N EMPIRICAL VALIDATION OF HIGH SENSITIVITY OF MODEL ACCURACY TO THE ANALOG NOISE OF TOP MaxNNScore EXPERTS

We conduct an extensive experiment to validate the high sensitivity of the model accuracy to the analog noise of the top MaxNNScore experts. To measure the sensitivity of model accuracy to the analog noise of each expert, we only map the expert of interest in analog, while the rest of the experts are mapped in digital. We measure the average accuracy under this mapping for each expert separately on the benchmark LLM tasks. The results are presented in Figure 6-13 for OLMoE. As observed in these figures, the experts yielding noticeable accuracy degradation are also the top MaxNNScore experts. In contrast, the noise-aware ENS baseline, which directly measures the experts' output degradation under analog noise, fails to include some of these critical experts (e.g., expert 27 in layer 4) in its top-k ranking, as the metric ignores the experts' contributions to model accuracy. Note that using the accuracy-level sensitivity itself for expert ranking is prohibitively expensive, as the process requires separate repetitive evaluation of accuracy over multiple random seeds for each expert of the model. For example, our experiment on OLMoE requires 320 V100 GPUs for more than 12 hours. In contrast, the proposed MaxNNScore-based identification does not require any inference on calibration data.

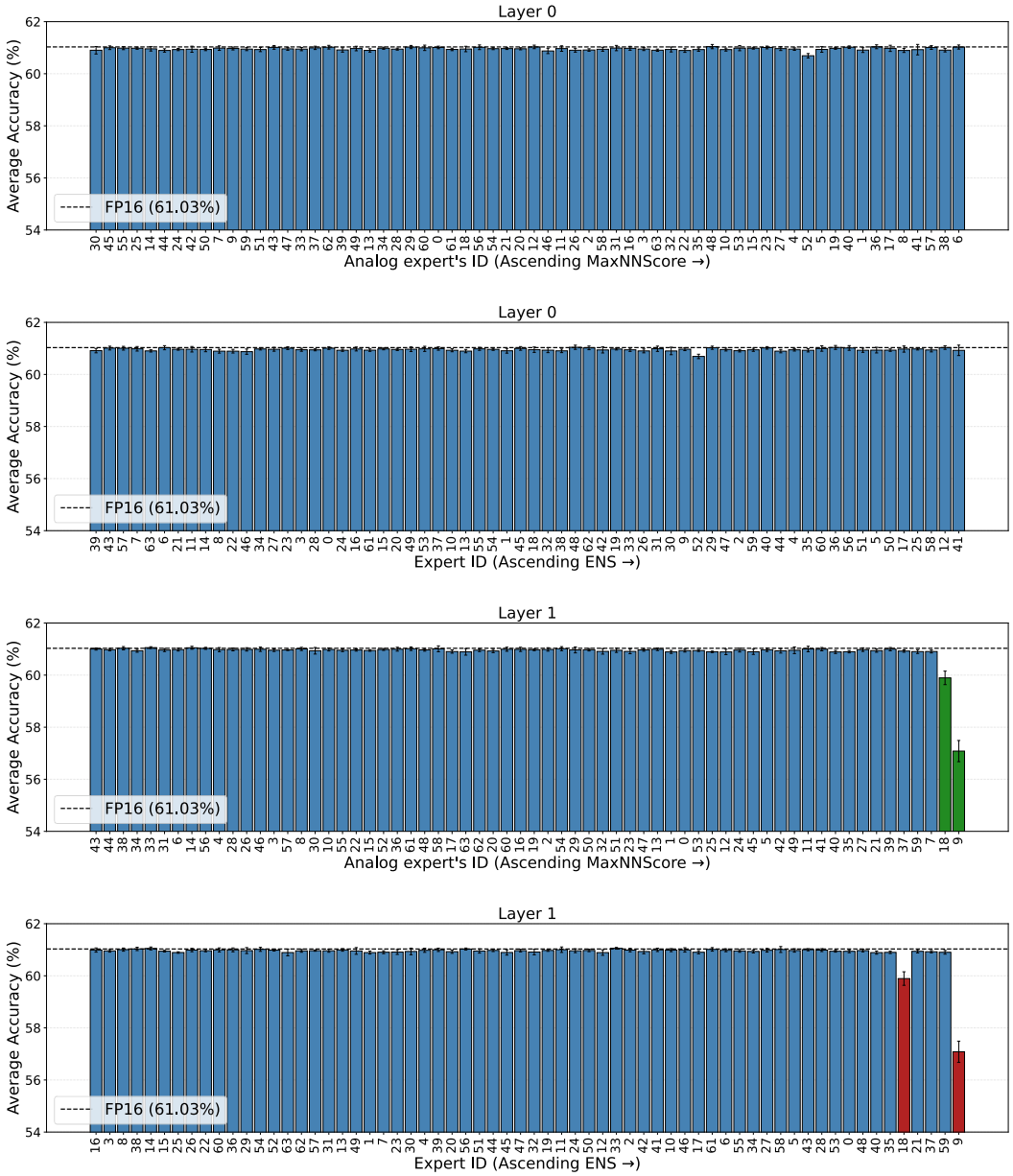


Figure 6: **Model accuracy-level analog noise sensitivity in OLMoE (Layers 0–1).** The noticeable accuracy degradation occurs for expert 18, 9 of layer 1. MaxNNScore requires the top 2 experts, while ENS requires the top 5 experts in digital to include these two experts.

