Appendix A
Defining our metrics

In this Appendix, we provide a brief overview of the spindle metrics used in this paper: F1-score, recall, precision, RMS score and spindle density.

Characterizing the performance of any spindle detector can be done in two different ways:

1) comparing to a ground truth spindle detection for the same data and report metrics of detection performance. This requires a trustworthy ground truth which can be used for comparison to compute metrics like the F1-score;
2) showing evidence that the spindles detected are in fact spindles and that their distribution approximates the expected values for humans, with metrics such as spindle density and RMS score of detected spindles.

F1-score, recall and precision are commonly used metrics in classification tasks. These metrics are especially useful when the class distributions are imbalanced which leads to other common metrics like the accuracy being biased towards the most common class. We choose these metrics as they do not take into account the True Negatives in their computation as opposed to other metrics like specificity, which would be biased by the rarity of spindles during sleep and would not be a good indicator of performance. However, these metrics require comparison to some ground truth. When such a ground truth is not available, we opt to report RMS score in sigma power and spindle density.

A. F1-Score

F1-score is a metric that combines both precision and recall into a single value. It is particularly useful in scenarios where the classes are imbalanced. The formula for F1-score is given by:

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

where precision is the ratio of true positive predictions to the total number of positive predictions, and recall is the ratio of true positive predictions to the total number of actual positive instances.

B. Recall

Recall, also known as sensitivity or true positive rate, measures the ability of a classifier to correctly identify positive instances out of all actual positive instances. The formula for recall is given by:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

where True Positive (TP) represents the number of correctly identified positive instances, and False Negative (FN) represents the number of positive instances incorrectly classified as negative.

C. Precision

Precision measures the proportion of true positive predictions out of all positive predictions made by the classifier. The formula for precision is given by:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

where True Positive (TP) represents the number of correctly identified positive instances, and False Positive (FP) represents the number of negative instances incorrectly classified as positive.

Given that our objective is to detect and stimulate spindles for CLS, we use a by-event evaluation of performance. This means that an event is considered a True Positive if our model detects a spindle the ground truth spindles, as opposed to making sure that every single sample is correctly identified as a spindle or non-spindle.

D. RMS Score

The RMS (Root Mean Square) score is a metric we defined in the context of this study to assess the quality of candidate spindle detections. It quantifies the sigma activity at a specific time compared to a baseline period. It is calculated by dividing the RMS value of the signal filtered in the sigma band (11-16 Hz) at the time of detection (from 0 to 0,5s post detection) by the RMS value of the same filtered signal 2 second prior to the detection (from -2 to -1.5s pre detection). As spindles, when occurring in trains, are often distant by 4 seconds, measuring spindle activity 2s prior to the current detection ensures a neutral baseline value.

Let $x(t)$ denote the sigma signal at time t within the specific time-window of interest. The sigma power of the segment is then calculated as the Root Mean Square (RMS) of the filtered signal:

$$RMS = \sqrt{\frac{1}{T}\sum_{t=1}^{T} x(t)^2} \tag{3}$$

where T is the length of the segment.

Finally, let RMS_{post} represent the RMS value of the sigma signal in the 0.5 seconds following a spindle detection, and RMS_{pre} represent the RMS value in a 0.5 seconds-long time window occuring 2 s prior to the same detection. The RMS score (RMSscore) is calculated as follows:

$$\text{RMSscore} = \frac{RMS_{\text{post}}}{RMS_{\text{pre}}} \tag{4}$$

E. Spindle Density

Sleep spindle density is a critical metric utilized in sleep research due to its ability to capture the frequency of spindle occurrences within a specified period, providing valuable insights into the temporal distribution of spindle activity. Unlike metrics solely based on spindle presence or absence, spindle density offers a more comprehensive understanding of spindle dynamics by accounting for variations in spindle occurrence over time. Mathematically, spindle density (SD) is computed as the number of spindles (N_{spindles}) detected within a defined epoch duration (T_{epoch}), typically expressed per unit of time (e.g., per minute). Therefore, the spindle density (SD) is calculated as follows:

$$SD = \frac{N_{\text{spindles}}}{T_{\text{epoch}}} \tag{5}$$

where N_{spindles} represents the total number of spindles detected within the epoch duration T_{epoch}.

