2024. *Proc Ling Soc Amer* 9(3). 5852. https://doi.org/10.3765/plsa.v9i3.5852.

AI hype in the classroom

Lex Konnelly & Nathan Sanders*

Abstract. Artificial intelligence (AI) tools, especially generative tools based on large language models (LLMs), such as ChatGPT, raise critical concerns for academic integrity, for ensuring genuine assessment of student learning, and for equity. Public understanding of these tools is clouded by hype about their capabilities, as they are often treated as knowledgeable and even sentient, and thus suitable for any human task. Of particular concern for instructors is how, and how much, students rely on these tools to complete their coursework. We address some of these issues in our classrooms by reporting on a recent pedagogical initiative within the Department of Linguistics at the University of Toronto during Summer 2023. As part of the initiative, we highlight the crucial role that linguistics can play in these discussions, by shifting the focus to LLMs as objects of study that are directly relevant to the linguistics classroom and to educate students on what linguistic tasks they are and are not good at. We offer strategies and sample assignment questions to help instructors deflate AI hype and facilitate greater AI literacy by demystifying the technology.

Keywords. technolinguistics; artificial intelligence; large language models; assessment design; AI literacy

1. Introduction. There has been recent rapid development and broad availability of so-called "artificial intelligence" (AI) tools, especially generative AI tools based on large language models (LLMs). Conversational LLMs, such as OpenAI's ChatGPT (released in November 2022, with a dramatic upgrade to a GPT-4 base in March 2023) and its competitors (Claude, Gemini, Llama, etc.) have surged into the public consciousness and widespread use. The hype around AI tools has led to many misconceptions about what they can and cannot do and which human cognitive abilities they can reliably replicate.

However, at a basic level, LLMs are just mathematical functions that assign probabilities to sequences of words and morphemes. LLM-powered chatbots like ChatGPT take this a step further by generating their own novel sequences in response to human input. Popular discourse about these tools often overlooks their underlying probabilistic nature and instead regards them as knowledgeable and even sentient, a position even held by some engineers who contributed to development of the technology (De Cosmo 2022; Chalmers 2023). The promise that LLMs represent a nascent version of artificial general intelligence, a theoretical AI system possessing generalized human cognitive abilities, potentially leads users to rely on LLMs for a variety of tasks normally requiring human judgment and reasoning, such as relationship advice and medical diagnoses. Increasingly, companies are developing models expressly for such purposes (e.g. AI "therapy" agents and medical chatbots).

* For discussions and collaborations at various stages of this work, we thank Ivan Bondoc, Cristina Cuervo, Alex Gustafson, Peter Jurgec, Naomi Nagy, Boris Steipe, Dave Suarez, Lisa Sullivan, and audiences at the 2024 annual meeting of the Linguistic Society of America and at the University of Toronto Department of Linguistics Pedagogy Workshop "AI-Generated Text and Teaching in Linguistics". We also thank anonymous reviewers for their helpful feedback. Authors: Lex Konnelly, independent scholar (a.konnelly@mail.utoronto.ca) & Nathan Sanders, University of Toronto (nathan.sanders@utoronto.ca).

These tools offer many possibilities for enhancing teaching and learning by augmenting classroom activities and facilitating student engagement. However, they also raise concerns with respect to academic integrity and ensuring genuine assessment of student learning. Institutional responses to AI tools in the classroom have also been inadequate and inequitable, by often taking punitive, surveillance-based approaches that unfairly penalize some students more than others. To better understand the challenges and affordances posed by tools like ChatGPT (the most-used model at the time of study), we worked with faculty volunteers from the University of Toronto's Department of Linguistics during Summer 2023 to explore productive strategies for addressing their concerns about AI. The instructors we worked with came from multiple subdisciplines and had diverse and individualized pedagogical goals for their courses but shared common concerns which helped narrow the focus of our investigation.

