Scalable Argumentative Writing Support: The Efficacy of Small Open-Source Models in K-12 Writing Instruction

Venkata Nagaraju Buddarapu
North Carolina State University
vbuddar@ncsu.edu

Damilola Babalola
North Carolina State University
djbabalo@ncsu.edu

Collin Lynch
North Carolina State University
cflynch@ncsu.edu

ABSTRACT

Writing is an essential skill across grade levels and disciplines. However instructors face critical challenges in providing constructive real-time feedback due to high workloads. While AI services may offer scalable writing support, institutional data privacy and costs remain significant barriers to widespread adoption. This study evaluates the efficacy of small open-source LLMs ($\leq$ 8B parameters) in generating rubric-aligned pedagogical feedback for student essays. Using the PERSUADE corpus (n=25,996), we benchmark small open-source LLMs(SLMs) including Mistral, Gemma, Phi, and LLaMA against proprietary baselines. Our results indicate that these models achieve 68–78% of proprietary performance in feedback quality. We conclude that SLMs represent a viable foundation for scalable, privacy-preserving essay instruction.

Keywords

Automated Writing Evaluation, Large Language Models, Open Source AI, Feedback Quality, Data Privacy.

1. INTRODUCTION

The integration of Artificial Intelligence in education, specifically in **Automated Writing Instruction (AWI) [12]**, offers a potentially transformative solution to the growing crisis in literacy education. Currently, K-12 settings face a dual challenge: declining student writing skills [13] and significant, widespread teacher shortages [36]. This environment has exacerbated a critical "feedback gap," where the demand for personalized, iterative guidance far outpaces the capacity of human educators to provide it.

Researchers have demonstrated that proprietary AI services such as GPT-4 [24] and Claude Sonnet [6] possess the interactive capabilities required to address these complex pedagogical tasks [2]. However, the adoption of these models is often precluded by high operational costs, reliance on external infrastructure, and legitimate data privacy concerns. Consequently, there is an urgent need for SLMs that can serve as private, resource-efficient engines for AWI, capable of providing high-order structural and content feedback at scale.

Within the domain of AWI, generating rubric-aligned feedback for argumentative essays remains a significant challenge. To be effective, models must transcend surface-level linguistic corrections and engage with the "argumentative core" of a text. We evaluate several SLMs such as Mistral, Gemma, Phi, and LLaMA on feedback generation and lead extraction using the PERSUADE corpus. Traditionally, assessing pedagogical feedback requires expert human evaluation, a process that is subjective, time-consuming, and difficult to scale. To mitigate this, we utilize GPT-4o as an automated judge, benchmarking these models against proprietary baselines such as GPT-4 and Claude 3.5 Sonnet (Sonnet 3.5) across four dimensions: specificity, constructiveness, tone, and actionability. Beyond proxy measurements, our study explores how lead extraction provides an additional performance signal for feedback quality when human-gold standards are unavailable. By analyzing the relationship between structural discourse detection accuracy and GPT-4o's feedback quality signal, we investigate whether these SLMs can bridge the feedback gap, successfully replicating the multi-dimensional guidance typically provided by human educators. While SLMs vary in tone and constructiveness, they demonstrate competitive performance in actionability. We further discuss enhancing these capabilities via LoRA [17] and ORPO [15] fine-tuning. This study addresses two primary research questions:

- **RQ1:** How effectively can SLMs replicate the feedback quality of proprietary models across key rubric dimensions?
- **RQ2:** To what extent does a model's accuracy in identifying the "argumentative core" (lead extraction) predict the quality of its generated feedback?

2. RELATED WORK

To address this pedagogical crisis, we must reconcile the reasoning capabilities of LLMs with the constraints of privacy-preserving, local deployment. Below, we review the evolution of language modeling and the current methodologies for Automated Writing Instruction (AWI) that define this research space

Proprietary vs. SLMs: While proprietary models like GPT-4

Venkata Nagaraju Buddarapu, Collin Lynch, and Damilola Babalola. Scalable Argumentative Writing Support: The Efficacy of Small Open-Source Models in K-12 Writing Instruction. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 765–769. International Educational Data Mining Society (2026).

© 2026 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.

https://doi.org/10.5281/zenodo.21040119

and Sonnet 3.5 set the state-of-the-art for reasoning and dialogue [25, 5], their adoption in K-12 is limited by high costs and data privacy risks [9]. This has catalyzed research into SLMs such as Llama-3, Mistral, Gemma, and Phi-3 [35, 19, 11, 14]. Optimization techniques like Quantization and Low-Rank Adaptation (LoRA) [16] further enable these models to match proprietary performance on domain-specific tasks while maintaining local, private deployment.

