

# Designing Answer-Aware LLM Hints to Scaffold Deeper Learning in K–12 Programming Education

Sahana Bhaskar  
Department of Computer Science  
Virginia Tech  
Blacksburg, Virginia, USA  
sahanab@vt.edu

Sally Hamouda  
Department of Computer Science  
Virginia Tech  
Blacksburg, Virginia, USA  
shamouda@vt.edu

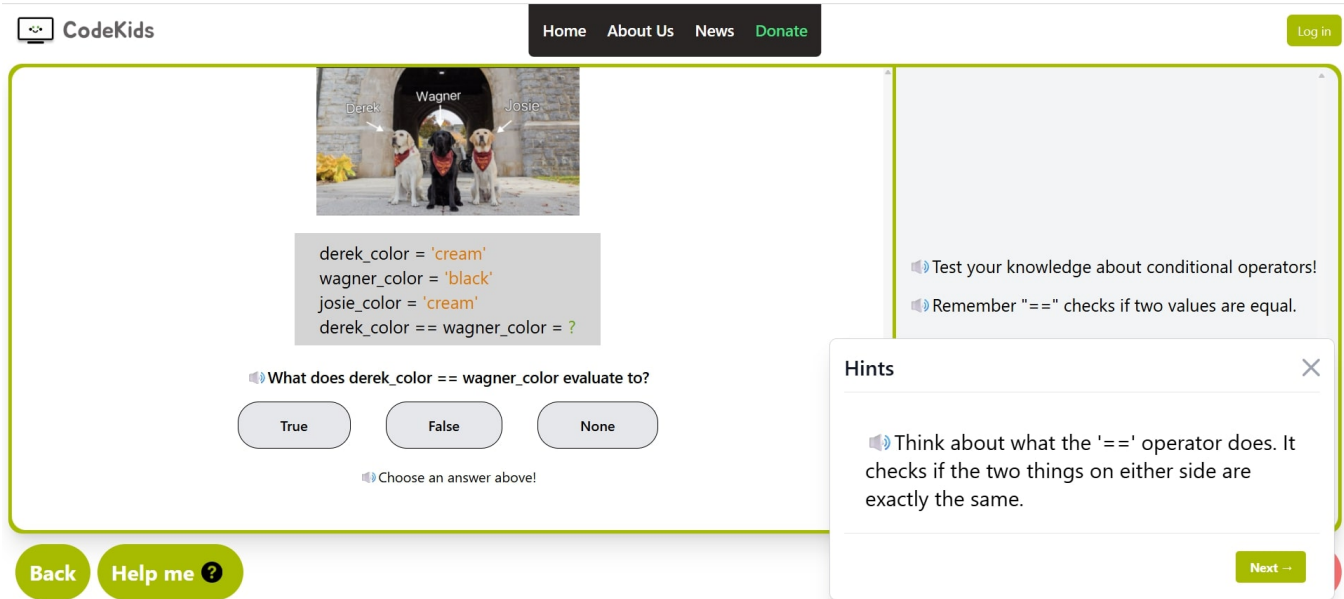


Figure 1: CodeKids hinting system in action.

## Abstract

*Motivation and Background.* Many K–12 students struggle with programming concepts. While LLMs offer scalable, timely support, overly direct answers can reduce reasoning and engagement [8], prompting the question: How can LLMs support learning without encouraging overreliance?

In our study with 105 students, 31.4% showed misconceptions about variable assignment and data types, and in another survey, only 20% correctly solved conditional problems. This highlights the need for scaffolding to address conceptual gaps in K–12 programming.

To address these gaps, we designed an answer-aware hint generation system using LLMs to support learning without reducing cognitive demand. We developed the system for *CodeKids*—an open-source, curriculum-aligned platform built with Virginia Tech and

local public schools. It helps students practice grade-level programming through interactive activities, using LLM-generated hints to guide thinking without revealing answers [1, 11].

Based on Vygotsky’s Zone of Proximal Development [12], our approach balances support and autonomy through structured prompting that preserves productive struggle.

*Methodology.* Building on research showing that machine learning supports K–12 learners without compromising cognitive development [15], we implemented a mindful answer-aware prompting approach [5, 7] grounded in two principles. The first principle, cognitive scaffolding, draws from ZPD and ITS research [10, 12], and ensures hints progress from general to specific while preserving learner autonomy. The second principle, technical safeguards, applies semantic similarity thresholds and constraint-based prompting to prevent answer leakage [13].

The system is deployed across 12 advanced *CodeKids* books covering core topics like variables, data types, conditionals, loops, and logical operators. Hints are concise, pedagogically sound, and generated by GPT-4 when students request help or load a page. Each request includes the topic, question, answer choices, and correct answer sent to the LLM, enabling context-aware adaptation to the activity and content.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

ICER 2025 Vol. 2, Charlottesville, VA, USA

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1341-5/25/08

<https://doi.org/10.1145/3702653.3744323>

Our prompt design constrains hints to one sentence, emphasizes conceptual clarity, and gradually increases specificity to preserve student agency. This aligns with research on scaffold types—such as sense-making, elaboration, and motivational cues—that support self-regulated learning [9].

To support diverse learners, the system includes text-to-speech for reading hints aloud. Our approach combines learning sciences and prompt engineering to foster scalable support, student agency, and conceptual understanding.

