Demonstrating the Potential of LLM AI to Provide High-Quality Narrative Feedback to Students and Parents/Caregivers on Writing Quality Aligned with LDC-SCALE Analytic Student Rubrics

Today's students and parents/caregivers unfortunately do not often receive actionable information on what students should do to improve deeper learning performance as demonstrated by the quality of their writing in response to complex disciplinary text. In practice, the sheer volume of providing narrative feedback on student disciplinary writing is often too large for teachers to provide each student with frequent or adequate feedback on each round of their writing, much less provide accurate and useful feedback and guidance to parents/caregivers that enables at home support regardless of their home language, socio-economic status, or other demographics. This feedback, in the form of unbiased, curriculum-embedded, formative assessment, has been found to drive more scalable student improvement than any other resource.[1] Yet, the scalable ability to provide frequent, expert, accurate assessments of the quality of writing to all students and parents/caregivers, along with specific feedback on what is needed to improve deeper learning performance and how to achieve these improvements, does not currently exist in the commercial or nonprofit marketplace.

LDC seeks to analyze the feasibility of using a Large Language Model (LLM) to generate analytic rubric-based scores and narrative feedback on deeper student writing in response to complex text that (1) are at least as accurate as teacher-provided feedback, (2) teachers confirm to be relevant, high-quality, and useful for their students, and (3) parents/caregivers find useful to help them understand the strengths and weaknesses in their student's writing as well as how they can actionably support their student in improving their writing skills.

This feasibility study is a critical and foundational step in LDC's plans to partner with SRI International to build a Generative Artificial Intelligence (AI) engine for assessing deeper learning in student work. The LDC-SRI partnership seeks to utilize SRI's machine learning (ML) and natural language processing (NLP) tools combined with LLM artificial intelligence to build such an engine. This report describes the feasibility study and highlights key learnings regarding the potential of AI (as tested out-of-the-box, without SRI enhancements) to score and provide feedback on student writing, as well as implications and next steps for future work.

Context

Students in high-need schools—and their parents/caregivers—too rarely receive analytic, detailed, and actionable feedback about their students' authentic disciplinary writing (Science, Social Studies, and ELA). One reason for this is that many teachers, particularly those in LDC's target urban and rural communities working with high-need students and high student-to-teacher ratios, struggle to keep up with the demands to score, grade, and provide accurate and frequent deeper learning feedback to what is often 100 or more students. In addition, accurately calibrating teachers to LDC's national analytic student work rubric, designed in collaboration with the

[1] Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Routledge.

Stanford Center for Assessment Learning and Equity (SCALE)[2], is manual and quite time intensive. This means overburdened teachers, with even less time available to undergo the calibration process, cannot fully leverage the power of the SCALE student rubric for their own and their students' benefit.

In practice, particularly for teachers serving 100+ students at a time, by the time students finish their argumentative or informational/explanatory disciplinary writing products, the teacher has missed multiple opportunities to address skill deficits because they were unable to generate, compile, and organize the formative data revealing student learning difficulties within the reading and writing processes. Likewise, students have also lost multiple opportunities to take ownership of their learning, build on formative feedback from the teacher, and practice the reading and writing skills essential for 21st-century learning. All students require and deserve this type of feedback, giving them access to deeper learning opportunities for writing reflection and growth.

Research confirms that providing students with formative feedback is one of the most powerful instructional interventions for improving student learning outcomes.[3] In ranking various instructional interventions according to their effect sizes, Hattie found that providing students with formative feedback had a .90 effect on student learning outcomes, second only to effective teachers—which has not proven scalable.[4] Thus, solutions are needed that equip educators with formative assessment resources in their classrooms in a way that increases teacher capacity to provide high-quality feedback frequently and consistently.

