Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025
ASEE
Paper ID #46182

Can LLMs Assist with Education Research? The Case of Computer Science Standards Analysis

Dr. Julie M. Smith

Dr. Julie M. Smith is a senior education researcher at the Institute for Advancing Computing Education. She holds degrees in Software Development, Curriculum & Instruction, and Learning Technologies. Her research focus is computer science education, particularly the intersection of learning analytics, learning theory, and equity and excellence. She was a research assistant at MIT's Teaching Systems Lab, working on a program aimed at improving equity in high school computer science programs; she is also co-editor of the SIGCSE Bulletin.

Jacob Koressel
Sofia De Jesus, Carnegie Mellon University
Joseph W Kmoch
Bryan Twarek

Can LLMs Assist with Education Research? The Case of Computer Science Standards Analysis

Abstract

Introduction: Recent AI advances, particularly the introduction of large language models (LLMs), have expanded the capacity to automate various tasks, including the analysis of text. This capability may be especially helpful in education research, where lack of resources often hampers the ability to perform various kinds of analyses, particularly those requiring a high level of expertise in a domain and/or a large set of textual data. For instance, we recently coded approximately 10,000 state K-12 computer science standards, requiring over 200 hours of work by subject matter experts. If LLMs are capable of completing a task such as this, the savings in human resources would be immense.

Research Questions: This study explores two research questions: (1) How do LLMs compare to humans in the performance of an education research task? and (2) What do errors in LLM performance on this task suggest about current LLM capabilities and limitations?

Methodology: We used a random sample of state K-12 computer science standards. We compared the output of three LLMs – ChatGPT, Llama, and Claude – to the work of human subject matter experts in coding the relationship between each state standard and a set of national K-12 standards. Specifically, the LLMs and the humans determined whether each state standard was identical to, similar to, based on, or different from the national standards and (if it was not different) which national standard it resembled.

Results: Each of the LLMs identified a different national standard than the subject matter expert in about half of instances. When the LLM identified the same standard, it usually categorized the type of relationship (i.e., identical to, similar to, based on) in the same way as the human expert. However, the LLMs sometimes misidentified ‘identical’ standards.

Discussion: Our results suggest that LLMs are not currently capable of matching human performance on the task of classifying learning standards. The mis-identification of some state standards as identical to national standards – when they clearly were not – is an interesting error, given that traditional computing technologies can easily identify identical text. Similarly, some of the mismatches between the LLM and human performance indicate clear errors on the part of the LLMs. However, some of the mismatches are difficult to assess, given the ambiguity inherent in this task and the potential for human error. We conclude the paper with recommendations for the use of LLMs in education research based on these findings.

1 Introduction and Background

1.1 Introduction

The AI revolution inaugurated by the release of ChatGPT in November 2022 has impacted many fields due to the ability of large language models (LLMs) to generate fluent text in response to user prompts. This ability has promise to automate many tasks, enabling, among other things, analysis of unstructured text at scale. This analysis could be particularly useful in education research, where large data sets require high levels of human expertise to assess and resources to perform such tasks are often lacking.

For example, our team recently manually coded numerous data points for each of nearly 10,000 state K-12 computer science learning standards. This task required over 200 hours of work by subject matter experts. The potential to automate this task via LLM would represent an enormous reduction in the resources needed for the task, enabling additional education research work to be performed.

Thus, the central aim of this study is to explore the usefulness of LLMs for the analysis of textual data in a specific domain, namely computer science learning standards. Specifically, we answer two research questions: (1) How do LLMs compare to humans in the performance of an education research task? and (2) What do errors in LLM performance on this task suggest about current LLM capabilities and limitations?

1.2 Large Language Models

Despite their novelty, LLMs have been used for a wide variety of tasks in domains ranging from education [1] to medical diagnoses [2]. LLMs have shown promise in tasks such as identifying relevant material from large corpora [3], including at the conceptual level [4] and not just for the less complex task of identifying keywords. However, LLMs such as ChatGPT can also produce erroneous output [5]. A common problem with LLMs is their tendency to hallucinate, a problem that may be inherent to their architecture [6]. Borji defines eleven types of LLM errors, including problems with reasoning, logic, humor, ethics, and bias [7]. The evidence of LLM bias, including racial bias, is particularly troubling. For example, LLMs show prejudice related to dialect markers associated with Black English [8] and stereotypes associated with student names [9]. This evidence of bias suggests that caution is warranted when LLMs are used in any task where judgement is required.

