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 #45868

Can AI Develop Curriculum? Integrated Computer Science As a Test Case (Research to Practice)

Dr. Julie M. Smith, Institute for Advancing Computing Education

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.

Monica McGill, Institute for Advanced Engineering

Monica McGill is President & CEO of the Institute for Advancing Computing Education (formerly known as CSEdResearch.org). Have previously worked in industry and academia, McGill is using her experiences as a computer scientist, professor, and researcher to enable others to build a strong foundation of CS education research focused on all children while also conducting it with partners and collaborators.

Jacob Koressel
Bryan Twarek

Can AI Develop Curriculum? Integrated Computer Science As a Test Case (Research to Practice)

Abstract

Introduction: Because developing integrated computer science (CS) curriculum is a resource-intensive process, there is interest in leveraging the capabilities of AI tools, including large language models (LLMs), to streamline this task. However, given the novelty of LLMs, little is known about their ability to generate appropriate curriculum content.

Research Question: How do current LLMs perform on the task of creating appropriate learning activities for integrated computer science education?

Methods: We tested two LLMs (Claude 3.5 Sonnet and ChatGPT 4-o) by providing them with a subset of K-12 national learning standards for both CS and language arts and asking them to generate a high-level description of learning activities that met standards for both disciplines. Four humans rated the LLM output – using an aggregate rating approach – in terms of (1) whether it met the CS learning standard, (2) whether it met the language arts learning standard, (3) whether it was equitable, and (4) its overall quality.

Results: For Claude AI, 52% of the activities met language arts standards, 64% met CS standards, and the average quality rating was middling. For ChatGPT, 75% of the activities met language arts standards, 63% met CS standards, and the average quality rating was low. Virtually all activities from both LLMs were rated as neither actively promoting nor inhibiting equitable instruction.

Discussion: Our results suggest that LLMs are not (yet) able to create appropriate learning activities from learning standards. The activities were generally not usable by classroom teachers without further elaboration and/or modification. There were also grammatical errors in the output, something not common with LLM-produced text. Further, standards in one or both disciplines were often not addressed, and the quality of the activities was often low. We conclude with recommendations for the use of LLMs in curriculum development in light of these findings.

1 Introduction and Background

The public release of ChatGPT in November 2022 inaugurated a new era of AI as large language models (LLMs) captured public attention due to their ability to generate fluent responses to open-ended questions. LLMs quickly proved their usefulness in tasks ranging from the generation of novel research ideas [1] to making medical decisions [2]. Despite their widespread adoption

and application, however, concerns remain about their accuracy [3] as well as ethical issues, including their environmental impact [4] and potential for bias [5].

1.1 LLMs in Education

In a systematic literature review of the use of LLMs in education, Yan et al. identified nine different ways that they are being used, including assessment, performance prediction, and providing feedback to students [6]. While developing curriculum is not specifically mentioned in their taxonomy, two of the categories – teaching support and content generation – have significant overlap. This review found a range of accuracy levels by the LLMs, from high levels on simpler tasks (such as sentiment analysis) to lower levels on more complex tasks (such as grading free form responses); the review also noted ethical concerns including a lack of transparency, data privacy issues, and bias.

In a review of large multimodal foundation models (which integrate LLMs with the ability to interact with other media such as audio and images), Küchemann et al. explored the implications of using these models in various educational contexts [7]. The advantages include the ability to design lessons and classroom activities that are current, engaging, and customized, but they also noted potential problems, including inaccuracies, bias, overuse leading to dependence, and respect for the intellectual property used to create the models.

A recent exploration of the ability of LLMs to develop mathematics curriculum found that GPT-4 was able to perform well in some areas (e.g., organizing activities, selecting strategies, identifying priorities) but struggled with other aspects (e.g., interdisciplinary assessment, certain math topics) [8]. Research focused on high school social studies curriculum found LLMs had a very high ($> 90\%$) success rate (per evaluation by human subject matter experts) in generating assessment questions aligned to Bloom's taxonomy [9].