At present, the scalable ability to provide frequent expert, accurate, nationally-calibrated feedback and guidance to all students on where/what they need to learn next to improve deeper learning performance does not currently exist. An LDC-SRI AI tool would solve this problem. Recent studies have begun to demonstrate the effectiveness of AI to improve student rubric scoring and feedback, and ongoing work to develop next generation AI tools has shown new approaches, similar to what LDC plans in its partnership with SCALE and SRI, have excelled in the most recent machine learning competitions looking at analysis of student writing.[5] When built, the LDC-SRI AI mechanism will provide students and their parents/caregivers with actionable information tied to LDC's standards-driven rubrics. The parent- and student-friendly feedback will focus on how students can improve reading and writing comprehension using the SCALE rubrics, i.e. analytic rubrics that specify the skills and subskills students need to master to progress and accelerate their learning.

---

[2] SCALE teacher certification is used in 10 states.

[3] Wiliam, D. (2011). *Embedded Formative Assessment* (p. 36). National Educational Service.

[4] Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Routledge.

[5] https://www.kaggle.com/competitions/feedback-prize-2021/overview; Measuring Reading Comprehension Is Hard. Can AI and Adaptive Tools Help?, A. Klein (Edweek March 2023); Kelbadov and Madnani (2020). Automated Evaluation of Writing – 50 Years and Counting. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7796–7810.

As a first step, LDC tested the following main hypothesis through this feasibility study: If an LLM is prompted to generate student- and parent-friendly feedback tied to LDC student work rubrics in response to student writing, will teachers and parents/caregivers find the feedback to be, in general, at least as useful as teacher-authored feedback (while also being a far faster and easier tech-instantaneous way for teachers to regularly and consistently generate that feedback).

Research Questions

The study addressed the following questions:

1. To what extent can out-of-the-box LLM AI (without additional SRI ML/NLP modulation of LLM output) currently be used to accurately score student work using LDC-SCALE analytic rubrics?
2. To what extent can out-of-the-box LLM AI currently generate useful narrative feedback to students that is tied to LDC-SCALE analytic rubrics?
3. Are parents/caregivers currently being adequately supported to understand and support their students' literacy skills?
4. To what degree can out-of-the-box LLM AI currently generate information about student writing that parents/caregivers find to be (a) of similar value as information generated by teachers, and (b) useful in helping them understand their student's performance and how they can help their student improve?

Methods

To answer these questions, LDC produced scores and associated feedback on student work samples from both teachers and out-of-the-box LLM AI, then engaged teachers and parents/caregivers in reviewing both sets of scores and providing feedback on their relative accuracy, clarity, and utility. Specifically, LDC carried out the following steps:

1. LDC identified a set of sample schools to obtain student work samples and engage teachers and parents/caregivers. The sample included 9 high-need schools, 6 urban and 3 rural located in New York, Michigan, and Wisconsin, with which LDC had no prior engagement. The student population at sample schools averaged 71% free- and reduced-price lunch, 76% non-white, and 15% English language learners. LDC intentionally selected high-need schools given students in these schools and their parents/caregivers often do not receive sufficient or actionable feedback about their writing quality, yet yearn for it acutely. Teachers in these schools usually also have a greater need for technology-enabled supports that can help mitigate their student support workload. The intention was that any solution presented by an LLM AI be designed at a minimum to serve this student population well, therefore initial testing started in the population the grant most wanted to serve.

2. LDC then compiled 35 samples of student work across grades 6-8 written in response to LDC, SCALE-validated performance tasks. Examples of tasks include *6th Grade ELA Literary Analysis: "Raymond's Run"*[6] and *8th Grade ELA Literary Analysis: "The Necklace."*[7] These tasks all holistically meet the "Exemplary" curriculum alignment criteria outlined by SCALE's *LDC Curriculum Alignment Rubric*[8] as well as in that rubric's specific categories of content and literacy skills, clarity and coherence, texts, and writing product. These tasks included standards-aligned LDC-SCALE validated student work rubrics that can be used to assess and provide feedback on the disciplinary writing produced by students.