In addition to the RMS score, spindle density serves as a valuable metric for evaluating the efficacy of our spindle detection algorithm in terms of both the quality and quantity of detected spindles, removing the necessity for a ground truth reference for comparison.

Appendix B
Portiloop Hardware Specifications

TABLE IV
TECHNICAL SPECIFICATIONS (HTTPS://CORAL.AI/PRODUCTS/DEV-BOARD-MINI

Component	**Specification**
CPU	MediaTek 8167s SoC (Quad-core Arm Cortex-A35)
ML Accelerator	Google Edge TPU coprocessor: 4 TOPS (int8); 2 TOPS per watt
RAM	2 GB LPDDR3
Flash Memory	8 GB eMMC

Appendix C
Validation Results

This Appendix describes the process employed to select the optimal model for cross-validation, considering the 24-hour training time constraint. It is crucial to reiterate that the primary focus of this investigation lies in evaluating the effectiveness of adaptation methods, rather than achieving the absolute best possible final model performance. Consequently, the inherent quality of the final model holds less significance compared to the improvements observed through the application of each adaptation method.

Table V presents the comprehensive results for each fold, encompassing both the chosen evaluation metrics: sleep staging accuracy and sleep spindle detection F1-score. The epoch that maximizes the sum of these two metrics is selected as the optimal model for subsequent analysis.

TABLE V
VALIDATION RESULT FOR EACH FOLD

	Spindle F1-score	Sleep-staging Accuracy	Combined (sum)
Fold 1	0.5982	86.14	1.460
Fold 2	0.5051	93.09	1.436
Fold 3	0.4902	88.59	1.376
Fold 4	0.4263	87.78	1.304
Fold 5	0.4913	86.87	1.360
Average	**0.5022**	**88.49**	**1.387**

APPENDIX D
THRESHOLD DISTRIBUTION

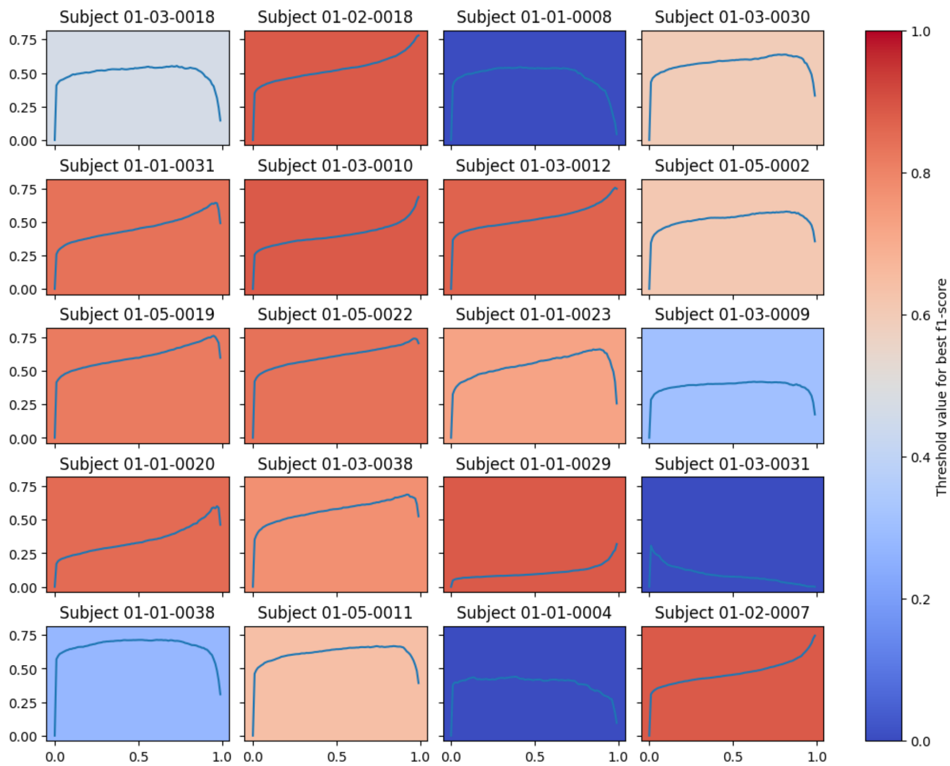

Fig. 6. Plot of the F1-score depending on threshold for the entire night for 20 random subjects. Although the model is trained to label either 0 or 1, the threshold of 0.5 is rarely the best threshold for any subject as was the case with the previous Portiloop model [21].