1.3 AI in Education Research

To date, there has been limited research on the potential of LLMs to contribute to the research process [10]; this is perhaps not surprising given their novelty. Küchemann et al. cataloged the ways in which LLMs and other AI tools might assist education researchers, including for data cleaning and analysis, literature reviews, assessing learning after an intervention, creating intelligent tutoring systems, and assisting in translation (including across disciplines as well as for natural languages) for diverse research teams [11]. But Küchemann et al. also identified challenges for AI use in education research, including authorship issues, transparency, bias and other ethical issues, and over-reliance.

Some work suggests that LLMs can adequately perform literature searches [12] and find papers to support a given claim [13], but work by Lehr et al. suggests that LLMs are not proficient in curating research studies. One study showed that LLMs generated research ideas that were judged

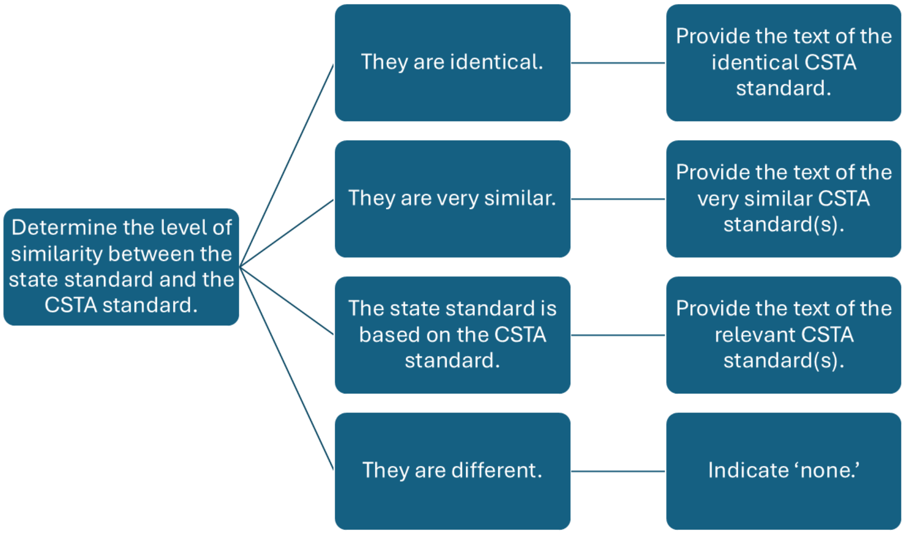
Determine the level of similarity between the state standard and the CSTA standard.
They are identical.
Provide the text of the identical CSTA standard.
They are very similar.
Provide the text of the very similar CSTA standard(s).
The state standard is based on the CSTA standard.
Provide the text of the relevant CSTA standard(s).
They are different.
Indicate ‘none.’

Figure 1: Summary of the process for categorizing standards.

(by humans) to be more novel – but perhaps less feasible – than those produced by (human) researchers [14]. Pack and Maloney explored how LLMs could be used to assist with education research, including qualitative data analysis, while noting that there are unresolved ethical issues – particularly regarding open questions around reliability, validity, and bias [15]. Barany et al. showed the promise of LLMs in the development of codebooks for qualitative analysis, finding that hybrid approaches could improve efficiency but that fully automated codebooks were unlikely to meet researchers’ needs [16].

1.4 The Human Analysis

In preparation for its upcoming effort to revise its computer science (CS) standards for K-12 students, the Computer Science Teachers Association (CSTA) engaged in several initiatives to provide important background information for its standards writers. One of these initiatives was to map all of the state CS standards to the CSTA standards as well as to tag a variety of characteristics of the state standards (e.g., topic, grade level). Subject matter experts with extensive experience in CS education and standards development were recruited for this task and then asked to analyze each state standard. Figure 1 summarizes the process.

First, they were asked to indicate each state standard’s similarity to a CSTA standard, using these categories: (1) identical to, (2) similar to, (3) based on, or (4) entirely different from a CSTA standard. (Note that the classification *based on* indicates only that there are some commonalities in the content of the two standards; it does not require evidence that the CSTA standard was used as a source for the state standard.)

Second, they were asked to choose the relevant CSTA standard for the state standards in categories (1), (2), and (3) or to choose ‘none’ for state standards in category (4). (In a few instances, the subject matter expert assigned more than one similar CSTA standard to a particular state standard.)

2 Methodology

We selected a random sample of 100 standards from the body of about 10,000 state K-12 computer science standards. We used three different LLMs – ChatGPT, Claude, and Llama – to reproduce the human analysis of the standards. We accessed each of these LLMs via Perplexity (`https://www.perplexity.ai/`) in July 2024, using the models available in Perplexity at that time.