3. LDC enlisted 14 teachers from sample schools to consensus-score student work samples.[9] Teachers were provided with the LDC-SCALE student work rubric and asked to independently score pieces of student work using the rubric as well as draft student-friendly feedback and parent-friendly feedback on the student writing sample. Teachers then came together in a total of 7 pairs to discuss their assessment and create consensus scores and feedback for each paper based on their conversations. Each LDC-SCALE rubric contains 6 scoring elements, meaning teacher-pairs engaged in 210 independent analytic scoring events across the full set of student work samples. Notably, teachers were not familiar with the rubric prior to scoring and were not trained or calibrated scorers.

4. While teachers scored student work samples, LDC simultaneously prompted ChatGPT to score these same samples, providing a score on the LDC-SCALE rubric, student-friendly feedback, and parent-friendly feedback on the student work. The same 210 independent analytic scoring events were completed by ChatGPT.

5. LDC compared the numeric scores provided by teacher-pairs and ChatGPT to assess how similar the two score sources were, as well as the relative directional differences between the scores where they existed. LDC also conducted focus groups with participating teachers to get their assessment of the ChatGPT scoring and feedback.

6. Finally, LDC recruited a sample of 13 parents/caregivers from participating schools to provide feedback on the relative clarity and utility of the feedback provided by both teachers and ChatGPT.[10] Parents/caregivers were first given a survey asking them to share their experiences with the frequency, quality, and usefulness of the feedback they currently receive from teachers about their student's writing. LDC then conducted focus groups where parents/caregivers were shown the parent-friendly feedback on student work samples from both teachers and ChatGPT, without being told which feedback

[6] https://s.ldc.org/u/emmvxfflvwcqj7xbje4opsgwa
[7] https://s.ldc.org/u/8v8gcjjek7c4s6sibfk3jgfd7
[8] https://coretools.ldc.org/resources/a85ed86c-9fae-475e-a433-f2e7cbaf2d99
[9] LDC's National Writing Project (NWP) regional partners made the opportunity available to ELA/Social Studies teachers at the middle school level. Teachers volunteered to participate. Teachers were given writing samples to assess from students in the same grade level and discipline as they teach.
[10] Participating teachers circulated a volunteer opportunity for parents within their schools.

originated from which source. Parents/caregivers were asked to share their opinions about the strengths and weaknesses of each set of feedback, how understandable the feedback was, and what else they need in order for this type of feedback to be most helpful and actionable in supporting their student's writing.

Results

Question 1: To what extent can out-of-the-box LLM AI (without additional SRI ML/NLP modulation of LLM output) currently be used to accurately score student work using LDC analytic rubrics?

Looking across the 210 scoring events, LDC found that on average the ChatGPT and teacher scores differed by 0.62 points on any given scoring element. Taking into account the direction of the scoring difference, ChatGPT was more likely to score student work slightly higher on the rubric than teachers. When considering the specific element of the rubric, LDC did not discern any clear patterns in the magnitude or direction of the scoring difference, meaning ChatGPT was not consistently more or less accurate, as compared to teachers, at scoring certain types of rubric elements.

Several factors contribute to an understanding of the relative effectiveness of ChatGPT in scoring student work accurately against LDC's rubrics. First, researchers at the Stanford Center for Assessment, Learning, and Equity (SCALE) who have partnered with LDC in creating and validating the student work rubrics advise that a teacher score difference of less than or equal to 0.5, as compared to master-calibrated teachers' consensus scores, demonstrates essentially effective calibration. Reaching this standard of a consistent 0.5 or less difference is indicative of a teacher having received sufficient training and practice on how to score using the rubric to be deemed able to produce accurate and reliable scores.

