WIP: Expanding AI-Driven Concept Inventory Analysis in Computer Science Education

Abstract

Generative Artificial Intelligence (AI) offers transformative opportunities for education, especially in concept-based learning assessments. This paper presents an AI-powered tool designed to analyze student responses to concept inventory questions in entry-level computer science courses. By combining natural language processing (NLP) with traditional assessments, the tool provides personalized feedback, fosters metacognitive skills, and aggregates insights to guide instructional strategies. Although focused on computer science, the approach is scalable to other disciplines where concept assessments play a role in gauging and improving student comprehension.

Introduction

Concept inventories have been instrumental in assessing conceptual understanding in STEM education. The Force Concept Inventory (FCI), developed by Hestenes et al. in 1992, was among the pioneering tools that highlighted the significance of evaluating students' grasp of fundamental physics concepts[1]. Since then, numerous concept inventories have been created across various STEM disciplines to diagnose misconceptions and measure learning gains[2]. These tools have become integral in both educational research and classroom practice, providing valuable insights into students' conceptual frameworks. It is worth mentioning that, in educational assessment, a *concept inventory* is a standardized tool comprising validated multiple-choice questions aimed at diagnosing students' understanding of fundamental concepts and identifying common misconceptions within a specific domain[3]. Alternatively, a *concept assessment* refers to a broader range of evaluation methods, including open-ended questions, designed to gauge students' grasp of particular concepts, often tailored to specific instructional contexts. The term *concept inventory* is often preferred in research settings due to its standardization and reliability, facilitating comparisons across different studies and populations[3]. However, in classroom settings, *concept assessment* may be favored for its flexibility and adaptability to diverse teaching objectives.

These tools are traditionally designed as multiple-choice assessments aimed at diagnosing common misconceptions and evaluating students' grasp of fundamental concepts. Traditionally, these assessments are conducted with carefully crafted multiple choice questions. The format of the questions and the possible answers are designed specifically to capture misconceptions students may have on a given topic. More specifically, along with the correct choice, incorrect choices (known as 'distractors') are crafted such that students who have a common misconception would select these answers. This is in contrast to a more generic multiple choice assessment

where incorrect answers may be provided with less thoughtful consideration. A major advantage of traditional concept assessments is their scalability, as multiple-choice questions can be efficiently graded even in large-scale settings. Additionally, incorrect answers provide valuable insights for both students and instructors by highlighting common misconceptions. Nonetheless, the format of multiple choice assessments, even carefully crafted concept assessments limit the depth of insights into student reasoning and conceptual mastery. More specifically, multiple-choice assessments have notable limitations in accurately capturing a student's true understanding. One of the primary concerns is the potential for students to guess correct answers. For example, in a five-option multiple-choice question, a student has a one in five chance of guessing the correct answer without truly possessing mastery of the underlying concept[4].

In contrast, short-answer responses provide more robust insights into student comprehension. By requiring students to generate responses independently, short-answer formats reduce the likelihood of guessing and demonstrate a deeper engagement with the material[5]. This approach allows educators to evaluate not only the correctness of an answer but also the reasoning and thought process behind it, which is essential for assessing higher-order thinking skills[5]. Moreover, short-answer assessments enable instructors to identify a breadth of misconceptions and even uncover additional misconceptions that aren't captured in the finite multiple choices. This can lead to more effective feedback and targeted instructional interventions[6]. Overall, short-answer assessments serve as a more accurate tool for evaluating conceptual mastery, promoting critical thinking, and reducing the impact of guessing when compared to multiple-choice formats. A significant limitation of short answer responses is in the workload required to adequately grade these types of assessments. Generative AI and natural language processing (NLP) provide a unique opportunity to address these limitations. By integrating AI into concept inventory analysis, we can augment traditional assessments with richer insights while maintaining scalability.

This work introduces an AI-powered tool designed to analyze short-answer responses to concept inventory questions, identifying patterns in student reasoning and providing targeted feedback. The tool's implementation demonstrates its potential to enhance both individual learning and instructional strategies. Students benefit from personalized feedback and Socratic-style reflection questions, while instructors gain aggregated data on cohort-wide performance. Although this paper focuses on computer science education, the methodology is adaptable to other disciplines, such as statics, thermodynamics, chemistry, etc..