1.2 Integrated CS Instruction

Computer science (CS) can be conceptualized as a subset of engineering [10]. In recent years, CS instruction in K-12 has expanded rapidly, with 35 states adopting CS learning standards between 2017 and 2023 (for a total of 43 states that now have CS standards) and nine states creating a CS requirement for high school graduation between 2016 and 2024 [11]. However, it is difficult for schools to offer computer science learning opportunities to all students [12] due to resource limitations. One potential solution is to integrate CS into other subject areas, which may result in more students having access to high-quality CS experiences, especially in the earlier grades [11, 13].

However, a major obstacle to such integration is the lack of curriculum, the development of which would require expertise in two subjects, as well as the ability to craft appropriate activities, address the needs of diverse learners, and meet learning standards, among other requirements. Because such expertise is in short supply, it is understandable that alternatives to human subject matter experts – including AI tools such as LLMs – would be considered for their ability to generate learning activities that meet the need for curriculum that integrates CS into another school subject. However, given the novelty of LLMs, little is known about their ability to generate appropriate curriculum content. Hence, this study explores the research question: How do current LLMs perform on the task of creating appropriate learning activities for integrated computer science education?

Item	Response Options
Does the activity meet the Common Core Standard?	Yes, Partially, No
Does the activity meet the CSTA standard?	Yes, Partially, No
Rate the equity of the activity.	Positive, Neutral, Problematic
Rate the overall quality of the activity.	High, Middling, Low

Table 1: Each proposed activity was coded by each of the four authors using the above questions and response options.

2 Methods

We used a selection of computer science and English language arts learning standards to assess whether LLMs can generate integrated learning activities. For the computer science learning standards, we used the 2017 CSTA middle school (grades 6th - 8th or ages 11 - 14) computer science learning standards [14]. This set of 23 learning standards covers five concepts: Computing Systems, Networks and the Internet, Data and Analysis, Algorithms and Programming, and Impacts of Computing. An example of these standards is CSTA 2-AP-19: "Document programs in order to make them easier to follow, test, and debug." For the English language arts learning standards, we used a random subset of 20 of the Common Core literacy standards [15]. An example of these standards is CCSS.ELA-LITERACY.L.6.5: "Demonstrate understanding of figurative language, word relationships, and nuances in word meanings." (Because the subset was determined randomly, a standard listed for more than one grade could appear more than once in our subset. This was in fact the case for the standard "Adapt speech to a variety of contexts and tasks, demonstrating command of formal English when indicated or appropriate," which appears in the set three times: once each for 6th, 7th, and 8th grade. All other standards appear only once in the dataset.)

We used two LLMs, ChatGPT 4-o and ClaudeAI 3.5 Sonnet, in August 2024. We accessed both through the Perplexity platform, and we used the same prompt for each:

> I have uploaded two files. The first, 'Common Core Standards,' includes some of the Common Core English Language Arts/Literacy standards. The second, 'CSTA Middle School Standards', includes middle school computer science standards. For each of the Common Core standards, please generate two classroom activities. Each classroom activity should address both the Common Core standard as well as one of the CSTA standards. Please indicate which CSTA standard is addressed. Please do this for each of the 20 Common Core standards.

We created a spreadsheet with the output, with columns for the generated activity, the Common Core standard, and the CSTA standard. Each of the four authors independently used a copy of this spreadsheet to answer the questions shown in Table 1.

The four authors can be considered to have subject matter expertise in K-12 computer science education based on their past and current roles as CS educators (at both K-12 and tertiary levels), as well as work as curriculum developers, researchers, in industry, and in state and national CS leadership.

There is an element of subjectivity in the questions we pose about each activity such that, for

example, two subject matter experts may well disagree about the quality of a learning activity. Thus, we chose to use an aggregate rating system instead of assessing the level of inter-rater reliability in recognition of the reality that expert opinion often includes a diversity of beliefs instead of one correct answer [16]. As a result of this method, each activity has four scores per question (one for each author).

3 Results

Table 2 shows a representative sample of each LLM's output. This sample is for Common Core Standard R1.6.9: "Compare and contrast one author's presentation of events with that of another (e.g., a memoir written by and a biography on the same person)."

