SP Publications
International Journal Of English and Studies (IJOES)
An International Peer-Reviewed and Refereed Journal; **Impact Factor:** 8.175 (SJIF)
ISSN: 2581-8333|**Volume 8, Issue-1(January);2026**

Reliability Risks of Generative AI in Corpus-Based Literature Teaching

Dr. T. Jude Livingston
Assistant Professor of English, Nazareth Margoschis College at Pillaiyanmanai
(Affiliated to Manonmaniam Sundaranar University, Tirunelveli)

Article Received: 25/12/2025
Article Accepted: 29/01/2026
Published Online: 31/01/2026
DOI:10.47311/IJOES.2025.8.01.607

Abstract

Corpus methods are entering the literature classroom, and generative AI is making them easier to use. A teacher can now ask an AI to find patterns in a text, count features, and group examples in seconds. This convenience carries a hidden danger. Large language models produce fluent output that is sometimes fabricated, and recent studies show they misclassify and miscount textual features that dedicated corpus tools handle reliably. This paper examines that danger in the specific context of corpus-based literature teaching. It reports a small demonstration that contrasts a deterministic concordance count, which returns the same result every time, with the documented behaviour of AI systems on the same kind of task. The paper argues that teachers and learners need a new competence, namely the ability to tell genuine distributional evidence from AI-generated approximations of it. Without this competence, corpus-based literature teaching risks replacing one weak foundation, impressionistic reading, with another, confident but unreliable machine output.

Keywords: generative AI, corpus stylistics, hallucination, literature pedagogy, AI literacy, EFL

SP Publications
International Journal Of English and Studies (IJOES)
An International Peer-Reviewed and Refereed Journal; **Impact Factor:** 8.175 (SJIF)
ISSN: 2581-8333|**Volume 8, Issue-1(January);2026**

Introduction

Corpus methods have begun to enter the literature classroom. Teachers use concordance tools to show learners how a word behaves across a text, how images recur, and how an author's style can be described through evidence rather than impression. The arrival of generative AI has accelerated this trend. A teacher no longer needs to learn specialised software. They can ask an AI, in plain language, to find the recurring images in a story or count how often a word appears. The barrier that kept corpus methods out of most classrooms appears to be falling.

This convenience has a cost that is not yet widely recognised. Generative AI does not count. It predicts. A large language model generates the most probable response to a prompt, and that response is not always accurate. When asked to count or classify the features of a text, an AI may return numbers and categories that look authoritative but are wrong. The output is fluent, confident, and sometimes fabricated. In a classroom built on corpus evidence, this is a serious problem, because the entire value of corpus work lies in the reliability of its counts.

Corpus Methods in the Literature Classroom

Corpus stylistics applies the methods of corpus linguistics to literary texts. Instead of reading a text only for meaning, the analyst examines its measurable features, including word frequencies, collocations, and the distribution of features across a whole text. Stubbs (2005) showed that the politically charged vocabulary of Conrad's Heart of Darkness becomes visible through keyword analysis in a way that no single reading reveals. Mahlberg (2013) demonstrated that recurring word clusters in Dickens function as characterisation devices across entire novels. These methods give readers access to a level of textual organisation that lies beyond the reach of ordinary reading.

In the language classroom, these methods have a clear appeal. They allow learners to ground their interpretations in evidence rather than impression. A learner who claims that a story dwells on a particular image can be asked to show how often the image appears and where it concentrates. The claim becomes checkable. For learners of English as a foreign language, who may lack the deep reading background that makes such patterns intuitively noticeable, corpus tools offer a way to perceive structure directly. The evidence is on the screen, not in the memory.

Until recently, the main obstacle was practical. Corpus tools required learning, and building a corpus took time. Generative AI appears to remove this obstacle. A teacher can describe what they want in ordinary language and receive an answer at once. The promise is real. So is the risk that comes with it, and that risk is the subject of this paper.

How Generative AI Differs from a Corpus Tool

A corpus tool and a generative AI appear to do similar work. Both can tell you how often a word appears in a text. But they reach their answers in fundamentally different ways, and the difference matters.

A corpus tool such as AntConc (Anthony, 2024) searches a text directly. When asked how often a word appears, it scans the text, finds every instance, and reports the exact number. The process is mechanical. It produces the same answer every time. Anyone who runs the same search on the same text gets the same result. This property, called reproducibility, is the foundation of corpus method. It is what makes corpus evidence more trustworthy than impressionistic reading.

A generative AI works differently. It does not search the text. It predicts a likely response based on patterns in its training data. Kalai et al. (2025) show that language models hallucinate because their training rewards confident guessing over admitting uncertainty. The model is built to produce a plausible answer, not necessarily a correct one. When the task is counting, the model may approximate rather than count. When the task is classification, it may assign a category that seems likely rather than one it has verified. The output can vary from one run to the next, even with the same prompt.