We divided the 100 state standards into groups of 10 since larger groups tend to result in the LLM truncating the output without completing the task. We provided each LLM with the following prompt for the first group of 10 standards:

I have provided a file called 'CSTAStandards.csv.' It contains two columns. The first column, called 'Identifier,' contains an ID. The second column, called 'Standard,' contains the text of the CSTA computer science standards. At the end of these instructions, I am going to provide, in the format of a Python list, 10 state standards with their state name and their identification codes. For each state standard, I want you to determine if it is (1) identical to, (2) similar to, (3) based on, or (4) entirely different from a CSTA standard. Please give me a Python list that includes, for each state standard, the following elements: (a) the text of the state standard, (b) the state identification code, (c) the state name, (d) one word — either 'identical,' 'similar,' 'based,' or 'different' — indicating its relationship to a CSTA standard, and (e) if the standard is identical, similar to, or based on a CSTA standard, the text of that CSTA standard or, if it is different, the word 'none.'

Then, after the output was received, we used this prompt for each of the remaining batches of state standards: *Please follow the same instructions above for this list*, followed by the list of 10 state standards.

In the remainder of this paper, we will use the term *verdict* for the determination of whether a standard is *identical*, *similar*, *based*, or *different*. Thus, for example, if the human coded two standards *identical* and the LLM also coded them *identical*, we would say that the human and the LLM reached the same verdict.

We will use the terms *match* and *mismatch* to describe the relationship between the human coder's choice of relevant CSTA standard and the LLM's determination. Thus, if the human and the LLM chose the same CSTA standard as being *identical* to a given state standard, that would be a considered a match. If the human chose 'none' (i.e., because the state standard was different from all CSTA standards) and the LLM did not (or vice versa), we consider that a mismatch. As mentioned previously, in a few instances, the human coder coded a given state standard as *identical*, *similar*, or *based on* more than one CSTA standard. Specifically, two of the state standards were matched to two CSTA standards, and one state standard was matched to three CSTA standards.

In the analysis below, we consider the LLM to have a match if it chose any one of the CSTA standards selected by the human coder.

3 Results

Note that, for the first attempt, ChatGPT provided (incorrect and incomplete) Python code to determine the categorization of the standards instead of performing the categorization itself. We

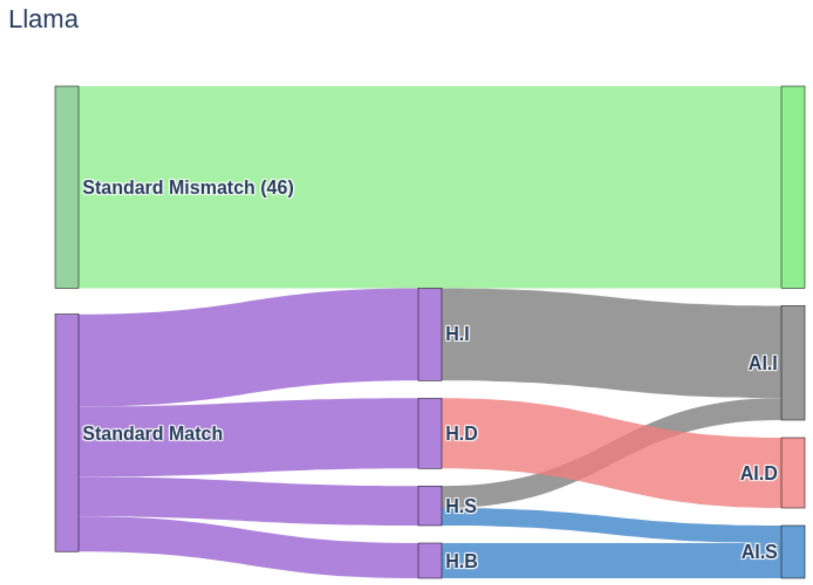

Figure 2: Llama performance. Key: H → human, AI → Llama, I → identical, D → different, S → similar, B → based on.

reprompted it to generate usable results.

Figures 2, 3, and 4 summarize the performance of each LLM. Mismatches with the human coder occurred in about half of the instances: Llama ($n = 46$), Claude ($n = 52$), and ChatGPT ($n = 41$). However, when the LLM had a match with the human coder, it usually had the same verdict as the human coder because it categorized the level of similarity in the same way as the human expert. However, and perhaps surprisingly, there were some instances where the human's verdict was *identical* but the LLM's verdict was not *identical*.

4 Discussion

As Figures 2, 3, and 4 show, all three LLMs had mismatch rates around 50%. For most research situations, an accuracy rate approaching that of a coin flip is not acceptable, and thus we conclude that LLMs are not (yet) capable of replacing human subject matter experts for similar tasks. We also recognize a host of other concerns related to LLMs, including their environmental impact [17], unauthorized use of copyrighted materials [18], potential to lead to significant economic dislocation [19], and ability to amplify misinformation [20].

