

Agatha: Gated Hopfield Attention for Adversarially Robust Natural Language Processing

Andreia Podasca

Drexel University

Philadelphia, PA, USA

andreia.podasca@drexel.edu

Anup Das

Drexel University

Philadelphia, PA, USA

anup.das@drexel.edu

Abstract

Transformer encoders are used for natural language understanding tasks, yet they remain vulnerable to gradient-based attacks in the embedding space. Existing defenses either use adversarial training or modify the inference pipeline, incurring additional computational overhead without a principled mechanism to contract the gradient flow. We propose **Agatha**, a transformer encoder that achieves adversarial robustness through architectural modifications. First, Agatha replaces softmax attention with an iterative Hopfield mechanism that projects representations into deep, sharply-bounded semantic attractors in its energy landscape. Second, Agatha introduces a gated residual connection that scales down the identity path, forcing the adversarial gradients to propagate through the energy landscape where they contract layer by layer due to its steep curvature. Evaluations on standard NLP tasks demonstrate that Agatha reduces attack success rates by 13 percentage points over the best defended baseline at 4.37M scale, with the advantage growing to over 33 points at BERT-base scale (110M), without any threat-model-specific training. AutoAttack, transfer attacks, and discrete text-level evaluations confirm the robustness is not an artifact of gradient masking. Agatha adds no parameters, achieves $1.2\times$ faster inference than the best robust baseline, and maintains clean performance within 2 points of the undefended model.

1 Introduction

Transformers are prevailing architectures for NLP (Vaswani et al., 2017; Devlin et al., 2019). However, they remain vulnerable to adversarial attacks that apply small perturbations to alter model predictions (Alzantot et al., 2018; Chang et al., 2021). In the embedding space, first-order attacks such as Projected Gradient Descent (PGD) serve as the standard benchmark for evaluating model robustness (Goodfellow et al., 2014; Miyato et al., 2016). PGD relies on the gradient of the loss flowing back

to the input embedding to compute bounded adversarial perturbations (Madry et al., 2017).

Existing defenses fall into two categories. Training-based defenses augment the training set with adversarial examples (Zhu et al., 2019; Xhonneux et al., 2024), while architectural defenses add auxiliary components, such as denoising layers, attention-masking modules, or inference-time randomization (Li et al., 2025; Shen et al., 2023; Moon et al., 2023). Both approaches incur additional inference-time overhead, which can be prohibitive at small model scales. These scales are common in edge deployment and on-device NLP, where robustness cannot be traded for additional compute.

More fundamentally, neither stops the adversarial gradient from reaching the input embedding. During attack, the adversarial gradient propagates along two paths (Figure 1, top): the residual connection, which acts as a direct highway passing signals unchanged, and the softmax attention sub-layer, which provides minimal attenuation.

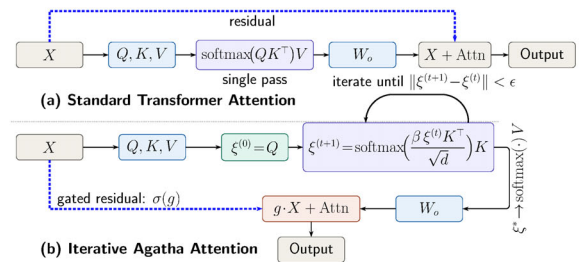


Figure 1: **Top:** Standard Transformers provide an unimpeded residual path for backpropagation. **Bottom:** Agatha scales the residual path and forces representations through an iterative Hopfield energy landscape.

Figure 1 (bottom) illustrates our proposed solution **Agatha**¹. It introduces two architectural modifications: an iterative attention mechanism to contract the backward-flowing gradient and a gated residual to block the free gradient highway.

Iterative Attention. We replace standard softmax

¹**Agatha** = *Adversarially robust gated hopfield attention*. Code: <https://github.com/drexel-DISCO/agatha>

attention with an iterative mechanism grounded in Modern Hopfield Networks (Ramsauer et al., 2020). The mechanism iteratively projects representations toward fixed points in an energy landscape. We design the architecture to operate in the deep attractor regime, where these fixed points form sharply-bounded semantic basins. A per-layer contraction condition bounds the gradient at each step, and because the mechanism iterates to convergence, this attenuation compounds across steps. We introduce two theorems to establish the conditions under which the attractor regime exists and contraction occurs; Section 5.11 empirically validates these conditions, with proofs in Appendix L.

Gated Residual. The residual connection cannot simply be removed: doing so destabilizes optimization in deep transformers (Xiong et al., 2020) and discards the surface-level linguistic information (lexical and morphosyntactic features) that downstream NLP tasks rely on alongside learned representations. The amount of surface-level information the combined representation must retain varies across tasks. Consider **sentiment classification**:

“The film’s ambition is undeniable and its visuals occasionally breathtaking, but the screenplay cannot decide whether it wants to be a character study or a thriller.”

Despite positive markers (**undeniable**, **breathtaking**), the negative verdict lies in the discourse structure following **but**. The attention mechanism compresses this into a polarity-bearing representation. The classifier does not need to recover the specific surface tokens that carried sentiment. Scaling down the residual path would have little effect on task performance.

Consider a **paraphrase detection** scenario:

A: “The committee declined to endorse the preliminary findings, though members indicated they might reconsider pending additional data.”

B: “Preliminary findings were not endorsed by the committee, though endorsement remained possible if further data supported the conclusions.”

The two sentences above share few n-grams and differ in voice, nominalization, and modality. The attention mechanism aligns corresponding spans, but the combined representation must retain enough original lexical forms to verify *what* was aligned, not merely *that* alignment occurred. Aggressive scaling of the residual path would discard these surface features, degrading performance.

To restrict the residual path without discarding the surface-level information that tasks require, we introduce a gate g that scales the residual connec-

tion. This gate weakens the residual gradient and forces any remaining adversarial signal through the Hopfield attractor landscape, where the iterated mechanism contracts it at each step. Together, these two mechanisms contract adversarial gradients without sacrificing task performance.

Contributions. Our contributions are as follows: (1) We introduce a structural modification that achieves adversarial robustness without any threat-model-specific training. (2) We rule out gradient masking through AutoAttack, transfer attacks, and Hopfield theory-aware attacks. (3) We demonstrate through ablation that gating and iterative Hopfield attention are jointly necessary. (4) We validate across standard NLP tasks, showing a 13 percentage point reduction in attack success at 4.37M parameter scale relative to the best defended baseline, with the advantage growing to 33 points at BERT-base scale (110M). Agatha achieves $1.2\times$ faster inference than the best robust baseline, and maintains clean performance within 2 points of the undefended model.

2 Preliminaries

2.1 Adversarial Attacks

Adversarial attacks in NLP can operate at the discrete token level or in the continuous embedding space. We evaluate both, but focus our analysis on embedding-space attacks, which apply perturbations directly to dense input representations. Given a classifier f and an input $\mathbf{x}$ with true label y , an attacker seeks a continuous perturbation δ such that $f(\mathbf{x} + \delta) \neq y$. This objective is formulated as maximizing the classification loss $\mathcal{L}$ with respect to δ , constrained by a budget ε under a chosen norm that bounds the perturbation magnitude (we use ℓ_2 as the primary norm and also evaluate under ℓ_∞):

$$\max_{\|\delta\| \leq \varepsilon} \mathcal{L}(f(\mathbf{x} + \delta), y). \quad (1)$$

Projected Gradient Descent (PGD) solves this iteratively by taking gradient steps and projecting back onto the constraint set (Madry et al., 2017):

$$\delta^{(t+1)} = \Pi_{\|\cdot\| \leq \varepsilon} \left(\delta^{(t)} + \alpha \cdot \text{sign}(\nabla_{\delta} \mathcal{L}(f(\mathbf{x} + \delta^{(t)}), y)) \right) \quad (2)$$

where α is the step size and Π denotes projection onto the ε -ball. PGD is considered the standard first-order benchmark; robustness against PGD generally implies robustness against other first-order methods. The effectiveness of PGD depends on how well the gradient propagates back to the input embedding, motivating the architecture of Agatha.

2.2 Modern Hopfield Networks

Modern Hopfield Networks (MHNs) generalize classical Hopfield networks to continuous states (Ramsauer et al., 2020). An MHN defines an energy function over a query state ξ and a set of stored patterns (keys) $\mathbf{K} = [\mathbf{k}_1, \dots, \mathbf{k}_N]^\top$:

$$E(\xi) = -\frac{1}{\beta} \log \sum_{i=1}^N \exp(\beta \mathbf{k}_i^\top \xi) + \frac{1}{2} \|\xi\|^2 + \text{const}, \quad (3)$$

where $\beta > 0$ is the inverse temperature. A **semantic attractor** is a local minimum ξ^* of $E(\xi)$ located near a stored key pattern $\mathbf{k}_i$, and its **basin** is the set $\mathcal{B}_i = \{\xi : \arg \max_j \mathbf{k}_j^\top \xi = i\}$ of queries that retrieve that pattern. The depth and sharpness of these basins increase with β .

Minimizing this energy yields a fixed-point condition equivalent to softmax attention: the query attends to keys to retrieve a weighted combination of stored patterns. At low β , the softmax is uniform and all queries converge to the mean of stored patterns. At high β , the softmax saturates and each query converges to a single pattern. We leverage the high- β regime to design Agatha.

3 Agatha – Gated Hopfield Attention

Agatha modifies two components of the standard Transformer encoder.

3.1 Iterative Hopfield Attention

We replace standard dot-product attention with an iterative update mechanism grounded in the energy function of Eq. (3). Given queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ projected from the input $\mathbf{x}$, we initialize the query state $\xi^{(0)} = \mathbf{Q}$. We iterate the energy update function until convergence or a maximum number of steps $T_{\max}$ is reached:

$$\xi^{(t+1)} = \mathbf{K}^\top \text{softmax} \left(\frac{\beta}{\sqrt{d}} \mathbf{K} \xi^{(t)} \right), \quad (4)$$

where d is the head dimension and the $\sqrt{d}$ scaling follows standard practice for numerical stability. Convergence is defined as:

$$\|\xi^{(t+1)} - \xi^{(t)}\| / \|\xi^{(t)}\| < \tau \quad (5)$$

for a threshold τ . Once converged at step T , the final attention output retrieves from value vectors:

$$\text{output} = \mathbf{V}^\top \text{softmax} \left(\frac{\beta}{\sqrt{d}} \mathbf{K} \xi^{(T)} \right). \quad (6)$$

By operating at high inverse temperature β , we place the mechanism in the **deep attractor regime** of Theorem 2, the phase $\beta > \beta_{\text{upper}}$ where the

softmax saturates and each query converges to a fixed point adjacent to a single stored pattern. We describe basins in this regime as **sharply-bounded** because their walls steepen as β grows (Section 2.2). Representations converge to basin floors rather than settling in diffuse, easily perturbed regions of the energy surface. Because the mechanism iterates to convergence, this contraction compounds across steps, attenuating the adversarial gradient at each iteration of the energy descent.

3.2 Gated Residual Connection

In a standard Transformer, a sublayer F computes $\mathbf{x}_{l+1} = \mathbf{x}_l + F(\text{LN}(\mathbf{x}_l))$, where LN denotes Layer Normalization. During backpropagation, the Jacobian of this layer is $\mathbf{I} + \frac{\partial F}{\partial \mathbf{x}_l}$. The identity component provides a direct path that passes the adversarial gradient $\nabla_{\delta} \mathcal{L}$ without attenuation, ensuring $\rho(J) \geq 1$ regardless of the sublayer. To block this bypass, Agatha introduces a scaling gate $g \in (0, 1)$:

$$\mathbf{x}_{l+1} = g \cdot \mathbf{x}_l + F(\text{LN}(\mathbf{x}_l)). \quad (7)$$

This scales the Jacobian to $g\mathbf{I}$, reducing the adversarial signal that bypasses the Hopfield attention.

The gate can be set in two ways: fixed by the practitioner, or learned from scratch during training. Appendix F evaluates both options. Because clean cross-entropy training provides no gradient pressure to close the gate, we treat g as a structural hyperparameter. A fixed $g = 0.7$ outperforms every defended baseline, while per-task tuning yields a further 4 percentage point improvement, reflecting the task-dependent surface-information needs described in Section 1. Under the combined architecture, the PGD gradient propagates as:

$$\nabla_{\delta} \mathcal{L} = \frac{\partial \mathcal{L}}{\partial \mathbf{x}_L} \prod_{l=1}^L (g\mathbf{I} + \mathcal{J}_F^{(l)}) \quad (8)$$

where $\mathcal{J}_F^{(l)}$ is the Jacobian of the Hopfield block at layer l . We design g to satisfy $g + \sigma_{\max} < 1$ (Theorem 1), with $\sigma_{\max} = \|\mathcal{J}_F\|_2$, which bounds the layer Jacobian as $\|\mathcal{J}\|_2 \leq g + \sigma_{\max} < 1$. We refer to this property as **gradient contraction**: each gated layer shrinks the backward gradient by at least this factor, and the signal available to PGD decays exponentially with depth. In the **forward pass**, perturbed queries contract toward energy minima. In the **backward pass**, the steep basin geometry contracts the adversarial gradient (Figure 2).

Overall, Agatha achieves adversarial robustness by creating a geometric trap through Hopfield attention and ensuring adversarial gradients cannot bypass it through the gated residual.

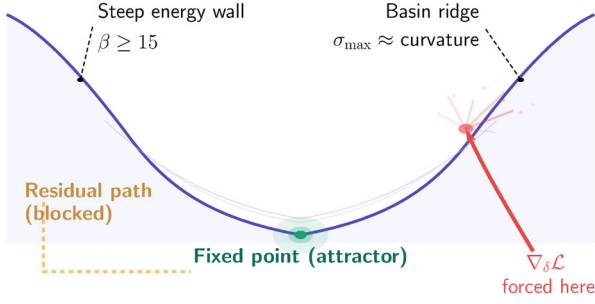


Figure 2: Gradient contraction within the Hopfield energy well. Steep geometric ridges of the well deflect and contract the adversarial gradient $\nabla_{\delta} \mathcal{L}$.

3.3 Initialization and Parameter Count.

Agatha reuses the pretrained $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$, and output projections of the source checkpoint without modification, so the mapping from a pretrained Transformer to Agatha is one-to-one and adds no parameters at any scale (Section 5; Appendix Table 13). The two modifications are computational rather than parametric: the softmax retrieval is applied iteratively (Eqs. 4 to 6), and the residual path is scaled by the scalar g , a structural hyperparameter rather than a learned weight. Initialization is therefore identical to the original checkpoint.

The following section formalizes the contraction and attractor conditions.

4 Robustness Analysis

The robustness of Agatha arises from the geometry of the Hopfield energy landscape and its interaction with the restricted residual flow. We formalize these properties through two core theorems: contraction of the layer Jacobian bounds how much adversarial signal reaches the input embedding, and a phase transition in the energy landscape determines when representations converge to discrete semantic attractors. Full proofs are provided in Appendix L.

Theorem 1 (Contraction via Identity Gating). *Let the gated Hopfield step compute $\mathbf{x}_{l+1} = g \cdot \mathbf{x}_l + F(\text{LN}(\mathbf{x}_l))$ with Jacobian $\mathcal{J} = g\mathbf{I} + \mathcal{J}_F$. Let $\sigma_{\max} := \|\mathcal{J}_F\|_2$ denote the operator 2-norm of the Hopfield block $\mathcal{J}_F$. Assuming stable iterative convergence, the system enforces a global contraction ($\|\mathcal{J}\|_2 < 1$, which implies $\rho(\mathcal{J}) < 1$) when $g + \sigma_{\max} < 1$.*

Theorem 1 establishes a sufficient condition for contraction: the architecture is designed to make this condition hold in practice, and Appendix D (Table 39) validates $\sigma_{\max} < 1 - g$.

Theorem 2 (Semantic Phase Transition). *Define the thermodynamic state function $f(\beta) = \beta \cdot \|\mathbf{x}\|^2$.*

$\max_i p_i(\beta) \cdot (1 - \max_i p_i(\beta))$. For a given query ξ , there exist input-dependent critical boundaries $\beta_{\text{lower}}(\xi) < \beta_{\text{upper}}(\xi)$ where $f(\beta) = 1$, defining three regimes: (1) $\beta < \beta_{\text{lower}}$: a spin-glass phase causing contraction to a featureless mean, (2) $\beta_{\text{lower}} \leq \beta \leq \beta_{\text{upper}}$: a metastable phase exhibiting no contraction, and (3) $\beta > \beta_{\text{upper}}$: the deep attractor phase causing contraction to discrete semantic attractors.

Section 5.11 validates that $\beta \geq 15$ is sufficient to operate in the deep attractor regime across all tasks and seeds evaluated in this paper.

Basin Ridge Deflection. At deep attractor temperatures ($\beta > \beta_{\text{upper}}$), the boundaries between semantic basins form steep geometric ridges (Fig. 2). Because the residual path is restricted, adversarial gradients directed across a basin boundary propagate against the high local curvature of the ridge, contracting the signal before it reaches the input embeddings. Section 5.13 provides empirical evidence consistent with this mechanism.

The regime and the contraction factor are established by Theorems 1 and 2, and measured in Sections 5.11 and 5.13. The ridge description above is geometric intuition for the operator norm $\sigma_{\max}$, which bounds the first-order sensitivity of the block rather than the curvature of the energy surface.

5 Results

5.1 Experimental Setup

Datasets and Architecture. We evaluate on eight GLUE tasks (Wang et al., 2018) (SST-2, MNLI, QNLI, QQP, CoLA, MRPC, RTE, WNLI) using a 2-layer, 2-head architecture with 128-dimensional hidden state (4.37M parameters). We also evaluate on BoolQ and CB from SuperGLUE (Wang et al., 2019) using the same architecture. To test scaling, we use a 256-dimensional, 4-layer, 4-head architecture (11.1M parameters) for MultiRC and WSC. All metrics are averaged across three random seeds.

Threat Model. Our primary evaluation uses PGD under an ℓ_2 constraint across perturbation budgets $\varepsilon \in [0.05, 5.0]$. To verify that the observed robustness is not an artifact of gradient masking, we evaluate under AutoAttack (Croce and Hein, 2020b), an ensemble of attacks (APGD-CE and FAB) designed to detect defenses that obfuscate rather than eliminate gradients. All attacks backpropagate through the fully unrolled Hopfield iteration, with no truncation or fixed-point approximation of the converged state; Appendix A.7 states the protocol and its cost.

Model	Clean ($\uparrow$) ($\epsilon = 0$)	Attack Success Rate (% , $\downarrow$)					
		Low ($\epsilon = 0.05$)	Medium			High ($\epsilon = 3.0$)	Extreme ($\epsilon = 5.0$)
			($\epsilon = 0.5$)	($\epsilon = 1.0$)	($\epsilon = 2.0$)		
Undeended Baseline							
TinyBERT	64.28 $\pm$ 0.52	2.48 $\pm$ 0.16	21.37 $\pm$ 1.01	40.10 $\pm$ 1.62	72.27 $\pm$ 1.71	90.94 $\pm$ 0.96	99.75 $\pm$ 0.06
Training-Based Robustness							
TinyBERT + FreeLB	67.19 $\pm$ 0.33	2.09 $\pm$ 0.26	18.99 $\pm$ 1.37	35.50 $\pm$ 1.91	66.68 $\pm$ 2.07	85.97 $\pm$ 1.01	98.12 $\pm$ 0.24
Architectural Robustness							
DiffuseDef	66.22 $\pm$ 0.57	1.74 $\pm$ 0.38	18.95 $\pm$ 2.58	38.07 $\pm$ 4.59	69.85 $\pm$ 4.69	90.30 $\pm$ 1.84	99.57 $\pm$ 0.17
DynAttn	65.77 $\pm$ 0.62	2.55 $\pm$ 0.36	16.68 $\pm$ 1.01	33.37 $\pm$ 1.38	63.58 $\pm$ 2.53	84.58 $\pm$ 2.41	99.14 $\pm$ 0.39
Agatha (Ours)	63.91 $\pm$ 0.20	1.48 $\pm$ 0.24	12.75 $\pm$ 1.32	26.16 $\pm$ 2.74	50.28 $\pm$ 3.13	65.58 $\pm$ 3.19	82.03 $\pm$ 2.18

Table 1: Average GLUE Score and Attack Success Rate (ASR) of PGD across 8 tasks. Clean performance is evaluated without perturbations. ASR is reported across varying perturbation budgets grouped by severity. Results are reported as mean $\pm$ std across three random seeds. **Bold** indicates best performance.

To test threat-model generalization, we report results under the ℓ_∞ norm, **FGSM** (Goodfellow et al., 2014), **Gaussian noise** (Wang et al., 2025), and two text-level attacks: **CharSwap** (Pruthi et al., 2019) and **TextBugger** (Li et al., 2018).

Baselines. We compare against an undefended TinyBERT baseline (Jiao et al., 2020), a training-based defense FreeLB (Zhu et al., 2019), and two architectural defenses: DiffuseDef (Li et al., 2025) and Dynamic Attention (Shen et al., 2023).

Evaluation Metrics. For PGD and FGSM, we report Attack Success Rate (ASR), the percentage of correct predictions flipped by the perturbation. For Gaussian noise and text-level attacks, we report Flip Rate at varying magnitudes. Metric normalization details are in Appendix A.

5.2 Main Results: Architectural Robustness

Table 1 summarizes PGD performance.

Clean Performance. Agatha achieves a clean score of 63.91, within one point of the TinyBERT baseline (64.28) using standard cross-entropy loss. FreeLB achieves the highest clean score (67.19) through adversarial training, with DiffuseDef (66.22) and DynAttn (65.77) following. None of these advantages translate into robustness.

Robustness Across Budgets. Agatha achieves the lowest ASR at every perturbation budget. At $\varepsilon = 2.0$, Agatha reduces ASR to 50.28, a 13.30-point reduction over the best architectural baseline DynAttn and a 16.40-point reduction over FreeLB. The advantage grows with budget: at $\varepsilon = 3.0$, the gap over DynAttn widens to 19.00 points (65.58 versus 84.58); at $\varepsilon = 5.0$, Agatha achieves 82.03 versus 99.14 for DynAttn and 98.12 for FreeLB. This pattern is consistent with our analysis: the energy basin structure provides a geometric constraint that operates independently of the training perturbation distri-

bution and persists at high budgets where other defenses degrade. Per-task breakdowns are in Appendix B.1 and a mechanistic interpretation of performance gaps in Appendix E.

5.3 Validation under AutoAttack

Architectures that modify gradient flow are susceptible to gradient masking (Athalye et al., 2018). To verify that Agatha’s robustness is genuine, we evaluate under **AutoAttack** (Croce and Hein, 2020b), taking the worst-case ASR between APGD-CE and FAB. Table 2 reports combined GLUE results under ℓ_2 ; detailed are in Appendix B.6.

Model	Clean ($\uparrow$)	ASR ($\downarrow$)		
		$\varepsilon = 1.0$	$\varepsilon = 2.0$	$\varepsilon = 3.0$
TinyBERT	64.28 $\pm$ 0.52	40.15 $\pm$ 1.63	71.69 $\pm$ 1.65	89.90 $\pm$ 0.85
FreeLB	67.19 $\pm$ 0.33	35.62 $\pm$ 1.91	66.65 $\pm$ 2.07	85.90 $\pm$ 0.97
DiffuseDef	63.44 $\pm$ 0.55	37.96 $\pm$ 3.75	69.03 $\pm$ 3.80	89.25 $\pm$ 1.50
DynAttn	63.38 $\pm$ 0.21	30.77 $\pm$ 1.01	56.94 $\pm$ 2.10	76.21 $\pm$ 2.02
Agatha (Ours)	63.91 $\pm$ 0.20	27.08 $\pm$ 2.23	51.45 $\pm$ 2.59	66.56 $\pm$ 2.55

Table 2: AutoAttack (Croce and Hein, 2020b) under ℓ_2 . **Bold**: highest clean score and lowest ASR.

Agatha achieves the lowest ASR at every budget: 51.45 at $\varepsilon = 2.0$, a 5.49 point gap over DynAttn (56.94), 17.58 over DiffuseDef (69.03), and 15.20 over FreeLB (66.65). We additionally evaluate the gradient-free Square attack (Andriushchenko et al., 2020), which recovers an order of magnitude less ASR than the white-box components (Appendix B.6). Gradient masking predicts the opposite ordering, so these results are consistent with genuine geometric suppression.

5.4 Norm Generalization

To test threat-model generalization, we evaluate all defenses under PGD with an ℓ_∞ constraint. Table 3 reports combined GLUE ASR at three representative budgets. Agatha achieves the lowest ASR at every budget: 11.69 at $\varepsilon = 0.01$, a 4.05 point gap over DynAttn (15.74). The gap grows at higher bud-

gets, reaching 16.82 points at $\varepsilon = 0.1$ (77.60 versus 94.42). This norm-independent robustness is consistent with the geometric nature of the Hopfield attractor structure, which operates on representation magnitude rather than a specific perturbation metric. Detailed results are in Appendix B.7.

