Fast and Low-Cost Genomic Foundation Models via Outlier Removal

Haozheng Luo*, Chenghao Qiu*, Maojiang Su, Zhihan Zhou, Zoe Mehta, Guo Ye, Jerry Yao-Chieh Hu, Han Liu

Northwestern University, Computer Science

https://arxiv.org/abs/2505.00598

ICML2025

Problem: Transformer-based Genomic Foundation Models (GFMs) encounters outlier-inefficient in quantization and fine-tuning.

Proposal: Fast and Low-Cost Genomic Foundation Models (termed GERM) via outlier-removal architecture and continual learning.

- Serves as an outlier-free model structure to address and mitigate outliers introduced by pretrained models and low-rank adaptation,
- Retains and improves the desirable properties of GFMs in quantization and low-rank adaptation,
- All DNABERT fine-tuning tasks finish in only 5 minutes on a single NVIDIA GeForce RTX 2080 Ti GPU.
- Achieves average performance improvements of **37.98%** in finetuning and **64.34%** in quantization.

Outlier: In GFMs, tokens or activations that disproportionately influence the attention mechanism with:

- Tokens with little or no meaningful information receive disproportionately high attention weights.
- Recurring nucleotide patterns are overemphasized by $\mathrm{Softmax}$.

GFMs: Large-scale pretrained models designed for modeling and analysing genomic sequences.

- Trained on massive genomic datasets
- **Classification models**: e.g., `DNABERT-2`, `Nucleotide Transformer (NT)`, `HyenaDNA`
- **Generative models**: e.g., `Evo`, `GenomeOcean`
- Larger GFMs, especially generative models, require substantial computational resources for deployment and fine-tuning.

- We propose a new GFM architecture GERM by replacing the $\mathrm{Softmax}$ in the attention mechanism with $\mathrm{Softmax}_1$ to achieve the Quantization Robustness and Fast Low-rank Adaptation.

$$\mathrm{Softmax}_1(S) := \frac{\exp(S)}{1 + \sum_{i=1}^{L} \exp(S_i)},$$

- The original OutEffHop method requires training from scratch; we propose a trade-off variant, GERM-T , to achieve sub-optimal performance with small-step continual learning.

Compare GERM with the vanilla attention on DNABERT-2 in quantization setting.

Model	#Bits	Quantization Method	MCC (↑)	Delta MCC (↓)	Avg Performance Drop (↓)	Avg. Kurtosis (↓)	Max inf. norm (↓)
Official	16W/16A	-	66.11	-	-	39.68	53.61
DNABERT-2	16W/16A	-	59.11	7.00	-	270.90	61.64
	8W/8A		33.60±0.41	32.51	43.81%		
	8W/8A	SmoothQuant	36.51±0.02	45.37	38.63%		
	6W/6A		20.74±0.04	45.37	66.18%		
	4W/4A		-1.03±0.06	67.06	101.24%		
	8W/8A	Outlier	25.26±0.02	40.85	57.60%		
	6W/6A		27.84±0.28	38.27	52.71%		
	8W/8A	OmniQuant	49.92±0.05	16.19	15.76%		
	6W/6A		48.47±0.14	17.64	18.61%		
	4W/4A		2.94±0.19	63.17	94.78%		
GERM	16W/16A	-	59.73	6.38	-	**21.29**	**10.62**
	8W/8A		57.30±0.08	8.81	**3.77%**		
	8W/8A	SmoothQuant	56.65±0.15	9.46	4.82%		
	6W/6A		56.48±0.07	9.63	**5.45%**		
	4W/4A		20.05±0.00	46.06	**69.44%**		
	8W/8A	Outlier	45.87±0.08	20.24	**25.23%**		
	6W/6A		40.57±0.56	25.54	36.27%		
	8W/8A	OmniQuant	55.99±0.09	10.12	5.95%		
	6W/6A		55.70±0.03	10.41	**6.41%**		
	4W/4A		49.42±0.00	16.69	**17.17%**		
GERM-T	16W/16A	-	59.30	6.81	-	251.40	28.49
	8W/8A		38.38±0.15	27.73	35.27%		
	8W/8A	SmoothQuant	57.52±0.00	8.59	**3.01%**		
	6W/6A		30.34±0.04	35.77	48.83%		
	4W/4A		0.22±0.00	65.89	99.63%		
	8W/8A	Outlier	42.57±0.05	23.54	28.31%		
	6W/6A		46.02±0.06	20.06	**22.34%**		
	8W/8A	OmniQuant	56.80±0.12	9.31	**4.21%**		
	6W/6A		55.41±0.00	10.71	6.57%		
	4W/4A		3.86±0.00	62.25	93.49%		

Results: GERM achieves an average performance improvement of **64.34%** in PTQ experiments. Similarly, GERM-T shows an average performance improvement of **43.04%** over the same baseline.