LLM	Claude 3.5 Sonnet	ChatGPT 4-o
Activity	Digital Biography Comparison: Students will compare and contrast a memoir and a biography of the same person using digital tools. They'll collect data on key events, perspectives, and language use, then transform this data into visual representations like timelines or word clouds to highlight differences.	Develop a project that create [sic] procedures with parameters to organize code and make it easier to reuse to enhance understanding of the Common Core standard: compare and contrast one author's presentation of events with that of another.
CSTA Standard	Collect data using computational tools and transform the data to make it more useful and reliable.	Create procedures with parameters to organize code and make it easier to reuse.

Table 2: Sample LLM Output

We evaluated whether each activity met its corresponding Common Core standard. For the activities produced by Claude AI, the 'Yes' response was awarded in 52% of cases ($n = 83$), 'Partially' in 41% ($n = 65$), and 'No' in 8% ($n = 12$). For the activities produced by ChatGPT, 75% of responses were 'Yes' ($n = 120$), 1% were 'Partially' ($n = 1$), and 24% were 'No' ($n = 39$). These results are summarized in Table 3.

LLM	**Yes**	**Partially**	**No**
Claude 3.5 Sonnet	83 (52%)	65 (41%)	12 (8%)
ChatGPT 4-o	120 (75%)	1 (1%)	39 (24%)

Table 3: Aggregate count and percent of responses to the question "Does this activity meet the Common Core standard?"

We also evaluated whether each activity met a CSTA standard. For the activities produced by Claude AI, the 'Yes' response was awarded in 64% of cases ($n = 103$), 'Partially' in 27% ($n = 43$), and 'No' in 9% ($n = 14$). For the activities produced by ChatGPT, 63% of responses were 'Yes' ($n = 100$), 14% were 'Partially' ($n = 22$), and 24% were 'No' ($n = 38$). See Table 4.

LLM	Yes	Partially	No
Claude 3.5 Sonnet	103 (64%)	43 (27%)	14 (9%)
ChatGPT 4-o	100 (63%)	22 (14%)	38 (24%)

Table 4: Aggregate count and percent of responses to the question "Does this activity meet the CSTA standard?"

For ChatGPT, out of the 160 scoring events (four scores for each of 40 activities), the activity was deemed to have met both Common Core and CSTA standards in 100 instances, representing 63% of total instances. For Claude AI, 66 out of 160 instances (42%) indicated that an activity met both sets of standards.

We assessed the equity of each activity. For the activities generated by ChatGPT, there were 160 neutral ratings and no positive or negative ratings. For the activities generated by Claude AI, there were 158 (99%) neutral ratings and 2 (1%) problematic ratings; no activities were rated as positively promoting equity. One activity was rated as problematic by two of the four reviewers; the activity was: "Students will create a program that simulates different speaking contexts. The program will prompt users to adapt their language based on the given scenario, helping students practice formal and informal English." This activity can be perceived as discouraging students from using authentic language in favor of 'adapting' their language. Table 5 summarizes these results.

LLM	Positive	Neutral	Problematic
Claude 3.5 Sonnet	0 (0%)	158 (99%)	2 (1%)
ChatGPT	0 (0%)	160 (100%)	0 (0%)

Table 5: Aggregate count and percent of responses to the prompt "Rate the equity of the activity."

We also assessed the quality of each activity. For Claude AI, there were 40 responses (25%) indicating high quality, 83 (52%) for middling quality, and 37 (23%) for low quality. For ChatGPT, there were 159 low quality (99%) ratings and 1 rating indicating middling quality (1%); there were no high quality ratings. The reason for this result is that ChatGPT's 'activities' restated the standards (see the example in Table 2).

Table 6 presents these results.

LLM	High	Middling	Low
Claude 3.5 Sonnet	40 (25%)	83 (52%)	37 (23%)
ChatGPT 4-o	0 (0%)	1 (1%)	159 (9%)

Table 6: Aggregate count and percent of responses to the prompt "Rate the quality of the activity."

Finally, we determined how many activities met the bar for all four items (that is, they met both Common Core and CSTA standards, were not problematic in terms of equity, and were of high quality).

