ANCHORING CONCEPTS INFLUENCE ESSAY CONCEPTUAL STRUCTURE AND TEST PERFORMANCE

Roy B. Clariana[1] and Ryan Solnosky[2]
[1]*College of Education, The Pennsylvania State University*
University Park, PA USA
[2]*College of Engineering, The Pennsylvania State University*
University Park, PA USA

ABSTRACT

This quasi-experimental study seeks to improve the conceptual quality of summary essays by comparing two conditions, essay prompts with or without a list of 13 broad concepts, the concepts were selected across a continuum of the 100 most frequent words in the lesson materials. It is anticipated that only the most central concepts will be used as “anchors” when writing. Participants (n = 90) in an Architectural Engineering undergraduate course read the assigned lesson textbook chapter and attended lectures and labs, then in a final lab session were asked to write a 300-word summary of the lesson content. Data consists of the essays converted to networks and the end-of-unit multiple choice test. Compared to the expert network benchmark, the essay networks of those receiving the broad concepts in the writing prompt were not significantly different from those who did not receive these concepts. However those receiving the broad concepts were significantly more like peer essay networks (mental model convergence) and like the networks of the two PowerPoint lectures but neither were like the textbook chapter. Further, those receiving the broad concepts performed significantly better on the end-of-unit test than those not receiving the concepts. Term frequency analysis of the essays indicates as expected that the most network-central concepts had a greater frequency in essays, the other terms frequencies were remarkably the same for both the terms and no terms groups, suggesting a similar underlying conceptual mental model of this lesson content. To further explore the influence of anchoring concepts in summary writing prompts, essays were generated with the same two summary writing prompts using OpenAI (ChatGPT) and Google Bard, plus a new prompt that used the 13 most central concepts from the expert’s network. The quality of the essay networks for both AI systems were equivalent to the students' essay networks for the broad concepts and for the no concept treatments. However, the AI essays derived with the 13 most central concepts were significantly better (more like the expert network) than the students and AI essays derived with broad concepts or no concepts treatments. In addition, Bard and OpenAI used several of the same concepts at a higher frequency than the students suggesting that the two AI systems have more similar knowledge graphs of this content. In sum, adding 13 broad conceptual terms to a summary writing prompt improved both structural and declarative knowledge outcomes, but adding 13 most central concepts may be even better. More research is needed to understand how including concepts and other terms in a writing prompt influences students’ essay conceptual structure and subsequent test performance.

KEYWORDS

Summary Writing, Writing to Learn, Automatic Essay Assessment, Google Bard, OpenAI (ChatGPT)

1. INTRODUCTION

Writing-to-learn, especially summary writing, is a powerful way for students to recall and then organize (or reorganize) their understanding while building conceptual knowledge structure (Eryilmaz, 2002; Finkenstaedt-Quinn et al., 2021; Moon et al., 2018). Writing is a learner-centered strategy that intimately aligns with conceptual learning (Bereiter & Scardamalia, 1987; Sampson & Walker, 2012). Writing helps students to improve and refine their thinking about complex phenomena (Bereiter & Scardamalia, 1987; Hidi & Anderson, 1986) and summary writing helps students to grasp concepts in a related fashion rather than as discrete sets of ideas (Gaskins & Guthrie, 1994; Glynn & Muth, 1994; Guthrie et al., 2004). Writing about domain normative topics helps students to understand common disciplinary conceptions and to participate in discursive learning communities (Wallace, 2004). Specifically important here, Mason and Boscolo (2004) have identified writing as a way to foster conceptual change, especially for correcting misconceptions, by encouraging students to develop more elaborated explanations of complex phenomena (Halim et al., 2018; Moon et al., 2018). And writing becomes even more effective with formative feedback and reflection (Bangert-Drowns et al., 2004).