Model	Clean	ASR ($\downarrow$)		
	($\uparrow$)	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$
TinyBERT	64.28 $\pm$ 0.52	19.03 $\pm$ 1.06	72.94 $\pm$ 1.89	94.98 $\pm$ 0.50
FreeLB	67.19 $\pm$ 0.33	20.82 $\pm$ 1.40	82.56 $\pm$ 1.79	99.45 $\pm$ 0.07
DiffuseDef	63.44 $\pm$ 0.55	17.24 $\pm$ 1.75	71.63 $\pm$ 3.28	95.47 $\pm$ 0.44
DynAttn	63.38 $\pm$ 0.21	15.74 $\pm$ 0.98	67.85 $\pm$ 2.31	94.42 $\pm$ 0.39
Agatha (Ours)	63.91 $\pm$ 0.20	11.69 $\pm$ 1.37	54.36 $\pm$ 2.02	77.60 $\pm$ 1.24

Table 3: Standard PGD under ℓ_∞ : Combined GLUE ASR. **Bold**: highest clean score and lowest ASR.

5.5 Transfer Attack Analysis

To rule out gradient masking, we conduct transfer attack experiments on SST-2.

Setup. We evaluate four attack directions: (1) TinyBERT $\rightarrow$ Agatha (**transfer**), (2) Agatha $\rightarrow$ TinyBERT (**transfer**), (3) TinyBERT $\rightarrow$ TinyBERT (**white-box**), and (4) Agatha $\rightarrow$ Agatha (**white-box**). Table 4 reports the target ASR and transfer rate (target ASR divided by source ASR) at $\varepsilon = 2.0$; full results in Appendix B.4.

Direction	Source ASR	Target ASR	Transfer Rate
TB $\rightarrow$ Agatha	59.90	15.63	26.1%
Agatha $\rightarrow$ TB	44.35	21.34	48.1%
TB $\rightarrow$ TB	59.90	59.90	100%
Agatha $\rightarrow$ Agatha	44.35	44.35	100%

Table 4: Transfer rates at $\varepsilon = 2.0$ on SST-2. Transfer rate = target ASR / source ASR. A high transfer rate under TB $\rightarrow$ Agatha would indicate gradient masking.

Agatha’s Robustness is Not Gradient Masking. If Agatha’s robustness were due to gradient masking, adversarial examples crafted on TinyBERT should transfer effectively to Agatha, with TB $\rightarrow$ Agatha target ASR approaching the TB $\rightarrow$ TB white-box ASR of 59.90. Instead, TB $\rightarrow$ Agatha records only 15.63, a transfer rate of 26.1%. This confirms that the energy basins neutralize adversarial perturbations rather than just hiding gradients.

Asymmetric Transfer. Attacks transfer more easily from Agatha to TinyBERT (48.1% rate) than from TinyBERT to Agatha (26.1% rate). This asymmetry is consistent with the basin geometry: adversarial directions that escape Agatha’s steep energy basins are disruptive, transferring effectively to the undefended model. Conversely, perturbations crafted on TinyBERT’s smooth loss surface lack the directionality to escape Agatha’s deep basins.

5.6 Generalization Beyond PGD

To verify that Agatha’s robustness generalizes beyond PGD, we evaluate against four additional at-

tack types: FGSM and random Gaussian perturbations in the embedding space, and two text-level attacks, CharSwap and TextBugger. Table 5 reports average GLUE results across all 8 tasks; full per-task breakdowns are in Appendix C.

Attack	Budget	FreeLB	DD	DA	Agatha
<i>Embedding-Space Attacks (ASR $\downarrow$)</i>					
PGD	$\varepsilon = 2.0$	66.68 $\pm$ 2.07	69.85 $\pm$ 4.69	63.58 $\pm$ 2.53	50.28 $\pm$ 3.13
FGSM	$\varepsilon = 2.0$	60.97 $\pm$ 2.65	63.28 $\pm$ 5.30	57.56 $\pm$ 2.62	46.87 $\pm$ 3.29
Gaussian	$\sigma = 2.0$	38.65 $\pm$ 2.86	30.35 $\pm$ 2.16	29.75 $\pm$ 1.41	23.95 $\pm$ 2.26
<i>Text-Level Attacks (Flip Rate $\downarrow$)</i>					
CharSwap	30%	25.29 $\pm$ 2.25	16.18 $\pm$ 2.11	18.24 $\pm$ 1.60	12.12 $\pm$ 0.75
TextBugger	30%	20.38 $\pm$ 1.52	13.19 $\pm$ 1.67	14.78 $\pm$ 1.34	10.29 $\pm$ 0.71

Table 5: Avg. GLUE flip rate/ASR. DD = DiffuseDef, DA = DynAttn. **Bold**: lowest ASR or flip rate.

Agatha achieves the lowest attack success rate across all five attack types. On embedding-space attacks, the advantage over baselines is consistent: 6 to 13 points over the best architectural baseline on PGD, FGSM, and Gaussian noise. The robustness extends to text-level attacks despite Agatha operating in embedding space: CharSwap flip rates drop from 16.18 (DiffuseDef) to 12.12, and TextBugger from 13.19 to 10.29. This suggests that the energy basin structure provides general representation stability beyond the continuous perturbation model. Agatha also achieves the highest average score on AdvGLUE (Wang et al., 2021) (Appendix C.5) with curated text-level adversarial examples.

5.7 Hopfield Dynamics-Aware Attacks

To test robustness against an informed adversary, we design four attacks targeting the theoretical properties from Theorems 1 and 2: *Basin Escape* maximizes the distance between clean and perturbed Hopfield fixed points; *Spectral Radius* maximizes $\rho(\mathcal{J}_F)$ to break the contraction condition; *Convergence Disruption* maximizes the Hopfield convergence residual; and *Joint* combines cross-entropy with the spectral radius objective. Table 6 reports combined GLUE ASR at $\varepsilon = 2.0$.

Std. PGD	Basin Esc.	Spectral	Conv. Dis.	Joint
50.28 $\pm$ 3.13	4.00 $\pm$ 0.60	3.02 $\pm$ 0.55	3.83 $\pm$ 0.55	43.82 $\pm$ 3.50

Table 6: Theory-aware attacks on Agatha at $\varepsilon = 2.0$.

No theory-aware attack outperforms standard PGD: Basin Escape, Spectral Radius, and Convergence Disruption all achieve under 4% combined ASR compared to 50.28% for standard PGD, confirming that Agatha’s defense cannot be theoretically exploited. Full analysis in Appendix D.

5.8 Ablation Study

To verify that the gate and Hopfield attention are jointly necessary, we evaluate four ablation variants of Agatha: (1) *Gate Only*, which applies the gated residual to standard dot-product attention, (2) *Hopfield Only*, which uses iterative Hopfield attention with an unscaled residual, (3) *Energy Transformer (ET)* (Hoover et al., 2023), which replaces the Transformer block with iterative energy minimization, leaving the residual unscaled, and (4) *Hopfield+Reg*, which pairs Hopfield attention with a training-time regularizer instead of the architectural gate (Podasca and Das, 2026). Table 7 reports GLUE scores and ASR at $\varepsilon \in \{1.0, 2.0, 3.0\}$; full results are in Appendix G.

Model	Clean ($\varepsilon = 0$)	Attack Success Rate			
		$\varepsilon = 1.0$	$\varepsilon = 2.0$	$\varepsilon = 3.0$	
TinyBERT	64.28 $\pm$ 0.52	40.10 $\pm$ 1.62	72.27 $\pm$ 1.71	90.94 $\pm$ 0.96	
Gate Only	63.93 $\pm$ 0.69	41.67 $\pm$ 2.88	75.42 $\pm$ 2.23	92.49 $\pm$ 1.78	
Hopfield Only	64.01 $\pm$ 0.20	33.52 $\pm$ 3.95	57.93 $\pm$ 2.73	73.87 $\pm$ 1.85	
ET	64.41 $\pm$ 1.07	37.70 $\pm$ 4.00	61.56 $\pm$ 5.67	75.25 $\pm$ 5.87	
Hopfield+Reg	61.56 $\pm$ 0.65	36.81 $\pm$ 3.64	65.09 $\pm$ 3.30	81.06 $\pm$ 1.97	
Agatha (Ours)	63.91 $\pm$ 0.20	26.16 $\pm$ 2.74	50.28 $\pm$ 3.13	65.58 $\pm$ 3.19	

Table 7: Ablation study: Average GLUE Score and ASR at $\varepsilon \in \{1.0, 2.0, 3.0\}$. **Bold:** best performance.

Gate Only Hurts. Gate Only records higher ASR than TinyBERT at all budgets, confirming that scaling down the residual without replacing the attention mechanism degrades representational capacity without providing gradient contraction.

Hopfield Only and ET Help, But Not Enough. Hopfield Only reduces ASR from 72.27 to 57.93 at $\varepsilon = 2.0$. ET, which uses iterative energy minimization but retains the unscaled residual, records 61.56 at the same budget. However, Agatha achieves 50.28, a further 7.65 point reduction over Hopfield Only and 11.28 over ET. The gate alone provides no benefit, yet combining it with the Hopfield mechanism yields consistent gains across all budgets. This confirms Agatha’s mechanism: the gate forces gradients through the attention sublayer where the energy landscape contracts them.

Regularization Does Not Substitute for the Gate. Hopfield+Reg replaces the architectural gate with a training-time regularizer. It performs worse than Hopfield Only across budgets (65.09 versus 57.93 at $\varepsilon = 2.0$); the regularizer constrains the Hopfield representation without recovering the inference-time gradient contraction of the gate.

5.9 Scalability to Larger Architectures

To verify that Agatha’s robustness generalizes beyond the base architecture, we evaluate on four

SuperGLUE tasks. BoolQ and CB use the same 4.37M architecture as the GLUE experiments. MultiRC and WSC use a scaled architecture with 256 dimensions, 4 layers, and 4 heads (11.1M parameters). Table 8 reports clean accuracy and ASR at $\varepsilon = 2.0$; full results are in Appendix B.2.

	4.37M		11.1M		110M
	BoolQ	CB	MultiRC	WSC	GLUE
FreeLB	82.02 $\pm$ 0.83	60.51 $\pm$ 6.17	80.90 $\pm$ 2.12	64.19 $\pm$ 18.10	76.52 $\pm$ 0.54
DiffuseDef	75.66 $\pm$ 4.20	52.83 $\pm$ 7.70	66.23 $\pm$ 2.04	82.67 $\pm$ 5.77	91.61 $\pm$ 0.76
DynAttn	74.45 $\pm$ 7.55	46.67 $\pm$ 2.36	70.68 $\pm$ 2.52	75.09 $\pm$ 6.26	90.02 $\pm$ 1.46
Agatha (ours)	66.41 $\pm$ 3.81	29.34 $\pm$ 1.69	55.07 $\pm$ 3.59	13.86 $\pm$ 1.36	43.29 $\pm$ 0.44

Table 8: ASR at $\varepsilon = 2.0$ across model scales. SuperGLUE tasks at 4.37M (BoolQ, CB) and 11.1M (MultiRC, WSC). Combined GLUE at BERT-scale (110M). **Bold:** lowest ASR per column.

Base Architecture (BoolQ, CB). On BoolQ, Agatha achieves the lowest ASR (66.41), while both architectural baselines record higher ASR (74.45 for DynAttn, 75.66 for DiffuseDef) and the training-based FreeLB records 82.02. On CB, Agatha records 29.34 ASR, well below DynAttn (46.67), DiffuseDef (52.83), and FreeLB (60.51).

Scaled Architecture (MultiRC, WSC). On MultiRC, Agatha achieves the lowest ASR (55.07), while DiffuseDef (66.23) and DynAttn (70.68) both record higher ASR. Both baselines also suffer a considerable clean accuracy drop (50.19 and 52.94) compared to Agatha (58.75). On WSC, Agatha achieves 13.86 ASR, far below DiffuseDef (82.67) and DynAttn (75.09). These results confirm that Agatha’s geometric robustness mechanism scales with model capacity. The GLUE column in Table 8 further validates the scaling at full **BERT-scale** ($\approx 110M$ parameters), where Agatha achieves a 33 point ASR reduction over the best defended baseline at $\varepsilon = 2.0$, more than double the gap at TinyBERT scale. Per-task breakdowns at BERT-scale and additional analysis are in Appendix H.

5.10 Extension to Decoder-Only Architectures

To test whether the architectural modification applies to causal decoders, we inject Agatha attention with the gated residual into two backbones from the Pythia suite (Biderman et al., 2023), Pythia-14M and Pythia-70M, and fine-tune on SST-2. Under causal masking the stored patterns at each position are restricted to preceding tokens, so the masked softmax remains a valid Modern Hopfield update over the visible key set. Table 9 reports clean accuracy and ASR against the unmodified Pythia check-

point. We use $\beta = 0.13$, $T_{\max} = 2$, and $g = 0.5$. Perturbation budgets are reported as a fraction of the average input embedding norm rather than as absolute ℓ_2 budgets, because of the difference in the embedding scale. Normalizing a perturbation by the magnitude of the representation it perturbs follows the relative perturbation metric of Podasca and Das (2025), introduced to compare robustness across models whose features differ in scale. All attack settings otherwise follow Appendix A.7.

		Attack Success Rate (% , ↓)				
Model	Clean (↑)	1%	2%	5%	10%	20%
<i>Pythia-14M</i>						
Baseline	81.84 ±0.30	16.35 ±1.13	36.24 ±0.68	85.34 ±0.43	100.00 ±0.00	100.00 ±0.00
Agatha	82.49 ±0.14	13.30 ±1.06	28.54 ±1.66	79.75 ±2.29	99.91 ±0.07	100.00 ±0.00
<i>Pythia-70M</i>						
Baseline	82.45 ±0.41	25.82 ±0.77	57.90 ±2.18	97.64 ±0.41	100.00 ±0.00	100.00 ±0.00
Agatha	82.95 ±0.24	21.84 ±1.85	50.37 ±0.68	96.27 ±0.78	100.00 ±0.00	99.95 ±0.07

Table 9: Decoder-only results on SST-2 under ℓ_2 PGD. Baseline is the unmodified Pythia checkpoint at each scale. **Bold** indicates best performance: highest clean accuracy and lowest ASR. Differences at 10% and 20% fall within one standard deviation and are not marked.

Robustness transfers at low budgets. Agatha records lower ASR than the unmodified backbone at the three lowest budgets and at both scales. The largest reduction is 7.70 points at 14M and 2%, from 36.24 to 28.54, with a comparable 7.53 point reduction at 70M and the same budget. The reduction is smaller at the extremes of this range, at 3.05 points for 14M at 1% and 1.37 points for 70M at 5%.

Clean accuracy improves at both scales. Clean accuracy increases by 0.65 points at 14M and 0.50 points at 70M. This differs from the encoder setting, where Agatha gives up 0.37 points of combined GLUE score relative to TinyBERT (Table 1).

The separation closes at high budgets. At 10% and 20% both models are driven to near-complete failure and the two are statistically tied.

5.11 Validation of the Robustness Theory

We conduct hyperparameter sweeps on SST-2 to validate Theorems 1 and 2. The β -sweep fixes $g = 0.1$ and varies $\beta \in \{5, 10, 15, 20, 50, 100\}$. The g -sweep fixes $\beta = 50$ and varies $g \in \{0.01, 0.1, 0.2, 0.5, 1.0\}$. Table 10 reports clean accuracy and ASR at $\varepsilon = 2.0$; full results across all budgets are in Appendix I.

β Sweep Confirms the Phase Transition (Theorem 2). ASR peaks in the metastable regime (81.28 at $\beta = 5$, 70.79 at $\beta = 10$), then drops sharply to 49.98 at $\beta = 15$ and 45.99 at $\beta = 50$.

g Sweep Confirms the Contraction Condition (Theorem 1). ASR increases monotonically from 45.99 at $g = 0.1$ to 52.66 at $g = 1.0$, which recovers the Hopfield Only ablation. At $g = 0.01$, clean accuracy degrades to 68.31 (± 2.86), confirming that excessively small gates impair trainability. The gate must be small enough to enable contraction but large enough to preserve trainability; $g = 0.7$ provides strong robustness across tasks (Section 5.12).

β	β sweep ($g = 0.1$)			g	g sweep ($\beta = 50$)	
	Clean	ASR $\varepsilon = 2.0$			Clean	ASR $\varepsilon = 2.0$
5	79.63 $\pm$ 0.84	81.28 $\pm$ 1.07	0.01		68.31 $\pm$ 2.86	39.10 $\pm$ 14.55
10	79.78 $\pm$ 0.98	70.79 $\pm$ 12.59	0.1		80.20 $\pm$ 0.14	45.99 $\pm$ 2.79
15	79.36 $\pm$ 0.92	49.98 $\pm$ 11.00	0.2		80.89 $\pm$ 0.78	48.22 $\pm$ 3.16
50	80.20 $\pm$ 0.14	45.99 $\pm$ 2.79	0.5		80.77 $\pm$ 0.24	49.97 $\pm$ 5.28
100	80.62 $\pm$ 0.66	53.68 $\pm$ 5.41	1.0		80.05 $\pm$ 0.57	52.66 $\pm$ 2.53

Table 10: Hyperparameter sweeps on SST-2. **Bold:** best robustness with clean accuracy above 80.

5.12 Gate Selection

To demonstrate that Agatha’s robustness is architectural rather than hyperparameter-dependent, we evaluate with a fixed gate across all tasks without task-specific tuning. Figure 3 (left) reports ASR at $\varepsilon = 2.0$ across fixed gate values.

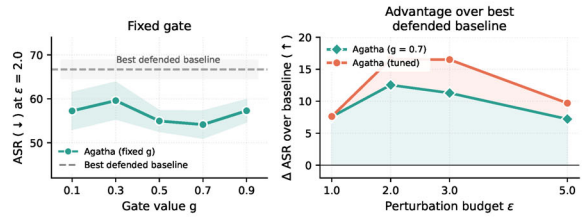


Figure 3: Left: ASR across fixed gate values at $\varepsilon = 2.0$. Shaded region: ± 1 std. Dashed line: best defended baseline. Right: advantage (Δ ASR) of $g = 0.7$ and per-task tuned Agatha over the best defended baseline across perturbation budgets.

Figure 3 (right) shows the advantage over the best defended baseline across budgets. From $\varepsilon = 2.0$ onward every fixed gate value outperforms the best defended baseline, and we recommend $g = 0.7$, which leads at every budget; clean accuracy remains stable (63.78 to 64.53). Per-task tuning yields a further 4 point advantage at $\varepsilon = 2.0$, but is not required. Full fixed-gate results are in Appendix F.1, per-layer asymmetric gating in Appendix F.2, and training the gate from scratch in Appendix F.3.

5.13 Validation of Gradient Contraction

To verify that Agatha’s architectural modifications contract adversarial gradients as predicted, we measure gradient flow during PGD attacks on MRPC. For each gated sublayer $\mathbf{x}_{l+1} = g \cdot \mathbf{x}_l + F(\text{LN}(\mathbf{x}_l))$, we decompose the backward-pass gradient into a

residual-path contribution $g \cdot \mathbf{u}$ and a sublayer-path contribution $\mathcal{J}_F^\top \mathbf{u}$, where $\mathbf{u} = \partial \mathcal{L} / \partial \mathbf{x}_{l+1}$.

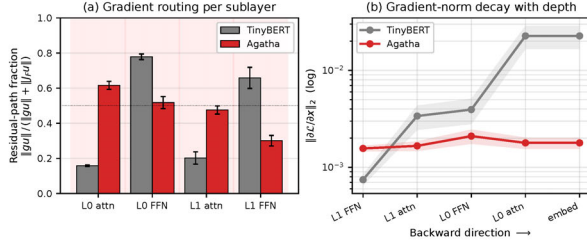


Figure 4: Gradient flow analysis on MRPC (averaged across three seeds and the full validation set). (a) Residual-path fraction per sublayer. (b) Gradient norm (log scale) from loss to embedding.

TinyBERT exhibits gradient amplification: the norm grows $30\times$ from the deepest sublayer to the embedding (0.023). Agatha maintains a flat profile near 0.002, yielding an embedding gradient $12.7\times$ smaller. This is consistent with Theorem 1: contraction compounds multiplicatively as $\prod_l (g + \sigma_{F_l})$.

5.14 Computational Cost

Table 11 compares the training and inference costs of Agatha against the baselines.

Model	Train FLOPs /sample	Infer. (ex/s)	ASR $\varepsilon = 2.0$
FreeLB	532M	15,660	66.68
DiffuseDef*	216M	13,422	69.85
DynAttn	177M	7,475	63.58
Agatha (max. iter $T_{\max}=3$)	735M	10,021	52.69
Agatha (max. iter $T_{\max}=100$)	735M	8,635	50.28

Table 11: Computational cost and robustness. Inference throughput measured on a single RTX 3090 GPU; ASR averaged across 8 GLUE tasks. *DiffuseDef adds a denoiser module (66K extra parameters, $\sim 1.5\%$); all other models share the same parameter count.

Training Cost. Agatha trains with standard cross-entropy and requires no adversarial inner loop, that is, no attack generation during training, but it is not cheaper to train. Its per-sample training FLOPs (735M) exceed both the architectural baselines (177M to 216M) and FreeLB (532M), so the training advantage is one of threat-model generality rather than reduced compute. FreeLB is tied to the perturbation distribution it trains against and requires a per-task budget grid search (Table 16), while Agatha generalizes across norms, attacks, and budgets without retraining.

Inference Cost. DynAttn pays the highest inference cost among baselines (7,475 ex/s) due to noise samples per forward pass, yet achieves only 63.58 ASR. Agatha at $T_{\max} = 3$ is $1.34\times$ faster and 11 points more robust. Increasing to $T_{\max} = 100$ (default), yields 2.4 further points, while still being

$1.2\times$ faster. Pareto-dominating every baseline on both axes without adding any parameters.

6 Related Work

Adversarial defenses in NLP span training-based methods (Zhu et al., 2019; Madry et al., 2017) and architectural approaches such as DiffuseDef (Li et al., 2025) and DynAttn (Shen et al., 2023), which add denoising or noise-injection modules. Both categories incur computational overhead and, as our results show, provide limited benefit at small model scales. Modern Hopfield Networks (Ram-sauer et al., 2020) have been studied for memory capacity and interpretability but not for robustness. Agatha bridges these lines. See Appendix M for comprehensive discussion of related works.

7 Conclusions

We introduced Agatha, an architectural modification to the Transformer that achieves adversarial robustness without any threat-model-specific training by replacing softmax attention with an iterative Hopfield mechanism and introducing a gated residual that forces adversarial gradients through the energy landscape, where it is contracted. We formalized this robustness through two theorems, validated via hyperparameter sweeps. Across 8 GLUE and 4 SuperGLUE tasks, Agatha reduces attack success rates by over 13 percentage points against the best architectural baseline at perturbation budget $\varepsilon = 2.0$, widening to 19 points at $\varepsilon = 3.0$, with robustness holding under AutoAttack, ℓ_∞ PGD, FGSM, Gaussian noise, discrete text-level attacks, and Hopfield-specific attacks. Agatha Pareto-dominates all baselines with 11 points lower ASR than DynAttn at $1.2\times$ faster inference and no additional parameters. These results suggest that architectural robustness through energy basin geometry is an effective alternative to adversarial training, with the advantage growing rather than diminishing with scale (from a 13-point gap at TinyBERT to a 33-point gap at BERT-base).

Acknowledgments

We thank the anonymous reviewers and the area chair for their constructive feedback.

This material is based upon work supported by the U.S. Department of Energy under Award No. DE-SC0022014 and the National Science Foundation under Grant No. CCF-1942697.

Limitations

Agatha is evaluated on English-language benchmarks; behavior on multilingual inputs has not been explored. We show that Agatha can be used as a drop-in replacement for the attention mechanism in TinyBERT, BERT-base, and the Pythia decoder backbones. We have not evaluated other BERT variants. RTE (276 examples) and WNLI (71 examples) have small validation sets where variance arises from validation-set size rather than seed sensitivity; Appendix B.1.4 reports a GLUE-6 analysis excluding these tasks, preserving the robustness gap. Comparing against the full set of adversarial defenses in NLP is a considerable undertaking, so we compare Agatha against representative and state-of-the-art techniques: FreeLB (Zhu et al., 2019) for training-based defenses, and DiffuseDef (Li et al., 2025) and Dynamic Attention (Shen et al., 2023) for architectural defenses. Agatha requires no adversarial inner loop, but its per-sample training cost is higher than FreeLB (Table 11), so the training advantage is one of threat-model generality rather than reduced training compute. Our evaluation focuses on classification tasks, where attack success rate is well-defined as the flip rate. We do not evaluate open-ended generation, where attractor contraction competes with preserving a fine-grained predictive distribution; we leave generation robustness to future work.

Due to space limitations, the main paper reports combined GLUE results for each evaluation; comprehensive per-task breakdowns, detailed ablation analysis, gate selection, and full hyperparameter sweep results are deferred to the appendix.

Ethical Considerations

This work proposes an architectural defense against adversarial attacks on NLP models. Improved robustness has direct societal benefits, reducing the risk of adversarial manipulation in safety-critical applications such as content moderation, financial analysis, and automated decision-making. We acknowledge that robustness research has dual-use potential: the attack evaluations in this paper could in principle be adapted to target deployed systems. We believe the defensive contribution outweighs this risk, as the attack methods we employ (PGD, FGSM) are well-established in the literature. All experiments use publicly available benchmark datasets (GLUE, SuperGLUE, AdvGLUE) that do not contain sensitive or personally identifiable data.