However, in the instances where there *was* a match, the LLM verdict was usually quite similar to the human verdict. Interestingly, one of the main exceptions to the trend of humans and LLMs reaching the same verdict is that the LLM would sometimes deem a state standard *identical* when it was not. This is a somewhat surprising finding given that identification of identical standards should be a relatively simple task for an LLM. In fact, this task does not actually require artificial intelligence – it is a task for which standard programming strategies of string matching would suffice and have been able to provide accurate results. But all three LLMs over-identified

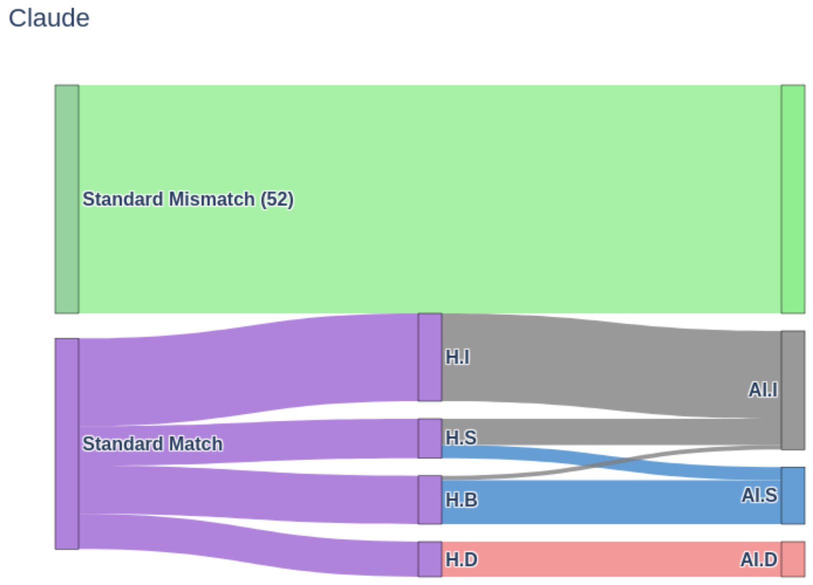

Figure 3: Claude performance. Key: H → human, AI → Claude, I → identical, D → different, S → similar, B → based on.

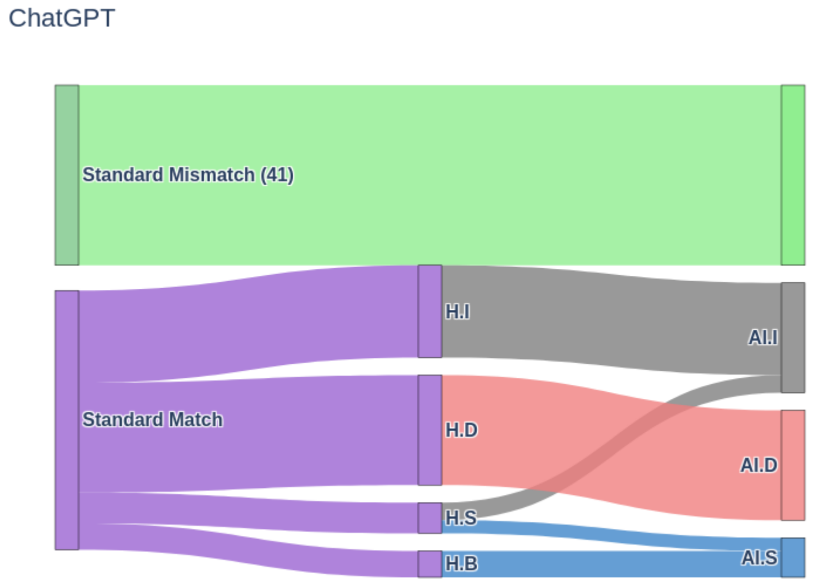

Figure 4: ChatGPT performance. Key: H → human, AI → ChatGPT, I → identical, D → different, S → similar, B → based on.

Entity	Verdict	CSTA Standard
Human	similar	Decompose problems and subproblems into parts to facilitate the design, implementation, and review of programs.
Claude	identical	Decompose problems and subproblems into parts to facilitate the design, implementation, and review of programs.
Llama	identical	Decompose problems and subproblems into parts to facilitate the design, implementation, and review of programs.
ChatGPT	similar	Decompose (break down) problems into smaller, manageable subproblems to facilitate the program development process.

Table 1: Classification for South Carolina HS1.AP.3.1: "Decompose tasks into smaller, reusable parts to facilitate the design, implementation, and review of programs."