Specifically, in the realm of computer science education, a significant amount of work has been done on developing and validating concept inventories (CIs)[7]. These CIs are designed to identify specific misconceptions students hold about key concepts, and they have been developed for various topics, including digital logic, CS1, and data structures[8]. The development of these CIs often involves a Delphi process to determine important concepts, along with interviews and analysis of student misconceptions to create effective questions and distractors[9]. The goal is to create reliable measures of student understanding that can be used to inform curriculum and pedagogy[8,10]. A systematic review of computer science CIs identified 33 total CIs, 12 of which are currently validated[10]. Concept inventories are validated assessments that measure students' understanding and require rigorous development, with the Basic Data Structures Inventory (BDSI) as a prime example[11]. A literature review also uncovers challenges of CI development in computer science, including pre-test limitations and the need to consider various programming

languages[10]. Some researchers are exploring non-traditional methods for developing CIs[10] including Adaptive Tool-Driven Conception Generation (ATCG) which offers an alternative approach, using crowdsourcing with students and machine learning, demonstrating that student-generated assessments can be comparable to expert-designed ones[9]. The Foundational CS1 Assessment (FCS1) and the Second CS1 Assessment (SCS1) are language-independent assessments for CS1 knowledge[12]. It's noted that a key challenge is that some assessments, like the FCS1, were not originally referred to as concept inventories despite their nature[10]. In this work, we have focused on the SCS1 as a proving ground of the strategy due to its wide scale availability.

The use of large language models (LLMs), such as GPT-4[13], in education is another prominent theme, particularly in generating programming exercises[14]. Studies have found that LLMs can generate clear and relevant exercises, but these are often easier than requested and may lack depth in thematic personalization[14]. Additionally, LLMs are being explored for their ability to provide feedback on programming assignments, create code explanations, and generate educational resources[15]. However, there are concerns about potential biases in LLMs and the need for guardrails when using them in educational settings[14]. It is noted that benchmarking is an important aspect of evaluating LLMs in computing education[15]. In the field of chemistry education, there is an emphasis on assessing students' conceptual knowledge using tools like concept inventories[16]. The impact of AI tools, like ChatGPT, on chemistry education is being studied with an analysis of AI-generated laboratory reports across various chemistry courses[16]. It has been found that while ChatGPT can generate lab reports, these often have issues such as fabricating references, inconsistent results, and lacking the nuanced experimental details that are important in chemistry lab work[16].

While AI-based strategies do have limitations like hallucinations, Retrieval Augmented Generative (RAG) AI systems mitigate the issue of AI hallucination by using developer-supplied content and can aid in designing complex tasks and evaluating student reasoning[17]. Additionally, recent works show that despite the limitations of traditional multiple-choice tests AI can enhance them by analyzing student explanations[18]. AI-based analyst bots like Dewey, can quickly analyze student responses, providing valuable feedback to instructors to improve teaching, and can be used to adapt concept inventories by analyzing student textual responses[18].

Overall, the literature suggests that CIs are useful for identifying and addressing student misconceptions, and that LLMs have potential as tools to help in computing education, however they may require careful integration and consideration for how they can be used most effectively and ethically.

Methods

Our tool is composed of several interrelated components designed to create a seamless workflow from question creation to student analysis and feedback. The architecture of the tool is summarized in Figure 1.

At its foundation is a database that houses all the assessment items, including each question's text, multiple-choice options, and metadata such as correct answers, underlying concepts, and known misconceptions. This database is flexible enough to store additional attributes, such as difficulty

