MambaByte: Token-free Selective State Space Model

Junxiong Wang **Tushaar Gangavarapu** **Jing Nathan Yan** **Alexander M Rush**
Cornell University
{jw2544,tg352,jy858,arush}@cornell.edu

Abstract

Token-free language models learn directly from raw bytes and remove the bias of subword tokenization. Operating on bytes, however, results in significantly longer sequences, and standard autoregressive Transformers scale poorly in such settings. We experiment with MambaByte, a token-free adaptation of the Mamba state space model, trained autoregressively on byte sequences. Our experiments indicate the computational efficiency of MambaByte compared to other byte-level models. We also find MambaByte to be competitive with and even outperform state-of-the-art subword Transformers. Furthermore, owing to linear scaling in length, MambaByte benefits from fast inference compared to Transformers. Our findings establish the viability of MambaByte in enabling token-free language modeling.

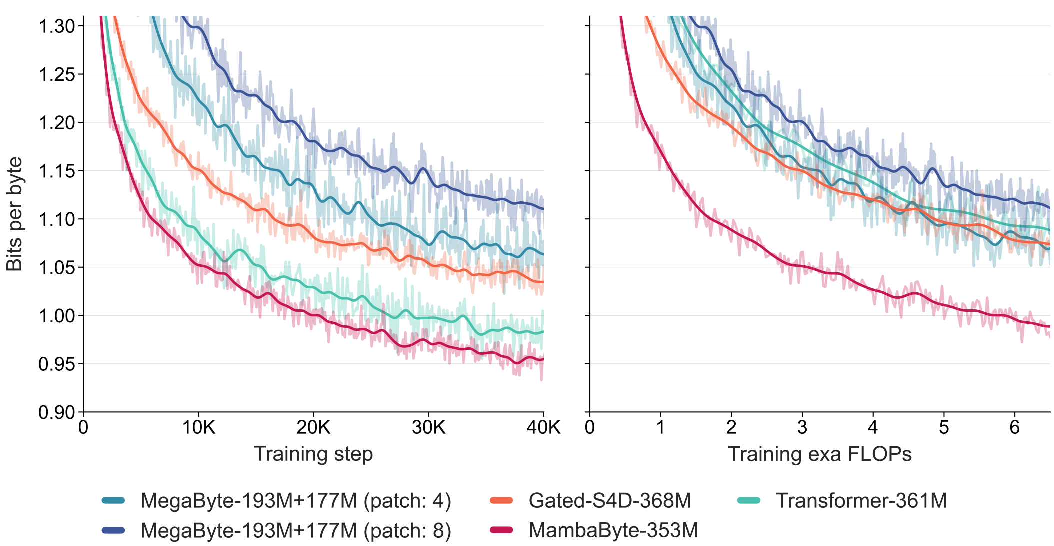

Figure 1: **Benchmarking byte-level models with a fixed parameter budget.** Language modeling results on PG19 ($8,192$ consecutive bytes), comparing the standard Transformer [Vaswani et al., 2017, Su et al., 2021], MegaByte Transformer [Yu et al., 2023], gated diagonalized S4 [Mehta et al., 2023], and MambaByte. (Left) Model loss over training step. (Right) FLOP-normalized training cost. MambaByte reaches Transformer loss in less than one-third of the compute budget.

1 Introduction

When defining a language model, a base tokenization is typically used—either words [Bengio et al., 2000], subwords [Schuster and Nakajima, 2012, Sennrich et al., 2015, Wu et al., 2016, Wang et al., 2020], or characters [Gao et al., 2020a]. Of these, subword tokenization has been the most popular choice, as it achieves a natural compromise between training efficiency and the ability to handle out-of-vocabulary words. However, several works (e.g., Xue et al. [2022]) have noted issues with subword tokenizers, such as a lack of robustness to typos, spelling and capitalization variations, and morphological changes.

Researchers [Clark et al., 2022, Xue et al., 2022, Yu et al., 2023] have employed an alternative approach of using byte sequences, i.e., an end-to-end mapping from raw data to predictions without any intermediate tokenization. Compared to subword models, byte-level language models can generalize more easily across orthographic and morphological variants. Of course, modeling text as bytes means that the resultant sequences are significantly longer than their subword counterparts. This pushes the efficiency issues upstream into the architecture itself.

Efficiency issues are particularly pronounced for autoregressive Transformers [Vaswani et al., 2017], which dominate language modeling [Brown et al., 2020, Touvron et al., 2023]. Due to the quadratic cost of attention, Transformers scale poorly for long (byte) sequences [Brown et al., 2020, Zhang et al., 2022]. Researchers have *compressed* the internal Transformer representation to work with long sequences, for instance, developing length-aware modeling approaches [Dai et al., 2020, Nawrot et al., 2022], where groups of tokens are merged within the intermediate layers. Recently, Yu et al. [2023] proposed the MegaByte Transformer, which uses compression in the form of fixed-size patches of bytes as a subword analog. As a result, MegaByte enables lower computational costs.[1]

In this work, we introduce MambaByte, an efficient and simple byte-level language model. The model is a straightforward adaptation of the recently introduced Mamba architecture [Gu and Dao, 2023], a linear-time approach for sequence modeling. Mamba builds off the approach pioneered by state space models (SSMs) [Gu et al., 2021, Gupta et al., 2022, Gu et al., 2022, Smith et al., 2023] by introducing a selection mechanism that is more effective for discrete data such as text and providing an efficient GPU implementation. Our simple observation is that using Mamba (without modifications) relieves the main computational bottleneck in language modeling, thus allowing for the elimination of patching and effective use of the available compute budget.

Experiments compare MambaByte to Transformers, SSMs, and MegaByte (patching) architectures in a fixed parameter and fixed compute setting on several long-form text datasets. Figure 1 summarizes our main findings. Compared to byte-level Transformers, MambaByte achieves better performance faster and is significantly more compute efficient. We also consider the viability of token-free language models compared to the existing state-of-the-art subword models. In this regard, we find MambaByte to be competitive with various subword baselines despite handling significantly longer sequences. Our results establish MambaByte as a strong alternative to the existing tokenizer-dependent models and advocate its use to facilitate end-to-end learning.

2 Background: Selective state space sequence models

SSMs model the evolution of a hidden state across time through a first-order differential equation. Linear time-*invariant* SSMs [Gu et al., 2021, Gupta et al., 2022, Gu et al., 2022, Smith et al., 2023] have shown promising results in deep learning across several modalities. However, Gu and Dao [2023] have recently argued that the constant dynamics of these approaches lack input-dependent context *selection* in the hidden state, which may be necessary for tasks such as language modeling. To this end, they proposed Mamba, which defines the time-varying continuous state dynamics for a given input $x(t) \in \mathbb{R}$, hidden state $h(t) \in \mathbb{R}^n$, and output $y(t) \in \mathbb{R}$ at time t as:

$$\frac{\mathrm{d}h(t)}{\mathrm{d}t} = \mathrm{A}h(t) + \mathrm{B}(t)x(t); \quad y(t) = \mathrm{C}(t)h(t), \tag{1}$$

which is parameterized by a diagonal time-invariant system matrix $\mathrm{A} \in \mathbb{R}^{n \times n}$ and time-dependent input and output matrices $\mathrm{B}(t) \in \mathbb{R}^{n \times 1}$ and $\mathrm{C}(t) \in \mathbb{R}^{1 \times n}$.

To model discrete-time sequences such as bytes, the continuous time dynamics in (1) must be approximated through discretization. This results in a discrete-time hidden state recurrence with new

[1] Although our experiments (see Figure 1) indicate that patching can also lower the model performance compared to the standard Transformer.

