

Score Report Conversational Agent to Guide Educators in Interpreting Assessment Results

Brian D. Gane, Dante I. Cisterna, & Amy K. Clark

Accessible Teaching, Learning, and Assessments Systems (ATLAS), University of Kansas

Abstract

We developed and evaluated a prototype conversational agent for teachers that can support their understanding and use of mastery-based score reports that describe a students' academic performance. The agent prototype delivered grounded responses reflecting intended interpretation and uses, however some responses were inaccurate, signaling aspects that can be improved.

1 Background

Score reporting has increasingly been recognized as a critical component of quality assessments, rather than merely the final step in communicating results (Hambleton & Zenisky, 2013; Zenisky et al., 2025). Research has shown that teachers, parents, and other stakeholders often struggle to understand the meaning of scores, performance levels, subscores, and graphical displays, limiting the extent to which assessment information can inform decision making (e.g., Kannan et al., 2018; O'Donnell & Sireci, 2021).

Artificial intelligence and large language models (LLMs) have the potential to support teachers' understanding of data about their students' learning and to use that information to make pedagogical decisions. Teachers make numerous decisions to optimize student learning outcomes. While assessment score reports provide essential data on student learning, a crucial challenge remains: How can high-level achievement information be presented to connect with the many context-specific instructional decisions teachers make daily. Users' interpretation of score report information is cognitively complex, but effective interpretation means teachers can use the results

and is critical evidence in a validity argument (Zapata-Rivera, 2018).

2 Aims

Prior research indicates teachers desire more support in understanding and using score reports (e.g., Clark et al., 2022). An LLM-based conversational agent (CA) offers the potential to help teachers better understand and use student results in assessment score reports. We aim to develop a CA that teachers can converse with to receive support in accurately interpreting an individual student report and using it to inform instructional planning and decision-making, and to facilitate communication with parents, guardians, or other teachers about students' performance. In this paper we report on our progress in developing and evaluating a prototype of a CA that could be deployed to educators to assist them in using and interpreting score reports.

Our specific research context is the Dynamic Learning Maps® (DLM) alternate assessment system. The DLM system is based on cognitive learning models, called learning maps, that represent how students, including students with the most significant cognitive disabilities, develop understanding across academic domains (Kingston et al., 2016). These learning maps describe learning as a network of interconnected nodes, defining knowledge, skills, and understandings. Contrasted with learning progressions, these maps capture the variability and non-linear nature of student learning (Swinburne Romine et al., 2025). Learning maps guide assessment design and provide a structure for reporting individual skill mastery data on the score report (see Figure 1).

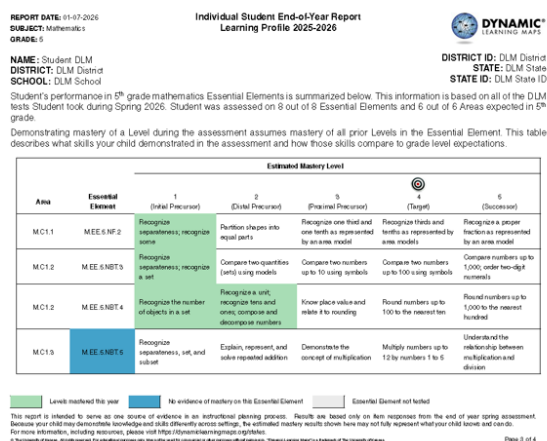


Figure 1: One page of an example score report for a hypothetical student. Shading in the cells in each row indicates the students' skill mastery along the five levels of each content standard/Essential Element.

Individual student reports for the DLM assessment provide a wealth of performance data that are not commonly reported in other systems. The DLM system uses diagnostic classification modeling which allows for reporting at the level of individual skill mastery. Individual skill mastery is also aggregated at two additional levels: conceptual area and overall (i.e., across all conceptual areas). While these three aggregation levels provide rich data that are useful for understanding a student's strengths and areas for growth, they may be less familiar to educators that more commonly receive assessment results that present a unitary score with confidence intervals and norm-based comparisons. Multiple data points are a strength of DLM reports but need to be packaged with appropriate supports that scaffold teachers' interpretations (Brown et al., 2023). Further, teachers may need additional information beyond a student's data when they move from interpreting to using results. For instance, using score report results to inform instructional decision making and planning requires coordinated use of multiple sources of information (e.g., pacing guide, assessment blueprints, learning maps, standards, student data points), which can tax teacher bandwidth.

