Chapter 5
Anchoring Concepts: Conceptual Structure and Test Performance

Roy B. Clariana and Ryan L. Solnosky

Abstract How do anchoring concepts in writing prompts influence essays? Undergraduate students (n = 90) read an assigned textbook chapter and attended lectures and labs, and then were asked to write a 300-word summary of the lesson content. Data consisted of the essays converted to networks and the end-of-unit multiple choice test. Comparing similarity to the expert network benchmark, the essay networks of those who received anchoring concepts were not significantly different from those who did not. However, those who received the anchoring concepts were significantly more like their peers' essay networks (mental model convergence) and were more like the networks of the two PowerPoint lectures. Furthermore, those receiving the anchoring concepts performed significantly better on the end-of-unit test than those who did not. Term frequency analysis shows that the most network-central concepts had the greatest frequency in the essays, the other terms' frequencies were remarkably the same for both the anchoring concepts and no concepts groups, suggesting a similar group-level underlying conceptual mental model of this lesson content. To further explore the influence of anchoring concepts in writing prompts, essays were generated with the same two writing prompts using OpenAI (ChatGPT) and Google Bard (i.e., Gemini). The quality of the essay networks for both AI systems were equivalent to the students' essay net- works. More research is needed to understand how including anchoring concepts in a writing prompt (i.e., prompt engineering) influences students' essay conceptual structure and subsequent test performance.

Keywords Summary writing · Writing to learn · Automatic writing assessment · Google bard · OpenAI (ChatGPT) · Prompt engineering

R. B. Clariana (*) · R. L. Solnosky
The Pennsylvania State University, University Park, PA, USA
e-mail: rbc4@psu.edu; rls5008@psu.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

P. Isaias et al. (eds.), *Artificial Intelligence for Supporting Human Cognition and Exploratory Learning in the Digital Age*, Cognition and Exploratory Learning in the Digital Age, https://doi.org/10.1007/978-3-031-66462-5_5

5.1 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 closely 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 those concepts in a related fashion (Gaskins & Guthrie, 1994; Glynn & Muth, 1994; Guthrie et al., 2004). Writing about domain normative topics helps students to understand common disciplinary conceptions and to more effectively participate in discursive learning communities (Wallace, 2004). Specifically important here, Mason and Boscolo (2004) have identified writing as a way to foster conceptual change by encouraging students to develop elaborated explanations of complex phenomena (Halim et al., 2018; Moon et al., 2018), especially for correcting misconceptions. And writing becomes even more effective with formative feedback and reflection (Bangert-Drowns et al., 2004).

This investigation considers whether including a list of lesson anchoring concepts in the writing prompt (following Cavallo et al., 2003) will influence essay quality. Here, anchoring concepts refer to any term or set of terms that would be semantic associates of the full set of terms in a specific conceptual space of interest that can serve to point towards and bound the semantic space of interest (i.e., the lesson content of interest).

Using anchoring concepts to clarify a writing prompt is consistent with prompt directedness in the concept map literature (Ruiz-Primo et al., 2001). Including a list of concepts in a writing prompt does not require much extra effort and could have at least three benefits, specifically these 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 and synonyms 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 approach using OpenAI with these three questions (https://platform.openai.com/playground):

1. *What is a bank*? (no anchoring concepts)—Reply: 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. *What is a bank*? *try to use these concepts: cash, receipt, teller, ATM*—Reply: 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

cash at the teller or use an ATM to deposit or withdraw money. When making a deposit, customers receive a receipt to show their transaction was completed, and

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

The response to the first question (no list of anchoring concepts) was about banks as financial institutions and this meaning 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. Thus, the added concepts anchored the replies to separate distinct AI conceptual spaces in the AI knowledge graph.

Does adding anchoring concepts to a writing prompt matter in students' essays? 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 three extra anchoring concepts, "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 trained raters. Students who received these three extra concepts in the writing prompt had twice as many misunderstandings (21 vs. 11) and these misunderstandings persisted across the course as measured using the same essay prompt at Posttest 1 (20 vs. 11) and at Posttest 2 (15 vs. 14). Including 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 that did improve over time.