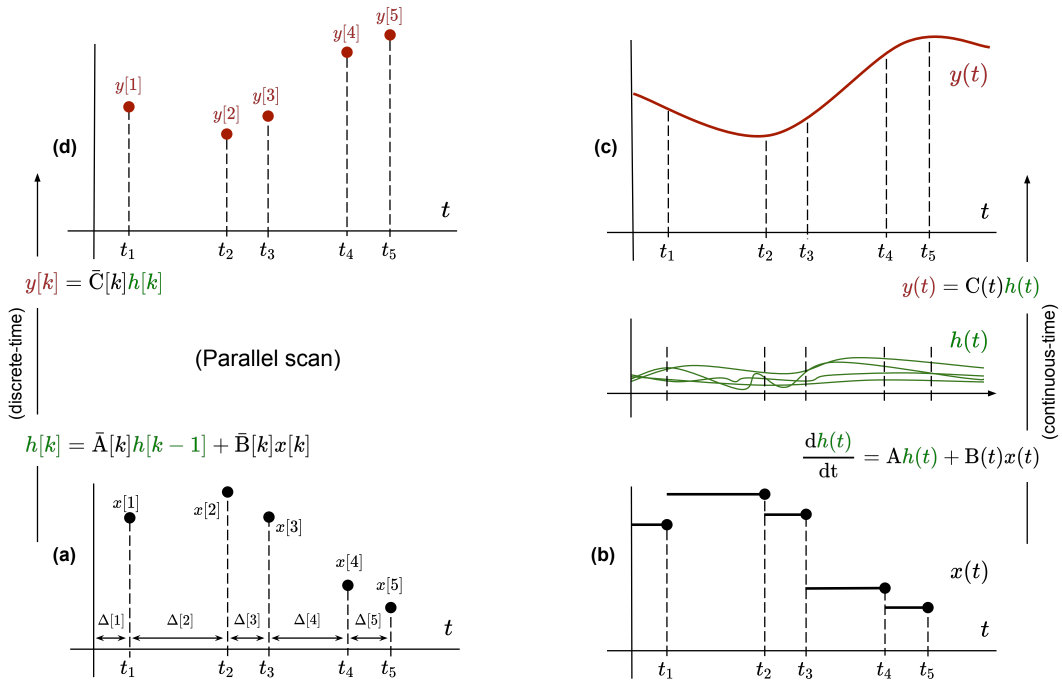

Figure 2: **Illustration of the Mamba SSM.** (a) The discrete-time input $x[k]$, along with input-selective $\Delta[k]$. (b) The continuous-time signal $x(t)$. (c) Mathematically, the SSM transforms the continuous-time $x(t)$ through an n-dimensional hidden state (here, $n = 4$) using parameters A and $\mathrm{B}(t)$, which is then mapped to the output $y(t)$ using $\mathrm{C}(t)$. (d) Practically, we compute $y[k]$ using a discrete-time parallel scan at the steps defined by $\Delta[k]$ and discrete-time matrices $\overline{A}[k]$, $\overline{B}[k]$, and $\overline{C}[k]$. At inference, we run the recurrence directly.

matrices at each timestep, $\overline{\mathrm{A}}$, $\overline{\mathrm{B}}$, and $\overline{\mathrm{C}}$, such that

$$h[k] = \overline{\mathrm{A}}[k]h[k-1] + \overline{\mathrm{B}}[k]x[k]; \quad y[k] = \overline{C}[k]h[k]. \tag{2}$$

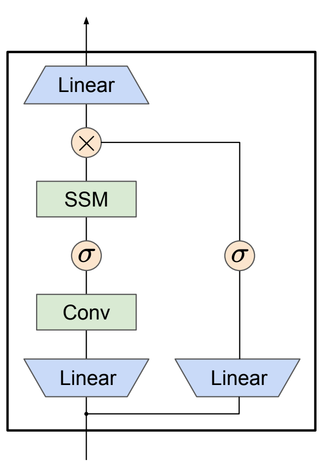

Figure 3: **Mamba block.** σ indicates Swish activation [Ramachandran et al., 2017].

Observe that (2) resembles a linear version of a recurrent neural network and can be applied in this recurrent form during language model generation. The discretization requires a timestep, $\Delta[k]$, for each input position, corresponding to treating $x[k] = x(t_k)$ for $t_k = \sum_{j=1}^{k} \Delta[j]$. The discrete-time matrices $\overline{\mathrm{A}}$, $\overline{\mathrm{B}}$, and $\overline{\mathrm{C}}$ can then be computed from $\Delta[k]$. Figure 2 illustrates how Mamba models discrete sequences.

In Mamba, the SSM terms are input-selective, i.e., B, C, and Δ are defined as functions of the input $x[k] \in \mathbb{R}^d$:

$$\Delta[k] = \mathrm{softplus}(W_\Delta(W_R x[k]); \quad \mathrm{B}(t_k) = W_{\mathrm{B}} x[k], \tag{3}$$

where $W_{\mathrm{B}} \in \mathbb{R}^{n \times d}$ (C is similarly defined), $W_\Delta \in \mathbb{R}^{d \times r}$ and $W_R \in \mathbb{R}^{r \times d}$ (for some $r \ll d$) are learnable weights, and softplus ensures positivity. Note that the SSM parameters A, B, and C are identical for each input dimension d, but the timesteps Δ are distinct; this results in a hidden state size of $n \times d$ per timestep k. (See Appendix D for specifics on discretization and selectivity.)

Mamba embeds this SSM layer into a full neural network language model. Specifically, the model utilizes a stack of gated layers inspired by the previous gated SSM [Mehta et al., 2023]. Figure 3 shows the Mamba architecture combining the SSM layer with a gated neural network.

Parallel scans for linear recurrences. At training time, we have access to the entire sequence x, allowing us to compute the linear recurrence more efficiently. Smith et al. [2023] demonstrated the use of work-efficient parallel scans [Blelloch, 1990] for efficiently computing the sequential recurrence in linear SSMs. For Mamba, we first map the recurrence to a sequence of L tuples, with $e_k = (A_k, b_k) := (\overline{\mathrm{A}}[k], \overline{\mathrm{B}}[k]x[k])$, then define an associative operator $\bullet$ such that $e_j \bullet e_k = (A_k A_j, A_k b_j + b_k)$. Finally, we apply a parallel scan to compute the sequence $[(\overline{\mathrm{A}}[1], h[1]), (\overline{\mathrm{A}}[2]\overline{\mathrm{A}}[1], h[2]), \ldots]$. In general, this requires $\mathcal{O}(T_\bullet \log_2(L))$ time, using $L/2$ processors, where $T_\bullet$ is the cost of a matrix-matrix multiplication. Noting $\overline{\mathrm{A}}$ to be a diagonal matrix, the linear recurrence can be computed parallelly in $\mathcal{O}(n \log_2(L))$ time and $\mathcal{O}(nL)$ space. A parallel scan with a diagonal matrix is also efficient in operation, requiring $\mathcal{O}(nL)$ FLOPs.

3 Experimental setup

Our experiments compare MambaByte to other byte-level Transformers and SSMs. All our models employ the same training recipes (see Appendix C for details). We utilize a set of diverse long-form text datasets: PG19 [Rae et al., 2020], Stories [Trinh and Le, 2018], Books [Gao et al., 2020b], ArXiv [Gao et al., 2020b], and Code [Gao et al., 2020b]. Dataset sizes and average document lengths are included in Appendix A.

Table 1: **Relative training FLOPs by model size.** All MegaByte models use a patch size of 8.