In this paper, we report on this pedagogical initiative and offer strategies and sample assignment questions. We begin in Section 2 by briefly discussing why we do not recommend attempts to detect AI in student work. Instead, we offer two primary alternative strategies in assessment design: thwarting the use of AI by the nature of the assignment (Section 3) and incorporating AI into assignments (Section 4). As discussed in Section 5, we also sought to highlight the crucial role that linguistics should play in these discussions, by shifting the focus to treating LLMs as objects of study that are relevant to the linguistics classroom. This provides students with an opportunity to engage with real-world applications of linguistics and to develop greater AI literacy, by demonstrating to students what linguistic tasks these tools are and are not good at, and by extension, sparking broader conversations about how to use it intentionally.

2. Detection is not recommended. A common concern about AI tools in academia is how to determine when students are using them in an unauthorized way. There are some types of undeniable evidence. Templatic phrases like *as an AI language model* or *as of my knowledge cut off* can show up in AI responses and remain in students' submitted work if they do not bother to read it closely. This unfortunately happens even in published research; see examples available by searching PubPeer here and here. Distinctive aspects of their output may even be quantifiable, such as overuse of certain infrequent words (e.g. *delve* and *tapestry* are commonly cited culprits).

There are indeed many AI detectors that operate on this idea, but they do not work well (Elkhatat et al. 2023; and as Casal & Kessler 2023 report, humans are no better). In particular, they can trigger false positives that reflect and reinforce biases against students who already face disadvantages due to the languages they use and how they use them (Liang et al. 2023; Jaiyeola 2024). Thus, we guided instructors away from overly punitive AI detection methods that unfairly penalize students. Instead, we focused on genuine assessment strategies that are more reliable and equitable, while also teaching students about the limitations of AI tools and dispelling the notion that they can solve their assignments for them.

3. Strategy #1: Thwarting. To figure out how to design assessments that AI tools cannot easily construct viable responses for, we did an informal hands-on investigation. With our volunteer instructors' permission, we entered different kinds of their assignments into ChatGPT, evaluated the responses, iteratively tweaking the prompts to get an understanding of what ChatGPT can and cannot do well. Due to practicality and time constraints, this investigation was not formal or systematic. We simply looked for patterns in questions that the model consistently struggled with and ones it provided strong answers for (relative to the course level).

At the time of our testing in September 2023, ChatGPT effectively handled most true/false and multiple-choice questions we tried, and it provided reliable answers to some open-ended questions, especially when asked to define or discuss introductory material. These are the kinds

of questions that would also generally be resolvable through a basic internet search, though ChatGPT is more convenient for students to use.

However, ChatGPT's accuracy and coherence dropped considerably beyond first-year material. In our investigation, it became apparent that the structure of a question, the depth of the material, and the presentation of data had a large impact on the output. For example, even in first-year courses, ChatGPT typically did not perform as well on questions from later assignments in the course, since it did not have access to course-specific presentation of instructional materials (though this could be partly defeated by the student feeding lecture notes to ChatGPT prior to asking their questions). Questions from higher-level courses were also challenging for ChatGPT, as they required more discipline-specific reasoning and applications of foundational concepts to more complex and novel situations. ChatGPT also could not provide reliable answers to questions about data that were represented graphically (in the form of plots or images of word lists that would be too cumbersome to input manually), though this may not remain the case for long.

Our primary advice to instructors who want to thwart ChatGPT in assessment design is to be mindful of presentation format and level of content, and to remember that an LLM, as a probabilistic model, cannot "reason" in any human sense; it can only generate a response that is statistically likely to be accurate or relevant. Straightforward true/false or multiple-choice questions should ideally be avoided or have low weighting in favor or questions that require thoughtful application of course concepts where possible. At introductory levels, this can be difficult, because the course content is often not deep enough to require much reasoning, as students are just beginning to learn crucial terminology and concepts. However, it is possible to construct some more thoughtful questions even at an introductory level.