So, should we add anchoring concepts or not to writing prompts? A purpose of this current investigation is to inform the development and use of a browser-based writing-to-learn tool called Graphical Interface of Knowledge Structure (GIKS) that provides immediate structural feedback as a network of concepts (Trumpower & Sarwar, 2010). For example, Wang et al. (2024) 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 focus concepts that were the 14 central high degree concepts in the expert network or full concepts that provided all 26 concepts in the network (e.g., central and peripheral). Participants in an undergraduate Architecture Engineering course (n = 68) completed a 2-week lesson module on Building with Timber and Wood, and then wrote a 300-word summary essay 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 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

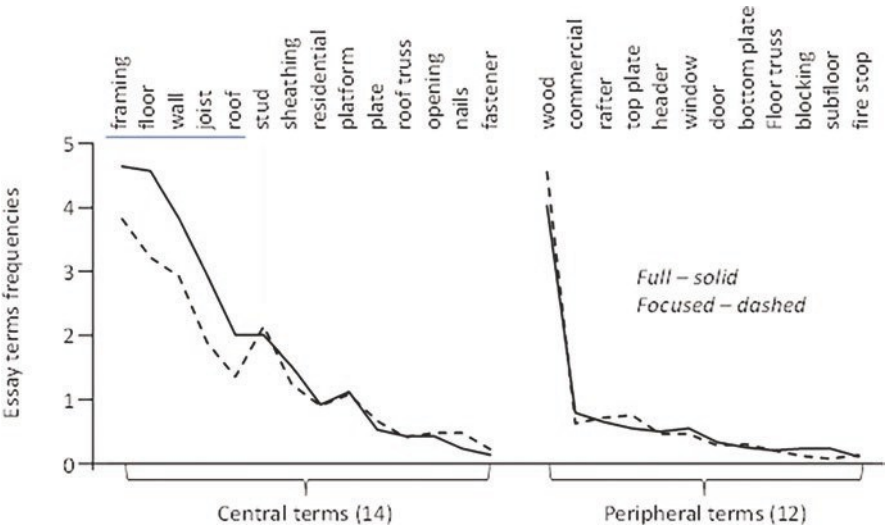

Fig. 5.1 Essay word frequencies of the Central and Peripheral network concepts from Wang et al. (2024)

concepts showed a higher frequency across the essays for the Full compared to the Focus condition (see Fig. 5.1).

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

Because of the anticipated benefits and the likely influence on essays (sometimes perhaps negative) of including a list of concepts in a summary writing prompt, it is critical for the ongoing research and development of GIKS software to determine whether to include a list of anchoring concepts or not in the essay 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., 2024), to explore this, in this investigation we created a list of concepts that intentionally spans the lesson space including highly central, central, peripheral, and highly peripheral, intending to replicate the highly central concept frequency findings from Wang et al. (2024).

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 in order 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 student's mental models as conceptual knowledge structure.

5.2 Participants, Materials, and Results

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

Participants in this quasi-experimental investigation are undergraduate students (N = 110, 24% female) in the course Building Documentation and Modelling 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 2 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 anchoring concepts essay prompt (final sample n = 52) while those on Thursday received the no concepts (control) prompt control (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 anchoring 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 of the high frequency concepts were included from the list of 13 anchoring 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 consisted of essay network similarity measures (as common link percent), end-of-unit multiple-choice test performance, and essay descriptive data (i.e., word frequencies). 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 (such as 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., the 26 expert network +9 anchoring concepts). The students' essay networks similarity to five different referent networks were compared for the List and No List groups (see Table 5.1). 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 anchoring 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 for the anchoring concepts group). In addition, the anchoring 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 anchoring concepts or not for similarity to the textbook chapter network (p = .228).

And finally, on the end-of-unit subtest that aligned with the lesson content, the anchoring concepts group (mean rank = 50.6) outperformed the no concepts group (mean rank = 38.5), H (df 1) = 4.687, p = .030, but not on the end-of-unit subtest that covered the other lessons in the module, H (df 1) = 0.430, p = .43).

