Designing Human-Centric Foundation Models

Narendra Patwardhan[1,2,*], Shreya Shetye[2], Lidia Marassi[1], Monica Zuccarini[1], Tannistha Maiti[2] and Tarry Singh[2]

[1]University of Naples Federico II, Via Claudio 21, 80125 Naples, Italy
[2]Deepkapha AI Research, Street Vaart ZZ n° 1.d,9401GE, Assen, The Netherlands

Abstract
In recent years, generative AI has made remarkable strides in algorithmic innovation, enhancing zero-shot generalization for downstream tasks. However, challenges of accessibility and extensibility hinder their use as foundation models, primarily due to scale dependence. This article investigates the potential of sustainability and programmability principles in architectural design to create a more context-aware and user-focused AI ecosystem, promoting the responsible development of foundation models.

Keywords
Foundation Models, Human-centric Design, Transformer Models, Programmability

1. Introduction

Artificial intelligence (AI) has made significant progress in recent years, particularly in the area of generative modeling [1, 2, 3]. These models can generate new and original content, such as images and text, based on a set of inputs. While these models have shown promising results in a variety of downstream tasks, they often rely on scaling up the compute and raw data to achieve such performance.[4, 5, 6]

However, training on web-scale corpora with limited filtering can lead to unforeseen failure modes and a lack of semantic context, which can raise ethical concerns such as bias and discrimination.[7] As AI systems are increasingly being adopted in various domains, such as personalized healthcare and social media content recommendation, they must prioritize human-centered principles and operate ethically and fairly.

While the impressive computational scale-up achieved by these models is undeniably an engineering feat, it is not a sustainable approach in the long run due to its significant energy consumption and carbon footprint. Scaling up comes with drawbacks that go beyond environmental concerns and also affect accessibility and extensibility. The sheer scale of these models makes them impractical for inference on standard hardware, thus limiting their accessibility to researchers and developers. Moreover, the reliance on scale undermines their effectiveness as

HCAI4U 2023: Workshop on User Perspectives in Human-Centred Artificial Intelligence, September 20, 2023, Turin, Italy.

[*] Corresponding author.

narendraprakash.patwardhan@unina.it (N. Patwardhan); shreya.shetye@deepkapha.com (S. Shetye); lidia.marassi@unina.it (L. Marassi); monica.zuccarini@unina.it (M. Zuccarini); tannistha.maiti@deepkapha.com (T. Maiti); tarry.singh@deepkapha.com (T. Singh)
0000-0002-4807-5664 (N. Patwardhan); 0009-0006-8134-5466 (L. Marassi); 0000-0003-4170-1295 (T. Singh)

CEUR Workshop Proceedings (CEUR-WS.org)

foundation models, as they encounter various challenges including limited control over output, brittleness, domain dependence, inefficiencies in training, and a lack of customizability for users.

In this paper, we delve into the potential of sustainability and programmability principles in architectural design to shape the future of foundation models. In Section 2, we focus on improving the accessibility of these models by addressing their computational complexity and exploring sustainable alternatives for internal components. We examine techniques for updating AI models in real-time without retraining to address unforeseen failure modes and adapt to changing definitions of ethics and propriety.

In Section 3, our focus shifts to the crucial aspect of extensibility in the design of foundation models. We explore how programmability plays a vital role in enabling users to customize and exert control over these models. Additionally, we delve into innovative approaches that facilitate context-length generalization. Moreover, we emphasize the importance of tailoring AI systems to cater to individual needs and values, thus fostering a more personalized and adaptive experience.

Finally, in Section 4, we summarize our findings and discuss the implications of our work. We hope to provide valuable insights into designing and deploying such models and help guide future research and development efforts in this area.

2. Sustainability Principles in Architectural Design

Bommasani Et al. [8] coined the term foundation model to refer to any model that is trained on a broad amount of data and is suitable for adaptation to a variety of downstream tasks. This is in contrast to task-specific models, which focus on learning optimal representations for a single problem based on a limited amount of data and aren't suitable for adaptation to other tasks without considerable retraining. With the advent of large-scale language models, much of the research in the field of AI (especially in the subdomain of NLP) has shifted focus from developing task-specific models to effectively adapting foundation models to downstream tasks.

As the models increase in scale, the optimization demands a proportional increase in data. To meet this need curation processes have seen a similar shift from their traditionally human-led nature to heuristics-based automated filtering. To overcome the prohibitive cost of label generation, most foundation models adopt self-supervised learning (SSL), which formulates the pretraining task as a property of raw unannotated data such as next-word prediction or masked reconstruction. [9, 10]

An unfortunate side effect of automated curation based on web crawling (which is further augmented by the use of SSL) is that imperfections in curation heuristics lead the resultant model to exhibit biases, show memorization of private data, vulnerability to adversarial assaults. [7] In the following subsections, we first explore the key components that form the backbone of the current crop of foundation models and possible alterations to make them sustainable. We then identify essential additions to base transformer architecture that decouples modality modeling from knowledge modeling.