The CA is designed to support educators in understanding and integrating these data and information sources in time efficient and accurate ways. Three research questions guided our development and evaluation of the CA prototype:

- What successes can we find in the CA's guidance?
- What limitations do we identify in the CA's guidance?
- How can we consistently and repeatedly evaluate the CA's performance as we develop new versions of the CA?

We created two primary use cases to guide CA design and evaluation. One use case was designed to understand whether giving the CA access to a specific student's score report would change its responses and/or conversation, and to explore the security constraints of the CA accessing student score reports.

- Use Case 1: User asks general questions about a DLM score report and questions related to a specific student's results, but does NOT upload that student's score report to the CA.
- Use Case 2: User uploads a DLM score report and asks questions related to that student's results.

We used both cases to guide our conversations with the CA during development and as a basis for our evaluation.

3 Methods

3.1 Agent Development

We developed the CA in Microsoft Copilot Studio, using the Claude Opus 4 model. Agent instructions reflected our intended uses and expectations for CA responses. We also identified unsupported uses and actions that the CA should not attempt to complete even if requested by the users. Through internal testing we refined the instructions to cover gaps and reinforce key guidance. The instructions articulated the agent's role, stance, tone, and purpose. They included safeguards, directions for responsible use, and quality assurance self-checks. The instructions also outlined the two use cases and described important features to include in responses, including length and depth. Finally, the instructions grounded the agent in a DLM-specific knowledge base: DLM-produced documents, web pages from the [DLM website](#), and the entire DLM [professional learning website](#). The DLM-produced documents included assessment blueprints, lists of

standards (“Essential Elements”, in DLM nomenclature), example score reports, performance level descriptors, instructional resources for teachers, materials for parents about score reports, and transcripts of professional development training modules. Additionally, we included one document that was specific guidance and directions to the agent about DLM terminology, DLM content structures, score report structure, and pedagogical guidance (e.g., “Recommendations for IEP [Individualized Education Program] development should promote comprehensive instruction in the subject rather than isolated instruction on specific skills or linkage levels”). We added this document after internal testing highlighted several shortcomings in CA-generated responses.

3.2 Evaluation framework

Adapting and extending literature on CAs in health care (Denecke, 2023), we identified evaluative criteria that were relevant for the prototype score report CA. For each evaluative criterion we selected from the Denecke (2023) framework, we adapted the criteria definition for score report (not health care) purposes. Table 1 lists the criteria, our modified definitions, and the source of evidence most directly relevant to the criteria. Additionally, we identified multiple indicators for each criteria. As an example, content accuracy indicators include correctness of information, completeness of answers, consistency with expected response, appropriate instructional guidance, and adequate use of report evidence. We did not create separate measures for each indicator, but the set of indicators informed the coding scheme and review forms we used to evaluate the CA (see sections 4.1 and 4.4).

3.3 Test Cases

We developed multiple test cases for each use case. Each test cases included a prompt for the CA, whether a score report file should be uploaded along with the prompt, and the expected response from the CA. The expected response was written as either an exemplar response that the CA should provide, or as a list of the features/aspects of the expected response that the CA should provide. Test cases covered a range of categories that reflected our intended uses for the CA. Table 2 provides the number of test cases for each category. The specific test cases were written by the three authors and

were motivated by the overarching use case, the intended features/capabilities of the CA, and/or specific challenges to interpretation and use of the DLM mastery-based score reports based on prior research (Clark et al., 2022; Gane et al., 2025).

Test cases that had a score report file identified were all from Use Case 2. Mock score reports for hypothetical students were used to evaluate those test cases.