Table 5.1 Kruskal–Wallis findings for each measure

	Students' essay network similarity (as % common links)				
	to expert	to peers	Chp. 11	PP #1	PP #2
Kruskal-Wallis H	.160	5.987	1.452	5.498	8.940
df	1	1	1	1	1
Asymp. Sig. (p =)	.689	**.014**	.228	**.019**	**.003**
No list (mean rank)	44.21	37.62	41.62	37.95	35.87
List (mean rank)	46.44	51.26	48.34	51.02	52.54

Bonferroni correction applied

5.2.2 Comparing Word Frequencies of Student Essays and AI Essays

Student essay concept frequencies align with the findings above from Wang et al. (2024) that providing a list of anchoring concepts in the prompt increases concept frequency of only a few of the most 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-anchoring concepts, floor and roofing, that were not in the list but that are highly central concepts in the instructor's expert network (see top panel of Fig. 5.2). To further explore the influence of providing lists of concepts in the writing prompt, 40 AI essays were generated using OpenAI playgound (text-davinci-003, temperature = .7, https://platform.openai.com/playground) and Google Bard (Language Model for Dialogue Applications, https://bard.google.com/). 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 13 list concepts plus the 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 (notice solid lines above dashed lines in the middle and bottom panels of Fig. 5.2). Also, although the two AI systems are distinctly different from each other, there are considerable similarities for term frequencies for the two AI systems 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 in the bottom panel of Fig. 5.2). Perhaps the two AI systems have a similar knowledge graph of this content.

Because of the clear influence of the most central lesson concepts (i.e., high degree nodes in the expert network), 20 more AI essays were generated with OpenAI and Bard using a new list of the 13 most central concepts in the expert network (e.g., matches the Focus condition strategy used by Wang et al., 2024) including: concrete, connection, construction, deck, design, fire proofing, floor, members, metal, roof, shape, span, and steel (the four underlined concepts were in the initial list of 13 used above). Then all students and AI essay networks were compared to the expert network as links-in-common percent overlap (see Fig. 5.3).

Performance of each arranged in order from high to low are: Bard with expert Concepts (M = .22, SD = .04), OpenAI with expert Concepts (M = .20, SD = .08), OpenAI No Concepts (M = .13, SD = .06), Student with broad Concepts (M = .13, SD = .06), Student No Concepts (M = .13, SD = .06), Bard No Concepts (M = .11, SD = .05), Bard with broad Concepts (M = .10, SD = .03), and OpenAI with broad concepts (M = .07, SD = .030). Note that using the initial 13 broad anchoring concepts in the AI writing prompt to derive the AI essays with both AI systems generally had a negative effect on the AI essays similarity to the expert; in contrast, using 13 central expert network anchoring concepts had a strong positive effect,