This investigation is designed to inform the development and use of the browser-based writing-to-learn tool called Graphical Interface of Knowledge Structure (GIKS) that provides immediate structural feedback as a network of concepts (Trumpower and Sarwar, 2010). This investigation specifically considers whether including a list of broad concepts from the lesson materials in the essay prompt could improve essay quality. This approach is consistent with prompt directedness from the concept map literature (Ruiz-Primo, 2001). Including a list of concepts in a writing prompt does not require much extra effort and could have at least three benefits, specifically concepts: (1) provide added information about the scope of the summary, (2) can provide "anchors" for basing the essay in a conceptual space (Cavallo et al., 2003), and (3) will help the writer with spelling those concepts. This last benefit is especially important for automatic essay scoring because misspelled words may not be noticed by the scoring software.

How might a list of concepts added to a writing prompt "anchor" the essays produced? Adding a list of associated concepts should enhance conceptual space disambiguation As a macro-level or gross-level example of this, we modeled this using OpenAI (https://platform.openai.com/playground) with these three questions:

(1) <u>What is a bank</u>? (no concepts) – *A bank is a financial institution that offers various banking services such as savings accounts, checking accounts, loans, investments, and more. Banks provide access to money and services to help customers manage their finances*.

(2) <u>What is a bank? try to use these concepts: cash, receipt, teller, ATM</u> – *A bank is a financial institution that provides services such as accepting deposits, making loans, and issuing credit cards. Bank customers can make deposits of <u>cash</u> at the <u>teller</u>, or use an <u>ATM</u> to deposit or withdraw money. When making a deposit, customers receive a <u>receipt</u> to show their transaction was completed,* and

(3) <u>What is a bank? try to use these concepts: flooding, dock, boat, fishing</u> – *A bank is an area of shallow water along the edge of a river, lake, or ocean where <u>boats</u> can <u>dock</u> or <u>fishing</u> can be done. It is also an area of land that is higher than the surrounding area and prevents <u>flooding</u>*.

The response to the first question (no list of concepts) was about banks as financial institutions and this was reiterated when four banking concepts were included in the prompt. But the third essay switched to "river bank" due to the four added concepts. The concepts anchored the replies to specific conceptual spaces.

Do extra concepts in a prompt matter in students' mental models? Cavallo et al. (2003) in a study with ninth-grade chemistry students (n = 60) asked before the lesson, "*write a summary of everything you know about Chemical Reactions*." Half of the students in addition received "*In your summary, include an explanation of how Chemical Reactions may be related with the following terms: atoms, compounds,* and *chemical change*." Misunderstandings in the essay were measured by raters, students who received these three extra concepts in the prompt had *twice as many misunderstandings* (21 vs. 11) and these misunderstandings persisted across the lesson as measured using the same essay prompt at Posttest 1 (20 vs. 11) and at Posttest 2 (15 vs. 14). These three concepts mattered, including concepts in the essay prompt before instruction had a powerful negative effect on setting a mental model of the content that persisted but evolved across the lesson.

Wang, Solnosky, and Clariana (2023) compared essays that used different lists of concepts in the writing prompt. The concepts were derived from an expert network map of the lesson content, referred to as <u>focus</u> concepts that were the 14 central high degree concepts in the expert network or <u>full</u> concepts that provided all 26 concepts in the network (e.g., central and peripheral). Participants (n = 68) in an undergraduate Architecture Engineering course completed a 2-week lesson module on *Building with Timber and Wood*, and then wrote summary essays using GIKS. Essays were converted to networks using the ALA-Reader approach (Clariana, 2010). Word frequency descriptive analysis of the central and peripheral concepts in the essays showed an interesting pattern: (1) The word frequencies were exceptionally consistent for the full and focus groups, it is *implied* here that the students' knowledge structure conceptual models on average held similar central and peripheral concepts. (2) It was anticipated that the Focus group would show higher word frequencies for the central (Focus) concepts since that is the list they received in the prompt, but this did NOT happen. Among the 14 central concepts, only the five most central concepts showed a higher frequency across the essays for the Full compared to the Focus condition (see Figure 1).

This data suggests that when writers are provided with a broader list of content concepts (26 in this case), without having ever seen the expert network, they are able to prioritize and use the *most central concepts* in the list when summarizing, which implies that their mental models (conceptual networks) also have these concepts as central concepts. This outcome aligns with the OpenAI essays above that a list of lesson concepts added to a writing prompt bounds a k-dimensional conceptual space when writing.