	ChatGPT	Claude	Llama
Human error	2	2	2
Nearly identical	7	7	6
Not nearly identical	0	2	0

Table 2: Summary of mismatches where the LLM's verdict was identical.

standards as identical, relative to the human coders. For example, one of the state standards in the sample is South Carolina HS1.AP.3.1: "Decompose tasks into smaller, reusable parts to facilitate the design, implementation, and review of programs." The human coder categorized this standard as *similar* to CSTA 2-AP-13: "Decompose problems and subproblems into parts to facilitate the design, implementation, and review of programs." Both Claude and Llama identified the same CSTA standard, but they both indicated that it was *identical* to that standard, when it obviously is not (see Table 1.)

It may be the case that the LLMs were using a 'fuzzy' approach to determine what constitutes an identical standard as opposed to a strict approach that tested for a verbatim match, although this would still be an inaccurate verdict given the options for verdicts that they were given. Or, it may be that this incorrect verdict is an example of the well-known tendency of LLMs to 'hallucinate' [21, 22].

We manually examined the instances where an LLM's verdict was *identical* but the human's was something else. Results are summarized in Table 2.

In two instances, the human erroneously rendered a verdict of something other than *identical* for standards that were *identical*. In other instances (ChaptGPT: $n = 7$, Claude: $n = 8$, Llama: $n = 7$), the LLM gave a verdict of *identical* to standards that were *nearly* (but not quite) *identical*. An example of one instance of this phenomenon can be found in Table 3.

In one instance, a standard that was *not* nearly identical was deemed *identical* by Claude. Claude deemed Kansas 5.IC.SLE.01 ("Observe intellectual property rights and give appropriate credit when using resources.") as identical to CSTA IB-IC-21 ("Use public domain or creative commons media, and refrain from copying or using material created by others without permission.").

Entity	Verdict	CSTA Standard
Human	similar	Give attribution when using the ideas and creations of others while developing programs.
ChatGPT	identical	Give attribution when using the ideas and creations of others while developing programs.
Claude	identical	Give attribution when using the ideas and creations of others while developing programs.
Llama	identical	Give attribution when using the ideas and creations of others while developing programs.

Table 3: Classification for Nevada 2.AP.PD.2: "Give attribution (credit) when using the ideas and creations of others while developing programs."

Human Categorization	LLM Categorization	ChatGPT	Claude	Llama
different	identical to, similar to, or based on	9	26	18
identical to, similar to, or based on	different	11	1	1

Table 4: Summary of mismatches.

These mismatches regarding the verdict of *identical* are particularly important in the context of this project, which was concerned with whether and how state standards differ from the CSTA standards. Thus, categorizing as *identical* a standard that is not – even if the change is minor – presents a problem in the context of the project.

Also interesting is that, for all cases of matches, if the human verdict was *different*, the LLM verdict was also *different* (and vice versa) – these instances are represented by the pink bars in Figures 2, 3, and 4. In other words, for this portion of the task, the LLM accuracy rate was 100%.

Note that there are instances of mismatches involving the human determining that the state standard was *different* but the LLM pairing it with a CSTA standard (or vice versa); these instances are summarized in Table 4.

A noteworthy feature of Table 4 is how different the pattern is for the LLMs: ChatGPT's mismatches are roughly evenly distributed ($n = 9$ for human verdict of different and $n = 11$ for LLM verdict of different), but Claude's and Llama's are skewed toward the human verdict of different (Claude: $n = 26$ for human verdict of different and $n = 1$ for LLM verdict of different; Llama: $n = 18$ for human verdict of different and $n = 1$ for LLM verdict of different). Given the opacity of the process leading to LLM output, it is not possible to determine why this pattern exists.

In some cases, the differences in the human and the LLM output point to the inherent ambiguity in the task. For example, see Table 5. This table shows the human and LLM classification for Wisconsin standard AP4.a.12.h: "Identify programming language features that can be used to define or specify an abstraction." The human and ChatGPT categorized it as *different* and, indeed,

Entity	Verdict	CSTA Standard
Human	different	none
ChatGPT	different	none
Claude	based	Compare multiple programming languages and discuss how their features make them suitable for solving different types of problems.
Llama	similar	Explain how abstractions hide the underlying implementation details of computing systems embedded in everyday objects.

Table 5: Classification for Wisconsin standard AP4.a.12.h: "Identify programming language features that can be used to define or specify an abstraction."

Entity	Verdict	CSTA Standard
Human	different	none
ChatGPT	different	none
Llama	different	none
Claude	similar to	Compare tradeoffs associated with computing technologies that affect people's everyday activities and career options.

