Integrating External Competitions into Project-Based Learning for AI Education: The Development of TinySlime at NIT, Toba College

Kazuaki Shiraishi* [a], Kaito Ogasawara [a] and Hibiki Otsu [a]

[a] National Institute of Technology, Toba College/ Department of Informatics and Mechanical Engineering, National Institute of Technology, Toba, Japan

Kazuaki Shiraishi * (siraisi@toba-cmt.ac.jp)

At the National Institute of Technology, Toba College, our Department of Informatics and Mechanical Engineering champions Project Based Learning (PBL) integrated with external competitions (e.g., World AI Competition YAMAGUCHI 2022 U18 2nd Place, DCON2024) to cultivate advanced AI competencies. This paper presents a two-fold educational strategy: initially, it highlights how this approach enabled students to develop "TinySlime," a high-performing, 1.1B parameter Japanese Large Language Model (LLM). The model's strong benchmark performance against larger LLMs, its on-premise operational capability minimizing data-leakage risks, and its public availability on Hugging Face collectively evidence the students' profound understanding of contemporary LLM challenges and their technical proficiency acquired through our pedagogical methods. Subsequently, this study details an educational intervention where the student-developed TinySlime was utilized as a primary teaching tool in an AI module. We conducted a comparative analysis with a traditional teaching approach, assessing pedagogical impact through student surveys. Results revealed that the TinySlime-enhanced module led to significantly higher student-reported cognitive engagement, motivation, authenticity of learning, and overall satisfaction (p < .001 across all categories). Specific benefits included an enhanced understanding of LLM principles and increased confidence in applying AI skills. This research underscores the efficacy of a student-centered approach where learners become creators, and their innovations are leveraged to enrich the educational experiences of their peers, demonstrating a powerful cycle of learning and application in AI education.

Keywords: *Project-Based Learning (PBL), AI Education, Large Language Models (LLMs), Educational Technology*

Introduction

The effective cultivation of advanced technological competencies, particularly in rapidly evolving domains like Artificial Intelligence (AI), is a primary objective for contemporary engineering education. At the Department of Informatics and Mechanical Engineering, National Institute of Technology, Toba College, we employ Project Based Learning (PBL) as a foundational pedagogical strategy. This approach is significantly enhanced by actively encouraging student participation in external technology competitions, including prestigious events such as the World AI Competition YAMAGUCHI 2022 (where our students achieved 2nd Place in the U18 Division) and DCON2024. Such engagement is designed to provide students with authentic challenges, thereby fostering deep technical expertise, innovative problem-solving, and an understanding of current industry and research landscapes.

The development of Large Language Models (LLMs) represents a forefront of AI research, characterized by rapid advancements in model scale and performance (Bai & Bai, 2023; Radford & Wu, 2019), often explained by scaling laws (Kaplan & McCandlish, 2020). However, the immense computational demands and a scarcity of efficient, high-performing models for specific languages like Japanese present substantial hurdles, particularly within educational contexts with limited resources (Fujii & Nakamura, 2024). It is within this challenging yet opportunity-rich environment that our educational approach demonstrates its efficacy. As a direct outcome of their engagement in our contest-centric PBL curriculum, students at our institution developed "TinySlime," a compact Japanese language model. The technical sophistication of TinySlime serves as compelling evidence of the students' profound understanding of contemporary LLM challenges and methodologies. For instance, their work drew upon established architectures like Llama 2 (Touvron & Martin, 2023) and incorporated current techniques such as Chat Vector (Huang & Li, 2023), reflecting an awareness of advancements seen in influential small models like the Phi-series (Gunasekar & Zhang, 2023; Hughes, 2023; Abdin & Jacobs, 2024) and TinyLlama (Zhang & Zeng, 2024). Despite its modest 1.1 billion parameters and development using limited hardware (two RTX 3090s and one RTX A6000), TinySlime achieved notable performance on Japanese benchmarks (e.g., Kurihara & Kawahara, 2022; Suzuki & Suzuki, 2020; elyza, n.d.). This accomplishment, especially its ability to match or surpass larger contemporary models, underscores the high level of technical skill and current knowledge students acquired. Furthermore, their decision to make TinySlime publicly available on Hugging Face and design it for on-premise operation—thereby mitigating data leakage risks and promoting accessibility—demonstrates their practical understanding of real-world deployment considerations.