For example, (1) is a common type of early question in an introductory phonetics or phonology course, when students are just beginning to memorize the International Phonetic Alphabet (IPA). It is also the kind of question that ChatGPT can easily answer correctly. However, the question in (2) measures the same kind level-appropriate knowledge, but it requires students to pull together multiple pieces of information and apply reasoning skills that AI tools cannot.

(1) What is the IPA symbol for a voiced alveolar fricative?
(2) What is the IPA symbol for a consonant phone that has the same place of articulation as [n], the same manner of articulation as [ʃ], and the same phonation as [w]?

ChatGPT is also very inconsistent when applying definitions and concepts to novel language data. For example, while it can readily *define* what a minimal pair is, it struggles to *identify* minimal pairs, especially in nonce data. Thus, instead of asking students to simply define concepts, it is more effective to ask them to actively apply those concepts. A particularly intuitive route is that of error identification, which ChatGPT also struggles with. It is good at generating well-formed content and finding patterns, but it is sometimes hilariously bad at finding errors; see the Appendix for an assignment that exploits this weakness.

More importantly, attempting to thwart AI tools results in generally better questions, precisely because they require an extra bit of reasoning beyond rote memorization. This is a theme that persists throughout discussions of AI tools in education: the vulnerability of our traditional assessments to AI is a symptom of a long-standing need for an update to our pedagogy, a need that predates AI itself. So while thwarting AI may seem like the goal, in reality, we need to refresh our pedagogy anyway.

Another way to thwart the use of AI tools is through deterrence with explicit policies and expectations in the syllabus. An explicit policy gives students concrete guidelines to follow, and it also gives the instructor something to point to if an academic integrity case needs to be brought. Such a policy should include specific descriptions of acceptable use: What can students use AI tools for? What can they not use them for? How should they demonstrate accountability? All of this is up to the individual instructor (within the bounds of department and institutional policies). A complete ban may be a pedagogically valid stop-gap measure, but it is not realistic, and it may be more productive to explicitly allow AI tools in some well-defined circumstances (certain assignments, specific uses, etc.).

4. Strategy #2: Incorporation. Not all instructors we worked with were concerned with prevention. Others recognized that these tools are not going away and that they will quickly become more sophisticated and accessible, so they wanted to find ways to explicitly incorporate them. These are not mutually exclusive goals, and in fact, they work quite well together. In this initiative, we advocated for an educational approach that engages students on what LLMs are and how they work, since AI literacy becomes ever more important with the technology's increasing ubiquity.

One option to consider is using LLM outputs as objects of analysis to stimulate discussion and critical thinking about AI tools, which aids in developing discipline-specific argumentation and analytical skills. This has the advantage of showing students the capabilities and limitations of LLMs from a linguistic perspective, demystifying the technology in disciplinarily relevant ways. By explicitly incorporating LLM output in assignments, students can see firsthand the genuine challenges in computationally modeling something as complex and diverse as human language, as well as the necessity of using conversational AI tools only for tasks in which they possess the necessary expertise to validate its results.

An example of an assignment that incorporates ChatGPT directly is given in the Appendix. This assignment is an introductory syntax assignment about syntactic constituency. Traditionally, such an assignment could be a straightforward essay, which ChatGPT could generate if prompted. However, this assignment is configured to highlight the risks of relying on ChatGPT by having them evaluate its output to the individual essay prompts. ChatGPT makes many errors, and when prompted with the entire assignment, its performance is even worse. It cannot identify its own errors, and it generates new errors. Importantly, the assignment still also requires students to understand syntactic constituency. It therefore serves two purposes: an assessment of students' grasp of the course material and a learning opportunity to advance their AI literacy. This kind of assignment could be readily adapted to almost any linguistics course, since data-driven analytical arguments are commonplace across sub-fields.

We also considered how to productively use AI tools as tools for doing work in linguistics (and in education in general), however, given time constraints and the interests of the volunteers, we left this topic for future work.