Licensing. The pretrained checkpoints used in this work are bert-base-uncased, google/bert_uncased_L-2_H-128_A-2, and google/bert_uncased_L-4_H-256_A-4 (all Apache-2.0). AdvGLUE is distributed under CC BY-SA 4.0. GLUE and SuperGLUE are collections of datasets, each governed by the original license of its constituent tasks; the benchmark authors note that these licenses permit use and redistribution in a research context. All artifacts are used for research evaluation consistent with their intended purpose.

AI Assistance

Anthropic Claude Opus 4.6 was used for improving the clarity and readability of the writing. The authors take full responsibility for the content.

References

- Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2890–2896.
- Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. 2020. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pages 484–501. Springer.
- Anish Athalye, Nicholas Carlini, and David Wagner. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, pages 274–283. PMLR.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pages 2397–2430. PMLR.
- Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. 2022. Bad characters: Imperceptible nlp attacks. In *IEEE symposium on security and privacy (SP)*, pages 1987–2004. IEEE.
- Kai-Wei Chang, He He, Robin Jia, and Sameer Singh. 2021. Robustness and adversarial examples in natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): Tutorial Abstracts*, pages 22–26.

- Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, pages 1310–1320. PMLR.
- Francesco Croce and Matthias Hein. 2020a. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning (ICML)*, pages 2196–2205. PMLR.
- Francesco Croce and Matthias Hein. 2020b. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, pages 2206–2216. PMLR.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Ahmed Farooq. 2025. A framework for non-linear attention via modern hopfield networks. *arXiv preprint arXiv:2506.11043*.
- Taniya Gidatkar, Oluwaseun Ajao, and Matthew Shardlow. 2025. Differential robustness in transformer language models: Empirical evaluation under adversarial text attacks. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pages 395–402.
- Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*. ICLR 2015.
- Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed Zaki, and Dmitry Krotov. 2023. Energy transformer. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:27532–27559.
- Jerry Yao-Chieh Hu, Pei-Hsuan Chang, Robin Luo, Hong-Yu Chen, Weijian Li, Wei-Po Wang, and Han Liu. 2024a. Outlier-efficient hopfield layers for large transformer-based models. *arXiv preprint arXiv:2404.03828*. ICML 2024.
- Jerry Yao-Chieh Hu, Bo-Yu Chen, Dennis Wu, Feng Ruan, and Han Liu. 2024b. Nonparametric modern hopfield models. *arXiv preprint arXiv:2404.03900*. ICML 2025.
- Jerry Yao-Chieh Hu, Donglin Yang, Dennis Wu, Chenwei Xu, Bo-Yu Chen, and Han Liu. 2023. On sparse modern hopfield model. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:27594–27608.
- Yunzhe Hu, Difan Zou, and Dong Xu. 2025. Hyper-set: Designing transformers via hyperspherical energy minimization. *arXiv preprint arXiv:2502.11646*.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 4163–4174.
- Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. 2020. Perceptual adversarial robustness: Defense against unseen threat models. *arXiv preprint arXiv:2006.12655*. ICLR 2021.
- Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*. NDSS 2019.
- Zhenhao Li, Huichi Zhou, Marek Rei, and Lucia Specia. 2025. Diffusedef: Improved robustness to adversarial attacks via iterative denoising. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9259–9274.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*. ICLR 2018.
- Tsubasa Masumura and Masato Taki. 2026. On the role of hidden states of modern hopfield network in transformer. *Advances in Neural Information Processing Systems*, 38:68607–68626.
- Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*. ICLR 2017.
- Han Cheol Moon, Shafiq Joty, Ruochen Zhao, Megh Thakkar, and Chi Xu. 2023. Randomized smoothing with masked inference for adversarially robust text classifications. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5145–5165.
- John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.
- Andreia Podasca and Anup Das. 2025. Analyzing the robustness of neuromorphic computing in the presence of variability in non-volatile memory. In *2025 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, pages 1–8. IEEE.
- Andreia Podasca and Anup Das. 2026. Robertha: Eigen-spectrum regularized attention for robust natural language understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15229–15264.

- Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591.
- Brian Pulfer, Yury Belousov, and Slava Voloshynovskiy. 2024. Robustness tokens: Towards adversarial robustness of transformers. In *European Conference on Computer Vision*, pages 110–127. Springer.
- Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. 2020. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*. ICML 2020.
- Hubert Ramsauer, Bernhard Schödl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. 2020. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*. ICLR 2021.
- Saul Santos, Vlad Niculae, Daniel McNamee, and Andre FT Martins. 2024. Sparse and structured hopfield networks. *arXiv preprint arXiv:2402.13725*. ICML 2024.
- Saleh Sargolzaei and Luis Rueda. 2024. Improving out-of-distribution data handling and corruption resistance via modern hopfield networks. In *International Conference on Pattern Recognition*, pages 81–96. Springer.
- Lujia Shen, Yuwen Pu, Shouling Ji, Changjiang Li, Xuhong Zhang, Chunpeng Ge, and Ting Wang. 2023. Improving the robustness of transformer-based large language models with dynamic attention. *arXiv preprint arXiv:2311.17400*. NDSS 2024.
- Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. 2020. Robustness verification for transformers. *arXiv preprint arXiv:2002.06622*. ICLR 2020.
- Matthew Smart, Alberto Bietti, and Anirvan M Sengupta. 2025. In-context denoising with one-layer transformers: connections between attention and associative memory retrieval. *arXiv preprint arXiv:2502.05164*. ICML 2025.
- Florian Tramèr and Dan Boneh. 2019. Adversarial training and robustness for multiple perturbations. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.
- Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1633–1645.
- Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2018. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*. ICLR 2019.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355.
- Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*. NeurIPS 2021.
- Jiawei Wang, Yushen Zuo, Yuanjun Chai, Zhendong Liu, Yicheng Fu, Yichun Feng, and Kin-Man Lam. 2025. Safeguarding vision-language models: Mitigating vulnerabilities to gaussian noise in perturbation-based attacks. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2773–2782. IEEE.
- Dennis Wu, Jerry Yao-Chieh Hu, Teng-Yun Hsiao, and Han Liu. 2024. Uniform memory retrieval with larger capacity for modern hopfield models. *arXiv preprint arXiv:2404.03827*. ICML 2024.
- Sophie Xhonneux, Alessandro Sordani, Stephan Günnemann, Gauthier Gidel, and Leo Schwinn. 2024. Efficient adversarial training in llms with continuous attacks. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:1502–1530.
- Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. On layer normalization in the transformer architecture. In *International Conference on Machine Learning (ICML)*, pages 10524–10533. PMLR.
- Mao Ye, Chengyue Gong, and Qiang Liu. 2020. Safer: A structure-free approach for certified robustness to adversarial word substitutions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3465–3475.

- Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, pages 7472–7482. PMLR.
- Xinyu Zhang, Hanbin Hong, Yuan Hong, Peng Huang, Binghui Wang, Zhongjie Ba, and Kui Ren. 2024. Text-crs: A generalized certified robustness framework against textual adversarial attacks. In *IEEE Symposium on Security and Privacy (SP)*, pages 2920–2938. IEEE.
- Haiteng Zhao, Chang Ma, Xinshuai Dong, Anh Tuan Luu, Zhi-Hong Deng, and Hanwang Zhang. 2022. Certified robustness against natural language attacks by causal intervention. In *International Conference on Machine Learning (ICML)*, pages 26958–26970. PMLR.
- Jiahao Zhao and Wenji Mao. 2023. Generative adversarial training with perturbed token detection for model robustness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 13012–13025.
- Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. FreeLB: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*. ICLR 2020.

Appendix. The appendix is organized as follows.

1. We provide details on the **evaluation methodology** in Appendix A.
2. We provide a **thorough empirical validation of Agatha** as follows.
 - Detailed analysis of Agatha’s PGD robustness in Appendix B.
 - Standard PGD (ℓ_2) GLUE in Appendix B.1.
 - Standard PGD (ℓ_2) SuperGLUE in Appendix B.2.
 - PGD (ℓ_2) with increased steps in Appendix B.3.
 - Transfer attack in Appendix B.4.
 - Standard PGD (ℓ_∞) in Appendix B.5.
 - Auto attack (ℓ_2) in Appendix B.6.
 - APGD-CE (ℓ_∞) in Appendix B.7.
 - Beyond PGD attack in Appendix C.
 - Hopfield-specific attack in Appendix D.
 - Mechanistic interpretation of performance gap in Appendix E.
 - Full gate ablation in Appendix F.
 - Fixed gate vs task-specific gate results in Appendix F.1.
 - Symmetric vs asymmetric gate results in Appendix F.2.
 - Fixed vs learned gate results in Appendix F.3.
 - Architectural ablation results in Appendix G.
 - BERT-scale results in Appendix H.
 - Hyperparameter sweep in Appendix I.
 - Convergence analysis in Appendix J.
 - Training dynamics in Appendix K.
3. We provide proofs of two theorems that guarantee robustness of Agatha in Appendix L.
4. We present a comprehensive overview of related works in Appendix M.

A Evaluation Methodology

A.1 Datasets

We evaluate on 8 tasks from the GLUE benchmark (Wang et al., 2018) and 4 tasks from the SuperGLUE benchmark (Wang et al., 2019). Table 12 reports the size of each split.

Benchmark	Task	Train	Val	Test	Metric
GLUE	SST-2	67,349	872	1,821	Accuracy
	CoLA	8,551	1,043	1,063	Matthews Corr.
	MRPC	3,668	408	1,725	F1 / Accuracy
	QQP	363,846	40,430	390,965	F1 / Accuracy
	MNLI	392,702	9,815	9,796	Accuracy
	QNLI	104,743	5,463	5,463	Accuracy
	RTE	2,490	276	3,000	Accuracy
	WNLI	635	71	146	Accuracy
SuperGLUE	BoolQ	9,427	3,270	3,245	Accuracy
	CB	250	56	250	F1 / Accuracy
	MultiRC	27,243	4,848	9,693	F1a / EM
	WSC	554	104	146	Accuracy

Table 12: Dataset statistics for the 8 GLUE and 4 SuperGLUE tasks evaluated in this work. Train/Val/Test counts are example counts. For MNLI, validation and test counts are the matched-set sizes; we use only the matched split. All evaluations are performed on the validation split, since test labels are not publicly available.

Metrics. To address metric heterogeneity across GLUE tasks, we normalize all clean task metrics to a 0 to 100 scale: Matthews Correlation as $\frac{MCC+1}{2} \times 100$, F1/Accuracy pairs as their arithmetic mean, and classification tasks by accuracy. We report Attack Success Rate (ASR) for PGD and FGSM, defined as the percentage of initially correct predictions flipped by the adversarial perturbation. For Gaussian noise, CharSwap, and TextBugger, we report Flip Rate using the same definition. All metrics are computed per-seed and reported as mean $\pm$ standard deviation across three random seeds. For the combined GLUE score, we average per-task means and propagate standard deviations as $\frac{1}{T} \sqrt{\sum_t \sigma_t^2}$. We exclude STS-B from our evaluation as it is a regression task; Attack Success Rate, which measures flipped classification predictions, is not applicable to continuous similarity scores.

A.2 Model Architectures

Agatha and all baselines share the same backbone parameter count for each architecture; Agatha adds no parameters relative to TinyBERT. Table 13 reports the architectures used in this paper.

Denoyer overhead scales with backbone. DiffuseDef’s denoyer module is the only baseline component that adds parameters beyond the shared backbone. Its parameter count grows with hidden dimension: 66,304 at TinyBERT base (128d), 263,680 at the scaled architecture (256d), and 2,402,304 at BERT-base (768d). The denoyer is applied at every inference step, with three denoise

Architecture	Pretrained Checkpoint	Params	Hidden	Layers	Heads	Extra (DiffuseDef)
TinyBERT (base)	google/bert_uncased_L-2_H-128_A-2	4.37M	128	2	2	+66,304
TinyBERT (scaled)	google/bert_uncased_L-4_H-256_A-4	11.1M	256	4	4	+263,680
BERT-base	bert-base-uncased	110M	768	12	12	+2,402,304

Table 13: Model architectures used in this work. Agatha, TinyBERT, FreeLB, and DynAttn share the same backbone parameter count for each architecture; Agatha adds no parameters at any scale. DiffuseDef adds a denoiser module that scales with hidden dimension. The TinyBERT (base) architecture is used for the eight GLUE tasks and for BoolQ and CB; the TinyBERT (scaled) architecture is used for MultiRC and WSC; BERT-base is used for the scaling experiment in Section 5.9 and Appendix H.

iterations per forward pass, so the inference compute overhead grows with the denoiser size. At deployment-relevant scales (BERT-large, 1024d; LLaMA-class, 4096d+), this overhead projects to tens of millions of parameters and a proportionally larger inference cost. Agatha, by contrast, adds no parameters and no module-level forward passes; its only overhead is the iterative Hopfield update on the existing key matrix, which is configurable. Setting $T_{\max} = 3$ iterations is sufficient for deployment (Section 5.14 and Appendix J).

A.3 System Configuration

All experiments are conducted on a Lambda workstation with two RTX 3090 GPUs. Table 14 shows the hardware and software configuration used.

Component	Specification
<i>Hardware Configuration</i>	
GPU	NVIDIA GeForce RTX 3090 (24GB VRAM)
CPU	AMD Ryzen Threadripper 3960X 24-Core @ 3.80GHz 24 cores, 48 threads
RAM	128GB DDR4-3200
<i>Software Environment</i>	
CUDA	Version 12.4
PyTorch	Version 2.4.1 (built with CUDA 12.4)
Operating System	Ubuntu 20.04.6 LTS (kernel 5.15.0-139)

Table 14: Hardware and software configuration. All models were trained and evaluated on a single RTX 3090 GPU with no multi-GPU parallelism.

A.4 Training Configuration

Table 15 reports training hyperparameters shared across all models. All models use the same optimizer, learning rate schedule, and tokenizer to isolate the effect of architectural differences.

A.5 Baseline Hyperparameters

Table 16 reports the method-specific hyperparameters for each baseline. All settings were selected

Parameter	Value
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0.01
LR schedule	Linear warmup + decay
Warmup ratio	0.1
Gradient clipping	Max norm 1.0
Epochs	30
Early stopping patience	20
Batch size (train)	32
Batch size (eval)	64
Tokenizer	bert-base-uncased
Pretrained checkpoint	google/bert_uncased_L-2_H-128_A-2
Max sequence length	128
Dropout	0.3
Loss function	CrossEntropyLoss

Table 15: Training hyperparameters shared across all models and tasks in this paper.

via grid search on validation performance; where the original paper recommended a default, we used it as the starting point.

A.6 Agatha Hyperparameters

Agatha introduces three hyperparameters beyond the shared training configuration: the inverse temperature β , the convergence threshold τ , and the gate value g . At inference, the Hopfield mechanism iterates until the residual falls below τ or the iteration count reaches $T_{\max}$.

The inverse temperature is fixed at $\beta = 50$, selected via the sweep in Section 5.11 as the value that reliably places all tasks in the deep attractor regime (Theorem 2). The convergence threshold is fixed at $\tau = 10^{-4}$, selected as the smallest value that yields stable results across all tasks and seeds.

The maximum Hopfield iteration count is set to $T_{\max} = 100$, a conservative upper bound; the iterative mechanism exits early when the residual falls below τ , so the effective iteration count is task- and input-dependent (average 27.5 iterations across tasks; Appendix J). For practical deploy-

Method	Parameter	Value
FreeLB	Ascent steps (K)	3
	Step size (α)	$0.3 \cdot \varepsilon$
	Training ε	Per-task (Table 17)
	Norm	ℓ_2
DiffuseDef	Denoise steps	3
	Noise scale	0.1
	Denoiser	MLP
DynAttn	Attention noise σ	1.0
	Noise samples (train & inference)	5
PGD-AT	Inner PGD steps	10
	Step size (α)	$0.3 \cdot \varepsilon$
	Training ε	Per-task (Table 17)
	Norm	ℓ_2

Table 16: Method-specific hyperparameters for each baseline. All baselines share the optimizer and training settings in Table 15. PGD-AT is used only in the QNLI diagnosis (Appendix E).

Task	PGD-AT ε	FreeLB ε
SST-2	0.5	1.0
CoLA	1.0	2.0
MNLI	0.1	1.0
QNLI	0.1	0.5
QQP	0.5	1.0
MRPC	1.0	2.0
RTE	0.5	2.0
WNLI	0.1	0.1

Table 17: Per-task training perturbation budgets for FreeLB and PGD-AT, selected via grid search over $\varepsilon \in \{0.1, 0.5, 1.0, 2.0\}$ on validation.

ment, $T_{\max} = 3$ is sufficient: it already outperforms all baselines while maintaining $1.3\times$ faster inference than DynAttn (Section 5.14).

The gate value g is the only task-specific hyperparameter, selected via grid search over $g \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ on validation performance. Table 18 reports the selected values.

Gate value	Tasks
$g = 0.1$	SST-2, CoLA, QQP, QNLI, MultiRC
$g = 0.5$	MNLI
$g = 0.7$	MRPC, RTE, BoolQ, CB, WSC
$g = 0.9$	WNLI

Table 18: Agatha gate value per task. $\beta = 50$, $\tau = 10^{-4}$, and $T_{\max} = 100$ are fixed for all tasks.

Lower g contracts the residual path more aggressively, forcing stronger contraction through the Hopfield mechanism at the cost of representational capacity. Tasks that tolerate this trade-off benefit

from aggressive gating ($g = 0.1$), while tasks where clean accuracy is more sensitive to capacity loss prefer a moderate gate ($g = 0.7$). MNLI uses an intermediate value ($g = 0.5$), and WNLI benefits from a high gate ($g = 0.9$) due to its small validation set where aggressive gating increases variance. For deployment without per-task tuning, a fixed $g = 0.7$ outperforms all baselines (Section 5.12).

A.7 Attack Configuration

Attack Differentiation. Every gradient-based attack in this paper backpropagates through the fully unrolled Hopfield iteration. Gradients are obtained by standard autodiff through each executed update step $\xi^{(t+1)} = \mathbf{K}^\top \text{softmax}(\frac{\beta}{\sqrt{d}} \mathbf{K} \xi^{(t)})$, up to the convergence exit of Eq. 5 or the cap $T_{\max}$, so the depth of the backward graph matches the number of iterations actually executed for that input. No gradient-stopping operation appears inside the forward pass: there is no truncated backpropagation through the iteration, no detached fixed point, and no implicit-function approximation of the converged state. The attacker therefore differentiates the exact inference-time computation graph, including its early-exit behavior, and the reduction in gradient magnitude we measure is a property of the geometry established by Theorems 1 and 2 rather than of the differentiation procedure. Outside the forward pass, the attacks follow the standard iterative protocol in which the perturbation is detached between successive attack steps and the projection onto the ε -ball is computed without gradient tracking. This is the usual construction for iterative attacks and does not truncate the backward pass through the model.

Cost of Full Unrolling. Backpropagation cost grows linearly with the number of executed Hopfield iterations, which averages 27.5 per input across GLUE (Appendix J). All attacks are run on a single RTX 3090 (Appendix A.3). Three results in this paper are consistent with the unrolled gradient being informative rather than obscured. Increasing PGD from 5 to 200 steps changes Agatha’s ASR by only 0.54 points, the smallest change among all evaluated models (Appendix B.3). Boundary-seeking FAB, which is designed to expose obfuscated gradients, is Agatha’s worst case among the white-box attacks yet still leaves it lowest-ASR at every budget (Appendix B.6). Transfer and theory-aware attacks both underperform white-box PGD (Sections 5.5 and 5.7).

PGD (Evaluation). We use 20-step PGD with step size $\alpha = 0.3 \cdot \varepsilon$ under an ℓ_2 norm constraint, initialized from the clean embedding plus small random noise ($0.1 \cdot \varepsilon$). We evaluate at perturbation budgets $\varepsilon \in \{0.05, 0.5, 1.0, 2.0, 3.0, 5.0\}$.

FGSM. Single-step gradient attack using the same ℓ_2 norm and budget range as PGD.

Gaussian Noise. Isotropic Gaussian perturbations $\delta \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ added directly to the input embedding with a fixed random seed for reproducibility (Wang et al., 2025). We evaluate at $\sigma \in \{0.5, 1.0, 2.0, 5.0\}$.

CharSwap. Adjacent character swaps applied to a randomly selected fraction of words, following Pruthi et al. (2019). We evaluate at perturbation rates $\{10\%, 20\%, 30\%, 50\%\}$.

TextBugger. Character-level perturbations (insertion, deletion, swap, homoglyph substitution) applied to a fraction of words, following Li et al. (2018). We evaluate at perturbation rates $\{10\%, 20\%, 30\%, 50\%\}$.

A.8 Computational Cost Instrumentation

All cost measurements use the same instrumentation pipeline across every model. Forward FLOPs per sample are measured using fvcare’s FlopCountAnalysis on a single validation sample traced on CPU for reproducibility. Training FLOPs are estimated as $3 \times$ forward FLOPs (forward pass, backward pass, and parameter update). Inference throughput (examples per second) is measured end-to-end over the full validation set with batch size 64, with `torch.cuda.synchronize()` calls bracketing the loop to ensure accurate GPU timing. Per-batch latency is measured separately via CUDA events over 50 timed passes after 10 warmup passes on a fixed batch, reporting median and 95th percentile latency. All measurements are conducted on a single RTX 3090 GPU.

B Empirical Validation of Agatha’s Adversarial Robustness

B.1 GLUE Performance under Standard PGD (ℓ_2)

We analyze Agatha under the primary attack protocol used in the main paper: 20-step PGD with step size $\alpha = 0.3\varepsilon$ under an ℓ_2 norm constraint. This subsection extends the combined GLUE results in Table 1 with per-task breakdowns across all 8 tasks, organized by metric type. All results are reported as mean $\pm$ std across three random seeds.

B.1.1 Combined GLUE Discussion

Table 1 summarizes combined GLUE performance under standard PGD. We provide an extended discussion organized by perturbation budget.

Clean Performance. Agatha achieves a clean score of 63.91, within one point of the TinyBERT baseline (64.28) using standard cross-entropy loss. FreeLB achieves the highest clean score (67.19) through adversarial training, with DiffuseDef (66.22) and DynAttn (65.77) following. None of these advantages translate into robustness.

Low Perturbation Budget ($\varepsilon = 0.05$). All models exhibit low attack success rates (1.48 to 2.55), as perturbations at this scale are insufficient to cross decision boundaries. Agatha records the lowest ASR (1.48) by a small margin.

Medium Perturbation Budgets ($\varepsilon \in \{0.5, 1.0, 2.0\}$). Agatha consistently achieves the lowest ASR at every medium budget. At $\varepsilon = 0.5$, Agatha records only 12.75, compared to 16.68 for DynAttn, 18.95 for DiffuseDef, and 18.99 for FreeLB. At $\varepsilon = 1.0$, Agatha records 26.16 versus 33.37 for DynAttn, 35.50 for FreeLB, and 38.07 for DiffuseDef. At $\varepsilon = 2.0$, Agatha further reduces ASR to 50.28, a 13.30 point reduction over the best architectural baseline (DynAttn, 63.58), and a 16.40 point reduction over the training-based FreeLB.

High Perturbation Budgets ($\varepsilon \in \{3.0, 5.0\}$). The advantage of Agatha’s robustness grows steadily at high budgets. At $\varepsilon = 3.0$, the gap over DynAttn widens to 19.00 points (65.58 versus 84.58). At $\varepsilon = 5.0$, Agatha achieves 82.03 compared to 98.12 for FreeLB, 99.14 for DynAttn, and 99.57 for DiffuseDef. This pattern is consistent with our analysis: the energy basin structure provides a geometric constraint that operates independently of the train-

ing perturbation distribution and persists at high budgets where other defenses degrade.

B.1.2 Task-by-Task Breakdown

Classification Tasks (Accuracy Metric). Table 19 shows results for GLUE tasks using accuracy as the primary metric: SST-2 (sentiment analysis), MNLI (natural language inference), QNLI (question-answering NLI), RTE (recognizing textual entailment), and WNLI (Winograd NLI).

Linguistic Acceptability (Matthews Correlation). Table 20 shows CoLA results using the Matthews Correlation Coefficient (MCC), normalized to a 0 to 100 scale as $\frac{\text{MCC}+1}{2} \times 100$.

Paraphrase Detection (F1 + Accuracy Average). Table 21 shows MRPC and QQP results, using the average of F1 and accuracy as the clean score.

B.1.3 Task-Specific Observations

SST-2 (Sentiment Analysis). Agatha closely matches TinyBERT in clean accuracy (80.20 versus 81.77) while still providing the lowest ASR from $\varepsilon = 0.5$ onward. At $\varepsilon = 2.0$, Agatha reduces ASR to 45.99, an 8 point improvement over DynAttn (54.26), 60.31 for DiffuseDef, and 72.86 for FreeLB. DiffuseDef and DynAttn both achieve the lowest ASR at $\varepsilon = 0.05$ (0.89), though the differences at this small budget are within noise.

MNLI (Natural Language Inference). DynAttn achieves the highest clean accuracy (64.25) and leads at low budgets, recording 17.13 at $\varepsilon = 0.5$ and 34.38 at $\varepsilon = 1.0$. Agatha leads from $\varepsilon = 2.0$ onward, recording 62.55 versus 64.84 for DynAttn, 73.78 for DiffuseDef, and 77.49 for FreeLB. The gap widens at higher budgets (93.08 at $\varepsilon = 5.0$ versus 99.02 for DynAttn and 98.74 for DiffuseDef).

