Sophia Doan

30 July 2026

Teaching Responsible AI to the Next Generation

Abstract

Elementary school students now encounter large language models (LLMs) such as ChatGPT and Gemini through homework help, search engines, and voice assistants, often before they have any framework for evaluating what these systems produce. Existing AI literacy curricula teach responsible use and introduce the idea of bias, but they generally stop short of explaining how language generation actually works. This paper argues for a linguistics-based framework that pairs simplified natural language processing (NLP) concepts – tokenization and next-word prediction – with the linguistic categories of syntax, semantics, and pragmatics. Analyzing AI-generated language through this lens can help students understand why a chatbot's answer might read smoothly and still be wrong, off-topic, or hollow. Built for elementary classrooms, the framework distills technical material that is typically reserved for older students and situates it within existing AI literacy efforts, preparing children to question AI output rather than accept it at face value.

Introduction

ChatGPT, Gemini, and similar large language models have moved language-based AI out of research labs and into ordinary classrooms, where students already turn to them for homework help, brainstorming, and casual conversation (Long and Magerko 1). Elementary school students are part of that shift, yet they rarely have the conceptual tools to make sense of how these systems generate the words on their screen or to judge whether those words deserve their trust. Left unaddressed, that gap matters: children who cannot distinguish a fluent answer from an

accurate one are poorly positioned to use AI responsibly as it becomes further embedded in schoolwork and daily life. This paper argues that a linguistics-based approach, one that treats AI-generated language as language and analyzes it accordingly, gives elementary students a more realistic and durable understanding of what these systems can and cannot do.

Background

AI literacy curricula for primary and secondary schools have multiplied over the past several years. AI4K12, a joint initiative of the Association for the Advancement of Artificial Intelligence and the Computer Science Teachers Association, organizes its K-12 guidelines around five "big ideas": perception, representation and reasoning, learning, natural interaction, and societal impact ("AI4K12"). The nonprofit aiEDU has built a similar track record, reaching hundreds of thousands of students with curricula that now include fifteen-minute "Elementary AI Explorations" activities for grades three through five (aiEDU). MIT's Day of AI, developed by the Responsible AI for Social Empowerment and Education initiative, has drawn thousands of participants worldwide since 2022 through hands-on, age-banded lessons, including modules built specifically for early elementary classrooms (Ouellette). These programs share a common emphasis: using AI responsibly, recognizing where it shows up in daily life, and understanding that it can encode bias.

What they tend to leave out is closer to the machinery: the mechanism by which LLMs actually generate text. Long and Magerko define AI literacy as a set of competencies that let people critically evaluate AI systems and interact with them effectively, not merely operate them (2). Meeting that bar requires more than knowing that a chatbot can be biased; it requires some sense of how the output was produced in the first place. LLMs generate hallucinated, overconfident, or biased text largely because they are built to predict the next most probable unit

of text given patterns in their training data, a design choice with direct consequences for reliability (Kalai et al.). Concepts from linguistics – syntax, semantics, and pragmatics – help explain why that design can produce text that sounds right without being right. Syntax governs whether a sentence is grammatically well formed, semantics governs what it means, and pragmatics governs whether that meaning fits its context (Yule). An AI-generated response can pass the first test and fail the other two, remaining fluent while wandering off-topic or stating something false with total confidence. These distinctions are well established in the field of linguistics, yet they receive little direct attention in existing K-12 AI literacy materials, which suggests real room for linguistic analysis in AI education generally and in elementary classrooms specifically.

Linguistics-Based Framework for AI Literacy

A linguistics-based framework rests on a simple premise: understanding AI-generated language requires understanding how language works and how AI systems process it, together. Neither piece alone gets students very far. Knowing that a chatbot can occasionally be "wrong" does not explain what wrong looks like or why it happens, and knowing that models predict units of text does not, by itself, help a nine-year-old evaluate a specific paragraph in front of them. Combining both gives students a working method: read the AI response, then ask whether it holds together grammatically, whether it says something true, and whether it actually answers the question that was asked. This works something like learning basic car maintenance. A driver does not need to rebuild an engine to notice when something is wrong, but a little knowledge of how the parts fit together helps them recognize when to bring in a mechanic rather than trust the dashboard blindly. AI literacy can serve a similar function, offering not technical mastery but enough of a mental model to know when a system's confident output deserves a second look.