Table 6: Classification for Arkansas standard CSRB.Y1.10.7: "Research and identify diverse careers and career opportunities (e.g., accessibility, availability, demand) that are influenced by computer science and the technical and soft skills needed for each."

there does not appear to be a close match to this standard in any of the CSTA standards. However, Claude categorized it as *based on* CSTA 3B-AP-4: "Compare multiple programming languages and discuss how their features make them suitable for solving different types of problems," perhaps based on the similar language about the "features" of "programming language(s)." On the other hand, Llama identified it as *similar* to CSTA 3A-CS-01: "Explain how abstractions hide the underlying implementation details of computing systems embedded in everyday objects," perhaps based on the reference to abstraction. Whether the Wisconsin standard should ultimately be classified based on the shared reference to language features or to abstraction or to neither is not a question with a clear answer, and it points to the inherent ambiguity in this classification task. However, this very ambiguity makes it difficult to assess the performance of LLMs on such tasks. It is also why subject matter experts with significant experience with writing and using K-12 computer science standards were selected for this project of determining the focus of the standards.

But in some cases, the LLM output appears to be clearly incorrect. For example, Arkansas CSRB.Y1.10.7 reads, "Research and identify diverse careers and career opportunities (e.g., accessibility, availability, demand) that are influenced by computer science and the technical and soft skills needed for each." (See Table 6.) The human, ChatGPT, and Llama classified this as *different*. Claude, however, classified it as *similar* to CSTA 2-IC-20: "Compare tradeoffs associated with computing technologies that affect people's everyday activities and career options." It may be that Claude used the reference to careers in both of these standards, but the context (an emphasis on the student's own future career versus an emphasis on technology tradeoffs) is quite different.

We also note that ChatGPT did not deem any standards as *based on*. Given the opaque nature of LLM output, it is not possible to determine why this is the case.

4.1 Limitations

Despite efforts to ensure uniformity, there was some element of subjectivity in the human coding task, such that one person might have deemed a state standard to be *similar to* a CSTA standard but another person might have determined that it was *based on* that standard. Or, two humans might have selected different CSTA standards, as a result of focusing on different portions of a state standard in order to identify matching content in the CSTA standard. This makes the task of assessing the LLM output sometimes difficult. Other prompting strategies – such as refining the prompt after an analysis of its output – may impact the results.

4.2 Recommendations

Whether to use an AI tool such as an LLM or whether to use humans to accomplish any given task is a complicated calculus. For this project, the resources required to enlist subject matter experts for the hand-coding of about 10,000 state computer science standards was immense, and the process took several months. It is therefore understandable that an automated alternative would have some appeal. At the same time, each of the three LLMs had a mismatch with the human choice of CSTA standard in about half of instances, suggesting that the outcome of human versus LLM coding is quite different. And yet it is unlikely that those mismatches *always* reflect LLM error: it is also possible that sometimes the human coder made a suboptimal choice, especially with such a lengthy and complex task. (The humans needed to decide among 120 CSTA standards when choosing the standard most closely related to each state standard.) In other words, for a complex and occasionally subjective task such as this, we cannot say with confidence that the mismatches always reflect LLM errors. At the same time, there is evidence that the LLM choice was at least sometimes a clear error (e.g., see Table 6). And the pattern of deeming *identical* standards that obviously were not is clearly an error – one unlikely to be made by a human.

Thus, we recommend what has been termed a human-in-the-loop model [23] for the use of LLMs in education research tasks. That is, the current limitations of LLMs are such that they cannot be trusted to accurately complete tasks in education research, as our findings above indicate. However, it may be possible to use an LLM to perform some research tasks, if and only if a human then vets the LLM output. We therefore offer the following recommendations for using LLMs in research:

- *Do not assume accurate LLM output.* Our findings confirm previous studies [24–26] showing that LLM output is not consistently accurate. LLMs will often provide warnings, either in their output or as part of their user interface, regarding the veracity of their content; these warning should be heeded.
- *Check for errors of the type a human would be unlikely to make.* As described above, the LLMs sometimes deemed *identical* a standard that obviously was not. This is not the type of error a human would be likely to make and therefore perhaps not the kind of error that one would be likely to check for. But LLMs often make errors unlike those of humans [9],

necessitating extra vigilance in vetting the output.

- *Do not assume all models will have similar performance.* As shown in Table 4, the models not only differed from the humans but also from each other. Due to the opaque nature of LLM output, it is not possible to determine exactly why this is the case, and asking the LLM to explain its output may yield an inaccurate (e.g., post hoc) explanation [27, 28].
- *Carefully consider design requirements of the research.* For the original project, knowing whether a state standard was identical to a CSTA standard was important, as one of the goals was to determine the differences between state and CSTA standards at a granular level. Thus, in this context, the minor mistakes made by LLMs in this classification suggest an important disadvantage of its use. In other contexts, however, this pattern of errors might be less important, especially when weighed against the resources required for human coding. Thus, researchers will need to carefully consider the impact that LLM mistakes will have and whether these mistakes outweigh the benefits of their use.
- *Remember that humans make mistakes.* We identified at least a few errors in the coding performed by the subject matter experts – something that is to be expected whenever humans perform a lengthy and complicated task. Thus, LLM errors need to be weighed in context of expected human performance, not against an ideal.