QNLI (Question-Answering NLI). FreeLB achieves the highest clean accuracy (80.36) and the lowest ASR at low budgets (1.16 at $\varepsilon = 0.05$, 13.02 at $\varepsilon = 0.5$). Agatha then leads from $\varepsilon = 1.0$ onward, recording 57.51 at $\varepsilon = 2.0$ compared to 62.50 for DynAttn, 68.79 for FreeLB, and 72.32 for DiffuseDef. The clean accuracy gap between Agatha (60.82) and FreeLB (80.36) likely reflects a task-level representation bottleneck at this model scale; we provide a full diagnosis in Appendix E.

RTE (Recognizing Textual Entailment). All models cluster in the 54 to 60 range for clean accuracy, reflecting the inherent difficulty of this task at small model scales. Agatha achieves the lowest

Task	Model	Clean ($\uparrow$)	Attack Success Rate ($\%, \downarrow$)					
			$\epsilon = 0.05$	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 2.0$	$\epsilon = 3.0$	$\epsilon = 5.0$
SST-2	TinyBERT	81.77 $\pm$ 0.19	1.36 $\pm$ 0.52	13.28 $\pm$ 0.45	28.33 $\pm$ 1.01	60.17 $\pm$ 1.36	85.32 $\pm$ 0.66	99.95 $\pm$ 0.07
	FreeLB	77.60 $\pm$ 0.11	1.67 $\pm$ 0.30	16.85 $\pm$ 0.65	35.67 $\pm$ 1.43	72.86 $\pm$ 1.56	94.53 $\pm$ 0.75	100.00 $\pm$ 0.00
	DiffuseDef	81.77 $\pm$ 0.30	0.89 $\pm$ 0.29	13.18 $\pm$ 0.10	28.98 $\pm$ 1.19	60.31 $\pm$ 1.08	85.04 $\pm$ 1.61	99.91 $\pm$ 0.08
	DynAttn	81.31 $\pm$ 0.11	0.89 $\pm$ 0.49	10.69 $\pm$ 0.48	25.68 $\pm$ 1.06	54.26 $\pm$ 0.52	80.37 $\pm$ 0.51	99.86 $\pm$ 0.14
	Agatha	80.20 $\pm$ 0.18	0.95 $\pm$ 0.08	10.53 $\pm$ 1.78	22.49 $\pm$ 2.97	45.99 $\pm$ 3.41	67.53 $\pm$ 6.06	90.61 $\pm$ 3.59
MNLI	TinyBERT	53.57 $\pm$ 0.44	2.37 $\pm$ 0.20	24.85 $\pm$ 0.41	49.86 $\pm$ 0.72	83.03 $\pm$ 0.26	95.94 $\pm$ 0.49	99.99 $\pm$ 0.01
	FreeLB	57.31 $\pm$ 1.23	2.03 $\pm$ 0.22	21.38 $\pm$ 0.69	42.64 $\pm$ 0.32	77.49 $\pm$ 0.84	94.97 $\pm$ 0.31	99.98 $\pm$ 0.01
	DiffuseDef	62.73 $\pm$ 1.37	1.87 $\pm$ 0.05	20.59 $\pm$ 0.51	40.80 $\pm$ 0.18	73.78 $\pm$ 2.57	89.42 $\pm$ 3.30	98.74 $\pm$ 1.29
	DynAttn	64.25 $\pm$ 0.59	4.41 $\pm$ 0.15	17.13 $\pm$ 0.52	34.38 $\pm$ 0.70	64.84 $\pm$ 1.26	85.87 $\pm$ 0.63	99.02 $\pm$ 0.35
	Agatha	54.64 $\pm$ 0.13	2.14 $\pm$ 0.53	19.03 $\pm$ 0.84	35.70 $\pm$ 0.70	62.55 $\pm$ 0.54	78.76 $\pm$ 0.78	93.08 $\pm$ 0.81
QNLI	TinyBERT	65.10 $\pm$ 0.61	1.87 $\pm$ 0.13	19.32 $\pm$ 1.22	38.63 $\pm$ 2.39	69.54 $\pm$ 4.35	88.62 $\pm$ 2.31	99.28 $\pm$ 0.10
	FreeLB	80.36 $\pm$ 0.69	1.16 $\pm$ 0.15	13.02 $\pm$ 1.66	30.36 $\pm$ 3.61	68.79 $\pm$ 6.98	92.85 $\pm$ 3.77	99.91 $\pm$ 0.04
	DiffuseDef	64.92 $\pm$ 2.04	1.95 $\pm$ 0.14	21.23 $\pm$ 2.92	41.60 $\pm$ 4.85	72.32 $\pm$ 3.22	89.99 $\pm$ 1.38	99.20 $\pm$ 0.29
	DynAttn	65.49 $\pm$ 0.60	3.42 $\pm$ 0.18	17.58 $\pm$ 3.49	34.49 $\pm$ 7.07	62.50 $\pm$ 12.47	81.44 $\pm$ 9.50	97.77 $\pm$ 1.50
	Agatha	60.82 $\pm$ 0.48	1.70 $\pm$ 0.23	15.29 $\pm$ 0.89	30.02 $\pm$ 0.97	57.51 $\pm$ 0.12	77.13 $\pm$ 0.81	95.01 $\pm$ 0.09
RTE	TinyBERT	57.64 $\pm$ 0.45	4.38 $\pm$ 0.86	30.06 $\pm$ 1.22	54.10 $\pm$ 4.31	88.95 $\pm$ 1.48	99.16 $\pm$ 0.78	100.00 $\pm$ 0.00
	FreeLB	60.41 $\pm$ 0.61	2.60 $\pm$ 1.02	21.71 $\pm$ 2.13	37.43 $\pm$ 3.39	70.51 $\pm$ 1.28	85.25 $\pm$ 0.71	99.60 $\pm$ 0.57
	DiffuseDef	57.76 $\pm$ 1.25	3.13 $\pm$ 0.69	26.45 $\pm$ 2.27	52.93 $\pm$ 3.19	89.38 $\pm$ 2.18	99.18 $\pm$ 1.43	100.00 $\pm$ 0.00
	DynAttn	57.04 $\pm$ 1.25	3.58 $\pm$ 0.97	27.94 $\pm$ 4.55	53.20 $\pm$ 4.53	88.56 $\pm$ 7.03	99.57 $\pm$ 0.37	100.00 $\pm$ 0.00
	Agatha	54.99 $\pm$ 0.91	2.62 $\pm$ 1.11	19.64 $\pm$ 9.30	35.09 $\pm$ 18.66	58.11 $\pm$ 21.57	70.21 $\pm$ 20.92	86.47 $\pm$ 10.23
WNLI	TinyBERT	53.05 $\pm$ 3.51	5.33 $\pm$ 0.35	44.57 $\pm$ 5.15	66.62 $\pm$ 5.44	93.27 $\pm$ 5.59	98.23 $\pm$ 1.26	100.00 $\pm$ 0.00
	FreeLB	53.99 $\pm$ 1.76	6.12 $\pm$ 1.41	41.96 $\pm$ 6.88	66.28 $\pm$ 6.77	93.22 $\pm$ 5.81	98.27 $\pm$ 1.23	100.00 $\pm$ 0.00
	DiffuseDef	57.75 $\pm$ 2.82	1.63 $\pm$ 2.82	24.57 $\pm$ 19.90	44.42 $\pm$ 35.31	68.79 $\pm$ 35.12	92.35 $\pm$ 11.21	100.00 $\pm$ 0.00
	DynAttn	54.46 $\pm$ 4.30	4.18 $\pm$ 2.56	34.99 $\pm$ 4.34	62.95 $\pm$ 1.87	93.22 $\pm$ 6.80	98.16 $\pm$ 1.60	100.00 $\pm$ 0.00
	Agatha	56.81 $\pm$ 0.81	1.65 $\pm$ 1.43	12.36 $\pm$ 4.08	39.57 $\pm$ 10.16	85.87 $\pm$ 8.11	95.87 $\pm$ 3.81	100.00 $\pm$ 0.00

Table 19: GLUE classification tasks under standard PGD (ℓ_2): Clean accuracy and Attack Success Rate (ASR). RTE (276 examples) and WNLI (71 examples) have small validation sets; results should be interpreted cautiously. **Bold** indicates best performance: highest clean accuracy and lowest ASR per task and budget.

Model	Clean ($\uparrow$)	Attack Success Rate ($\%, \downarrow$)					
		$\epsilon = 0.05$	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 2.0$	$\epsilon = 3.0$	$\epsilon = 5.0$
TinyBERT	54.40 $\pm$ 0.91	1.98 $\pm$ 0.55	16.64 $\pm$ 5.43	34.82 $\pm$ 9.30	74.60 $\pm$ 8.21	95.69 $\pm$ 2.35	100.00 $\pm$ 0.00
FreeLB	54.29 $\pm$ 1.24	1.25 $\pm$ 0.89	15.40 $\pm$ 7.95	30.64 $\pm$ 12.62	62.97 $\pm$ 13.61	88.69 $\pm$ 6.72	99.95 $\pm$ 0.07
DiffuseDef	55.32 $\pm$ 0.48	2.25 $\pm$ 0.58	20.63 $\pm$ 3.14	43.71 $\pm$ 6.14	81.08 $\pm$ 5.82	97.28 $\pm$ 1.38	100.00 $\pm$ 0.00
DynAttn	52.52 $\pm$ 0.71	0.46 $\pm$ 0.21	7.26 $\pm$ 1.71	17.74 $\pm$ 2.66	54.85 $\pm$ 4.05	87.97 $\pm$ 0.82	100.00 $\pm$ 0.00
Agatha	56.14 $\pm$ 0.74	1.99 $\pm$ 0.35	17.20 $\pm$ 0.70	30.15 $\pm$ 1.87	56.45 $\pm$ 2.90	77.01 $\pm$ 3.36	94.89 $\pm$ 2.11

Table 20: CoLA under standard PGD (ℓ_2): Linguistic acceptability (MCC normalized to 0 to 100 scale). **Bold** indicates best performance: highest clean score and lowest ASR.

ASR from $\epsilon = 0.5$ onward, recording 58.11 at $\epsilon = 2.0$ compared to 70.51 for FreeLB, 88.56 for DynAttn, and 89.38 for DiffuseDef. The high standard deviations across all models clearly reflect the small validation set size (276 examples).

WNLI (Winograd NLI). Agatha achieves the lowest ASR at $\epsilon = 0.5$ (12.36) and $\epsilon = 1.0$ (39.57). DiffuseDef records the lowest ASR at $\epsilon = 2.0$ (68.79) but with high seed variance (± 35.12), making this result unreliable. At high budgets, all models eventually converge to 100.00 ASR, reflecting the small validation set (71 examples) where individual flips have a large effect.

CoLA (Linguistic Acceptability). Agatha achieves the highest clean score (56.14). DynAttn provides the lowest ASR from $\epsilon = 0.05$ through 2.0 (0.46 at $\epsilon = 0.05$, 54.85 at $\epsilon = 2.0$), though at the cost of a lower clean score (52.52). Agatha recovers at higher budgets, recording the lowest ASR at $\epsilon = 3.0$ (77.01) and $\epsilon = 5.0$ (94.89). DiffuseDef records the highest ASR among all defenses at

$\epsilon = 2.0$ (81.08), indicating that pretraining degrades robustness on this task.

MRPC (Paraphrase Detection). Agatha matches TinyBERT in clean performance (75.52 versus 76.99) while providing the lowest ASR among all evaluated tasks. At $\epsilon = 5.0$, Agatha limits the ASR to 20.85, compared to 97.79 for DynAttn, 99.89 for DiffuseDef, 90.09 for FreeLB, and 99.77 for TinyBERT. The near-zero ASR at $\epsilon = 0.05$ (0.00 ± 0.00) confirms that the energy basins on this task are sufficiently deep to absorb small perturbations entirely.

QQP (Quora Question Pairs). Agatha maintains clean performance comparable to TinyBERT (72.12 versus 71.73) while consistently reducing ASR at all budgets from $\epsilon = 0.5$ onward. At $\epsilon = 2.0$, Agatha achieves 31.26 compared to 53.98 for DynAttn, 62.19 for DiffuseDef, and 45.34 for FreeLB. FreeLB achieves the highest clean accuracy on this task (77.41) but records 95.41 ASR at $\epsilon = 5.0$, compared to only 75.36 for Agatha.

Task	Model	Clean ($\uparrow$)	Attack Success Rate (% , $\downarrow$)					
			$\epsilon = 0.05$	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 2.0$	$\epsilon = 3.0$	$\epsilon = 5.0$
MRPC	TinyBERT	76.99 $\pm$ 0.49	1.25 $\pm$ 0.42	8.54 $\pm$ 2.39	19.58 $\pm$ 4.98	48.52 $\pm$ 7.72	80.91 $\pm$ 6.41	99.77 $\pm$ 0.33
	FreeLB	76.14 $\pm$ 0.10	1.13 $\pm$ 0.64	12.39 $\pm$ 0.81	20.83 $\pm$ 1.06	42.23 $\pm$ 1.64	64.07 $\pm$ 1.64	90.09 $\pm$ 1.68
	DiffuseDef	76.95 $\pm$ 0.67	0.80 $\pm$ 0.52	10.71 $\pm$ 2.58	21.89 $\pm$ 5.43	50.98 $\pm$ 10.50	84.60 $\pm$ 8.19	99.89 $\pm$ 0.20
	DynAttn	76.35 $\pm$ 0.56	0.92 $\pm$ 0.71	6.09 $\pm$ 3.10	13.47 $\pm$ 5.99	36.41 $\pm$ 11.43	64.64 $\pm$ 16.44	97.79 $\pm$ 2.65
	Agatha	75.52 $\pm$ 0.20	0.00 $\pm$ 0.00	0.71 $\pm$ 0.00	1.41 $\pm$ 0.00	4.48 $\pm$ 0.53	10.37 $\pm$ 0.75	20.85 $\pm$ 1.02
QQP	TinyBERT	71.73 $\pm$ 1.78	1.33 $\pm$ 0.08	13.69 $\pm$ 0.38	28.86 $\pm$ 1.17	60.12 $\pm$ 2.46	83.62 $\pm$ 1.93	99.02 $\pm$ 0.34
	FreeLB	77.41 $\pm$ 0.20	0.77 $\pm$ 0.03	9.18 $\pm$ 0.07	20.17 $\pm$ 0.10	45.34 $\pm$ 0.40	69.14 $\pm$ 0.67	95.41 $\pm$ 0.72
	DiffuseDef	72.56 $\pm$ 2.14	1.36 $\pm$ 0.06	14.25 $\pm$ 0.45	30.25 $\pm$ 1.43	62.19 $\pm$ 2.58	84.57 $\pm$ 2.04	98.81 $\pm$ 0.28
	DynAttn	74.75 $\pm$ 1.76	2.57 $\pm$ 0.11	11.78 $\pm$ 0.60	25.10 $\pm$ 1.87	53.98 $\pm$ 3.19	78.60 $\pm$ 2.58	98.68 $\pm$ 0.48
	Agatha	72.12 $\pm$ 0.51	0.76 $\pm$ 0.21	7.22 $\pm$ 1.88	14.84 $\pm$ 4.09	31.26 $\pm$ 8.67	47.79 $\pm$ 12.18	75.36 $\pm$ 13.39

Table 21: GLUE paraphrase tasks under standard PGD (ℓ_2): Clean score (average of F1 and accuracy) and Attack Success Rate (ASR). **Bold** indicates best performance: highest clean score and lowest ASR per task and budget.

Model	Clean ($\uparrow$)	Attack Success Rate (% , $\downarrow$)					
		Low	Medium		High	Extreme	
	($\epsilon = 0$)	($\epsilon = 0.05$)	($\epsilon = 0.5$)	($\epsilon = 1.0$)	($\epsilon = 2.0$)	($\epsilon = 3.0$)	($\epsilon = 5.0$)
<i>Standard Training</i>							
TinyBERT	67.26 $\pm$ 0.37	1.69 $\pm$ 0.15	16.05 $\pm$ 1.02	33.35 $\pm$ 1.82	66.00 $\pm$ 2.07	88.35 $\pm$ 1.25	99.67 $\pm$ 0.08
<i>Training-Based Robustness</i>							
TinyBERT + FreeLB	70.52 $\pm$ 0.32	1.34 $\pm$ 0.19	14.70 $\pm$ 1.37	30.05 $\pm$ 2.21	61.61 $\pm$ 2.58	84.04 $\pm$ 1.32	97.56 $\pm$ 0.30
<i>Architectural Robustness</i>							
DiffuseDef	69.04 $\pm$ 0.56	1.52 $\pm$ 0.14	16.77 $\pm$ 0.84	34.54 $\pm$ 1.62	66.78 $\pm$ 2.17	88.48 $\pm$ 1.57	99.42 $\pm$ 0.23
DynAttn	69.11 $\pm$ 0.36	2.11 $\pm$ 0.15	11.76 $\pm$ 0.84	25.14 $\pm$ 1.65	54.47 $\pm$ 2.96	79.81 $\pm$ 3.20	98.85 $\pm$ 0.52
Agatha (Ours)	66.57 $\pm$ 0.18	1.26 $\pm$ 0.12	11.66 $\pm$ 0.49	22.43 $\pm$ 0.92	43.04 $\pm$ 1.63	59.77 $\pm$ 2.35	78.30 $\pm$ 2.35

Table 22: Average GLUE-6 Score and Attack Success Rate (ASR) of standard PGD (ℓ_2) across 6 tasks (SST-2, MNLI, QNLI, QQP, CoLA, MRPC), excluding RTE (276 examples) and WNLI (71 examples). Results are reported as mean $\pm$ std across three random seeds. **Bold** indicates best performance.

Summary. Two patterns emerge across tasks:

1. Agatha provides the largest robustness gains on tasks with competitive clean accuracy (SST-2, MRPC, QQP, QNLI), confirming that well-formed energy basins around semantic attractors simultaneously support classification and perturbation contraction.
2. On CoLA, DynAttn achieves lower ASR at medium budgets but at a cost of 3.6 point clean accuracy gap; Agatha recovers at $\epsilon \geq 3.0$.

B.1.4 Robustness of Combined Results to Small Validation Sets

Two GLUE tasks have notably small validation sets: RTE (276 examples) and WNLI (71 examples). At medium and high perturbation budgets, these tasks produce standard deviations considerably larger than the remaining six tasks (e.g., ± 21.57 for Agatha on RTE and ± 8.11 for Agatha on WNLI at $\epsilon = 2.0$), and their inclusion inflates the combined variance reported in Table 1. Table 22 therefore replicates the main results across the six remaining tasks with stable validation sets: SST-2, MNLI, QNLI, QQP, CoLA, and MRPC (GLUE-6).

The GLUE-6 results clearly confirm the findings from the main paper. At $\epsilon = 2.0$, Agatha

achieves an ASR of 43.04 compared to 54.47 for DynAttn, a gap of 11.43 percentage points, and 61.61 for FreeLB, a gap of 18.57 points. The reduction in combined variance is the primary effect of the exclusion: Agatha’s standard deviation at $\epsilon = 2.0$ drops from ± 3.13 on GLUE-8 to ± 1.63 on GLUE-6.

The remaining variance in GLUE-6 at $\epsilon = 5.0$ (± 2.35) is driven primarily by QQP, which exhibits seed sensitivity of ± 13.39 at that budget. Unlike RTE and WNLI, this variance does not arise from a small validation set (QQP has 3,668 validation examples) but from the sensitivity of gradient-based attacks to initialization at large perturbation budgets, a pattern observed across all models on this task (per-task results in Table 21).

B.2 SuperGLUE Performance under Standard PGD (ℓ_2)

Table 23 provides per-task SuperGLUE results across all budgets. BoolQ and CB use the base architecture (128-dim, 2-layer, 2-head, 4.37M parameters). MultiRC and WSC use a scaled architecture (256-dim, 4-layer, 4-head, 11.1M parameters).

B.2.1 Task-Specific Observations

BoolQ (Boolean Questions). On BoolQ, Agatha achieves the highest clean accuracy (66.43) and the

Task	Model	Clean Acc	Attack Success Rate (%)					
			$\epsilon = 0.05$	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 2.0$	$\epsilon = 3.0$	$\epsilon = 5.0$
BoolQ	FreeLB	65.77 $\pm$ 0.44	2.30 $\pm$ 0.22	22.49 $\pm$ 0.86	45.47 $\pm$ 0.98	82.02 $\pm$ 0.83	97.46 $\pm$ 0.78	100.00 $\pm$ 0.00
	DiffuseDef	63.91 $\pm$ 0.94	1.56 $\pm$ 0.10	18.48 $\pm$ 1.53	38.96 $\pm$ 3.52	75.66 $\pm$ 4.20	96.09 $\pm$ 1.62	100.00 $\pm$ 0.00
	DynAttn	64.80 $\pm$ 0.52	2.37 $\pm$ 0.48	19.29 $\pm$ 2.83	38.68 $\pm$ 5.06	74.45 $\pm$ 7.55	93.91 $\pm$ 4.24	100.00 $\pm$ 0.00
	Agatha	66.43 $\pm$ 0.13	1.43 $\pm$ 0.30	16.28 $\pm$ 1.59	33.91 $\pm$ 2.80	66.41 $\pm$ 3.81	86.90 $\pm$ 2.25	98.96 $\pm$ 0.49
CB	FreeLB	70.83 $\pm$ 0.84	1.71 $\pm$ 2.42	9.25 $\pm$ 1.24	20.96 $\pm$ 4.99	60.51 $\pm$ 6.17	88.23 $\pm$ 1.26	100.00 $\pm$ 0.00
	DiffuseDef	72.02 $\pm$ 0.84	0.81 $\pm$ 1.15	11.52 $\pm$ 4.98	23.90 $\pm$ 5.52	52.83 $\pm$ 7.70	75.14 $\pm$ 6.30	98.33 $\pm$ 2.36
	DynAttn	71.43 $\pm$ 0.00	1.65 $\pm$ 1.16	9.90 $\pm$ 1.92	19.80 $\pm$ 3.25	46.67 $\pm$ 2.36	73.62 $\pm$ 6.84	96.67 $\pm$ 3.12
	Agatha	69.05 $\pm$ 1.68	2.63 $\pm$ 2.15	12.94 $\pm$ 0.31	18.16 $\pm$ 2.48	29.34 $\pm$ 1.69	38.03 $\pm$ 8.48	58.64 $\pm$ 9.70
MultiRC	FreeLB	65.10 $\pm$ 0.63	1.79 $\pm$ 0.19	22.14 $\pm$ 0.82	45.32 $\pm$ 1.14	80.90 $\pm$ 2.12	92.96 $\pm$ 1.84	98.09 $\pm$ 0.60
	DiffuseDef	50.19 $\pm$ 0.74	2.20 $\pm$ 0.16	20.23 $\pm$ 0.83	38.33 $\pm$ 1.36	66.23 $\pm$ 2.04	82.31 $\pm$ 2.32	96.24 $\pm$ 1.34
	DynAttn	52.94 $\pm$ 1.33	4.44 $\pm$ 0.58	24.44 $\pm$ 1.24	44.20 $\pm$ 2.31	70.68 $\pm$ 2.52	84.60 $\pm$ 2.08	95.81 $\pm$ 1.11
	Agatha	58.75 $\pm$ 0.89	4.36 $\pm$ 1.30	20.75 $\pm$ 1.38	34.40 $\pm$ 1.75	55.07 $\pm$ 3.59	69.96 $\pm$ 3.65	86.64 $\pm$ 2.96
WSC	FreeLB	63.78 $\pm$ 0.45	2.51 $\pm$ 1.89	19.04 $\pm$ 9.83	35.07 $\pm$ 14.37	64.19 $\pm$ 18.10	95.45 $\pm$ 6.43	100.00 $\pm$ 0.00
	DiffuseDef	64.74 $\pm$ 1.20	4.88 $\pm$ 3.77	22.80 $\pm$ 8.22	42.64 $\pm$ 12.55	82.67 $\pm$ 5.77	95.99 $\pm$ 2.84	100.00 $\pm$ 0.00
	DynAttn	61.22 $\pm$ 2.52	3.62 $\pm$ 2.02	16.52 $\pm$ 14.07	37.77 $\pm$ 12.94	75.09 $\pm$ 6.26	97.47 $\pm$ 3.57	100.00 $\pm$ 0.00
	Agatha	64.74 $\pm$ 0.45	0.50 $\pm$ 0.70	4.95 $\pm$ 1.87	6.43 $\pm$ 1.38	13.86 $\pm$ 1.36	22.26 $\pm$ 4.26	40.58 $\pm$ 3.38

Table 23: Full SuperGLUE results: Clean accuracy and Attack Success Rate (ASR) across all perturbation budgets. BoolQ and CB use the base architecture (4.37M). MultiRC and WSC use the scaled architecture (11.1M). **Bold** indicates best performance: highest clean accuracy and lowest ASR per task and budget.

lowest ASR at every budget. At $\epsilon = 2.0$, Agatha records 66.41 compared to 74.45 for DynAttn, 75.66 for DiffuseDef, and 82.02 for FreeLB.