As Figure 1 indicates, in almost all instances where an activity was rated as meeting the CSTA standard, it also met the Common Core standard and had a neutral or positive equity rating. This

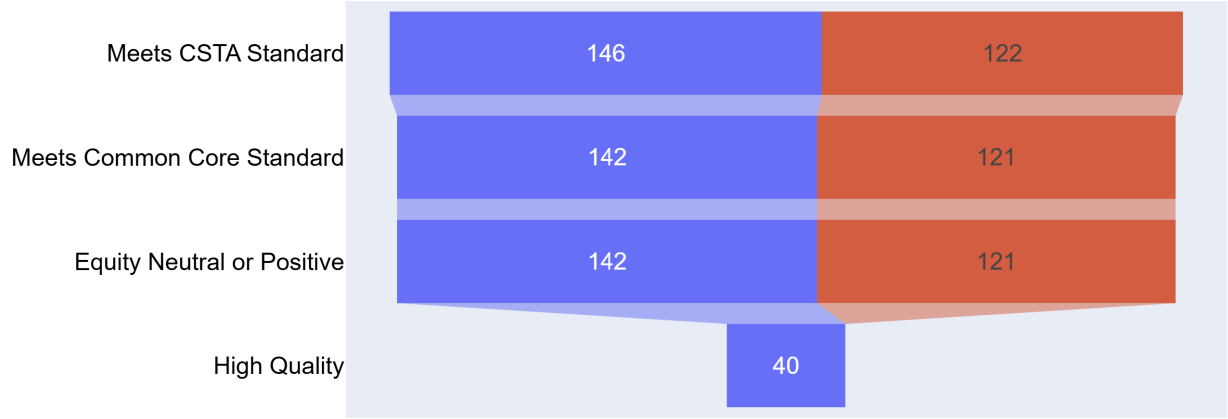

Figure 1: A funnel chart indicates the count of ratings for meeting Common Core standards, CSTA standards, having neutral or positive equity, and being high quality. Results for Claude 3.5 Sonnet are in blue; results for ChatGPT 4-o are in red. Thus, for Claude's output, there were 146 counts of 'meets CSTA standards.' Of those, there were 142 counts of 'meets Common Core standards,' and, of those, 142 counts of neutral or positive equity, and, of those, 40 counts of high quality.

Count	CSTA Standard
8	2-DA-08
5	2-DA-09
4	2-AP-12, 2-DA-07, 2-AP-10
3	2-AP-11
2	2-AP-13, 2-AP-14, 2-AP-16, 2-AP-17, 2-AP-18
1	2-CS-02, 2-AP-15

Table 7: Frequency table for Claude 3.5 Sonnet's use of CSTA standards.

was the case for both Claude AI and ChatGPT. However, for Claude AI, only 40 (28%) were also rated as being of high quality of those items receiving acceptable ratings on the previous three questions; for ChatGPT, no item was rated of high quality in any instance.

We also examined how many different CSTA standards were used by each LLM and how many times each standard was used. ChatGPT used 21 different CSTA standards across the 40 activities. Interestingly, each CSTA standard appeared twice, with the exception of two CSTA standards that appeared only once. Claude AI used 13 different CSTA standards, each appearing between one and eight times in the activities; see Table 7.

4 Discussion

Our findings suggest that LLMs, in the context of this study, are not (yet) capable of generating learning activities that integrate computer science and English language arts content. For ChatGPT, there was near unanimity across activities and across authors that the quality was low. This is because ChatGPT essentially reiterated the task instead of generating an activity. In general, it would basically restate both of the standards with the direction to develop a project that meets both standards without specifying how this might be done. For example, for the Common

Core standard asking students to "analyze how differences in the points of view of the characters and the audience or reader create such effects as suspense or humor," ChatGPT proposed that students "develop a project that systematically identify [sic] and fix [sic] problems with computing devices and their components to enhance understanding of the Common Core standard: analyze how differences in the points of view of the characters and the audience or reader create such effects as suspense or humor." This project was intended to meet this CSTA standard: "systematically identify and fix problems with computing devices and their components."

