

From Decision Boundaries to State Discrimination: A Quantum Reframing of Malware Classification

www.SpecterAI.ai

January 25, 2026

Abstract

Classical malware classification systems are fundamentally framed as geometric optimization problems: executable samples are mapped into high-dimensional feature spaces, and decision boundaries are learned to separate benign from malicious behavior. While this paradigm has produced practical results, it exhibits well-known failure modes in adversarial and high-dimensional regimes, including boundary fragility, sensitivity to perturbation, and performance plateaus despite increased model complexity. In this work, we explore an alternative framing inspired by quantum decision theory. Rather than treating classification as boundary fitting, we reformulate the problem as one of state discrimination, where samples are encoded as states and decisions emerge through optimal measurement. This reframing does not claim universal speedup or accuracy gains; instead, it provides a principled way to reason about error limits, robustness, and failure modes that are obscured in classical formulations. Using an EMBER-inspired feature model and simulated quantum state representations, we demonstrate how this perspective clarifies the structural limits of classical malware detection and motivates more trustworthy decision-making in security-critical environments.

1 Introduction

Modern malware detection pipelines rely heavily on supervised machine learning models trained on engineered feature sets. Datasets such as EMBER exemplify this approach by representing executables as fixed-length vectors derived from byte statistics, opcode frequencies, section metadata, entropy measures, and import information. Classification is then performed by learning a decision surface that partitions this feature space into benign and malicious regions.

Despite continued architectural innovation, these systems face persistent challenges. As feature dimensionality increases, distance metrics lose semantic meaning, correlations between features grow dense, and decision boundaries become increasingly sensitive to small input perturbations. These issues are exacerbated in adversarial settings, where attackers explicitly seek to exploit boundary fragility through obfuscation, packing, or minor behavioral modifications.

This work argues that many of these limitations arise not from insufficient data or model capacity, but from the underlying geometric framing of the classification problem itself. We explore an alternative perspective rooted in quantum decision theory, where classification is treated as a problem of distinguishing states under optimal measurement rather than fitting a surface through feature space.

2 Classical Decision Boundaries in High-Dimensional Feature Spaces

In the classical paradigm, a malware classifier seeks a function

$$f: \mathbb{R}^n \rightarrow \{0, 1\}$$

where each sample is represented as a point in an n -dimensional feature space. The classifier implicitly or explicitly learns a decision boundary defined by

$$f(x) = \text{sign}(w^\top x + b),$$

or a nonlinear generalization thereof.

As n grows large, several geometric pathologies emerge. Distances between points concentrate, making nearest-neighbor intuition unreliable. Feature correlations introduce redundancy that inflates apparent dimensionality without adding discriminative power. Decision boundaries become highly curved and sensitive to small perturbations, resulting in false positives and false negatives even when semantic behavior is unchanged.

These effects are not failures of optimization but consequences of imposing Euclidean geometry on spaces where meaningful structure is sparse and adversarially manipulated.

3 Boundary Fragility and Perturbation Sensitivity

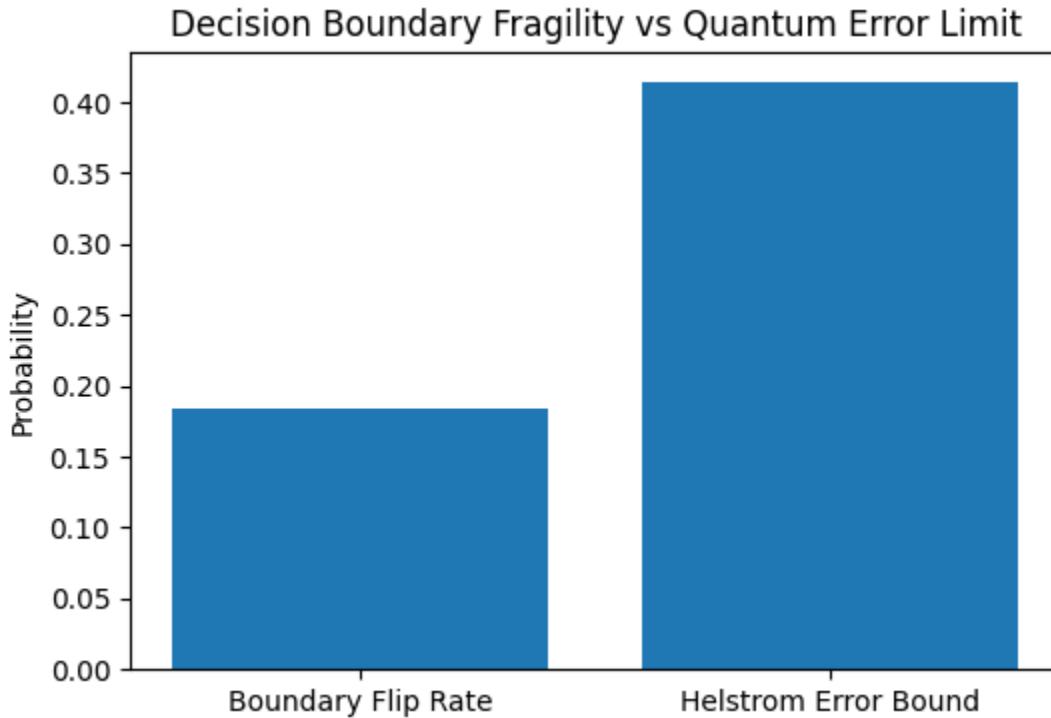
A defining characteristic of boundary-based classifiers is their sensitivity to small perturbations. Let x denote a feature vector and δ a small perturbation such that $\|\delta\| \ll \|x\|$. In practice, it is frequently observed that