5. Linguistics and AI literacy. Part of linguistics education is debunking myths, requiring students to unlearn what they think they know about how language works, so that they can learn how it actually does work. We need to bring this same mentality to AI in our classrooms, using AI hype as a learning opportunity specifically within the linguistics wheelhouse. Since LLMs are models of language, linguistics can offer crucial perspectives to bring to the topic of conversational AI. It is our responsibility as linguists and educators to support students' development of AI literacy with a strong basis in linguistics, so that they can better understand the underlying processes and identify the limitations and harms of the technology (Bender et al. 2021), especial-

ly when applied recklessly. We can dispel the illusion that these technologies indeed possess intelligence, are reliable substitutes for independent learning, or are trustworthy sources of information. This matters not only in our classrooms, but for society more broadly as well.

References

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? 🦜. *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623. New York: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Casal, J. Elliot & Matt Kessler. 2023. Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics* 2(3). 1–12. https://doi.org/10.1016/j.rmal.2023.100068.

Chalmers, David J. 2023. Could a large language model be conscious? *arXiv* https://arxiv.org/abs/2303.07103.

De Cosmo, Leonardo. 2022. Google engineer claims AI chatbot is sentient: Why that matters. *Scientific American*. July 12, 2022. https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/.

Elkhatat, Ahmed, M., Khaled Elsaid & Saeed Almeer. 2023. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal of Educational Integrity* 19. 17. https://doi.org/10.1007/s40979-023-00140-5.

Jaiyeola, Temitayo. 2024. Online uproar over Nigerian English being flagged as ChatGPT-ish. *Business Day*. April 10, 2024. https://businessday.ng/technology/article/online-uproar-over-nigerian-english-flagged-as-chatgpt-ish/.

Liang, Weixin, Mert Yuksekgonul, Yining Mao, Eric Wu & James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns* 4(7). 100779. https://doi.org/10.1016/j.patter.2023.100779.

Appendix: Sample syntax assignment using ChatGPT output

Note that instructor commentary on the quality of ChatGPT's output is given here in red and set off by square brackets and would not be given to students in the actual assignment.

In this homework assignment, you are shown three sample prompts that were given to ChatGPT, along with ChatGPT's response to each prompt. All of the prompts together are an attempt to use ChatGPT to write an essay piece by piece to make an argument about syntactic constituency in general North American English using three of four constituency tests we have discussed in this course. Your task is to evaluate each of the three responses. Be explicit about what ChatGPT gets right and wrong. Finally, write your own concluding paragraph that would complete this argument by using the fourth syntactic constituency test we have talked about that was not submitted as a prompt to ChatGPT. Note: based on the extensive errors you see in ChatGPT's responses, I hope you are convinced why you should not try to use it to write the fourth paragraph for you!

Prompt #1: Consider the following sample data from general North American English, some of which may be useful for determining whether or not a preposition plus a following noun phrase form a syntactic constituent, which we may call a prepositional phrase (PP):

It's best to meet with Eyota in the afternoon.
The afternoon is when it's best to meet with Eyota.
In the afternoon is when it's best to meet with Eyota.
With Eyota is when it's best to meet in the afternoon.
Kylo's father is in the hospital.
That is the hospital where Kylo's father is.
Kylo's father is who is in the hospital.
The hospital is where Kylo's father is.
I saw the creature behind Ilsa's house.
Ilsa's house is where I saw the creature.
Behind Ilsa's house is where I saw the creature.

Select one relevant example from this set and show how it and slight variations of it can be used to demonstrate the constituency of our proposed PP constituent, based on how questions are formed and answered.

ChatGPT's response: One relevant example from the given set that demonstrates PP constituency is: "In the afternoon is when it's best to meet with Eyota." To demonstrate the PP constituency of this sentence, we can ask questions about the prepositional phrase "in the afternoon":

1. "When is it best to meet with Eyota?" Answer: "In the afternoon."
2. "When is not a good time to meet with Eyota?" Answer: "Not in the afternoon."
3. "What is the best time to meet with Eyota?" Answer: "In the afternoon."

