 CC-BY-NC

Against the use of GPTZero and other LLM-output detection tools

These last few months have been a VERY interesting time to be me: a person whose background is in computational linguistics, cognitive science, and psycholinguistics who ended up becoming a writing professor. The advent of ChatGPT last November has meant that suddenly, everyone is talking about what "AI" systems like ChatGPT (which I will refer to from here forward as a large language model, or LLM, rather than as "AI") mean for us as teachers of writing, and I have been spending a lot of time in various forums debunking misguided ideas about it and trying my best to explain things in a way that will be generative and productive instead of riling everyone up. That's my goal in this piece, too.

To be very clear from the outset: I'm neither strongly "for" nor "against" using LLMs in the writing classroom; I think there are use-cases that could be quite pedagogically valuable AND I think there are good reasons to be skeptical about the value of bringing LLM-generated text into the classroom (and there are very legitimate ethical concerns as well). But I also don't think LLMs are going away, so regardless of whether we do or don't use LLMs in our classrooms, we do need to be talking about them with our students, and helping them to develop critical AI literacy. What I also think we desperately need right now are studies that look at how the use of tools like ChatGPT at different stages of the writing process impact developing writers, and I hope to be involved in doing some of that work.

But this piece is not about whether we should be inviting ChatGPT into our classrooms. Instead, I want to address an approach to thinking about LLM-generated writing that I think is absolutely toxic: the promotion of detection tools like GPTZero, which claim to be able to identify whether a text was produced by an LLM or a human, as a way of determining whether students are producing their own writing. On one level, my resistance to these tools is based on the same concerns as I have about TurnItIn and other "plagiarism-detection" tools: while I understand the concerns that drive instructors towards such tools, I think setting up a skeptical, adversarial approach toward one's own students just isn't good pedagogy.

But GPTZero invites and exacerbates an even more insidious attitude toward student writing. I wrote about this in a thread on Mastodon, which went...well, "viral" isn't the right word, but it certainly got a lot more attention than anything else I've ever posted there. So I want to unpack these ideas a bit more here, where I don't have to worry about character limits and getting the threading right (which I failed to do on Mastodon; I wrote more posts than show up in the thread I'm linking to. Here's another. And another.)

Basically, what motivates me is that I have seen way too many people uncritically accepting the hype in BOTH directions — about the capabilities of LLMs and about the capabilities of automated detectors. I started experimenting with GPTZero because I kept hearing people claim that it was going to "solve" the concerns that keep cropping up about how teachers will know whether their students have written their own writing or used ChatGPT or some other LLM to produce it. The primary response I saw, when the student who developed GPTZero announced it to the public, was relief, in the sense of "oh, thank goodness, now we have a tool we can use when we suspect

that a student is turning in ChatGPT-generated writing," which of course takes for granted that this tool is in fact a foolproof way of determining the "truth" regarding a particular text's provenance; in other words, the relief is premised on accepting the hype about the GPTZero's detection abilities.

But alongside that relief, I'm also noticing a trend toward people saying that we don't actually need these tools, because it is "obvious" when someone has submitted work that was generated by an LLM. The most common framing I see is "I'll know that my student used AI to generate their paper if the paper has proper spelling and grammar (because my students don't produce writing with proper spelling and grammar)". And this makes me want to scream. It's so typical of the "damned if you do, damned if you don't" bind that we so often put students in, especially those students who are non-native speakers or speakers of marginalized dialects of English: we insist that students produce a particular variety of standardized English in their writing, penalize them if they don't, but then also treat them with skepticism when they DO...and this is precisely the mindset that tools like GPTZero reinforce, I think.

I don't have any doubts at all about *whose* writing is most likely to be viewed with suspicion and thus sent through a tool like GPTZero, because it's precisely the same sets of students we already subject to this kind of skepticism and policing. My biggest concern is that schools will listen to the hype and decide to use automated detectors like GPTZero and put their students through "reverse Turing Tests", and I know that the students that will be hit hardest are the ones we already police the most: the ones who we think "shouldn't" be able to produce clear, clean prose of the sort that LLMs generate. The non-native speakers. The speakers of marginalized dialects. So I've been pushing against every suggestion that we should adopt these tools in academic settings.