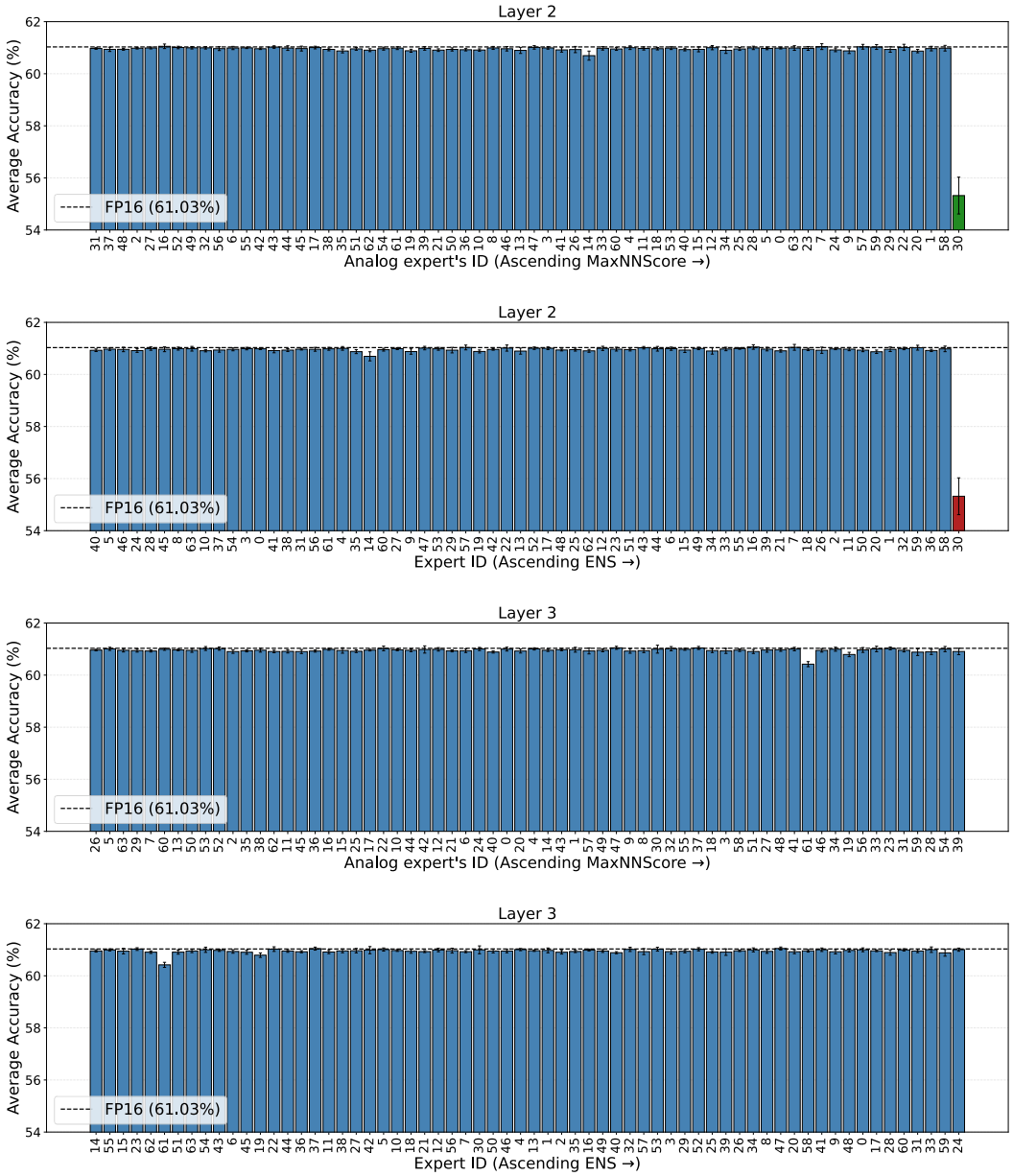


Figure 7: **Model accuracy-level analog noise sensitivity in OLMoE (Layers 2–3).** The noticeable accuracy degradation occurs for expert 30 of layer 2.