Automated Writing Instruction (AWI): AWI has transitioned from shallow, feature-based scoring (e.g., e-rater [7]) to deep generative tutoring [20, 31]. Effective AWI relies on *argument mining* the segmentation of discourse elements like leads, claims, and evidence [30, 21]. While compact models can achieve near-GPT-4 accuracy in discourse classification [27], their capacity to translate this structural detection into high-order pedagogical feedback remains an active area of investigation.

Evaluating Feedback Quality: Pedagogical feedback is multidimensional, requiring assessment across specificity, tone, and actionability [32]. Given the subjectivity and cost of human expert grading and recent work utilizes Large-Model Judges (e.g., G-Eval [22]). We utilize GPT-4 as a reference signal. We treat this as a reasonable baseline while remaining cautious regarding its direct generalization to pedagogical feedback tasks.

Despite progress in discourse classification, the pedagogical reliability of SLMs remains unproven. This study bridges this gap by evaluating a diverse roster of SLMs using the PERSUADE corpus. We investigate lead extraction as a structural proxy for feedback utility measured by specificity, constructiveness, tone, and actionability benchmarked against a GPT-4o reference judge. Isolating lead extraction provides a clean baseline for structural awareness, as effective hooks are strong predictors of essay coherence [10], while avoiding the confounding complexities of multi-span relational tasks [29].

3. METHODS

In this section, we describe the experimental workflow, which integrates model selection and prompt engineering with an evaluation framework.

3.1 Dataset and Model Selection

We utilize the PERSUADE (Persuasive Essays for Rating, Selecting, and Understanding Argumentative Discourse) corpus ($n = 25,996$) [10]. This dataset is particularly suited for our study as it contains student writing across multiple prompts and genres, with each essay featuring a gold-standard *Lead*, the foundational "hook" or thesis providing the ground truth for our structural extraction tasks. The corpus represents a diverse range of student proficiencies (Grades 6–12), as detailed in Table 1. We benchmark a roster of SLMs against proprietary baselines to evaluate the trade-offs between computational efficiency and feedback performance with model specifications outlined in Table 2.

3.2 Prompt Design:

Our pipeline utilized three distinct zero-shot prompt architectures. For *Lead Extraction*, we employed a strict *verbatim constraint*, instructing models to transcribe the exact introductory string without paraphrasing or error correction to ensure metrics reflected architectural alignment. For *Generative Feedback*, we utilized a *Role-Context-Task* framework, anchoring models as expert tutors and providing PERSUADE 2.0 rubric definitions to standardize formative output. Finally, for *Feedback Scoring*, the GPT-4o judge evaluated the student essays and generated feedback using granular Likert-scale rubrics (1–5) detailed in Section 3.3.

Table 1: PERSUADE 2.0 Corpus

Category	Details / Value
Total Essays (N)	25,996
Sample Size (n)	1,000 (Randomly Sampled)
Unique Prompts	15 (e.g., Electoral College)
Grade Levels	6–12 (Middle/High School)
Avg. Length	350–500 words
Elements	Lead, Position, Claim, Counterclaim, Rebuttal, Evidence, Conclusion

3.3 Evaluation Framework

To standardize the automated assessment, the GPT-4o judge applied a 1–5 Likert scale across four pedagogical dimensions developed in consultation with writing instructors. These definitions, detailed below, prioritize revision utility and student motivation:

Specificity: Measures the transition from vague, boilerplate phrases (Score 1) to explicit referencing or quoting of the student's unique text (Score 5).

Constructiveness: Evaluates whether the model moves beyond binary signaling (Score 1) to provide a developmental roadmap explaining the rhetorical "why" behind suggestions (Score 5).

Tone: Assesses the shift from punitive or clinical language (Score 1) to an encouraging, "growth mindset" voice (Score 5).

Actionability: Quantifies the presence of concrete next steps; a Score 5 requires a clear "revision roadmap" with at least two feasible, immediate tasks.

To measure structural identification(lead extraction), we employed two metrics: **SBERT Cosine Similarity** [28] for semantic overlap, and **BLEU** (Bilingual Evaluation Understudy) [26] to measure n-gram precision between the extracted leads and the ground truth. Overall, the reported values for both extraction metrics and feedback quality represent the mean scores calculated across the sampled essays.