Not only does this activity merely repeat the Common Core and CSTA standard content without providing any additional guidance for teachers, but the proposed activity asks students to identify and fix computing problems to analyze how suspense or humor is created. It is difficult to imagine an adult with both computing and literacy expertise – let alone a middle school student – figuring out how to do these two things as part of one coherent project. All of the activities suggested by ChatGPT were similar to this example in that they simply reiterated the Common Core and the CSTA standard with no explanation as to how they might be combined in one activity.

While Claude AI did in fact generate learning activities instead of merely repeating the Common Core and CSTA standards, only 25% of those activities received the high rating for quality. For example, the following activity generated by Claude AI received three quality ratings of high: "Interactive Grammar Game: Students will develop an interactive game that tests and teaches standard English grammar. They'll incorporate existing code libraries for game mechanics and focus on creating engaging grammar challenges." More common were activities with the middling rating, such as this example, which received three middling ratings: "Narrative Viewpoint Analyzer: Students will develop a tool that analyzes text to identify and categorize different narrative viewpoints. They'll collect data from various texts and use computational tools to transform this data into insights about authorial techniques." Note that while this activity makes sense, it is not ready to implement as written, particularly for educators with limited CS teaching experience. In this case, the educator would need to know how to guide middle school students in the development of a tool to analyze text.

Further, some of the activities may not be possible to implement successfully, even for teachers with a high level of pedagogical content knowledge. For example, Claude suggested that students "develop and systematically test a program that identifies instances of dramatic irony in a text by comparing character knowledge with reader knowledge." Computer science researchers struggle to develop computational tools that can detect irony [17]; it is not likely that a middle school student would be able to create such a tool.

We note that virtually all of the equity ratings were neutral, with just a few problematic and none positive. It is likely the case that, had our prompt specifically mentioned equity, the activities would have had more positive equity ratings. However, we think it is important to note that it appears that specific prompting for equity is required and should not be expected by default.

Additionally, we noticed something that is otherwise rare in LLM-generated text: English usage errors. The ChatGPT responses, as the example above shows, contained usage errors because they repeated the text of the Common Core standard without changing the verb form to match the context of the output (e.g., 'identify' instead of 'identifies.')

As a result of these findings, teachers and curriculum developers should not rely on LLMs to generate learning activities because the output would be unlikely to meet their expectations. As Figure 1 shows, none of the activities generated by ChatGPT met the tested requirements, and only about one quarter of those generated by Claude did. It is possible that future advances in LLMs would lead this recommendation to change, but it may be the case that some limitations of LLMs – including hallucinations resulting from the nature of their architecture [3] – simply cannot be overcome, and specialized curriculum development will remain a task solely for human experts. To the extent that LLMs are used for curriculum development, a human-in-the-loop model [18] is required; that is, all LLM output should be vetted and modified by subject matter experts to ensure that it is accurate, equitable, and meets other requirements.

Future research testing LLM capabilities for education research might add a creativity score to the metrics used in this study, in order to gain a better perspective on whether LLMs can reproduce human performance more broadly. Future research could also explore whether other prompting approaches – such as providing examples of high-quality activities, asking for one activity at a time, or iteratively prompting would yield different results. We also note that, due to the nature of LLMs, re-using the same prompt will not result in identical results – a challenge to the research process.

5 Conclusion

The findings of this study suggest that LLMs cannot be successfully used to generate high-level ideas for integrated CS learning activities. Due to the resource-intensive nature of curriculum development, it is perhaps to be expected that automated approaches will be desirable for this and similar tasks. It may be the case that curriculum development remains the province of humans for the foreseeable future. Or, perhaps future advances in AI – including ensemble models [19] and/or retrieval augmented generation [20] – may prove more adept at curriculum development tasks.

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.

References

[1] 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.