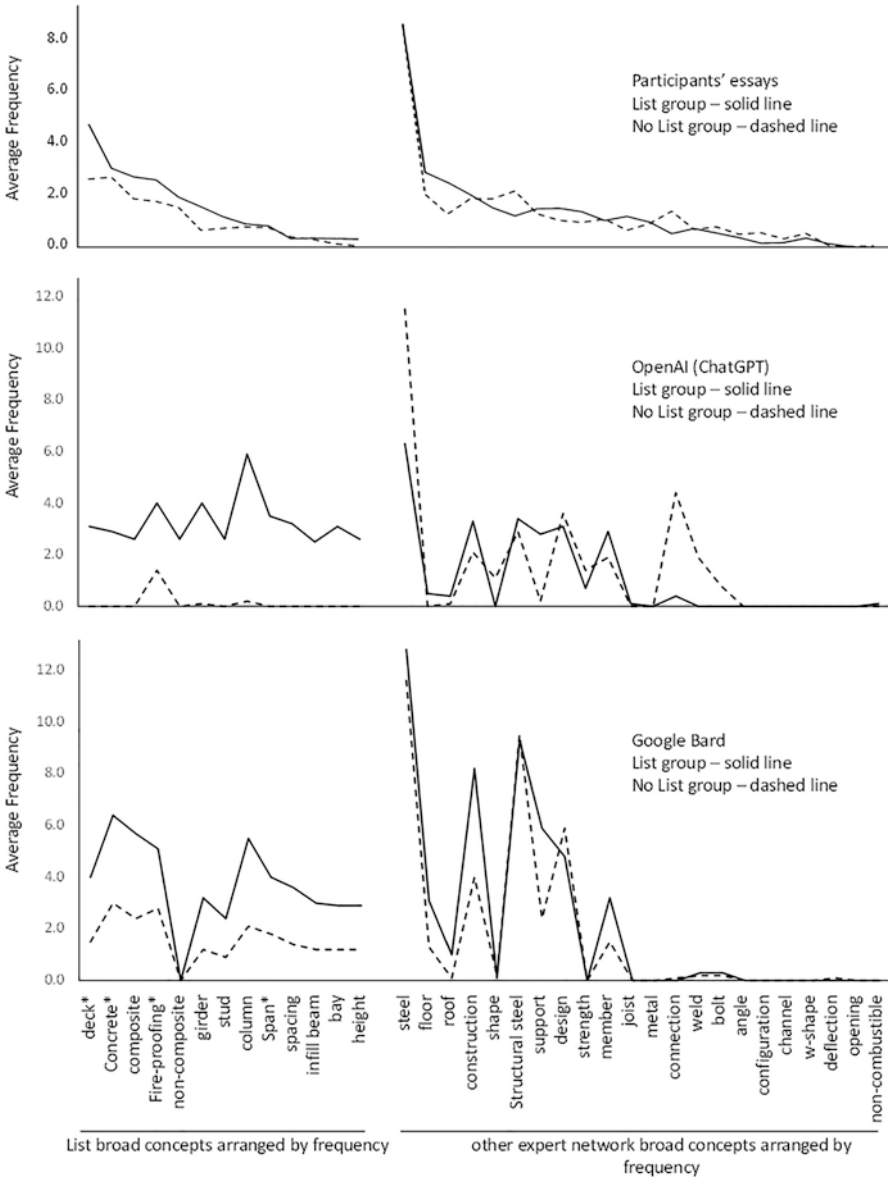

Fig. 5.2 Students' essay word frequencies of students for the 13 broad anchoring concepts (left) and 22 other expert concepts (right)

thus including central anchoring concepts in an essay prompt has a substantive impact on AI output responses.

How do the AI essay networks compare to the students' essay networks? 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,

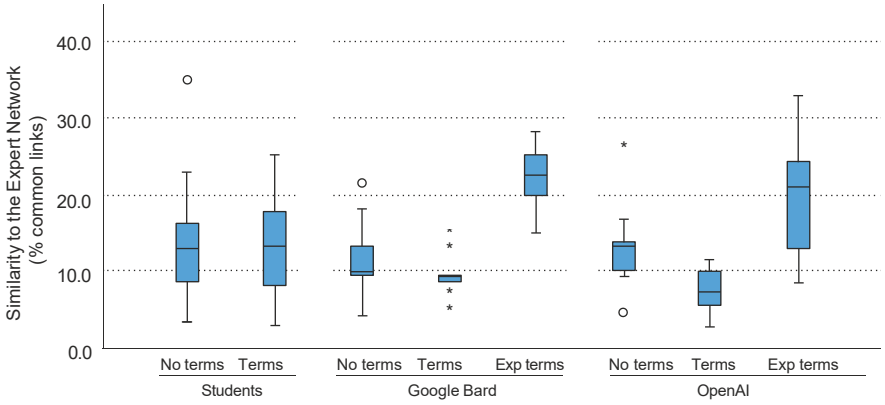

Fig. 5.3 Box plots of the similarity of each group to the expert network (as % common links)

Asymptotic (2-sided test) p < .001. Five pair-wise comparisons were significant (Bonferroni correction applied) including:

- Bard with expert concepts >

  OpenAI with broad concepts (k = 100.100, p < .000),
  Bard broad concepts (k = −68.400, *p* = .014),
  Student No concepts (k = −58.621, *p* = .005),
  Student broad concepts (k = −55.274, *p* = .007)], and

- OpenAI with expert concepts > OpenAI with broad concepts (k = 78.750, *p* = .002).