And to be clear, I do understand why people would want such a tool like GPTZero to exist — I understand why people would want to be assured that what they are reading are the words of a person, and not the output of an LLM. (See the recent issues at Clarkesworld; basically, there has been an absolute deluge of what seem to be rapidly-produced LLM-generated submissions, and the speed at which such submissions can be generated is drowning out legitimate submissions, and this is creating a need to figure out which submissions are legitimate so that they aren't stuck sifting through impossibly huge piles of junk submissions.) I, too, dread the future in which the internet is swamped in SEO-optimized LLM-generated bullshit, and can all too easily see how quickly that will make much of what we do online impossible. And as a writing teacher, I'm obviously concerned that if students use text-generation systems to produce their papers, they will be missing out on the process that helps them to build their skills as writers, and missing out on all of the ways that writing can be a tool for deeper thinking.

But: Everything I know about how LLMs work and how humans produce language (and I know a non-trivial amount about both, given the background I have) tells me that you will never, ever be able to build something that can *reliably* distinguish between the two based on a sample of text. And sure enough, in my experiments, GPTZero fails miserably. So many false positives, including significant chunks of *my* writing. (I promise that literally none of the text in this post was generated by ChatGPT or any other LLM. It's all me. But how can I prove that?)

Why do I say that what GPTZero claims to be able to do is not possible? Here, I want to dive a little deeper into both how LLMs like ChatGPT work and how humans process language. Let's start with the LLMs. I'm going to try to keep things relatively simple here — getting too deep into the math isn't actually necessary for the point I'm making, and frankly, I'm not confident I can explain the math clearly anyway. In essence, these systems are an elaborate kind of "autocomplete": they are very, very good at playing the game of "guess the next word" and as a result, can generate text that sounds very fluent. But thinking of it as "autocomplete" makes it a little too easy to dismiss some of the complexity going on under the hood that makes it so that the output of LLMs cannot be reliably distinguished from human-written text. What informs the "guess" that the LLM is making about the next word is something much, much richer than simple transition probabilities; it's not a simple probability estimate based on how frequently word X occurs after word Y. To elaborate on what I mean here, I'm going to try to explain some of what's going on "under the hood", with the goal of keeping things accessible for writing instructors who don't share my background.

The LLM that underlies ChatGPT is GPT-3; "GPT" stands for "Generative Pre-Trained Transformer". A transformer model is an innovation on a more conventional neural network that adds in "attention layers", allowing the model to selectively "pay attention" to different positions of an input string. So, for example, if the model is "learning" from an example sentence, different layers of the neural network can "attend" to different words in the sentence, and can learn about the relationships between different parts of the sentence, not just those parts that are immediately adjacent. Models like GPT-3 have an enormous number of "parameters", which allow the model to "attend" to many, many aspects of the "context" in which a particular word appears. This architecture gives transformer models a way of flexibly representing a LOT of information about the context in which different words appear, and that flexibility is a big part of what gives transformer models the power they have to generate such fluent-sounding text.

These models are "pre-trained" (that's the "P" in "GPT") through a process called "masking", in which a word is removed from a sequence of text; it's a game of "fill in the blank" played billions of times over. The model has to use whatever statistics it has developed based on the examples it has already processed to predict what that masked word should be, and by doing this over and over and adjusting the parameters of the model in order to make the "right" answer be the one more likely to be produced in similar contexts the next time, it eventually "learns" which features to pay attention to in the context, and this information gets represented in its model.

One type of data representation you get from this process is something called a "word embedding." You can think of each word as being represented as a very, very long list (in math terms, it's a "vector") of number values that represent aspects of the context in which that word is typically found. One of the things that makes GPT-3 (the transformer that underlies ChatGPT) so powerful is that these vectors are *incredibly* long. But since they're vectors, and we know how to do math with vectors, we can do math with word embeddings. And this math yields some interesting stuff that we might describe as something like semantics. In essence, the vectors represent coordinates in a massively multidimensional space (not the kind of space you can imagine, because alas, most of us humans can visualize and make sense of 3-4 dimensions, tops,

and here, we're talking about several orders of magnitude more dimensions than that). And so, for any pair of words, you can compare their "embedding" vectors to generate a measurement of how "similar" they are: if their coordinates in this massively multidimensional space are close to each other, they're more similar.

These "similarity" measures capture a surprising amount of meaning, which we can see when we start trying to add and subtract vectors from each other. For example, we can generate analogies by subtracting one word from another, then adding another word, as in **bigger - big + cold = colder** or **CU - copper + gold = AU** or **sushi - Japan + Germany = bratwurst**. That last example perhaps already has you thinking: are these embeddings just capturing biased stereotypes? And the answer to that is yes, but only because those stereotypes are "embedded" in the language that the LLM is modeling. (LLMs are an unflattering mirror! For more discussion of this, see the section about LLMs in Brian Christian's book, "The Alignment Problem".)