APPENDIX E
MODEL STRUCTURE

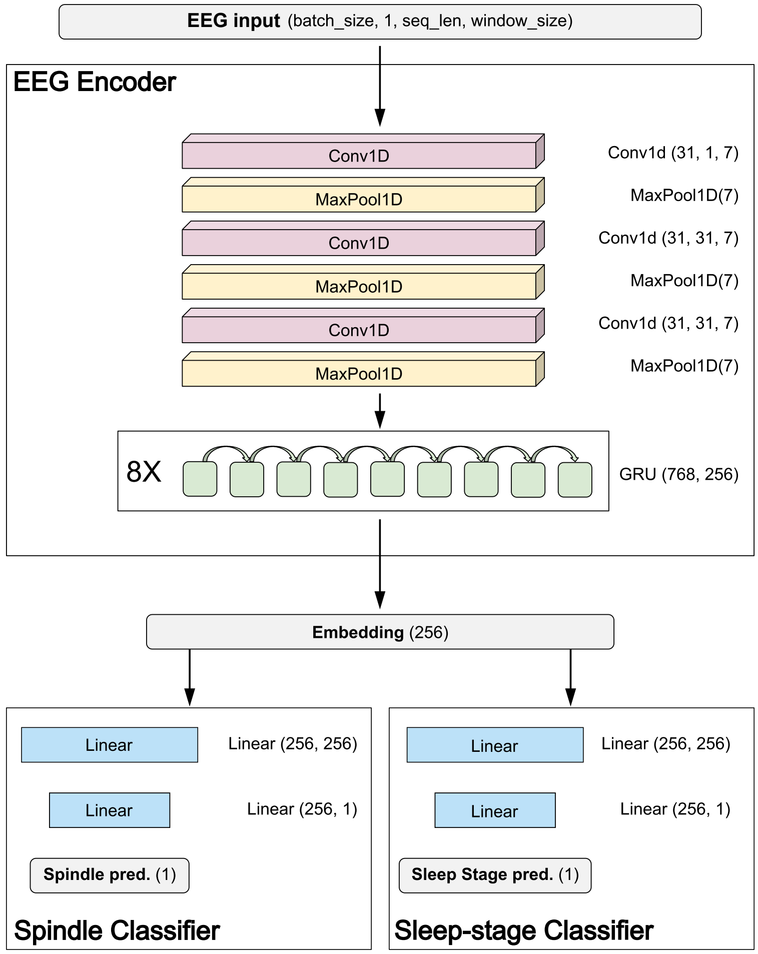

Fig. 7. Dual-Task Model Architecture

This Appendix provides a comprehensive explanation of the dual-task model architecture illustrated in Figure 7. The model is built using the PyTorch deep learning framework [56], and it incorporates various layers to process the input electroencephalogram (EEG) data to achieve the two objectives of sleep stage classification and online sleep spindle detection. Here is the detailed description of each layer used:

- Conv1D Layers (Conv1d): This sequence of convolutional layers with 1-dimensional kernels is responsible for extracting features from the raw EEG data. The number of filters and kernel sizes used in these layers (denoted as *Conv1D(inChannels, outChannels, kernelSize)*) are crucial for capturing relevant temporal and spectral features from the EEG signal.
- MaxPool1D Layers (MaxPool1d): These layers perform downsampling along the temporal dimension of the data, reducing its dimensionality while preserving important features. The kernel size controls the amount of downsampling applied (denoted as *MaxPool1d(kernelSize)*).
- GRU Layer (GRU): This Gated Recurrent Unit (GRU) layer is a type of recurrent neural network (RNN) that effectively captures temporal dependencies within the EEG data. The number of units in the GRU layer (denoted as *GRU(inputsize, hiddenSize)* determines its capacity to learn complex temporal relationships.
- Linear Layers (Linear): A sequence of fully-connected linear layers performs further feature extraction and transformation on the combined representation. The number of units in each linear layer (denoted as *Linear(inFeatures, outFeatures)* for input and output dimensions respectively) determines its complexity and capacity to learn higher-level features.