2.1. Key Components of Foundation Models

The fundamental constituent of all modern foundation models is the transformer block based on the idea of self-attention [11], which allows the model to focus on certain parts of the input sequence when processing it. The transformer block consists of self-attention and feedforward layers. One of the key advantages of transformers is their ability to handle long input sequences and internal parallelization, which allows them to be trained efficiently on large datasets.

We will go through each of the components of the transformer block and discuss the possible alternatives to make them more sustainable. The original formulation of the transformer is presented with an encoder-decoder architecture. Modern variants such as GPTs focus on decoder-only architecture. This simplification allows for straightforward implementations and removes the overhead of architectural choices such as early or late fusion, number of context injections, etc. However, the encoder-decoder architecture allows caching of the context. This is particularly useful when downstream applications can request multiple generations based on the same input.

It is commonly thought that attention is the bottleneck in transformer-based models. Numerous papers have attempted to reduce the computational complexity of the attention mechanism by providing approximations to the full attention matrix at the cost of a slight decrease in accuracy. [12, 13, 14, 15] As these methods only focus on FLOPs and ignore the overhead of memory access patterns, they often fail to achieve wall-clock speedups. Coupled with the inability to achieve the same accuracy as the standard attention, these methods have not been widely adopted.

Flash Attention [16] is an exact attention algorithm that reduces the number of memory reads/writes within different types of accelerator memories (HBM and SRAM) with the help of tiling. It has been found to offer 15% wall-clock speedup over the standard implementations of BERT-large (seq. len. 512), 3x speedup for GPT-2 (seq. len. 1k), and 2.4x speedup for the long-range arena (seq. len. >1k). Since our primary concern is to reduce energy consumption, wall-clock timing is a more appropriate metric to look at than a notion of complexity detached from the hardware. We thus recommend the adoption of hardware-specific versions of Flash Attention as a sustainable alternative to approximate attention mechanisms described above.

It has been found that the effect of quadratic cost for attention minimizes with scale as it only adds $O(dL)$ FLOPs per token per layer, whereas matrix multiplications (in the form of linear layers) add $O(d^2)$[6]. A common misconception is that the parameter count directly correlates with energy consumption. However, higher-order mechanisms such as routing allow scaling up the parameter count without a proportional increase in energy consumption [17]. Routing introduces sparsity which results in developing independent paths that act together as a mixture of experts. We recommend adopting routing as a sustainable alternative to the standard linear layers, as it enables the model to capture the diverse information content from larger datasets without requiring additional processing power. By incorporating routing into our model architecture, we can enhance both sustainability and efficiency while maintaining the ability to leverage the valuable information contained in extensive datasets.

In an orthogonal direction, Jaegle Et al. [18] searching for universal architectures proposed the use of cross-attention to decouple the computational complexity of transformer blocks from the input sequence length. This allowed them to consume multimodal data without any architectural changes. An extension to this work soon followed to decouple this complexity from output sequence length as well [19]. This decoupling from both ends allows a network to treat multimodal data in a unified manner.

[20] provided formulation for [19] to be suitable for self-supervised learning. We recommend considering the adoption of the general architectural structure of the perceiver family as a sustainable alternative to the established self-attention (SA)-based encoder-decoder and decoder-only architectures. This architectural approach has shown promising potential in terms of reducing inference time latency.

2.2. Essential Additions to Foundation Models

Training large models incur a significant cost in terms of the environment, finances, and time. Unfortunately, the current foundation models require periodic retraining lest they become outdated in their predictions. Knowledge about world events and new inventions is an important indirect context that contributes to the quality of the generated text.

Based on the scale, even the cost and resources required for retraining can be significant. Consumers often resort to non-robust methods such as prompt tuning for incorporating specific knowledge to avoid fine-tuning. However, since the context length for most models is limited, prompt tuning is only effective in limited cases. Instilling expert knowledge in trained models has been an area of active research.

Boregeaud Et al. showed that the quality of generation can be improved with additional context obtained from retrieval [21]. Instead of increasing the size of the model and training on more data, their method gives models the ability to directly access a large database during the forward pass by using a small, frozen network and doing the nearest neighbor lookup.

This method (RETRO) can be used with models trained from scratch as well as be retrofitted after the initial training has been done. While RETRO hasn't yet been widely adopted it provides a promising direction for service providers to perform hotfixes to the model, an essential feature in mature products.