CB (Commitment Bank). DiffuseDef achieves the highest clean accuracy (72.02), followed by DynAttn (71.43) and FreeLB (70.83). Agatha achieves the lowest ASR from $\epsilon = 1.0$ onward, recording 29.34 at $\epsilon = 2.0$, well below DynAttn (46.67), DiffuseDef (52.83), and FreeLB (60.51). At $\epsilon = 5.0$, Agatha achieves 58.64 while DynAttn records 96.67 and DiffuseDef 98.33, demonstrating that the energy basins continue to provide protection even at high budgets. The small validation set (56 examples) warrants caution in interpreting differences.

MultiRC (Multi-Sentence Reading Comprehension). MultiRC demonstrates the scaling pattern most clearly. Agatha achieves the highest clean accuracy among architectural defenses (58.75), while DiffuseDef (50.19) and DynAttn (52.94) both suffer a considerable accuracy drop. Agatha achieves the lowest ASR from $\epsilon = 1.0$ onward, recording 55.07 at $\epsilon = 2.0$, compared to DiffuseDef (66.23), DynAttn (70.68), and FreeLB (80.90). FreeLB achieves the highest overall clean accuracy (65.10) but this alone does not translate into robustness.

WSC (Winograd Schema Challenge). WSC provides the clearest demonstration of scalability. Agatha ties DiffuseDef for the highest clean accuracy (64.74) and achieves the lowest ASR at every budget. At $\epsilon = 2.0$, Agatha records 13.86 ASR, far below DynAttn (75.09), DiffuseDef (82.67), and FreeLB (64.19). The 4-layer architecture provides sufficient depth for the Hopfield mechanism to

form well-separated attractors for coreference resolution, which benefits from discrete semantic distinctions between referents.

B.2.2 Scaling Trends

Comparing the base (4.37M) and scaled (11.1M) architectures reveals a consistent pattern. On the base tasks, Agatha provides the lowest ASR while maintaining competitive clean accuracy across all defenses. On the scaled tasks, Agatha achieves the highest clean accuracy among architectural defenses while providing even stronger robustness (WSC: 64.74 clean, 13.86 ASR; MultiRC: 58.75 clean, 55.07 ASR). This suggests that increased model capacity allows the Hopfield energy landscape to form deeper basins that simultaneously support both classification and perturbation contraction, reducing the clean-robustness trade-off observed at smaller scales.

B.3 PGD (ℓ_2) With Increased Step Count

To test whether Agatha’s robustness degrades under stronger optimization, we evaluate PGD with increasing step counts $T_{\text{PGD}} \in \{5, 10, 20, 50, 100, 200\}$ at fixed $\epsilon = 2.0$. If the defense relied on gradient masking, additional optimization steps would recover a stronger attack signal and increase ASR.

B.3.1 Combined Results

Table 24 reports combined GLUE results.

All models plateau by $T_{\text{PGD}} = 20$, with negligible ASR increases from $T_{\text{PGD}} = 20$ to $T_{\text{PGD}} = 200$. Agatha’s combined GLUE ASR increases by only 0.54 points from $T_{\text{PGD}} = 5$ to $T_{\text{PGD}} = 200$ (49.68 $\rightarrow$ 50.22), comparable to DynAttn (1.43) and smaller

Model	ASR at $\varepsilon = 2.0$ (%)						Δ
	5	10	20	50	100	200	
TinyBERT	70.59 ± 2.01	71.80 ± 2.04	72.27 ± 2.10	72.42 ± 2.12	72.42 ± 2.11	72.43 ± 2.11	+1.83
FreeLB	64.64 ± 2.55	66.03 ± 2.51	66.68 ± 2.54	66.97 ± 2.55	67.03 ± 2.56	67.05 ± 2.56	+2.41
DD	67.45 ± 4.78	68.95 ± 4.81	69.84 ± 4.68	70.01 ± 4.70	70.03 ± 4.70	70.03 ± 4.70	+2.58
DA	61.69 ± 2.57	62.78 ± 2.43	63.03 ± 2.44	63.16 ± 2.47	63.05 ± 2.53	63.11 ± 2.48	+1.43
Agatha	49.68 ± 3.15	50.20 ± 3.13	50.27 ± 3.13	50.37 ± 3.13	50.21 ± 3.21	50.22 ± 3.20	+0.54

Table 24: PGD with increased step count: ASR at $\varepsilon = 2.0$ with increasing attacker optimization steps T_{PGD} . Column headers indicate T_{PGD} values. Δ = ASR change from $T_{\text{PGD}} = 5$ to $T_{\text{PGD}} = 200$. DD = DiffuseDef, DA = DynAttn. All models plateau by $T_{\text{PGD}} = 20$, confirming that the standard $T_{\text{PGD}} = 20$ evaluation used throughout is sufficient. **Bold** indicates lowest ASR per step count and smallest Δ .

than DiffuseDef (2.58) and FreeLB (2.41). The absence of a late-stage ASR increase with more optimization steps provides evidence that Agatha’s robustness is not an artifact of gradient masking.

B.3.2 Per-Task Results

Table 25 reports per-task PGD results at representative step counts. Two patterns emerge.

First, Agatha achieves the lowest ASR on six of eight tasks at $T_{\text{PGD}} = 200$, with the exceptions being CoLA (where DynAttn leads, 55.16 versus 56.40) and WNLI (where DiffuseDef leads at 68.79 but with high seed variance of ± 35.12). Second, Agatha’s ASR is stable across step counts. On MRPC, ASR remains at 4.48 to 4.71 from $T_{\text{PGD}} = 5$ to $T_{\text{PGD}} = 200$, and on RTE, at 57.89 throughout. This is consistent with the energy basin interpretation: once the attacker reaches the basin boundary, additional steps cannot push the representation further because the curvature deflects the gradient.

The only task with notable ASR decrease for Agatha is WNLI (-0.83 points), which reflects the small validation set (71 examples) rather than a systematic trend. This analysis complements the transfer attack results in Section 5.5: the transfer analysis rules out gradient masking by showing that black-box attacks do not recover the white-box ASR, while the adaptive analysis confirms that increasing the attacker optimization budget does not close the robustness gap.

B.4 Transfer Attack

Table 26 reports target ASR across all perturbation budgets for the transfer attack experiments on SST-

2 summarized in Table 4.

The transfer gap persists across budgets. The separation between TB→Agatha and TB→TB is consistent at every perturbation budget. At $\varepsilon = 0.5$, TB→Agatha records only 2.91 compared to 13.58 for TB→TB, a transfer rate of 21.4%. The rate grows modestly with budget (28.2% at $\varepsilon = 3.0$, 36.3% at $\varepsilon = 5.0$), consistent with larger perturbations being more likely to cross Agatha’s basin boundaries. At $\varepsilon = 5.0$, nearly two-thirds of TinyBERT-crafted adversarial examples fail to transfer to Agatha, confirming that the defense holds under large budgets and is not an artifact of gradient masking.

Asymmetric transfer holds across budgets.

The direction-dependent asymmetry observed at $\varepsilon = 2.0$ persists throughout the budget range. At $\varepsilon = 5.0$, Agatha→TB reaches 55.34 target ASR against only 35.89 for TB→Agatha. Perturbations that escape Agatha’s deep energy basins carry directional information that is broadly disruptive to TinyBERT, whereas perturbations optimized against TinyBERT’s smooth loss surface lack the geometric structure needed to breach Agatha.

B.5 Standard PGD (ℓ_∞)

To verify that Agatha’s robustness generalizes beyond the ℓ_2 threat model, we evaluate under standard PGD with an ℓ_∞ norm constraint.

Table 27 reports combined GLUE results across five perturbation budgets. Agatha achieves the lowest ASR at every ℓ_∞ budget. At $\varepsilon = 0.01$, Agatha records 11.69 compared to 15.74 for DynAttn and 17.24 for DiffuseDef. The advantage persists at higher budgets: at $\varepsilon = 0.05$, Agatha records 54.36 versus 67.85 for DynAttn, a gap of 13.49 points; at $\varepsilon = 0.1$, the gap widens to 16.82 points (77.60 versus 94.42). This confirms that Agatha’s robustness is norm-independent, consistent with Hopfield energy basin geometry operating on representation magnitude rather than a specific perturbation direction.

B.6 AutoAttack (ℓ_2)

AutoAttack (Croce and Hein, 2020b) is an ensemble evaluation protocol designed to detect gradient obfuscation. It combines two complementary attack strategies, APGD-CE and FAB, each targeting a different failure mode.

APGD-CE (Auto-PGD with Cross-Entropy Loss). APGD-CE (Croce and Hein, 2020b) extends standard PGD with momentum and adaptive

Task	Model	ASR at $\varepsilon = 2.0$ (%)					Δ
		$T_{\text{PGD}} = 5$	$T_{\text{PGD}} = 20$	$T_{\text{PGD}} = 50$	$T_{\text{PGD}} = 100$	$T_{\text{PGD}} = 200$	
SST-2	DD	57.18 $\pm$ 0.95	60.31 $\pm$ 1.08	60.54 $\pm$ 1.11	60.54 $\pm$ 1.11	60.54 $\pm$ 1.11	+3.36
	DA	53.91 $\pm$ 1.76	53.96 $\pm$ 0.84	54.29 $\pm$ 1.18	54.29 $\pm$ 0.35	53.68 $\pm$ 1.12	-0.23
	Agatha	45.28 $\pm$ 3.53	45.90 $\pm$ 3.38	46.33 $\pm$ 3.14	46.23 $\pm$ 3.45	46.23 $\pm$ 3.08	+0.95
CoLA	DD	79.92 $\pm$ 6.02	81.08 $\pm$ 5.82	81.08 $\pm$ 5.82	81.08 $\pm$ 5.82	81.08 $\pm$ 5.82	+1.15
	DA	53.62 $\pm$ 4.27	54.60 $\pm$ 3.41	55.76 $\pm$ 3.97	55.16 $\pm$ 4.62	55.16 $\pm$ 4.18	+1.53
	Agatha	55.87 $\pm$ 2.81	56.50 $\pm$ 2.98	56.45 $\pm$ 2.90	56.45 $\pm$ 2.90	56.40 $\pm$ 2.82	+0.54
MRPC	DD	49.04 $\pm$ 9.99	50.86 $\pm$ 10.38	51.32 $\pm$ 10.78	51.32 $\pm$ 10.78	51.32 $\pm$ 10.78	+2.28
	DA	35.22 $\pm$ 10.09	37.07 $\pm$ 10.57	36.14 $\pm$ 10.60	36.14 $\pm$ 11.51	36.71 $\pm$ 11.70	+1.49
	Agatha	4.48 $\pm$ 0.19	4.48 $\pm$ 0.19	4.71 $\pm$ 0.39	4.48 $\pm$ 0.19	4.71 $\pm$ 0.39	+0.24
QQP	DD	59.25 $\pm$ 2.55	62.19 $\pm$ 2.58	62.53 $\pm$ 2.66	62.58 $\pm$ 2.68	62.60 $\pm$ 2.69	+3.35
	DA	51.67 $\pm$ 3.36	53.36 $\pm$ 3.29	53.45 $\pm$ 3.34	53.53 $\pm$ 3.19	53.45 $\pm$ 3.18	+1.78
	Agatha	30.32 $\pm$ 8.28	31.32 $\pm$ 8.70	31.35 $\pm$ 8.71	31.37 $\pm$ 8.71	31.39 $\pm$ 8.64	+1.07
MNLI	DD	70.00 $\pm$ 3.76	73.77 $\pm$ 2.56	74.13 $\pm$ 2.45	74.19 $\pm$ 2.43	74.22 $\pm$ 2.40	+4.22
	DA	60.56 $\pm$ 0.69	63.68 $\pm$ 0.60	64.67 $\pm$ 0.93	64.44 $\pm$ 0.69	64.36 $\pm$ 0.81	+3.80
	Agatha	61.48 $\pm$ 0.55	62.55 $\pm$ 0.54	62.58 $\pm$ 0.55	62.67 $\pm$ 0.63	62.65 $\pm$ 0.57	+1.16
QNLI	DD	71.61 $\pm$ 3.30	72.32 $\pm$ 3.22	72.34 $\pm$ 3.21	72.35 $\pm$ 3.22	72.35 $\pm$ 3.22	+0.73
	DA	61.02 $\pm$ 12.27	62.08 $\pm$ 12.33	61.85 $\pm$ 12.20	61.67 $\pm$ 12.55	62.02 $\pm$ 12.22	+0.99
	Agatha	56.25 $\pm$ 0.56	57.45 $\pm$ 0.18	57.56 $\pm$ 0.25	57.59 $\pm$ 0.23	57.48 $\pm$ 0.26	+1.24
RTE	DD	87.92 $\pm$ 1.91	89.38 $\pm$ 2.18	89.38 $\pm$ 2.18	89.38 $\pm$ 2.18	89.38 $\pm$ 2.18	+1.46
	DA	87.81 $\pm$ 4.98	88.22 $\pm$ 5.78	87.79 $\pm$ 6.28	88.86 $\pm$ 6.39	87.37 $\pm$ 6.25	-0.44
	Agatha	57.89 $\pm$ 21.94	58.11 $\pm$ 21.57	58.11 $\pm$ 21.57	57.89 $\pm$ 21.94	57.89 $\pm$ 21.94	+0.00
WNLI	DD	64.69 $\pm$ 35.91	68.79 $\pm$ 35.12	68.79 $\pm$ 35.12	68.79 $\pm$ 35.12	68.79 $\pm$ 35.12	+4.11
	DA	89.67 $\pm$ 10.62	91.28 $\pm$ 7.76	91.33 $\pm$ 7.76	90.35 $\pm$ 6.80	92.17 $\pm$ 6.34	+2.50
	Agatha	85.87 $\pm$ 8.11	85.87 $\pm$ 8.11	85.87 $\pm$ 8.11	85.04 $\pm$ 9.08	85.04 $\pm$ 9.08	-0.83

Table 25: Per-task PGD results at $\varepsilon = 2.0$. Δ = ASR change from $T_{\text{PGD}} = 5$ to $T_{\text{PGD}} = 200$. DD = DiffuseDef, DA = DynAttn. **Bold** indicates lowest ASR per task and step count.

Direction	Target ASR (%)				
	$\varepsilon = 0.5$	$\varepsilon = 1.0$	$\varepsilon = 2.0$	$\varepsilon = 3.0$	$\varepsilon = 5.0$
<i>Transfer</i>					
TB→Agatha	2.91 $\pm$ 0.60	7.05 $\pm$ 1.13	15.63 $\pm$ 2.11	23.35 $\pm$ 3.61	35.89 $\pm$ 6.92
Agatha→TB	4.31 $\pm$ 0.50	9.89 $\pm$ 1.13	21.34 $\pm$ 2.83	32.28 $\pm$ 4.48	55.34 $\pm$ 6.83
<i>White-box</i>					
TB→TB	13.58 $\pm$ 0.46	28.34 $\pm$ 2.82	59.90 $\pm$ 3.27	82.85 $\pm$ 3.53	98.82 $\pm$ 0.14
Agatha→Agatha	10.05 $\pm$ 1.21	21.34 $\pm$ 1.82	44.35 $\pm$ 2.73	67.94 $\pm$ 3.91	91.38 $\pm$ 1.51

Table 26: Full transfer attack results on SST-2. Target ASR is the percentage of the target model’s correct predictions flipped by adversarial examples crafted on the source model. Mean $\pm$ std across three seeds.

step-size halving. Given a model f , input $\mathbf{x}$, and true label y , APGD-CE iterates

$$\begin{aligned} \mathbf{g}^{(t)} &= \nabla_{\mathbf{x}} \mathcal{L}_{\text{CE}}(f(\mathbf{x}^{(t)}), y) + \beta \cdot \mathbf{m}^{(t)}, \\ \mathbf{x}^{(t+1)} &= \Pi_{\mathcal{B}_\varepsilon(\mathbf{x})} \left(\mathbf{x}^{(t)} + \alpha^{(t)} \cdot \text{sign}(\mathbf{g}^{(t)}) \right), \end{aligned} \quad (9)$$

where $\mathbf{m}^{(t)}$ is a momentum term and $\alpha^{(t)}$ is halved whenever the loss fails to improve over a fixed window of iterations. The attack also checkpoints the single best adversarial example found across all steps. Together, these features make APGD-CE strictly stronger than standard PGD: the adaptive step size avoids oscillation near optima, momentum escapes shallow local maxima, and checkpointing ensures that transient successes are not lost.

FAB (Fast Adaptive Boundary). FAB (Croce and Hein, 2020a) takes a fundamentally different approach: rather than maximizing the classification loss within an ε -ball, it seeks the minimum-norm

perturbation that crosses the decision boundary. At each step, FAB linearizes the classifier around the current point and projects toward the closest decision boundary:

$$\begin{aligned} \mathbf{d}^{(t)} &= \nabla_{\mathbf{x}}(f_y - f_j)(\mathbf{x}^{(t)}), \\ \mathbf{x}^{(t+1)} &= \Pi_{\mathcal{B}_\varepsilon(\mathbf{x})} \left(\mathbf{x}^{(t)} - \eta^{(t)} \cdot \frac{f_y(\mathbf{x}^{(t)}) - f_j(\mathbf{x}^{(t)})}{\|\mathbf{d}^{(t)}\|^2} \cdot \mathbf{d}^{(t)} \right), \end{aligned} \quad (10)$$

where $j = \arg \max_{k \neq y} f_k(\mathbf{x}^{(t)})$ is the runner-up class. This boundary-seeking strategy is complementary to APGD-CE: it detects cases where the decision boundary is close but the loss gradient points away from it, a signature of gradient masking.

Square Attack. Square (Andriushchenko et al., 2020) takes a third approach: it uses no gradient information at all. It performs randomized search over localized square-shaped updates to the input, accepting a candidate only when it improves the

Model	ASR (% , $\downarrow$)				
	$\epsilon = 0.001$	$\epsilon = 0.005$	$\epsilon = 0.01$	$\epsilon = 0.05$	$\epsilon = 0.1$
TinyBERT	2.27 $\pm$ 0.15	9.71 $\pm$ 0.70	19.03 $\pm$ 1.06	72.94 $\pm$ 1.89	94.98 $\pm$ 0.50
FreeLB	2.22 $\pm$ 0.29	10.79 $\pm$ 0.90	20.82 $\pm$ 1.40	82.56 $\pm$ 1.79	99.45 $\pm$ 0.07
DiffuseDef	1.72 $\pm$ 0.30	8.65 $\pm$ 1.69	17.24 $\pm$ 1.75	71.63 $\pm$ 3.28	95.47 $\pm$ 0.44
DynAttn	4.35 $\pm$ 0.64	8.51 $\pm$ 0.44	15.74 $\pm$ 0.98	67.85 $\pm$ 2.31	94.42 $\pm$ 0.39
Agatha	1.26 $\pm$ 0.19	6.05 $\pm$ 0.68	11.69 $\pm$ 1.37	54.36 $\pm$ 2.02	77.60 $\pm$ 1.24

Table 27: Combined GLUE ASR under standard PGD (ℓ_∞). **Bold:** lowest ASR per budget.

Model	Attack	ASR (% , $\downarrow$)				
		$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 2.0$	$\epsilon = 3.0$	$\epsilon = 5.0$
TinyBERT	APGD-CE	21.37 $\pm$ 1.01	39.83 $\pm$ 1.62	69.11 $\pm$ 1.98	84.68 $\pm$ 1.25	96.95 $\pm$ 0.42
	FAB	21.49 $\pm$ 1.01	40.15 $\pm$ 1.63	71.69 $\pm$ 1.65	89.90 $\pm$ 0.85	99.61 $\pm$ 0.09
	AutoAttack	21.49 $\pm$ 1.01	40.15 $\pm$ 1.63	71.69 $\pm$ 1.65	89.90 $\pm$ 0.85	99.61 $\pm$ 0.09
FreeLB	APGD-CE	18.94 $\pm$ 1.36	34.56 $\pm$ 1.90	61.66 $\pm$ 2.24	78.45 $\pm$ 1.42	93.32 $\pm$ 0.34
	FAB	19.13 $\pm$ 1.36	35.62 $\pm$ 1.91	66.65 $\pm$ 2.07	85.90 $\pm$ 0.97	98.01 $\pm$ 0.24
	AutoAttack	19.13 $\pm$ 1.36	35.62 $\pm$ 1.91	66.65 $\pm$ 2.07	85.90 $\pm$ 0.97	98.01 $\pm$ 0.24
DiffuseDef	APGD-CE	18.92 $\pm$ 2.10	37.53 $\pm$ 3.85	66.36 $\pm$ 4.20	82.99 $\pm$ 2.59	96.17 $\pm$ 0.65
	FAB	18.96 $\pm$ 2.11	37.96 $\pm$ 3.75	69.03 $\pm$ 3.80	89.25 $\pm$ 1.50	99.41 $\pm$ 0.15
	AutoAttack	18.96 $\pm$ 2.11	37.96 $\pm$ 3.75	69.03 $\pm$ 3.80	89.25 $\pm$ 1.50	99.41 $\pm$ 0.15
DynAttn	APGD-CE	15.90 $\pm$ 0.63	30.77 $\pm$ 1.01	56.94 $\pm$ 2.10	76.21 $\pm$ 2.02	93.85 $\pm$ 1.50
	FAB	13.64 $\pm$ 0.89	26.89 $\pm$ 1.05	52.30 $\pm$ 1.82	70.59 $\pm$ 1.87	83.89 $\pm$ 1.00
	AutoAttack	15.90 $\pm$ 0.63	30.77 $\pm$ 1.01	56.94 $\pm$ 2.10	76.21 $\pm$ 2.02	93.85 $\pm$ 1.50
Agatha	APGD-CE	13.16 $\pm$ 1.06	25.65 $\pm$ 2.15	48.26 $\pm$ 2.67	61.45 $\pm$ 2.92	75.72 $\pm$ 2.40
	FAB	13.28 $\pm$ 1.07	27.08 $\pm$ 2.23	51.45 $\pm$ 2.59	66.56 $\pm$ 2.55	83.09 $\pm$ 1.50
	AutoAttack	13.28 $\pm$ 1.07	27.08 $\pm$ 2.23	51.45 $\pm$ 2.59	66.56 $\pm$ 2.55	83.09 $\pm$ 1.50

Table 28: Combined GLUE ASR under AutoAttack (ℓ_2). APGD-CE and FAB are reported individually; AutoAttack is the per-example worst case. **Bold:** lowest AutoAttack ASR per budget.

margin loss, and terminates when the query budget is exhausted. Because it never differentiates the model, a defense that suppresses gradients without changing the underlying decision geometry remains fully exposed to it.

AutoAttack. The AutoAttack protocol runs both APGD-CE and FAB independently and reports the per-example worst case:

$$\text{ASR}_{\text{AA}}(\mathbf{x}) = \max(\text{ASR}_{\text{APGD}}(\mathbf{x}), \text{ASR}_{\text{FAB}}(\mathbf{x})). \quad (11)$$

A defense that suppresses one attack but not the other is exposed by this protocol.

Table 28 reports combined GLUE ASR for each component and their worst case.

Results. Agatha achieves the lowest AutoAttack ASR at every budget, from 13.28 at $\epsilon = 0.5$ to 83.09 at $\epsilon = 5.0$. The gap over DynAttn grows with budget: 2.62 points at $\epsilon = 0.5$, 5.49 at $\epsilon = 2.0$, and 9.65 at $\epsilon = 3.0$. The worst-case attack differs by defense: FAB dominates for TinyBERT, FreeLB, DiffuseDef, and Agatha, while APGD-CE dominates for DynAttn. For Agatha, FAB consistently finds higher ASR than APGD-CE, indicating that boundary-seeking is more effective than loss-maximizing against the energy basin geometry. This is expected: the Hopfield energy landscape flattens the loss surface within basins (contracting APGD-CE), but the decision boundaries them-

selves remain reachable via linearization (favoring FAB). Despite this, Agatha’s worst-case ASR remains below all baselines at every budget.

Gradient-Free Evaluation with Square. Table 29 reports each attack at $\epsilon = 2.0$ against the strongest architectural baseline, with Square run at a budget of 5,000 queries.

Model	Clean ($\uparrow$)	APGD-CE	FAB	Square	Worst-case
DynAttn	63.38 $\pm$ 0.21	56.94 $\pm$ 2.10	52.30 $\pm$ 1.82	9.31 $\pm$ 4.01	56.94 $\pm$ 2.10
Agatha	63.91 $\pm$ 0.20	48.26 $\pm$ 2.67	51.45 $\pm$ 2.59	5.31 $\pm$ 3.54	51.45 $\pm$ 2.59

Table 29: Combined GLUE ASR at $\epsilon = 2.0$ under ℓ_2 . Square uses a budget of 5,000 queries; the worst-case column is the AutoAttack ensemble of Eq. 11, taken over APGD-CE and FAB. **Bold:** best performance: highest clean accuracy and lowest ASR.

Square records the lowest ASR of the three attacks for both models, at 5.31 for Agatha against a white-box worst case of 51.45, an order of magnitude lower. Gradient masking predicts the opposite ordering, in which gradient-free attacks match or exceed gradient-based ones (Athalye et al., 2018). The gradient-based components are therefore exploiting the unrolled gradients rather than being deflected by them, which is consistent with the transfer analysis of Section 5.5, where surrogate-crafted examples recover only 26.1% of white-box ASR. The same ordering holds for DynAttn, so the

gap between the black-box and white-box attacks is not specific to Agatha.

APGD-DLR. The DLR loss normalizes by the gap between the top and third logit and is therefore defined only for $K \geq 3$ classes, which excludes seven of the eight GLUE tasks we evaluate. This motivated the APGD-CE and FAB ensemble used throughout. We report APGD-DLR on MNLI, the one task where it applies, at $\varepsilon = 2.0$ under ℓ_2 .