This paper, however, extends beyond documenting the development of TinySlime as an indicator of student learning. It primarily focuses on a subsequent educational initiative where this student-developed artifact was purposefully integrated as a pedagogical tool to enhance AI education for their peers. We implemented a module utilizing TinySlime as a central teaching resource and compared its educational impact against existing teaching approaches. This novel educational practice included dedicated Q&A sessions with the original student developers of TinySlime, fostering a unique peer-learning environment. To deepen the students' understanding of core concepts, such as the Transformer algorithm—foundational to models like TinySlime (related concepts in Ouyang & Wu, 2022; Schulman & Wolski, 2017; Dao & Fu, 2022)—we conducted interactive group work and hands-on workshops where students directly engaged with and manipulated the model.

Therefore, this study reports on a two-tiered educational strategy. Firstly, it presents the development of TinySlime as a testament to the effectiveness of PBL enriched by external contest participation in imbuing students with advanced, research-informed AI capabilities. Secondly, and more centrally, it evaluates the pedagogical value of leveraging such a student-created technological artifact as a tangible and relatable resource for teaching complex AI concepts to other students. Through this investigation, we aim to illuminate how integrating student-led innovation into the curriculum can create a virtuous cycle, fostering both advanced skill acquisition and enriched learning experiences within AI education.

Materials and Methods or Pedagogy

Educational Context: Fostering AI Competencies at the National Institute of Technology, Toba College

This study was conducted within the Department of Informatics and Mechanical Engineering at the National Institute of Technology, Toba College. Our curriculum emphasizes PBL as a primary pedagogical approach to cultivate practical engineering skills. A distinctive feature of our PBL strategy is the active encouragement and support for student participation in external technology and AI competitions, such as the World AI Competition YAMAGUCHI 2022 (where our students achieved 2nd Place in the U18 Division) and DCON2024. These competitions provide authentic, challenging environments that motivate students to acquire and apply cutting-edge knowledge beyond standard coursework.

Evidencing Advanced Student Capabilities: The TinySlime LLM Project

The development of "TinySlime," a 1.1 billion parameter Japanese language model, by students engaged in this PBL and contest-driven framework serves as a significant exemplar of the advanced technical competencies cultivated. This student-led project is not the primary focus of the current educational paper, but its technical execution and outcomes provide crucial evidence of the students' sophisticated understanding of contemporary LLM development challenges and state-of-the-art methodologies.

Demonstrated Understanding of LLM Development: The students embarked on this project to address the dual challenges of high computational costs associated with large LLMs (Kaplan & McCandlish, 2020) and the specific need for efficient Japanese language models. Their preliminary experiments involved a rigorous comparison of existing models like Llama 2 (Touvron & Martin, 2023), TinyLlama (Zhang & Zeng, 2024), and ELYZA-japanese-Llama-2 (elyza, n.d.), and various open-source datasets to determine optimal choices for their resource-constrained environment (Figures 1 & 2, Tables 1 & 2). This selection process itself demonstrated a mature research approach.

Technical Implementation and Optimization: For continual pre-training, students selected high-quality, well-filtered datasets (e.g., augmxnt/shisa-pretrain-en-ja-v1) and employed techniques like Data Selection for Language Models via Importance Resampling (DSIR) to maximize data diversity under computational limitations. They meticulously configured hyperparameters (e.g., micro-batch size, gradient accumulation, AdamW optimizer settings, cosine learning rate scheduler) and utilized tools like FlashAttention 2 (Dao & Fu, 2022) and DeepSpeed (Holmes & Tanaka, 2024) to enhance training efficiency and manage VRAM usage on their hardware (three GPUs equivalent to GeForce RTX 3090s).