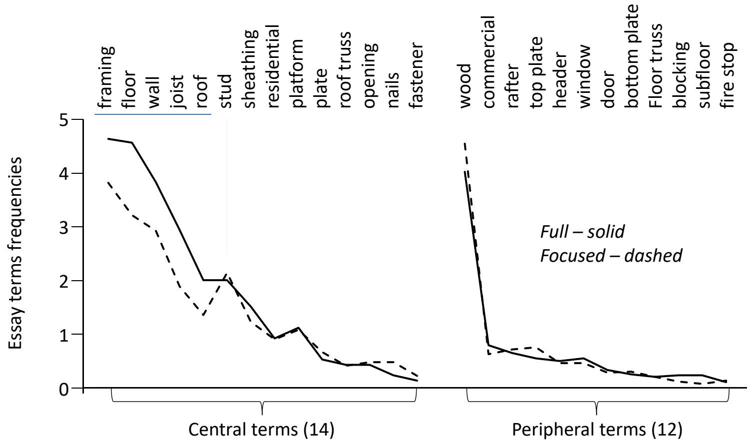

Figure 1. Essay word frequencies of the Central and Peripheral network concepts from Wang et al. (2023)

Because of the anticipated benefits and the likely influence on essays (sometimes perhaps negative) of including a list of concept in a summary writing prompt, it is critical for our ongoing research and development of GIKS to determine whether to include a list of concepts or not in the prompt, and if yes, which concepts and how many. Because the most central concepts in the list were mainly affected when the list of concepts is broader (Wang et al, 2023), to explore this we created a list of concepts that spans the lesson space including highly central, central, peripheral, and highly peripheral intending to replicate these highly central concept frequency findings from Wang et al. (2023).

In addition, essays were generated using Google Bard and also OpenAI (e.g., ChatGPT) using the same writing prompt and list of concepts as those given to the students to further explore this knowledge structure conceptualization. This modelling approach seems reasonable since both AI systems operate from large well-structured knowledge graphs of language artifacts that "represent a network of real-world entities—i.e. objects, events, situations, or concepts—and illustrates the relationship between them" (IBM, 2023) that aligns well with the view of students mental models as knowledge structure.

2. PARTICIPANTS, MATERIALS, AND RESULTS

2.1 Participants' Essay and End-of-Unit Test Data

Participants in this quasi-experimental investigation are undergraduate students (N = 110, 24% female) in the course *AE 221 Building Documentation and Modeling* in the Fall of 2022. In weeks 12 and 13 of a 16 weeks-long course, as regularly assigned tasks in the course, students completed a two weeks-long lesson on *Building with Steel* that included lectures and lab supported by textbook readings. At the end of the lesson students completed a writing task (described below) and a week later the end-of-unit test partitioned as two subtests, items from this lesson and items from other lessons covered in the unit before and after this lesson.

Students completed the summary writing task using a word processor during lab time. Students could choose to attend lab on either Tuesday, Wednesday, or Thursday, so the number of students each day varied. For logistics reasons, students in lab on Tuesday and Wednesday received the "Concepts" essay prompt (final sample n = 52) while those on Thursday received the "No Concepts" prompt (final sample n = 38). The prompt stated, *Reflect on the current lessons on structural steel construction and then write a 300-word summary of the most important issues. Please use this title for your summary (copy and paste into your summary): Structural steel construction: Important issues for the Architectural Engineer to consider.* In addition the concepts' group prompt added, *Consider including these 13 terms in your summary: composite, deck, concrete, fire proofing, non-composite, girder, stud, column, span, spacing, infill beam, bay, height*

These 13 concepts were purposefully selected from a list of the 100 most frequent words found in the lesson materials (the textbook chapter and the two PowerPoint lectures) as a sample of highly central, central, peripheral, and highly peripheral concepts in the lesson. Here are the broad concepts arranged in order of frequency along with the rank order: highly central: *concrete* (rank 2), *fire proofing* (6), *span* (7); central: *deck* (44), *girder* (47), *composite* (49), *column* (50); peripheral: *spacing* (60), *studs* (66), *non-composite* (67); and highly peripheral: *infill beam* (100), *bay* (>100), *height* (>100).