The classification models generate a single floating-point value as output. To ensure these outputs range between 0 and 1, a sigmoid activation function is applied as the final layer. This transformation allows for a probabilistic interpretation of the model's predictions. A predefined threshold is then employed to convert the continuous output into a binary classification (positive or negative).

APPENDIX F
REPEATED MEASURES ANOVA - SLEEP STAGING CONFIGURATION * AGE

Here, we show the full results of the ANOVA test performed between the results of SLA7 compared to LA7 depending on each sleep staging configuration to determine the significance of sleep staging in the computation of our online ground truth spindles.

TABLE VI
WITHIN SUBJECTS EFFECTS

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	η^2
Sleep Staging	Greenhouse-Geisser	2.877	1.862	1.545	111.437	$< .001$	0.247
Sleep Staging * Age Cat	Greenhouse-Geisser	0.795	1.862	0.427	30.807	$< .001$	0.068
Residuals	Greenhouse-Geisser	3.434	247.675	0.014			

Note: the assumption of sphericity is violated so we use the Greenhouse-Geisser sphericity correction.

TABLE VII
BETWEEN SUBJECTS EFFECTS

Cases	Sum of Squares	df	Mean Square	F	p	η^2
Age Cat	0.885	1	0.885	32.110	$< .001$	0.076
Residuals	3.666	133	0.028			

TABLE VIII
POST HOC COMPARISONS - AGE CAT * SLEEP STAGING

		Mean Difference	SE	t	p_{bonf}	p_{holm}
Older, GroundTruth	Younger, GroundTruth	0.016	0.023	0.681	1.000	0.779
	Older, None	0.213	0.019	10.926	$< .001$	$< .001$
	Younger, None	0.118	0.023	5.149	$< .001$	$< .001$
	Older, Online	0.303	0.019	15.539	$< .001$	$< .001$
	Younger, Online	0.101	0.023	4.412	$< .001$	$< .001$
Younger, GroundTruth	Older, None	0.197	0.023	8.591	$< .001$	$< .001$
	Younger, None	0.103	0.020	5.226	$< .001$	$< .001$
	Older, Online	0.287	0.023	12.505	$< .001$	$< .001$
	Younger, Online	0.086	0.020	4.365	$< .001$	$< .001$
Older, None	Younger, None	-0.095	0.023	-4.123	$< .001$	$< .001$
	Older, Online	0.090	0.019	4.613	$< .001$	$< .001$
	Younger, Online	-0.112	0.023	-4.859	$< .001$	$< .001$
Younger, None	Older, Online	0.185	0.023	8.037	$< .001$	$< .001$
	Younger, Online	-0.017	0.020	-0.862	1.000	0.779
Older, Online	Younger, Online	-0.201	0.023	-8.774	$< .001$	$< .001$

APPENDIX G
REPEATED MEASURES ANOVA - ADAPTATION CONFIGURATIONS * AGE FOR SINGLE NIGHT EXPERIMENTS

TABLE IX
WITHIN SUBJECTS EFFECTS

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	η^2
Config	Greenhouse-Geisser	0.148	2.002	0.074	37.581	$< .001$	0.027
Config * Age Cat	Greenhouse-Geisser	0.025	2.002	0.013	6.375	0.002	0.004
Residuals	Greenhouse-Geisser	0.505	256.312	0.002			

Note: the assumption of sphericity is violated so we use the Greenhouse-Geisser sphericity correction.

TABLE X
BETWEEN SUBJECTS EFFECTS

Cases	Sum of Squares	df	Mean Square	F	p	η^2
Age Cat	0.136	1	0.136	3.654	0.058	0.024
Residuals	4.773	128	0.037			