While the absolute difference on average in scores between ChatGPT and the sample teachers in this study is slightly higher than the desired 0.5 or less, this result is promising. Notably, the teachers participating in the study were not trained nor considered calibrated on the rubric, meaning we would expect there to be a higher rate of inaccuracy in their scores as compared to calibrated teachers who would be using the LDC rubrics in practice. Further, ChatGPT was used "out of the box," meaning for the purposes of this study LDC was not able to provide a "supervisory layer" by using SRI International's ML/NLP tools. The addition of ML/NLP has been shown to improve LLM accuracy and reduce bias (racial, ethnic, socioeconomic, etc.).[11] Therefore, neither set of scores in this study benefited from the kind of training required to improve scoring accuracy, and that would be done with future iterations of an LDC-SRI AI tool. Given the difference of 0.62

[11] AI is coming to schools, and if we're not careful, so will its biases. A. Perry, Brookings Institution (2020). Lancaster, D., (2023, April) Ethical Considerations and Addressing Biases in ChatGPT-like AI Solutions, LinkedIn. https://www.linkedin.com/pulse/ethical-considerations-addressing-biases-chatgpt-like-dean-lancaster/.

found with “untrained” scorers, we are optimistic that adding these supervisory layers in the future would produce LLM AI-based scores that fall within the desired <=.05 range.

Additionally, it is important to note that the score difference seen here may not necessarily be interpreted as an inaccuracy of ChatGPT scoring since we cannot claim the teacher scores to be accurate given their lack of calibration. From this test we can only speak to the relative difference in scores between these two untrained sources. However, even so, the value of the difference taken in context of the SCALE-determined marker of calibration leads towards great optimism about the potential for accurate scores from future iterations of LLM AI.

Question 2: To what extent can out-of-the-box LLM AI currently generate useful narrative feedback to students that is tied to LDC analytic rubrics?

For AI-generated feedback to be a viable solution for both increasing the amount of useful information parents/caregivers and students receive about student writing, and freeing up time required of teachers to regularly and consistently provide such feedback, it must be seen in the eyes of teachers as accurate, high-quality, and aligned to definitions of student performance. To gather information on teacher views about AI-generated scores and narrative feedback on student writing aligned with LDC’s rubrics, LDC had teacher-pairs score student work samples and create student- and parent-facing feedback. Teachers then reviewed the AI-generated scores and feedback on the same writing and compared to their own.

Overall, teachers found the quality of AI-generated scores and feedback to be sometimes “hit or miss.” In some cases they found it to be accurate, meaning similar to the scores and feedback teachers had created with their partner, and actionable for students. One teacher noted after reviewing the AI feedback, “I didn’t find myself changing a whole lot, and if I did it was just adding.” At the same time, the AI scores did not always match or nearly match the teacher-generated scores. In some cases teachers saw patterns in score discrepancies, noting for example that AI tended to score a bit higher. In other cases, it wasn’t clear what was driving the differences in scores.

Focusing on the narrative feedback provided by AI, teachers similarly reported mixed views, sometimes praising it and other times noting where it lacked important context or might not be understandable, or relatable, for students. Several teachers gave examples of AI feedback being too “heady” or drawing on “literature theory,” in many cases using language that might not be appropriate for the student grade level. One teacher remarked, “Some of it was specific and helpful, and then other times I’m like, ‘this is too much.’” Another teacher, when describing a question posed by ChatGPT for what a student might consider in revising their work, said, “That’s a senior level sociology class question, not a seventh grade read a short story and respond question.” Note, this reflection by teachers may be in part a result of some teachers not fully understanding grade level rigor expectations and/or lowering expectations to meet student skill levels, rather than an indicator of AI having too rigorous expectations. In LDC’s experience training

teachers in manual student work calibration, involving more than 20,000 student writing samples scored manually by teachers, teachers' understandings of grade level rigor expectations were often far below expected standards. Addressing the propensity for students to receive scoring or feedback that is targeted towards lower-than-expected rigor levels is a particular challenge this project seeks to address in supporting teacher (and parent/student) learning expectations.

In reflecting on the value of the AI-generated feedback, teachers honed in on the need for both objectivity and subjectivity in grading, especially when grading writing. They again provided contrasting views, liking the extent to which AI brought more objectivity, and often uniformity, to grading, while at the same time lamenting the loss of subjectivity. One teacher captured this duality when saying, "Being objective when you're grading, we're trying to achieve that when we're using rubrics to level the playing field and we have specific things we want to look at, the same way, for every student. But it's very difficult to be objective, because we have to take into consideration where they've come from and, for each individual, how they started as a writer. The subjectivity can be a good thing coming from a teacher standpoint because we know them as humans, and AI is pulling that piece out."