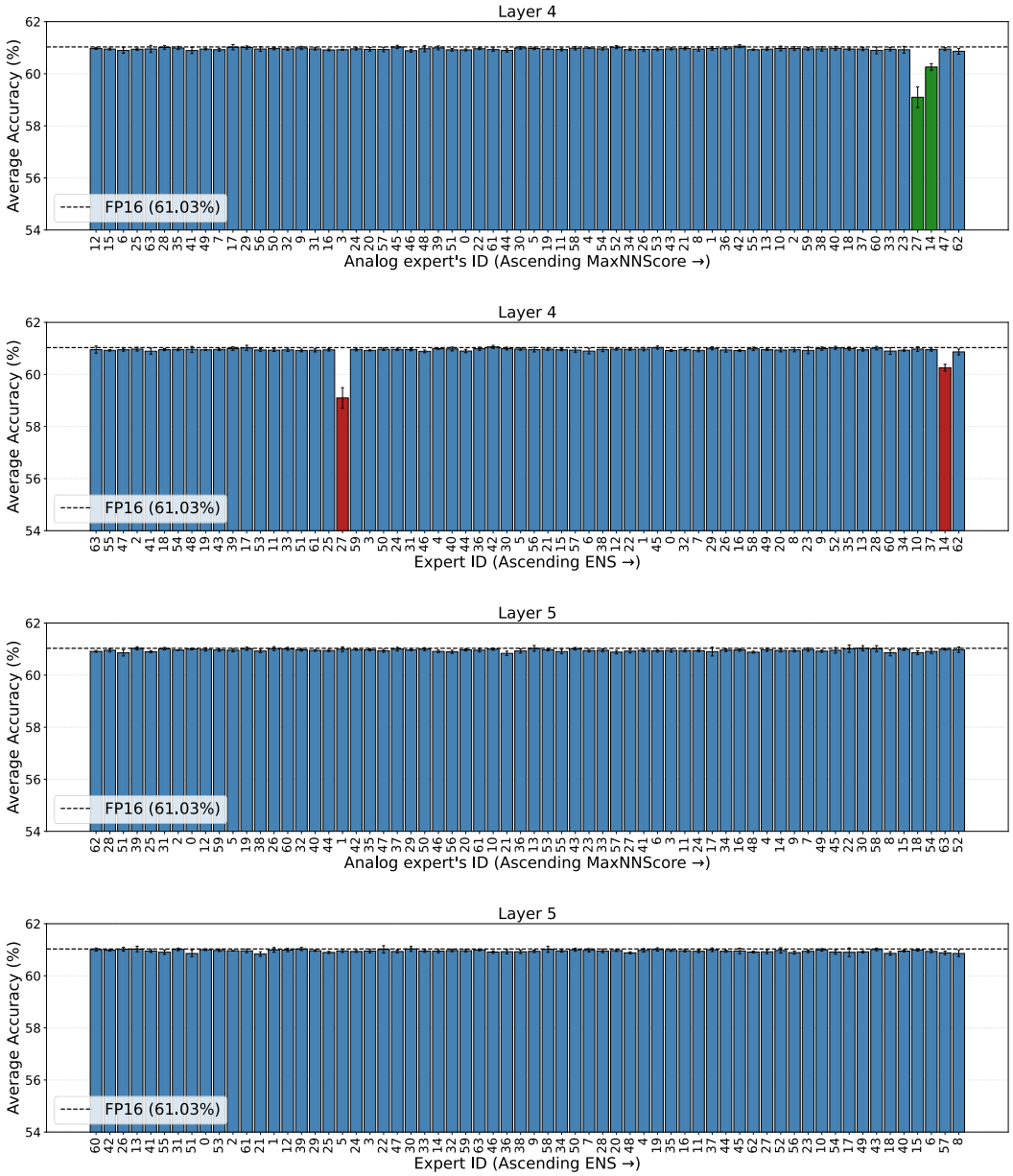


Figure 8: **Model accuracy-level analog noise sensitivity in OLMoE (Layers 4–5).** The noticeable accuracy degradation occurs for experts 27 and 14 of layer 4. MaxNNScore requires the top 4 experts, while ENS requires the top 46 experts in digital to include these two experts.

