Optimization, Adaptation and Applications of Large Language Models

David Ponce

Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián (Spain)
University of the Basque Country (EHU), Faculty of Informatics, Manuel Lardizabal pasealekua 1, 20018 Donostia-San Sebastián (Spain)

Abstract

The exponential growth in the size of Large Language Models (LLMs) has led to a paradigm shift in Natural Language Processing (NLP), demonstrating unprecedented capabilities across diverse natural language understanding and generation tasks. Despite their remarkable performance, these models present substantial computational and environmental challenges due to their massive parameter counts, requiring significant resources for both training and deployment phases. This doctoral thesis aims to explore three complementary aspects of LLMs: (i) model compression and optimization; (ii) parameter-efficient adaptation for down-streaming tasks; and (iii) exploring the application of language models across a diverse spectrum of NLP tasks, with particular emphasis on leveraging efficient architectures such as Small Language Models (SLMs). This work aims to establish optimal trade-offs between model performance and computational efficiency, thereby contributing to the development of more accessible and environmentally sustainable language technologies without compromising task-specific efficacy.

Keywords

Efficient Large Language Models, Small Language Models, Parameter-Efficient Adaptation

1. Justification of the proposed research

Large Language Models based on the Transformer architecture have significantly influenced the field of Natural Language Processing, demonstrating notable capabilities across various tasks. The landscape of these models continues to evolve, with developments from both commercial and research organizations. Meta's Llama family [1] (now in its third generation with models ranging from 8B to 70B parameters), DeepSeek-R1 models [2] (671B parameters), and the EU-supported EuroLLM initiative [3] (developing models specifically for European languages) represent substantial investments in language model technology. In parallel, efforts to address linguistic diversity have resulted in models like ALIA [4] (focused on Spanish and other Iberian languages) and Latxa [5] (targeting the Basque language).

Despite their technological achievements, the scaling trajectory of these models presents substantial challenges. The computational requirements to train and adapt LLMs have reached unprecedented levels, with associated environmental impacts that raise serious sustainability concerns. A single training run for a large-scale model can generate carbon emissions equivalent to five times the lifetime emissions of an average car [6]. Furthermore, the infrastructure required for both training and deployment effectively restricts advanced language technology development and limits its application in industrial settings where computational resources are constrained, particularly for small and medium companies that cannot access high-performance computing clusters or afford the operational costs associated with large model inference.

In response to these challenges, research interest has shifted toward Small Language Models as viable alternatives to their resource-intensive counterparts. Recent developments in this domain include Microsoft's Phi series [7], Huggingfaces's SmolLM [8] from 135M up to 1.7B parameters, and compact variants of established models such as Llama3.2 (1B/3B) [1] and EuroLLM (1.7B) [3]. These compact

Doctoral Symposium on Natural Language Processing, 25 September 2025, Zaragoza, Spain.
adponce@vicomtech.org (D. Ponce)
https://zolastro.github.io/ (D. Ponce)
0009-0008-7434-5868 (D. Ponce)

models aim to provide alternatives for deployment in resource-constrained environments, though they typically demonstrate lower performance than their larger counterparts, particularly on complex reasoning tasks.

Various efficiency techniques have emerged to mitigate the computational demands of adapting language models to specific tasks or domains. Parameter-efficient methods like Low-Rank Adaptation (LoRA) [9] enable model specialization with minimal computational overhead by introducing small trainable components while keeping most parameters frozen. Quantization reduces memory requirements by representing weights with fewer bits [10], while knowledge distillation [11] transfers capabilities from larger models to more compact architectures. Pruning eliminates redundant parameters, creating sparser networks that maintain performance with reduced computational needs [12]. Training optimizations including mixed precision training [13] and curriculum learning [14] further reduce resource requirements while potentially improving model quality. These approaches collectively offer practical solutions for deploying advanced language technologies in resource-constrained environments without compromising essential capabilities.