**Evaluation.** We evaluated semantic hint alignment using sentence embeddings: 98.1% of hints scored  $\geq 0.30$  in content alignment and 44.2%  $\geq 0.20$  in answer alignment, indicating strong relevance with minimal over-reliance. GPT-4, used as an LLM-as-a-judge due to its  $>85\%$  agreement with human ratings [14], gave an average score of 0.958 for hints on convergence, pedagogical value, and context. Combining LLM and cosine scores (0.7/0.3), we computed a Hint Quality Score of 0.749 [3]. To assess real-world impact, we developed surveys to collect feedback on clarity, usefulness, and learning [4].

**Ongoing Work and Vision.** We are investigating hint convergence across LLMs (e.g., Claude 3, Gemini 1.5 Pro) and exploring alternative prompting strategies to improve diversity. Future work includes personalizing hints through difficulty adaptation and embedding-based models for curriculum-aligned scaffolding [6], reducing reliance on proprietary LLMs, and incorporating retrieval-augmented generation (RAG) for contextualization [2].

## CCS Concepts

• **Social and professional topics** → K-12 education; • **Applied computing** → Interactive learning environments; • **Computing methodologies** → Machine learning; Natural language generation.

## Keywords

K-12 Education, Interactive Learning Environments, Hint Generation, Large Language Models, Natural Language Generation

### ACM Reference Format:

Sahana Bhaskar and Sally Hamouda. 2025. Designing Answer-Aware LLM Hints to Scaffold Deeper Learning in K-12 Programming Education. In *ACM Conference on International Computing Education Research V.2 (ICER 2025 Vol. 2), August 03–06, 2025, Charlottesville, VA, USA*. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3702653.3744323>

## References

- [1] Anubhav Jangra, Jamshid Mozafari, Adam Jatowt, and Smaranda Muresan. 2024. Navigating the Landscape of Hint Generation Research: From the Past to the Future. *arXiv preprint arXiv:2404.04728* (2024). arXiv:2404.04728 [cs.CL] <https://arxiv.org/abs/2404.04728> Version 2, submitted on 26 Nov 2024.
- [2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint arXiv:2005.11401* (2020). arXiv:2005.11401 [cs.CL]
- [3] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-Judge. *arXiv preprint arXiv:2411.16594* (2024). arXiv:2411.16594 [cs.AI]
- [4] Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence* 6 (2024), 100199. doi:10.1016/j.caeai.2023.100199
- [5] Jamshid Mozafari, Florian Gerhold, and Adam Jatowt. 2024. Using Large Language Models in Automatic Hint Ranking and Generation Tasks. *arXiv preprint arXiv:2412.01626* (2024). arXiv:2412.01626 [cs.CL] <https://arxiv.org/abs/2412.01626> Version 1, submitted on 2 Dec 2024.
- [6] Konstantinos Pliakos, Isidoros Perikos, and Ioannis Hatzilygeroudis. 2019. Item Response Theory + Machine Learning: A Personalized Cold-Start Recommendation Framework for Adaptive Learning. In *International Conference on Artificial Intelligence in Education*. Springer, 261–273.
- [7] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2025. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *arXiv preprint arXiv:2402.07927* (2025). arXiv:2402.07927 [cs.AI] <https://arxiv.org/abs/2402.07927> Version 2, submitted on 16 Mar 2025.
- [8] Nischal Sapkota and Jack Bondurant. 2024. Assessing Concepts, Procedures, and Cognitive Demand of ChatGPT-generated Mathematical Tasks. *arXiv preprint arXiv:2404.05411* (2024).
- [9] Steffen Steinert, Karina E Avila, Stefan Ruzika, Jochen Kuhn, and Stefan Küchermann. 2024. Harnessing large language models to develop research-based learning assistants for formative feedback. *Smart Learning Environments* 11, 1 (2024), 1–15. doi:10.1186/s40561-024-00354-1
- [10] Jannatul Tithi, Matthew Rowe, and Stefan Harrer. 2025. The Promise and Limits of LLMs in Constructing Proofs and Hints for Intelligent Tutoring Systems. *arXiv preprint arXiv:2505.04736* (2025).
- [11] Virginia Tech News. 2024. Coding careers of the future: CS students team up with elementary schools. <https://news.vt.edu/articles/2024/02/coe-cs-coding-careers-of-the-future.html> Accessed: 2025-05-12.
- [12] Lev Semenovich Vygotsky. 1980. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA.
- [13] Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2024. Progressive-Hint Prompting Improves Reasoning in Large Language Models. *arXiv preprint arXiv:2304.09797* (2024). arXiv:2304.09797 [cs.CL] <https://arxiv.org/abs/2304.09797> Version 6, submitted on 7 Oct 2024.
- [14] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), Track on Datasets and Benchmarks*. arXiv:2306.05685 [cs.CL] <https://arxiv.org/abs/2306.05685> Version 4, submitted on 24 Dec 2023.
- [15] Xiaofei Zhou, Kaixin Li, Abdul Moid Munawar, and Zhen Bai. 2021. Scaffolding Design to Bridge the Gaps between Machine Learning and Scientific Discovery for K-12 STEM Education. In *Proceedings of the 2021 Interaction Design and Children Conference (IDC '21)*. ACM, Athens, Greece, 1–6. doi:10.1145/3459990.3465194 ACM, New York, NY, USA.