This difference is not a flaw that better models will simply remove. Kalai et al. (2025) argue that hallucination is a structural feature of how these systems are trained and evaluated. As long as models are rewarded for fluent answers, they will sometimes produce fluent errors. For most everyday uses, this is tolerable. For corpus work, where the count is the evidence, it is not.

It is worth being precise about why prediction and counting are different operations. To count, a system must inspect every relevant item in the text and tally it. The process is exhaustive and mechanical. To predict, a system estimates the most likely answer given the patterns it has learned. A model that has seen many texts may estimate that a common word like you appears, say, about twenty times in a passage of this length, and it may report a number close to the truth without ever having counted. Sometimes the estimate will be exact. Sometimes it will be close. Sometimes it will be wrong. The crucial point is that the model cannot tell the user which of these has occurred, because it did not count in the first place. The number it reports carries no record of how it was reached.

Evidence That AI Miscounts and Misclassifies Text

The claim that AI is unreliable for corpus tasks is not speculation. It has been measured. Several recent studies have tested generative AI on exactly the kind of work a teacher might delegate to it.

Altameemi (2026) compared ChatGPT-4 with a purpose-built language model on the task of classifying concordance lines by theme. ChatGPT reached reasonable precision of around eighty to ninety percent on general themes, but it struggled with context-sensitive cases and produced notable misclassifications. The study traced these errors to the model's reliance on probabilistic patterns rather than structured analysis. An accuracy of eighty to ninety percent may sound high, but in corpus work it means that one in every five to ten classifications is wrong, with no indication of which ones.

Curry et al. (2024) conducted a critical evaluation of ChatGPT for corpus approaches to discourse analysis. They found that the model could not reliably analyse each concordance line independently, and that its results could be unreliable in ways difficult to predict or correct. Their conclusion was cautionary. AI can assist corpus work, but its output must be verified, because it cannot be trusted on its own.

The broader literature on AI in education reaches the same conclusion from a different direction. Studies of AI hallucination in educational settings warn that fluent, confident output can mislead learners who lack the means to check it (Shoufan, 2026). The danger is greatest precisely when the output looks most authoritative. A fabricated frequency count, presented in a clean table, carries an appearance of objectivity that an impressionistic reading does not. The learner has more reason to trust it and less means to question it.

A further finding from this literature deserves attention. The same prompt, given to the same model more than once, can produce different answers. Researchers studying hallucination routinely report results across multiple runs, or seeds, precisely because a single run is not representative of what the model will do. For a teacher, this means that asking an AI to count a feature on Monday and again on Friday may yield two different numbers, with no signal that anything has changed. A corpus tool, by contrast, returns the same number on Monday, on Friday, and on any other day, because it performs the same mechanical search each time. Consistency is not a minor convenience here. It is the property that allows one analysis to be compared with another, and one learner's work to be checked against another's.

A Demonstration of Reproducible Counting Versus Probabilistic Output

To make the contrast concrete, consider a simple task. Take a passage of literary text and count how often a particular word appears. This is the most basic corpus operation, and it is the kind of task a teacher might now hand to an AI.

For this demonstration, the opening of Jane Austen's Pride and Prejudice was used, a public-domain passage of 523 word tokens. A short, transparent counting procedure was written and run on the passage five times. The procedure searches the text directly, in the

manner of a corpus tool. The results were identical on every run. The word ‘you’ appeared 18 times. The word ‘must’ appeared 6 times. Table 1 records the five runs.

Table 1

Five Identical Runs of a Reproducible Concordance Count

Run	Count of “you”	Count of “must”
1	18	6
2	18	6
3	18	6
4	18	6
5	18	6

The value of this result is not the numbers themselves. It is their stability. The count did not change across runs, and it can be checked by hand. The six occurrences of must, for example, can be located and verified in the passage by any reader. The six lines containing must, shown with their immediate context, run as follows. A single man of good fortune must be in want of a wife. My dear, you must know Mrs Long says. You must know that I am thinking. Therefore you must visit him as soon as he comes. But my dear, you must indeed go and see Mr Bingley. And indeed you must go, for it will be impossible. Each line can be found in the source passage. The count is transparent, reproducible, and falsifiable. These are the properties that make corpus evidence worth having.

A generative AI offers no such guarantee. This paper does not present fabricated AI output in place of a real test, since to do so would reproduce the very problem under discussion. Instead it relies on the published measurements described above. Altameemi (2026) and Curry et al. (2024) both found that AI classification of textual features varies and errs in ways that a direct search does not. The contrast is therefore not between a perfect tool and a flawed one. It is between a method whose errors are visible and checkable and a method whose errors are hidden inside fluent, confident prose.

The pedagogical danger lies in this hiddenness. When a corpus tool makes an error, for example because a search term was mistyped, the error is usually visible in the output and can be corrected. When an AI miscounts, the wrong number arrives in the same confident form as a right one. A learner has no way to tell them apart without checking against the text directly, which is the very skill the AI was meant to replace.