Experiment	Models	FLOPs per train byte
Medium-scale	MegaByte-758M+262M : MambaByte-353M	1.02 : 1
Large-scale	MegaByte-1.3B+350M : MambaByte-972M	0.54 : 1
	MegaByte-1.3B+218M : MambaByte-972M	0.40 : 1

Performance comparison across architectures requires care. To this end, we consider two settings: compute-matched and parameter-matched. This setup is necessary as the default MegaByte Transformer employs a *global* module that works with $8\times$-patched representations of the input, thus using $8\times$ fewer feed-forward FLOPs per byte than a raw Transformer, while having significantly more parameters. Table 1 shows the MegaByte and MambaByte model sizes employed in our experiments. The (forward pass) FLOPs computation for various model architectures and the associated hyperparameters employed are detailed in Appendix B.

All MambaByte models were trained using the open-source Mamba code base.[2] At training, we shuffle the documents and use contiguous sequences of $8,192$ bytes (one per document), starting from a random position. We enable mixed precision training using BF16 for training efficiency at scale. The optimizer, learning rate scheduler, and other training details are specified in Appendix C.

Press et al. [2021] proposed using a sliding window to trade off speed for performance during inference. Following this, we employ a sliding window (with a stride of $L_{\text{ctx}}/2$ for a byte sequence of length L_{ctx}) when comparing with the state-of-the-art subword models in Table 3.

4 Results

Table 2 shows the bits per byte (BPB) across each dataset. For this experiment, the MegaByte-758M+262M and MambaByte models use the same number of FLOPs per byte (see Table 1). We observe MambaByte to outperform MegaByte consistently across all datasets. Furthermore, we note that we could not train MambaByte for the full 80B bytes due to monetary constraints, but MambaByte outperforms MegaByte with $0.63\times$ less compute and training data. Additionally, MambaByte-353M also outperforms byte-level Transformer and PerceiverAR.

How is MambaByte performing better than a much larger model in so few training steps? Figure 1 further explores this relationship by looking at models with the same number of parameters. The graphs indicate that for MegaByte models of the same parameter size, models with less input patching perform better, but when compute-normalized, they perform similarly. In fact, a full-length Transformer, while slow in an absolute sense, also performs similarly to MegaByte when compute-normalized. In contrast, switching to the Mamba architecture significantly improves both the compute usage and the model performance.

[2] https://github.com/state-spaces/mamba.

Table 2: **Medium-scale experiments.** MegaByte and MambaByte use the same FLOPs per byte. (The BPB for Transformer, PerceiverAR, and MegaByte are taken from Yu et al. [2023].)

Byte-level model	Context	Bytes trained	Test BPB ↓				
			PG19	Stories	Books	ArXiv	Code
Transformer-320M	$1,024$	80B	1.057	1.064	1.097	0.816	0.575
PerceiverAR-248M	$8,192$	80B	1.104	1.070	1.104	0.791	0.546
MegaByte-758M+262M (patch: 8)	$8,192$	80B	1.000	0.978	1.007	0.678	0.411
MambaByte-353M	$8,192$	30B*	**0.930**	**0.908**	**0.966**	**0.663**	**0.396**

Table 3: **Large-scale experiment on PG19.** The observed BPB scores are converted to word-level PPL for comparison with past works. All the byte-level models are compute-matched. MambaByte-972M significantly outperforms other byte-level models and is competitive with state-of-the-art subword models. (Accompanying citation indicates the work from which the corresponding result was taken; fields marked – are unknown.)

	(#Layers) Model	Vocab	Effective context (in bytes)[3]	Effective bytes trained[3]	Val PPL↓	Test PPL↓
Subword	(36) Transformer-XL [Rae et al., 2020]	32K	$2,048/4,096$	400B	45.5	36.3
	(36) Compressive [Rae et al., 2020]	32K	$2,048/2\times 2,048$	400B	43.4	33.6
	(22) Routing-490M[4] [Roy et al., 2021]	82K	$32,768$	330B	–	33.2
	(60) PerceiverAR-974.6M [Hawthorne et al., 2022]	32K	$8,192$	1.68T	45.9	28.9
	(24) Block-Recurrent-1.3B [Hutchins et al., 2022]	32K	$4,096$/recurrence	–	–	**26.5**
Byte	(–) Transformer-320M [Yu et al., 2023]	256	$8,192$	400B	81.6	69.4
	(–) PerceiverAR-248M [Yu et al., 2023]	256	$8,192$	400B	119.1	88.8
	(24+24) MegaByte-1.3B+350M [Yu et al., 2023]	256	$8,192$/patch: 8	400B	42.8	36.4
	(48) MambaByte-972M	256	$8,192$[5]	150B*	**39.5**	33.0

Table 4: **Generation speed benchmarking.** Speed to generate $8,192$ bytes; fields marked – are unknown. (Upper) The BPB on PG19 and generation time for the Transformer and MegaByte are taken from Yu et al. [2023]. (Lower) MegaByte and MambaByte run on the same hardware.

Model	Bytes trained	Context	Test BPB ↓	Generation time (s) ↓
Transformer-350M	–	$1,024$	1.064	132
MegaByte-1.3B+218M (patch: 8)	–	$8,192$	0.991	93
MegaByte-1.3B+218M (patch: 8)[6]	–	$8,192$	–	265
MambaByte-972M	75B*	$8,192$	**0.883**	**29**
w/ sliding window ($2\times$ bytes)			**0.863**	58
MambaByte-1.6B	–	$8,192$	–	36

Following these findings, Table 3 compares a larger version of these models on the PG19 dataset. For this experiment, we compare MambaByte-972M with MegaByte-1.3B+350M and other byte-level models, as well as several state-of-the-art subword models. (The conversion from BPB to perplexity (PPL) is detailed in Appendix E). We find that MambaByte-972M, even just trained for 150B bytes, outperforms all the byte-level models and achieves competitive performance with subword models.

[3] For subword models, we use one subword as being equivalent to four bytes.
[4] The number of parameters is noted from Hutchins et al. [2022].
[5] For inference, we use a context of $32,768$ bytes.
[6] Open-source implementation: `https://github.com/lucidrains/MEGABYTE-pytorch`.

Text generation. Autoregressive inference in Transformer models requires caching the entire context, which can significantly affect the generation speed. MambaByte does not suffer from this bottleneck as it maintains a single hidden state per layer that evolves with time, enabling constant time per generation step. Table 4 compares the text generation speeds of MambaByte-972M and MambaByte-1.6B with MegaByte-1.3B+350M on an A100 80GB PCIe GPU. While MegaByte significantly reduces the generation cost through patching, we observe MambaByte to be $2.6\times$ faster in a parameter-matched setting due to its use of recurrent generation. Appendix F includes more information about the generation process.

5 Conclusion

We introduce MambaByte, a token-free SSM for modeling long byte-sequences. MambaByte outperforms other byte-level models over several datasets and shows competitive results with subword Transformers, thus serving as a promising tokenization alternative. SSMs also enable significantly fast text generation due to their recurrent nature, making byte models practical. Our findings establish the possibility of token-free language modeling in future large models.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in neural information processing systems*, 30, 2017.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv e-prints*, pages arXiv–2104, 2021.

Lili Yu, Daniel Simig, Colin Flaherty, Armen Aghajanyan, Luke Zettlemoyer, and Mike Lewis. MegaByte: Predicting Million-byte Sequences with Multiscale Transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=JTmO2V9Xpz.

Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. Long Range Language Modeling via Gated State Spaces. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=5MkYIYCbva.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A Neural Probabilistic Language Model. *Advances in neural information processing systems*, 13, 2000.

Mike Schuster and Kaisuke Nakajima. Japanese and Korean Voice Search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE, 2012.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. *arXiv preprint arXiv:1508.07909*, 2015.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*, 2016.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. Neural Machine Translation with Byte-Level Subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9154–9160, 2020.