$$f(x) \neq f(x + \delta),$$

even when δ corresponds to benign transformations such as minor packing changes or metadata variation.

In the accompanying notebook, this phenomenon is quantified by measuring the boundary flip rate under small Gaussian perturbations applied to EMBER-like feature vectors. The resulting instability highlights a core limitation: decision boundaries encode historical clustering rather than intrinsic behavioral distinctions.

Figure 1: Comparison of classical boundary flip rate under perturbation and the quantum Helstrom error bound.



4 Quantum-Inspired Reframing: Feature Vectors to States

The quantum reframing begins by abandoning the assumption that samples should be treated as static points in feature space. Instead, each feature vector $x \in \mathbb{R}^n$ is mapped to a normalized state vector

$$|\psi_x\rangle = \frac{x}{\|x\|},$$

which may be interpreted as a pure state in a Hilbert space.

In this formulation, information is not encoded solely in individual coordinates but in the relationships between components. Superposition and interference allow correlated structure to be represented implicitly, rather than approximated by explicit geometric surfaces.

Importantly, this encoding does not introduce randomness or noise. It provides a structured way to represent uncertainty and correlation without forcing them into Euclidean distance metrics.

5 State Discrimination and the Helstrom Bound

Once samples are represented as states, classification becomes a problem of state discrimination. Given two states $|\psi_0\rangle$ and $|\psi_1\rangle$ with prior probabilities π_0 and π_1 , the minimum achievable error probability under optimal measurement is given by the Helstrom bound:

$$P_e = \frac{1}{2} \left(1 - \sqrt{1 - 4\pi_0\pi_1 |\langle\psi_0|\psi_1\rangle|^2} \right).$$

This bound represents a fundamental limit. Unlike classical error rates, it cannot be reduced by additional training, architectural complexity, or hyperparameter tuning. It depends solely on the intrinsic distinguishability of the underlying states.

In the notebook, class-mean benign and malicious states are constructed and their overlap is used to compute the Helstrom bound. This provides a principled reference point against which classical boundary behavior can be compared.

6 What Improves and What Does Not

The quantum reframing offers specific, constrained advantages. It provides robustness to small perturbations by tying error to state overlap rather than boundary location. It clarifies the distinction between model failure and data indistinguishability. It offers interpretable limits on achievable performance that are absent in classical formulations.

However, it does not provide universal speedup, guaranteed accuracy improvements, or reductions in data requirements. Training costs, state preparation complexity, and measurement overhead remain significant considerations. When states are fundamentally indistinguishable, no method—classical or quantum—can overcome that limitation.

This disciplined scope is essential. The value of the reframing lies in conceptual clarity and robustness, not in exaggerated performance claims.

7 Implications for Malware Detection

In adversarial domains such as malware detection, trust and interpretability are as important as raw accuracy. A classifier that fails unpredictably under minor perturbations is operationally dangerous, even if its average performance appears strong.

By reframing classification as state discrimination, we gain a framework that exposes intrinsic limits, highlights robust decision criteria, and avoids the illusion that increasingly complex boundaries necessarily yield better security outcomes.

8 Conclusion

This work demonstrates that many persistent challenges in malware classification stem

from the classical geometric framing of the problem. By adopting a quantum-inspired perspective rooted in state discrimination and optimal measurement, we gain a more principled understanding of robustness, error limits, and failure modes. While this reframing does not promise universal improvement, it offers a structurally sound foundation for trustworthy decision-making in high-dimensional, adversarial environments.

Reproducibility

All experiments presented here are reproducible using the accompanying Jupyter notebook. The dataset is synthetic but EMBER-inspired, and all quantum behavior is simulated mathematically without reliance on quantum hardware.