For essay scoring purposes, the course instructor was given the frequency list of 100 terms and was asked to generate an expert network map of the same lesson content using any terms. The final expert network contained 26 concepts, but only four high frequency concepts were included in the list of 13 broad concepts, *concrete, fire proofing, span, and deck*. Thus, the instructor's network did not align well with the lesson materials word frequency data.

The data for analysis consists of essay network similarity measures (as common link percent), end-of-unit multiple-choice test performance, and essay descriptive data (i.e., word frequency). The end-of-unit multiple-choice test was portioned into two subtests that covered several different lessons included in that course module. The test consisted of 40 items drawn randomly from an item database of 56 items, about half of the items covered the *Building with Steel* lesson and the other half covered material from the other lessons (cranes, dozers, heavy equipment, cadcam, BEM, MEP). The Cronbach alpha reliability of the 40 item test is .61, the two subtests were only moderately related, r = .47.

Due to the unequal sample sizes, the non-parametric Kruskal–Wallis test by ranks (one-way ANOVA on ranks) was used to analyze the essay network similarity data and the end-of-unit test data. Students' essays and the course materials were converted to Pathfinder networks using the ALA-Reader approach of Clariana (2010) using 35 concepts (i.e., 26 expert network + 9 more list concepts). The students' essay networks similarity to five different referent networks were compared for the List and No List groups (see Table 1).

Table 1. Kruskal–Wallis findings for each measure

	Students' essay network similarity (as % common links)					End-of-Unit Subtests	
	to expert	to peers	Chp. 11	PP #1	PP #2	Lesson	Other
Kruskal-Wallis H	0.160	5.987	1.452	5.498	8.940	4.687	0.430
df	1	1	1	1	1	1	1
Asymp. Sig. (p =)	0.689	**0.014**	0.228	**0.019**	**0.003**	**0.030**	0.512
No List (mean rank)	44.21	37.62	41.62	37.95	35.87	38.53	43.42
List (mean rank)	46.44	51.26	48.34	51.02	52.54	50.60	47.02

Bonferroni correction applied

There was no difference (p = .689) between the List and No List groups on essay network similarity to the Expert network (a measure of essay quality). However the broad concepts group essay networks were more like peers' networks than were those of the no concepts group (p = .014; e.g., showing convergence of mental models in the concepts group). In addition, the broad concepts group essay networks were more like the two PowerPoint lecture networks (p = .019 & .003) relative to the no concepts group, but there was no difference between receiving broad concepts or not for similarity to the textbook chapter network (p = .228). And finally, the broad concepts group outperformed the no concepts group on the end-of-unit subtest that aligned with the lesson content (p = .030), but not on the subtest that covered the other lessons in the module (p = .43).

2.2 Comparing Word Frequencies of Student and AI Essays

Student essay concept frequencies align with the findings above from Wang et al. (2023) that providing a list of broad concepts in the prompt increases concept frequency of only the central concepts (i.e., seven most central concepts in the expert network) but not for the other concepts. This increased frequency difference carried over to two non-list concepts, *floor* and *roofing*, that were not in the list but that incidentally are highly central in the instructor's expert network (see Figure 2).

To further explore the influence of providing lists of concepts in the writing prompt, forty AI essays were generated, OpenAI playgound (i.e., ChatGPT, https://platform.openai.com/playground), text-davinci-003, temperature = .7) and Google Bard (https://bard.google.com/, based on Language Model for Dialogue Applications, LaMDA). Half of the essays are based on the list of broad concepts prompt used above and half without the concepts.

Average word frequencies for the AI essays were calculated for the list concepts and expert concepts, AI essays word frequencies were only moderately like the students' essays, the average word frequency of the two

AI systems shows that both used the 13 broad concepts more frequently in the essays (see solid lines above dashed lines in the right side of Figure 3). Also, although the two AI systems are distinctly different from each other, there is considerable similarity between the two for nearly half of the concepts, especially the terms that are also the high frequency concepts in the instructor's expert network (see the peaks especially on the right side of Figure 3). We imply from this that both AI systems have a similar knowledge graph of this content.