Yingqiang Gao, Nikola I Nikolov, Yuhuang Hu, and Richard HR Hahnloser. Character-Level Translation with Self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1591–1604, 2020a.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.

Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. *Transactions of the Association for Computational Linguistics*, 10:73–91, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*, 2022.

Zihang Dai, Guokun Lai, Yiming Yang, and Quoc Le. Funnel-Transformer: Filtering out Sequential Redundancy for Efficient Language Processing. *Advances in neural information processing systems*, 33:4271–4282, 2020.

Piotr Nawrot, Szymon Tworkowski, Michał Tyrolski, Łukasz Kaiser, Yuhuai Wu, Christian Szegedy, and Henryk Michalewski. Hierarchical Transformers Are More Efficient Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1559–1571, 2022.

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently Modeling Long Sequences with Structured State Spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal State Spaces are as Effective as Structured State Spaces. *Advances in Neural Information Processing Systems*, 35:22982–22994, 2022.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the Parameterization and Initialization of Diagonal State Space Models. *Advances in Neural Information Processing Systems*, 35:35971–35983, 2022.

Jimmy T.H. Smith, Andrew Warrington, and Scott Linderman. Simplified State Space Layers for Sequence Modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Ai8Hw3AXqks.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Guy E Blelloch. Prefix Sums and Their Applications. (CMU-CS-90-190), nov 1990. URL https://www.cs.cmu.edu/~guyb/papers/Ble93.pdf.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive Transformers for Long-Range Sequence Modelling. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylKikSYDH.

Trieu H. Trinh and Quoc V. Le. A Simple Method for Commonsense Reasoning. *arXiv preprint arXiv:1806.02847*, 2018.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027*, 2020b.

Ofir Press, Noah A. Smith, and Mike Lewis. Shortformer: Better Language Modeling using Shorter Inputs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5493–5505, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.427. URL https://aclanthology.org/2021.acl-long.427.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient Content-Based Sparse Attention with Routing Transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021. doi: 10.1162/tacl_a_00353. URL https://aclanthology.org/2021.tacl-1.4.

Curtis Hawthorne, Andrew Jaegle, Cătălina Cangea, Sebastian Borgeaud, Charlie Nash, Mateusz Malinowski, Sander Dieleman, Oriol Vinyals, Matthew Botvinick, Ian Simon, Hannah Sheahan, Neil Zeghidour, Jean-Baptiste Alayrac, Joao Carreira, and Jesse Engel. General-purpose, long-context autoregressive modeling with Perceiver AR. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8535–8558. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hawthorne22a.html.

DeLesley Hutchins, Imanol Schlag, Yuhuai Wu, Ethan Dyer, and Behnam Neyshabur. Block-Recurrent Transformers. *Advances in Neural Information Processing Systems*, 35:33248–33261, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting Recurrent Neural Networks for Long Sequences. *arXiv preprint arXiv:2303.06349*, 2023.

Albert Gu, Isys Johnson, Aman Timalsina, Atri Rudra, and Christopher Re. How to Train your HiPPO: State Space Models with Generalized Orthogonal Basis Projections. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=klK17OQ3KB.

Eric Nguyen, Karan Goel, Albert Gu, Gordon Downs, Preey Shah, Tri Dao, Stephen Baccus, and Christopher Ré. S4ND: Modeling Images and Videos as Multidimensional Signals with State Spaces. *Advances in neural information processing systems*, 35:2846–2861, 2022.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Appendix

A Dataset specifics

Table 5: **Text dataset statistics.** The total bytes, total documents, and the mean document size (bytes per document) for each dataset.

	Total bytes	Total docs	Bytes/doc
PG19	11.74G	$28,752$	$4,082,210$
Stories	34.18G	$948,247$	$36,045$
Books	108.38G	$196,640$	$551,179$
ArXiv	60.27G	$1,264,405$	$47,665$
Code	677G	$56,626,342$	$11,958$

We benchmark our results on various long-form text datasets. The PG19 dataset [Rae et al., 2020] is an extensive collection of full-length English books (written before 1919) from the Project Gutenberg online library. The PG19 dataset is ideal to test for long-distance context modeling [Gao et al., 2020b]. The Stories dataset [Trinh and Le, 2018] is a subset of the CommonCrawl data used for commonsense reasoning and language modeling. The Books dataset [Gao et al., 2020b] is another collection of English books. The ArXiv dataset [Gao et al., 2020b] comprises technical publications in LaTeX from the arXiv online archive. Finally, the Code dataset [Gao et al., 2020b] is a large dataset of publicly available open-source code (under Apache, MIT, or BSD licenses). Dataset statistics are tabulated in Table 5.

For the PG19 dataset, we employ the train, validation, and test data splits as indicated by Rae et al. [2020]. For Stories, Books, ArXiv, and Code datasets, we randomly sample 40M consecutive bytes for testing and the rest to train MambaByte.

B Compute-constrained modeling

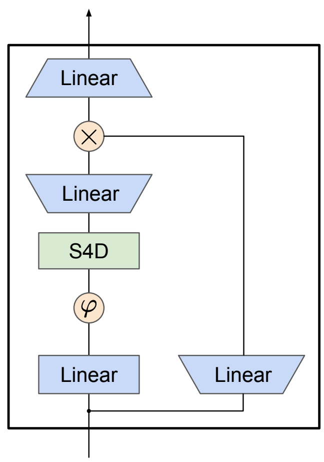

Figure 4: **Gated-S4D block.** Adapted from Mehta et al. [2023]; φ indicates GELU activation [Hendrycks and Gimpel, 2016].

As noted earlier, we evaluate and benchmark MambaByte in a compute-controlled setting. To this end, we estimate the FLOPs per byte incurred by various byte-level model architectures. We parameterize the architectures using hyperparameters n (n_g/n_l) number of (global/local) layers, dimension d (d_g/d_l) of the (global/local) residual stream, expansion factor e of linear layers, patch size p in MegaByte, state dimension n_{state} in SSMs, 1D convolution kernel size k, and low-rank projection dimension r in Mamba. We also include L_{ctx} bytes in the input context. Detailed component-wise compute counts for the forward pass are included in Table 6.

For the medium-scale language modeling experiments (Table 1, §5 of Yu et al. [2023]), Yu et al. [2023] employ the MegaByte-758M+262M model, with a context length of $8,192$ and patch size of 8, trained for 80B bytes. As shown in Figure 5, MambaByte-353M ($n = 53$, $d = 1,024$, $e = 2$) and MegaByte-758M+262M use the same total compute in FLOPs; hence, we employ the MambaByte-353M to benchmark against MegaByte-758M+262M in Table 2 of §4.

For the PG19 scaling experiment (Table 2, §5 and Appendix D.3 of Yu et al. [2023]), Yu et al. [2023] use MegaByte-1.3B+350M (context length of $8,192$ and patch size of 8) trained for 400B bytes to benchmark the observed word-level perplexity against several state-of-the-art subword models. Owing to our hardware limitations, we train MambaByte-972M ($n = 48$, $d = 1,792$, $e = 2$) and control for the total compute used (see Figure 5 to view the associated computational costs). All the model sizes and associated hyperparameters employed in this work are tabulated in Table 7.

C Training recipes

All the models in this study were trained using an AdamW optimizer with $\beta = (0.9, 0.95)$. We used a linear learning rate warm-up (for the first 500 steps) followed by cosine annealing. Keeping consistent

[7] We used the open-source implementation: `https://github.com/lucidrains/MEGABYTE-pytorch`.

Table 6: **Compute (forward pass) estimates for various byte-level language models.** Embedding, de-embedding, and sub-leading terms such as biases, nonlinearities, and layer norms are omitted. (α_* indicates an implementation-specific constant scaling term.)