3.4 Lead Extraction as a proxy

We prioritize lead extraction as the primary proxy for feedback quality indicator because the lead serves as a critical rhetorical "anchor," requiring the model to establish global context from the outset. While we plan to investigate broader discourse elements such as evidence and counterclaims in future work, the ability to craft an introductory "hook" that previews an argumentative trajectory is a hallmark of rhetorical control [8]; thus, its successful extraction

signals a model's capacity to integrate long-range dependencies and align initial framing with a global thesis.

4. RESULTS

We evaluated model performance across essays using an NVIDIA A100 GPU. The evaluation focused on two primary dimensions: **Lead Extraction** (RQ1), assessing structural identification, and **Feedback Quality** (RQ2), assessing pedagogical utility.

As shown in Table 3, SLMs specifically **LLaMA-3.1-8B** and **Phi-4-mini** surpassed GPT-4o in extraction accuracy. This suggests that structural identification is a mature capability for models $\geq$4B parameters. In the feedback phase, Phi-4-mini proved the most capable SLM, achieving $\sim$68% of the actionability scores recorded by Claude Sonnet.

While SLMs demonstrated high structural accuracy, Table 4 highlights a disparity in feedback substance. SLMs often offer correct but generic advice compared to the targeted revision steps provided by Sonnet 3.5. Furthermore, Phi models occasionally repeated student text verbatim, suggesting that while 8B models are adequate for extraction, tonal nuance and creative synthesis require further tuning.

Notably, we found a correlation (p=0.71) between lead extraction accuracy and feedback generation quality. This suggests that a model's ability to identify lead components can be an indicator of its overall proficiency in providing helpful feedback on argumentative essays such as [10].

5. DISCUSSION

This study demonstrates that compact SLMs represent a viable foundation for scalable, privacy-preserving writing instruction, achieving functional adequacy by matching $\sim$70% of proprietary model performance. While our findings underscore the viability of these models, the observed performance disparities present two critical imperatives for future research.

First, consistent with established correlations between parameter scale and the reasoning capabilities necessary for nuanced feedback [23], our analysis showed a sharp decline when moving from 4-billion to 2-billion parameters. Consequently, future work must investigate whether high-quality training data (synthetic or organic) and specialized instruction tuning (e.g., via LoRA [16] [37]) can offset these parameter deficits to preserve pedagogical efficacy on smaller hardware.

Second, as shown in Table 3, a clear performance hierarchy emerges: LLaMA-3.1-8B and Phi-4-mini dominate the open-source cohort in Lead Extraction (0.87) and Actionability (3.37), respectively, while Gemma-2B represents a performance floor with a collapse in extraction precision (0.22 Cosine). These disparities, contextualized by Table 2, raise critical questions regarding the interplay between architectural "thinness" and pedagogical utility; specifically, it remains to be determined whether the transition from Multi-Query Attention (MQA) [4] and shallow layer depths to Grouped-Query Attention (GQA) [3] and 32-layer profiles is the primary driver of a model's ability to sustain long-range dependencies in argumentative text. While SLMs nearly match proprietary benchmarks in Tone, the persistent "Actionability Gap" suggests that structural awareness alone may not suffice for complex instruction, necessitating further research into whether the observed floor in Gemma-2B is a consequence of its specific attention bottleneck or if parameter density remains the ultimate ceiling for replicating human-level revision roadmaps. Furthermore, does the 8k context window adequate for the PERSUADE corpus constrain holistic feedback for the longer, multi-document assignments prevalent in higher education?

6. LIMITATIONS

The SLMs performance reflects zero-shot capabilities rather than the upper bounds achievable through further prompt engineering and domain-specific fine-tuning. Despite the promising performance of SLMs, several constraints qualify these findings. First, our evaluation relies on the PERSUADE corpus, which consists primarily of argumentative essays from middle and high school students. Consequently, further research is required to evaluate these models across other genres such as narrative or expository writing and within university-level academic discourse to ensure the generalizability of our findings. Second, while GPT-4o serves as a functional proxy for human judgment, its efficacy must be validated in authentic classroom settings for inherit proprietary biases that favor "polished" prose, potentially overlooking subtle pedagogical nuances.

7. CONCLUSION

In conclusion, while SLMs demonstrate significant potential for providing scalable and private writing instruction, their deployment requires targeted research into architecture, datasets, and specialized fine-tuning. By bridging the performance gap between compact models and larger proprietary counterparts, we can develop accessible, high-fidelity tools that provide expert-level pedagogical support. Ultimately, these models offer a viable path toward mitigating instructor workloads while ensuring high-quality, private writing instruction is accessible at scale.

8. REFERENCES

[1] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, and Y. Zhang. Phi-4 technical report, 2024.

