One Textbook Is All You Need

Alyssa Sawyer*
Harvey Mudd College
asawyer@hmc.edu

Iraj Moradi*
Pitzer College
imoradi@students.pitzer.edu

Lucas Welch*
Claremont McKenna College
lwelch25@cmc.edu

Jingxiu Zhao
Scripps College
jzhao5793@scrippscollege.edu

Sophia Huang
Claremont McKenna College
shuang24@cmc.edu

Nethmin Liyanage
Pitzer College
nliyanag@students.pitzer.edu

Mike Izbicki
Claremont McKenna College
mizbicki@cmc.edu

ABSTRACT

The best language models are trained on more than 1 trillion tokens of English language text. Most languages, however, do not have such large training datasets available. We investigate an extremely data-limited regime where only 80,000 tokens of text are available in the form of a high-quality Latin textbook. We also introduce a new dataset for evaluating Latin models that contains over 5,000 high-quality human annotated questions and answers that were originally designed to assess human learning. We find that the small, high-quality textbook data is sufficient to improve the performance of language models on this new dataset.

KEYWORDS

latin, large language models, datasets, cognitive science

ACM Reference Format:
Alyssa Sawyer, Iraj Moradi, Lucas Welch, Jingxiu Zhao, Sophia Huang, Nethmin Liyanage, and Mike Izbicki. 2023. One Textbook Is All You Need. In *Proceedings of Southern California Data Science Day (KDD'23)*. ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Large Language Models don't have a strong understanding of under-represented languages due to the lack of requisite training data[12]. Previous efforts increased model performance on low-resource languages by using news[15], synthetic data[2], classical texts[3], and multilingual common crawl data[13].

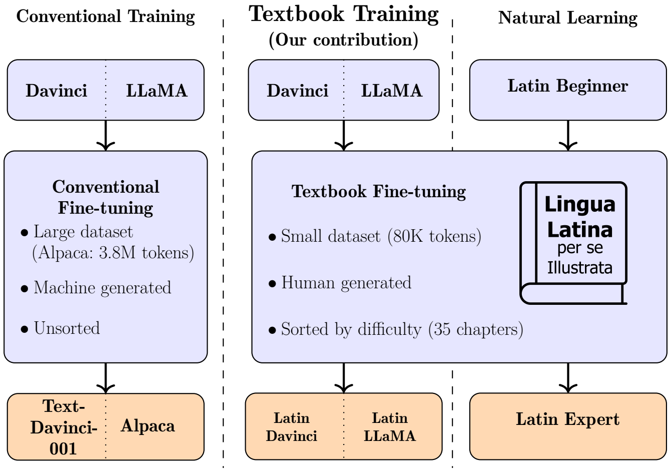

Figure 1: Our textbook training method equips LLMs to learn languages with minimal amounts of data. We compare our training process with conventional LLM fine-tuning and human learning.

Our contribution introduces the idea of *textbook training* for low-resource languages, where a small amount of high-quality text from a textbook is used in fine-tuning to improve coherency. Textbook data has been used to train an LLM from scratch[5], however, it had 80x more data compared to our singular textbook.

Our method of textbook training is inspired by the idea of *natural learning*[1], a form of example-based human learning popular with language educators. We aim to mirror human cognitive efforts to lower the computational cost of training. Figure 1 illustrates these concepts.

We fine-tune the LLaMA[19] and Davinci[4] models on a textbook to generate Latin-Davinci and Latin-LLaMA. Latin is an ideal candidate for this experiment due to its under-representation in common training datasets, as well as the high caliber of available textbooks. LLaMA has limited Latin exposure, primarily from English Wikipedia, while Davinci has slightly more.

OpenAI has published results of its models on several AP tests[16], however, other researchers have raised concerns

*Authors contributed equally to this research.

KDD'23, August 6-10, 2023, Long Beach, CA
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

about the results due to possible data contamination[9, 11, 14]. Thus, our end goal is for Latin-Davinci and Latin-LLaMA (Figure 1) to get a 5 on the AP Latin test.

2 EXPERIMENTS

We fine-tune Davinci and LLaMA on all 35 chapters of the *Lingua Latina per se Illustrata* textbook[6, 7], which we predict will enhance Latin performance efficiently with its focused content on grammar and vocabulary. This textbook follows a natural learning style, as its chapters are written solely in Latin, as a narrative. Our fine-tuning dataset size is substantially smaller than conventional, as seen in the table below.

Model Type	Model	Data Size (tokens)
Base	LLaMA-30B/60B[19]	1,400,000,000,000
Base	LLaMA-7B/13B[19]	1,000,000,000,000
Base	GPT-3[4]	300,000,000,000
Base	phi-1[5]	7,000,000,000
Base	Latin-BERT[3]	642,700,000
Fine-tuned	Alpaca[18]	3,800,000
Fine-tuned	Latin-Davinci/LLaMA	80,017

To evaluate the models' performance, we took questions from the quizzes at the end of each chapter(pensums), totaling 5,951 cloze-style questions. There is no overlap between the pre-training of LLaMA and Davinci and the testing data, as the quiz answers are neither in the textbook nor available online. Pensum style A asks the student to fill in a missing word ending, while pensum style B asks the student to fill in an entire word. Below shows a pensum A task, where the missing word ending, colored in red, needs to be correctly identified.

Q: `Iūlius pater Mārc~ est.`

To score the model, for a given question we gave it multiple versions of the quiz sentence with four different replacements for a missing word or word ending. We identified the model's choice by whichever sentence had the lowest perplexity[10], an evaluation strategy also used in an LLM benchmark on Mandarin Chinese[17].