Instruction Tuning and Chat Capabilities: To imbue TinySlime with chat capabilities, the students astutely applied the Chat Vector technique (Huang & Li, 2023) to transfer abilities from a DPO-tuned base model (TinyLlama-Chat). This choice allowed them to bypass the need for extensive instruction tuning data, effectively leveraging existing advanced models—a pragmatic solution given resource constraints. They further employed imitation learning, referencing the Phi-family of models (Gunasekar & Zhang, 2023; Hughes, 2023), using synthetic data generated by Mixtral-8x7B-v0.1.

Performance and Evaluation: The students rigorously evaluated TinySlime using established Japanese benchmarks, including the JP Language Model Evaluation Harness (focusing on JGLUE tasks like Kurihara & Kawahara, 2022; Suzuki & Suzuki, 2020) and ELYZA-task-100. The results indicated that their 1.1B parameter model performed exceptionally well, outperforming some larger models and demonstrating

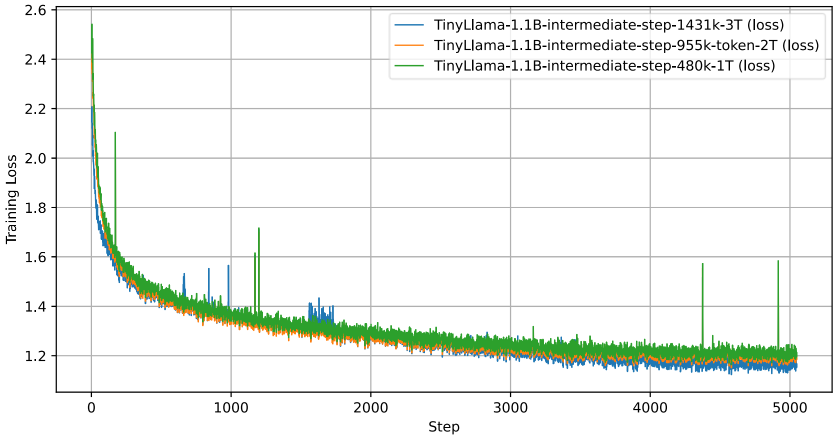

Figure.1 The training loss curves for each model are shown. From this figure, one can observe a decreasing trend in the loss during the model's training process. The lower training loss indicates that the model fits the data better, decreasing the loss as the training progresses. In particular, TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T has the lowest loss value compared to the other models, which confirms that learning is progressing.

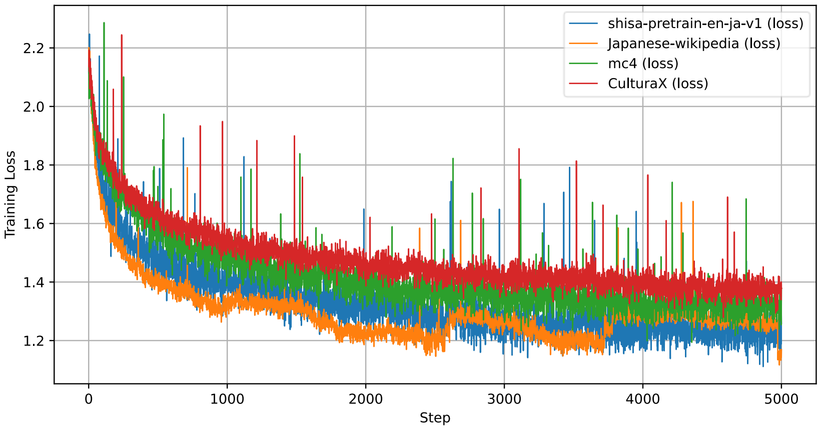

Figure.2 The training loss curves for each training dataset are shown. This figure compares the changes in losses during the model's training process using different data sets. The lower training loss indicates that the model fits the data better, and we can see how the loss decreases as the training progresses. In particular, the augmxnt/shisa-7b-v1 dataset shows the fastest loss decrease compared to the other datasets, confirming that the training progresses efficiently.

