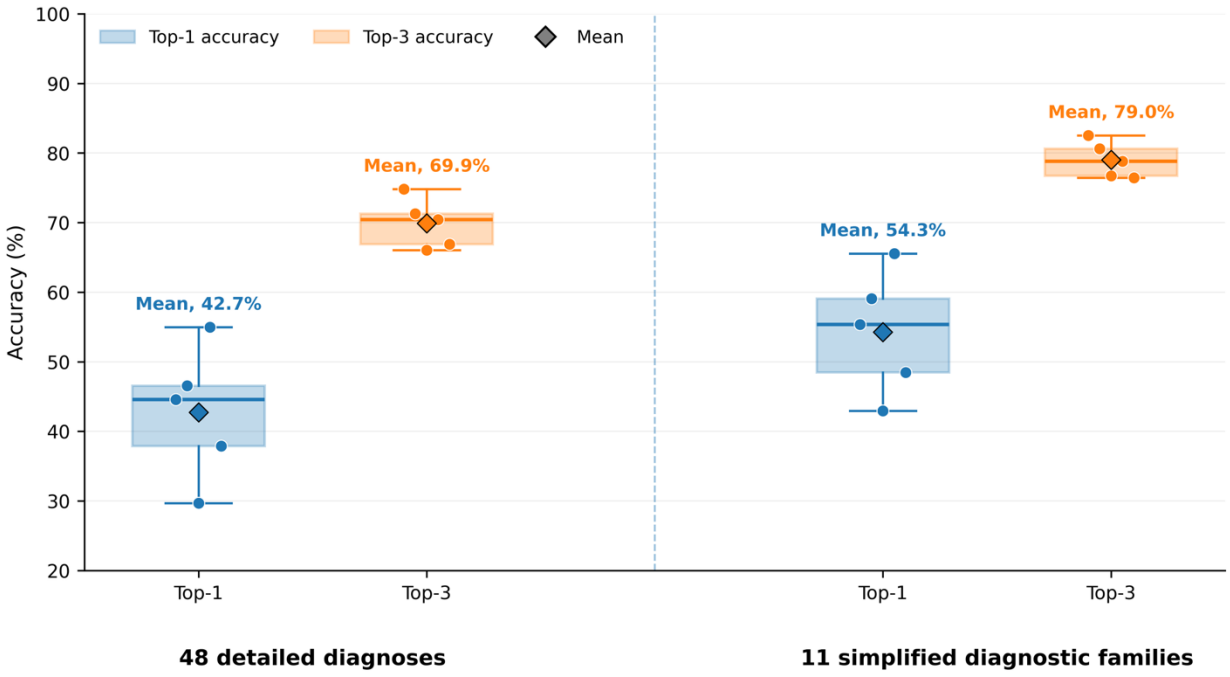
Supplementary Methods and Validation of Component CNN Models

The GPT Fusion framework integrates structured prediction outputs from two independently developed convolutional neural network (CNN) models: a multimodal ResNet-50 trained on the MILK10K dataset [1] and the SIIM-90 CNN ensemble originally developed for the SIIM-ISIC Melanoma Classification Challenge [2]. Because pretrained weights for the MILK10K ResNet-50 model were not publicly available, the model was reproduced using the authors' published code, dataset, and training protocol. The reproduced implementation achieved performance comparable to the originally reported results (eFigure 1 and eTable 1), supporting the faithful reproduction of the published model and its use as the ResNet-50 component of the GPT Fusion framework. The ResNet-50 model was evaluated on the original five MILK10K validation folds to verify faithful reproduction of the published model, whereas binary melanoma detection was compared between ResNet-50 and the SIIM-90 CNN ensemble using a 460-lesion evaluation cohort employed in our previous studies [3].