This doctoral thesis investigates optimization, adaptation, and application for language models with the objective of enabling their effective deployment in computationally constrained environments. This research aims to identify practical trade-offs between model performance and computational efficiency. The work focuses on three complementary approaches: (i) optimization of training methodologies through data curation and architectural refinements; (ii) parameter-efficient adaptation techniques that minimize computational overhead during task specialization; (iii) exploring the application of language models across a diverse spectrum of NLP tasks, with particular emphasis on leveraging efficient architectures and compact models, which facilitate the practical deployment of these technologies in industrial settings where computational resources are limited. These approaches address the need for more accessible and sustainable language model technologies that maintain adequate performance while reducing environmental impact.

2. Background

2.1. Training Optimization

The training phase represents the most computationally intensive component in the lifecycle of large language models, with contemporary models requiring staggering amounts of computation. Substantial research efforts have focused on reducing training duration without compromising model quality through data-centric methods and computational efficiency techniques.

Data curation has emerged as a critical determinant of model efficiency. Dodge et al. [15] demonstrated that careful data selection yields models outperforming those trained on substantially larger but less refined datasets. Gunasekar et al. [16] showed models trained on high-quality "textbook-like" data achieve performance comparable to those trained on vastly larger web-scraped corpora. Recent contributions from Allal et al. [17] demonstrated benefits in mixing textual data with code and mathematical content when training SLMs, addressing their heightened sensitivity to data noise [18, 7]. Curriculum learning [14] presents examples to models in a structured progression from simple to complex instances, demonstrating improved convergence rates across multiple domains.

Mixed Precision Training [13] uses lower precision numerical formats for most computational operations while selectively using higher precision for numerically critical operations, typically yielding 2-3x throughput improvements while maintaining model convergence through techniques like loss scaling.

2.2. Inference Optimization

Inference efficiency is critical for practical deployment scenarios, particularly for latency-sensitive applications. Key approaches include structural modifications and representational optimizations.

Model pruning eliminates redundant parameters according to various saliency criteria. In transformer-based models, structured pruning has shown particular efficacy demonstrating that up to 50% of attention heads can be removed with minimal performance degradation [12]. Muralidharan et al. [19] advanced this field with a structured approach involving cycles of pruning, knowledge transfer, and weight adjustment. Layer Collapse [20] enables model size reduction by collapsing rear layers into prior ones while preserving model structure, maintaining over 80% of task performance at 25-30% pruning ratios and outperforming existing structured pruning methods.

Low-rank factorization decomposes weight matrices into products of smaller matrices, exploiting the inherent low-rank nature of neural network parameters. Wang et al. [21] demonstrated that attention matrices can be effectively approximated through such decompositions, reducing computational requirements and memory footprint.

Knowledge distillation [11] transfers information from a large "teacher" model to a compact "student" model. Gu et al. [22] showed this can preserve much of the in-context learning capabilities of LLMs in smaller models. Recent implementations in Gemma-2 [23] and LaMini-GPT [24] have applied advanced distillation variants for resource-constrained environments.

Quantization reduces parameter and activation precision from training standards to lower bit-width formats. Recent developments have focused on mixed-precision approaches, enabling sub-8-bit quantization with minimal accuracy impact. Models like Qwen [25] and StableLM [26] demonstrate quantization's effectiveness.

2.3. Parameter-Efficient Adaptation Techniques

Parameter-Efficient Fine-Tuning (PEFT) methodologies modify only a small subset of parameters while maintaining comparable performance to full fine-tuning, addressing the prohibitive costs of conventional approaches for LLMs.

Adapter-based methods incorporate specialized modules that compress and expand internal representations. During adaptation, only these modules are trained while the base model remains frozen, reducing trainable parameters by 95-99%. Strategic placement of adapters has achieved near full fine-tuning performance with as few as 0.1% of trainable parameters [9]. Prompt-tuning [27] modifies input representation space rather than internal model parameters, prepending optimizable continuous vectors to input embeddings. This approach shows scaling properties where performance approaches full fine-tuning as model size increases. Prefix-tuning [28] generalizes prompt tuning by incorporating optimizable vectors at each transformer layer, enabling more expressive adaptation while maintaining parameter efficiency. Low-Rank Adaptation [9] approximates weight updates through low-rank decompositions, significantly reducing memory requirements. Recent extensions include QLoRA [29], combining quantization with LoRA, and AdaLoRA [30], optimizing rank allocation across model components. LOMO [31] offers an alternative approach for full parameter fine-tuning with limited resources.