Combining multiple modalities, such as text, images, audio, and video, can improve the performance and generalization ability of large AI models. This is because multimodal data provides more comprehensive information, is more robust to errors in any individual modality, and better reflects the complexity of real-world situations. For example, if an image is blurry or distorted, the accompanying text may still provide enough information to accurately identify the object.

Networks that use a common structure to process independent modalities have been shown to have synergistic effects on generalization. While current state-of-the-art models for computer vision are trained to predict a fixed set of categories, in real-world scenarios, the categories are often not known in advance and the model must be able to adapt to new categories as well as provide a null prediction when the category differs from the known set. Because the cardinality of text is far larger than the categories in classification datasets, models that share a common representation space can be applied in a zero-shot manner [22] or even used to identify and label new categories.

To enable adaptation to new modalities, it is crucial to consider tokenization as part of the architectural design process. Tokenization refers to the process of breaking down the input modality into smaller units, such as words or subword units for text, to facilitate processing by transformer models. One approach that has shown promise is byte-based tokenization [23]. Instead of relying solely on word-level tokenization, byte-based tokenization allows for a more fine-grained decomposition and is a natural embedding space across modalities. Byte-based tokenization can handle unseen categories well, however, trades computation for memory requirements as the sequence length increases by an order of magnitude compared to other representations. An approach by [24] overcomes this limitation by segmenting the byte-based sequence into patches and applying a local submodel to have intra-patch information pooling and a global submodel to have inter-patch information routing.

3. Programmability for User-Focused AI

In this section, we explore various aspects of programmability that empower users with customization, control, and alignment with their individual needs and values. We discuss the concept of empowering users through customization and control, highlighting diffusion models and pluggable control as effective techniques. Additionally, we delve into two innovative approaches for context length generalization, namely ALiBi, and RoPe, which utilize advanced positional embeddings. Lastly, we address the significance of aligning AI systems with individual needs and values, highlighting the adaptation capabilities of LoRA. By examining these aspects, we aim to shed light on the potential of programmability to enhance the user experience and enable AI systems to better meet the diverse requirements of individuals.

3.1. Pluggable Control

A new class of methods has rapidly captured the image generation landscape, called diffusion models. [25, 2] These models can be trained in a semi-supervised manner, where the model is trained on a combination of labeled and unlabelled data. Diffusion generates input-output pairs by iteratively adding Gaussian noise to the input till the result itself resembles Gaussian noise. The model is then trained to invert this process. Due to the inherent stability of optimization that stems from supervised learning and the excellent quality of the generated samples, diffusion models have largely replaced the more traditional generative adversarial networks (GANs). [26, 27, 28]

While initially, diffusion models were costly to train and obtain generations from at high resolution, recent work has shifted diffusion to be an internal component of a larger network and work with latent representations [3]. Due to open-sourcing of the weights of this model (known as StableDiffusion), it has become viable to produce high-resolution images even on a consumer-grade GPU or with some additional time, on a CPU.

Diffusion-LM [29] provides a natural extension of diffusion models to the domain of NLP. Instead of producing a single token at a time and then finding the chain that has a maximum probability, Diffusion-LM work with spans. Diffusion-LM allows the use of arbitrary differentiable modules as a guidance mechanism. This makes them quite attractive for removing toxicity and stopping the leakage of sensitive information. While the research on these models is still in its infancy, through our future work we aim to explore their use in the context of instilling reliable and responsible behavior in the resultant generations.

A new technique, introduced on top of diffusion models but applicable to most generative models is inversion [30]. By providing a few instances of a concept not present in the corpus, inversion learns its optimal representation in the input domain. This allows for the generation of new instances and combinations with existing concepts. Inversion provides a critical programmable interface to the model that we should thrive to enable within our models.

3.2. Context Length Generalization

Context-length generalization is crucial in transformer models as it enables the ability to process and understand larger contexts in downstream applications such as chatbots and summarization. Transformer models are typically trained on sequences of up to 4k tokens, which limits their capacity to handle longer contexts. Since transformers are permutation invariant, the choice of positional embeddings plays a significant role in determining their ability to generalize across different context lengths [31].

ALiBi [32] adopts a different strategy. It introduces a penalty proportional to the distance between query-key attention scores, rather than using positional embeddings. This biasing mechanism enables ALiBi to effectively handle longer contexts by discouraging excessive attention to distant tokens and promoting a better focus on relevant information within the given context. Rotary positional embeddings (presented in [33]) have been shown to handle larger contexts by changing their base at inference time [34].

3.3. Low Resource Fine-tuning

Fine-tuning large models has traditionally been a challenging task in AI research. The complexity and size of these models often make fine-tuning computationally expensive and resource-intensive. Moreover, the need for large-scale labeled datasets for fine-tuning can be a bottleneck, as collecting and annotating such datasets is time-consuming and costly. However, the introduction of LoRA [35] (low-rank adaption) has brought significant simplification to the process of fine-tuning large models. LoRA and its subsequent extension [36] utilizes low-rank decomposition to efficiently update the weights of the pre-trained model during fine-tuning. This approach reduces the computational requirements and memory footprint, making it more feasible to fine-tune large models on limited resources.