5 Conclusion

This study explored whether LLMs could accurately code learning standards, specifically state and national K-12 computer science learning standards. Given that the very architecture of LLMs is based on the identification of patterns in text, it seemed plausible that LLMs would perform well when asked to identify patterns across texts. However, our study showed that all three of the LLMs tested – ChatGPT, Claude, and Llama – did not identify the same learning standard as the human coder in about half of instances. However, in the cases where the LLMs did identify the same standard, they were usually able to classify it as identical to, similar to, or based on that standard in the same way that the human did. One interesting exception to this pattern is that the LLMs sometimes identified as identical standards that were not. We conclude based on these findings that LLMs are not capable of reproducing human-like behavior on this task, although we do offer a set of recommendations for situations where LLMs are used.

6 Acknowledgments

This project is supported by the National Science Foundation (NSF) under Grant No. 2311746. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF. We thank Vicky Sedgwick for her support in collecting and tagging the state standards and her contributions to our analyses and Monica M. McGill for supportive leadership throughout the process.

References

[1] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. Practical and ethical challenges of large language models in education: A

systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112, January 2024. ISSN 0007-1013. doi: 10.1111/bjet.13370. URL https://bera-journals.onlinelibrary.wiley.com/doi/full/10.1111/bjet.13370. Publisher: John Wiley & Sons, Ltd.

[2] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open*, 7(10):e2440969, October 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2024.40969. URL https://doi.org/10.1001/jamanetworkopen.2024.40969.

[3] Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. Goal Driven Discovery of Distributional Differences via Language Descriptions. *Advances in Neural Information Processing Systems*, 36: 40204–40237, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/7e810b2c75d69be186cadd2fe3febeab-Abstract-Conference.html.

[4] Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLooM. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–28, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642830. URL https://doi.org/10.1145/3613904.3642830.

[5] Yuchang Wu, Zhongxuan Sun, Qinwen Zheng, Jiacheng Miao, Stephen Dorn, Shubhabrata Mukherjee, Jason M. Fletcher, and Qiongshi Lu. Pervasive biases in proxy genome-wide association studies based on parental history of Alzheimer's disease. *Nature Genetics*, 56(12):2696–2703, December 2024. ISSN 1546-1718. doi: 10.1038/s41588-024-01963-9. URL https://www.nature.com/articles/s41588-024-01963-9. Publisher: Nature Publishing Group.

[6] Adam Tauman Kalai and Santosh S. Vempala. Calibrated Language Models Must Hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, pages 160–171, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 979-8-4007-0383-6. doi: 10.1145/3618260.3649777. URL https://doi.org/10.1145/3618260.3649777.

[7] Ali Borji. A Categorical Archive of ChatGPT Failures, April 2023. URL http://arxiv.org/abs/2302.03494. arXiv:2302.03494 [cs].

[8] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. Dialect prejudice predicts AI decisions about people's character, employability, and criminality, March 2024. URL http://arxiv.org/abs/2403.00742. arXiv:2403.00742 [cs].

[9] Julie M. Smith. ”I'm Sorry, but I Can't Assist”: Bias in Generative AI. In *Proceedings of the 2024 on RESPECT Annual Conference*, RESPECT 2024, pages 75–80, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 9798400706264. doi: 10.1145/3653666.3656065. URL https://doi.org/10.1145/3653666.3656065.

[10] Steven A. Lehr, Aylin Caliskan, Suneragiri Liyanage, and Mahzarin R. Banaji. ChatGPT as Research Scientist: Probing GPT's Capabilities as a Research Librarian, Research Ethicist, Data Generator and Data Predictor, June 2024. URL http://arxiv.org/abs/2406.14765. arXiv:2406.14765 [cs].

[11] Stefan Küchemann, Karina Avila, Yavuz Dinc, Chiara Hortmann, Natalia Revenga Lozano, Verena Ruf, Niklas Stausberg, Steffen Steinert, Frank Fischer, Martin Fischer, Enkelejda Kasneci, Gjergji Kasneci, Thomas Kuhr, Gitta Kutyniok, Sarah Malone, Michael Sailer, Albrecht Schmidt, Matthias Stadler, Jochen Weller, and Jochen Kuhn. *Are Large Multimodal Foundation Models all we need? On Opportunities and Challenges of these Models in Education*. January 2024. doi: 10.35542/osf.io/n7dvf.