TABLE XI
POST HOC COMPARISONS - AGE CAT * CONFIG

		Mean Difference	SE	t	p_{bonf}
Older, Baseline	Younger, Baseline	−0.017	0.018	−0.978	1.000
	Older, Threshold	−0.022	0.006	−3.440	0.018
	Younger, Threshold	−0.060	0.018	−3.384	0.025
	Older, Fine-tuning	0.011	0.006	1.746	1.000
	Younger, Fine-tuning	−0.010	0.018	−0.582	1.000
	Older, Combined	−0.007	0.006	−1.042	1.000
	Younger, Combined	−0.059	0.018	−3.297	0.034
Younger, Baseline	Older, Threshold	−0.004	0.018	−0.241	1.000
	Younger, Threshold	−0.043	0.006	−6.685	$< .001$
	Older, Fine-tuning	0.028	0.018	1.597	1.000
	Younger, Fine-tuning	0.007	0.006	1.102	1.000
	Older, Combined	0.011	0.018	0.609	1.000
	Younger, Combined	−0.041	0.006	−6.445	$< .001$
Older, Threshold	Younger, Threshold	−0.039	0.018	−2.165	0.894
	Older, Fine-tuning	0.033	0.006	5.186	$< .001$
	Younger, Fine-tuning	0.011	0.018	0.637	1.000
	Older, Combined	0.015	0.006	2.398	0.475
	Younger, Combined	−0.037	0.018	−2.078	1.000
Younger, Threshold	Older, Fine-tuning	0.071	0.018	4.002	0.003
	Younger, Fine-tuning	0.050	0.006	7.787	$< .001$
	Older, Combined	0.054	0.018	3.014	0.084
	Younger, Combined	0.002	0.006	0.240	1.000
Older, Fine-tuning	Younger, Fine-tuning	−0.021	0.018	−1.201	1.000
	Older, Combined	−0.018	0.006	−2.788	0.156
	Younger, Combined	−0.070	0.018	−3.916	0.004
Younger, Fine-tuning	Older, Combined	0.004	0.018	0.213	1.000
	Younger, Combined	−0.048	0.006	−7.546	$< .001$
Older, Combined	Younger, Combined	−0.052	0.018	−2.928	0.110

Appendix H
Repeated Measures ANOVA - Spindle Density of adaptation configuration * Experiment Type

This Appendix presents the detailed results of our ANOVA analysis comparing the spindle density of various adaptation configurations (Baseline, Threshold, WeightAveraging, and Train) with the two experiment types (Random and SameSubject).

TABLE XII
Within Subjects Effects

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	η^2
Config	Greenhouse-Geisser	1673.858	1.887	887.022	10.634	$< .001$	0.042
Config * experiment_type	Greenhouse-Geisser	226.324	1.887	119.935	1.438	0.240	0.006
Residuals	Greenhouse-Geisser	18101.349	217.011	83.412			

Note: the assumption of sphericity is violated so we use the Greenhouse-Geisser sphericity correction.

TABLE XIII
Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η^2
experiment_type	9.216	1	9.216	0.054	0.817	2.324×10^{-4}
Residuals	19646.473	115	170.839			

TABLE XIV
Post Hoc Comparisons - experiment_type * Config

		Mean Difference	SE	t	p_{holm}
Random, Baseline	SameSubject, Baseline	0.520	2.760	0.188	1.000
	Random, Threshold	6.858	1.000	6.860	$< .001$
	SameSubject, Threshold	5.310	2.760	1.924	0.878
	Random, WeightAveraging	10.629	1.000	10.632	$< .001$
	SameSubject, WeightAveraging	7.146	2.760	2.589	0.211
	Random, Train	4.148	1.000	4.149	0.001
	SameSubject, Train	6.808	2.760	2.466	0.283
SameSubject, Baseline	Random, Threshold	6.338	2.760	2.296	0.424
	SameSubject, Threshold	4.790	2.957	1.620	1.000
	Random, WeightAveraging	10.109	2.760	3.662	0.007
	SameSubject, WeightAveraging	6.626	2.957	2.241	0.462
	Random, Train	3.628	2.760	1.314	1.000
	SameSubject, Train	6.288	2.957	2.126	0.581
Random, Threshold	SameSubject, Threshold	-1.548	2.760	-0.561	1.000
	Random, WeightAveraging	3.771	1.000	3.772	0.005
	SameSubject, WeightAveraging	0.288	2.760	0.104	1.000
	Random, Train	-2.710	1.000	-2.711	0.155
	SameSubject, Train	-0.050	2.760	-0.018	1.000
SameSubject, Threshold	Random, WeightAveraging	5.318	2.760	1.927	0.878
	SameSubject, WeightAveraging	1.836	2.957	0.621	1.000
	Random, Train	-1.162	2.760	-0.421	1.000
	SameSubject, Train	1.498	2.957	0.507	1.000
Random, WeightAveraging	SameSubject, WeightAveraging	-3.482	2.760	-1.262	1.000
	Random, Train	-6.481	1.000	-6.483	$< .001$
	SameSubject, Train	-3.821	2.760	-1.384	1.000
SameSubject, WeightAveraging	Random, Train	-2.998	2.760	-1.086	1.000
	SameSubject, Train	-0.338	2.957	-0.114	1.000
Random, Train	SameSubject, Train	2.660	2.760	0.964	1.000