4. Conclusion

In conclusion, integrating sustainability and programmability principles into the architectural design of foundation models holds the potential to address the challenges of accessibility and extensibility, fostering a context-aware and user-focused AI ecosystem. By prioritizing responsible development, we can empower users with customization and improve efficiency. It is crucial for researchers and practitioners to embrace these principles and collaboratively design foundation models that promote a more inclusive and responsible AI future.

Acknowledgments

This research was made possible thanks to the generous support of the SIMAR GROUP s.r.l., Monte Urano (FM, Italy) and NextGenerationEU, which provided a Ph.D. scholarship to the lead author.

References

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv preprint arXiv:2204.06125 (2022).

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[5] O. Lieber, O. Sharir, B. Lenz, Y. Shoham, Jurassic-1: Technical details and evaluation, White Paper. AI21 Labs 1 (2021).

[6] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, et al., Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model, arXiv preprint arXiv:2201.11990 (2022).

[7] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610–623.

[8] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021).

[9] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training, OpenAI (2018).

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[12] S. Wang, B. Z. Li, M. Khabsa, H. Fang, H. Ma, Linformer: Self-attention with linear complexity, arXiv preprint arXiv:2006.04768 (2020).

[13] N. Kitaev, Ł. Kaiser, A. Levskaya, Reformer: The efficient transformer, arXiv preprint arXiv:2001.04451 (2020).

[14] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[15] B. Chen, T. Dao, E. Winsor, Z. Song, A. Rudra, C. Ré, Scatterbrain: Unifying sparse and low-rank attention, Advances in Neural Information Processing Systems 34 (2021) 17413–17426.

[16] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, C. Ré, Flashattention: Fast and memory-efficient exact attention with io-awareness, arXiv preprint arXiv:2205.14135 (2022).

[17] W. Fedus, B. Zoph, N. Shazeer, Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, The Journal of Machine Learning Research 23 (2022) 5232–5270.

[18] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, J. Carreira, Perceiver: General perception with iterative attention, in: International conference on machine learning, PMLR, 2021, pp. 4651–4664.

[19] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, et al., Perceiver io: A general architecture for structured inputs & outputs, arXiv preprint arXiv:2107.14795 (2021).

[20] C. Hawthorne, A. Jaegle, C. Cangea, S. Borgeaud, C. Nash, M. Malinowski, S. Dieleman, O. Vinyals, M. Botvinick, I. Simon, et al., General-purpose, long-context autoregressive modeling with perceiver ar, arXiv preprint arXiv:2202.07765 (2022).

[21] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, et al., Improving language models by retrieving from trillions of tokens, in: International conference on machine learning, PMLR, 2022, pp. 2206–2240.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[23] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, C. Raffel, Byt5: Towards a token-free future with pre-trained byte-to-byte models, Transactions of the Association for Computational Linguistics 10 (2022) 291–306.

[24] L. Yu, D. Simig, C. Flaherty, A. Aghajanyan, L. Zettlemoyer, M. Lewis, Megabyte: Predicting million-byte sequences with multiscale transformers, arXiv preprint arXiv:2305.07185 (2023).

[25] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, B. Guo, Vector quantized diffusion model for text-to-image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10696–10706.

[26] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, Advances in Neural Information Processing Systems 34 (2021) 8780–8794.

[27] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, arXiv preprint arXiv:2010.02502 (2020).

[28] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Advances in Neural Information Processing Systems 33 (2020) 6840–6851.

[29] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, T. B. Hashimoto, Diffusion-lm improves controllable text generation, arXiv preprint arXiv:2205.14217 (2022).

[30] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, D. Cohen-Or, An image is worth one word: Personalizing text-to-image generation using textual inversion,

arXiv preprint arXiv:2208.01618 (2022).

[31] R. Csordás, K. Irie, J. Schmidhuber, The devil is in the detail: Simple tricks improve systematic generalization of transformers, arXiv preprint arXiv:2108.12284 (2021).

[32] O. Press, N. A. Smith, M. Lewis, Train short, test long: Attention with linear biases enables input length extrapolation, arXiv preprint arXiv:2108.12409 (2021).

[33] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, Y. Liu, Roformer: Enhanced transformer with rotary position embedding, arXiv preprint arXiv:2104.09864 (2021).

[34] G. Georgi, llama.cpp, https://github.com/ggerganov/llama.cpp, 2023.

[35] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).

[36] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, arXiv preprint arXiv:2305.14314 (2023).