References

[1] P. Drineas, R. Kannan, and M. Mahoney. *Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication.* SIAM Journal on Computing, 2006.

[2] I. Ipsen and T. Wentworth. *Importance Sampling for a Monte Carlo Matrix Multiplication Algorithm, with Application to Information Retrieval.* SIAM Journal on Scientific Computing, 2011.

[3] N. Halko, P.-G. Martinsson, and J. A. Tropp. *Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions.* SIAM Review, 2011.

[4] G. Metere. *Low-Rank GEMM: Efficient Matrix Multiplication via Low-Rank Approximation with FP8 Acceleration.* arXiv:2511.18674, 2025.

[5] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. *LoRA: Low-Rank Adaptation of Large Language Models.* 2021.

[6] Y. Koren, R. Bell, and C. Volinsky. *Matrix Factorization Techniques for Recommender Systems.* Computer, 2009.