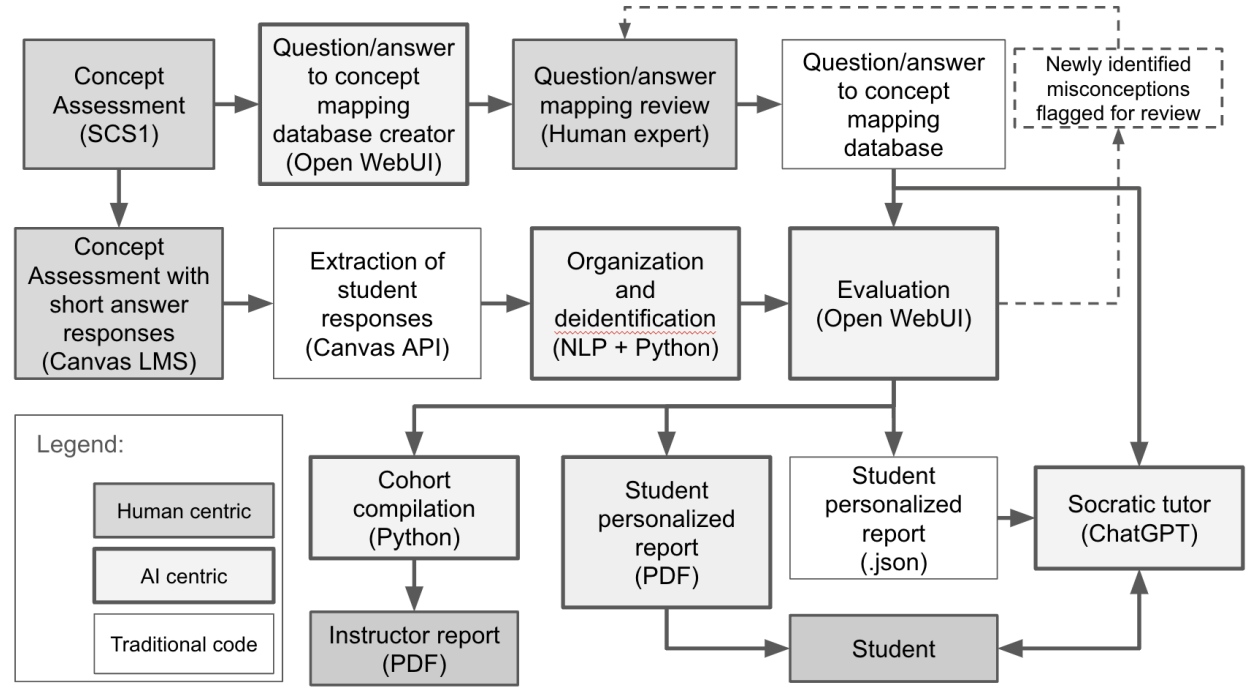

Figure 1: Overview of the tool architecture, illustrating the workflow from question creation to feedback generation.

levels or tags for specific content areas in the future. To build this database, we employed a custom Large Language Model (LLM) assistant tasked with mapping potential answers of concept questions to their corresponding concepts and misconceptions. Hosted on a secure, locally managed OpenWebUI[19] server running Llama3.2-8B[20], this approach ensured that proprietary data, such as the SCS1 concept assessment, was not exposed to third-party services like OpenAI or Google or Microsoft. The LLM served as an initial drafter of the database content, which was subsequently validated by human experts to ensure accuracy and quality prior to deployment. This hybrid approach highlights the balance between leveraging advanced AI tools and safeguarding proprietary educational resources. Additionally, this approach dramatically reduced the time necessary to create a mapping of question answers/distractors to the underlying core concepts.

Students interact with the tool through a streamlined, minimalistic quiz interface that prioritizes clarity and usability inside the Canvas Learning Management System[21]. In a traditional multiple-choice concept assessment quiz, each question is presented sequentially, and students are tasked with identifying the correct answer from the bank of possible answers. In addition to the correct answer, the other answers are aimed at common misconceptions, with the idea being that the concept inventory-style quiz allows the uncovering of common misunderstandings if a student chooses an incorrect answer. These are commonly referred to as “distractors”. We postulate this approach can be improved if students are required not only to select an answer but also to provide a brief written explanation of their reasoning. Furthermore, students can be asked to identify one incorrect answer and explain why it is incorrect. This dual explanation is beneficial for gaining deeper insights into students’ conceptual understandings and virtually eliminates the possibility of

students guessing correct answers without proficiency in the underlying concept(s). The combination of short-answer responses to multiple-choice concept inventory-style questions allows for a more representative evaluation of an individual's true understanding of a particular concept. However, short-answer responses are much more time-consuming to grade than simple multiple-choice tests, which can be graded with near-perfect accuracy by automated systems. Recent advances in practical usability and the overall quality of Large Language Models allow for a degree of automated evaluation of short-answer responses that has simply not been practical until now.

Once students submit their responses on the Canvas LMS, a custom Python script pulls the results using the Canvas API. These results are deidentified and organized for the next step in the process. Anonymized student submissions are then passed to the system's Nature Language Processing (NLP)-driven backend to process the data. Employing a locally hosted model, such as a Llama2-8B instance running via Ollama, the system first uses nature language processing to parse the student responses into the selected letter answer and the accompanying explanation. The system then compares each student's selected answer and written explanation against the database of correct concepts and known misconceptions for a single question. If the explanation aligns with the correct conceptual pathways indicated in the database, the system marks the response as fully correct. Conversely, if the explanation aligns with a known misconception, the system flags the response as incorrect or incomplete along with a brief synopsis of the apparent underlying misconcept(s). If a student's response appears to uncover a misconcept that is not contained in the database then the system will flag the response for human expert evaluation. This new misconcept can be inserted into and thus continuously improve the mapping database. This approach is akin to the class-sourcing approach of developing concept assessment questions[9]. The process continues for all questions in the set. The output of this automated grading script is an organized file identifying a student's apparent conceptual understandings and misunderstandings from the assessment. This output information is saved as a JSON file and used as input for three additional stages in the overall process.