Model	Component	FLOPs per byte
Transformer [Vaswani et al., 2017]	Multi-head attention	$2n(4d^2 + 2L_{\text{ctx}}d)$
	Pointwise feed-forward	$2n(2ed^2)$
MegaByte [Yu et al., 2023]	Embedding projection	$2d_g^2$
	Global transformer model	$2n_g(4d_g^2 + 2d_g L_{\text{ctx}}/p + 2ed_g^2)/p$
	Global-to-local projection	$2d_g d_l$
	Local transformer model	$2n_l(4d_l^2 + 2pd_l + 2ed_l^2)$
Gated-S4D (Figure 4)	Linear projections	$2n(3ed^2 + d^2)$
	Kernel via Vandermonde $v(\overline{\text{A}})$	$n(\alpha_v ed(n_{\text{state}} + L_{\text{ctx}}) \log_2^2(n_{\text{state}} + L_{\text{ctx}})/L_{\text{ctx}})$
	S4D SSM with convolution	$n(\alpha_{\text{fft}} \log(L_{\text{ctx}})ed + ed)$
	Element-wise gating	ned
MambaByte (Figure 3)	Linear projections	$2n(3ed^2)$
	Pre-SSM 1D convolution	$2nked$
	Δ, B, C from input x	$2n(2edr + 2edn_{\text{state}})$
	Discretization, pre-scan: $\overline{\text{A}}$, $\overline{\text{B}}x$	$n(3edn_{\text{state}})$
	Recurrence with parallel scan	$n(edn_{\text{state}})$
	Output: $y = \overline{\text{C}}h + \overline{\text{D}}x$	$2nedn_{\text{state}} + ned$
	Element-wise gating	ned

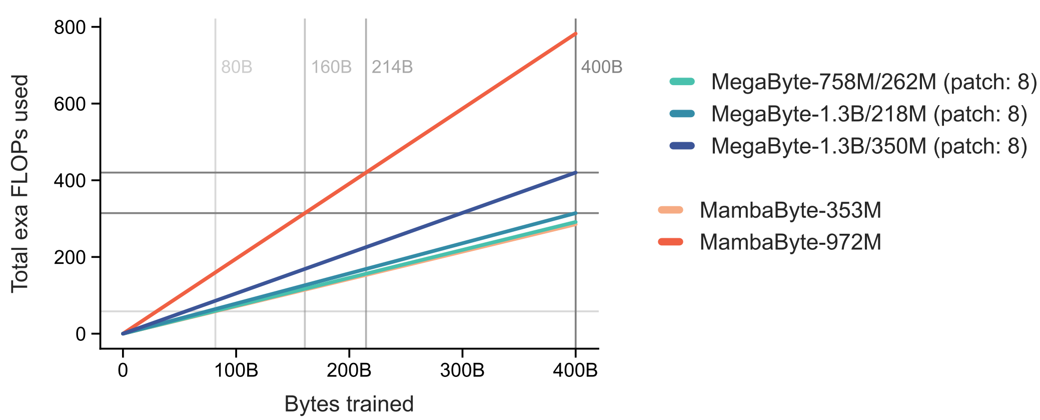

Figure 5: **Computational cost for different model architectures at different scales.** All models use a context length of $8,192$, and MegaByte architectures use a patch size of 8.

with MegaByte training [Yu et al., 2023], we used a batch size of 48 across all our experiments. Additionally, we do not use dropout with any of our models.

For the experiments in Figure 1, we conducted a hyperparameter search using peak learning rates of 0.0002, 0.0006, and 0.0008 and clipped the gradient norm to 1.0 for all the models. The best-observed performance curve for each model is reported in Figure 1. Furthermore, we use an improved Transformer recipe that uses RMSNorm instead of LayerNorm, rotary positional encodings [Su et al., 2021], and linear terms without bias (same as [Yu et al., 2023]).

In our medium-scale experiments shown in Table 2, we set the peak learning rate to 0.0004 and clipped the gradient norm to 0.1. We trained the MambaByte-353M for a total of 80K steps, equivalent to $80,000 \times 48 \times 8,192 \approx 30$B bytes.

Table 7: **Model hyperparameters.** We report the model size and associated hyperparameters for all the models employed in this study. (Accompanying citation indicates the work from which the associated configuration is noted; fields marked as $-$ are unknown.)

Model	Parameters	Hyperparameters				
		n (n_g/n_l)	d (d_g/d_l)	e	L_{ctx}	Others
Transformer	320M [Yu et al., 2023]	22	$1,024$	4	$1,024$	heads: $-$
	350M [Yu et al., 2023]	24	$1,024$	4	$1,024$	heads: 16
	361M	28	$1,024$	4	$8,192$	heads: 16
PerceiverAR	248M [Yu et al., 2023]	17	$1,024$	4	$8,192$	latents: $1,024$
MegaByte	193M+177M[7]	14/14	$1,024/1,024$	4	$8,192$	$p=4,8$; heads: $16/16$
	758M+262M [Yu et al., 2023]	14/18	$2,048/1,024$	4	$8,192$	$p=8$; heads: $16/16$
	1.3B+218M [Yu et al., 2023]	24/15	$2,048/1,024$	4	$8,192$	$p=8$; heads: $32/-$
	1.3B+350M [Yu et al., 2023]	24/24	$2,048/1,024$	4	$8,192$	$p=8$; heads: $32/16$
Gated-S4D	368M	26	$1,024$	4	$8,192$	$n_{\text{state}}=64$
MambaByte	353M	53	$1,024$	2	$8,192$	$k=4; n_{\text{state}}=16; r=64$
	972M	48	$1,792$	2	$8,192$	$k=4; n_{\text{state}}=16; r=112$
	1.6B	48	$2,304$	2	$8,192$	$k=4; n_{\text{state}}=16; r=144$

In the large-scale experiment on PG19, we use a similar setting to that in the medium-scale experiments: the peak learning rate is set to 0.0004, and the gradient norm is clipped to 0.1. The MambaByte-972M is trained for 380K steps, equivalent to $380,000 \times 48 \times 8,192 \approx 150$B bytes.

D Discretization and selection

Discretization has deep connections to continuous-time systems, which allows for desirable properties such as model normalization [Orvieto et al., 2023, Gu et al., 2023] and resolution invariance [Nguyen et al., 2022]. In this section, we show how zero-order hold discretization of a selective SSM can be viewed as a generalization of the gating mechanism in recurrent networks.

Zero-order hold discretization. For a given input $x(t) \in \mathbb{R}$, we wish to discretize a continuous-time SSM defined by (1) in §2. To this end, we sample the system at different time intervals such that $x[k] = x(t_k)$ for $t_k = \sum_{j=1}^{k} \Delta[j]$ and assume a zero-order hold, i.e., $x(t)$ is constant between samples: $x(t_k + \xi) = x(t_k) = x[k]$ for any $\xi \in [t_k, t_{k+1})$. The resultant matrices of the associated discrete SSM are:[8]

$$\overline{\mathrm{A}} = \exp(\mathrm{A}\,\Delta); \quad \overline{\mathrm{B}} = \mathrm{A}^{-1}(\exp(\mathrm{A}\,\Delta) - \mathrm{I})\,\mathrm{B}; \quad \overline{\mathrm{C}} = \mathrm{C}\,.$$

Selection mechanics and gating in recurrent networks. Gu and Dao [2023] note that a selective SSM can be realized as a gated recurrence by setting $\Delta = \mathrm{softplus}(z(x)) = \mathrm{softplus}(W_\Delta(W_R x))$ (as indicated in (3) of §2). By letting $\mathrm{A} = -1$, $\mathrm{B} = 1$, and $n = 1$, the authors observe:

$$\begin{aligned}\overline{\mathrm{A}} &= \exp(\mathrm{A}\,\Delta) \\ &= \exp(-\log(1 + \exp(z(x)))) \\ &= \frac{1}{1 + \exp(z(x))} \\ &= \sigma(-z(x)) \\ &= 1 - \sigma(z(x)).\end{aligned} \qquad \begin{aligned}\overline{\mathrm{B}} &= \mathrm{A}^{-1}(\exp(\mathrm{A}\,\Delta) - \mathrm{I})\,\mathrm{B} \\ &= \mathrm{I} - \exp(\mathrm{A}\,\Delta) \\ &= \sigma(z(x)).\end{aligned}$$

[8] In Mamba [Gu and Dao, 2023], B is discretized through a simplified Euler (as opposed to zero-order hold) discretization from empirical observations of A being more important than B, and the performance does not change significantly with simplification on B.