Additionally, Retrieval-Augmented Generation (RAG) offers a complementary approach that enhances model capabilities by retrieving relevant information from external knowledge sources without modifying model parameters, enabling domain adaptation through contextualization rather than fine-tuning [32].

2.4. Architecture Optimization and Small Language Models

Studies have revealed substantial redundancy in pretrained transformers, showing approximately 80% of attention heads can be removed with minimal performance impact [33]. Gromov et al. [34] demonstrated the high modeling capacity of deeper layers in generative language models, exploiting this through layer pruning and fine-tuning. Small Language Models challenge the "bigger is better" narrative. Schick and Schütze [35] showed relatively small models (60-350M parameters) could achieve competitive few-shot learning performance through carefully constructed prompting. Models like TinyLlama [36], MobileLLaMA [37], and Phi-4 [7] integrate various techniques optimizing neural networks while

limiting quality losses, demonstrating SLMs can handle specific tasks and exhibit emergent capabilities similar to larger models.

Alternative transformer formulations like Performer [38] and Linear Transformer [39] replace quadratic-complexity attention with linear approximations. Other optimization advances include Multi-Query Attention [40], Group-Query Attention [41], and FlashAttention [42], optimizing memory and inference speed through improved data access. The integration of architectural innovations with efficient adaptation techniques offers a promising direction, potentially delivering order-of-magnitude efficiency improvements compared to conventional methodologies.

Parallel to transformer optimization efforts, researchers have investigated fundamentally different architectural paradigms. Recurrent RWKV [43] combines RNN-style sequential processing with transformer-like parallelization, achieving linear scaling with sequence length and constant memory usage during inference. State Space Models such as Mamba [44] have demonstrated exceptional performance on long-sequence tasks while maintaining linear computational complexity through selective scanning mechanisms that efficiently capture long-range dependencies. Diffusion-based language models, exemplified by LLaDa [45], represent another promising research direction that adapts iterative denoising frameworks from computer vision to text generation, offering unique advantages in controllable text synthesis and generation diversity. These architectural alternatives complement transformer optimization approaches by addressing fundamental efficiency limitations through novel computational paradigms rather than parameter reduction alone.

3. Description of the proposed research and hypotheses

For the thesis described in this document, the following main objective has been defined: Research and development of optimization, adaptation, and application strategies for Large Language Models that establish optimal trade-offs between performance and computational efficiency.

3.1. Objectives

In the framework of this thesis, optimization techniques applied to LLMs will be investigated and compared with the aim of reducing the hardware requirements and computational resources demanded by these models during the training and inference phases. Therefore, the work will focus on efficient training and adaptation methods, as well as optimization and compression techniques for large language models, in order to deploy these models in a realistic production environment.

To meet the objective, the following tasks will be undertaken:

- **Optimization for the training and adaptation of large language models**: Given the extensive computational infrastructure resources necessary for the training and use of high-quality large language models, lower-cost alternatives will be investigated at different levels: (i) optimization of training data, based on data selection methods, Curriculum Learning, and preprocessing variants (word segmentation, casing, etc.); and (ii) architecture optimization, based on variants of the standard Transformer architecture in particular.
- **Efficient adaptation methods**: Different techniques will be explored, such as adaptation through fine-tuning, which involve adjusting the weights of neural networks in language models, and methods that only require minimal additional weight adjustments, without the need to adjust the base network weights completely, based on methods like prefix-learning, LoRA, or adapter-tuning.
- **Efficient deployment and inference of large language models**: Compression techniques for large language models will be investigated, such as pruning, distillation, or quantization, with the aim of deploying these models in limited computational environments.
- **Evaluation of language model capabilities across functional domains**: This work will assess how language models perform across varied NLP applications, examining their effectiveness in different contexts such as conversational AI, machine translation, text simplification and other specialized domain tasks while focusing on models optimized for computational efficiency.