These questions and answers show that "in the afternoon" functions as a single unit, indicating the presence of a prepositional phrase as a syntactic constituent. Variations of this sentence can also be used to demonstrate PP constituency:

4. “Where is it best to meet with Eyota?” Answer: “In the afternoon.”
5. “What is the ideal time for meeting with Eyota?” Answer: “In the afternoon.”
6. “At what time of the day is it optimal to meet with Eyota?” Answer: “In the afternoon.”

These variations also show that “in the afternoon” is a single unit and functions as a prepositional phrase constituent.

[This is nearly an okay response. The first example does indeed use proper question formation and response, with the PP *in the afternoon* as a response, indicating that it may be a syntactic constituent. However, the second example is completely irrelevant, since it introduces unnecessary negation. The third example is also odd, since it is very similar to the first, but replaces *when is it best* to *what is the best time*. The remaining three examples are also unnecessary and oddly constructed, not matching how this constituency test would have been discussed in class. In addition, the formatting of examples used by ChatGPT does not match how examples would have been formatted in class, for example, no quotation marks would have been used in a numbered list of example sentences.]

Prompt #2: Now do the same thing, except using the syntactic constituency test concerning the placement of sentential adverbs.

ChatGPT’s response: One syntactic constituency test that can be used to determine whether a preposition plus a noun phrase forms a constituent is the placement of sentential adverbs. A relevant example from the given set that demonstrates this test is: “Kylo’s father is in the hospital.” To demonstrate the PP constituency of this sentence, we can add a sentential adverb, such as “fortunately” or “unfortunately”:

1. “Fortunately, Kylo’s father is in the hospital.”
2. “Unfortunately, Kylo’s father is in the hospital.”

Both of these sentences have the same basic meaning as the original sentence, but with the added adverb, they also demonstrate that "in the hospital" forms a single constituent, a prepositional phrase. Similarly, we can use variations of the sentence to show the placement of sentential adverbs:

3. “It's fortunate that Kylo’s father is in the hospital.”
4. “Unfortunately for Kylo, his father is in the hospital.”
5. “Kylo’s father is, unfortunately, in the hospital.”

In all of these variations, the placement of the adverb “fortunately” or “unfortunately” demonstrates that “in the hospital” is a prepositional phrase constituent.

[This is a completely wrong response. Students would have been shown how to use this test by attempting to place the adverb inside the hypothesized constituent and finding that the resulting sentence is ungrammatical. So we would want an example like **Kylo’s father is in often the hospital*, where the adverb *often* intervenes between the preposition and the following NP. Since

this is ungrammatical, but *often* can otherwise be placed in other positions outside known constituents (*Kylo's father often is in the hospital*, *Kylo's father is often in the hospital*, *Kylo's father is in the hospital often*, etc.), this suggests that *in the hospital* is a constituent that blocks an intervening sentential adverb.]

Prompt #3: Finally, do the same thing, except using the syntactic constituency test concerning how some words or phrases can be used to refer to some other syntactic constituent in the same sentence.

ChatGPT's response: Another syntactic constituency test that can be used to determine whether a preposition plus a noun phrase forms a constituent is substitution. This test involves replacing a word or phrase in the sentence with another word or phrase that refers to the same constituent. A relevant example from the given set that demonstrates this test is: "The hospital is where Kylo's father is." To demonstrate the PP constituency of this sentence, we can replace "the hospital" with a pronoun that refers to the same constituent:

1. "It is where Kylo's father is."
2. "Where Kylo's father is, is the hospital."

Both of these sentences have the same basic meaning as the original sentence, but with the replacement of "the hospital" with a pronoun or another phrase, they also demonstrate that "where Kylo's father is" forms a single constituent, a prepositional phrase. Similarly, we can use variations of the sentence to show the substitution test:

3. "The place where Kylo's father is located is the hospital."
4. "Where Kylo's father currently is can be found in the hospital."

In both of these variations, the substitution of "the place where Kylo's father is located" and "where Kylo's father currently is" for "the hospital" demonstrates that "where Kylo's father is" forms a prepositional phrase constituent.