The technical half of the framework should stay light. Tokenization – the process by which a model breaks text into smaller units before processing it – and prediction – the process by which it selects likely next units based on patterns in its training data – are the only two mechanisms this paper proposes teaching directly, and even these should be introduced through analogy rather than formal definition. A simple demonstration works well here: give students a sentence with the last word missing and ask several of them, independently, to guess what comes next. The range and overlap of their guesses mirrors, in miniature, what an LLM is doing at a much larger scale, and it makes the idea of prediction concrete without requiring any discussion of embeddings, the attention mechanisms that let a model weigh which earlier words matter most (Vaswani et al.), or the reinforcement learning from human feedback that shapes a model's later behavior after an initial round of supervised fine-tuning (Ouyang et al.). Those latter concepts belong in a middle or high school extension of this framework, not in an elementary classroom.

The linguistic half does the interpretive work. Once students have some sense of how an AI response gets generated, they can apply syntax, semantics, and pragmatics as three separate checks on any piece of AI-generated text. A worksheet built around a single chatbot response, annotated for grammar, accuracy, and relevance to the prompt, gives students hands-on practice distinguishing categories that adults themselves often conflate. This step matters because fluency and comprehension are not the same thing, and a response that sounds confident is not automatically a response that is correct.

Ethical discussion should run through every lesson rather than sit in a unit of its own. AI-generated content reflects choices made well upstream of any individual response: what data a model was trained on, whose feedback shaped its later behavior, and what kind of answers that feedback rewarded. A user's own judgment is also shaped by automation bias, the

well-documented tendency to trust automated output more than the evidence in front of a person's own eyes warrants (Goddard et al. 121). Ending each module with an open discussion that connects a lesson's technical content to its social consequences keeps students from treating fluency, or confidence, as a stand-in for trustworthiness.

Discussion

A linguistics-based approach does not replace existing AI literacy curricula so much as fill a gap they share. AI4K12, aiEDU, and Day of AI all teach students to use AI responsibly and to recognize where bias might enter a system, but none walk students through why a fluent answer can still be inaccurate. Framing that question in linguistic terms gives elementary students a concrete, repeatable method rather than a vague warning to double-check everything.

That said, the framework remains untested. Children's mental models of how AI actually works shift substantially across the elementary and middle school years: a field study of students in grades three through eight found that younger children tend to attribute AI's outputs to some kind of built-in intelligence, while older children are more likely to describe AI as recognizing patterns in the data it was given (Dangol et al.). A framework pitched at "elementary school" as a single category may need real differentiation between, say, a first grader and a fifth grader, and that differentiation has not yet been worked out or piloted in an actual classroom. Simplifying tokenization and prediction for young learners also carries a real risk of overgeneralization, since a metaphor precise enough to be useful can just as easily be precise enough to be wrong. Future work should test how these concepts land with students at different ages and revise the curriculum design accordingly.

Conclusion

As AI becomes a fixture of how children do homework, search for information, and communicate, the case for teaching them how these systems actually produce language, rather than only how to use them politely, becomes harder to ignore. Combining simplified NLP concepts with linguistic analysis gives elementary students a method for questioning AI output instead of a list of rules to memorize. The framework proposed here is conceptual rather than field-tested, and it will need real classroom trials to find out whether its ideas land the way this paper hopes they will. But a version of AI literacy that treats fluency and accuracy as separate things, rather than treating one as evidence of the other, gives students a more honest picture of what a chatbot is doing when it answers them.

Works Cited

aiEDU. "About Us." aiEDU, www.aiedu.org/about. Accessed 30 July 2026.

"AI4K12." AI4K12, ai4k12.org. Accessed 30 July 2026.

Dangol, Aayushi, et al. "Children's Mental Models of AI Reasoning: Implications for AI Literacy Education." *Proceedings of the Interaction Design and Children Conference* (IDC '25), Association for Computing Machinery, 2025, doi:10.1145/3713043.3728856.

Goddard, Kate, et al. "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators." *Journal of the American Medical Informatics Association*, vol. 19, no. 1, 2012, pp. 121–27.

Kalai, Adam Tauman, et al. "Why Language Models Hallucinate." arXiv, 4 Sept. 2025, arxiv.org/abs/2509.04664.

Long, Duri, and Brian Magerko. "What Is AI Literacy? Competencies and Design Considerations." *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, 2020, pp. 1–16.

Ouellette, Katherine. "Day of AI Curriculum Meets the Moment." *MIT Open Learning*, 26 June 2023, openlearning.mit.edu/news/day-ai-curriculum-meets-moment-0.

Ouyang, Long, et al. "Training Language Models to Follow Instructions with Human Feedback." *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 27730–44.

Vaswani, Ashish, et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

Yule, George. *The Study of Language*. 8th ed., Cambridge University Press, 2020.