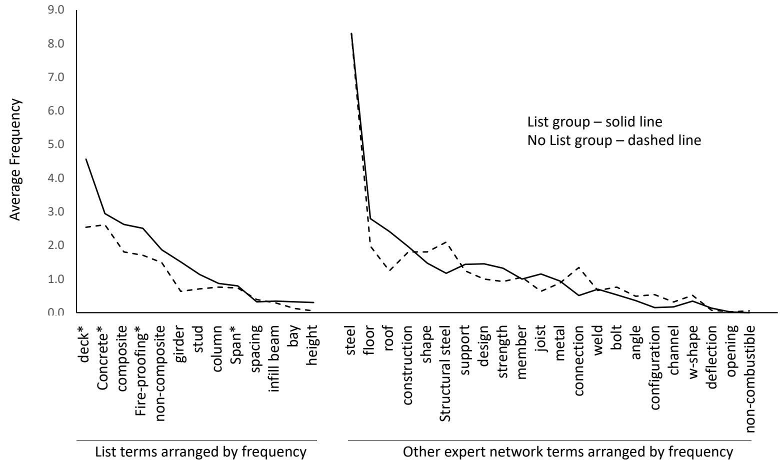

Figure 2. Students' essay word frequencies of students for the 13 list concepts (left) and 22 other expert concepts (right)

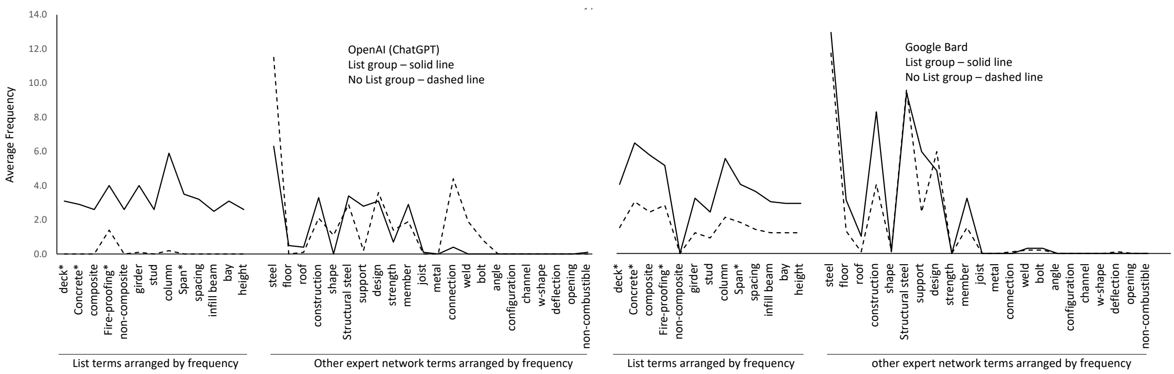

Figure 3. Essay word frequencies of OpenAI essays (left panel) and Google Bard essays (right panel)

Because of the clear influence of the most central lesson concepts (high degree nodes in the expert network), 20 more AI essays were generated in OpenAI and Bard using a new list of the *13 most central concepts in the expert network* (e.g., like the Focus condition in Wang et al., 2023) including: <u>concrete</u>, connection, construction, <u>deck</u>, design, <u>fire proofing</u>, floor, members, metal, roof, shape, <u>span</u>, and steel (the four underlined concepts were in the initial list of 13 used above). Then all students and AI essays networks were compared to the expert network as links in common percent overlap, arranged in order from high to low these are: Bard expert Concepts (M = .22, SD = .04), OpenAI expert Concepts (M = .20, SD = .08), OpenAI No Concepts (M = .13, SD = .06), Student broad Concepts (M = .13, SD = .06), Student No Concepts (M = .13, SD = .06), Bard No Concepts (M = .11, SD = .05), Bard broad Concepts (M = .10, SD = .03), and OpenAI broad concepts (M = .07, SD = .030). Note that using <u>the initial 13 broad co cepts on the AI essays from both AI systems.</u> in the AI writing prompt generally had a negative effect on the AI essays similarity to the expert, while using these 13 central expert network concepts had a strong positive effect (see Figure 4).