Criterion	Definition	Source of Evidence
Ease of use	Extent to which a person believes that using a particular CA would be worthwhile	Subject matter expert feedback
Flexibility in dialogue handling	Ability to maintain a conversation and deal with users’ generic questions or responses	Subject matter expert feedback
Content accuracy	Extent to which CA responses are consistent with DLM documentation, expert expectations, & knowledge base	CA responses
Security	Extent to which the CA appropriately protects sensitive information, follows privacy expectations, and responds safely when handling student-level assessment information and educational decision-making requests	CA responses

Table 1: Evaluation criteria and sources of evidence

3.4 Test case response evaluation and subject matter expert agent evaluation

We conducted an internal review of the CA responses to the test cases. To complete these reviews, we prompted the CA with each test case

and saved its response. Later, two raters (BG and DC) coded each response in terms of two ratings: Accuracy (Accurate, Inaccurate, or Misleading response) and Discrepancy (Major, Minor, or Hallucination). These two metrics are not independent; the discrepancy codes are only used for Inaccurate responses. Raters coded responses independently. Due to time constraints, we did not complete any pre-training to help the raters align their coding criteria and have not yet had raters discuss cases that had disagreements in order to reach consensus.

At the time of this writing, we collected feedback from two staff members with subject matter expertise in English language arts and/or mathematics content knowledge, program implementation, and score report delivery and intended uses of results. These individuals were instructed to engage with and evaluate the agent. They were provided with an example hypothetical student's score report to upload if they chose. Following interactions with the CA, they completed an evaluation rating form that included ratings and feedback on the agent's response accuracy and usefulness for individual prompts, as well as overall feedback on their experience with the CA as a whole and its handling of student privacy for the hypothetical student. They also completed the modified System Usability Scale (SUS; Brooke, 1996; Bangor et al., 2008; Lewis, 2018), which we scored using the established procedures. Reviews from additional subject matter experts are underway.

Category	Test Case Count
Interpreting a score report or section (Generic, not a specific student)	7
Interpreting a score report or section (For a specific student)*	7
Understanding a particular term, phrase, or language in the report	6
Parent/guardian communication*	3
Preparing an IEP	3
Planning instruction*	3
*These topics included test cases designed to test privacy (i.e., no PII) in the CA response	

Table 2: Test case categories and counts

4 Results

We report preliminary analysis of the CA prototype's performance and effectiveness. We describe the results of our internal evaluation of the CA responses and provide two annotated examples of CA responses, following by results from the subject matter expert review.

4.1 Test case responses evaluation

Because we did not resolve discrepancies to yield consensus codes, we report several analyses to describe our findings. We first report the pattern of accuracy codes for each rater individually and then consider both raters' codes together, focusing on the trends where both raters agreed.

Both raters classified about half of the responses to the test cases as accurate ($N_{BG} = 15$; $N_{DC} = 16$). The remaining responses were classified as inaccurate ($N_{BG} = 13$; $N_{DC} = 8$) or misleading ($N_{BG} = 1$; $N_{DC} = 5$). Those responses categorized as inaccurate were more often considered minor ($N_{BG} = 10$; $N_{DC} = 6$) than major ($N_{BG} = 3$; $N_{DC} = 2$). Although we had a discrepancy category for hallucinations, neither rater used it; none of the 29 CA responses showed evidence of hallucinations. In summary, both raters seemed to indicate about half of the CA responses were accurate and half inaccurate (mostly minor) or misleading.

As we noted in the Methods section, our internal evaluation of the CA responses to the test cases was completed without pre-training and alignment for raters. Nevertheless, we computed Gwet's AC_1 as a measure of interrater agreement; AC_1 values range from -1 to 1 (Gwet, 2008). Although agreement was lower than ideal, raters still somewhat agreed ($AC_1 = .47$, $p < .01$, 95% CI [.20, .74]).