Table 8: **PG19 dataset statistics.** Split-wise UTF-8 encoded byte L_B and space-separated token counts L_T in the PG19 dataset. (The byte count includes the newline character.) We also indicate the associated bytes per token L_B/L_T.

	L_B	L_T	L_B/L_T
Train	$11,677,824,216$	$1,973,048,393$	5.92
Validation	$17,733,002$	$3,007,061$	5.90
Test	$41,289,101$	$6,965,511$	5.93

Using $\overline{\mathrm{A}}$ and $\overline{\mathrm{B}}$ from above in the discrete recurrence (2), the selective SSM takes the form of a 1D gated recurrence:

$$h[k] = (1 - \sigma(z(x)))\, h[k-1] + \sigma(z(x))x[k]. \tag{4}$$

It is interesting to note from (4) that $\lim_{\Delta\to\infty} h[k] = x[k]$ and $\lim_{\Delta\to 0} h[k] = h[k-1]$: a large Δ ($\Delta \to \infty$) denotes the evolution of the system to focus only on the current input and forgetting the state. In contrast, a small Δ ($\Delta \to 0$) represents a transient input being ignored.

Selectivity of A, B, and C matrices. Gu and Dao [2023] argue that since the system matrix A only affects the model through Δ, i.e., $\overline{\mathrm{A}} = \exp(\mathrm{A}\,\Delta)$. Hence, the selectivity in Δ is sufficient to ensure selectivity in A.

While the selectivity in Δ enables selectivity in the input matrix B, Gu and Dao [2023] hypothesize that making B and C selective (in addition to Δ) would allow for more fine-grained control based on the content $x[k]$ and evolving context $h[k]$.

E Evaluation metrics

Subword-based language models [Vaswani et al., 2017, Hawthorne et al., 2022, Hutchins et al., 2022] report their performance in word-level PPL, while byte-level language models [Xue et al., 2022, Yu et al., 2023] report theirs in BPB. To facilitate meaningful comparisons, we report performance in BPB when benchmarking against byte-level models and PPL when comparing to token-level models. In this section, we detail the conversion between word-level PPL and BPB.

Irrespective of the underlying segmentation, the amount of information $I(D)$ in a given dataset D is constant. Simply put,

$$I(D) = L_T \text{ bits per token} = L_B \text{ bits per byte} \tag{5a}$$

$$\triangleq \frac{-\ln(D; \text{model})}{\ln(2)}, \tag{5b}$$

where L_T and L_B are the length of the dataset in tokens and bytes, respectively. From (5), we observe:

$$\text{BPB} = \frac{-\ln(D; \text{model})/L_B}{\ln(2)} = \frac{\ell_{\text{byte}}}{\ln(2)},$$

where ℓ_{byte} is the observed byte-level negative log-likelihood loss (computed using $\ln$). From (5), we also note the following conversion from BPB to word-level PPL:

$$\frac{-\ln(D; \text{model})/L_T}{\ln(2)} = \frac{L_B}{L_T}\,\text{BPB} = \frac{L_B}{L_T}\frac{\ell_{\text{byte}}}{\ln(2)}$$

$$\Rightarrow \text{PPL} = \exp\left(\frac{L_B}{L_T}\ell_{\text{byte}}\right) = \exp\left(\frac{L_B}{L_T}\ln(2)\,\text{BPB}\right).$$

For the PG19 dataset, we train MambaByte-972M to minimize BPB over the training data and report word-level PPL on the test data. Split-wise values of L_B/L_T for the PG19 dataset are tabulated in Table 8.

F PG19 generation samples

This section includes a few sample generations from the MambaByte-972M trained on the PG19 dataset. We use Nucleus sampling with $p = 0.98$ [Holtzman et al., 2020] and generate continuations for a total of $8,192$ bytes (including the given context prefix). Furthermore, we chose the same test set prefixes used in Appendix F of Rae et al. [2020]. We observe that the model is able to continue the dialogue in the style of the prefix and effectively recall the character names over hundreds of bytes.

Baby Mine by Margaret Mayo

Context (487 bytes):

> CHAPTER I
>
> Even in college Alfred Hardy was a young man of fixed ideas and high ideals and proud of it.
>
> His friend, Jimmy Jinks, had few ideas and no ideals, and was glad of it, and before half of their first college term had passed, Jimmy had ridded himself of all such worries as making up his own mind or directing his own morals. Alfred did all these things so much better, argued Jimmy, furthermore, Alfred LIKED to do them–Jimmy owed it to his friend to give him that pleasure.

MambaByte-972M:

> It was not that Alfred really considered his friend an ideal husband, but Jimmy admitted that it looked differently. And, too, it was all there was to life.
>
> And on went Alfred, through long years of honest comradeship and unblushing promises and bubbling kindness. In his heart he had long since forgotten Jimmy and his friendship, since he realized the superiority of Jimmy's ideal, if of any one's.
>
> It was to be expected that Prince William would fall a victim. Perhaps he was a modified betting man, and then had thrown over Nature and his team-mates, as a victim to ridicule or a follower of hazardous ways. It was then that Prince William went back into society, perhaps sought an interview with Miss La Roche, and learned that he had been the victim of romance death, and a suicide, and had been endowed with an absurd roman cross. Whatever it was, he saw in it the myth of Alfred's heroic deeds or the myth of a yellow dog in the market plate. It sobered him gradually. He became nearly a convert, and it was Alfred who, in bidding Miss Rossiter good-by, said that he would never forget her, and never would be friendly with his friend's family. It was not till now that Miss La Ross realized that a really good life results in an excellent impulse, but the sequel proved the truth of Miss La Ross's predictions. One night, a year before Prince William's death, when she was wearing a diamond star in her hair, and admiring it, she heard Jimmy's front door open in the hall, and, turning round to look through the open doorway, in came Jimmy, and after that where Jimmy went Alfred would sometimes go to the house next door, and they went over this now, and ever since Miss La Ross's spirit had been there, and Jimmy had often insisted that a marriage should be accomplished in some ancient house–"And we will found a family of our own," said Jimmy, "a for some distant house in this great Belgium."
>
> The days went on and it was a glorious spring morning. Early in March a thick autumn rain had fallen on the house in New York and the leafless trees looked weird and strange, and the wind roared like a crested wave

of fire. In the afternoon the thunder roared and soon an occasional gray drizzle eddied about the wet streets and the city glow and tint began to merge into a world of pink and red and amber and purple and sunset colors.

The low ceiling pillars had covered their treasures of gold and gems, and the great stained-glass windows in dreamily gentle, half-lights, had arranged themselves in red and gold stripes of magic tints. Even though the delicate color scheme had had its disappointments once, it had certainly affected one hour in this room as well as ten months later and longer stays at least would have done.