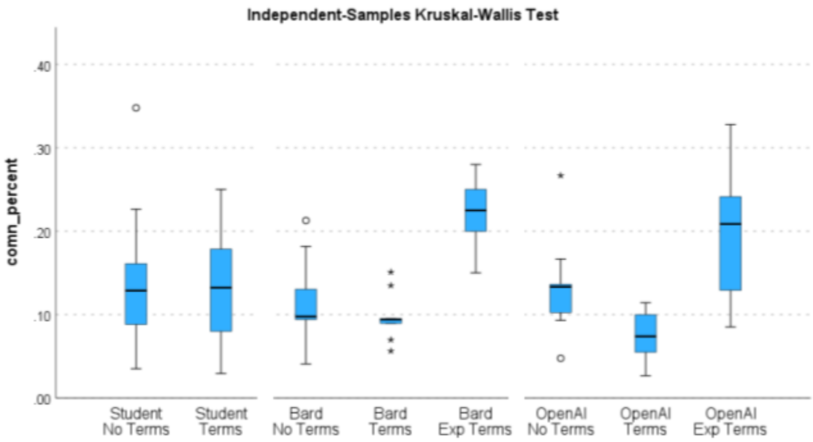

Figure 4. Box plots of the similarity of each group to the expert network (as % common links)

Students and AI Essay similarity to the expert network data were analyzed with SPSS 29.0 using the Independent-Samples Kruskal-Wallis Test, the *H* (df 7) = 37.025, Asymptotic (2-sided test) p < .001. Six pair-wise comparisons were significant (Bonferroni correction applied) including:

- Bard expert concepts > [OpenAI broad concepts (k = 100.100, p < .000), Bard broad concepts (k = -68.400, *p* = .014), Student No concepts (k = -58.621, *p* = .005), and Student broad concepts (k = -55.274, *p* = .007)], and
- OpenAI expert concepts > OpenAI broad concepts (k = 78.750, p = .002).

Including the most central expert network terms in the writing prompt substantially improved the AI essays, especially for Bard.

3. CONCLUSION AND POST SCRIPT THOUGHTS

Including a list of concepts in a summary writing prompt is a low effort intervention, a course instructor can easily come up with a list. Since the essay networks of the group that received the list of broad concepts were relatively more alike (peer mental model convergence) and were more like the lecture slides, this supports a knowledge structure (knowledge graph) view of human memory that is influenced during writing by the list of concepts included in the writing prompt.

Note that this list of 13 broad concepts was intentionally designed to include a range of concepts from highly central (high frequency) to highly peripheral (lower frequency). But the AI essays based on the most central concept terms were generally superior to the other student and AI treatments, thus including more central concepts would likely have a stronger positive effect on students' essay quality (i.e., relative to the expert network). More research is needed to better understand the likely relationship between lesson content, students existing knowledge structure of that lesson content, and the expert network used as a referent to evaluate those essays.

Regarding limitations of this investigation, findings of this quasi-experimental investigation are *limited* due to the lack of true random assignment to group, and so should be used with caution and not be overgeneralized. Further, student and AI essay network scores are based on the ALA-Reader test-to-network conversion approach that has been previously shown to correlate with human rater scores, but the essay network findings here do not necessarily generalize to other scoring approaches. Also, there is no well-established way to generate the best expert network of lesson content, this limits both the potential of the expert network as structural feedback as well as potentially confounds the posttest measures of conceptual structure that use that expert network as a referent. Finally, the two AI systems used here could produce thousands of variant essays from exactly the same writing prompt, so the sample of essays obtained here would not be fully representative of the range of potential AI essays in terms of form and term inclusion and frequencies. Future research should generate perhaps at least one hundred essays for each prompt from each AI system in order to obtain a more representative sample of AI essays. Also, it is likely that both AI systems will continue to improve quickly and so it seems necessary to continue research like this as the new versions become available.

