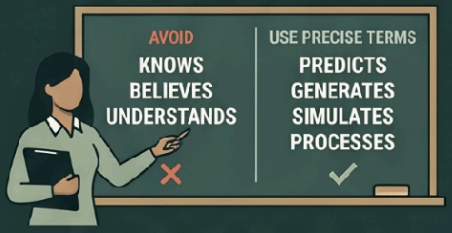
Computes the probable next word	Plan text or verify truth

EMERGENCE

Encodes finer scale regularities	Acquire understanding or consciousness

DEONTOLOGICAL HYGIENE FOR EDUCATORS

TERMINOLOGICAL HYGIENE



Avoid saying AI 'knows' or 'believes'; require students to use precise technical terms.

THE VERIFICATION IMPERATIVE



Always verify AI-generated syntheses before including them in teaching or assessment materials.

Gemini Notebook

Conclusions and Metacognitive Reflection Prompts

This resource has described the probabilistic functioning of large language models: autoregressive generation, few-shot learning, the distinction between form and meaning and emergent capabilities. The common thread is a thesis that is simple and demanding at the same time: these systems produce statistically plausible texts, not guaranteed assertions. The statistical mirror reflects the data, not the world. For the teacher, understanding this is the condition for using AI as a tool of enquiry, without delegating to it what the school must form: critical judgement.

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

- In which activity of your subject does a fluent text risk being mistaken for a true text, and how could you make the difference visible?
- How would you explain, with an example drawn from your subject, what it means that the model chooses "the most probable word" and not "the right word"?
- What task would you give students to bring out the difference between simulating learning (few-shot) and truly understanding?

Technical Metadata Sheet

Metadata Field	Specification / Parameter Value
Identifier	OER 2.2.1
Title	The Mechanics of Large Language Models: Deconstructing the 'Statistical Mirror'
Creator	University of Alicante - Spain
Description	The text makes comprehensible how a large language model generates text, deconstructing the metaphor of the statistical mirror. It explains autoregressive generation, few-shot learning and emergent capabilities, distinguishing formal coherence from factual accuracy. The didactic focus is on understanding why the model simulates rather than understands, and why scale produces statistical jumps and not mental faculties.
Keywords	large language models; autoregressive generation; statistical mirror; few-shot learning; emergent capabilities; factual accuracy; AI literacy; secondary school
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., 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 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/ 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	describe the autoregressive generation of text and explain why the model chooses each word on the basis of probability; distinguish formal coherence from factual accuracy, recognising in the 'statistical mirror' a simulation and not an understanding; recognise, on the basis of few-shot learning and emergent capabilities, that scale produces statistical jumps and not mental faculties.
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., 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

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/

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