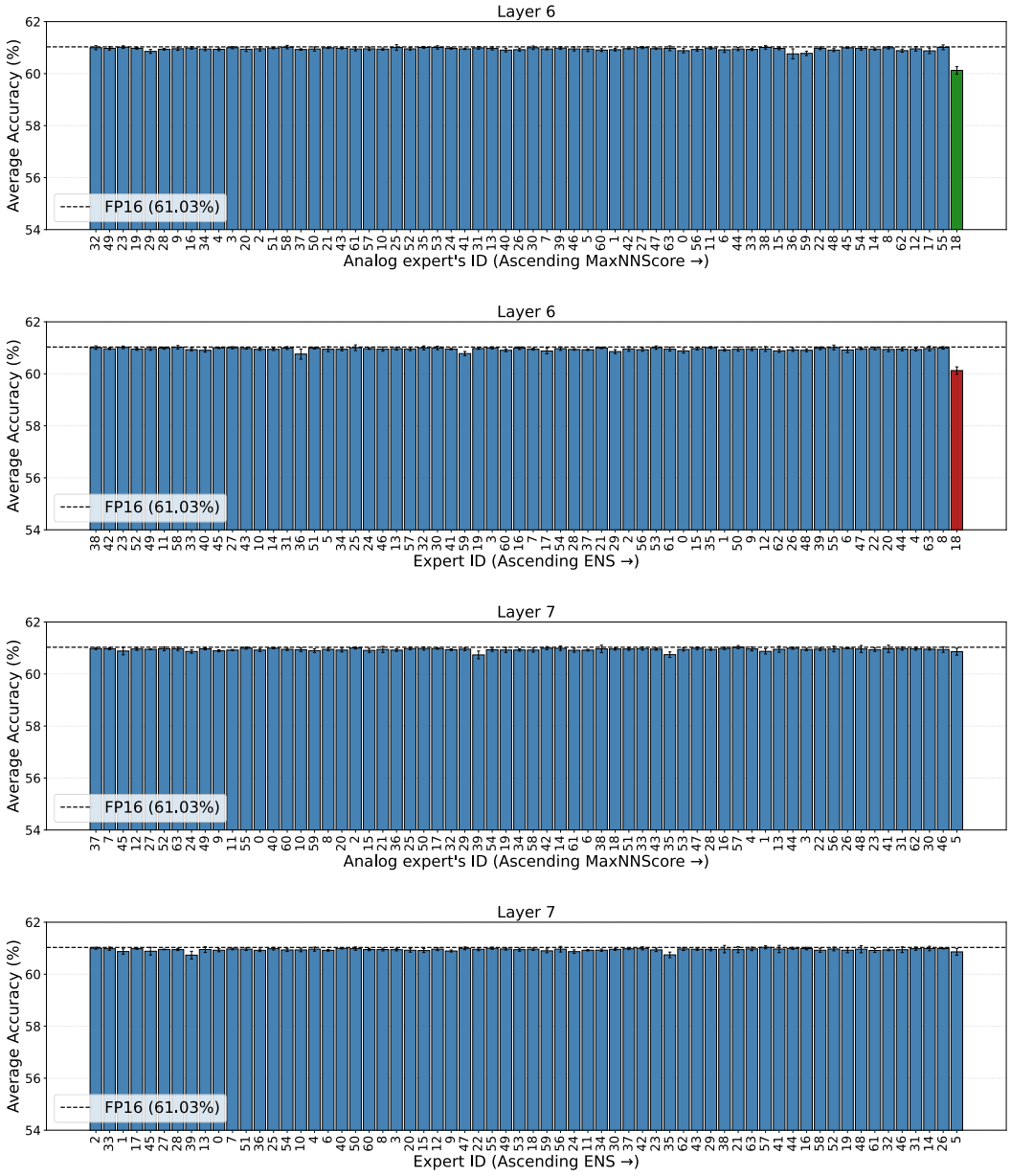


Figure 9: **Model accuracy-level analog noise sensitivity in OLMoE (Layers 6–7).** The noticeable accuracy degradation occurs for expert 18 of layer 6.