3.2. Hypothesis

This thesis is built upon the following hypotheses:

- **Efficiency-Performance Trade-off**: Small Language Models (1-5B parameters) optimized through selected compression and adaptation techniques can achieve performance comparable to much larger models on specific NLP tasks while requiring fewer computational resources.
- **Architecture Optimality**: The redundancy in standard transformer architectures can be systematically identified and eliminated, resulting in models with fewer parameters that may preserve nearly all of the original performance across common NLP benchmarks.
- **Data Leverage**: Carefully curated training and fine-tuning datasets can compensate for reduced model size, enabling more efficient models to match or exceed the performance of larger models trained on noisy or unfiltered corpora.

4. Methodology and Research Progress

4.1. Research Methodology

The research methodology adopts a multifaceted approach to investigate efficiency-performance trade-offs in language model optimization and deployment:

- **Continuous State-of-the-Art Literature Analysis**: A systematic and ongoing review of emerging research constitutes a foundational component of the methodology. This continuous monitoring is crucial given the rapidly evolving developments in efficient language modeling.
- **Cross-Domain Task Evaluation**: Language models will be systematically evaluated across diverse NLP tasks using established benchmarks for fair comparisons and reproducibility. This approach enables the identification of domains where optimized smaller models demonstrate competitive performance relative to their larger counterparts.
- **Parameter-Efficient Adaptation Exploration**: The comparative efficacy of adaptation techniques will be assessed for domain specialization. Additionally, Retrieval-Augmented Generation approaches will be investigated as complementary methods that enhance model capabilities without parameter modification.
- **Resource-Aware Compression Investigation**: Model compression techniques will be explored with consideration of available computational resources. Compression approaches will be assessed through comparative experiments measuring both task performance preservation and computational efficiency gains
- **Alternative Architectures**: Investigation of alternative architectures to the Transformer paradigm alongside automated architecture search methodologies to identify more efficient structural configurations.

4.2. Research Progress and Preliminary Findings

Progress in this research has resulted in several publications in conferences in the field:

- **Unsupervised Subtitle Segmentation with Masked Language Models** [46]: An unsupervised approach to subtitle segmentation using pretrained masked language models, predicting line endings and subtitle breaks based on punctuation likelihood. The method achieves competitive segmentation accuracy while preserving original text and complying with length constraints, validating that efficient masked language models could perform specialized text processing tasks without requiring large-scale generative models or supervised fine-tuning. This work was presented in the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023).

- **Split and Rephrase with Large Language Models** [47]: An evaluation of large language models on the task of splitting complex sentences into shorter grammatical ones while preserving meaning. The study includes prompting variants, domain shift analysis, and comparison of fine-tuned models with zero-shot and few-shot approaches, showing significant improvements over previous state-of-the-art with relatively small models and training datasets, while revealing that sentence splitting remains a challenging task even for large models. This research was presented in the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024).
- **Vicomtech@WMT 2024: Shared Task on Translation into Low-Resource Languages of Spain** [48]: Participation in WMT 2024 Shared Task addressing translation into Aragonese, Aranese, and Asturian. This study notably demonstrated that smaller, specialized models could compete with LLMs on low-resource translation tasks, while also proposing the application of LLMs for backtranslation generation to improve training data for smaller models. The results were presented in the Proceedings of the Ninth Conference on Machine Translation (WMT 2024).
- **Automating Easy Read Text Segmentation** [49]: An investigation of methods for automating Easy Read text segmentation, including masked and generative language models and constituent parsing. The study includes automatic and human evaluations in three languages, analyzing strengths and weaknesses of proposed alternatives under resource limitations. The study demonstrated that smaller encoder-only models consistently surpassed the quality of generative decoder-only models while significantly reducing the risk of hallucinations. This research was presented in the Findings of the Association for Computational Linguistics: EMNLP 2024.