From an objectivity standpoint, teachers praised AI for its ability to consistently apply a set of standards and address "grading exhaustion." AI was able to provide fresh terms, and new ways of providing feedback, yet ultimately do so in a way that provided desired consistency. Teachers noted how difficult it is to bring this level of objectivity and fresh eyes to the 50th or 100th essay they are grading.

From a subjectivity standpoint, teachers noted the importance of context and differentiation when grading writing, in particular related to a student's growth and approach to writing. In their feedback they regularly take into account information such as how far a student has progressed and what specific improvements can be seen when compared to prior writing. One teacher made the point that students are not uniform, therefore they do not need uniform feedback. Instead, they need differentiated feedback that is aware of and responsive to both student and school culture and context. Several teachers remarked that the current version of AI-generated feedback was a great starting point, acknowledging that a middle ground between uniformity and differentiation is ultimately where writing assessment should be. One teacher summed up this sentiment by saying, "Somewhere in between AI and me is the real score."

Interestingly, the teachers engaged in our pilot described being fascinated by the feedback produced by ChatGPT, raising several questions about the future of AI, how much the public really knows about and understands the origin of and use of ChatGPT and similar tools, and what the use of AI could mean for the education system in coming years. Focus group members understood the potential of AI to get smarter in the future, openly wondering if the capacity existed, or soon would, to make it improve in the areas they identified as weaknesses in the test scores and feedback. And, if AI could serve as a support to them in the future—grounds for future Cambiar funded research.

In fact, during our focus groups teachers described how ChatGPT feedback had led to their own learning, an interesting byproduct they were not expecting. For example, one teacher described how reviewing the AI feedback helped him see themes across student papers that, had the feedback been for his individual students, would have sparked ideas about additional instruction he would have provided. Another teacher explained how his investigation into why his score differed from ChatGPT's led to reflection and learning on his part. "When I looked at the feedback that ChatGPT gave and the rationale for it, I could understand why and then it had me going back and looking at the paper and then agreeing with ChatGPT." In this vein, teachers wondered about other ways AI-generated feedback can be useful to them in providing differentiated support, for example by helping them more readily create a plan for individualized interventions for students. This finding in particular dovetails with LDC's broader desire to use an LLM and analytic rubric scores to then connect teachers, students, and parents/caregivers with additional instruction (targeted disciplinary literacy lessons in LDC's online CoreTools platform) to enable students to both address next proximal learning and also build on existing student asset strengths. With this next generation of LDC's LLM AI, instead of a teacher's "great job," AI can connect a student to progressively more advanced literacy learning.

Question 3: Are parents/caregivers currently being adequately supported to understand and support their students' literacy skills?

To inform broader questions about the potential value of LLM AI-generated feedback for both teachers and parents/caregivers, LDC first sought to understand the extent to which parents/caregivers are currently receiving information about the quality of their students' writing, as well as whether they would like more opportunities to receive this feedback. Survey results confirmed hypotheses that parents/caregivers in high-need schools are not receiving sufficient amounts of this type of data and feedback and would like to receive more.

	% of Parents/Caregivers Who Agree or Strongly Agree
I regularly see examples of the writing my student does for school assignments.	23%
I regularly see data indicating how close my student's writing meets expectations of academic standards.	38%
I regularly see teacher feedback provided to my student regarding how they can improve their writing to meet academic standards.	31%
I would like to more frequently see examples of the writing my student does for school assignments.	69%

I would like to more frequently see data indicating how close my student's writing meets expectations of academic standards.	77%
I would like to more frequently see teacher feedback provided to my student regarding how they can improve their writing to meet academic standards.	62%
If I were more frequently given access to my student's writing, data indicating how close they are to meeting academic standards, and teacher feedback regarding how they can improve their writing, I would spend more time with my student to help them improve their writing.	69%