Table.1 The results of comparing the performance of different models are shown. The horizontal axis represents the different models, and the vertical axis shows the performance evaluation index. As can be seen from the table, the TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T model shows the highest performance, outperforming the other models.

base model name	**JCommonsenseQA** (3-shot)	**JNLI** (3-shot)	**MARC-ja** (0-shot)	**JSQuAD** (2-shot)	**jaqket-v2** (1-shot)	**xwinograd** (1-shot)	**mgsm** (0-shot)	**AVERAGE** (5-shot)
TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T	22.52	30.01	59.15	52.17	21.20	63.61	1.20	35.69
TinyLlama/TinyLlama-1.1B-intermediate-step-955k-token-2T	24.04	30.59	52.69	57.81	21.23	59.85	1.20	35.34
TinyLlama/TinyLlama-1.1B-intermediate-step-480k-1T	19.75	31.50	50.14	49.23	20.80	60.06	1.60	33.30

capabilities comparable to certain 7B parameter models . They also conducted safety evaluations, showing the model's capacity for ethical responses inherited via Chat Vector and learned through synthetic data, referencing concepts like RLHF (Ouyang & Wu, 2022; Rafailov & Sharma, 2023).

The student's ability to navigate these complex technical choices, optimize for limited resources, and achieve competitive performance underscores their learning depth and proficiency in state-of-the-art LLM development. The model's availability on Hugging Face and its on-premise operational capability further highlight their practical problem-solving skills, addressing issues of accessibility and data security.

Pedagogical Intervention: Utilizing Student-Developed TinySlime as a Teaching Tool

Building on the successful development of TinySlime as an outcome of our PBL and contest-driven approach, we designed a pedagogical intervention to leverage this student-created artifact as a teaching tool for a subsequent cohort of students in an AI-related module.

Participants and Design: The intervention took place within a cross-grade PBL course in the Department of Informatics and Mechanical Engineering, involving 15 student participants. A comparative approach was adopted, where one group of students (the "LLM group") experienced the TinySlime-enhanced curriculum, while a control group (the "Existing group") received instruction through our established methods for teaching similar AI concepts.

Intervention Components: The intervention for the LLM group was delivered as a structured, four-session module designed to provide concrete, hands-on engagement. Session 1 began with a Q&A session featuring TinySlime's student developers, focusing on their design choices and development challenges. This was followed by students setting up the model on their local machines. In Sessions 2 and 3, students engaged in practical, hands-on exploration by running the TinySlime model, experimenting with various prompts to understand its capabilities and limitations, and observing operational aspects like system resource usage. The module culminated in Session 4, a group discussion where students shared and compared their findings. Instead of a formal report, this session focused on a collaborative reflection on the model's unique characteristics, the challenges of LLM development, and the value of a smaller, specialized model.

To assess the impact of this intervention, a survey was administered to both the LLM group and the Existing group at the end of the module. This Data Collection and Evaluation phase utilized a questionnaire that measured students' perceptions across five key dimensions: Cognitive Engagement, Motivation, Collaboration, Authenticity of Learning Experience, and Overall Satisfaction, with items rated on a Likert scale.

Results and Discussion

Reliability of Survey Instrument

The survey instrument demonstrated good to excellent internal consistency for both the LLM and Existing method groups across all measured categories. Cronbach's alpha coefficients were as follows: Cognitive (LLM: 0.906, Existing: 0.752), Motivation (LLM: 0.934, Existing: 0.869), Collaboration (LLM: 0.931, Existing: 0.877), Authenticity (LLM: 0.961, Existing: 0.907), and Satisfaction (LLM: 0.800, Existing: 0.970). These values indicate that the scales used were reliable measures of the intended constructs.

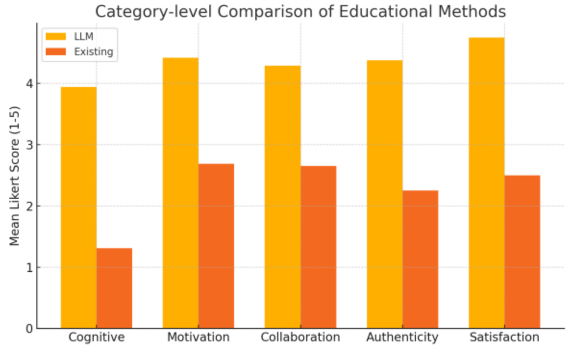

Figure 3 Category-Level Comparison Of Educational Methods

Comparative Analysis of Student Perceptions: LLM-Enhanced vs. Existing Pedagogy

The survey data revealed statistically significant and meaningful differences in student perceptions between the group taught using the TinySlime-enhanced pedagogy (LLM group) and the group taught using existing methods (Existing group).

