Name of Certifying Engineer(s): Parker Zhang

Email of Certifying Engineer(s): quic\_xiaopeng@quicinc.com

Name(s) of System Under Test: Low Power AI Subsystem on Snapdragon Next Gen Mobile Platform

Division (check one):

* Open
* Closed

Category (check one):

* Available
* Preview
* Research, Development, and Internal (RDI)

Benchmark(s) (check all that apply):

* Visual Wake Words
* Keyword Spotting
* Anomaly Detection
* Image Classification
* Streaming Wakeword

Please fill in the following table adding lines as necessary:

|  |  |  |
| --- | --- | --- |
| System Under Test Name | Benchmark | Accuracy/AUC (FP/FN for SWW) |
| eAI on Snapdragon® Next Gen MTP | Visual Wake Words | 83.5% Accuracy |
| eAI on Snapdragon® Next Gen MTP | Keyword Spotting | 91.3% Accuracy |
| eAI on Snapdragon® Next Gen MTP | Anomaly Detection | 0.86 AUC |
| eAI on Snapdragon® Next Gen MTP | Image Classification | 85.5% Accuracy |

For each SUT, is the benchmark Accuracy/AUC target met? (Not a requirement for the Open division) (check all that apply):

* Yes (Visual Wake Words … 80% Accuracy)
* Yes (Keyword Spotting … 90% Accuracy )
* Yes (Anomaly Detection … 0.85 AUC)
* Yes (Image Classification ... 85% Accuracy)
* Yes (Streaming Wakeword... FP ≤ 8; FN ≤ 8)
* No, for some combination of benchmark, scenario and SUT

For each SUT and benchmark, did the submission run on the whole validation set in accuracy mode? (check one):

* Yes
* No

For each SUT and benchmark, does the submission use either the MLCommons Runner or the EEMBC Runner? (check one)

* Yes
* No

For each SUT and benchmark, is the same code run in accuracy and performance modes? (check one)

* Yes
* No

Are the weights calibrated using data outside of the official calibration set? (check one)

* Yes
* No

What numerics does the submission use? (check all that apply)

* INT4
* INT8
* INT16
* UINT8
* UINT16
* FP11
* FP16
* BF16
* FP32
* Other, please specify:

What backend does the submission use? (check all that apply)

* Vendor backend, please name: Qualcomm AI Stack
* TF-Lite Micro
* Micro TVM
* Other, please specify:

Which of the following caching techniques does the submission use? (check all that apply, ideally none, except for streaming benchmark):

* Caching Inputs between iterations
* Caching responses between iterations
* Caching intermediate computations between iterations

Which of the following techniques does the submission use? (check all that apply, ideally none if submitting to the closed division.)

* Quantization aware training
* Wholesale weight replacement
* Weight supplements
* Discarding non-zero weight elements
* Pruning
* Modifying weights during the timed portion of an inference run
* Hard coding the total number of queries
* None of the above

Is the submission congruent with all relevant MLPerf rules?

* Yes
* No

If the answer to the above question is no, please explain:

For each SUT, have you filled out the JSON system description file?

* Yes
* No

For each SUT, does the submission accurately reflect the real-world performance of the SUT?

* Yes
* No

Does your submission include the following: (check all that apply)

* System description file
* Code that implements the benchmarks
* Code/scripts that train the model(s) (Open Division)
* Metadata that describes each system-implementation combination tested
* Scripts that set up and execute each system implementation tested
* Result logs for each system implementation tested
* This Checklist