Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Small but Significant: On the Promise of Small Language Models for Accessible AIED

Yumou WEI[a], Paulo CARVALHO[a] & John STAMPER[a]
[a]*Human-Computer Interaction Institute, Carnegie Mellon University, USA*
{yumouw, pcarvalh, jstamper}@andrew.cmu.edu

Abstract: GPT has become nearly synonymous with large language models (LLMs), an increasingly popular term in AIED proceedings. A simple keyword-based search reveals that 61% of the 76 long and short papers presented at AIED 2024 describe novel solutions using LLMs to address some of the long-standing challenges in education, and 43% specifically mention GPT. Although LLMs pioneered by GPT create exciting opportunities to strengthen the impact of AI on education, we argue that the field's predominant focus on GPT and other resource-intensive LLMs (with more than 10B parameters) risks neglecting the potential impact that small language models (SLMs) can make in providing resource-constrained institutions with equitable and affordable access to high-quality AI tools. Supported by positive results on knowledge component (KC) discovery, a critical challenge in AIED, we demonstrate that SLMs such as Phi-2 can produce an effective solution without elaborate prompting strategies. Hence, we call for more attention to developing SLM-based AIED approaches.

Keywords: Small Language Models, Accessible AIED, KC Discovery

1. Introduction

It is an exciting time for AIED. Technological breakthroughs in large language models (LLMs) (Brown et al., 2020) have provided unprecedented opportunities for AIED researchers and practitioners to solve some of the long-standing challenges in the field (Kasneci et al., 2023). The excitement is aptly exemplified by the community's fast adoption of LLMs in AIED research—of the 76 long and short papers accepted for AIED 2024, 61% (47 papers) describe innovative solutions using LLMs, as revealed by a simple keyword-based search in the proceedings (Olney et al., 2024). Among the ever-expanding constellation of available LLMs, the GPT family, including ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023), appears to be the community's favorite: 33 of the 47 papers (70%) adopting LLMs also mention GPT. Although LLMs pioneered by GPT herald exciting possibilities to reinforce AI's positive influence on education, we argue that the community's predominant focus on GPT and other similar resource-intensive gigantic language models (with more than ten billion parameters) risks neglecting the critical impact that small language models (SLMs) can make in creating equitable and accessible education central to the mission of AIED.

The definition of SLMs is constantly changing as new technologies emerge to shape the landscape of language models. The BERT model (Devlin et al., 2019) in its largest configuration, for example, has 340 million parameters—an overwhelming amount in 2018 but only a fraction by today's standard. In relation to the current state of the art, we consider a language model small if it has fewer than ten billion parameters and requires modest hardware resources, such as a consumer-grade GPU. Canonical examples of SLMs include Llama-2 7B (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023), and Phi-2 (Javaheripi et al., 2023). Phi-2, a lightweight but capable model that has only 2.7B parameters, might be a particularly good fit for the AIED community and the range of problems we are trying to address. Trained on high-quality "textbook-like" data (Gunasekar et al., 2023), Phi-2 subsumes deep knowledge about various academic disciplines and aligns better with educational contexts, which require precision and reliability, than other SLMs trained on mixed-quality data sourced from the

Internet. Its smaller size also enables local deployment on consumer-grade hardware, desirable for most educational settings where computational resources are limited.

Educational institutions operate under distinct constraints that make their AI implementation needs different from those of commercial environments. Budget limitations, technical infrastructure, privacy requirements, and equity considerations all influence technology adoption in educational settings (Reich & Ito, 2017). GPT-scale LLMs typically require substantial computational resources for local deployment or incessant API costs for cloud access, not affordable to all teachers or students (Kasneci et al., 2023). SLMs, however, only require a fraction of the resources entailed by LLMs and can be deployed on modest hardware at a much lower cost—Phi-2's 2.7 billion parameters only require about 5.4 GB of memory for storage with a 16-bit representation of floating-point numbers[1], which can fit comfortably to a consumer-grade GPU.