Overall Category-Level Comparison: As illustrated in Figure 3, the LLM group reported significantly higher mean scores across all five pedagogical dimensions: Cognitive Engagement (LLM M=3.94, SD=0.96 vs. Existing M=1.31, SD=0.45; p < .001), Motivation (LLM M=4.42, SD=0.83 vs. Existing M=2.69, SD=0.99; p < .001), Collaboration (LLM M=4.29, SD=0.87 vs. Existing M=2.65, SD=1.01; p < .001), Authenticity (LLM M=4.38, SD=0.85 vs. Existing M=2.25, SD=0.98; p < .001), and Satisfaction (LLM M=4.75, SD=0.41 vs. Existing M=2.50, SD=1.05; p < .001) .

These findings suggest a strong positive impact of incorporating the student-developed LLM and associated activities into the learning experience.

Specific Item-Level Insights: Further analysis at the item level highlighted several key areas where the LLM-enhanced approach excelled. Students in the LLM group reported:

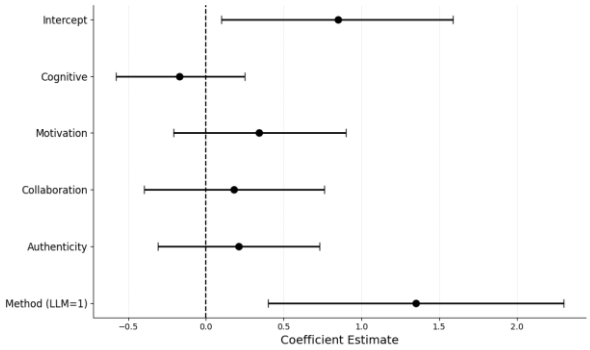

Figure 4 Regression Coefficients with 95% CI

Significantly greater understanding of LLM fundamental principles (e.g., tokenizers, attention mechanisms) after exercises with TinySlime (Cohen’s d = 2.64).

Markedly higher confidence in explaining the relationship between parameter count and model performance (Cohen’s d = 3.26) and in their ability to fine-tune an LLM in the future (Cohen’s d = 2.95).

A stronger conviction that LLM learning would be beneficial for their future careers (Cohen’s d = 1.67).

Greater enjoyment and intrinsic motivation, as indicated by higher ratings on items such as "found the experiments interesting" (Cohen’s d = 2.14) and "felt a personal desire to engage in the tasks rather than feeling compelled" (Cohen’s d = 1.61).A perception of the learning environment as more authentic and closer to real-world development practices, particularly regarding the on-premise operation of TinySlime (Cohen’s d = 2.25) and consideration of computational/data leakage risks (Cohen’s d = 2.12).

Higher overall satisfaction with the seminar and a stronger desire to participate in similar PBL and contest-integrated classes in the future (Cohen’s d = 2.83 and 2.69, respectively).

Predictors of Student Satisfaction: An Ordinary Least Squares (OLS) regression analysis was conducted to identify factors predicting overall student satisfaction

(Figure 4). The results indicated that the pedagogical method itself (Method LLM=1) was a significant positive predictor of satisfaction (Coefficient = 1.347, 95% CI [0.399, 2.296]). This suggests that, holding other factors constant, participation in the LLM-enhanced module directly contributed to higher student satisfaction. Motivation and Authenticity also showed positive trends, although their confidence intervals were wider. Because the study involved only 15 participants, statistical power and generalizability are limited.

Conclusions

This study from the National Institute of Technology, Toba College, describes our two-part educational strategy focused on student-led AI innovation. First, we found that PBL combined with participation in external technology competitions helps students develop impressive technical skills. Our students' development of the TinySlime language model, showcasing their grasp of contemporary AI challenges, advanced methodologies, and rigorous evaluation practices, stands as strong evidence of this educational outcome.