Less than a quarter of parents/caregivers surveyed reported they regularly see examples of their student's school-based writing, and only roughly one-third regularly see teacher feedback regarding how students can improve their writing (31%) or other data regarding whether their student's writing meets academic standards (38%). Yet, parents/caregivers were clear in their desire to have greater access to all of these artifacts related to student writing quality: 69% would like to more frequently see samples of their student's writing; 62% would like more opportunities to receive teacher feedback on how their student's writing can be improved, and 77% would like more frequent access to data on how their student's writing aligns with academic standards. The survey was administered to parents/caregivers prior to showing them examples of both teacher- and AI-generated feedback on student writing. After seeing these examples, parents/caregivers confirmed they do not regularly receive feedback of this kind, with 72% reporting they never get this type of detailed feedback and 26% saying they sometimes get this feedback.

Importantly, parents/caregivers indicated that if they had more frequent access to student writing as well as feedback about its quality and how it can be improved, they would utilize this information proactively with their students. Seven in ten parents/caregivers reported they would spend more time with their students helping them to improve their writing if they had access to this feedback.

While representing a very small sample of parents/caregivers, these results nonetheless provide support for the hypothesis that should LLM AI tools be created that are able to generate accurate and parent/caregiver-friendly feedback on student writing quality, while aiding teachers' ability to provide more frequent feedback, parents/caregivers in high-need school would both value having access to this data and would use it to further support their student's academic growth and writing skills.

Question 4: To what degree can out-of-the-box LLM AI currently generate information about student writing that parents/caregivers find to be (a) of similar value as information generated by teachers, and (b) useful in helping them understand their student's performance and how they can help their student improve?

Parents/caregivers were asked to review parent-friendly feedback on student writing created by both teachers and AI, without knowing the source of each piece of feedback. They were then asked to reflect on the strengths and weaknesses of each type of feedback and share which they preferred, and why, with a focus on the extent to which the feedback would support their efforts to help students improve their writing. Across the sample, *73% of parents/caregivers indicated the feedback generated by AI was better overall in helping them understand the strengths and challenges of the student writing sample they reviewed, and in helping them understand how they could support the student to improve their writing.*

Overall, parents/caregivers had largely similar perceptions of the value and qualities of the teacher and AI-generated feedback. For both types, parents/caregivers generally found the feedback to be helpful, well organized, and easy to understand. While many parents/caregivers noted both types of feedback provided specific suggestions for ways students could improve the writing sample, others noted a need for more concrete guidance and examples in both forms of feedback. In both cases, parents/caregivers praised the feedback for providing positive reflections on student writing as well as pointing to areas of needed improvement, both for setting a positive tone and for helping parents/caregivers understand the strengths of student writing in addition to the weaknesses. The consensus among parents/caregivers was that the AI-generated feedback was similar in value, and met the study threshold of "at least as useful" as teacher-generated feedback. One parent/caregiver summed this up by saying, "I think both versions provided enough feedback to be able to improve the writing."

When prompted to describe the differences between the two types of feedback, parents/caregivers noted the teacher-generated feedback was often more vague, and broad, providing slightly less specific or actionable examples of how students could improve the writing. One parent/caregiver described the teacher-created feedback by saying, "This feedback did not have the specific questions for deeper understanding of the text and how to improve it the way [the AI-generated] version did." In contrast, parents/caregivers noted the AI-generated feedback was somewhat more technical, and potentially less accessible for parents/caregivers and students, than the teacher-generated feedback. Similar to teacher views, some parents/caregivers wondered if the AI-generated feedback was not fully aligned with the grade level of the students in terms of the questions posed and suggestions for improvement.