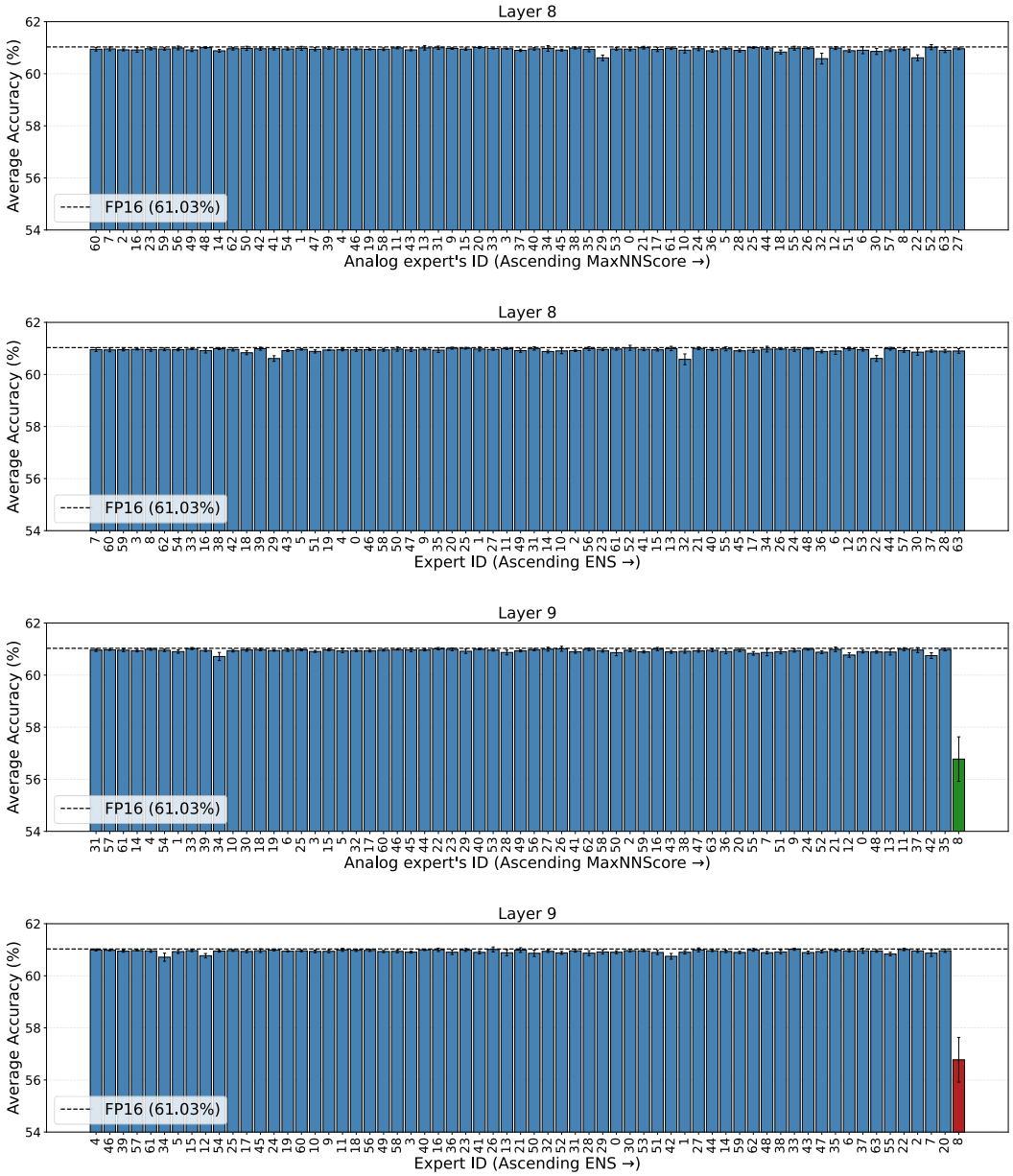


Figure 10: **Model accuracy-level analog noise sensitivity in OLMoE (Layers 8–9).** The noticeable accuracy degradation occurs for expert 8 of layer 9.