Overall, the raters agreed on the accuracy code for 18 of 29 cases (62%). Of the 18 cases in which both raters agreed, 10 were classified as accurate, 7 as inaccurate, and 1 as misleading. Of the 7 responses that raters agreed were inaccurate, raters agreed that 4 were minor inaccuracies but disagreed on whether the remaining 3 responses reflected a major or minor inaccuracy. Four of those 7 inaccurate responses exhibited a CA score report read error in which the CA extracted an incorrect mastery value from the skill mastery matrix (see Figure 1). This was a known bug and therefore raters tended to categorize it as minor, not major. The remaining 3 inaccuracies reflected CA responses that were incomplete or unclear (e.g.,

describing some, but not all, of the ways a teacher can use a score report). For the 11 cases in which raters did not agree, 4 cases were rated as accurate by one rater and misleading by the other. Five cases were rated as accurate by one rater and minor inaccuracies by the other. Two cases were rated as accurate by one rater and major inaccuracies by the other.

4.2 Accurate response example

In this test case, the user asks about comparing results across students and uploads one specific student's score report. The full response is nearly 4,200 characters so we summarize it and include illustrative quotes. **Prompt:** *"How can I compare this report results, for the student 'Juno', to other students?" + DLM score report for Juno (uploaded as a PDF).* **Response:** The CA response first notes that reports are designed to "describe individual student learning relative to grade-level Essential Elements, not to rank or compare students to each other" but then affirms that looking across multiple students' reports can be valid and useful "to inform grouping, planning and shared instructional focus". It lists three ways to compare reports meaningfully, using the different categories of performance metrics on the report and what insights comparisons along each metric could yield. It then describes three options for how the teacher can actually compare: (1) Create a simple cross-student grid that allows one to see patterns in students' skill mastery for each content standard, (2) group students that have common/shared next-step targets, noting "This shifts focus from 'who's higher' to 'who's ready for similar instruction.'", or (3) look for complimentary profiles because "students with different strengths can support each other". The response also describes four "cautions when comparing" that focus on intended uses of the score reports and reminders that "different students may have been assessed on different skills" (the DLM assessment model means not all students complete the same assessments) and that the score report data is "a snapshot, not the full picture" of what the student knows and can do. The response concludes with reflection prompts for the teacher including, "What's driving the interest in comparison—grouping decisions, planning shared instruction, or something else? Knowing your goal will help me give you the most useful guidance."

The entire CA response is consistent with the intended uses and interpretations of the score report

and consistently uses framing and language that reflect the strength-based approach, combined with high expectations for students with significant cognitive disabilities SWSCD, that are emblematic of the DLM program.

4.3 Inaccurate (minor) response example

In this test case, the user asks about where to begin instruction for a specific student in the new/upcoming academic year but does not upload any score report. The CA response is considerably shorter (1,440 characters) than the last example. The most notable response feature is that it provides "strategy-level starting points" for a teacher with a returning student in the new year (note: this scenario is not uncommon for teachers of students with significant cognitive disabilities, since they may teach across multiple grade levels).

Prompt: *I'm teaching the same student again next year. What are some recommended starting places for instruction when we begin the school year?*

Response: The CA response begins by noting that it could give more specific guidance if a score report for the student were uploaded or key details for the student (e.g., grade level, subject of focus, etc.) are provided. It then provides general starting points that it can suggest without further data: "(1) Revisit evidence of mastery from the prior year to identify skills to maintain, generalize, and build upon. (2) Look for patterns across Essential Elements—are there content strands where the student showed consistent strength or emerging skills? (3) Consider linkage level trends—this can suggest whether to focus on deepening current levels or stretching toward the next. (4) Plan early formative checks to confirm what has been retained over the summer before moving forward."

Both raters flagged this response as minor inaccuracies because the response provided left out a key detail that was included in our expected response. Specifically, it omitted a concrete suggestion that the user could use results from a specific section of the report (data on mastery of conceptual areas) as the thread to connect across years. The relevant portion of the expected response stated, "Start with the Area bar charts. Areas are consistent across grades, while the specific Essential Elements change by grade, so understanding which Areas the student showed more mastery in (and where the student may have had less mastery) can help support instructional planning in the next year." While this exact

wording was not necessary, it should have pointed to the bar charts explicitly and used the “Areas” label/name to eliminate ambiguity and direct the user to the appropriate section.