Most notably, several parents/caregivers described a strength of the AI-generated feedback, as compared to the teacher-generated feedback, as the way the AI feedback provided concrete ways for parents/caregivers to engage with students to help drive improvements to their writing. One parent/caregiver noted, "It was great at giving the parent and student ways to fix the issues of the paper. I also like the way that they set up suggestions for the parent to help engage in conversation with their student." Other parents/caregivers described liking the discussion prompts provided by the AI-generated feedback and "the specific questions to ask and guidelines for how to support the students with deeper feedback." The desire for this type of guidance was named by parents/caregivers as a quality of the type of feedback they would like to receive and that would

ultimately be most helpful to them in supporting improvement to their student's writing, along with details about not only the writing prompt but the expectations for the assignment, and feedback that is directly connected to and provided along with the student writing sample, similar to comments written in the margin of a document.

Implications & Conclusion

Parents/caregivers in high-need schools are not regularly receiving detailed, actionable feedback about the quality of their student's writing and how to help support students in making improvements. They are eager for this feedback, and report, should they receive it, they would readily use it to help their student continue to improve their writing. Yet, teachers in these schools have higher student loads and fewer resources, making it harder to find the time to consistently provide this type of deeper learning feedback on student writing.

This study, while small scale, demonstrated the potential of AI to provide high quality, rubric-aligned, actionable narrative feedback to both students and parents/caregivers, serving to not only equip both with the information needed to support ongoing student learning and improvement, but also provide teachers with a critical time-saving resource that can bolster their other efforts to drive student learning.

Out-of-the-box AI scores using the LDC-SCALE student writing rubric were similar, on average, to scores of untrained teachers. While scoring accuracy would need to be improved to meet LDC and SCALE thresholds for accuracy (e.g. calibrated), LDC is confident their next generation AI can do this. The key to this would be combining the LLM out of the box with SRI's proprietary ML/NLP algorithms, building on SRI's current use of their ML/NLP in conjunction with an LLM (roughly 50/50) in their multi-million dollar Departments of Defense and Homeland Security contracts. LDC is confident that with additional investment this tool could be built and tested, and in so doing garner high levels of educator confidence in a combined solution (a "walled garden" or supervisory layer that improves ChatGPT's good-but-not-always-great output).

Perhaps more importantly, looking beyond numerical rubric scores, teachers and parents/caregivers found the AI-generated feedback to be similar in quality to teacher-generated feedback. While both the AI- and teacher-generated feedback were found to have varying strengths and weaknesses, the AI feedback was deemed at least as useful for parents/caregivers and students in understanding the quality of student writing as feedback created by teachers. In fact, the AI feedback was preferred by three-quarters of the parents/caregivers in this sample for helping them understand the strengths and challenges of the student writing sample and in helping them understand how they could support the student to improve their writing. In particular, parents/caregivers noted a significant strength of the AI-generated feedback, as compared to teacher-generated feedback, was the extent to which it provided prompts and other concrete ways for how parents/caregivers can engage with students and use the feedback to help support improvements to their writing. This finding, which speaks to the promise of AI in

supporting improved deeper learning performance, can be readily extended to other applications beyond providing narrative feedback. Notably, LDC's plans for a next generation of smarter AI, designed specifically to fully leverage LDC's suite of tools and resources, will link analytic rubric scores to teacher curricular resources, connecting students (and parents/caregivers) to progressively more advanced learning opportunities.

These results from an untrained AI tool speak to the great potential for the quality and utility of feedback that could be generated through an LDC-SRI partnership that utilizes SRI's machine learning and natural language processing tools, combined with a large language model artificial intelligence, to build an engine specifically designed for these purposes.

With additional funding to build this engine, LDC can address the questions and challenges raised by the teachers and parents/caregivers in this study about the current form of AI feedback, ensuring the next generation of this tool addresses their feedback to the extent possible. Areas for exploration include: whether AI can be built to track, and take into account, student writing over time, e.g. how far a student has progressed, and what specific improvements can be seen when compared to prior writing; ensuring language used in feedback is accessible for parents/caregivers and students; ensuring feedback is well aligned with student grade level expectations; ensuring feedback to students and teachers mitigates socioeconomic and multilingual potential biases, and maximizing the extent to which concrete guidance for improvement, directed towards both students and parents/caregivers, is provided, and then connected to additional literacy curricular resources to continue to build on student strength and/or address their current but addressable literacy challenges.