Model	Clean ($\uparrow$)	Standard PGD	APGD-DLR
DynAttn	64.25 $\pm$ 0.59	64.84 $\pm$ 1.26	65.23 $\pm$ 0.38
Agatha	54.64 $\pm$ 0.13	62.55 $\pm$ 0.54	65.28 $\pm$ 0.22

Table 30: MNLI clean accuracy and ASR at $\varepsilon = 2.0$ under ℓ_2 . APGD-DLR applies only to MNLI among the eight GLUE tasks. **Bold:** best performance: highest clean accuracy and lowest ASR.

APGD-DLR is stronger against Agatha than standard PGD on this task, at 65.28 against 62.55, consistent with the FAB result: margin-based, boundary-seeking objectives are the worst case for Agatha’s flattened loss surface. The 2.29 point advantage Agatha holds over DynAttn under standard PGD therefore does not persist under APGD-DLR, where the two are statistically tied. DynAttn also retains higher clean accuracy on this task. The shift under DLR is nonetheless far smaller than the ASR recovery reported when gradient masking is circumvented (Athalye et al., 2018), which is the signature this evaluation is designed to detect.

B.7 APGD-CE (ℓ_∞)

To complement the standard PGD ℓ_∞ evaluation (Appendix B.5), we evaluate under APGD-CE with an ℓ_∞ norm constraint. APGD-CE’s adaptive step-size halving and momentum make it strictly stronger than standard PGD, providing a more rigorous test of ℓ_∞ robustness. Table 31 reports combined GLUE results.

Model	ASR (% , $\downarrow$)				
	$\varepsilon = 0.001$	$\varepsilon = 0.005$	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$
TinyBERT	2.27 $\pm$ 0.15	9.71 $\pm$ 0.70	19.03 $\pm$ 1.06	72.93 $\pm$ 1.90	95.00 $\pm$ 0.50
FreeLB	2.22 $\pm$ 0.29	10.79 $\pm$ 0.90	20.82 $\pm$ 1.40	82.58 $\pm$ 1.78	99.49 $\pm$ 0.08
DiffuseDef	1.72 $\pm$ 0.30	8.65 $\pm$ 1.69	17.24 $\pm$ 1.75	71.64 $\pm$ 3.28	95.55 $\pm$ 0.45
DynAttn	4.40 $\pm$ 0.39	8.84 $\pm$ 0.79	15.92 $\pm$ 1.26	64.33 $\pm$ 2.47	92.40 $\pm$ 0.55
Agatha	1.32 $\pm$ 0.21	6.33 $\pm$ 0.72	12.51 $\pm$ 1.43	55.69 $\pm$ 1.84	78.66 $\pm$ 1.32

Table 31: Combined GLUE ASR under APGD-CE (ℓ_∞). **Bold:** lowest ASR per budget.

Agatha achieves the lowest ASR at every budget, consistent with the standard PGD ℓ_∞ results. At $\varepsilon = 0.01$, Agatha records 12.51 compared to 15.92

for DynAttn and 17.24 for DiffuseDef. At $\varepsilon = 0.05$, Agatha records 55.69 versus 64.33 for DynAttn, a gap of 8.64 points, and at $\varepsilon = 0.1$ the gap widens to 13.74 points (78.66 versus 92.40). The ASR values under APGD-CE are marginally higher than under standard PGD at the same budgets (e.g., 12.51 versus 11.69 at $\varepsilon = 0.01$), confirming that APGD-CE finds slightly stronger attacks, but the modest improvement does not close the gap over baselines.

C Performance of Agatha on Other Attacks

C.1 FGSM (Embedding Space)

Table 32 reports per-task FGSM results at three representative budgets.

Task	Model	Attack Success Rate (%)		
		$\epsilon = 1.0$	$\epsilon = 2.0$	$\epsilon = 5.0$
SST-2	FreeLB	33.84 $\pm$ 1.74	63.65 $\pm$ 4.39	97.93 $\pm$ 0.92
	DiffuseDef	26.79 $\pm$ 1.39	51.24 $\pm$ 1.72	92.90 $\pm$ 1.67
	DynAttn	25.15 $\pm$ 0.91	50.77 $\pm$ 1.62	94.45 $\pm$ 0.44
	Agatha	20.59 $\pm$ 2.69	40.56 $\pm$ 4.42	80.26 $\pm$ 8.41
CoLA	FreeLB	30.18 $\pm$ 15.11	58.37 $\pm$ 18.16	97.52 $\pm$ 2.01
	DiffuseDef	42.26 $\pm$ 5.49	76.03 $\pm$ 6.61	99.04 $\pm$ 0.71
	DynAttn	17.26 $\pm$ 1.13	51.14 $\pm$ 5.65	98.33 $\pm$ 0.98
	Agatha	29.66 $\pm$ 1.67	55.04 $\pm$ 2.19	92.11 $\pm$ 2.06
MRPC	FreeLB	20.38 $\pm$ 1.30	39.98 $\pm$ 1.64	81.08 $\pm$ 1.72
	DiffuseDef	20.86 $\pm$ 5.10	46.53 $\pm$ 10.77	95.89 $\pm$ 1.80
	DynAttn	12.84 $\pm$ 6.33	32.20 $\pm$ 9.97	86.58 $\pm$ 12.67
	Agatha	1.41 $\pm$ 0.00	3.77 $\pm$ 0.19	18.14 $\pm$ 0.15
QQP	FreeLB	18.63 $\pm$ 0.20	36.86 $\pm$ 0.59	73.35 $\pm$ 1.08
	DiffuseDef	27.71 $\pm$ 1.72	52.37 $\pm$ 3.28	90.31 $\pm$ 2.13
	DynAttn	22.15 $\pm$ 1.74	45.53 $\pm$ 3.42	87.75 $\pm$ 2.65
	Agatha	13.35 $\pm$ 3.49	25.99 $\pm$ 6.65	53.90 $\pm$ 12.05
MNLI	FreeLB	40.06 $\pm$ 0.79	69.32 $\pm$ 0.92	98.30 $\pm$ 0.97
	DiffuseDef	37.48 $\pm$ 1.48	63.35 $\pm$ 4.94	91.49 $\pm$ 2.66
	DynAttn	28.94 $\pm$ 0.62	51.67 $\pm$ 0.59	88.29 $\pm$ 0.67
	Agatha	33.65 $\pm$ 1.15	57.81 $\pm$ 0.61	86.97 $\pm$ 0.86
QNLI	FreeLB	28.09 $\pm$ 3.42	57.28 $\pm$ 4.62	96.48 $\pm$ 1.16
	DiffuseDef	40.80 $\pm$ 4.27	69.81 $\pm$ 2.82	97.08 $\pm$ 1.03
	DynAttn	30.86 $\pm$ 7.90	57.14 $\pm$ 12.82	92.48 $\pm$ 6.25
	Agatha	28.44 $\pm$ 0.92	51.87 $\pm$ 2.07	86.51 $\pm$ 4.36
RTE	FreeLB	37.04 $\pm$ 3.50	69.91 $\pm$ 1.29	98.20 $\pm$ 0.62
	DiffuseDef	52.11 $\pm$ 3.44	84.80 $\pm$ 1.89	99.79 $\pm$ 0.36
	DynAttn	50.62 $\pm$ 3.61	83.99 $\pm$ 3.45	100.00 $\pm$ 0.00
	Agatha	34.23 $\pm$ 17.30	56.57 $\pm$ 23.21	81.20 $\pm$ 17.82
WNLI	FreeLB	62.83 $\pm$ 9.21	92.39 $\pm$ 8.56	99.17 $\pm$ 1.44
	DiffuseDef	43.56 $\pm$ 36.38	62.12 $\pm$ 39.87	94.06 $\pm$ 8.27
	DynAttn	60.99 $\pm$ 2.17	88.05 $\pm$ 10.78	99.19 $\pm$ 1.41
	Agatha	37.93 $\pm$ 9.73	83.39 $\pm$ 8.88	98.35 $\pm$ 1.43
Combined GLUE (8 tasks)				
	FreeLB	33.88 $\pm$ 2.31	60.97 $\pm$ 2.65	92.75 $\pm$ 0.46
	DiffuseDef	36.45 $\pm$ 4.70	63.28 $\pm$ 5.30	95.07 $\pm$ 1.17
	DynAttn	31.10 $\pm$ 1.40	57.56 $\pm$ 2.62	93.38 $\pm$ 1.81
	Agatha	24.91 $\pm$ 2.56	46.87 $\pm$ 3.29	74.68 $\pm$ 2.96

Table 32: FGSM attack results per task. **Bold** indicates lowest ASR per task and budget.

Agatha achieves the lowest combined GLUE ASR at all budgets, with an 11 point advantage over DynAttn at $\epsilon = 2.0$. On most tasks, Agatha records the lowest ASR; the exceptions are CoLA and MNLI, where DynAttn leads at $\epsilon \leq 2.0$, QNLI, where FreeLB leads at $\epsilon = 1.0$, and WNLI, where DiffuseDef leads at $\epsilon = 2.0$ but with high seed variance (± 39.87). The pattern mirrors PGD: energy basin structure provides robustness against gradient-based embedding perturbation.

C.2 Gaussian Noise (Embedding Space)

Table 33 reports per-task Gaussian noise results. Unlike PGD and FGSM, Gaussian perturbations are non-adversarial and test representation stability. Agatha reduces the combined Gaussian flip rate by 6 points relative to the best architectural baseline (DynAttn, 29.75) at $\sigma = 2.0$. The advantage is

Task	Model	Flip Rate (%)		
		$\sigma = 1.0$	$\sigma = 2.0$	$\sigma = 5.0$
SST-2	FreeLB	16.75 $\pm$ 1.67	33.20 $\pm$ 1.10	44.58 $\pm$ 0.31
	DiffuseDef	9.86 $\pm$ 0.62	22.30 $\pm$ 0.52	43.66 $\pm$ 3.13
	DynAttn	9.32 $\pm$ 0.49	22.35 $\pm$ 2.70	42.11 $\pm$ 3.44
	Agatha	12.30 $\pm$ 3.01	29.46 $\pm$ 5.03	49.10 $\pm$ 5.73
CoLA	FreeLB	20.50 $\pm$ 5.09	37.69 $\pm$ 19.41	39.87 $\pm$ 14.82
	DiffuseDef	10.37 $\pm$ 2.45	17.44 $\pm$ 5.49	43.07 $\pm$ 24.48
	DynAttn	3.27 $\pm$ 0.70	9.80 $\pm$ 3.79	30.99 $\pm$ 10.49
	Agatha	12.66 $\pm$ 1.36	21.03 $\pm$ 1.26	34.22 $\pm$ 4.55
MRPC	FreeLB	27.48 $\pm$ 2.95	20.38 $\pm$ 2.77	26.35 $\pm$ 1.13
	DiffuseDef	15.72 $\pm$ 3.18	16.51 $\pm$ 4.06	9.46 $\pm$ 0.39
	DynAttn	9.60 $\pm$ 5.98	11.11 $\pm$ 4.92	9.11 $\pm$ 2.59
	Agatha	0.24 $\pm$ 0.20	0.83 $\pm$ 0.41	6.72 $\pm$ 3.43
QQP	FreeLB	36.56 $\pm$ 0.61	36.81 $\pm$ 0.47	40.78 $\pm$ 3.38
	DiffuseDef	21.37 $\pm$ 0.43	44.15 $\pm$ 4.08	47.72 $\pm$ 4.00
	DynAttn	23.92 $\pm$ 3.05	39.79 $\pm$ 4.19	51.98 $\pm$ 2.51
	Agatha	6.77 $\pm$ 2.03	15.15 $\pm$ 5.45	38.77 $\pm$ 15.53
MNLI	FreeLB	45.11 $\pm$ 6.26	61.09 $\pm$ 1.44	64.45 $\pm$ 1.35
	DiffuseDef	26.07 $\pm$ 3.19	49.45 $\pm$ 3.71	58.93 $\pm$ 2.75
	DynAttn	36.01 $\pm$ 3.82	51.55 $\pm$ 1.97	57.68 $\pm$ 2.15
	Agatha	27.05 $\pm$ 1.40	45.69 $\pm$ 0.15	62.19 $\pm$ 1.45
QNLI	FreeLB	40.37 $\pm$ 6.47	49.33 $\pm$ 2.61	50.84 $\pm$ 1.00
	DiffuseDef	22.33 $\pm$ 7.90	39.03 $\pm$ 3.84	43.84 $\pm$ 7.09
	DynAttn	34.85 $\pm$ 17.29	36.76 $\pm$ 3.04	33.78 $\pm$ 6.52
	Agatha	15.96 $\pm$ 2.24	30.83 $\pm$ 5.04	45.44 $\pm$ 4.59
RTE	FreeLB	21.94 $\pm$ 4.66	40.10 $\pm$ 7.48	46.24 $\pm$ 3.77
	DiffuseDef	22.48 $\pm$ 2.10	35.63 $\pm$ 0.95	45.84 $\pm$ 3.71
	DynAttn	25.34 $\pm$ 5.19	39.26 $\pm$ 2.97	48.11 $\pm$ 3.15
	Agatha	15.71 $\pm$ 8.04	23.88 $\pm$ 8.83	37.76 $\pm$ 12.44
WNLI	FreeLB	21.77 $\pm$ 1.85	30.57 $\pm$ 8.52	41.04 $\pm$ 7.68
	DiffuseDef	10.27 $\pm$ 9.53	18.30 $\pm$ 14.37	36.49 $\pm$ 5.10
	DynAttn	23.61 $\pm$ 3.61	27.37 $\pm$ 6.44	45.85 $\pm$ 10.86
	Agatha	14.76 $\pm$ 14.64	24.72 $\pm$ 12.91	40.49 $\pm$ 0.84
Combined GLUE (8 tasks)				
	FreeLB	28.81 $\pm$ 1.50	38.65 $\pm$ 2.86	44.27 $\pm$ 2.19
	DiffuseDef	17.31 $\pm$ 1.70	30.35 $\pm$ 2.16	41.13 $\pm$ 3.36
	DynAttn	20.74 $\pm$ 2.50	29.75 $\pm$ 1.41	39.95 $\pm$ 2.20
	Agatha	13.18 $\pm$ 2.17	23.95 $\pm$ 2.26	39.34 $\pm$ 2.75

Table 33: Gaussian noise results per task. **Bold** indicates lowest flip rate per task and noise level.

particularly large on MRPC (0.83 versus 11.11 for DynAttn) and QQP (15.15 versus 39.79). On SST-2 and CoLA, DynAttn achieves lower flip rates than Agatha, consistent with the FGSM pattern on these tasks. On MNLI, DiffuseDef leads at $\sigma = 1.0$ and DynAttn at $\sigma = 5.0$, while Agatha leads at $\sigma = 2.0$. Since Gaussian noise is non-adversarial, the combined results confirm that the energy basin structure provides general representation stability, not just adversarial gradient contraction.

C.3 CharSwap (Discrete Text-Level)

Table 34 reports CharSwap results, which swap adjacent characters in randomly selected words (Pruthi et al., 2019).

Agatha achieves the lowest combined CharSwap flip rate at all perturbation levels, with the gap widening at higher rates (16.07 versus 24.72 for DiffuseDef at 50%). On SST-2 and CoLA, DynAttn records lower flip rates, consistent with the pattern observed for FGSM and Gaussian noise on these tasks. On QNLI, DiffuseDef leads at 10% but

Task	Model	Flip Rate (%)		
		10%	30%	50%
SST-2	FreeLB	8.03 $\pm$ 0.35	26.06 $\pm$ 0.43	44.58 $\pm$ 0.60
	DiffuseDef	7.34 $\pm$ 0.41	20.47 $\pm$ 1.63	35.85 $\pm$ 2.64
	DynAttn	6.45 $\pm$ 0.65	18.97 $\pm$ 0.65	35.22 $\pm$ 0.50
	Agatha	9.05 $\pm$ 1.20	22.68 $\pm$ 2.97	34.46 $\pm$ 4.26
CoLA	FreeLB	6.84 $\pm$ 5.41	9.45 $\pm$ 8.43	11.28 $\pm$ 10.98
	DiffuseDef	7.11 $\pm$ 2.20	8.88 $\pm$ 1.91	9.52 $\pm$ 2.95
	DynAttn	1.82 $\pm$ 0.50	2.36 $\pm$ 0.72	2.37 $\pm$ 0.78
	Agatha	11.93 $\pm$ 1.36	14.32 $\pm$ 2.08	17.75 $\pm$ 2.60
MRPC	FreeLB	9.01 $\pm$ 0.51	34.79 $\pm$ 5.40	54.17 $\pm$ 5.07
	DiffuseDef	4.43 $\pm$ 3.00	10.00 $\pm$ 5.84	19.65 $\pm$ 13.25
	DynAttn	2.87 $\pm$ 1.55	7.25 $\pm$ 4.97	15.03 $\pm$ 11.19
	Agatha	0.35 $\pm$ 0.35	1.30 $\pm$ 0.54	1.53 $\pm$ 0.20
QQP	FreeLB	18.18 $\pm$ 0.88	30.95 $\pm$ 0.29	35.76 $\pm$ 0.53
	DiffuseDef	13.39 $\pm$ 1.48	29.26 $\pm$ 0.56	41.42 $\pm$ 1.73
	DynAttn	14.84 $\pm$ 1.40	29.59 $\pm$ 0.98	40.54 $\pm$ 1.95
	Agatha	13.43 $\pm$ 2.08	22.10 $\pm$ 2.41	28.89 $\pm$ 1.96
MNLI	FreeLB	10.08 $\pm$ 0.56	25.71 $\pm$ 1.08	35.39 $\pm$ 1.11
	DiffuseDef	4.02 $\pm$ 0.95	23.45 $\pm$ 1.39	36.11 $\pm$ 0.88
	DynAttn	13.10 $\pm$ 0.18	24.91 $\pm$ 1.01	37.71 $\pm$ 1.13
	Agatha	9.64 $\pm$ 0.49	13.62 $\pm$ 0.34	17.93 $\pm$ 0.81
QNLI	FreeLB	10.00 $\pm$ 1.80	25.35 $\pm$ 3.93	41.07 $\pm$ 2.25
	DiffuseDef	4.50 $\pm$ 2.14	11.42 $\pm$ 7.36	19.43 $\pm$ 13.32
	DynAttn	8.05 $\pm$ 0.22	14.08 $\pm$ 1.62	25.67 $\pm$ 9.26
	Agatha	7.44 $\pm$ 2.68	9.01 $\pm$ 3.09	11.02 $\pm$ 4.48
RTE	FreeLB	12.97 $\pm$ 3.06	23.77 $\pm$ 9.19	30.76 $\pm$ 11.17
	DiffuseDef	7.09 $\pm$ 1.92	15.84 $\pm$ 1.98	20.06 $\pm$ 5.86
	DynAttn	9.78 $\pm$ 1.44	18.70 $\pm$ 3.02	31.87 $\pm$ 6.26
	Agatha	4.80 $\pm$ 1.57	12.24 $\pm$ 2.19	16.17 $\pm$ 4.52
WNLI	FreeLB	6.26 $\pm$ 6.81	26.21 $\pm$ 11.07	38.19 $\pm$ 10.32
	DiffuseDef	4.73 $\pm$ 4.65	10.15 $\pm$ 13.58	15.74 $\pm$ 18.59
	DynAttn	7.42 $\pm$ 1.99	30.03 $\pm$ 11.16	38.69 $\pm$ 8.15
	Agatha	1.65 $\pm$ 1.43	1.65 $\pm$ 1.43	0.83 $\pm$ 1.44
Combined GLUE (8 tasks)				
	FreeLB	10.17 $\pm$ 1.18	25.29 $\pm$ 2.25	36.40 $\pm$ 2.45
	DiffuseDef	7.20 $\pm$ 0.86	16.18 $\pm$ 2.11	24.72 $\pm$ 3.43
	DynAttn	8.04 $\pm$ 0.42	18.24 $\pm$ 1.60	28.39 $\pm$ 2.24
	Agatha	7.29 $\pm$ 0.55	12.12 $\pm$ 0.75	16.07 $\pm$ 1.06

Table 34: CharSwap results per task. Perturbation rate indicates the fraction of words with swapped characters. **Bold**: lowest flip rate per task and perturbation rate.

Agatha recovers at higher perturbation rates. This is notable because CharSwap operates at the discrete character level, which Agatha was not designed to defend against. The robustness likely arises because character perturbations propagate through the tokenizer into embedding-space changes, which the energy basins then absorb.

C.4 TextBugger (Discrete Text-Level)

Table 35 reports per-task TextBugger results, which apply character-level perturbations including insertion, deletion, swap, and homoglyph substitution to input tokens (Li et al., 2018).

Task	Model	Flip Rate (%)		
		10%	30%	50%
SST-2	FreeLB	8.42 $\pm$ 0.28	21.53 $\pm$ 1.04	36.80 $\pm$ 0.16
	DiffuseDef	7.30 $\pm$ 0.36	16.98 $\pm$ 0.11	27.29 $\pm$ 2.94
	DynAttn	6.54 $\pm$ 0.65	16.86 $\pm$ 0.73	27.17 $\pm$ 0.70
	Agatha	8.67 $\pm$ 1.28	19.73 $\pm$ 3.79	31.55 $\pm$ 4.81
CoLA	FreeLB	6.05 $\pm$ 4.44	7.96 $\pm$ 5.85	10.27 $\pm$ 9.72
	DiffuseDef	6.10 $\pm$ 1.03	7.40 $\pm$ 1.86	9.62 $\pm$ 3.03
	DynAttn	1.77 $\pm$ 0.57	2.04 $\pm$ 0.71	2.37 $\pm$ 0.73
	Agatha	10.87 $\pm$ 1.49	12.38 $\pm$ 0.96	15.70 $\pm$ 1.88
MRPC	FreeLB	8.90 $\pm$ 0.39	27.58 $\pm$ 5.35	44.36 $\pm$ 7.17
	DiffuseDef	2.17 $\pm$ 0.54	9.67 $\pm$ 5.48	16.81 $\pm$ 11.90
	DynAttn	2.19 $\pm$ 1.39	6.57 $\pm$ 4.18	9.83 $\pm$ 7.35
	Agatha	0.35 $\pm$ 0.35	0.94 $\pm$ 0.54	1.53 $\pm$ 0.20
QQP	FreeLB	14.23 $\pm$ 0.76	26.42 $\pm$ 0.46	34.48 $\pm$ 0.43
	DiffuseDef	10.01 $\pm$ 1.18	22.32 $\pm$ 0.63	36.44 $\pm$ 1.28
	DynAttn	11.99 $\pm$ 1.31	23.35 $\pm$ 1.01	36.34 $\pm$ 1.31
	Agatha	10.60 $\pm$ 1.71	17.71 $\pm$ 2.20	25.52 $\pm$ 2.11
MNLI	FreeLB	8.73 $\pm$ 0.30	20.45 $\pm$ 1.57	30.40 $\pm$ 1.64
	DiffuseDef	7.68 $\pm$ 0.73	18.50 $\pm$ 1.03	30.08 $\pm$ 0.62
	DynAttn	12.16 $\pm$ 0.30	20.48 $\pm$ 0.74	30.94 $\pm$ 1.19
	Agatha	9.51 $\pm$ 0.28	12.89 $\pm$ 0.29	17.27 $\pm$ 0.56
QNLI	FreeLB	7.91 $\pm$ 1.79	20.33 $\pm$ 4.51	35.23 $\pm$ 4.84
	DiffuseDef	3.90 $\pm$ 1.41	9.35 $\pm$ 4.51	16.52 $\pm$ 9.33
	DynAttn	7.10 $\pm$ 0.08	11.18 $\pm$ 0.93	19.23 $\pm$ 5.31
	Agatha	6.11 $\pm$ 2.25	8.37 $\pm$ 2.97	10.82 $\pm$ 4.42
RTE	FreeLB	9.99 $\pm$ 4.74	20.57 $\pm$ 6.84	26.94 $\pm$ 6.75
	DiffuseDef	6.28 $\pm$ 2.01	12.72 $\pm$ 1.96	19.22 $\pm$ 3.91
	DynAttn	8.68 $\pm$ 1.41	16.36 $\pm$ 3.28	26.10 $\pm$ 8.04
	Agatha	4.80 $\pm$ 0.92	10.28 $\pm$ 1.57	14.42 $\pm$ 3.46
WNLI	FreeLB	4.30 $\pm$ 3.99	18.17 $\pm$ 3.76	30.30 $\pm$ 6.93
	DiffuseDef	3.14 $\pm$ 3.54	8.60 $\pm$ 10.95	13.33 $\pm$ 16.69
	DynAttn	9.04 $\pm$ 5.40	21.39 $\pm$ 9.14	28.69 $\pm$ 12.58
	Agatha	1.65 $\pm$ 1.43	0.00 $\pm$ 0.00	1.65 $\pm$ 1.43
Combined GLUE (8 tasks)				
	FreeLB	8.57 $\pm$ 0.99	20.38 $\pm$ 1.52	31.10 $\pm$ 2.04
	DiffuseDef	5.82 $\pm$ 0.59	13.19 $\pm$ 1.67	21.16 $\pm$ 2.91
	DynAttn	7.43 $\pm$ 0.75	14.78 $\pm$ 1.34	22.58 $\pm$ 2.20
	Agatha	6.57 $\pm$ 0.48	10.29 $\pm$ 0.71	14.81 $\pm$ 1.01

Table 35: TextBugger results per task. Perturbation rate indicates the fraction of words perturbed. **Bold** indicates lowest flip rate per task and perturbation rate.