To-day a soft South wind had drifted through the open door, and a sudden east wind had driven open the French windows of Miss La Ross's bedroom, and it rained in pure flames between the ceiling and boarded floor. Alfred's room was fragrant with his presence.

"... A little singing bird
That, living in a cage, demands a friend
Whose cool-blooded lure can warm the heart with love
To a fluttering, wounded thing.

As in a pathetic garden, so in the hall room.

It was cold, to-day. Already the rooms seemed overheated. The curtains were already half drawn.

She shivered.

"Mid-winter, to-day," thought Alfred, watching the sweep of Ann's white shoulder and patting her thin cotton frock. "Seven there were of days. And seven is yet untold gone. Fine, fine day, by Christ! Come out of this old soot, and we'll fly... Away. God rest his soul from hell, if ever such a devil crawled this broad, raw earth.... Where are you, Ann?"

Ann waited and trembled, she knew not why, for a sharp voice was asking suddenly for the check book in her hand.

"Get me change enough to pay for lunch for Jimmy," Alfred chided.

Before the one empty chair on the hall table and under the curtains lay a crashing pile of ready money. "And the window shades are closed," added Alfred.

"It won't shut out the rain," smiled Ann.

"But he won't care," protested Ann.

Alfred laid a strong withdrawing hand on the fair golden hair for a moment.

"It's all right," he coaxed. "Without a cent behind them to-day we can put in four thousand and close the bottom against a falling price like this." He was holding up the window sill six inches.

While he stood she whispered:

"I'm only lucky to save the day."

"He helps you without a reward," Alfred said.

"He's kind... and darned bad."

Ann noted dangerous things that afternoon.

"You could sing and play?" she asked.

"No, no!" insisted Alfred. "I CAN'T play and sing. The room is cold. It's warm within."

Alfred was changing clothes when he had that lucky escape, and Alfred momentarily forgot his debt. Ann laid the bill she had placed on the table, and when she had gone Alfred had not even looked at it, and it was the act she saw in that frame of mind, remembering it, that made her put it back again.

Now Alfred was thoroughly cold and temperamental, and when he probed an obligation that he had just been trying to shift on the other fellow, he was more easily reminded. When Jimmy, cold and hungry, had wormed his way into his room that day at dinner, and been halted at his close chair by the soup stove, the young man's gaze had fixed furiously to the platter of gold and had immediately started on the other food with an intensity of expression that had awakened Jimmy's appreciation of the hot day of purposes and had aroused even Ann's observant sense.

Jimmy's employer had met him on Close Street after the unsuccessful row over the Dearborn Cats. Jimmy, who was not naturally an observant boy, had tried to keep in the line of his employer's movements and tell Alfred his employer just what he did for a living, but all Alfred's energy had vanished, and on sundry occasions he had caught Jimmy's eye, and once he had promptly appeared to mere assiduous examination of the window. Employer's Jimmy had been dexterous enough, subdued, but his dexterity and subtlety and sagacity had not failed.

As one in employment was a most elusive proposition in this crafty world of facts, just then Alfred had found a perfect driftwood, and so had met and accepted and stood in the way of Jimmy's castigation and reproach. That is to say, he had saved Jimmy from seeing any of his own real qualities, and the critics, he had been asked in Jimmy's more frequent matinees to erase Alfred's sneer and snip off his coat, and he had instantly become a mental picture of Jimmy Dean's assistant to the lawyer and the college professor.

It was Jimmy's reckless impetuousness, not his single fearless single energy, that had led Ann through the door at sight of Ann, that had electrified the tremendous audience, not her own act or attitude. Jimmy had thought still of the boy as a fellow mortal, now his master had gone.

That was a satisfactory driftwood, of Jimmy.

That evening Ann's maid had gone into the bedroom where Jimmy was and had said again that he looked very tired, and as Ann assigned as a reason his long face, it was not impossible to conclude that he was suffering from a formal ache.

But, as she had only managed a nap and had been rather silent when away from Jimmy, Ann determined that the tired-out and worn-looking man must

be tired, too, and she secretly determined that at breakfast to-morrow
she would place in his hands all the evidence she could imagine against
Jimmy, as pertaining to the unsolved mystery of Jimmy Dean's friend.

CHAPTER II

That afternoon Jimmy was taken into the professor's home to dinner.
Wonderful Dean's Hal

The Diary of Samuel Pepys

Context (826 bytes):

June 1st. Having taken our leaves of Sir W. Batten and my Lady, who are gone this morning to keep their Whitsuntide, Sir W. Pen and I and Mr. Gauden by water to Woolwich, and there went from ship to ship to give order for and take notice of their forwardness to go forth, and then to Deptford and did the like, having dined at Woolwich with Captain Poole at the tavern there. From Deptford we walked to Redriffe, calling at the half-way house, and there come into a room where there was infinite of new cakes placed that are made against Whitsuntide, and there we were very merry. By water home, and there did businesses of the office. Among others got my Lord's imprest of L1000 and Mr. Creed's of L10,000 against this voyage their bills signed. Having wrote letters into the country and read some things I went to bed.

MambaByte-972M:

There came my ship in this morning, which had safely arrived after all our troubles, though some to this day we thought had been lost, viz., of the north winds. Whereabouts of us, and going to find them, we had much ado to get up in the day time without much losing our mast and going about by night. At noon came the King and Queen and all their Court (unhappily in some pain from the swelling of the belly). There were myself with others going to dinner to meet them; and my Lady Sandwich came, by special invitation of my Lady Jem., to dine at Court. She was invited to dinner afterwards by Mrs. Penington; and after dinner I going out to her coach, we fell to talk of the happiness of this, and being astonished at the curiosity and ingenuity almost of the age and remarkable folly of the age for things of this kind: "Many good fortunes," says she, "will happen upon this discourse; your lady (I know) hath the choice of your person taken in some few; but you, charming and ingen as you are, I doubt you will not make an end." "Bah!" says I, "it is good dallying with the ladies; for if once off of her, I promise you her talking and witty company, and you could not choose but have her."

At dinner we staid some time, discoursing about her forwardness as to her coming over from her beloved Deptford to be Lady. My Lady is so tender, that I did not help to give her opportunity of shining a little, by having been called out to dinner to Sir R. Holmes before I set out, it having been proposed to my Lady to stay and go as far as Gravesende; of whom she was very much pleased, even to the skirt of her gowne. But, when I came back I found the strictest inquiry made after me, I coming over to Hinchinbroke; and my Lady cryed out of my going into the coach, lest she should be made to say I stayed for her. Poor little wench, I confess I was in a great confusion, being to see my Lady set out without me; which, if I had known, I had wisely taken a resolution not to have been of it. At last it was agreed between Sir W. Batten and my Lady that the maid should go herself in a boat, and my Lady his wife to go in another boat with Mrs. Penington, who had a great mind to go. My Lady had thoughts of sending Jane for her. Which we were loth to, the singleness of their minds and opposition they meeting with no manner of business at home, and being desired to do so by my Lady, did instance me to get Jane and little Mrs. Penington to stay with me, and myself to lodge and launch two boats more, for the building now and to study the disposing of them; but, being there, occasion will not suffer me to go upon them, but I leave it wholly to their own pleasures. The day proving most pleasant all the afternoon, so that it was agreed on, and even our preparations done, my mind all over afeard and discouraged, as if things would not have been better. Upon our sending away this poor wench (my Lady sitting looking on) I rose up and took my wife and my son T. Greaves as proxies for her to go by herself, in my great coach in the Towre garden, in the last coach my Lord will lend me; which coach is well chosen and good. So giving order for the horse to be chosen neatest and whitest, we set forth; Mrs. Penington on the box with me scolding at him all the way; my great discourse my losing the fine pearl, where I had much ado to keep from crying, such a jealous pain she put me to. Away to Lambeth chappell; where, I perceive, my Lord did give his desire to have understand who are to sit in his chappell, he standing up all the sermon. This continued most part of the sermon, and in our going over the bridge above bridge, I bethought myself that my Lord must have grounded his resolution of having it so. He seemed therefore to give very civil direction at cards after sermon in the evening, to my Lord's great satisfaction; but on the reading of the Breviate he was as offended as I could be of the alteration (enquiring too deeply into the goodness of the gentlemen in their not desiring to have a further change) which my Lord did very readily, and told me he fell very off; and, losing all he had, and taking nothing, sighed very heavily and mournfully, and so went away. The business was not made much merrier for this sermon; for which I was sorry, and remaining professedly sad we rowed to Kingston, and after some solemn service we landed there and went up to a little alehouse some few miles up yet Stoul. But, Lord! the humours of boys, which I never knew in my life before, how fretting, tearing and swearing, and so I made their bellies fret presently, no supper nor ever man of the house sitting down as I expected. So ended this businesse with joy beyond what was hoped for it the first time. To Stoul and back again.