Eval Sentences	Davinci		LLaMA-13B	
	Base	Fine	Base	Fine
`Iūlius pater Mārcus est.`	**12.54**	2.96	**121.15**	11.86
`Iūlius pater Mārcum est.`	22.63	2.81	217.30	13.39
`Iūlius pater Mārcī* est.`	14.91	**1.44**	198.13	**7.46**
`Iūlius pater Mārcōrum est.`	16.35	2.88	123.84	10.46

The table above illustrates how the fine-tuned models identify the correct version of Marcus; in Latin there are many forms of nouns. The base models' failure in this specific example could be attributed to how in English, the most common borrowed word ending from Latin is the nominative case (the `-us` ending).

Latin's macron system, the horizontal lines above some vowels, can alter the meaning of a sentence. Experiments fine-tuning on chapters 1 to 5 with and without macrons yielded similar results. To save compute we used the text without macrons, as they increase token count.

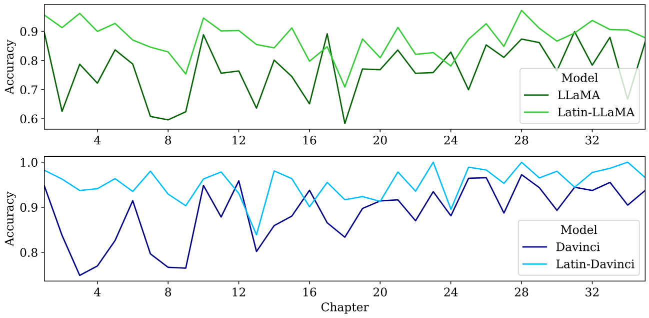

Figure 2: Fine-tuning Davinci and LLaMA on chapters 1-35 increases performance across multiple quiz styles. Model performance is variable across chapters due to differing content.

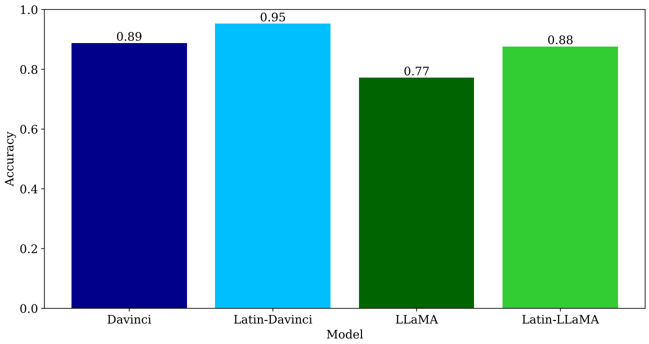

Figure 3: Fine-tuning models with textbook training outperforms the original models across all quiz styles and 35 chapters.

As shown in Figures 2 and 3, Latin-Davinci and Latin-LLaMA outperform the base models. For LLaMA fine-tuning, we used LoRA[8], showing that computationally efficient methods of fine-tuning can improve language performance.

We limited the number of answers to four. It significantly decreased the cost and time of the evaluation, reducing OpenAI API costs by more than 70x. The later chapters should have lower performance, as increasingly complex Latin grammar concepts are being tested. However, the multiple choice nature of the evaluation along with more context clues leads to these later chapters being inflated in accuracy. As our preliminary results are promising, we plan to implement other evaluation techniques in order to more accurately address the free-response nature of these questions.

REFERENCES

[1] J Scott Armstrong. 2010. Natural learning in higher education. *Available at SSRN 1928831* (2010).

[2] Abhijeet Awasthi, Nitish Gupta, Bidisha Samanta, Shachi Dave, Sunita Sarawagi, and Partha Pratim Talukdar. 2023. Bootstrapping Multilingual Semantic Parsers using Large Language Models. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

[3] David Bamman and Patrick J. Burns. 2020. Latin BERT: A Contextual Language Model for Classical Philology. arXiv:2009.10053 [cs.CL]

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[5] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks Are All You Need. arXiv:2306.11644 [cs.CL]

[6] Ørberg Hans, H. 2005. *Exercitia Latina I: Exercises for Familia Romana (Lingua Latina) (Latin Edition)* (paperback ed.). Focus. 144 pages. https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=1585102121

[7] Ørberg Hans, H. 2011. *Lingua Latina per se Illustrata, Pars I: Familia Romana (Latin Edition)* (paperback ed.). Focus. 328 pages. https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=1585104205

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[9] Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop Uploading Test Data in Plain Text: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks. arXiv:2305.10160 [cs.CL]

[10] Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America* 62, S1 (1977), S63–S63.

[11] Benjamin Marie. 2023. The Decontaminated Evaluation of GPT-4. https://benjaminmarie.com/the-decontaminated-evaluation-of-gpt-4/

[12] Nikita Moghe, Evgeniia Razumovskaia, Liane Guillou, Ivan Vulić, Anna Korhonen, and Alexandra Birch. 2022. MULTI3NLU++: A Multilingual, Multi-Intent, Multi-Domain Dataset for Natural Language Understanding in Task-Oriented Dialogue. *arXiv preprint arXiv:2212.10455* (2022).

[13] Amirkeivan Mohtashami, Mauro Verzetti, and Paul K. Rubenstein. 2023. Learning Translation Quality Evaluation on Low Resource Languages from Large Language Models. arXiv:2302.03491 [cs.CL]

[14] Arvind Narayanan and Sayash Kapoor. 2023. GPT-4 and professional benchmarks: the wrong answer to the wrong question. https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks

[15] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 116–126. https://doi.org/10.18653/v1/2021.mrl-1.11

[16] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[17] Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino Linguistic Evaluation of Large Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4606–4634. https://aclanthology.org/2022.emnlp-main.305

[18] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

[19] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]