One argument that justifies the higher costs of GPT-scale LLMs is their superior performance in various tasks. However, we argue that the more affordable and accessible SLMs can also deliver impressive results if we manage to exploit their potential adequately. In Section 3, we present a case study of knowledge component (KC) discovery (Koedinger et al., 2012), a critical challenge in AIED, and describe our unique solution using Phi-2 (Wei et al., 2025). Our approach makes creative use of Phi-2 as a probability machine to measure question similarity and applies a clustering algorithm to identify questions belonging to the same KC; results on two datasets show that instructors can better predict student performance using the KCs generated by our approach than using those produced by experts or the more powerful GPT-4o. These positive findings from the case study reinforce our position that **small language models such as Phi-2 can provide effective solutions to critical AIED problems and hold great promise as a catalyst for inclusive, personal, and ethical education equitably accessible to teachers and students**.

2. Background

2.1 The Rise of Large Language Models in Education

The field of education has tremendously benefited from the advances in natural language processing (NLP) in recent decades, which have evolved from rule-based approaches to statistical methods and eventually to neural-network models (Litman, 2016). Early educational applications used relatively simple NLP techniques for tasks such as automated essay scoring (Shermis & Burstein, 2013); more recent work, however, uses advanced language models to tackle increasingly complex challenges in education.

Introduced in 2017, the Transformer architecture (Vaswani et al., 2017) enables researchers to build more sophisticated language models with enhanced language understanding and generation capabilities. Together with more efficient hardware and better available corpora, this architectural innovation spurred the development of models with progressively larger parameter counts—some prominent milestones include GPT-3 (Brown et al., 2020) (175B parameters), PaLM (Chowdhery et al., 2022) (540B parameters), and GPT-4 (OpenAI, 2023) (estimated 1.76T parameters). These gigantic language models have demonstrated remarkable capabilities across various educational applications, including but not limited to hint creation (Pardos & Bhandari, 2023), question generation (Sarsa et al., 2022), and KC discovery (Moore et al., 2024).

Concomitant to the development of more capable models is the emphasize of scaling—increasing model size, training data, and computational resources—as the primary mechanism for improving model performance (Kaplan et al., 2020). This scaling law suggests that many unexpected capabilities can emerge as model size increases, with larger models generally outperforming smaller ones across diverse tasks (J. Wei et al., 2022). While the successful application of the scaling law has nearly depleted the available benchmarks to

[1] 2.7B float numbers require 16×2.7B=43.2B bits, which translate to 43.2B/8B = 5.4 gigabytes if 8 bits make 1 byte.

measure the progress of LLMs, urging the development of the "Humanity's Last Exam"[2], it has also raised the computational and financial requirements that prevent resource-constrained educational institutions from equitably using LLMs, and necessitated stricter, more private access to the source code and training data that could have helped researchers build more effective AIED tools. Moreover, the community's widespread predilection for large and even larger models can exacerbate the danger of overlooking the impact that SLMs can make in providing effective and accessible AIED solutions.

2.2 The Potential of Small Language Models in Education

In contrast to the scaling efforts, researchers have also developed smaller and more efficient models that challenge the dominance of scaling as the only way to attain good performance. More recently, models like Phi-2 (2.7B parameters) have demonstrated that careful data curation and innovative training methodologies can produce surprisingly capable models at significantly smaller scales (Javaheripi et al., 2023).

Developed by Microsoft Research, Phi-2 is an epitome of efficient language models. This SLM is built on the standard Transformer decoder-only architecture and is trained with the conventional next-token prediction objective. What makes it special, however, is not architectural innovations but the unique training methodology used. Unlike many larger models trained on vast but heterogeneous corpora sourced from the Internet, Phi-2 was trained predominantly on what the researchers call "textbook-quality data" (Gunasekar et al., 2023)—carefully curated content with an emphasis on educational materials, synthetic texts designed for reasoning capabilities, and filtered web content with high educational value.