Including the most central expert network terms in the writing prompt substantially improved the AI essay quality relative to the expert, especially for Bard (now called Gemini).

Is biological sex a factor? Traditional test formats in STEM courses (i.e., multiple choice, visual-spatial, and computation test items) substantially favours males over females (Maccoby & Jacklin, 1974: males > females with an effect size = .4 to .5). In contrast, females tend to outscore males on tests of verbal ability such as summary writing, females > males with an effect size = .25. An Education Testing Service (ETS) meta-analysis by Cole (1997) reported that across many studies females tend to excel over males in writing tasks. How do the male and female essays compare? A Kruskal–Wallis test by ranks of essay network similarity to the expert network showed no difference between male and female essay quality, the difference between the rank totals of 38.7 (female) and 48.7 (male) was not significant, H (df 1) = 2.94, *p* = .087.

5.3 Conclusion and Postscript Thoughts

The data in this investigation highlights the importance of essay term frequencies as a perhaps critical measure that is both descriptive and prescriptive. For example, Yeari and Lantin (2021) coined the term centrality deficit to describe the text representation of poor comprehenders, particularly of central ideas, who construct a low-quality, poorly connected text representation during reading. Existing and future theories of knowledge structure and associated data collection approaches should give central ideas a more central role.

Pedagogically, including a list of anchoring 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 anchoring 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 concepts included in the writing prompt.

Note that the 13 broad anchoring concepts were 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 in the expert's network were generally superior to the students' and the other AI essay networks. 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 understand the likely relationship between lesson content, student's existing knowledge structure of that lesson content, and the expert network used as a referent to evaluate those essays.

Regarding limitations of this study, 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 text-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 essay scoring approaches such as human raters with rubrics.

Also, this investigation clearly shows that because there is no well-established way to establish the best expert network of lesson content, this limits both the potential of the expert network as structural feedback as well as potentially confounds posttest measures of conceptual structure that use an expert network as the referent for generating links in common scores. Substantial additional thought, theory, and research are needed to determine how to establish an expert network referent.

Finally, the two AI systems used here could produce thousands of similar but variant essays using the same writing prompt, so the sample of essays obtained here are only partially representative of the range of potential AI essays in terms of form, term inclusion, and concept frequencies. Although time consuming and thus costly, future research should generate perhaps at least one hundred essays for each prompt with each AI system in order to obtain a large representative sample of AI essays.

In addition, it is likely that both AI systems will continue to improve quickly and so it seems critical to continue research like this as the new AI versions become available.

Regarding final reflections, modelling essay responses with AI systems can provide rapid insights about the likely effects of different writing prompt conditions on students' knowledge structure. For example, how would students and OpenAI handle a list of concepts that are non sequitur of the content and writing prompt? Consider this essay response below based on the same writing prompt as used 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 form an inference (see in italics above).

What about opinion essays? We asked both AI systems their opinion of this present investigation 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]?"

> *OpenAI response*: "I am also studying the ways in which different networks can be used to create better models for understanding and predicting human behaviour. 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."

> *Google Bard response*: "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 et al., 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. 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 over time and experiences to convergence of individual's 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 anchoring concepts 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.

Acknowledgements This material is based upon work supported by the U.S. National Science Foundation under award No. (2215807). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation.

References

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research, 74*(1), 29–58. https://doi.org/10.3102/00346543074001029

Bereiter, C., & Scardamalia, M. (1987). An attainable version of high literacy: Approaches to teaching higher-order skills in reading and writing. *Curriculum Inquiry, 17*(1), 9–30. https://doi.org/10.1080/03626784.1987.11075275

Cavallo, A. M. L., McNeely, J. C., & Marek, E. A. (2003). Eliciting students' understandings of chemical reactions using two forms of essay questions during a learning cycle. *International Journal of Science Education, 25*(5), 583–603. https://doi.org/10.1080/09500690210145774

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 diag- nostics and systematic analysis of knowledge* (pp. 117–130). Springer. https://doi. org/10.1007/978-1-4419-5662-0_7