5. Research Elements Proposed for Discussion

The research presented in this thesis raises several questions that would benefit from scholarly discussion and expert feedback at the symposium:

- **Parameter Redundancy and Efficient Architecture Design**: While empirical evidence demonstrates the redundancy of parameters and layers through successful application of pruning and other compression techniques, challenges remain in training efficient models from scratch without this redundancy. This raises questions about whether the redundancy is necessary for the training process itself, or if alternative architectural designs and training methodologies could yield inherently more efficient models.
- **Knowledge Preservation Metrics in Model Compression**: Standard metrics for model compression evaluation such as KL divergence of vocabulary distributions or cosine similarity of hidden representations may not fully capture a model's capabilities. Models compressed following these metrics sometimes demonstrate unexpected performance divergences on knowledge-intensive tasks. This suggests the need for more comprehensive evaluation metrics that accurately measure preservation of different types of capabilities during the compression process.
- **Fine-tuning Stability of Instruction-Tuned Models**: Fine-tuning models tuned for instruction following has shown to be particularly sensitive, sometimes leading to a decline in generalization quality. This phenomenon raises questions about optimal adaptation strategies for these aligned models, including the appropriate volume and diversity of fine-tuning data, suitable learning rates, and mechanisms to preserve general capabilities while enhancing domain-specific performance.

6. Conclusions

This thesis aims to advance the field of Natural Language Processing by investigating optimization, adaptation, and application strategies for Large Language Models that balance performance with computational efficiency. By exploring compression techniques, parameter-efficient fine-tuning methods, and architectural modifications, the research seeks to address growing concerns regarding computational requirements and the environmental impact of increasingly large models.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[2] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, arXiv preprint arXiv:2501.12948 (2025).

[3] P. H. Martins, P. Fernandes, J. Alves, N. M. Guerreiro, R. Rei, D. M. Alves, J. Pombal, A. Farajian, M. Faysse, M. Klimaszewski, P. Colombo, B. Haddow, J. G. de Souza, A. Birch, A. F. Martins, Eurollm: Multilingual language models for europe, Procedia Computer Science 255 (2025) 53–62. URL: https://www.sciencedirect.com/science/article/pii/S1877050925006210. doi:https://doi.org/10.1016/j.procs.2025.02.260, proceedings of the Second EuroHPC user day.

[4] A. Gonzalez-Agirre, M. Pàmies, J. Llop, I. Baucells, S. D. Dalt, D. Tamayo, J. J. Saiz, F. Espuña, J. Prats, J. Aula-Blasco, M. Mina, A. Rubio, A. Shvets, A. Sallés, I. Lacunza, I. Pikabea, J. Palomar, J. Falcão, L. Tormo, L. Vasquez-Reina, M. Marimon, V. Ruíz-Fernández, M. Villegas, Salamandra technical report, 2025. URL: https://arxiv.org/abs/2502.08489. arXiv:2502.08489.

[5] J. Etxaniz, O. Sainz, N. Miguel, I. Aldabe, G. Rigau, E. Agirre, A. Ormazabal, M. Artetxe, A. Soroa, Latxa: An open language model and evaluation suite for Basque, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 14952–14972. URL: https://aclanthology.org/2024.acl-long.799/. doi:10.18653/v1/2024.acl-long.799.

[6] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3645–3650. URL: https://aclanthology.org/P19-1355/. doi:10.18653/v1/P19-1355.

[7] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, et al., Phi-3 technical report: A highly capable language model locally on your phone, arXiv preprint arXiv:2404.14219 (2024).

[8] L. B. Allal, A. Lozhkov, E. Bakouch, L. von Werra, T. Wolf, Smollm - blazingly fast and remarkably powerful, 2024.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., ICLR 1 (2022) 3.

[10] A. Bhandare, V. Sripathi, D. Karkada, V. Menon, S. Choi, K. Datta, V. Saletore, Efficient 8-bit quantization of transformer neural machine language translation model, arXiv preprint arXiv:1906.00532 (2019).

[11] G. E. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, CoRR abs/1503.02531 (2015). URL: http://arxiv.org/abs/1503.02531. arXiv:1503.02531.