Compare GERM with the vanilla attention on NT 2.5B in quantization setting.

Model	#Bits	Quantization Method	MCC	Delta MCC	Average Performance Drop
NT-2.5B-multi	16W/16A	-	56.98	-	-
	6W/6A	-	18.52	38.46	67.50%
	4W/4A	-	1.39	55.59	97.56%
	6W/6A	Outlier	50.23	6.75	11.85%
	4W/4A		40.74	16.24	28.50%
	6W/6A	SmoothQuant	47.23	9.75	17.11%
	4W/4A		35.16	21.82	38.29%
	6W/6A	OmniQuant	49.55	7.43	13.04%
	4W/4A		43.63	13.35	23.43%
GERM (NT-2.5B-multi)	16W/16A	-	57.16	-0.18	-
	6W/6A	-	45.96	11.2	**19.59%**
	4W/4A	-	42.48	14.68	**25.68%**
	6W/6A	Outlier	52.24	4.92	8.61%
	4W/4A		49.00	8.16	**14.28%**
	6W/6A	SmoothQuant	51.95	5.21	**9.11%**
	4W/4A		48.15	31.09	15.76%
	6W/6A	OmniQuant	52.55	4.61	8.07%
	4W/4A		49.26	7.90	**13.82%**
GERM-T (NT-2.5B-multi)	16W/16A	-	56.82	0.16	-
	6W/6A	-	32.58	24.24	42.66%
	4W/4A	-	10.49	46.33	81.54%
	6W/6A	Outlier	52.14	4.68	**8.24%**
	4W/4A		46.24	10.58	18.62%
	6W/6A	SmoothQuant	51.61	5.21	9.17%
	4W/4A		48.12	8.70	**15.31%**
	6W/6A	OmniQuant	52.43	4.39	**7.73%**
	4W/4A		47.28	9.54	16.79%

Results: GERM achieves an average performance improvement of **50.83%** in PTQ experiments. Similarly, GERM-T shows an average performance improvement of **36.73%** over the same baseline.

Compare GERM with the vanilla attention on DNABERT-2 in low-rank adaptation setting.

Models	Low-Rank Adaptation Method	MCC (↑)	Delta MCC different (↓)	Avg Performance Drop (↓)	Avg. kurtosis(↓)	Max inf. norm(↓)
DNA BERT-2	Full	59.11	7.00	-	270.90	61.41
	LoRA	50.91±1.67	15.2	13.87%	-	219.20
	QLoRA	50.65±0.13	15.46	14.31%	292.85	53.91
	LoftQ	50.76±0.06	15.31	14.05%	299.18	54.18
GERM	Full	59.73	6.38	-	**21.29**	**10.62**
	LoRA	57.27±0.70	8.84	**4.12%**	-	**19.41**
	QLoRA	53.16±0.21	12.95	**10.99%**	**34.29**	**27.27**
	LoftQ	53.11±0.08	13.00	**11.08%**	**33.02**	**27.41**
GERM-T	Full	59.30	6.81	-	251.40	28.49
	LoRA	55.60±0.28	10.51	6.23%	-	140.86
	QLoRA	51.05±0.07	15.06	13.90%	287.95	53.92
	LoftQ	51.20±0.13	14.91	13.65%	286.16	53.35

Results: GERM achieves an average performance improvement of **37.98%** in low-rank adaptation compared to DNABERT-2 model. Similarly, GERM-T shows an average performance improvement of **20.01%** over the same baseline.

Compare GERM with the vanilla attention on NT 2.5B in low-rank adaptation setting.

Model	Fine-Tuning Method	MCC	Delta MCC	Average Performance Drop
NT-2.5B-multi	Full	56.98	-	-
	LoRA	53.50	3.48	6.11%
	QLoRA	52.29	4.69	8.19%
	LoftQ	52.89	4.09	7.17%
GERM (NT-2.5B-multi)	Full	57.16	-0.18	-
	LoRA	55.98	1.18	**2.06%**
	QLoRA	55.52	1.64	**2.87%**
	LoftQ	55.80	1.36	**2.38%**
GERM-T (NT-2.5B-multi)	Full	56.82	0.16	-
	LoRA	55.24	1.58	<u>2.78%</u>
	QLoRA	53.32	3.50	<u>6.16%</u>
	LoftQ	53.74	3.08	<u>5.42%</u>

Results: GERM achieves an average performance improvement of **66.02%** in low-rank adaptation. Similarly, GERM-T shows an average performance improvement of **34.56%** over the same baseline.