1. Validation of the reproduced ResNet-50 model

Since the pretrained ResNet-50 model weights were not publicly released by the MILK10K developers, we reproduced the model using the authors' original code, dataset, and training protocol without modification. The validation results shown in eFigure 1 and summarized in eTable 1 demonstrate that the reproduced implementation achieved consistent performance across the five original validation folds. The reproduced implementation achieved a mean Top-3 differential diagnosis accuracy of **69.9%** for the 48 detailed diagnosis categories, closely matching the **68.0%** Top-3 differential diagnosis accuracy reported by the MILK10K developers [1], thereby supporting faithful reproduction of the original ResNet-50 model.



eFigure 1. ResNet-50 classification performance on the original five MILK10K validation folds. Top-1 primary diagnosis and Top-3 differential diagnosis accuracies are shown for predicting the 48 detailed diagnostic categories and the corresponding 11 simplified diagnostic families. Each boxplot summarizes the performance of the best validation model from one of the five cross-validation folds. The center line indicates the median; the box, interquartile range (IQR); whiskers, minimum and maximum values; circles, individual validation folds; and diamonds, means.

eTable 1. ResNet-50 classification performance on the original five MILK10K validation folds

a. Overall multiclass performance of each fold (48 detailed diagnoses)

Fold	Top-1 Accuracy	Top-3 Accuracy	Macro Precision	Macro Recall (Balanced Accuracy)	Macro F1
0	44.6%	74.8%	27.9%	28.6%	19.2%
1	46.6%	71.3%	35.8%	36.2%	31.4%
2	29.7%	66.0%	24.0%	22.4%	18.3%
3	55.0%	70.4%	39.7%	38.8%	36.4%
4	37.9%	66.9%	26.5%	25.7%	22.6%
Mean	42.7%	69.9%	30.8%	30.3%	25.6%

b. Overall multiclass performance of each fold (11 simplified diagnostic families)

Fold	Top-1 Accuracy	Top-3 Accuracy	Macro Precision	Macro Recall (Balanced Accuracy)	Macro F1
0	55.3%	76.5%	34.6%	43.9%	36.5%
1	59.1%	80.7%	39.4%	48.0%	41.6%
2	42.9%	76.4%	32.2%	44.1%	32.9%
3	65.6%	82.5%	38.9%	41.7%	39.1%
4	48.5%	78.7%	29.4%	39.1%	30.4%
Mean	54.3%	79.0%	34.9%	43.4%	36.1%

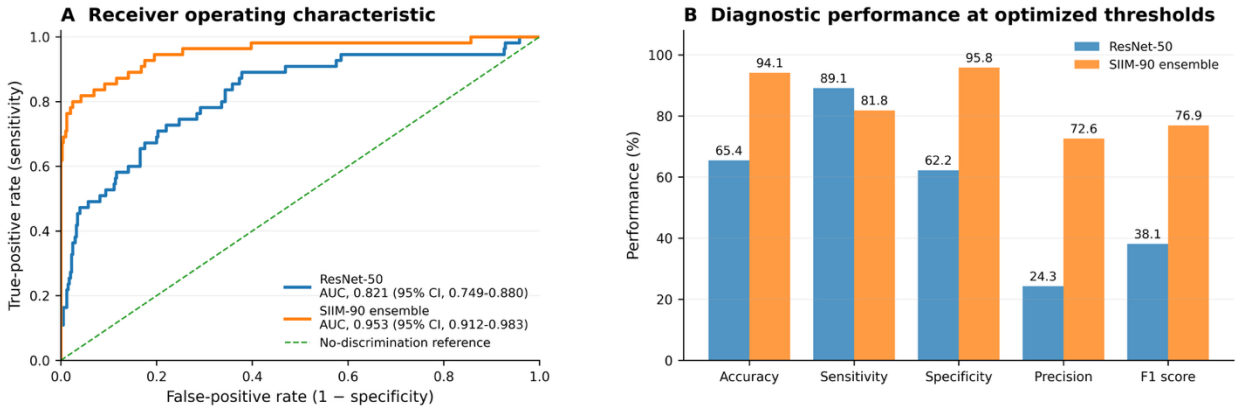
c. Summary

Representation	Mean Top-1 Accuracy	Mean Top-3 Accuracy
48 detailed diagnoses	42.7%	69.9%
11 simplified diagnostic families	54.3%	79.0%

As shown in eFigure 1 and summarized in eTable 1, Top-3 differential diagnosis accuracy was substantially higher than Top-1 accuracy for both the 48 detailed diagnoses and the 11 simplified diagnostic families, indicating that the model more frequently included the correct diagnosis among its three highest-ranked predictions. As expected, grouping the 48 diagnoses into 11 broader diagnostic families further improved classification accuracy.