This unique training methodology, which ranks data quality higher than quantity, results in an efficient SLM that is particularly useful for educational applications. In competitive benchmarks that evaluate reasoning skills in math (GSM8k (Cobbe et al., 2021)) and coding (HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021)), Phi-2 substantially outperformed Mistral 7B (Jiang et al., 2023) and Llama-2 13B (Touvron et al., 2023), which are 1.6× and 3.8× larger than Phi-2. Compared to the 25× larger Llama-2 70B (Touvron et al., 2023), Phi-2 achieved significantly better performance in coding and demonstrated comparable reasoning skills in math (Javaheripi et al., 2023). In the MMLU benchmark (Hendrycks et al., 2021), which assesses language model knowledge in 57 academic subjects, Phi-2 outperformed Llama-2 13B (54.8) and achieved a score (56.7) comparable to that achieved by Mistral 7B (60.1).

From a computational efficiency perspective, Phi-2 also offers distinct advantages for educational applications. Requiring approximately 5.4 GB of memory for storage (with additional memory for inference), Phi-2 can be deployed on consumer-grade hardware with modest requirements (the conventional 16-GB GPU), enabling local inference without cloud infrastructure dependencies. This flexibility in deployment helps reduce the first digital divide (Attewell, 2001) that prevents resource-constrained schools from using the latest AI tools, and protects student privacy (Prinsloo & Slade, 2017) by not requiring student data to be shared with a third party.

Phi-2's solid results on academic benchmarks and modest requirements on computer hardware make it a competitive alternative to gigantic language models that entail substantial computational resources and provoke critical privacy concerns. Its extensive pre-training on high-quality textbook-like data makes Phi-2 particularly tuned to educational applications. In what follows, we describe a concrete case study in which we creatively used Phi-2 to design a KC discovery algorithm that outperformed instructional experts and its GPT counterpart.

3. Case Study: Knowledge Component Discovery

Representing specific concepts or skills that students acquire through learning to perform a task or solve a problem, knowledge components (KCs) are essential elements in the KLI framework (Koedinger et al., 2012) that help instructors assess student learning. Traditionally, instructional experts are elicited to participate in Cognitive Task Analysis (CTA) (Clark et al.,

[2] https://agi.safe.ai/

2008) to identify the KCs associated with each assessment item, but CTA incurs considerable time and labor cost (Stamper et al., 2011) even when applied to moderately sized question banks. The accelerating adoption of AI in education aggravates the burden on instructional designers, who are overwhelmed by the growing amount of AI-generated questions that each needs to be analyzed by hand.

To address this challenge, a recent approach (Moore et al., 2024) uses GPT-4 (OpenAI, 2023) to extract KCs from multiple-choice questions (MCQs). The authors devised elaborate prompting strategies to ask GPT-4 to simulate instructional experts or textbook authors. Although in an evaluation study, the majority of the three participants preferred GPT-generated KCs to those designed by experts for more than 60% of the evaluated questions, this approach produced KC labels with slightly different wording for questions that instructors think should belong to the same KC (Moore et al., 2024). In our replication of their study using more advanced GPT-4o, the most intelligent non-reasoning LLM offered by OpenAI, we obtained 614 unique KC labels for 630 MCQs from the same e-learning dataset[3] used by Moore et al., 2024. The large number of KC labels comparable to the number of questions suggests that some labels can be merged. In fact, we discovered that GPT-4o had produced unnecessarily refined labels (e.g., "Analyze CTA", "Analyze CTA in E-learning", and "Analyze CTA methodologies") that could have been merged.

In our recent work proposing a new KC discovery method called KCluster (Wei et al., 2025), we demonstrate that exploiting the native potential of a language model as a "probability machine" rather than the more conventional text generation capabilities can lead to a strong KC discovery algorithm even with SLMs such as Phi-2. The core idea is that language models can induce a novel measure of question similarity, which a clustering algorithm can use to identify groups of similar questions that are likely to share the same KC. Specifically, for an arbitrary pair of questions $\boldsymbol{q_s}$ and $\boldsymbol{q_t}$ from a large collection of questions, we evaluate the change in log-probability of $\boldsymbol{q_s}$ with and without the presence of $\boldsymbol{q_t}$:

$$\Delta\,(\boldsymbol{q_s}, \boldsymbol{q_t}) \stackrel{\text{def}}{=} \mathbf{log\,Pr}(\boldsymbol{q_s} \mid \boldsymbol{q_t}) \;-\; \mathbf{log\,Pr}(\boldsymbol{q_s})$$

and define a novel measure of question similarity called "congruity":

$$\text{Congruity}\,(\boldsymbol{q_s}, \boldsymbol{q_t}) \stackrel{\text{def}}{=} \frac{\mathbf{1}}{\mathbf{2}}\,[\Delta\,(\boldsymbol{q_s}, \boldsymbol{q_t}) + \Delta\,(\boldsymbol{q_t}, \boldsymbol{q_s})]$$

The formula for Δ is mathematically equivalent to the pointwise mutual information (PMI) (Church & Hanks, 1989) between two questions (instead of words). The idea was nevertheless inspired by word collocations. We postulate that if one question increases the likelihood of another question appearing, the two questions are congruent and likely to relate to the same KC; on the other hand, if one question barely changes or even decreases the probability that another question occurs, the two questions are incongruent and likely to belong to different KCs. SLMs such as Phi-2, which were *specifically* trained to calculate next-token probabilities, are exceptional at evaluating conditional log-probabilities of the form $\mathbf{log\,Pr}(\boldsymbol{q_s} \mid \boldsymbol{q_t})$. An algorithm that uses Phi-2 to calculate various required probabilities is described in the original paper (Wei et al., 2025).

We evaluated our approach against instructional experts and our replication of the previous study (Moore et al., 2024) using more advanced GPT-4o, on two datasets collected in a graduate e-learning course taught by two different instructors in 2022 and 2023[4]. A common practice to compare different KC discovery approaches is to fit an Additive Factors Model (AFM) (Cen et al., 2007) with the KCs generated by each method to student response data; a better KC discovery approach should allow an instructor to predict student responses with a lower root mean square error (RMSE). On the 2022 dataset, KCluster generated 114 KCs (comparable to the 101 KCs in the best expert KC model) and achieved an RMSE of 0.4220, outperforming both experts (0.4235) and GPT-4o (0.4395); likewise, on the 2023 dataset, KCluster generated 92 KCs (comparable to the 75 KCs in the best expert KC model) and scored 0.4066, leading both experts (0.4075) and GPT-4o (0.4101). We did not merge similar KC labels generated by KCluster unless they were identical. Notably, GPT-4o, a highly capable LLM, performed the worst on the two distinct datasets. This strengthens our claim

[3] https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=5426

[4] https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=5843

that **SLMs can also deliver superior results if their potential is adequately exploited**.

4. Conclusion

Through this forward-looking paper, we did not argue that the AIED community should eschew LLMs in favor of their more efficient counterparts, nor did we suggest that SLMs are capable of everything LLMs can do. Similar to many AIED researchers, we share the excitement about the complementary development of both lines of NLP research and their potential application to education. However, to empower teachers and students for an equitable future, the promise of SLMs in providing accessible AIED solutions is not to be neglected. As shown in the case study and more in our recent paper (Wei et al., 2025), an innovative exploitation of SLM's potential can deliver superior results than the standard use of LLMs based on intensive prompt engineering. **We urge the AIED community to reconsider, next time when making a convenient API call to an LLM, whether it endangers the accessibility to the target audience, who may actually benefit from an SLM**.

References

Attewell, P. (2001). Comment: The first and second digital divides. *Sociology of Education*, 74(3), 252–259. Retrieved March 12, 2025, from http://www.jstor.org/stable/2673277

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., & Sutton, C. (2021). Program synthesis with large language models.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Cen, H., Koedinger, K. R., & Junker, B. (2007). Is over practice necessary? improving learning efficiency with the cognitive tutor through educational data mining. *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, 511–518.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., ... Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*. https://arxiv.org/abs/2107.03374

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... Fiedel, N. (2022). Palm: Scaling language modeling with pathways.

Church, K. W., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. *27th Annual Meeting of the Association for Computational Linguistics*, 76–83.

Clark, R. E., Feldon, D., van Merriënboer, J. J. G., Yates, K., & Early, S. (2008). Cognitive task analysis. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd, pp. 577–593).