One stage is to pass the student's organized quiz results to a script that generates guiding questions and other resources designed to encourage the student to improve their reasoning. These guiding questions emulate the Socratic method by prompting reflection and deeper engagement without directly providing the correct answer. This is provided to the student as a single PDF report file to help them improve on their own.

An additional stage uses the same output information to inform a custom GPT-based Socratic tutor. Students receive the JSON file containing their quiz results provided in a token-efficient format to load to a custom GPT available for free on the ChatGPT store. This GPT is pre-loaded with the expert-level concept mapping database and instructions to act as a Socratic-style tutor. The chatbot uses the individual's quiz results in concert with the expert database to help the student improve their understanding. For example, a student demonstrating proficiency in for-loops but struggling with while-loops might be prompted to explore the similarities and differences between the two concepts.

The third stage is to provide a single report to the class instructor. This report compiles the cohort's assessment results with a particular emphasis on common miconceptions.

Designed for scalability, this system supports large-scale course deployments while maintaining a personalized experience for each student. Ongoing improvements include the integration of audio transcription tools, such as Otter.ai, OpenAI Whisper API, or Eleven Labs, which will extend the tool's capabilities to include oral responses similar to interactive code reviews. These additions provide a richer, more interactive experience akin to oral examinations, blending scalable technology with individualized engagement[22].

Future plans

The next stages of this project aim to implement the proposed tool and conduct systematic studies in an attempt to evaluate its impact on student learning outcomes in first-year computing courses.

The study will begin with a cohort of students completing the SCS1 concept assessment with short answer responses to establish a baseline measurement of their conceptual understanding. This data will provide a reference point for evaluating subsequent and more advanced implementations.

Students will receive individualized PDF summaries generated by the tool's architecture, highlighting their conceptual strengths and recommended areas of further study based on their performance. Additionally, students will have access to the Socratic tutor chatbot designed to help them engage with concepts more deeply.

We will track the performance of students who engage with these tools using subsequent concept assessments and traditional evaluations in their first-year computing coursework. Our hypothesis is that students who use the tutor bot and/or review the personalized reports will demonstrate greater improvements in conceptual understanding compared to those who opt out of these resources. We acknowledge that there may be a self-selection bias among students who choose to engage with the tools, and we invite feedback from the engineering education community on strategies to address this issue.

Course instructors will receive compilation summaries that provide cohort-level insights into student performance and common misconceptions. We will solicit instructor feedback via surveys to assess whether these insights influenced their teaching approaches and if so, how.

To test the scalability and broader efficacy of the approach, a multi-course study will be conducted. All participating students across multiple first-year computing courses in different departments will take the SCS1 concept assessment. In selected courses, students will receive their personalized feedback and chatbot access, while others will serve as control groups. By comparing performance across these groups, we aim to isolate the effects of the feedback mechanisms on student learning outcomes.

In addition to measuring quantitative changes in student performance, we will collect qualitative feedback from students regarding the perceived usefulness of the concept assessments, personalized reports, and the chatbot. This dual approach will provide a richer understanding of how these tools impact learning and engagement.

The insights gained will contribute to refining the tool and addressing challenges such as

engagement bias and scalability. Through this strategy, we aim to assess the effectiveness of AI-driven personalized feedback in enhancing conceptual understanding, with the goal of demonstrating how generative AI can serve as a tool in fostering a more engaging and effective learning environment.

References

[1] David Hestenes, Malcolm Wells, and Gregg Swackhamer. Force concept inventory. *The Physics Teacher*, 30 (3):141–158, 1992. doi: 10.1119/1.2343497. URL https://doi.org/10.1119/1.2343497.

[2] Julie Libarkin. Concept inventories in higher education science, 2008. URL https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_072624.pdf.

[3] Michael W. Klymkowsky and Kathy Garvin-Doxas. *Concept Inventories: Design, Application, Uses, Limitations, and Next Steps*, pages 775–790. Springer International Publishing, Cham, 2020. ISBN 978-3-030-33600-4. doi: 10.1007/978-3-030-33600-4_48.

[4] Educational Resources Information Center (ERIC). Multiple-choice assessments and guessing: Limitations and considerations. *ERIC*, 2016. URL https://files.eric.ed.gov/fulltext/EJ1091824.pdf.