Appendix I
Repeated Measures ANOVA - Spindle Density of adaptation configuration * Night Number

TABLE XV
Within Subjects Effects

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	η^2
Config	Greenhouse-Geisser	10828.127	2.168	4994.211	34.993	$< .001$	0.134
Config * night_num	Greenhouse-Geisser	916.044	10.841	84.501	0.592	0.832	0.011
Residuals	Greenhouse-Geisser	34347.400	240.663	142.720			

Note: the assumption of sphericity is violated so we use the Greenhouse-Geisser sphericity correction.

TABLE XVI
Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η^2
night_num	1751.581	5	350.316	1.187	0.320	0.022
Residuals	32767.598	111	295.204			

TABLE XVII
Post Hoc Comparisons - Config

		Mean Difference	SE	t	p_{holm}
Baseline	Threshold	6.660	1.030	6.465	$< .001$
	Fine-tuning	4.417	1.030	4.288	$< .001$
	WeightAveraging	10.231	1.030	9.931	$< .001$
	Combined	11.981	1.030	11.629	$< .001$
	ClassifierOnly	8.354	1.030	8.109	$< .001$
Threshold	Fine-tuning	-2.243	1.030	-2.177	0.119
	WeightAveraging	3.571	1.030	3.466	0.003
	Combined	5.320	1.030	5.164	$< .001$
	ClassifierOnly	1.694	1.030	1.644	0.207
Fine-tuning	WeightAveraging	5.814	1.030	5.644	$< .001$
	Combined	7.563	1.030	7.342	$< .001$
	ClassifierOnly	3.937	1.030	3.821	0.001
WeightAveraging	Combined	1.749	1.030	1.698	0.207
	ClassifierOnly	-1.877	1.030	-1.822	0.207
Combined	ClassifierOnly	-3.627	1.030	-3.520	0.003

APPENDIX J
ANOVA - RMS SCORE COMPARED TO NIGHT NUMBER AND CONFIGURATION

TABLE XVIII
ANOVA - RMS_SCORE

Cases	Sum of Squares	df	Mean Square	F	p	η^2
night_num	14659.143	5	2931.829	970.807	$< .001$	0.003
config	48077.866	5	9615.573	3183.975	$< .001$	0.009
night_num * config	7133.757	25	285.350	94.487	$< .001$	0.001
Residuals	$5.161 \times 10^{+6}$	1708899	3.020			

TABLE XIX
POST HOC COMPARISONS - CONFIG

		Mean Difference	SE	t	p_{tukey}
WeightAveraging	ClassifierOnly	0.193	0.005	37.106	$< .001$
	Combined	-0.013	0.006	-2.287	0.199
	Fine-tuning	0.371	0.005	76.972	$< .001$
	Baseline	0.447	0.005	97.462	$< .001$
	Threshold	0.238	0.005	47.803	$< .001$
ClassifierOnly	Combined	-0.205	0.005	-37.559	$< .001$
	Fine-tuning	0.178	0.005	37.883	$< .001$
	Baseline	0.254	0.004	56.979	$< .001$
	Threshold	0.045	0.005	9.231	$< .001$
Combined	Fine-tuning	0.383	0.005	74.991	$< .001$
	Baseline	0.460	0.005	93.870	$< .001$
	Threshold	0.250	0.005	47.604	$< .001$
Fine-tuning	Baseline	0.076	0.004	18.992	$< .001$
	Threshold	-0.133	0.004	-29.936	$< .001$
Baseline	Threshold	-0.209	0.004	-49.887	$< .001$