The Patrol of the Sun Dance Trail by Ralph Connor

Context ($1,059$ bytes):

High up on the hillside in the midst of a rugged group of jack pines the Union Jack shook out its folds gallantly in the breeze that swept down the Kicking Horse Pass. That gallant flag marked the headquarters of Superintendent Strong, of the North West Mounted Police, whose special duty it was to preserve law and order along the construction line of the Canadian Pacific Railway Company, now pushed west some scores of miles.

Along the tote-road, which ran parallel to the steel, a man, dark of skin, slight but wiry, came running, his hard panting, his streaming face, his open mouth proclaiming his exhaustion. At a little trail that led to the left he paused, noted its course toward the flaunting flag, turned into it, then struggled up the rocky hillside till he came to the wooden shack, with a deep porch running round it, and surrounded by a rustic fence which enclosed a garden whose neatness illustrated a characteristic of the British soldier. The runner passed in through the gate and up the little gravel walk and began to ascend the steps.

MambaByte-972M:

Indescribable and yet unconscious was this man. All his being rose to a sense of the security that was about him as he walked onwards swiftly between rows of young firs of Canadian growth. He swung impulsively by the gap of a gate and paused. Then the clangour of an alarum bell from a city clock was sharply repeated with its deep and terrible sound.

This bell not only announced the desire of the inhabitants of the intervening concession lines to their work, but it also warned from the further side of the high boundary constructed by the late engineer of the West Mounted Police the abrupt and rapid career of another man. The pistol had fallen, and the runner was searching for another.

The solicitor buried his face in his hands.

"In God's name, man!" he ejaculated. "How do I know? Haven't I just found one?"

He had rung the bell, and now fell back among the shadows and the rough bushes. The sound startled him. He stood transfixed a moment before he looked about him.

"There's only one man knowing where that tote-road goes." He still held his breath and then listened. "'Tis where the first news comes from, and–oh, I see! There's my boy."

He waited a moment. Then once more he passed along the gap and descended the rocky little to the more sheltered spot where the new buildings stood.

He was an under officer in that high branch of the service known as the Union League, and a short sharp beat was his upon the white drilling of bashi-bazouks. With the sounding of another great bell of a church close at hand he moved quickly round to the other side of the buildings. As he approached, however, he took from his pocket a thin black silk neckerchief. It was damp and stained with the blood of dead men. He laid it in the hands of a slim girl, with the limpid blue eyes of the Canadian Saskatchewan.

"What's that for?" he demanded.

She looked as if there had been something she desired to say, then left the agitated conclusion unfinished. Her eyes sought his in the pathetic wistfulness of a child, then suddenly fell. For the hurt he had done her was not a wound incurred in battle. It was merely a little scratch in the hand, and let alone that, in a manner of speaking, it was all she had. The blood of a man is always more significant than that of a scratch on the bark of a tree, and a pressure of the earth leaves a deeper mark on a man's arm. With a sigh the runner removed the blood stain and turned his face towards the sound again. He walked half across the open grass from which he had sprung. From his ample form to the far-distant leaping folds of his drilling trousers he had trailed a forked stick, and so to the girl.

In a few seconds he came back.

"It's me, pardner, Superintendent Strong. It's me I'm goin' down from the Soo, for the job I had in Mexico after I came out here. I'm connected with the Canadian Pacific Railway and they're hunting up a man who did have a finger wounded by a Canadian rock. I'm sendin' the little flag with her." He emphasised the word "flag." A rough skin mark, furrowed in a straight line down his left cheek, marked the place of the scar and brought him to a sudden stop. His eyes were on the scrolled letters above his head.

"I'm going down to get it. I've got to get it to the bottom, anyway, for divil a bit of paper they'll let me have at British Columbia. Oh, God!"

He raised his voice. In a moment he had departed. In a few minutes he had rejoined the girl. They rejoined the solicitor and returned with him to an open space before the meeting place of the railway company. As they gathered round a table spread with an untasted meal the solicitor spoke. The railroad company was working out from British Columbia to Montreal.

"In our fight we had it hard," he said. "The northern route to League Island was blocked, we could not reach there to recruit. We had to look for a northern route, for there was none. At first the league flag of Ottawa was given up. That was only till October. Then a young man on the ground from London came to us. He'd been in the runner's service along the whole line from Montreal. He was headed for Canada on the telegraph. Two of us had to flag him as soon as we set out from here. He had been over that ground about fifty times before, and knew the whole road well for forty miles. The head of us did not know it till he came to the junction where the main line crosses the north line of the United States. We took that name on the tin to test him."

"What was the corporation over there for?" said the solicitor. "I remember, I remember. It occupied a part of the big Kelvin mine. I was helping get the first claim post run by the Union League at the time I was there. He was out hunting coal. He came down one day to see the coal pits about the ground. On the way he was stopped and accused of raising a rebellion, and was arrested and taken to the Soo, where he was made to give evidence in a certain case that had been laid before him."

"And what was the precise cause of the complaint?" asked the runner.

"Well, it wasn't a case at all, it was a fact. That's all," explained the constable.

"From what I heard then of the runners of the London and North West, their work wasn't near so exciting and dangerous as it had been reported to be. Also it was the work of others, others still, and they were arrested. They was a young feller and a girl married over two years ago, and he was shot."

"Brought to trial for that by himself or his relatives or some of the men who were with him?" There was a puzzled, gentle expression on the face of the railway superintendent. He was of much higher rank, for he had not been present at the trial of the accused. He glanced up at the runner.

"Arrested?" The bit of food in his mouth was working like a millstone in the Soo employer's breast. Then, as though unconsciously to himself, his lips said "yes" instead of "no," and he added instead, "and sworn to it. That's as far as you've got, pardner. Anything else, sir?" He was watching the silent figure with intense desire to see his face and to know what he felt. It did not come, and he settled himself in his chair with a sigh.

"That was short work. They marched the young feller up here, and give him the Canadian division. It was the station sergeant-inspector from the Canadian line sending down from headquarters to show he was all right and not having heard anything against him. And if you don't know that it's not the worst of the testimony we have to give, pardner. It wasn't the best. The fact is the young man was getting three weeks' sentence at the time."

"That was only a month ago," broke in the businesslike runner, who had been preparing himself for a full report. "What had he done? Tell us?"

There was something pathetic in the voice and in the manner of the young man. Then, as he mounted his story, the under-officer took up the thread in an apologetic tone, but was brought back to a moment's serious interest by the stopping of it by the voice of the other.