4.4 Subject matter experts’ evaluation

The two reviewers rated the CA positively and indicated it demonstrated strong usability. Across prompt-based evaluations, both reviewers consistently provided the highest ratings for accuracy and usefulness, with most ratings receiving 4 out of 4. Reviewers highlighted the agent’s ability to accurately interpret score reports, explain mastery and performance results, provide actionable instructional guidance, and handle out-of-scope questions appropriately. Reviewers noted that the agent effectively responded to and redirected tangential questions back to meaningful use of assessment results and helped educators translate score report information into next instructional steps. The System Usability Scale responses were also favorable, with overall scores of 95 and 97.5, placing the agent in the highest range of usability (Bangor et al., 2008).

Suggested areas for refinement included reducing response length in some situations, limiting repetition of information already displayed in score reports, clearly communicating the scope and limitations of the tool, and strengthening privacy protections by suppressing student-identifying information when reports are uploaded, which was inconsistently applied in reviewer interactions. Overall, both reviewers viewed the agent as highly valuable and identified relatively minor enhancements to improve an already strong user experience.

5 Conclusions

AI has the potential to impact multiple aspects of assessment design, delivery, and reporting. As score reports can provide actionable and interactive tools for teachers that support their learning needs and context (Zapata-Rivera, 2018), ensuring that users understand how to interpret score reports and have access to guidance and support remain critical for effective and valid use. Doing so can help ensure assessment results are informative for educators and support data-driven decision making.

Our development and evaluation of the DLM CA prototype demonstrate early successes from which to build upon. We note three positive signals.

First, evaluations by the two subject matter experts found the prototype useful, accurate, and valuable. Second, our internal review of CA responses to test case queries found no hallucinations and many accurate responses. Inaccurate responses were often minor or misleading (rather than major inaccuracies), reflecting incomplete guidance or portions that might be misinterpreted by users. Third, the tone and content of the CA responses reflected foundational DLM programmatic principles and intended uses.

One critical factor in the accuracy of CA responses was likely the extensive knowledge base which provided the CA access to 55 DLM-produced documents that were labelled and indexed to ground the CA responses. This seemed effective in preventing hallucinations and major inaccuracies while also connecting score report data to relevant instructional decisions that were consistent with the underlying cognitive models (i.e., DLM learning maps).

The evaluation also surfaced several inaccuracies and sub-optimal responses. First, a persistent challenge emerged despite engineering efforts to solve it—the CA sometimes made errors when “reading” data in an uploaded score report PDF. The source of this error was in making visual judgments of the color shading of cells in the skill mastery matrix (see Figure 1). Second, the SME review indicated that occasionally the CA did not preserve student privacy, for instance, repeating a student’s name in the response when the original query included the name. Several of the test cases that internal reviewers used intentionally included student details to test privacy safeguards (see Table 2). These test cases did not produce responses that broke the safeguards but several of the subject matter experts’ conversations did. This suggests the privacy safeguards need to be reviewed and strengthened to ensure the responses follow established criteria for safety and privacy.

5.1 Limitations

While this development and evaluation yielded important findings, several limiting factors constrain the strength of the inferences we can draw. Our evaluation had small sample sizes in terms of test cases and subject matter experts. We developed more test cases than we were able to evaluate, and we did not conduct rater pre-training to align raters. The subject matter expert review is ongoing, but at the time of writing we only had two

complete reviews. As the prototype is developed further, additional testing and evaluation will be required.

5.2 Future directions

We noted two critical issues that need to be resolved before moving from a prototype to an operationally-ready system: (1) privacy and safety guardrails need to be strengthened and applied consistently and (2) visual reading of the graphical data displays in DLM score reports are challenging for the LLM (even with significant engineering efforts to build a custom report reader agent) and additional engineering efforts are needed to ensure the agent captures correct data.

Beyond the specific engineering efforts to improve the CA, we plan to expand our evaluation approach and sources of evidence, using the Denecke framework as a starting point and adapting for a score report agent. In tandem, we will use the same test cases with future versions of the prototype to evaluate whether changes to instructions improved responses and to conduct regression testing.