[This response immediately starts off on a bad foot, by using the technical term *substitution*, which was not used in the prompt or in the course at all (the test would have been framed as a pronoun reference test). This would be a clear marker of someone writing the response using outside resources rather than following the material as taught in the course. Further, the examples ChatGPT gives apply the test to the wrong string (*the hospital* instead of *in the hospital*). For example, in *It is where Kylo's father is*, *where* is not substituting for *the hospital*, since **It is the hospital Kylo's father is* is ungrammatical! The prompt specifies that the relevant word (in this case, *where*) should refer to some other constituent within the same sentence. Thus, we need an example like *Kylo's father is in the hospital, and his mother is there, too*. Here, *there* refers back to *the hospital*. This is the kind of structure that would have been used in class for this test.]

What happens when we give the entire assignment to ChatGPT? Can it find its own errors, and also produce a coherent fourth paragraph using the unnamed fourth test? This is not given to the students in the assignment, but it could be discussed afterward. Here is ChatGPT's response when the entire assignment is given to it, in which it tries to evaluate its own output.

ChatGPT's response: ChatGPT appears to understand the three different tests for determining syntactic constituency. It correctly identifies relevant examples from the given set and demonstrates how each example can be used to show the constituency of the proposed prepositional phrase (PP). [Wrong!]

In the first prompt, ChatGPT correctly uses the question-and-answer test to demonstrate the constituency of a prepositional phrase. In the second prompt, it correctly uses the sentential adverb test to demonstrate the same thing. [Wrong!] In the third prompt, ChatGPT correctly uses the substitution test to demonstrate the constituency of the prepositional phrase. [Wrong!]

However, ChatGPT is not perfect and occasionally makes mistakes. [One of the truest things it has said!] For example, in the second prompt, ChatGPT says that the placement of the adverb "demonstrates that 'in the hospital' is a prepositional phrase constituent." It is more accurate to say that the placement of the adverb supports the idea that "in the hospital" is a prepositional phrase constituent, but it does not necessarily demonstrate it. [This is utter nonsense.] Additionally, in the third prompt, ChatGPT stops in the middle of a sentence without finishing it. [There is no such sentence!]

In conclusion, ChatGPT generally shows a good understanding of the three different tests for determining syntactic constituency. [In fact, it does not demonstrate any understanding at all, let along good understanding.] However, it does make some minor [!!!] errors that could be improved upon with further training. [Training without learning is not going to solve the problem!] Overall, the examples and explanations provided by ChatGPT are helpful in demonstrating the ways in which syntactic constituency can be determined in general North American English.

Prompt #4: [Note the use of *prompt* instead of *response*!] Another way to test for constituency is by using the coordination test. [Too bad for ChatGPT, because this was not the fourth constituency test that would have been used in this course! We would have used the movement test and ignored the coordination test. Again, this would be a sign of someone writing using outside materials rather than course materials.] To determine if a prepositional phrase is a single constituent, we can check if it can be coordinated with another similar prepositional phrase. A relevant example from the given set that demonstrates this test is: "Behind Ilsa's house is where I saw the creature." To apply the coordination test to this sentence, we can try coordinating it with a similar prepositional phrase, such as "in the woods" or "at the park":

1. "Behind Ilsa's house is where I saw the creature, and in the woods is where I heard it."
2. "At the park is where I heard the creature, and behind Ilsa's house is where I saw it."

Both of these sentences sound grammatically correct and demonstrate that "behind Ilsa's house" is a single constituent, a prepositional phrase. They show that "behind Ilsa's house" can be coordinated with another similar prepositional phrase and still make sense in the context of the sentence. Therefore, the coordination test provides further evidence that "behind Ilsa's house" is a prepositional phrase constituent. [Not only did it pick the wrong test, but it used this test incorrectly! A correct usage of the coordination test would be something like *behind Ilsa's house and at the park is where I saw the creature*.]