Secondly, and forming the core of this paper's investigation, the subsequent utilization of this student-developed artifact, TinySlime, as a pedagogical tool yielded significant enhancements in AI education. Students who learned with and through TinySlime reported markedly higher levels of cognitive engagement, motivation, perceived authenticity of the learning experience, and overall satisfaction compared to peers in a traditional learning environment. The hands-on interaction with a locally operable, peer-developed LLM and insights from its creators fostered a deeper understanding of complex AI concepts and boosted students' confidence in their own technological capabilities.

This research suggests that fostering a cycle where students transition from learners to creators and where their creations become valuable learning resources for others can create a uniquely effective and engaging educational ecosystem. While this study is based on a specific institutional context and a single iteration of the pedagogical intervention, the overwhelmingly positive results encourage further exploration of such student-centered, artifact-driven teaching methodologies in AI and other advanced technology domains. Future work could involve longitudinal studies to track the long-term impact on student skills and career trajectories and adapt this model to other complex engineering subjects.

Acknowledgements

We extend our sincere gratitude to the students of the Department of Informatics and Mechanical Engineering at the National Institute of Technology, Toba College, particularly the developers of TinySlime, Kaito Ogasawara and Hibiki Otsu, for their innovative work and their willingness to share their expertise with their peers. We also thank all the students who participated in the educational intervention and provided valuable feedback through the surveys.

References

Abdin, M. & Jacobs, S. A. (2024). Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2404.14219

Bai, J. & Bai, S. (2023). Qwen Technical Report. arXiv:2309.16609. Retrieved from http://arxiv.org/abs/2309.16609

Dao, T. & Fu, D. Y. (2022). FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. arXiv [cs.LG]. Retrieved from http://arxiv.org/abs/2205.14135

elyza. (n.d.). ELYZA-japanese-Llama-2-7b. Hugging Face. Retrieved June 18, 2024, from https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b

Fujii, K. & Nakamura, T. (2024). Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2404.17790

Gunasekar, S. & Zhang, Y. (2023). Textbooks Are All You Need. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2306.11644

Holmes, C. & Tanaka, M. (2024). DeepSpeed-FastGen: High-throughput Text Generation for LLMs via MII and DeepSpeed-Inference. arXiv [cs.PF]. Retrieved from http://arxiv.org/abs/2401.08671

Huang, S.-C. & Li, P.-Z. (2023). Chat Vector: A Simple Approach to Equip LLMs with Instruction Following and Model Alignment in New Languages. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2310.04799

Hughes, A. (2023). Phi-2: The surprising power of small language models. Microsoft Research. Retrieved June 18, 2024, from https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/

Kaplan, J. & McCandlish, S. (2020). Scaling Laws for Neural Language Models. arXiv [cs.LG]. Retrieved from http://arxiv.org/abs/2001.08361

Kurihara, K. & Kawahara, D. (2022). JGLUE: Japanese General Language Understanding Evaluation. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 2957-2966). Marseille, France: European Language Resources Association.

Ouyang, L. & Wu, J. (2022). Training language models to follow instructions with human feedback. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2203.02155

Radford, A. & Wu, J. (2019). Language Models are Unsupervised Multitask Learners. Retrieved June 18, 2024, from https://paperswithcode.com/paper/language-models-are-unsupervised-multitask

Rafailov, R. & Sharma, A. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv [cs.LG]. Retrieved from http://arxiv.org/abs/2305.18290

Schulman, J. & Wolski, F. (2017). Proximal Policy Optimization Algorithms. arXiv [cs.LG]. Retrieved from http://arxiv.org/abs/1707.06347

Suzuki, M. & Suzuki, J. (2020). JAQKET: クイズを題材にした日本語 QA データセットの構築. In 言語処理学会第 26 回年次大会. Retrieved from https://www.anlp.jp/proceedings/annual_meeting/2020/pdf_dir/P2-24.pdf

Touvron, H. & Martin, L. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2307.09288

Zhang, P. & Zeng, G. (2024). TinyLlama: An Open-Source Small Language Model. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2401.02385