Aggregate and Malware-Family Calibration in the Released EMBER2024 Detector

Tara Dixit
Stanford University
Stanford, CA
tdixit@stanford.edu

Abstract

The released EMBER2024 PE detector has strong overall scores, but aggregate metrics may hide errors for specific malware families. I evaluated the released LightGBM model on 540,000 PE records represented by 2,568 `threm ber` features. The data contain 360,000 Win32, 120,000 Win64, and 60,000 Dot_Net records. The released archives contained duplicated structural halves, so the evaluation kept the first documented half of each weekly member. The selected data have 539,940 unique hashes. Sixty hashes occur twice across member and week boundaries, with no label, family, file-type, or detector-input conflicts. The model reaches 0.9803222222 accuracy, 0.9981880554 ROC AUC, a 0.0147949965 Brier score, 0.0030850838 expected calibration error (ECE), and 0.0277246070 maximum calibration error (MCE) with 15 bins. These aggregate values are strong. Family results are different. Among families with at least 100 malicious records, `malicord` has a false-negative rate of 0.851852 and ECE of 0.602358. `lazzzy` and `rugmi` also have false-negative rates above 0.61. The family results are descriptive, not causal. They show that good aggregate calibration can coexist with large errors in smaller groups.

1 Introduction

EMBER2024 is a malware benchmark with data from several file types and time periods [1]. Its released PE detector is a LightGBM model [2]. This study asks a narrow question:

> *Can strong aggregate calibration hide high-confidence false negatives for specific malware families in the released EMBER2024 PE detector?*

Calibration describes whether a model's stated confidence matches its observed accuracy. A model can be well calibrated on average while giving poor results for a subgroup. This matters in malware detection because families differ in size, behavior, and representation in the data.

I evaluate the released model without retraining it. I report aggregate accuracy, ROC AUC, Brier score, ECE, and MCE. I then measure false-negative rate and ECE within malicious families. The main result is simple: the aggregate scores are strong, but several families have high false-negative rates and large calibration gaps.

2 Data

2.1 Released PE test records

The evaluation uses the PE part of EMBER2024. Table 1 shows the documented test counts. The 12 test weeks contain equal numbers of malicious and benign records overall.

Table 1: Evaluated PE records.

File type	Records per week	Total records
Win32	30,000	360,000
Win64	10,000	120,000
Dot_Net	5,000	60,000
Total	45,000	540,000

The three compressed source archives total 3,990,366,412 bytes. Controlled extraction produced 36 JSONL members totaling 17,655,416,292 bytes. Each weekly member contained two ordered structural halves. Corresponding rows matched on SHA-256, label, family, file type, week, and all 12 inputs used by the PE feature extractor. Only the `caps`, `mbc`, and `ttps` fields differed. Those fields are not detector inputs.

The selection keeps the first documented half of each member: 30,000 Win32, 10,000 Win64, or 5,000 Dot_Net rows. This is a fixed structural rule. It does not claim that the first half is better. The result has 539,940 unique hashes. Of these, 539,880 occur once and 60 occur twice across member and week boundaries. The repeated rows have no label, family, file-type, or detector-input conflicts.

2.2 Prepared inputs

The official `thrember` PE extractor produces 2,568 features per record. The released LightGBM model also reports 2,568 input features. Preparation preserved selection order and produced a `float32` feature matrix, an `int32` label vector, and aligned metadata. Inference produced one finite `float64` malicious-class probability per row.

3 Method

3.1 Aggregate measures

Predictions at or above 0.5 are malicious. Accuracy uses that threshold. ROC AUC measures ranking over both classes. The Brier score is the mean squared difference between a probability and its binary label [3].

ECE and MCE follow common fixed-bin summaries [4]. For a probability p, predicted-label confidence is the larger of p and $1 - p$. Confidence is divided into equal-width bins over $[0.5, 1.0]$. A bin gap is the absolute difference between its mean confidence and accuracy. ECE is the count-weighted mean gap. MCE is the largest nonempty-bin gap. The primary result uses 15 bins. Empty bins have zero count, add no weight to ECE, and are omitted from MCE.