Clariana, R. B., Tang, H., & Chen, X. (2022). Corroborating a sorting task measure of indi- vidual and of local collective knowledge structure. *Educational Technology Research and Development, 70*, 1195–1219. https://doi.org/10.1007/s11423-022-10123-x

Cole, N. S. (1997). *The ETS gender study: How females and males perform in educational settings*. Educational Testing Service. https://eric.ed.gov/?id=ED424337

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, 39*(10), 1001–1015. https://doi.org/10.1002/tea.10054

Finkenstaedt-Quinn, S., Petterson, M., Gere, A., & Shultz, G. (2021). Praxis of writing-to-learn: A model for the design and propagation of writing-to-learn in STEM. *Journal of Chemistry Education, 98*, 1548–1555. https://doi.org/10.1021/acs.jchemed.0c01482

Gaskins, I. W., & Guthrie, J. T. (1994). Integrating instruction of science, reading, and writing: Goals, teacher development, and assessment. *Journal of Research in Science Teaching, 31*(9), 1039–1056. https://doi.org/10.1002/tea.3660310914

Glynn, S. M., & Muth, K. D. (1994). Reading and writing to learn science: Achieving scientific literacy. *Journal of Research in Science Teaching, 31*(9), 1057–1073. https://doi.org/10.1002/tea.3660310915

Guthrie, J. T., Wigfield, A., Barbosa, P., Perencevich, K., Taboada, B., Ana, M., Davis, M., Scafiddi, N., & Tonks, S. (2004). Increasing reading comprehension and engagement through concept-oriented reading instruction. *Journal of Educational Psychology, 96*, 403–423. https://doi.org/10.1037/0022-0663.96.3.403

Halim, A. S., Finkenstaedt-Quinn, S. A., Olsen, L. J., Gere, A. R., & Shultz, G. V. (2018). Identifying and remediating student misconceptions in introductory biology via writing-to-learn assignments and peer review. *Life Sciences Education, 17*(28), 1–12. https://doi.org/10.1187/cbe.17-10-0212

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of Educational Research, 56*(4), 473–493. https://doi.org/10.2307/1170342

IBM. (2023). *What is a knowledge graph*? https://www.ibm.com/topics/knowledge-graph

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford University Press. https://psycnet.apa.org/record/1975-09417-000

Mason, L., & Boscolo, P. (2004). Role of epistemological understanding and interest in interpreting a controversy and in topic-specific belief change. *Contemporary Educational Psychology, 29*(2), 103–128. https://doi.org/10.1016/j.cedpsych.2004.01.001

Moon, A., Gere, A. R., & Shultz, G. V. (2018). Writing in the STEM classroom: Faculty conceptions of writing and its role in the undergraduate classroom. *Science Education, 109*, 1007–1028. https://doi.org/10.1002/sce.21454

Rahimi, F., & Abadi, A. B. T. (2023). ChatGPT and publication ethics. *Archives of Medical Research, 54*(3), 272–274. https://doi.org/10.1016/j.arcmed.2023.03.004

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the validity of cog- nitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment, 7*(2), 99–141. https://doi.org/10.1207/S15326977EA0702_2

Sampson, V., & 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, 34*(10), 1443–1485. https://doi.org/10.1080/09500693.2012.667581

Trumpower, D. L., & Sarwar, G. S. (2010). Effectiveness of structural feedback provided by pathfinder networks. *Journal of Educational Computing Research, 43*(1), 7–24. https://doi.org/10.2190/EC.43.1.b

Wallace, C. S. (2004). Framing new research in science literacy and language use: Authenticity, multiple discourses, and the "third space". *Science Education, 88*(6), 901–914. https://doi.org/10.1002/sce.20024

Wang, Y., Solnosky, R., & Clariana, R. B. (2024). *The effectiveness of full and focused structural feedback on students' knowledge structure and learning*. Technology Knowledge and Learning.

Yeari, M., & Lantin, S. (2021). The origin of centrality deficit in text memory and comprehension by poor comprehenders: A think-aloud study. *Reading and Writing, 4*, 595–625. https://doi.org/10.1007/s11145-020-10083-9