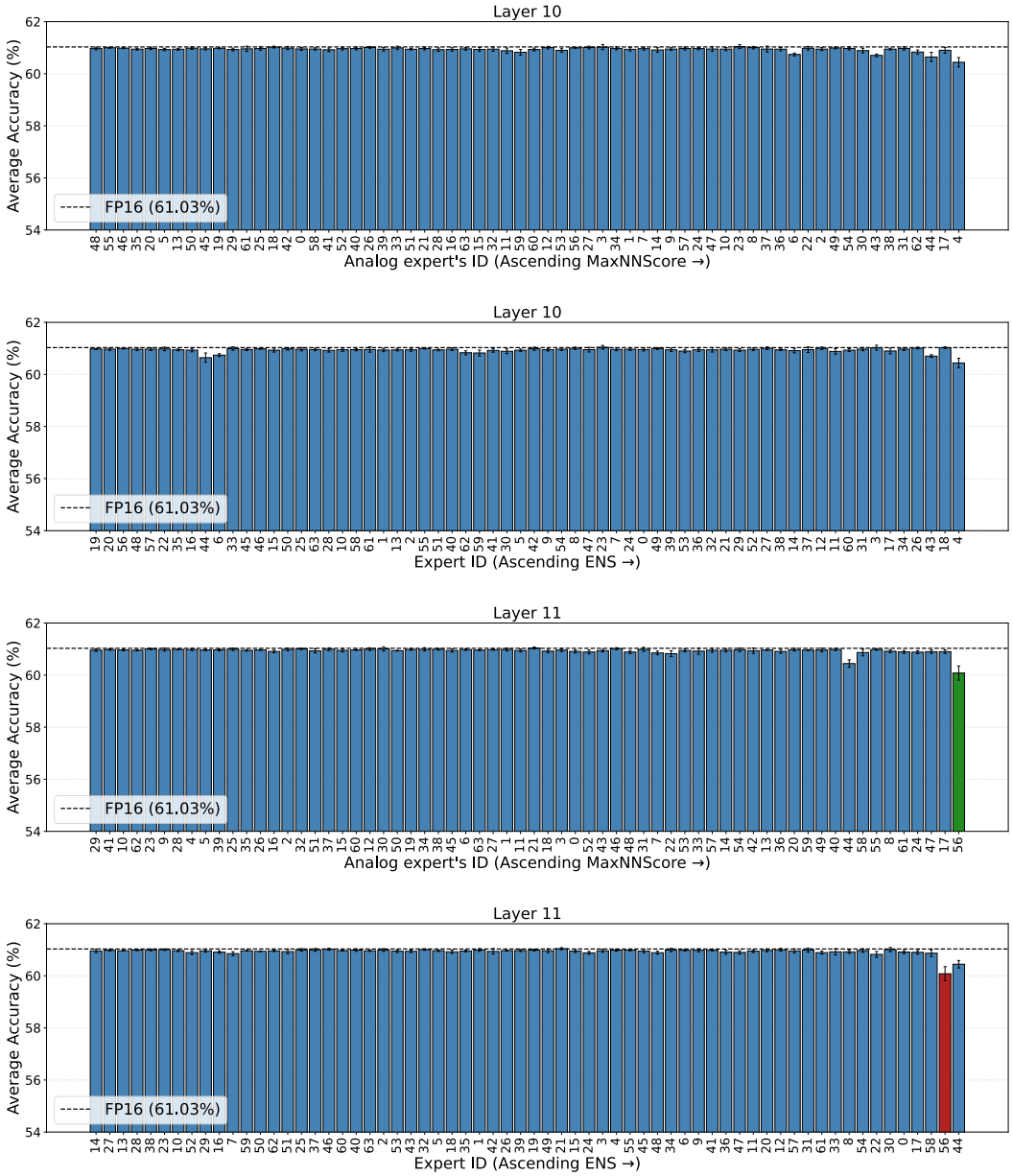


Figure 11: **Model accuracy-level analog noise sensitivity in OLMoE (Layers 10–11).** The noticeable accuracy degradation occurs for expert 56 of layer 11.