[12] H. Sajjad, F. Dalvi, N. Durrani, P. Nakov, On the effect of dropping layers of pre-trained transformer models, Computer Speech & Language 77 (2023) 101429.

[13] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, et al., Mixed precision training, arXiv preprint arXiv:1710.03740 (2017).

[14] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th annual international conference on machine learning, 2009, pp. 41–48.

[15] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, N. Smith, Fine-tuning pretrained

language models: Weight initializations, data orders, and early stopping, arXiv preprint arXiv:2002.06305 (2020).
[16] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, et al., Textbooks are all you need, arXiv preprint arXiv:2306.11644 (2023).
[17] L. B. Allal, A. Lozhkov, E. Bakouch, G. M. Blázquez, G. Penedo, L. Tunstall, A. Marafioti, H. Kydlíček, A. P. Lajarín, V. Srivastav, et al., Smollm2: When smol goes big–data-centric training of a small language model, arXiv preprint arXiv:2502.02737 (2025).
[18] D. Rolnick, A. Veit, S. Belongie, N. Shavit, Deep learning is robust to massive label noise, arXiv preprint arXiv:1705.10694 (2017).
[19] S. Muralidharan, S. Turuvekere Sreenivas, R. Joshi, M. Chochowski, M. Patwary, M. Shoeybi, B. Catanzaro, J. Kautz, P. Molchanov, Compact language models via pruning and knowledge distillation, Advances in Neural Information Processing Systems 37 (2024) 41076–41102.
[20] Y. Yang, Z. Cao, H. Zhao, Laco: Large language model pruning via layer collapse, arXiv preprint arXiv:2402.11187 (2024).
[21] S. Wang, B. Z. Li, M. Khabsa, H. Fang, H. Ma, Linformer: Self-attention with linear complexity, arXiv preprint arXiv:2006.04768 (2020).
[22] Y. Gu, L. Dong, F. Wei, M. Huang, MiniLLM: Knowledge distillation of large language models, in: The Twelfth International Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=5h0qf7IBZZ.
[23] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al., Gemma 2: Improving open language models at a practical size, arXiv preprint arXiv:2408.00118 (2024).
[24] M. Wu, A. Waheed, C. Zhang, M. Abdul-Mageed, A. F. Aji, Lamini-lm: A diverse herd of distilled models from large-scale instructions, arXiv preprint arXiv:2304.14402 (2023).
[25] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, et al., Qwen2. 5 technical report, arXiv preprint arXiv:2412.15115 (2024).
[26] M. Bellagente, J. Tow, D. Mahan, D. Phung, M. Zhuravinskyi, R. Adithyan, J. Baicoianu, B. Brooks, N. Cooper, A. Datta, et al., Stable lm 2 1.6 b technical report, arXiv preprint arXiv:2402.17834 (2024).
[27] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, arXiv preprint arXiv:2104.08691 (2021).
[28] X. L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, arXiv preprint arXiv:2101.00190 (2021).
[29] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, Advances in neural information processing systems 36 (2023) 10088–10115.
[30] Q. Zhang, M. Chen, A. Bukharin, N. Karampatziakis, P. He, Y. Cheng, W. Chen, T. Zhao, Adalora: Adaptive budget allocation for parameter-efficient fine-tuning, arXiv preprint arXiv:2303.10512 (2023).
[31] K. Lv, Y. Yang, T. Liu, Q. Guo, X. Qiu, Full parameter fine-tuning for large language models with limited resources, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 8187–8198. URL: https://aclanthology.org/2024.acl-long.445/. doi:10.18653/v1/2024.acl-long.445.
[32] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.
[33] F. Dalvi, N. Durrani, H. Sajjad, Y. Belinkov, A. Bau, J. Glass, What is one grain of sand in the desert? analyzing individual neurons in deep nlp models, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 6309–6317.
[34] A. Gromov, K. Tirumala, H. Shapourian, P. Glorioso, D. Roberts, The unreasonable ineffectiveness of the deeper layers, in: The Thirteenth International Conference on Learning Representations,

2025. URL: https://openreview.net/forum?id=ngmEcEer8a.

[35] T. Schick, H. Schütze, It's not just size that matters: Small language models are also few-shot learners, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2339–2352. URL: https://aclanthology.org/2021.naacl-main.185/. doi:10.18653/v1/2021.naacl-main.185.