So, for example, we find sexist biases in word embeddings, such that **doctor - male = nurse**. Tolga Bolukbasi and colleagues published an article back in 2016 looking at how we might go about debiasing the word embedding representations in LLMs, and current models like ChatGPT do attempt to reduce bias both through methods like the one presented in that 2016 paper and through Reinforcement Learning from Human Feedback, but there are limits to how effective this can be given how thoroughly these biases are "baked in" to our language. In an attempt to address the ethical concerns about offensive, violent, abusive, and downright toxic content that a model like ChatGPT could produce, developer OpenAI has also employed Kenyan workers making less than $2 an hour to rate toxic output in order to train a detection algorithm used by the model to ensure that its output is not toxic. (These "safety measures" are surprisingly easy to fool, however, with clever prompt design.)

The ethical issues I just raised are worth exploring with students, but in any case: I hope I've made it clear that when we're thinking about what an LLM "knows" (I put that word in scare quotes, because I don't think we mean the same thing when we talk about a model "knowing" something as we do when we say a person knows something), it's not JUST what we might think of as simple transition probabilities, knowing that *this* word follows *that* word X% of the time; the context is much, much richer than that. It is an open area of debate to what extent the contextual representations generated by the model map onto anything even remotely like our own internal semantic knowledge. I share a lot of Emily Bender and Alex Koller's skepticism about what LLMs could possibly "know" based simply on linguistic input that is not grounded, not linked to physical reality in some way. And yet as a cognitive scientist, I'm also quite compelled by the case made by my friend Steven Piantadosi (along with Felix Hill) for why there are good reasons to think that LLMs could be developing representations of "meaning" that aren't terribly dissimilar from our own human representations of meaning.

But in the end, regardless of exactly how we would describe what is happening under the hood in terms of what "knowledge" about semantics an LLM can be said to have and what is actually being represented about the context, the LLM's task is the simple one we started with: choose the word that is most likely to occur next, given the context.

Onto how humans process and produce language. Spoiler alert: it's not as different as you might think. Here, I'm also going to be simplifying significantly, obviously. But while predicting the next word is not ALL of what our minds are doing when we process and produce language, it's not an insignificant piece of the pie. We know from a great deal of research, some of which I discuss in my recent paper in the Journal of Teaching Writing (also see the discussion in the Piantadosi & Hill piece linked earlier), that human language processing is incremental. We are not waiting until we've finished listening to or reading a sentence to make sense of what that sentence is saying, but instead, we are using what we've already heard or read, along with all sorts of other knowledge (about the broader linguistic context, about the world, about the likely intentions of the person communicating with us, and so on) to predict what is going to come next, and when what comes next aligns with that prediction, we are able to process it more efficiently. And the knowledge that is informing those predictions is inherently probabilistic and associative in nature; it is learned from our experience as users of the language and as experiencers of the world.

When we produce language, we also do not necessarily wait until we've got the whole sentence planned to open our mouths (or to begin typing). A big piece of what helps us to continue speaking or typing from that point onward is our internal sense of what words naturally follow from what we've produced so far, based on our linguistic and world knowledge — that is, the prior linguistic context and other aspects of the context in which we are speaking or writing are narrowing the space of possible continuations (and this is helpful, because the space of possible words is so enormous that if we did not constrain it somehow, I'm not sure how we'd ever finish a sentence!). It's not that we can't ever produce something unexpected, but rather, that our own cognitive architecture is fundamentally aimed at developing expectations and predictions that narrow the space of possibilities to ease processing.

So, in the end, our own human linguistic knowledge is in no small part probabilistic and associative in nature, and we are good little prediction machines in terms of how we process and produce language, able to use both immediate context and more distant contextual knowledge to inform those predictions. And we have an LLM that is designed to be very good at predicting the next word from context, and current LLMs are able to represent (via the "embeddings" I talked about) a great deal of information about context, including both local and more distant context. Is it the *same* as the mental representations humans rely upon when we are producing language in context? Almost certainly not! Humans have intentions, we have goals, we have personal experiences, we have desires, we have senses, we are interacting in and can act upon the physical world...and while some of these things might be representable in *some* fashion by an LLM like ChatGPT as it does its vector math on the prompts we give it, it's not going to be exactly the same.