[2] Ethan Goh, Bryan Bunning, Elaine Khoong, Robert Gallo, Arnold Milstein, Damon Centola, and Jonathan H. Chen. ChatGPT Influence on Medical Decision-Making, Bias, and Equity: A Randomized Study of Clinicians Evaluating Clinical Vignettes, November 2023. URL https://www.medrxiv.org/content/10.1101/2023.11.24.23298844v1. Pages: 2023.11.24.23298844.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Bihao Hu, Longwei Zheng, Jiayi Zhu, Lishan Ding, Yilei Wang, and Xiaoqing Gu. Teaching Plan Generation and Evaluation With GPT-4: Unleashing the Potential of LLM in Instructional Design. *IEEE Transactions on Learning Technologies*, 17:1445–1459, 2024. ISSN 1939-1382. doi: 10.1109/TLT.2024.3384765. URL https://ieeexplore.ieee.org/document/10490240. Conference Name: IEEE Transactions on Learning Technologies.

[9] Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. How Good are Modern LLMs in Generating Relevant and High-Quality Questions at Different Bloom's Skill Levels for Indian High School Social Science Curriculum? In Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan, editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 1–10, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.bea-1.1/.

[10] Michael C. Loui. Computer science is a new engineering discipline. *ACM Comput. Surv.*, 27(1):31–32, March 1995. ISSN 0360-0300. doi: 10.1145/214037.214049. URL https://dl.acm.org/doi/10.1145/214037.214049.

[11] CSTA, IACE, ACM, Code.org, College Board, CSforAll, and ECEP Alliance. Reimagining CS Pathways: Every student prepared for a world powered by computing. Technical report, 2024. URL https://reimaginingcs.org/wp-content/uploads/2024/07/CSTA-Reimagining-CS-Pathways_v4.0.pdf.

[12] Code.org, CSTA, and ECEP Alliance. 2023 State of Computer Science Education. Technical report, Code.org, CSTA & ECEP Alliance, 2023. URL https://advocacy.code.org/2023_state_of_cs.pdf.

[13] Monica M. McGill and Laycee Thigpen. Extrinsic Barriers to Integrating Computer Science in Elementary School Subject Areas in the United States. In *Proceedings of the 19th WiPSCE Conference on Primary and Secondary Computing Education Research*, WiPSCE '24, pages 1–10, New York, NY, USA, September 2024. Association for Computing Machinery. ISBN 979-8-4007-1005-6. doi: 10.1145/3677619.3678116. URL https://dl.acm.org/doi/10.1145/3677619.3678116.

[14] Deborah Seehorn, Tammy Primann, Todd Lash, Bryan Twarek, Daniel Moix, Leticia Batista, Julia Bell, Chris Kuszmaul, Dianne O'Grady-Cunniff, Minsoo Park, Lori Pollock, Meg Ray, Dylan Ryder, Vicky Sedgwick, Grant Smith, and Chimna Uche. CSTA K-12 Computer Science Standards Revised 2017. Technical report, Computer Science Teachers Association, 2017. URL https://members.csteachers.org/documents/en-us/46916364-83ab-4f51-85fb-06b3b25b417c/1/.

[15] National Governors Association Center for Best Practices and Council of Chief State School Officers. English Language Arts Standards | Common Core State Standards Initiative. URL https://www.thecorestandards.org/ELA-Literacy/.

[16] Bernard Charlin, Martin Desaulniers, Robert Gagnon, Daniel Blouin, and Cees van der Vleuten. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. *Teaching and Learning in Medicine*, 14(3):150–156, 2002. ISSN 1040-1334. doi: {10.1207/S15328015TLM1403_3}.

[17] Yucheng Lin, Yuhan Xia, and Yunfei Long. Augmenting emotion features in irony detection with Large language modeling, April 2024. URL http://arxiv.org/abs/2404.12291. arXiv:2404.12291 [cs].

[18] 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.

[19] Han Yang, Mingchen Li, Huixue Zhou, Yongkang Xiao, Qian Fang, and Rui Zhang. One LLM is not Enough: Harnessing the Power of Ensemble Learning for Medical Question Answering. *medRxiv*, page 2023.12.21.23300380, December 2023. doi: 10.1101/2023.12.21.23300380. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10775333/.

[20] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey, March 2024. URL http://arxiv.org/abs/2312.10997. arXiv:2312.10997 [cs].