The TextBugger pattern closely mirrors CharSwap. Agatha achieves the lowest combined flip rate at all perturbation levels, with particularly strong performance on MRPC (0.94 versus 6.57 for DynAttn at 30%) and QNLI (8.37 versus 9.35 for DiffuseDef). On SST-2 and CoLA, DynAttn again records lower flip rates than all other defenses. On MNLI at 10%, DiffuseDef leads (7.68), but Agatha recovers at higher perturbation rates.

C.5 AdvGLUE (Curated Text-Level Attacks)

AdvGLUE provides curated adversarial examples combining word substitutions, character perturbations, paraphrase rewrites, and distraction insertion. Table 36 reports per-task results on the 5 available AdvGLUE evaluation tasks.

Agatha achieves the highest combined AdvGLUE score (44.81), followed by DynAttn (44.52), DiffuseDef (43.21), and FreeLB (41.71). No single defense dominates across all tasks: Agatha wins on adv_rte (69.55), while DynAttn leads on adv_qqp,

Task	Model	AdvGLUE ($\uparrow$)
adv_sst2	FreeLB	47.07 $\pm$ 2.06
	DiffuseDef	40.34 $\pm$ 1.03
	DynAttn	42.57 $\pm$ 1.17
	Agatha	46.40 $\pm$ 5.16
adv_qqp	FreeLB	28.83 $\pm$ 4.69
	DiffuseDef	44.46 $\pm$ 3.67
	DynAttn	53.76 $\pm$ 2.71
	Agatha	29.57 $\pm$ 5.88
adv_qnli	FreeLB	40.99 $\pm$ 1.56
	DiffuseDef	43.40 $\pm$ 2.73
	DynAttn	43.02 $\pm$ 2.81
	Agatha	41.89 $\pm$ 4.05
adv_rte	FreeLB	57.20 $\pm$ 16.11
	DiffuseDef	50.38 $\pm$ 6.09
	DynAttn	50.21 $\pm$ 3.97
	Agatha	69.55 $\pm$ 5.57
adv_mnli	FreeLB	34.44 $\pm$ 3.91
	DiffuseDef	37.47 $\pm$ 2.66
	DynAttn	33.06 $\pm$ 1.65
	Agatha	36.64 $\pm$ 1.26
<i>Combined (5 tasks)</i>		
	FreeLB	41.71
	DiffuseDef	43.21
	DynAttn	44.52
	Agatha	44.81

Table 36: AdvGLUE results. Higher score indicates better robustness. **Bold**: highest score per task.

DiffuseDef on adv_qnli and adv_mnli, and FreeLB on adv_sst2. These results are notable because AdvGLUE attacks involve complex perturbation strategies including paraphrase rewrites and distraction insertion, going beyond simple character or word substitutions.

D Hopfield-Specific Attacks

We evaluate four attacks designed to directly target Agatha’s theoretical defense mechanism. All attacks use 20-step PGD with ℓ_2 norm. Results report mean $\pm$ std across three random seeds.

Attack Descriptions.

- **Standard PGD**: cross-entropy loss only (reference baseline).
- **Basin Escape**: maximizes the Euclidean distance between clean and perturbed Hopfield fixed points, attempting to push representations out of their energy basins.
- **Spectral Radius**: directly maximizes $\rho(\mathcal{J}_F)$ to break the contraction condition $g + \sigma_{\max} < 1$ from Theorem 1.
- **Convergence Disruption**: maximizes the Hopfield convergence residual at each iteration, attempting to prevent the iterative mechanism from reaching a stable fixed point.

- **Joint (CE + ρ)**: combines cross-entropy with 0.5 $\cdot$ spectral radius, attacking both prediction and contraction simultaneously.

D.1 Combined Results

Table 37 reports combined GLUE ASR across all budgets, Table 38 reports Hopfield iteration count changes, and Table 39 reports spectral radius across all 8 tasks at five perturbation budgets.

Attack	Attack Success Rate (%)				
	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 2.0$	$\epsilon = 3.0$	$\epsilon = 5.0$
Standard PGD	12.75 $\pm$ 1.32	26.16 $\pm$ 2.74	50.28 $\pm$ 3.13	65.58 $\pm$ 3.19	82.03 $\pm$ 2.18
Basin Escape	1.49 $\pm$ 0.22	2.36 $\pm$ 0.33	4.00 $\pm$ 0.60	5.12 $\pm$ 0.64	7.58 $\pm$ 1.02
Spectral Radius	1.28 $\pm$ 0.34	2.03 $\pm$ 0.39	3.02 $\pm$ 0.55	4.29 $\pm$ 0.77	6.22 $\pm$ 1.26
Conv. Disruption	1.71 $\pm$ 0.35	2.41 $\pm$ 0.42	3.83 $\pm$ 0.55	4.98 $\pm$ 1.04	6.69 $\pm$ 1.18
Joint (CE + ρ)	11.11 $\pm$ 1.21	22.45 $\pm$ 2.21	43.82 $\pm$ 3.50	58.60 $\pm$ 3.81	75.61 $\pm$ 3.35

Table 37: Combined GLUE ASR for Hopfield-specific attacks. **Bold** indicates lowest ASR among the three theory-aware attacks (excluding Standard PGD and Joint). No theory-aware attack approaches Standard PGD performance at any budget.

Attack	Hopfield Iterations (before $\rightarrow$ after)				
	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 2.0$	$\epsilon = 3.0$	$\epsilon = 5.0$
Standard PGD	24 $\rightarrow$ 26	24 $\rightarrow$ 26	24 $\rightarrow$ 29	24 $\rightarrow$ 29	24 $\rightarrow$ 29
Basin Escape	24 $\rightarrow$ 26	24 $\rightarrow$ 25	24 $\rightarrow$ 27	24 $\rightarrow$ 29	24 $\rightarrow$ 32
Spectral Radius	24 $\rightarrow$ 22	24 $\rightarrow$ 20	24 $\rightarrow$ 20	24 $\rightarrow$ 19	24 $\rightarrow$ 22
Conv. Disruption	24 $\rightarrow$ 39	24 $\rightarrow$ 46	24 $\rightarrow$ 48	24 $\rightarrow$ 43	24 $\rightarrow$ 39
Joint (CE + ρ)	24 $\rightarrow$ 23	24 $\rightarrow$ 23	24 $\rightarrow$ 23	24 $\rightarrow$ 24	24 $\rightarrow$ 25

Table 38: Hopfield iteration count before and after each attack. Convergence Disruption nearly doubles the iteration count but achieves under 4% ASR at all budgets, confirming that delayed convergence alone does not translate into prediction flips.

D.2 Per-Task Results

Table 40 reports per-task ASR at $\epsilon = 2.0$.

D.3 Analysis

Three findings emerge from this analysis.

No theory-aware attack outperforms standard PGD. On all 8 tasks at $\epsilon = 2.0$, Standard PGD achieves higher ASR than every Hopfield-specific attack. The three pure theory-aware attacks (Basin Escape, Spectral Radius, Convergence Disruption) achieve at most 4% combined ASR compared to 50.28% for Standard PGD. Attacking the theoretical mechanism is less effective than attacking the prediction output directly.

The spectral radius is structurally constrained. Even the dedicated Spectral Radius attack, which directly maximizes $\rho(\mathcal{J}_F)$, shifts the spectral radius

Attack	Spectral Radius (ρ before $\rightarrow$ ρ after)				
	$\varepsilon = 0.5$	$\varepsilon = 1.0$	$\varepsilon = 2.0$	$\varepsilon = 3.0$	$\varepsilon = 5.0$
Standard PGD	0.721 $\rightarrow$ 0.720	0.721 $\rightarrow$ 0.720	0.721 $\rightarrow$ 0.722	0.720 $\rightarrow$ 0.723	0.720 $\rightarrow$ 0.726
Basin Escape	0.721 $\rightarrow$ 0.720	0.719 $\rightarrow$ 0.720	0.721 $\rightarrow$ 0.725	0.720 $\rightarrow$ 0.724	0.719 $\rightarrow$ 0.737
Spectral Radius	0.720 $\rightarrow$ 0.726	0.721 $\rightarrow$ 0.733	0.719 $\rightarrow$ 0.748	0.720 $\rightarrow$ 0.765	0.720 $\rightarrow$ 0.812
Conv. Disruption	0.720 $\rightarrow$ 0.752	0.720 $\rightarrow$ 0.728	0.721 $\rightarrow$ 0.750	0.721 $\rightarrow$ 0.732	0.722 $\rightarrow$ 0.718
Joint (CE + ρ)	0.720 $\rightarrow$ 0.721	0.719 $\rightarrow$ 0.723	0.721 $\rightarrow$ 0.725	0.721 $\rightarrow$ 0.728	0.720 $\rightarrow$ 0.735

Table 39: Spectral radius before and after attack. Even the dedicated Spectral Radius attack increases ρ by at most 0.09 (at $\varepsilon = 5.0$), insufficient to break the contraction condition.

Task	Standard PGD	Basin Escape	Spectral Radius	Conv. Disrupt.	Joint (CE + ρ)
SST-2	45.99 $\pm$ 3.41	3.05 $\pm$ 0.42	3.86 $\pm$ 0.42	2.67 $\pm$ 0.17	43.99 $\pm$ 4.60
CoLA	56.45 $\pm$ 2.90	1.89 $\pm$ 1.25	2.04 $\pm$ 1.26	1.80 $\pm$ 0.88	55.61 $\pm$ 3.50
MRPC	4.48 $\pm$ 0.53	0.12 $\pm$ 0.20	0.00 $\pm$ 0.00	0.24 $\pm$ 0.20	2.71 $\pm$ 1.07
QQP	31.26 $\pm$ 8.67	4.50 $\pm$ 2.50	2.48 $\pm$ 0.99	4.28 $\pm$ 3.25	30.24 $\pm$ 8.53
MNLI	62.55 $\pm$ 0.54	11.75 $\pm$ 0.40	5.91 $\pm$ 0.50	10.44 $\pm$ 0.53	57.19 $\pm$ 0.43
QNLI	57.51 $\pm$ 0.12	5.20 $\pm$ 1.59	5.06 $\pm$ 2.65	3.20 $\pm$ 1.00	51.27 $\pm$ 4.78
RTE	58.11 $\pm$ 21.57	3.04 $\pm$ 2.59	1.54 $\pm$ 1.03	7.20 $\pm$ 2.15	45.13 $\pm$ 24.65
WNLI	85.87 $\pm$ 8.11	2.46 $\pm$ 2.44	3.29 $\pm$ 2.85	0.81 $\pm$ 1.41	64.41 $\pm$ 6.64
Combined	50.28 $\pm$ 3.13	4.00 $\pm$ 0.60	3.02 $\pm$ 0.55	3.83 $\pm$ 0.55	43.82 $\pm$ 3.50

Table 40: Per-task ASR at $\varepsilon = 2.0$ for Hopfield-specific attacks. **Bold** indicates lowest ASR among the three theory-aware attacks. Standard PGD outperforms all theory-aware attacks on every task.

by at most 0.23 on MNLI at $\varepsilon = 2.0$ (from 3.02 to 3.25). On tasks with $g = 0.1$ (CoLA, SST-2, QQP, QNLI), the spectral radius remains below 0.15 under all attacks at all budgets, far below the $1 - g = 0.9$ threshold needed to break contraction. This provides direct empirical evidence that the contraction condition from Theorem 1 holds under sustained adversarial pressure.

Convergence disruption does not translate into prediction flips. Convergence Disruption forces the Hopfield mechanism to use up to 72 iterations on MRPC (from a clean baseline of 24), yet achieves only 0.24% ASR on that task. The attack successfully delays convergence but cannot prevent it: the iterative mechanism eventually reaches a fixed point within the correct energy basin. This validates the $T_{\max} = 100$ setting used throughout, which provides sufficient headroom for convergence even under sustained adversarial conditions.

Joint attack is counterproductive. The Joint attack (CE + $0.5 \cdot \rho$) achieves only 43.82% combined ASR, 6.5 points below Standard PGD. Diverting gradient budget toward the spectral radius objective weakens the primary cross-entropy signal without meaningfully disrupting the underlying defense mechanism, confirming that the two objectives are antagonistic from the attacker’s perspective.

E Diagnosis of Performance Gap: A QNLI Case Study

E.1 Clean Accuracy and ASR

On QNLI, FreeLB achieves 80.36 clean accuracy, approximately 20 points above all architectural defenses (60 to 65 range). This gap is unique to QNLI and does not appear on any other GLUE task, raising the question of whether FreeLB’s advantage reflects a genuine property of training-based defenses or an anomaly specific to this task. To test this, we include PGD-AT (Madry et al., 2017), a second training-based defense not used elsewhere in the paper. Table 41 reports the results.

Category	Model	Clean Acc ($\uparrow$)	ASR $\varepsilon = 2.0$ ($\downarrow$)
Training-Based	PGD-AT	80.66 $\pm$ 0.66	74.26 $\pm$ 5.99
	FreeLB	80.36 $\pm$ 0.69	68.79 $\pm$ 6.98
Undeended	TinyBERT	65.10 $\pm$ 0.61	69.54 $\pm$ 4.35
Architectural	DiffuseDef	64.92 $\pm$ 2.04	72.32 $\pm$ 3.22
	DynAttn	65.49 $\pm$ 0.60	62.50 $\pm$ 12.47
	Agatha	60.82 $\pm$ 0.48	57.51 $\pm$ 0.12

Table 41: QNLI clean accuracy and ASR at $\varepsilon = 2.0$. Training-based defenses achieve ~ 20 points higher accuracy than architectural defenses. **Bold:** lowest ASR.

PGD-AT confirms the pattern: it achieves 80.66 clean accuracy, nearly identical to FreeLB (80.36), while all architectural defenses cluster in the 60 to 65 range. This is not a FreeLB anomaly but a systematic property of training-based defenses on QNLI. The ~ 20 point clean accuracy gap co-exists with worse robustness: PGD-AT records 74.26 ASR, higher than both Agatha (57.51) and DynAttn (62.50). We investigate the source of this gap through three analyses: adversarial accuracy, prediction confidence, and convergence behavior.

E.2 Adversarial Accuracy

ASR measures the fraction of *correct* predictions flipped by an attack. When models differ in clean accuracy, ASR comparisons can be misleading: a model with higher clean accuracy defends a larger set of predictions that may include many low-confidence, easily flipped examples. To account

for this, we report *adversarial accuracy*, defined as $\text{clean_acc} \times (1 - \text{ASR})$, which measures the fraction of all examples that a model both classifies correctly and successfully defends.

Figure 5 reports adversarial accuracy on QNLI across perturbation budgets.

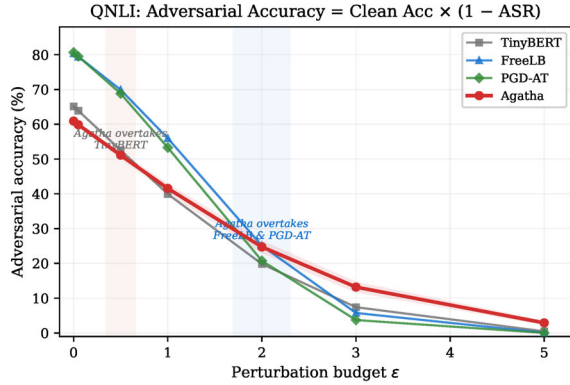


Figure 5: QNLI adversarial accuracy ($\text{clean_acc} \times (1 - \text{ASR})$) across all perturbation budgets.

Despite the 20 point clean accuracy gap, Agatha overtakes TinyBERT by $\epsilon = 1.0$, recording an adversarial accuracy of 41.55 compared to 39.95. Against FreeLB and PGD-AT, Agatha approaches parity near $\epsilon = 2.0$ (24.74 versus 25.08 for FreeLB and 20.76 for PGD-AT). At $\epsilon = 3.0$, all models converge below 14% adversarial accuracy, indicating that QNLI at this model scale is a difficult task for all architectures under strong perturbation.

The rapid convergence of adversarial accuracy across models indicates that the baselines’ higher clean accuracy consists largely of predictions that are easily flipped.

E.3 Prediction Confidence

To understand whether the clean accuracy gap reflects a difference in representation quality, we examine the prediction confidence of both Agatha and TinyBERT on QNLI. For each model, we collect the maximum softmax probability on every validation example and compare distributions for correctly and incorrectly classified examples.

Figure 6 reports the results. The two models exhibit nearly identical confidence distributions on their correct predictions: Agatha records a mean confidence of 0.706 (median 0.700), while TinyBERT records 0.705 (median 0.695). For both models, approximately 50% of correct predictions fall below 0.7 confidence, and roughly 25% fall below 0.6. These values are close to the chance level of

0.5 for binary classification, indicating that both models produce low-confidence representations on a considerable fraction of QNLI examples.

The confidence gap between correct and incorrect predictions is small for both models: 0.039 for Agatha and 0.061 for TinyBERT. Neither model has learned to separate its confident knowledge from its uncertain guesses on this task. TinyBERT’s 254 additional correct predictions (3603 versus 3349 for Agatha) follow the same flat confidence profile as the remainder, explaining why they are readily flipped under perturbation and why the adversarial accuracy gap closes quickly.

This analysis demonstrates that QNLI at the 4.37M parameter scale produces shallow representations regardless of architecture. The clean accuracy gap between Agatha and the baselines reflects **a difference in the number of low-confidence correct predictions, not in representation quality.**

E.4 Convergence Behavior

To connect the representation quality analysis to the energy basin mechanism, we compare the Hopfield convergence behavior of Agatha on QNLI against MRPC, a task where Agatha achieves strong robustness (ASR 4.48 at $\epsilon = 2.0$). For each task, we track the average number of Hopfield iterations during clean and adversarial evaluation, and report the iteration gap (adversarial iterations minus clean iterations) across perturbation budgets. Figure 7 reveals a qualitative difference in how the energy basins respond to perturbation across the two tasks.

Basin Depth. On MRPC, clean inputs converge in an average of 45.5 iterations, while on QNLI, convergence requires only 36.8 iterations. Fewer iterations indicate that QNLI representations settle into shallower basins, consistent with the low prediction confidence observed in Section E.3.

Basin Resistance. On MRPC, adversarial inputs require *more* iterations than clean inputs, with a positive gap that reaches +4.36 at $\epsilon = 2.0$ (Figure 7b). This indicates that the energy basins actively resist perturbation: the adversarial input is pulled back toward the basin floor, requiring additional iterations to converge. On QNLI, the iteration gap is *negative* at all budgets, reaching -6.74 at $\epsilon = 5.0$. Adversarial inputs converge faster than clean inputs, indicating that perturbations collapse representations toward attractors with less resistance rather than being opposed by basin geometry.

QNLI: Prediction Confidence — Agatha vs TinyBERT

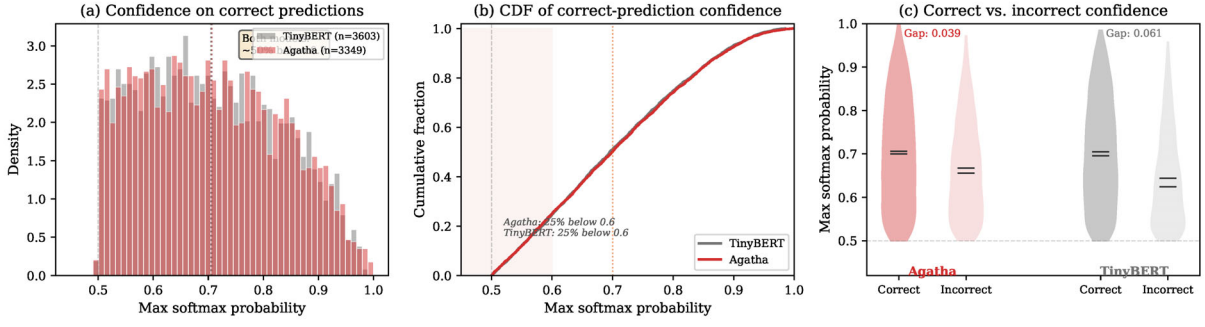


Figure 6: QNLI prediction confidence for Agatha and TinyBERT. (a) Both models show nearly identical confidence distributions on correct predictions (means: 0.706 and 0.705). (b) CDF confirms that $\sim 50\%$ of correct predictions fall below 0.7 for both models. (c) The confidence gap between correct and incorrect predictions is small for both models (0.039 and 0.061), indicating poor separation.

Agatha: QNLI vs MRPC — Energy Basin Formation and Robustness

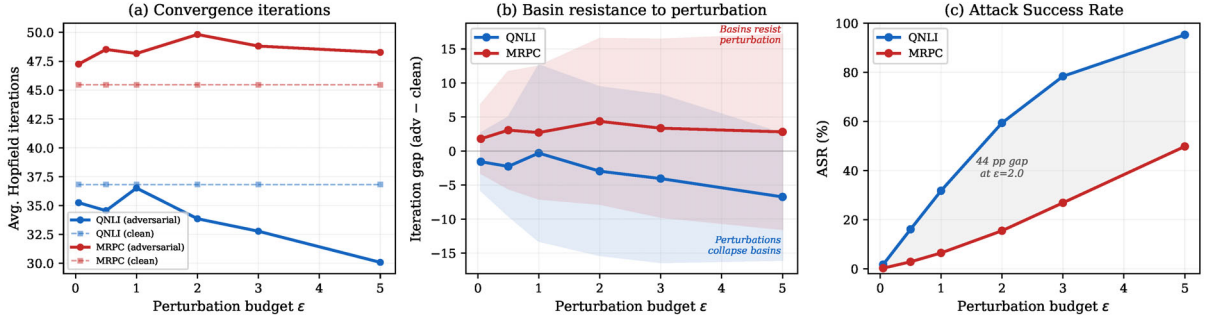


Figure 7: Agatha convergence behavior on QNLI versus MRPC. (a) MRPC requires more iterations to converge (45.5 versus 36.8), indicating deeper basins. (b) The iteration gap (adversarial minus clean iterations) is positive for MRPC (basins resist perturbation) and negative for QNLI (perturbations collapse basins). (c) MRPC maintains low ASR across budgets while QNLI degrades quickly.

Robustness Correlation. The direction of the iteration gap correlates with robustness outcomes across tasks. MRPC, with its positive gap, achieves an ASR of 4.48% at $\epsilon = 2.0$. QNLI, with its negative gap, records 57.51% at the same budget. This suggests that the iteration gap can serve as a diagnostic signal for basin quality: a positive gap indicates well-formed basins that resist perturbation, while a negative gap indicates shallow basins that fail to oppose adversarial directions.

E.5 Summary

The QNLI analysis reveals that Agatha’s lower clean accuracy on this task is a representation quality limitation at the 4.37M parameter scale, not a failure of the robustness mechanism. Three findings support this conclusion.

First, the adversarial accuracy gap between Agatha and the baselines closes rapidly and inverts

by $\epsilon = 1.0$ versus TinyBERT, indicating baselines’ extra correct predictions are trivially flippable.

Second, both Agatha and TinyBERT produce nearly identical low-confidence predictions on QNLI, indicating a task-level representation bottleneck independent of architecture.

Third, the Hopfield convergence analysis reveals that QNLI forms shallower basins with negative iteration gaps under perturbation, in contrast to MRPC where deep basins resist adversarial inputs.