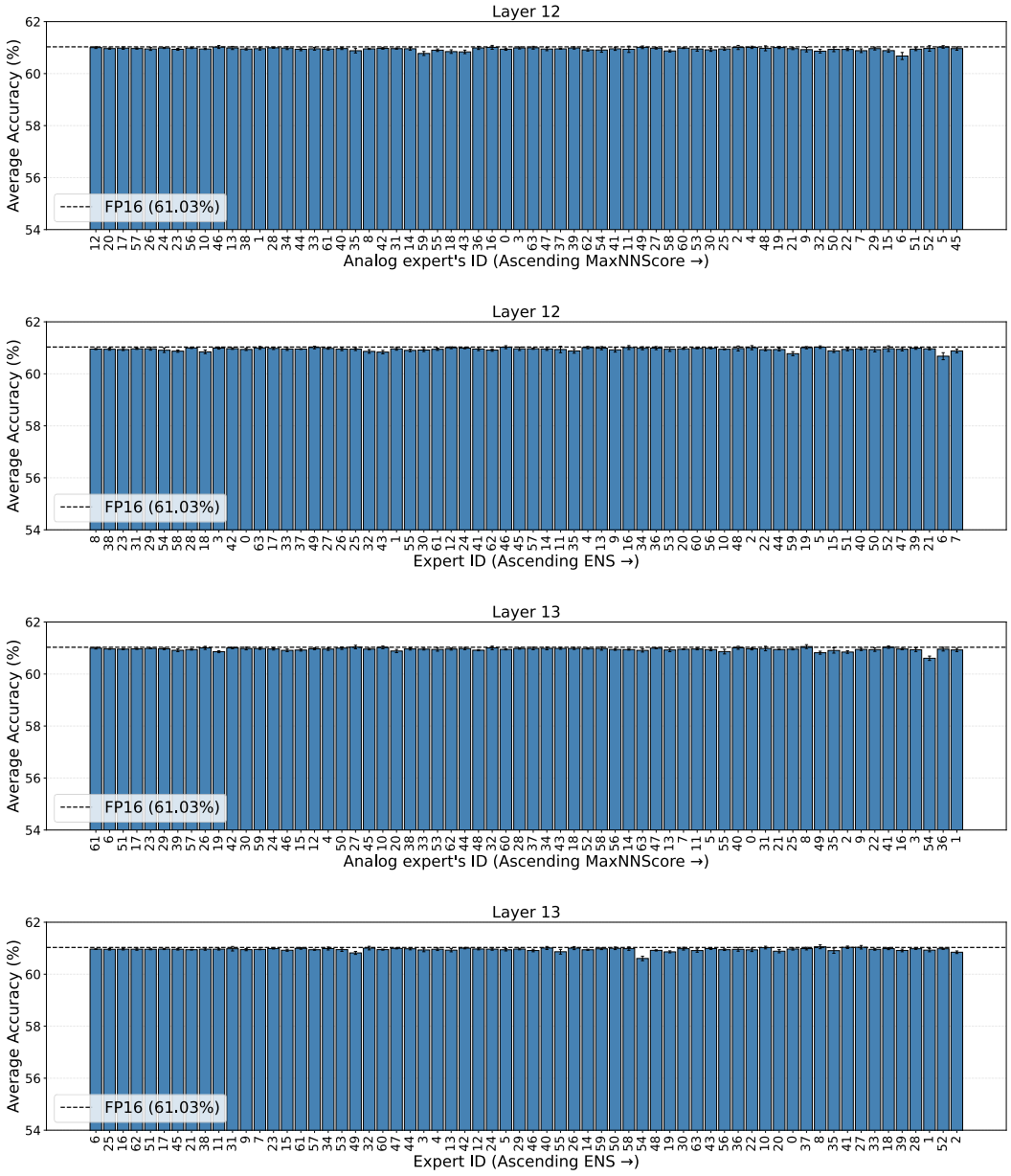


Figure 12: **Model accuracy-level analog noise sensitivity in OLMoE (Layers 12–13).**