But: I do think that it is fundamentally *close enough* that if we are trying to decide, *purely on the basis of a piece of text*, whether that text was generated by a person or by an LLM, we are highly likely to fall victim to both false positives and false negatives. I am far more concerned about false positives than I am about false negatives; I think that the potential harm of falsely identifying a person's writing as having been LLM-generated far outweighs the potential harm of falsely identifying LLM-generated writing as being human-written. If you're a student whose professor is accusing you of academic dishonesty, the consequences of having your writing inaccurately flagged as LLM-generated could be quite devastating if your professor or institution is one who

believes the hype around the infallibility of GPTZero. What is your recourse? How DO you actually prove that you are the writer?

On that note, what does GPTZero make of MY writing? It actually flags a decent amount of it as "more likely to have been generated by AI". (Here's an example from my JTW paper). I mostly find this fascinating. I've always been told that I write very well, and I'm curious about the extent to which that means that I'm very good at producing writing that is easy for readers to understand, and to what extent THAT ability is based on producing writing that is, on some level, "predictable" in the very sense that LLMs are designed to produce. And yet, I don't think I sound like a robot — I sound like "me"! Or at least, I sound like one version of me: the version who is attempting to clearly explain something to another person who doesn't share my knowledge, and working to do that in a way that's engaging and friendly. That's not the only "voice" I have, but it's my primary writing voice! In any case, I don't think I lack "voice" in my writing, which is the main thing that people point to when they say they can tell which writing was produced by an LLM and which writing was produced by a human.

But when I said I "mostly" found GPTZero's flags of my own writing to be fascinating, well...the other reaction I have is to start wondering whether these same people who confidently declare that they can tell whether a piece of text was written by a human or an LLM would declare my own writing to be generated by an LLM, and what on earth my recourse would be to prove them wrong? Is my "voice" too similar to ChatGPT's "voice"? Does it even make sense to say ChatGPT has "a voice"? And if ChatGPT *doesn't* have a "voice", and I *do* have a "voice", but our texts are as indistinguishable as GPTZero says that some of them are, what does that tell us about "voice"?

Interestingly, in many of the examples of my own writing that I put through GPTZero, the parts that were flagged as "likely AI" were the *ends* of things — the end of an introduction section of paper, where I laid out a roadmap for the remainder of the article; the end of a short essay I'd posted to Facebook, where I summed up what I'd argued in clear, simple language. I have some suspicions about why that might be. Anecdotally, I've also heard several neurodivergent folks say their writing gets flagged, which is something I'm very curious about. And my international students tell me that they are very concerned that the well-formed sentences they produce will be more likely to be viewed as potentially AI-generated, not just due to the aforementioned inherent skepticism that many people have towards fluent writing when produced by those we expect not to be fluent, but also because they worry they are following "scripts" that they learned in their English classes, and they're afraid that their more limited vocabulary will mean that their writing will sound as vague and generic as examples we've seen from ChatGPT.

If we think about the aspects of writing that are "predictable" in the sense that a detection algorithm like GPTZero is looking for, it's going to be the more "conventionalized" stuff: things like the "roadmapping" at the ends of introductions, and "signal phrases" at other parts of arguments, and transition phrases, and so forth. What ChatGPT aims to produce is fluent writing, and in so doing, it produces many of the explicit coherence strategies that we teach to students, like "Given-before-New", End Focus, various kinds of transitions and metadiscourse, and so on. I would not be at all surprised if those same strategies are more relied upon by the students who don't

inherently trust their internal sense of what "sounds right"; in other words, the neurodivergent, the non-native speakers, the speakers of non-standard dialects.

I also think that in our attempts to emphasize what makes human writing special and to point out the limitations of LLMs, we run the risk of dismissing the role that conventions and predictability play in making writing work well, especially certain types of writing. In a very real sense, language itself only works because of conventions: if we don't all agree that a particular way of putting words together produces a particular meaning, then it won't actually work to convey that meaning. Of course, if the language in a piece of writing is entirely predictable, it's going to be boring, but there's a real balance that every writer has to strike in order to be understood. I suspect that part of why some of my own writing gets flagged as "likely AI-generated" is that I am often writing from a position of explaining complicated topics to non-expert audiences, which I'm told I do very effectively. But this means that when I'm writing, I'm thinking about how I can make things easy for readers to process (which means, in part, that it will be more "predictable"), and it means that I place a high value on not using language that will be inaccessible to readers, which almost certainly means that I'm producing writing that, while it conveys information that required a great deal of careful thought on my part to process and organize, is not, based purely on surface-level statistics, all that different from the vague, generic-sounding bullshit that ChatGPT will helpfully supply.