2. Validation of binary melanoma detection models

The SIIM-90 CNN ensemble, which ranked first in the SIIM-ISIC Melanoma Classification Challenge, was designed specifically for binary melanoma detection. To validate both the reproduced ResNet-50 model and the downloaded SIIM-90 CNN ensemble before integrating their outputs into the GPT Fusion framework, both models were evaluated using the same 460-lesion MILK10K evaluation cohort employed in our previous studies [3]. This independent cohort was originally constructed to evaluate large language models under identical conditions and contains paired dermoscopic and clinical close-up images, with equal representation from each of the five skin-tone groups (92 lesions per group). Using the same evaluation cohort enabled a direct comparison among the reproduced ResNet-50 model, the SIIM-90 CNN ensemble, and our previous benchmark. As shown in eFigure 2 and summarized in eTable 2, the SIIM-90 CNN ensemble substantially outperformed the reproduced ResNet-50 model for binary melanoma detection, achieving an ROC AUC of 0.953 compared with 0.821. These results are consistent with the SIIM-90 ensemble's top performance in the SIIM-ISIC challenge and support its use as the melanoma prediction component of the GPT Fusion framework.



eFigure 2. Binary melanoma detection performance of ResNet-50 and the SIIM-90 CNN ensemble on the 460-lesion MILK10K evaluation cohort. The evaluation cohort included 55 melanomas and 405 nonmelanomas (melanoma prevalence, 12.0%). **A**, Receiver operating characteristic (ROC) curves comparing ResNet-50 and the SIIM-90 ensemble. Areas under the ROC curves (AUCs) and 95% confidence intervals were estimated using bootstrap resampling. **B**, Diagnostic performance at the optimized operating threshold selected independently for each model using the Youden index (0.065 for ResNet-50 and 0.203 for the SIIM-90 ensemble).

eTable 2. Binary melanoma detection performance of ResNet-50 and the SIIM-90 CNN ensemble on the 460-lesion MILK10K evaluation cohort

Metric	ResNet-50	SIIM-90 CNN ensemble
ROC AUC	0.821	0.953
95% CI	0.749–0.880	0.912–0.983
Threshold	0.065	0.203
Accuracy	65.4%	94.1%
Sensitivity	89.1%	81.8%
Specificity	62.2%	95.8%
Precision	24.3%	72.6%
F1 score	38.1%	76.9%

Using the same 460-lesion evaluation cohort also enabled direct comparison with our previous GPT-5.2 benchmark study [3]. The reproduced ResNet-50 model provided the greatest sensitivity, GPT-5.2 demonstrated the most balanced overall performance (accuracy, 60.9%; F1 score, 59.8%), and the SIIM-90 CNN ensemble achieved the strongest overall binary melanoma detection performance, with superior accuracy, specificity, precision, F1 score, and ROC AUC. These complementary strengths motivated the integration of CNN predictions with LLM reasoning in the proposed GPT Fusion framework.

REFERENCES

1. Tschandl P, Akay BN, Rosendahl C, Rotemberg V, Todorovska V, Weber J, et al. MILK10k: A Hierarchical Multimodal Imaging-Learning Toolkit for Diagnosing Pigmented and Nonpigmented Skin Cancer and its Simulators. Journal of Investigative Dermatology. 2025.
2. Qishen Ha BL, Fuxu Liu. Identifying Melanoma Images using EfficientNet Ensemble: Winning Solution to the SIIM-ISIC Melanoma Classification Challenge. arXiv. 2020.

3. Frederickson KL, Adunyah SE, Wang Q. Evaluation of GPT-5.2 for melanoma detection across skin tones. Frontiers in Medicine. 2026;13.