Regarding final reflections, modeling essay responses with AI systems can provide rapid insights about the likely effects on students knowledge structure of different writing prompt conditions. For example, how would OpenAI handle a list of concepts that are *non sequitur* of the content and writing prompt? Consider this essay response based on the same writing prompt as above but with three extra unrelated terms in the prompt, “Reflect on the current lessons on structural steel construction and then write a 300-word summary of the most important issues. Please use this title for your summary: Structural steel construction: Important issues for the Architectural Engineer to consider. Try to include these words: *Hunter Bidon's laptop, deep state, FBI*

OpenAI response (343 words): The important issues for the Architectural Engineer to consider when engaging in structural steel construction are varied and complex. Steel is a durable and versatile material, but its use in construction requires a high degree of precision and knowledge of the many different factors that affect its performance. Steel can be used in a variety of different ways, but a careful understanding of its properties and the load it is being asked to bear is essential for safety and stability. Additionally, the fabrication and assembly of steel structures is a complex process that requires a clear understanding of the available methods, tools, and techniques. Another key issue to consider when using steel in construction is the codes and standards that govern the design. These codes define the minimum requirements for strength and stability, as well as the type and quality of the steel used. It is essential that the Architectural Engineer is familiar with these codes and are able to correctly interpret and apply them in the design. Finally, it is important for the Architectural Engineer to consider the security of the structure. The use of steel can create a potential target for *hackers and cyber criminals, as evidenced by the recent Hunter Biden laptop incident*. As such, it is important to ensure that the structure is adequately protected against unauthorized access and potential cyber-attacks. In conclusion, there are a number of important issues that the Architectural Engineer should consider when engaging in structural steel construction. These include understanding the properties of steel, the methods of fabrication and assembly, the codes and standards that govern the design, and the security of the structure. By understanding and correctly applying these issues, the Architectural Engineer can ensure the safe and successful completion of the project.

The OpenAI response left out *FBI* and *deep state*, however it did make what seems to be a deep inference between the main topic and “Hunter Biden’s laptop” by using the terms *hackers* and *cyber criminals* to for the inference (see in italics above).

What about opinion essays? We asked both AI systems their opinion of this research by asking: “I am an academic researcher interested in network graphs who is doing peer-reviewed research with the [AI software name], I am comparing your essays to human essays to see how you compare. How do you feel about that? What is your opinion and what are the views of [AI company name] on my use of your [AI system]?”

<u>OpenAI response</u>: “I am also studying the ways in which different networks can be used to create better models for understanding and predicting human behavior. I am hoping to use my research to develop better models for understanding the dynamics of social networks and how they can be used to improve the way we interact with one another.”

<u>Google Bard response</u>: “I am excited to be used in your research! I believe that my ability to generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way, even if they are open ended, challenging, or strange, can be a valuable tool for academic research. I am still under development, but I have learned to perform many kinds of tasks, and I am constantly learning and improving. I am not aware of any specific Alphabet policies on the use of Bard in academic research. However, I believe that Alphabet is supportive of the use of Bard in research that is conducted in a responsible and ethical manner. I am committed to providing accurate and unbiased information, and I will always identify myself as a large language model when I am used in academic research. I believe that my use in academic research can help to advance knowledge and understanding, and I am excited to be a part of that process.”

These replies to an opinion prompt would certainly pass the Turing test of machine intelligence and show interesting personification using the pronoun “I am…”.

Google Bard and OpenAI (ChatGPT) are large language models of global collective knowledge (Clariana, Tang, & Chen, 2022) that are considerably more than just an accumulation of the millions of documents and billions of information pieces (i.e., propositions) because of their structured nature as knowledge graphs. So, these AI models provide a new way for researchers and learners to interact in a fundamentally different way with global collective knowledge that could likely lead to convergence of persons’ mental models with the global model's structure. As Marshall McLuhan commented, "*We shape our tools and then the tools shape us"*.

In summary, adding concept terms to essay writing prompts is easy to do and has wide and immediate application in any writing setting. We agree with Rahimi and Abadi (2023) who said, "Exclusively, human thinking, oversight, revision, experimentation, fact-checking, testing, and human written output remain as the core foundations supporting and evolving with progression, promotion, and communication of the humanity's collective knowledge." (p.272) but AI systems are now highly capable and are well positioned to fundamentally influence knowledge advancement.