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*. https://arxiv.org/abs/2110.14168

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), 4171–4186.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Giorno, A. D., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., & Li, Y. (2023). Textbooks are all you need.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding [arXiv:2009.03300]. *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2009.03300

Javaheripi, M., Bubeck, S., Abdin, M., Aneja, J., Bubeck, S., Mendes, C. C. T., Chen, W., Giorno, A. D., Eldan, R., Gopi, S., Gunasekar, S., Javaheripi, M., Kauffmann, P., Lee, Y. T., Li, Y., Nguyen,

A., de Rosa, G., Saarikivi, O., Salim, A., ... Zhang, Y. (2023, December). Phi-2: The surprising power of small language models.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavril, T., Lachaux, M.-A., Massiceti, D., Rio, J., Lambert, R., Bhosale, S., Aminov, S., Kool, W., Everett, R., ...Calandriello, J. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*. https://arxiv.org/abs/2310.06825

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. https://doi.org/https://doi.org/10.1016/j.lindif.2023.102274

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(5), 757–798. https://doi.org/https://doi.org/10.1111/j.1551-6709.2012.01245.x

Litman, D. (2016). Natural language processing for enhancing teaching and learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1). https://doi.org/10.1609/aaai.v30i1.9879

Moore, S., Schmucker, R., Mitchell, T., & Stamper, J. (2024). Automated generation and tagging of knowledge components from multiple-choice questions. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 122–133. https://doi.org/10.1145/3657604.3662030

Olney, A. M., Chounta, I.-A., Liu, Z., Santos, O. C., & Bittencourt, I. I. (Eds.). (2024). *Artificial Intelligence in Education: 25th International Conference*, AIED 2024, Recife, Brazil, July 8–12, 2024, Proceedings, Part II (Vol. 14830). Springer Nature Switzerland.

OpenAI. (2022, November). Chatgpt: Optimizing language models for dialogue [Accessed: March 2025]. https://openai.com/blog/chatgpt/

OpenAI. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Pardos, Z. A., & Bhandari, S. (2023). Learning gain differences between chatgpt and human tutor generated algebra hints.

Prinsloo, P., & Slade, S. (2017). An elephant in the learning analytics room: The obligation to act. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 46–55.

Reich, J., & Ito, M. (2017). From good intentions to real outcomes: Equity by design in learning technologies. In W. Fitzgerald, J. Burns, N. Sonwalkar, & J. Urry (Eds.), The digital learning challenge: Obstacles to educational uses of copyrighted material in the digital age (pp. 1–42). *The Digital Media; Learning Research Hub*. https://clalliance.org/publications/good-intentions-real-outcomes-equity-design-learningtechnologies/

Sarsa, S., Denny, P., Hellas, A., & Leinonen, J. (2022). Automatic generation of programming exercises and code explanations using large language models. *Proceedings of the 2022 ACM Conference on International Computing Education Research* - Volume 1, 27–43.

Shermis, M., & Burstein, J. (2013). Handbook of automated essay evaluation: Current applications and new directions. *Journal of Writing Research*, 5(2), 239–243.

Stamper, J. & Koedinger, K. (2011). Human-Machine Student Model Discovery and Improvement Using DataShop. In Biswas, G., Bull, S., Kay, J., and Mitrovic, A., eds., Artificial Intelligence in Education, 353–360. Berlin, Heidelberg: Springer. ISBN 978-3-642-21869-9.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Zhu, W. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*. https://arxiv.org/abs/2307.09288

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30 (nips 2017)* (pp. 5998–6008). Curran Associates, Inc.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models [Survey Certification]. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=yzkSU5zdwD

Wei, Y., Carvalho, P., & Stamper, J. (2025). KCluster: An LLM-based Clustering Approach to Knowledge Component Discovery. In Mills, C., Alexandron, G., Taibi, D., Bosco, G. L., and Paquette, L., eds., Proceedings of the 18th International Conference on Educational Data Mining, 228–240. Palermo, Italy: International Educational Data Mining Society.