3.2 Family measures

Family analysis uses malicious records only. A prediction below 0.5 is a false negative, so family false-negative rate equals one minus family accuracy. The primary table includes families with at least 100 malicious records. It excludes missing, blank, `unknown`, `none`, and `nan` family values without changing the remaining labels. Family ECE uses the same 15-bin predicted-label-confidence definition as the aggregate result.

The family estimates describe this dataset and model. They do not show that a family name caused an error. The minimum count limits the noisiest small groups, but it does not make every estimate stable.

4 Results

4.1 Aggregate results

Table 2 reports the verified results. The model is accurate and ranks malicious records well. The Brier score, ECE, and MCE are also low.

Figure 1 compares confidence and observed accuracy. Most of the mass is in the highest-confidence bins. The plotted points remain near the diagonal at the aggregate level.

Table 2: Aggregate results on 540,000 PE records.

Metric	Value
Accuracy	0.9803222222
ROC AUC	0.9981880554
Brier score	0.0147949965
15-bin ECE	0.0030850838
15-bin MCE	0.0277246070

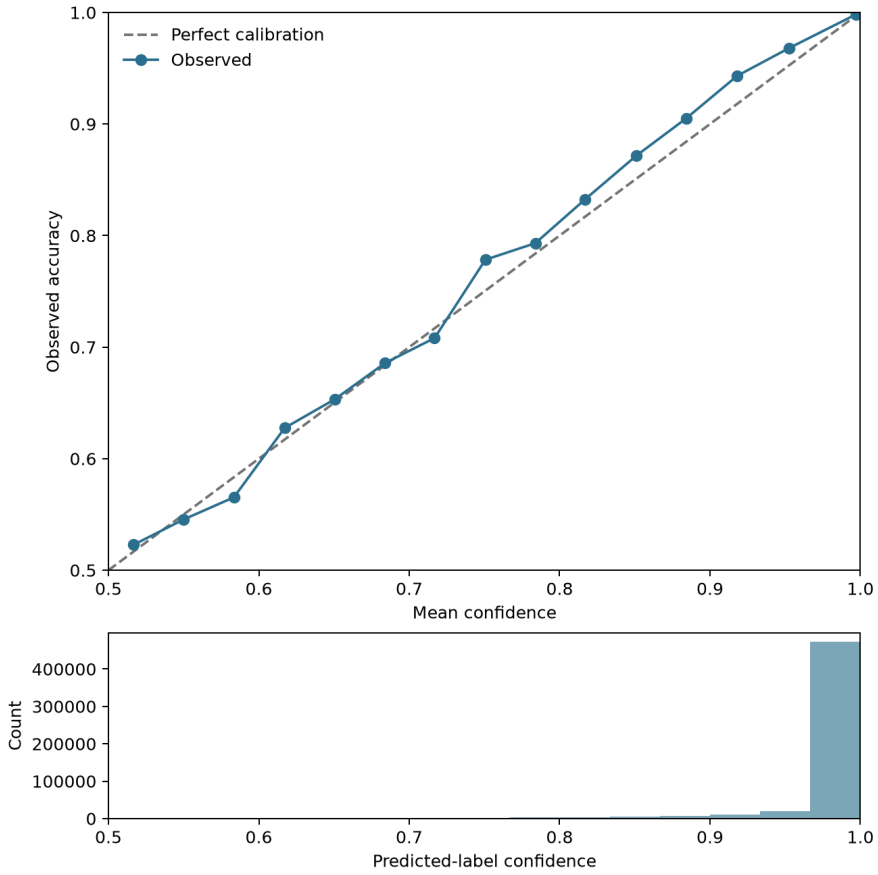

Figure 1: Aggregate reliability diagram with 15 equal-width confidence bins.