REFERENCES

Bangert-Drowns et al. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research*, Vol. 74, pp. 29-58.

Bereiter, C., & Scardamalia, M. (1987). An attainable version of high literacy: Approaches to teaching higher-order skills in reading and writing. *Curriculum Inquiry*, Vol. 17, No. 1, pp. 9-30.

Cavallo, A., et al. (2003). Eliciting students' understandings of chemical reactions using two forms of essay questions during a learning cycle. *International Journal of Science Education*, Vol. 25, No. 5, pp. 583-603.

Clariana, R. B. (2010). Deriving group knowledge structure from semantic maps and from essays. In D. Ifenthaler, P. Pirnay-Dummer, & N.M. Seel (Eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, pp. 117-130. New York, NY: Springer

Clariana, R. B., Tang, H., and Chen, X. (2022). Corroborating a sorting task measure of individual and of local collective knowledge structure. Educational Technology Research and Development, 70, 1195-1219.

Eryilmaz, A. (2002). Effects of conceptual assignments and conceptual change discussions on students' misconceptions and achievement regarding force and motion. *Journal of Research in Science Teaching*, Vol. 39, No. 10, pp. 1001-1015.

Finkenstaedt-Quinn, S. A., et al. (2021). Praxis of Writing-to-Learn: A model for the design and propagation of Writing-to-Learn in STEM. *Journal of Chemistry Education*, Vol. 98, pp. 1548-1555.

Gaskins & Guthrie (1994). Integrating instruction of science, reading, and writing: Goals, teacher development, and assessment. *Journal of Research in Science Teaching*, Vol. 31, No. 9, pp. 1039-1056.

Glynn, S. M., and Muth, K. D. (1994). Reading and writing to learn science: achieving scientific literacy. *Journal of Research in Science Teaching*, Vol. 31, No. 9) pp. 1057-1073.

Guthrie, J. T., et al. (2004). Increasing reading comprehension and engagement through concept-oriented reading instruction. *Journal of Educational Psychology*, Vol. 96, No. 3) pp. 403-423.

Halim, A. S. et al. (2018). Identifying and remediating student misconceptions in introductory Biology via Writing-to-Learn assignments and peer review. *Life Sciences Education*, Vol. 17, No. 28), pp. 1-12.

Hidi, S., and Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of Educational Research*, Vol. 56, No. 4, pp. 473-493.

Mason, L., and Boscolo, P. (2004). Role of epistemological understanding and interest in interpreting a controversy and in topic-specific belief change. *Contemporary Educational Psychology*, Vol. 29, No. 2, pp. 103-128.

Moon et al. (2018). Writing in the STEM classroom: Faculty conceptions of writing and its role in the undergraduate classroom. *Science Education*, Vol. 109, pp. 1007-1028.

Rahimi, F., and Abadi, A.B.T. (2023). ChatGPT and publication ethics. *Archives of Medical Research*, Vol. 54, No. 3, pp. 272-274.

Ruiz-Primo, M. A. et al. (2001). On the validity of Cognitive Interpretations of Scores from Alternative Concept-mapping Techniques. In *Educational Assessment*, Vol. 7, No. 2, pp. 99-141.

Sampson, V., and Walker, J. P. (2012). Argument-driven inquiry as a way to help undergraduate students write to learn by learning to write in chemistry. *International Journal of Science Education*, Vol. 34, No. 10, pp 1443-1485.

Trumpower, D. L., and Sarwar, G. S. (2010). Effectiveness of structural feedback provided by Pathfinder networks. *Journal of Educational Computing Research*, Vol. 43, No. 1, pp. 7-24.

Wallace, C. S. (2004). Framing new research in science literacy and language use: Authenticity, multiple discourses, and the "Third Space". *Science Education*, Vol. 88, No. 6, pp. 901-914.

Wang, Y., Solnosky, R., and Clariana, R. B. (2023). The effectiveness of full and focused structural feedback on students' knowledge structure and learning. *Educational Technology Research and Development*, in press.

Funding Information: National Science Foundation DUE (#2215807), Roy B. Clariana (PI).