[5] Turnitin. What makes effective test questions and answers for assessments?, 2024. URL https://bit.ly/40kSlQY.

[6] Bootstrep. Understanding short-answer questions: Assessment benefits and drawbacks, 2024. URL https://www.bootstrep.org/understanding-short-answer-questions-assessment-benefits-drawbacks.

[7] C. Taylor, D. Zingaro, L. Porter, K.C. Webb, C.B. Lee, and M. Clancy. Computer science concept inventories: past and future. *Computer Science Education*, 24(4):253–276, 2014. doi: 10.1080/08993408.2014.970779.

[8] Kevin C. Webb, Daniel Zingaro, Soohyun Nam Liao, Cynthia Taylor, Cynthia Lee, Michael Clancy, and Leo Porter. Student performance on the bdsi for basic data structures. *ACM Transactions on Computing Education*, 22(1):1–34, 2021. doi: 10.1145/3470654.

[9] Sam Saarinen, Shriram Krishnamurthi, Kathi Fisler, and Preston Tunnell Wilson. Harnessing the wisdom of the classes: Classsourcing and machine learning for assessment instrument generation. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19)*, page 7. ACM, 2019. doi: 10.1145/3287324.3287504.

[10] Murtaza Ali, Sourojit Ghosh, Prerna Rao, Raveena Dhegaskar, Sophia Jawort, Alix Medler, Mengqi Shi, and Sayamindu Dasgupta. Taking stock of concept inventories in computing education: A systematic literature review. In *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1 (ICER '23 V1)*, page 19. ACM, 2023. doi: https://doi.org/10.1145/3568813.3600120.

[11] Leo Porter, Daniel Zingaro, Soohyun Nam Liao, Cynthia Taylor, Kevin C Webb, Cynthia Lee, and Michael Clancy. Bdsi: A validated concept inventory for basic data structures. In *International Computing Education Research Conference (ICER '19)*, pages 1–9. ACM, 2019.

[12] Miranda C. Parker, Mark Guzdial, and Allison Elliott Tew. Uses, revisions, and the future of validated assessments in computing education: A case study of the fcs1 and scs1. In *Proceedings of the 17th ACM Conference on International Computing Education Research (ICER 2021)*, page 9. ACM, 2021. doi: https://doi.org/10.1145/3446871.3469744.

[13] OpenAI. Chatgpt. https://openai.com/chatgpt.

[14] Evanfiya Logacheva, Arto Hellas, James Prather, Sami Sarsa, and Juho Leinonen. Evaluating contextually personalized programming exercises created with generative ai. In *ACM Conference on International Computing Education Research V.1 (ICER '24 Vol. 1)*, page 19. ACM, 2024. doi: https://doi.org/10.1145/3632620.3671103.

[15] Murtaza Ali, Prerna Rao, Yifan Mai, and Benjamin Xie. Using benchmarking infrastructure to evaluate llm performance on cs concept inventories: Challenges, opportunities, and critiques. In *ACM Conference on International Computing Education Research V.1 (ICER '24 Vol. 1)*, page 17. ACM, 2024. doi: https://doi.org/10.1145/3632620.3671097.

[16] Joseph K. West, Jeanne L. Franz, Sara M. Hein, Hannah R. Leverentz-Culp, Jonathon F. Mauser, Emily F. Ruff, and Jennifer M. Zemke. An analysis of ai-generated laboratory reports across the chemistry curriculum and student perceptions of chatgpt. *J. Chem. Educ.*, 100:4351–4359, 2023. doi: https://doi.org/10.1021/acs.jchemed.3c00581.

[17] Melanie M Cooper and Michael W Klymkowsky. Let's not squander the affordances of llms for the sake of expedience: Using retrieval augmented generative ai chatbots to support and evaluate student reasoning. *J. Chem. Ed. in press*, 2024. In press.

[18] Michael W Klymkowsky and Melanie M Cooper. The end of multiple choice tests: using ai to enhance assessment. *arXiv preprint arXiv:2406.07481*, 2024. https://doi.org/10.48550/arXiv.2406.07481.

[19] OpenWebUI Contributors. Openwebui: A user interface for openai and other models. https://openwebui.com/.

[20] Meta AI. Llama 3.2-8b. https://ai.meta.com/llama, 2024.

[21] Inc. Instructure. Canvas learning management system. https://www.instructure.com/canvas.

[22] Lee Dong-Kyu. A gpt-based code review system for programming language learning, 2024. URL https://arxiv.org/abs/2407.04722.