4.2 Malware-family results

The family results give a different view. Table 3 lists six families with large false-negative rates. `malicord` has 138 false negatives among 162 malicious records. Its false-negative rate is 0.851852 and its ECE is 0.602358. `lazzzy` and `rugmi` also have false-negative rates above 0.61.

Table 3: Selected family results at the 100-record minimum.

Family	Malicious records	False-negative rate	15-bin ECE
`malicord`	162	0.851852	0.602358
`lazzzy`	179	0.620112	0.452050
`rugmi`	256	0.617188	0.447200
`goblin`	165	0.406061	0.230170
`penguish`	131	0.404580	0.278403
`softcnapp`	105	0.333333	0.237920

Of 270,000 malicious rows, 37,633 have no usable family label. The remaining 232,367 rows contain 3,786 family labels. At the primary minimum, 232 families and 198,854 malicious rows are eligible. Another 3,554 families, containing 33,513 rows, are below the minimum.

Figure 2 shows the leading family false-negative rates and ECE values. The two rankings overlap, but they are not the same metric.

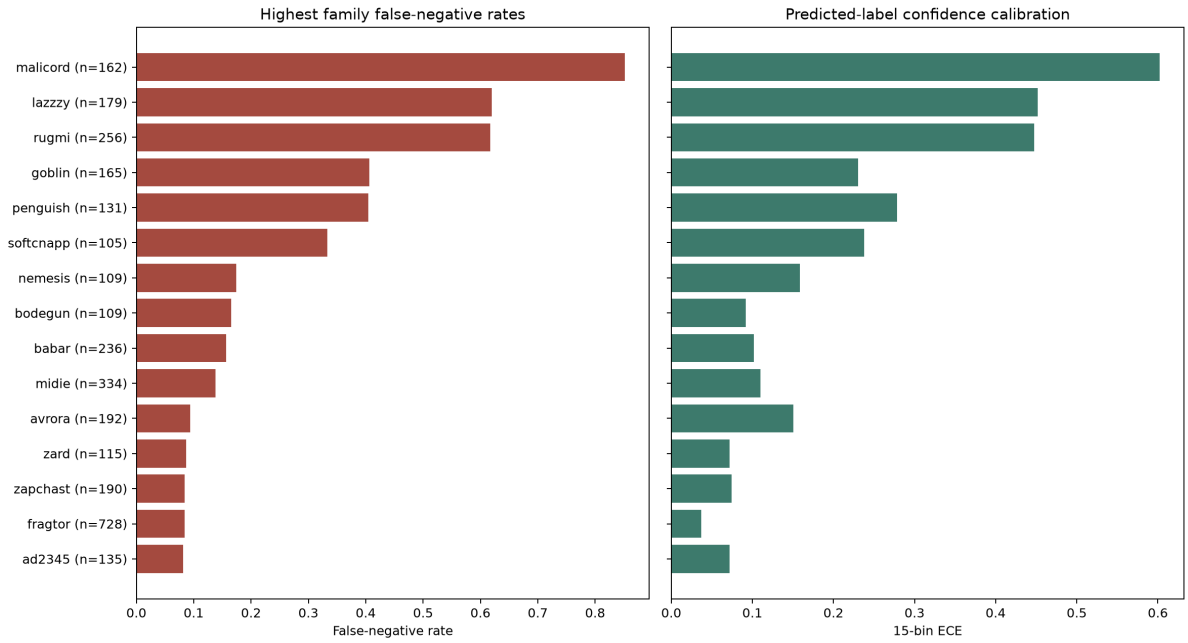

Figure 2: Largest family false-negative rates and calibration errors.

5 Sensitivity Checks

5.1 Calibration bins

Table 4 varies the number of aggregate confidence bins. ECE stays between 0.00297 and 0.00323. MCE rises as narrower bins expose larger local gaps. The same ten families appear in the top-ten family-ECE set for every bin count.

Table 4: Sensitivity to the number of confidence bins.