Is the solution to just be more creative and innovative and unconventional in our own writing? Isn't that what we want for our students? We may say that, and we may even mean it, but look: I'm a rather linguistically playful person (as my friends will attest) who truly delights in offbeat writing, and yet in many of the genres in which I write, that's not exactly viewed positively. We need to remember that we are asking students to do a job not entirely dissimilar to what ChatGPT is doing when we tell it to do something like "write a rhetorical analysis of Martin Luther King Jr's "I Have a Dream" speech in an academic style." Students, too, have to try to figure out from looking at examples of the genres we are asking them to produce what a rhetorical analysis "sounds like". Students, who are not yet fully members of the discourse communities we're asking them to write as if they are members of, understandably start by trying to emulate the discourse patterns they're seeing in the genre-specific examples from which they're learning. They're doing this with their full human selves, of course; they're doing it with intentions and goals and desires and senses that an LLM does not possess, at least not in the way we humans do, but the task itself is not fundamentally different. Is it any surprise that their writing may end up sounding pretty similar to what an LLM generates?

We are so accustomed to using writing as a way of assessing someone's understanding, and it's disorienting to realize that text with fairly similar surface-level features could be produced by someone with deep understanding of a topic who is working to distill the information into a text that can be easily understood by a reader, or by a student who is grappling with concepts they only partially understand and attempting to produce the kind of writing they're being asked to produce, or by an LLM that is simply predicting the next word given the context…but it's true. To be clear, I am not saying that these kinds of writing are *never* distinguishable from each other! I think there are many examples that are in fact "obviously" the product of an expert writer…but

given everything that I've written here about how LLMs work, I would be very, very careful about assuming there are any examples that are "obviously" the product of an LLM rather than a human.

This complicates so many things about how we think about the work we're doing as educators, but the answer is not to pretend that we, or tools like GPTZero, can reliably tell that a piece of writing was produced by an LLM rather than a person. The risk of false-positives is too high, and will disproportionately impact precisely the same students who already face the most skepticism about their writing skills. Instead, I hope that we will work with our students to develop critical AI literacy. On that front, I will offer a few suggestions:

1. I highly encourage writing teachers to test some of their own writing using the GPTZero app. Perhaps you will find that absolutely nothing in your writing is flagged; perhaps I truly am an anomaly who naturally produces writing that is very similar to the writing produced by LLMs. But I'm guessing that you WILL see some of your own writing flagged as "likely" AI-generated, and I think it's really important for you to see that and reflect on what that tells you about the reliability of these programs. The fact that there are ANY false positives means that this is a tool that should not ever be applied to student writing, except, perhaps, as an educational experience aimed at developing their critical AI literacy. Which leads me to...

2. If you DO decide to introduce your students to GPTZero, think very carefully about how you frame it. In my own classes, I have shown my students that GPTZero flags aspects of my own writing as AI-generated, and we've discussed the features of that writing that might be leading to those flags. We've also discussed many of the concerns I've already shared in this piece. In my class on professional writing, we've also experimented with testing more "formulaic" writing, like the kind often produced in "template"-driven cover letters, to see which aspects of the template get flagged, and we've played around with comparing drafts of letters that are closer to the template to ones that diverge more significantly and include more personal details. I'm very careful not to say that anything flagged by GPTZero is inherently bad writing, because I don't think that's true (and not *just* because it flags some of my own writing!); some amount of predictable text is almost certainly going to be present in a genre as constrained as a cover letter. But it has been an eye-opening way for students to see the effects of targeting their cover letters.

3. Finally, I offer here the note that I shared with my students at the beginning of the Spring '23 semester about LLMs like ChatGPT. In short, what I am trying to encourage in this note is for students to think critically about what these models can and cannot do, to think critically about what the media discussions of these tools are actually saying about students and the work they're doing, and to think critically about why they are in a college writing class and how learning actually happens. This is just a starting point, but my students have told me that they really appreciated that I shared this with them at the start of the semester, and it has prompted very productive discussions.

But: I will close by noting, with delightful irony, that when I pasted that note into GPTZero, it told me: "Your text is likely to be written entirely by AI." It wasn't — that text is all me! But how would I

prove it to you? And this, dear reader, is why we should never, ever use these AI-detection tools to police our students' writing.