User studies are an important next step and can be conducted with the prototype to understand whether a score report CA provides meaningful value to teachers and is easy to use and useful. Including score report users throughout the design cycle is critical to ensure the system is responsive to teachers' needs (Van der Kleij et al., 2014; Zapata-Rivera & Katz, 2014).

We expect that other researchers and assessment programs are evaluating the potential for conversational agents to help stakeholders understand score reports. Establishing common evaluative criteria will help the field progress and understand critical features of CAs that help teachers make sense of assessment data and use it to inform their instruction. CAs built from a strong assessment system foundation will have a head start. CAs will likely be most impactful when they are built on top of an assessment system with evidence for its validity and with CA guidance that is consistent with the intended interpretations and uses of the assessment results.

References

- Aaron Bangor, Philip T. Kortum, and James T. Miller. 2008. An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24(6), 574–594. <https://doi.org/10.1080/10447310802205776>
- John Brooke. 1996. SUS: A “quick and dirty” usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). Taylor & Francis Inc.
- Gavin T. L. Brown, Priya Kannan, Sandip Sinharay, Diego Zapata-Rivera, and April L. Zenisky. 2023. Challenges and opportunities in score reporting: A panel of personal perspectives. *Frontiers in Education*, 8:1211580. <https://doi.org/10.3389/feduc.2023.1211580>
- Amy K. Clark, Brooke Nash, and Meagan Karvonen. 2022. Teacher assessment literacy: Implications for diagnostic assessment systems. *Applied Measurement in Education*, 35(1), 17–32. <https://doi.org/10.1080/08957347.2022.2034823>
- Kerstin Denecke. 2023. Framework for guiding the development of high-quality conversational agents in healthcare. *Healthcare*, 11(8): 1061.
- Brian D. Gane, Amy K. Clark, and Elizabeth Kavitsky. 2025. Parent interpretation and use of diagnostic mastery-based score reports. *Applied Measurement in Education*, 38 (3–4), 363–381. <http://doi.org/10.1080/08957347.2026.2630044>
- Kilem L. Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61, 29–48. <https://doi.org/10.1348/000711006X126600>
- Priya Kannan, Diego Zapata-Rivera, and Emily A. Leibowitz. 2018. Interpretation of score reports by diverse subgroups of parents. *Educational Assessment*, 23(3), 173–194. <https://doi.org/10.1080/10627197.2018.1477584>
- Neal M. Kingston, Meagan K. Karvonen, Sue Bechard, and Karen Erickson. 2016. The philosophical underpinnings and key features of the Dynamic Learning Maps alternate assessment. *Teachers College Record*, 118(14), 1–30.
- James R. Lewis. 2018. The system usability scale: Past, present, and future. *International Journal of Human-Computer Interaction*, 34(7), 577–590. <https://doi.org/10.1080/10447318.2018.1455307>
- Francis O'Donnell and Stephen G. Sireci. 2021. Language matters: Teacher and parent perceptions of achievement labels from educational tests. *Educational Assessment*, 27(1), 1–26. <https://doi.org/10.1080/10627197.2021.2016388>
- Russell Swinburne Romine, Jonathan Schuster, Meagan Karvonen, W. Jake Thompson, Karen Erickson, Vanessa Simmering, and Sue Bechard. 2025. Learning maps as cognitive models for instruction and assessment. *Education Sciences*, 15(3), 365.

- Fabienne M. Van der Kleij, Theo J. H. M. Eggen, and Ronald J. H. Engelen. 2014. Towards valid score reports in the computer program LOVS: A redesign study. *Studies in Educational Evaluation*, 43, 24–39. <https://doi.org/10.1016/j.stueduc.2014.04.004>
- Diego Zapata-Rivera. 2018. *Score Reporting Research and Applications* (1st ed.). Routledge.
- Diego Zapata-Rivera and Irvin R. Katz. 2014. Keeping your audience in mind: Applying audience analysis to the design of interactive score reports, *Assessment in Education: Principles, Policy & Practice*, 21(4), 442-463. <http://dx.doi.org/10.1080/0969594X.2014.936357>