Bins	Aggregate ECE	Aggregate MCE	Family-ECE overlap
10	0.002967793	0.023373977	10/10
15	0.003085084	0.027724607	10/10
20	0.003177991	0.029541059	10/10
30	0.003227550	0.035004511	10/10

5.2 Family minimum

Changing the minimum family size changes both eligibility and the leading families. Table 5 compares each setting with the primary 100-record top ten. The 50-record rule admits 394 families. Only six of its top ten false-negative-rate families match the primary list. The 200-record rule admits 128 families, with only three matches. Small-family rankings should not be read as precise population estimates.

Table 5: Sensitivity to the minimum malicious family count.

Minimum	Families	Rows	FNR overlap	ECE overlap
50	394	210,067	6/10	6/10
100	232	198,854	10/10	10/10
200	128	184,149	3/10	3/10

6 Limits

This study has several limits:

- Family results are descriptive and do not establish a causal effect.
- Many malicious records have no family label. Family labels also come from the upstream data and may contain noise.
- Estimates for small families are unstable. The study reports no confidence intervals.
- Sixty hashes repeat across weeks. They have no observed conflicts, but a unique-hash sensitivity analysis was not run.
- The first structural half was selected by a fixed rule. The study does not claim that it is better than the second half.
- The evaluation covers PE records and one released model. It does not measure other EMBER2024 file types or models.

A separate SHAP pipeline used EMBER2018 malicious samples and a separately trained Random Forest. It therefore did not explain the evaluated EMBER2024 model. SHAP is a method for assigning feature contributions to predictions [5]. Model-aligned SHAP, retraining, and post-hoc calibration are not part of this study.

7 Reproducibility

The workflow pins the official source, dataset, and model repository. The exact identifiers are:

Source revision: `0ef753e81d98bf209f71b03cd331dfc190b5b54d`
Dataset revision: `3d23efef7c0f0b702c5024400cfff4c3744a3832`
Model revision: `e5b945dd90e1a1a1ec0cc07b3a17b52e9ba2d0c2`
Model SHA-256: `4252027863492ac138785c8c18576f43dad77d00faddc14e8c0072e8db419f99`

Downloaded files were checked before extraction. The selection manifest records every selected file and the repeated-hash profile. The preparation manifest records the row count, order, data types, feature count, sizes, and checksums. The inference manifest links the released model to the prepared inputs and predictions. The analysis manifest links those inputs to the small tracked tables and figures. The analysis code commit is `c63e5548d6c1eab93e57b99e292d1af68fa31318`.

The environment record describes a macOS arm64 Python 3.12 setup with LightGBM 4.7.0, `thrember` 0.1.0, and the Homebrew OpenMP runtime. The requirements snapshot is host-specific, not a universal cross-platform lock file. Synthetic tests cover archive checks, selection, preparation, inference, and metric definitions without requiring the large artifacts.

8 Conclusion

The released EMBER2024 PE model performs well on the full 540,000-record evaluation. Its accuracy is 98.03%, ROC AUC is 0.9982, and 15-bin ECE is 0.0031. These values do not describe every malware family. Several eligible families have false-negative rates above 0.40, and three are above 0.61.

The evidence supports a limited conclusion. Strong aggregate calibration can hide large family-level errors. These results identify groups for closer study, but they do not explain why the errors occur.

References

[1] Robert J. Joyce, Gideon Miller, Phil Roth, Richard Zak, Elliott Zaresky-Williams, Hyrum Anderson, Edward Raff, and James Holt. EMBER2024 - A Benchmark Dataset for Holistic Evaluation of Malware Classifiers. Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2025. https://arxiv.org/abs/2506.05074.

[2] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30, 2017. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

[3] Glenn W. Brier. Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review, 78(1):1–3, 1950. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. Proceedings of the 34th International Conference on Machine Learning, 70:1321–1330, 2017. https://proceedings.mlr.press/v70/guo17a.html.

[5] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30, 2017. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.