[36] P. Zhang, G. Zeng, T. Wang, W. Lu, Tinyllama: An open-source small language model, arXiv preprint arXiv:2401.02385 (2024).

[37] X. Chu, L. Qiao, X. Lin, S. Xu, Y. Yang, Y. Hu, F. Wei, X. Zhang, B. Zhang, X. Wei, et al., Mobilevlm: A fast, strong and open vision language assistant for mobile devices, arXiv preprint arXiv:2312.16886 (2023).

[38] K. M. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Q. Davis, A. Mohiuddin, L. Kaiser, D. B. Belanger, L. J. Colwell, A. Weller, Rethinking attention with performers, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=Ua6zuk0WRH.

[39] A. Katharopoulos, A. Vyas, N. Pappas, F. Fleuret, Transformers are rnns: Fast autoregressive transformers with linear attention, in: International conference on machine learning, PMLR, 2020, pp. 5156–5165.

[40] N. Shazeer, Fast transformer decoding: One write-head is all you need, arXiv preprint arXiv:1911.02150 (2019).

[41] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebron, S. Sanghai, GQA: Training generalized multi-query transformer models from multi-head checkpoints, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 4895–4901. URL: https://aclanthology.org/2023.emnlp-main.298/. doi:10.18653/v1/2023.emnlp-main.298.

[42] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, C. Re, Flashattention: Fast and memory-efficient exact attention with IO-awareness, in: A. H. Oh, A. Agarwal, D. Belgrave, K. Cho (Eds.), Advances in Neural Information Processing Systems, 2022. URL: https://openreview.net/forum?id=H4DqfPSibmx.

[43] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. Chung, L. Derczynski, X. Du, M. Grella, K. Gv, X. He, H. Hou, P. Kazienko, J. Kocon, J. Kong, B. Koptyra, H. Lau, J. Lin, K. S. I. Mantri, F. Mom, A. Saito, G. Song, X. Tang, J. Wind, S. Woźniak, Z. Zhang, Q. Zhou, J. Zhu, R.-J. Zhu, RWKV: Reinventing RNNs for the transformer era, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 14048–14077. URL: https://aclanthology.org/2023.findings-emnlp.936/. doi:10.18653/v1/2023.findings-emnlp.936.

[44] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, in: First Conference on Language Modeling, 2024. URL: https://openreview.net/forum?id=tEYskw1VY2.

[45] S. Nie, F. Zhu, Z. You, X. Zhang, J. Ou, J. Hu, J. Zhou, Y. Lin, J.-R. Wen, C. Li, Large language diffusion models, arXiv preprint arXiv:2502.09992 (2025).

[46] D. Ponce, T. Etchegoyhen, V. Ruiz, Unsupervised subtitle segmentation with masked language models, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 771–781. URL: https://aclanthology.org/2023.acl-short.67/. doi:10.18653/v1/2023.acl-short.67.

[47] D. Ponce, T. Etchegoyhen, J. Calleja, H. Gete, Split and rephrase with large language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 11588–11607. URL: https://aclanthology.org/2024.acl-long.622/. doi:10.18653/v1/2024.acl-long.622.

[48] D. Ponce, H. Gete, T. Etchegoyhen, Vicomtech@WMT 2024: Shared task on translation into low-resource languages of Spain, in: B. Haddow, T. Kocmi, P. Koehn, C. Monz (Eds.), Proceedings

of the Ninth Conference on Machine Translation, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 934–942. URL: https://aclanthology.org/2024.wmt-1.91/. doi:10.18653/v1/2024.wmt-1.91.

[49] J. Calleja, T. Etchegoyhen, D. Ponce, Automating easy read text segmentation, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 11876–11894. URL: https://aclanthology.org/2024.findings-emnlp.694/. doi:10.18653/v1/2024.findings-emnlp.694.