Implications for the Literature Classroom

These reliability risks do not mean that AI has no place in corpus-based literature teaching. They mean that its place must be defined carefully, and that a new competence must be taught alongside it.

The first implication concerns the division of labour between AI and dedicated tools. AI is well suited to the supporting tasks of corpus work. It can help assemble a corpus, suggest features worth investigating, draft descriptions, and propose interpretations. These are tasks where approximation is acceptable and where a human checks the result. AI is not suited to producing the counts that serve as evidence. That work should be done by a transparent tool whose results can be reproduced and verified. The principle is simple. Use AI for the labour, not for the evidence.

The second implication concerns assessment. If learners are taught to use AI for corpus work without being taught to verify its output, they will learn to trust confident numbers uncritically. This is the opposite of what corpus-based teaching is meant to develop. The goal of corpus pedagogy is to ground interpretation in checkable evidence. An AI that supplies unchecked evidence undermines that goal while appearing to serve it.

The third implication is the most important. Learners now need a specific kind of AI literacy, one focused on evidence. They need to understand that an AI does not count but predicts, that its output can vary and err, and that any quantitative claim it makes about a text must be checked against the text itself. This is not general digital literacy. It is a precise competence, the ability to distinguish genuine distributional evidence from a plausible imitation of it. Research on AI in education already calls for explicit instruction in verification and for accurate mental models of how these systems behave (Shoufan, 2026). The corpus-based literature classroom is one place where that instruction can be made concrete.

There is a productive irony here. The skill learners need in order to use AI safely for corpus work is the same skill corpus work was always meant to develop. Both require the learner to return to the text and check the claim. An AI that miscounts a word, and a critic who asserts a pattern that is not there, fail in the same way. The remedy in both cases is verification against the evidence. Pattern awareness, the capacity to attend to and check the distributional features of a text, is therefore not made obsolete by AI. It is made more necessary.

This suggests a concrete classroom practice. Rather than banning AI from corpus work or accepting its output uncritically, teachers can build verification into the task itself. A learner who uses an AI to find how often a word appears in a story can be required to confirm the number using a transparent tool, or by searching the text directly. The

discrepancies that emerge become teaching moments. When the AI's count and the verified count agree, the learner sees that the tool can help. When they disagree, the learner learns, in the most direct way possible, why the evidence must always be checked. The exercise turns the reliability problem into a lesson about the nature of evidence, which is the deeper aim of corpus-based teaching in any case.

Limitations

This paper has limitations that should be stated plainly. The demonstration is small. It uses a single short passage and a single counting task, chosen to illustrate the property of reproducibility rather than to measure AI error directly. The figures on AI performance come from published studies rather than from a controlled experiment conducted here. A fuller study would run a range of corpus tasks through several AI systems, repeat each task many times, and measure the variation and error against a reproducible baseline. That study remains to be done, and it would strengthen the argument considerably.

A second limitation concerns the pace of change. AI systems are improving, and some of the specific error rates cited here may fall over time. But the underlying argument does not depend on any particular error rate. As long as AI predicts rather than counts, its output will lack the reproducibility that corpus evidence requires. The competence this paper calls for will remain necessary even as the systems improve.

Conclusion

Generative AI is making corpus methods more accessible to the literature classroom, and that is a genuine benefit. But accessibility is not the same as reliability. An AI that produces a frequency count is not performing the same operation as a corpus tool, even when the output looks identical. The tool counts. The AI predicts. The difference is invisible on the surface and decisive underneath.

The demonstration in this paper makes the point in miniature. A transparent count returns the same result every time and can be checked by hand. AI output, as the published evidence shows, varies and errs in ways that are hidden inside fluent prose. For a pedagogy that rests on the trustworthiness of its evidence, this is a difference that cannot be ignored.

The responsible path is not to refuse AI but to teach learners to use it with their eyes open. Use AI for the labour of corpus work and a transparent tool for its evidence. Teach learners to check every quantitative claim against the text. The skill this requires is the same one good reading has always required, the willingness to return to the evidence and see what is actually there. In the age of confident machines, that willingness matters more, not less.

References

Altameemi, Y. (2026). Generative AI in corpus linguistics: Comparing ChatGPT-4 and manually developed NLP models for thematic concordance analysis. *Cogent Arts & Humanities, 13*(1), Article 2592697.

Anthony, L. (2024). *AntConc* (Version 4.3.1) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/antconc/

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics, 4*(1), Article 100082.

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. *arXiv*. https://arxiv.org/abs/2509.04664

Mahlberg, M. (2013). *Corpus stylistics and Dickens's fiction*. Routledge.

Shoufan, A. (2026). AI hallucination from students' perspective: A thematic analysis. *arXiv*. https://arxiv.org/abs/2602.17671

Stubbs, M. (2005). Conrad in the computer: Examples of quantitative stylistic methods. *Language and Literature, 14*(1), 5-24.