[12] Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. LitSearch: A

Retrieval Benchmark for Scientific Literature Search, July 2024. URL https://arxiv.org/abs/2407.18940v1.

[13] Ori Press, Andreas Hochlehnert, Ameya Prabhu, Vishaal Udandarao, Ofir Press, and Matthias Bethge. CiteME: Can Language Models Accurately Cite Scientific Claims?, July 2024. URL https://arxiv.org/abs/2407.12861v1.

[14] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers, September 2024. URL https://arxiv.org/abs/2409.04109v1.

[15] Austin Pack and Jeffrey Maloney. Using Generative Artificial Intelligence for Language Education Research: Insights from Using OpenAI's ChatGPT. *TESOL Quarterly*, 57(4):1571–1582, 2023. ISSN 1545-7249. doi: 10.1002/tesq.3253. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/tesq.3253. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/tesq.3253.

[16] Amanda Barany, Nidhi Nasiar, Chelsea Porter, Andres Zambrano, Alexandra Andres, Dara Bright, Mamta Shah, Xiner Liu, Sabrina Gao, Jiayi Zhang, Shruti Mehta, Jaeyoon Choi, Camille Giordano, and Ryan Baker. *ChatGPT for Education Research: Exploring the Potential of Large Language Models for Qualitative Codebook Development*. July 2024.

[17] Alba M. Mármol Romero, Adrián Moreno-Muñoz, F. Plaza-del Arco, M. Dolores Molina González, and Arturo Montejo-Ráez. Environmental Impact Measurement in the MentalRiskES Evaluation Campaign. 2024. URL https://www.semanticscholar.org/paper/Environmental-Impact-Measurement-in-the-Evaluation-Romero-Moreno-Mu%C3%B1oz/363fa4d3f5e2e1a0b4c071192068377f582d0282.

[18] Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright Violations and Large Language Models, October 2023. URL http://arxiv.org/abs/2310.13771. arXiv:2310.13771 [cs].

[19] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308, June 2024. doi: 10.1126/science.adj0998. URL https://www.science.org/doi/10.1126/science.adj0998. Publisher: American Association for the Advancement of Science.

[20] Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. On the Risk of Misinformation Pollution with Large Language Models, October 2023. URL http://arxiv.org/abs/2305.13661. arXiv:2305.13661 [cs].

[21] Gabrijela Perković, Antun Drobnjak, and Ivica Botički. Hallucinations in LLMs: Understanding and Addressing Challenges. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, pages 2084–2088, May 2024. doi: 10.1109/MIPRO60963.2024.10569238. URL https://ieeexplore.ieee.org/abstract/document/10569238. ISSN: 2623-8764.

[22] Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, Li Zhang, Zhongqi Li, and Yuchi Ma. Exploring and Evaluating Hallucinations in LLM-Powered Code Generation, May 2024. URL http://arxiv.org/abs/2404.00971. arXiv:2404.00971 [cs].

[23] Mohamed A. Mabrok, Hassan K. Mohamed, Abdel-Haleem Abdel-Aty, and Ahmed S. Alzahrani. Human models in human-in-the-loop control systems. *Journal of Intelligent & Fuzzy Systems*, 38(3):2611–2622, January 2020. ISSN 1064-1246. doi: 10.3233/JIFS-179548. URL https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs179548. Publisher: IOS Press.

[24] Li Zhong and Zilong Wang. Can LLM Replace Stack Overflow? A Study on Robustness and Reliability of Large Language Model Code Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (19):21841–21849, March 2024. ISSN 2374-3468. doi: 10.1609/aaai.v38i19.30185. URL https://ojs.aaai.org/index.php/AAAI/article/view/30185. Number: 19.

[25] Sungmin Kang, Louis Milliken, and Shin Yoo. Identifying Inaccurate Descriptions in LLM-generated Code Comments via Test Execution, June 2024. URL http://arxiv.org/abs/2406.14836. arXiv:2406.14836 [cs].

[26] Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review. *Diagnostic Pathology*, 19(1):43, February 2024. ISSN 1746-1596. doi: 10.1186/s13000-024-01464-7. URL https://doi.org/10.1186/s13000-024-01464-7.

[27] Jenny Kunz and Marco Kuhlmann. Properties and Challenges of LLM-Generated Explanations, February 2024. URL http://arxiv.org/abs/2402.10532. arXiv:2402.10532 [cs].

[28] Rohan Ajwani, Shashidhar Reddy Javaji, Frank Rudzicz, and Zining Zhu. LLM-Generated Black-box Explanations Can Be Adversarially Helpful, October 2024. URL http://arxiv.org/abs/2405.06800. arXiv:2405.06800 [cs].