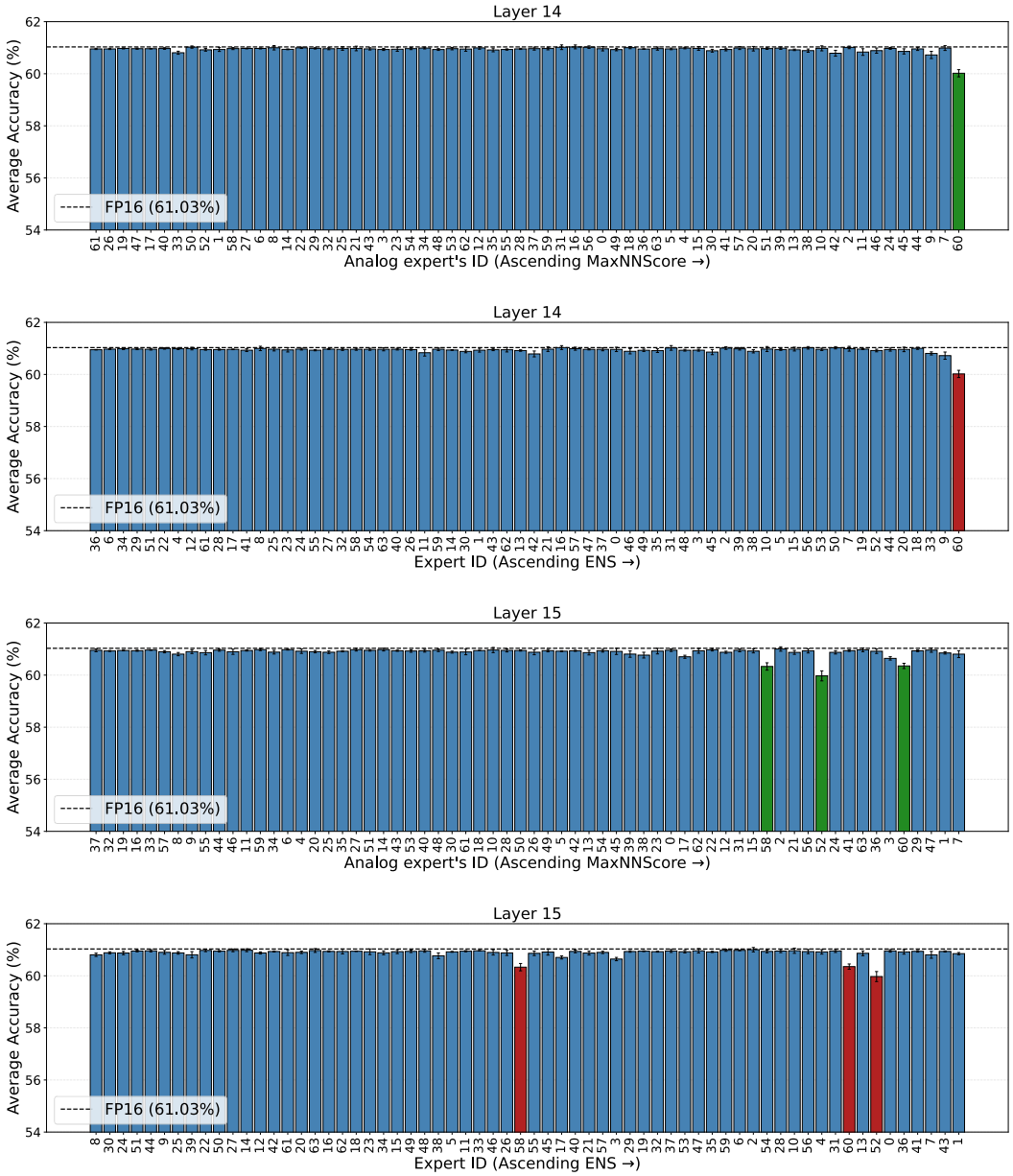


Figure 13: **Model accuracy-level analog noise sensitivity in OLMoE (Layers 14–15).**