[2] G. Adams, A. Fabbri, F. Ladhak, E. Lehman, and N. Elhadad. From sparse to dense: Gpt-4 summarization with chain of density prompting, 2023.

[3] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Fedus, S. Sumers, and A. Deshpande. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

[4] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints, 2023.

[5] Anthropic. Claude model card, 2023.

Table 2: Model Roster and Architectural Specifications

Model	Source	Params	Layers	Heads	Attention	Context
Proprietary Baselines						
Sonnet 3.5	Anthropic	–	–	–	–	200k
GPT-4o	OpenAI	∼175B*	–	–	–	128k
SLMs						
LLaMA-3.1-8B [34]	Meta	8B	32	32	GQA	128k
Mistral-7B [18]	Mistral AI	7B	32	32	GQA/SWA	128k
Phi-4-mini [1]	Microsoft	4.2B	32	24	GQA	128k
Phi-3-mini [1]	Microsoft	3.8B	30	32	GQA	128k
Gemma-7B [33]	Google	7B	26	28	MHA	8k
Gemma-2B [33]	Google	2B	18	1	MQA	8k

Table 3: Model Performance Comparison ($n = 1,000$). Best open-source results are bolded and colored.

	Lead Extraction		Feedback Quality (1–5)			
Model	**Cosine**	**BLEU**	**Spec.**	**Constr.**	**Tone**	**Action.**
Proprietary Baselines						
Sonnet 3.5	—	—	4.98	4.94	4.65	4.97
GPT-4o	0.82	0.52	—	—	—	—
SLMs						
LLaMA-3.1-8B	**0.87**	**0.57**	3.69	3.70	4.09	3.22
Phi-4-mini	0.85	0.50	**3.89**	**3.71**	4.09	**3.37**
Phi-3-mini	0.84	0.46	3.47	3.61	4.17	3.15
Mistral-7B	0.77	0.23	3.61	3.58	4.18	3.19
Gemma-7B	0.79	0.49	3.12	3.35	4.20	2.91
Gemma-2B	0.22	0.11	2.78	3.02	3.91	2.60

Table 4: Qualitative Comparison of Generated Feedback (ID: 2241005482)

Model	Feedback Excerpt
Sonnet 3.5	"To strengthen this, cite one specific local park mentioned in your town's budget to ground your argument."
Phi-4-mini	"The essay includes a clear position. However, add more evidence to support claims about the environment."
Gemma-2B	"Good job on this essay. Try to write more sentences and check your spelling next time."

[6] Anthropic. Claude 3.5Sonnet Model Card Addendum. Technical report, Anthropic, 2024.

[7] Y. Attali and J. Burstein. Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 2006.

[8] A. Becker. Student writing in the disciplines: Graduate students' navigating the learning of local rhetorical contexts. *Journal of Writing Research*, 7(3):345–382, 2016.

[9] R. e. a. Bommasani. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[10] S. Crossley, Y. Tian, P. Baffour, A. Franklin, M. Benner, and U. Boser. A large-scale corpus for assessing written argumentation: Persuade 2.0. *Assessing Writing*, 61:100865, 2024.

[11] G. DeepMind. Gemma: Lightweight open models by google deepmind, 2024.

[12] L. T. Frase. Knowledge, information, and action: Requirements for automated writing instruction. *Journal of Computer-Based Instruction*, 11(2):55–59, 1984.

[13] A. C. Graesser and D. S. McNamara. Literacy challenges in the 21st century: A perspective on writing and reading in digital environments. *Educational Psychologist*, 56(3):134–146, 2021.

[14] S. e. a. Gunasekar. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

[15] J. Hong, N. Lee, and J. Thorne. Orpo: Monolithic preference optimization without reference model, 2024.

[16] E. e. a. Hu. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[17] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021.

[18] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril,

T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023.

[19] E. e. a. Jiang. Mistral: A sparse mixture of experts for efficient language modeling. *arXiv preprint arXiv:2310.06825*, 2023.

[20] E. Kasneci, K. Sessler, H. Krosse, and G. Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

[21] Z. e. a. Ke. Automated assessment of student writing: The persuade corpus. In *Proceedings of LREC*, 2022.

[22] J. e. a. Liu. Gpteval: Nlg evaluation using gpt-4 with better prompting and calibration. *arXiv preprint arXiv:2309.03409*, 2023.

[23] L. C. Magister, J. Mallinson, J. Adamek, E. Malmi, and A. Severyn. Teaching small language models to reason. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada, July 2023. Association for Computational Linguistics.

[24] OpenAI. Gpt-4 technical report. *OpenAI*, 2023.