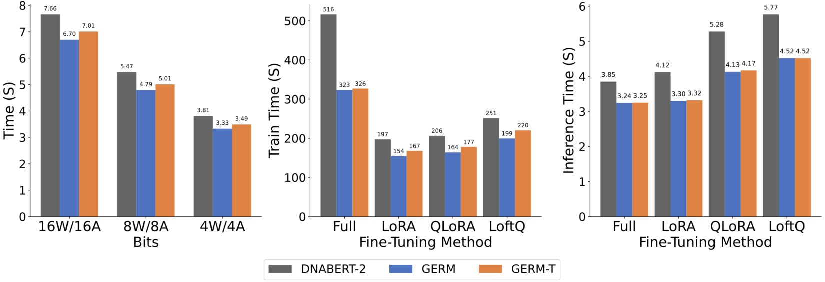
Compare GERM-T with the vanilla attention on various continual learning steps.

Method	Fine-Tuning Method	MCC (↑)	Avg Performance Drop (↓)
DNABERT-2	Full	59.11	-
GERM	Full	59.73	-
Out20k	Full	59.21	-
GERM-T	Full	59.30	-
Out100k	Full	60.56	-
DNABERT-2	LoRA	50.91	13.87%
GERM	LoRA	56.78	**4.94%**
Out20k	LoRA	54.75	7.53%
GERM-T	LoRA	55.60	6.24%
Out100k	LoRA	56.61	6.52%
DNABERT-2	QLoRA	50.65	14.31%
GERM	QLoRA	53.16	**11.00%**
Out20k	QLoRA	50.61	14.52%
GERM-T	QLoRA	51.05	13.91%
Out100k	QLoRA	51.24	15.39%
DNABERT-2	LoftQ	50.76	14.13%
GERM	LoftQ	53.11	**11.08%**
Out20k	LoftQ	50.94	13.97%
GERM-T	LoftQ	51.20	13.66%
Out100k	LoftQ	50.77	16.17%

Results: Our method outperforms the vanilla approach across all test sets. Also, we observe that GERM-T exhibits the most optimal performance drop during quantization and low-rank adaptation compared to other continual learning steps.

Experimental Studies: Comparison of Performance in Resource-Constrained Environments

Compare GERM with the vanilla attention on DNABERT-2 in resource-constrained setting.



Results: Both GERM and GERM-T achieve shorter full-rank fine-tuning times per epoch compared to DNABERT-2. Additionally, the model quantization latency for both GERM and GERM-T is lower than that of DNABERT-2, while delivering superior quantization performance.

Compare GERM with the vanilla attention on DNABERT-2 in CPU-only environments.

Method	Fine-Tuning Method	MCC (↑)	Time (sec.)	
			Train	Inference
DNABERT-2	LoRA	50.91	808.23	29.66
GERM	LoRA	**57.27**	**618.68**	**23.10**
GERM-T	LoRA	55.60	674.40	23.57
DNABERT-2	QLoRA	50.65	516.04	63.17
GERM	QLoRA	**53.16**	**358.34**	**45.28**
GERM-T	QLoRA	51.50	418.13	46.91

Results: Both GERM and GERM-T achieve shorter fine-tuning times per epoch compared to DNABERT-2, with the only exception being QLoRA when deployed, where the time is slightly longer.

Summary

- Fast and Low-Cost Genomic Foundation Models
  - Manages outliers in transformer-based GFMs.
  - Remove outlier in model pretraining and fine tuning period.
- Theoretical Enhancements
  - Provide expressive guarantee of low-rank adaption.
- Small-Step Continual Learning
  - Leverages continual learning to address the training-from-scratch limitation in [Hu et al., 2024].
  - Achieves sub-optimal yet effective performance.
- Empirical Performance of GERM
  - Achieves 92.14% lower average kurtosis and 82.77% lower maximum infinity norm $|\mathbf{x}|_\infty$, enabling robust quantization and fast low-rank adaptation.
  - Improves fine-tuning performance by 37.98% and quantization performance by 64.34% over the baseline.

Thank You!

Haozheng Luo*, Chenghao Qiu*, Maojiang Su, Zhihan Zhou, Zoe Mehta, Guo Ye, Jerry Yao-Chieh Hu, Han Liu

- hluo@u.northwestern.edu
- q1320460765@tju.edu.cn
- maojiangsu2030@u.northwestern.edu
- zhihanzhou2020@u.northwestern.edu
- zoe.mehta@vhhscougars.org
- guoye2018@u.northwestern.edu
- jhu@u.northwestern.edu
- hanliu@northwestern.edu
- https://github.com/MAGICS-LAB/GERM