EfficeintML.ai Lab 5 Report

Name:\_\_\_\_\_\_\_\_

1. Loop Unrolling (20pt): Please fill in the starter code in *kernel/template/loop\_unrolling.cc* to implement loop unrolling and run the `./evaluate.sh loop\_unrolling` to evaluate performance improvement.
   1. Please copy and paste your implementation in *kernel/template/loop\_unrolling.cc*: (15pt)

* 1. How does the performance in GOPs, achieved through loop unrolling on your computer, compare to the reference implementation? Please explain the performance difference. (5pt)

1. Multithreading (20pt): Please fill in the starter code in *kernel/template/multithreading.cc* to implement multithreading and run the `./evaluate.sh multithreading` to evaluate performance improvement.
   1. Please copy and paste your implementation in *kernel/template/multithreading.cc*: (15pt)
2. How does the performance in GOPs, achieved through multithreading on your computer, compare to the reference implementation? Please explain the performance difference. (5pt)

1. SIMD Programming (20pt): Please fill in the starter code in *kernel/template/simd\_programming.cc* to implement SIMD programming and run the `./evaluate.sh simd\_programming` to evaluate the performance improvement.
   1. Please copy and paste your implementation in *kernel/template/simd\_programming.cc*: (15pt)
   2. How does the performance in GOPs, achieved through SIMD programming on your computer, compare to the reference implementation? Please explain the performance difference. (5pt)
2. Multithreading with Loop Unrolling (20pt): Please fill in the starter code in *kernel/template/multithreading\_loop\_unrolling.cc* to implement multithreading and loop unrolling and run the `./evaluate.sh multithreading\_loop\_unrolling` to evaluate the performance improvement.
   1. Please copy and paste your implementation in *kernel/template/multithreading\_loop\_unrolling.cc*: (15pt)
   2. How does the performance in GOPs, achieved through multithreading and loop unrolling on your computer, compare to the reference implementation? Please explain the performance difference. (5pt)
3. Combination of All Techniques (20pt): Please fill in the starter code in *kernel/template/all\_techniques.cc* to implement all techniques above and run the `./evaluate.sh all\_techniques` to evaluate the performance improvement.
   1. Please copy and paste your implementation in *kernel/template/all\_techniques.cc*: (15pt)
   2. How does the performance in GOPs, achieved through all optimization techniques on your computer, compare to the reference implementation? Please explain the performance difference. (5pt)
4. Bonus (20pt): Any optimization techniques on your mind? Try to implement them to improve the performance further! If you can further improve the performance compared to the optimized kernel in [TinyChatEngine](https://github.com/mit-han-lab/TinyChatEngine), you can get bonus points here! Each percent of performance speedup equals one point (create a pull request in the repo and get verified by the TA), up to 20 points.