[25] OpenAI, J. Achiam, S. Adler, et al. Gpt-4 technical report, 2024.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[27] S. Ramachandran, G. Ilharco, Q. Xie, S. Agrawal, K.-W. Chang, and N. A. Smith. Persuaide: Generating and evaluating personalized feedback on student persuasive writing. In *Proceedings of the 2023 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2023.

[28] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[29] C. Stab and I. Gurevych. Annotating argument components and relations in persuasive essays. In J. Tsujii and J. Hajic, editors, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland, Aug. 2014. Dublin City University and Association for Computational Linguistics.

[30] C. Stab and I. Gurevych. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In *Argument Mining Workshop*, 2017.

[31] M. Stahl, L. Biermann, A. Nehring, and H. Wachsmuth. Exploring llm prompting strategies for joint essay scoring and feedback generation, 2024.

[32] M. Stevenson. A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition*, 42:1–16, 2016.

[33] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, J. Ferret, P. Liu, P. Tafti, A. Friesen, M. Casbon, S. Ramos, R. Kumar, C. L. Lan, S. Jerome, A. Tsitsulin, N. Vieillard, P. Stanczyk, S. Girgin, N. Momchev, M. Hoffman, S. Thakoor, J.-B. Grill, B. Neyshabur, O. Bachem, A. Walton, A. Severyn, A. Parrish, A. Ahmad, A. Hutchison, A. Abdagic, A. Carl, A. Shen, A. Brock, A. Coenen, A. Laforge, A. Paterson, B. Bastian, B. Piot, B. Wu, B. Royal, C. Chen, C. Kumar, C. Perry, C. Welty, C. A. Choquette-Choo, D. Sinopalnikov, D. Weinberger, D. Vijaykumar, D. Rogozińska, D. Herbison, E. Bandy, E. Wang, E. Noland, E. Moreira, E. Senter, E. Eltyshev, F. Visin, G. Rasskin, G. Wei, G. Cameron, G. Martins, H. Hashemi, H. Klimczak-Plucińska, H. Batra, H. Dhand, I. Nardini, J. Mein, J. Zhou, J. Svensson, J. Stanway, J. Chan, J. P. Zhou, J. Carrasqueira, J. Iljazi, J. Becker, J. Fernandez, J. van Amersfoort, J. Gordon, J. Lipschultz, J. Newlan, J. yeong Ji, K. Mohamed, K. Badola, K. Black, K. Millican, K. McDonell, K. Nguyen, K. Sodhia, K. Greene, L. L. Sjoesund, L. Usui, L. Sifre, L. Heuermann, L. Lago, L. McNealus, L. B. Soares, L. Kilpatrick, L. Dixon, L. Martins, M. Reid, M. Singh, M. Iverson, M. Görner, M. Velloso, M. Wirth, M. Davidow, M. Miller, M. Rahtz, M. Watson, M. Risdal, M. Kazemi, M. Moynihan, M. Zhang, M. Kahng, M. Park, M. Rahman, M. Khatwani, N. Dao, N. Bardoliwalla, N. Devanathan, N. Dumai, N. Chauhan, O. Wahltinez, P. Botarda, P. Barnes, P. Barham, P. Michel, P. Jin, P. Georgiev, P. Culliton, P. Kuppala, R. Comanescu, R. Merhej, R. Jana, R. A. Rokni, R. Agarwal, R. Mullins, S. Saadat, S. M. Carthy, S. Cogan, S. Perrin, S. M. R. Arnold, S. Krause, S. Dai, S. Garg, S. Sheth, S. Ronstrom, S. Chan, T. Jordan, T. Yu, T. Eccles, T. Hennigan, T. Kocisky, T. Doshi, V. Jain, V. Yadav, V. Meshram, V. Dharmadhikari, W. Barkley, W. Wei, W. Ye, W. Han, W. Kwon, X. Xu, Z. Shen, Z. Gong, Z. Wei, V. Cotruta, P. Kirk, A. Rao, M. Giang, L. Peran, T. Warkentin, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, D. Sculley, J. Banks, A. Dragan, S. Petrov, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, S. Borgeaud, N. Fiedel, A. Joulin, K. Kenealy, R. Dadashi, and A. Andreev. Gemma 2: Improving open language models at a practical size, 2024.

[34] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.

[35] H. e. a. Touvron. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[36] U.S. Department of Education. Teacher Shortage Areas. `https://www.ed.gov/teaching-and-administration/professional-development/teacher-shortage-areas`, 2025. Page last reviewed February 6, 2025.

[37] T. e. a. Wang. Orpo: Offline reinforcement learning with preference optimization, 2023. OpenAI Blog.