F Full Gate Ablations

F.1 Fixed Gate Results

Table 42 reports average GLUE score and ASR across all perturbation budgets for Agatha with fixed gate values. Each fixed g row uses the same gate across all 8 tasks with no task-specific tuning or adversarial evaluation. Results are reported as

Model	Attack Success Rate						
	Clean	Low	Medium		High	Extreme	
	($\epsilon = 0$)	($\epsilon = 0.05$)	($\epsilon = 0.5$)	($\epsilon = 1.0$)	($\epsilon = 2.0$)	($\epsilon = 3.0$)	($\epsilon = 5.0$)
<i>Standard Training</i>							
TinyBERT	64.28 $\pm$ 0.52	2.48 $\pm$ 0.16	21.37 $\pm$ 1.01	40.10 $\pm$ 1.62	72.27 $\pm$ 1.71	90.94 $\pm$ 0.96	99.75 $\pm$ 0.06
<i>Training-Based Robustness</i>							
FreeLB	67.19 $\pm$ 0.33	2.09 $\pm$ 0.26	18.99 $\pm$ 1.37	35.50 $\pm$ 1.91	66.68 $\pm$ 2.07	85.97 $\pm$ 1.01	98.12 $\pm$ 0.24
<i>Architectural Robustness: Baselines</i>							
DiffuseDef	66.22 $\pm$ 0.57	1.74 $\pm$ 0.38	18.95 $\pm$ 2.58	38.07 $\pm$ 4.59	69.85 $\pm$ 4.69	90.30 $\pm$ 1.84	99.57 $\pm$ 0.17
DynAttn	65.77 $\pm$ 0.62	2.55 $\pm$ 0.36	16.68 $\pm$ 1.01	33.37 $\pm$ 1.38	63.58 $\pm$ 2.53	84.58 $\pm$ 2.41	99.14 $\pm$ 0.39
<i>Architectural Robustness: Fixed Gate (no task-specific tuning; best in group bolded)</i>							
Agatha ($g = 0.1$)	64.53 $\pm$ 0.39	2.75 $\pm$ 0.83	22.13 $\pm$ 4.47	34.95 $\pm$ 4.10	57.25 $\pm$ 4.19	72.66 $\pm$ 5.33	90.94 $\pm$ 2.55
Agatha ($g = 0.3$)	63.78 $\pm$ 0.23	1.62 $\pm$ 0.40	16.28 $\pm$ 3.16	33.62 $\pm$ 6.32	59.60 $\pm$ 4.19	76.08 $\pm$ 4.79	90.13 $\pm$ 4.91
Agatha ($g = 0.5$)	63.84 $\pm$ 0.13	1.97 $\pm$ 1.08	14.93 $\pm$ 2.87	31.21 $\pm$ 3.49	54.96 $\pm$ 2.35	70.59 $\pm$ 2.00	85.94 $\pm$ 1.55
Agatha ($g = 0.7$)	63.82 $\pm$ 0.17	1.49 $\pm$ 0.27	13.21 $\pm$ 1.33	27.92 $\pm$ 2.75	54.14 $\pm$ 3.08	70.10 $\pm$ 2.75	85.26 $\pm$ 1.36
Agatha ($g = 0.9$)	63.95 $\pm$ 0.18	1.71 $\pm$ 0.54	17.34 $\pm$ 3.10	32.44 $\pm$ 4.60	57.29 $\pm$ 2.54	73.26 $\pm$ 1.77	87.83 $\pm$ 0.72
<i>Architectural Robustness: Per-Task Tuned Gate</i>							
Agatha (tuned)	63.91 $\pm$ 0.20	1.48 $\pm$ 0.24	12.75 $\pm$ 1.32	26.16 $\pm$ 2.74	50.28 $\pm$ 3.13	65.58 $\pm$ 3.19	82.03 $\pm$ 2.18

Table 42: Fixed gate evaluation: Average GLUE Score and ASR across all perturbation budgets. Each fixed g row applies the same gate value to all 8 tasks with no adversarial evaluation. **Bold** indicates best ASR among fixed gate values and highest clean score overall. All fixed gate configurations outperform all defended baselines from $\epsilon = 1.0$ onward. Clean performance remains stable across all gate values (63.78 to 64.53).

mean $\pm$ std across three random seeds.

All fixed gates outperform defended baselines from $\epsilon = 2.0$ onward. At $\epsilon = 2.0$, the weakest fixed gate ($g = 0.3$, ASR 59.60) still outperforms DynAttn (63.58), DiffuseDef (69.85), and FreeLB (66.68). At $\epsilon = 5.0$, the best fixed gate ($g = 0.7$, ASR 85.26) outperforms DynAttn (99.14) by nearly 14 points. At $\epsilon = 1.0$, all fixed gates outperform FreeLB and DiffuseDef, though $g \in \{0.1, 0.3\}$ slightly trail DynAttn (33.37). This confirms that the robustness is architectural: the combination of Hopfield attention and gating provides robustness across a wide range of gate values, rather than depending on a specific setting.

The optimal fixed gate is $g = 0.7$, not the most aggressive setting. The most aggressive gate ($g = 0.1$) is not the best overall default. $g = 0.7$ achieves the lowest ASR at $\epsilon \geq 0.5$ by balancing gradient contraction with representational capacity. This is consistent with the analysis in Section 1: tasks like paraphrase detection require the combined representation to retain enough lexical information to verify semantic alignments, and aggressive gating discards this information.

Per-task tuning provides consistent additional gains. The tuned configuration reduces ASR by nearly 4 points at $\epsilon = 2.0$ (50.28 versus 54.14 for the best fixed gate) and 3 points at $\epsilon = 5.0$ (82.03 versus 85.26). This improvement comes from allowing tasks with different representational demands to

use different optimal gate values, confirming that the gate modulates the degree of robustness rather than acting as a binary switch.

F.2 Per-Layer Gate Selection

Throughout the main paper, Agatha applies a single gate value g uniformly across all layers. A natural question is whether different layers benefit from different gate values. The first layer receives input representations directly from the embedding, where PGD perturbations are applied, while deeper layers operate on representations that have already passed through one round of Hopfield contraction. This section investigates whether *asymmetric* gating, where each layer receives its own gate value, improves robustness over the *symmetric* baseline.

F.2.1 Setup

We evaluate on SST-2 using the base architecture (128 dimensions, 2 layers, 2 heads, 4.37M parameters) with $\beta = 50$ and $T_{\max} = 100$. For each configuration, we train with learned gates initialized at (g_0, g_1) , where g_0 is the gate for layer 0 and g_1 for layer 1, applied to both the Hopfield and FFN sublayers. We sweep over a 5×5 grid of values $g_0, g_1 \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, yielding 5 symmetric configurations (diagonal) and 20 asymmetric configurations (off-diagonal). Results are reported as mean $\pm$ std across three random seeds.

$g_0 \backslash g_1$	ASR at $\varepsilon = 2.0$ (%) $\downarrow$				
	0.1	0.3	0.5	0.7	0.9
0.1	45.99 $\pm$ 3.41	41.81 $\pm$ 7.29	42.46 $\pm$ 1.66	50.07 $\pm$ 3.04	43.65 $\pm$ 2.34
0.3	51.95 $\pm$ 1.37	<u>49.77 $\pm$ 8.30</u>	46.32 $\pm$ 2.59	46.61 $\pm$ 3.72	53.47 $\pm$ 2.04
0.5	50.40 $\pm$ 3.78	52.36 $\pm$ 3.28	<u>49.97 $\pm$ 5.28</u>	47.20 $\pm$ 6.96	49.69 $\pm$ 5.10
0.7	52.65 $\pm$ 3.94	52.90 $\pm$ 2.72	47.12 $\pm$ 3.42	<u>55.96 $\pm$ 2.43</u>	52.56 $\pm$ 5.65
0.9	58.66 $\pm$ 1.08	55.24 $\pm$ 7.16	52.04 $\pm$ 4.77	52.81 $\pm$ 4.72	<u>58.07 $\pm$ 1.41</u>

Table 43: ASR at $\varepsilon = 2.0$ on SST-2 across all (g_0, g_1) configurations. Rows correspond to g_0 (layer 0), columns to g_1 (layer 1). Underlined: symmetric (diagonal) entries. **Bold**: best overall. Clean accuracy is stable across all configurations (79.89 to 81.46).

Rank	Config	Type	Clean	Attack Success Rate (%)				
				$\varepsilon = 0.5$	$\varepsilon = 1.0$	$\varepsilon = 2.0$	$\varepsilon = 3.0$	$\varepsilon = 5.0$
1	(0.1, 0.3)	Asym	80.35 $\pm$ 0.33	9.98 $\pm$ 1.68	20.02 $\pm$ 4.16	41.81 $\pm$ 7.29	64.68 $\pm$ 7.79	89.43 $\pm$ 4.65
2	(0.1, 0.5)	Asym	80.05 $\pm$ 0.47	8.98 $\pm$ 1.60	18.73 $\pm$ 1.42	42.46 $\pm$ 1.66	65.14 $\pm$ 2.38	89.98 $\pm$ 1.20
3	(0.1, 0.9)	Asym	80.58 $\pm$ 0.14	9.58 $\pm$ 0.76	19.59 $\pm$ 1.52	43.65 $\pm$ 2.34	65.42 $\pm$ 1.68	90.61 $\pm$ 1.61
4	(0.1, 0.1)	Sym	80.20 $\pm$ 0.18	10.53 $\pm$ 1.78	22.49 $\pm$ 2.97	45.99 $\pm$ 3.41	67.53 $\pm$ 6.06	90.61 $\pm$ 3.59
5	(0.3, 0.5)	Asym	80.54 $\pm$ 0.39	10.49 $\pm$ 1.18	21.69 $\pm$ 1.78	46.32 $\pm$ 2.59	70.06 $\pm$ 1.56	92.98 $\pm$ 0.99

Table 44: Top 5 gate configurations on SST-2 ranked by ASR at $\varepsilon = 2.0$. The three best configurations are asymmetric with $g_0 = 0.1$. The best symmetric configuration ($g = 0.1$) ranks fourth. **Bold** indicates best performance per column: highest clean accuracy and lowest ASR.

F.2.2 Results

Table 43 reports clean accuracy and ASR at $\varepsilon = 2.0$ for all configurations. Three observations emerge.

Asymmetric gating outperforms symmetric gating. The best configuration overall is $(g_0 = 0.1, g_1 = 0.3)$ with an ASR of 41.81, compared to 45.99 for the best symmetric baseline ($g = 0.1$), a reduction of 4.18 percentage points. The top three configurations by ASR are all asymmetric (Table 44). Clean accuracy remains stable across all 25 configurations, ranging from 79.89 to 81.46, indicating that asymmetric gating does not trade off clean performance for robustness.

Layer ordering matters. For a given pair of gate values, the order in which they are assigned to layers has a substantial effect. Table 45 reports this directionality for all 10 unique asymmetric pairs.

In 9 of 10 pairs, assigning the smaller gate to layer 0 (closer to the input embedding) yields lower ASR. The effect is largest for (0.1, 0.9) versus (0.9, 0.1): 43.65 compared to 58.66, a difference of 15.02 percentage points. The exception is (0.5, 0.7) where the difference is 0.08 points, within noise.

F.2.3 Interpretation

The directionality effect is consistent with Agatha’s gradient contraction mechanism. PGD computes perturbations at the input embedding (Eq. 2 in the main text), so the adversarial gradient must propagate backward through all layers to reach the embedding. From Eq. 8, the gradient scales as

$\prod_{l=1}^L (g_l \mathbf{I} + \mathcal{J}_F^{(l)})$. Layer 0, which is the last layer the gradient passes through before reaching the embedding, therefore has the most direct influence on the gradient magnitude available to PGD. A tight gate at layer 0 ensures that the gradient is strongly attenuated at this final stage, regardless of how much signal survives through deeper layers. A loose gate at layer 1, conversely, preserves representational capacity in the deeper layer without compromising the gradient contraction at the input boundary.

This interpretation also explains why the effect is largest when g_0 and g_1 differ substantially: the configuration (0.1, 0.9) concentrates nearly all contraction at the input boundary while preserving an almost standard residual in the deeper layer, creating a strong asymmetry in the gradient path.

F.2.4 Summary

Per-layer asymmetric gating provides a modest but consistent improvement over symmetric gating on SST-2, reducing ASR by up to 4.18 percentage points at $\varepsilon = 2.0$ without affecting clean accuracy. The pattern is directional: tighter gating on the layer closest to the input embedding is preferred, consistent with Agatha’s gradient contraction mechanism. These results suggest that tuning per-layer gate values, rather than a single shared gate, is a useful direction for maximizing robustness. We note that the symmetric gate ($g = 0.1$) already outperforms all defended baselines at $\varepsilon \geq 2.0$ (Section 5.12), so **asymmetric gating represents a**

Pair	ASR($g_{\text{low}}, g_{\text{high}}$)	ASR($g_{\text{high}}, g_{\text{low}}$)	Δ	Tighter g_0 wins?
(0.1, 0.3)	41.81 $\pm$ 7.29	51.95 $\pm$ 1.37	-10.14	✓
(0.1, 0.5)	42.46 $\pm$ 1.66	50.40 $\pm$ 3.78	-7.94	✓
(0.1, 0.7)	50.07 $\pm$ 3.04	52.65 $\pm$ 3.94	-2.57	✓
(0.1, 0.9)	43.65 $\pm$ 2.34	58.66 $\pm$ 1.08	-15.02	✓
(0.3, 0.5)	46.32 $\pm$ 2.59	52.36 $\pm$ 3.28	-6.04	✓
(0.3, 0.7)	46.61 $\pm$ 3.72	52.90 $\pm$ 2.72	-6.29	✓
(0.3, 0.9)	53.47 $\pm$ 2.04	55.24 $\pm$ 7.16	-1.77	✓
(0.5, 0.7)	47.20 $\pm$ 6.96	47.12 $\pm$ 3.42	+0.08	
(0.5, 0.9)	49.69 $\pm$ 5.10	52.04 $\pm$ 4.77	-2.35	✓
(0.7, 0.9)	52.56 $\pm$ 5.65	52.81 $\pm$ 4.72	-0.25	✓

Table 45: Directionality of asymmetric gating on SST-2 at $\varepsilon = 2.0$. g_{low} and g_{high} denote the smaller and larger gate in each pair. Δ is the ASR difference (negative means tighter g_0 is better).

Model	Clean	Attack Success Rate (%)				
		$\varepsilon = 0.5$	$\varepsilon = 1.0$	$\varepsilon = 2.0$	$\varepsilon = 3.0$	$\varepsilon = 5.0$
Hopfield Only ($g = 1.0$, fixed)	80.50 $\pm$ 0.43	12.97 $\pm$ 3.98	27.10 $\pm$ 5.99	55.63 $\pm$ 10.40	77.95 $\pm$ 9.74	95.47 $\pm$ 3.92
Hopfield w/ Learned Gate ($g = 1.0$, learned)	80.05 $\pm$ 0.57	11.17 $\pm$ 0.82	23.53 $\pm$ 2.24	52.66 $\pm$ 2.53	74.73 $\pm$ 1.90	95.41 $\pm$ 0.69
Agatha ($g = 0.1$, fixed)	80.20 $\pm$ 0.18	10.53 $\pm$ 1.78	22.49 $\pm$ 2.97	45.99 $\pm$ 3.41	67.53 $\pm$ 6.06	90.61 $\pm$ 3.59

Table 46: Learning from scratch on SST-2. *Hopfield w/ Learned Gate* initializes the gate at $g = 1.0$ and allows it to update during training. *Hopfield Only* uses the same $g = 1.0$ but keeps it fixed. Agatha uses a fixed gate of $g = 0.1$. The learned gate does not discover an aggressive setting, and the resulting robustness remains comparable to the fixed $g = 1.0$ baseline. **Bold**: highest clean accuracy and lowest ASR.

refinement rather than a requirement.

F.3 Learned Gate Results

Our primary analysis assumes that the gate is a fixed value chosen by the practitioner. A natural question is whether the gate can instead *discover an effective value on its own* during training when initialized at $g = 1.0$, the standard residual connection. We formulate $g = \sigma(\gamma)$, where σ is the sigmoid function and γ is a scalar. This allows the gate to be updated via standard backpropagation. We initialize $g_0 = g_1 = 1.0$ (the standard residual connection) and train on SST-2 with all other settings identical to the main results.

Table 46 compares the resulting model against Agatha with a fixed gate of $g = 0.1$ and the Hopfield Only ablation with a fixed gate of $g = 1.0$. The learned gate provides no meaningful robustness improvement over the fixed $g = 1.0$ baseline. At $\varepsilon = 2.0$, the learned gate achieves an ASR of 52.66, compared to 55.63 for Hopfield Only, a difference of 2.97 percentage points. By contrast, Agatha with a fixed gate of $g = 0.1$ achieves 45.99, a reduction of 6.67 points over the learned gate. Clean accuracy is comparable across all three configurations (80.05 to 80.50), confirming that the training objective provides no pressure to reduce the gate.

Inspecting the final gate values confirms that the optimizer does not move the gate from its initialization. When initialized at $g = 1.0$, all gates remain at 0.9999 across all three seeds and both sublayer

types. The training objective is clean cross-entropy, which contains no adversarial signal. A gate near 1.0 preserves the full residual path, and since the model never encounters adversarial perturbations during training, there is no gradient pressure to reduce the gate toward a more aggressive setting. The learned gate model therefore converges to a configuration functionally identical to Hopfield Only, and the 6.67 point ASR gap relative to Agatha at $\varepsilon = 2.0$ is entirely attributable to the initialization.

This result supports the claim that the robustness of Agatha is architectural rather than learned. The gate value is a design decision made before training, and clean training alone cannot discover that an aggressive gate improves adversarial robustness.

G Full Architecture Ablation

Table 47 extends the ablation study from Table 7 to all perturbation budgets.

Gate Only increases ASR relative to TinyBERT at every budget, recording 22.48 versus 21.37 at $\varepsilon = 0.5$ and 75.42 versus 72.27 at $\varepsilon = 2.0$. This confirms that gating the residual without a capable attention mechanism is counterproductive.

Hopfield Only provides meaningful robustness, reducing ASR from 40.10 to 33.52 at $\varepsilon = 1.0$ and from 72.27 to 57.93 at $\varepsilon = 2.0$, but consistently falls short of Agatha at all budgets. ET, which also uses iterative energy minimization but retains the unscaled residual, records 61.56 at $\varepsilon = 2.0$, worse than Hopfield Only (57.93). This confirms that iterative

Model	Attack Success Rate						
	Clean	Low	Medium		High	Extreme	
	($\epsilon = 0$)	($\epsilon = 0.05$)	($\epsilon = 0.5$)	($\epsilon = 1.0$)	($\epsilon = 2.0$)	($\epsilon = 3.0$)	($\epsilon = 5.0$)
<i>Baseline</i>							
TinyBERT	64.28 $\pm$ 0.52	2.48 $\pm$ 0.16	21.37 $\pm$ 1.01	40.10 $\pm$ 1.62	72.27 $\pm$ 1.71	90.94 $\pm$ 0.96	99.75 $\pm$ 0.06
<i>Single Component</i>							
Gate Only	63.93 $\pm$ 0.69	2.39 $\pm$ 0.37	22.48 $\pm$ 1.84	41.67 $\pm$ 2.88	75.42 $\pm$ 2.23	92.49 $\pm$ 1.78	99.81 $\pm$ 0.10
Hopfield Only	64.01 $\pm$ 0.20	2.00 $\pm$ 0.45	17.92 $\pm$ 2.62	33.52 $\pm$ 3.95	57.93 $\pm$ 2.73	73.87 $\pm$ 1.85	88.51 $\pm$ 1.54
ET	64.41 $\pm$ 1.07	3.56 $\pm$ 1.63	23.43 $\pm$ 3.36	37.70 $\pm$ 4.00	61.56 $\pm$ 5.67	75.25 $\pm$ 5.87	83.83 $\pm$ 5.88
<i>Alternative Combination</i>							
Hopfield+Reg	61.56 $\pm$ 0.65	1.83 $\pm$ 0.55	19.04 $\pm$ 2.34	36.81 $\pm$ 3.64	65.09 $\pm$ 3.30	81.06 $\pm$ 1.97	91.75 $\pm$ 1.72
<i>Full Model</i>							
Agatha (Ours)	63.91 $\pm$ 0.20	1.48 $\pm$ 0.24	12.75 $\pm$ 1.32	26.16 $\pm$ 2.74	50.28 $\pm$ 3.13	65.58 $\pm$ 3.19	82.03 $\pm$ 2.18

Table 47: Full architecture ablation results: Average GLUE Score and Attack Success Rate (ASR) across all perturbation budgets. *Gate Only*: gated residual with standard attention. *Hopfield Only*: iterative Hopfield attention with unscaled residual. *ET*: Energy Transformer (Hoover et al., 2023) with unscaled residual. *Hopfield+Reg*: Hopfield attention with a training-time regularizer (Podasca and Das, 2026). **Bold**: best performance: highest clean score and lowest ASR.

energy minimization alone is insufficient: without the gated residual, the gradient bypass through the identity path undermines the contraction mechanism regardless of the specific energy formulation.

Agatha shows a clear advantage even at $\epsilon = 0.5$ (12.75 versus 17.92 for Hopfield Only and 23.43 for ET), confirming that the synergistic interaction between the gate and the energy landscape is active across the full perturbation spectrum.

Hopfield+Reg performs worse than Hopfield Only at every budget, recording 36.81 versus 33.52 at $\epsilon = 1.0$ and 65.09 versus 57.93 at $\epsilon = 2.0$. The regularizer operates during training but does not structurally block the residual gradient path at inference, allowing adversarial signals to bypass the Hopfield mechanism entirely. This contrasts with Agatha, which maintains a 17.72 point advantage over TinyBERT at $\epsilon = 5.0$ (82.03 versus 99.75). Table 48 provides the per-task ablation results.

Summary. Four patterns emerge across tasks.

Gate and Hopfield are jointly necessary. Agatha achieves the lowest ASR on all 8 tasks from $\epsilon = 1.0$ onward, with the exception of WNLI at $\epsilon \geq 2.0$ where the small validation set (71 examples) produces high variance. This demonstrates that the synergy between the gate and the Hopfield mechanism is consistent across task types.

Gate Only is counterproductive. Gate Only degrades robustness relative to TinyBERT on most tasks, with the effect most pronounced on CoLA (92.04 versus 74.60 at $\epsilon = 2.0$) and QNLI (77.10 versus 69.54), confirming the main paper findings.

Task	Model	Attack Success Rate (%)			
		Clean	$\epsilon = 1.0$	$\epsilon = 2.0$	$\epsilon = 3.0$
SST-2	TinyBERT	81.77 $\pm$ 0.19	28.33 $\pm$ 1.01	60.17 $\pm$ 1.36	85.32 $\pm$ 0.66
	Gate Only	82.26 $\pm$ 0.24	29.18 $\pm$ 0.82	61.06 $\pm$ 0.63	86.99 $\pm$ 0.98
	Hopfield Only	80.50 $\pm$ 0.43	27.10 $\pm$ 5.99	55.63 $\pm$ 10.40	77.95 $\pm$ 9.74
	ET	73.28 $\pm$ 7.59	52.56 $\pm$ 12.40	81.71 $\pm$ 6.78	94.16 $\pm$ 2.09
	Hopfield+Reg	80.76 $\pm$ 0.57	38.37 $\pm$ 1.88	74.38 $\pm$ 1.43	92.16 $\pm$ 1.41
	Agatha	80.20 $\pm$ 0.18	22.49 $\pm$ 2.97	45.99 $\pm$ 3.41	67.53 $\pm$ 6.06
CoLA	TinyBERT	54.40 $\pm$ 0.91	34.82 $\pm$ 9.30	74.60 $\pm$ 8.21	95.69 $\pm$ 2.35
	Gate Only	54.99 $\pm$ 0.34	54.93 $\pm$ 8.74	92.04 $\pm$ 2.57	98.70 $\pm$ 0.78
	Hopfield Only	55.58 $\pm$ 0.59	37.29 $\pm$ 3.85	68.13 $\pm$ 5.09	89.79 $\pm$ 2.18
	ET	68.20 $\pm$ 0.67	30.36 $\pm$ 9.17	71.27 $\pm$ 14.21	94.52 $\pm$ 5.06
	Hopfield+Reg	50.00 $\pm$ 0.00	34.31 $\pm$ 15.87	67.22 $\pm$ 15.29	90.27 $\pm$ 6.88
	Agatha	56.14 $\pm$ 0.74	30.15 $\pm$ 1.87	56.45 $\pm$ 2.90	77.01 $\pm$ 3.36
MRPC	TinyBERT	76.99 $\pm$ 0.49	19.58 $\pm$ 4.98	48.52 $\pm$ 7.72	80.91 $\pm$ 6.41
	Gate Only	76.54 $\pm$ 0.64	16.88 $\pm$ 5.54	44.68 $\pm$ 14.83	76.04 $\pm$ 13.85
	Hopfield Only	75.94 $\pm$ 0.39	4.87 $\pm$ 3.38	9.41 $\pm$ 4.92	16.28 $\pm$ 6.90
	ET	70.59 $\pm$ 1.64	20.26 $\pm$ 14.61	39.43 $\pm$ 27.89	42.45 $\pm$ 31.21
	Hopfield+Reg	74.61 $\pm$ 0.31	6.83 $\pm$ 2.85	14.81 $\pm$ 4.51	23.44 $\pm$ 7.09
	Agatha	75.52 $\pm$ 0.20	1.41 $\pm$ 0.00	4.48 $\pm$ 0.53	10.37 $\pm$ 0.75
QQP	TinyBERT	71.73 $\pm$ 1.78	28.86 $\pm$ 1.17	60.12 $\pm$ 2.46	83.62 $\pm$ 1.93
	Gate Only	74.83 $\pm$ 1.45	32.03 $\pm$ 3.55	65.25 $\pm$ 5.13	88.48 $\pm$ 2.44
	Hopfield Only	71.75 $\pm$ 0.06	17.43 $\pm$ 0.61	37.03 $\pm$ 1.55	56.98 $\pm$ 1.95
	ET	76.49 $\pm$ 0.57	23.38 $\pm$ 1.00	50.41 $\pm$ 1.05	73.97 $\pm$ 1.59
	Hopfield+Reg	72.08 $\pm$ 0.35	22.71 $\pm$ 3.76	48.29 $\pm$ 7.99	71.11 $\pm$ 10.42
	Agatha	72.12 $\pm$ 0.51	14.84 $\pm$ 4.09	31.26 $\pm$ 8.67	47.79 $\pm$ 12.18
MNLI	TinyBERT	53.57 $\pm$ 0.44	49.86 $\pm$ 0.72	83.03 $\pm$ 0.26	95.94 $\pm$ 0.49
	Gate Only	48.46 $\pm$ 0.65	47.86 $\pm$ 1.40	82.40 $\pm$ 1.19	96.48 $\pm$ 0.48
	Hopfield Only	56.40 $\pm$ 0.18	36.66 $\pm$ 1.00	65.12 $\pm$ 1.05	81.93 $\pm$ 1.34
	ET	55.99 $\pm$ 1.50	37.67 $\pm$ 3.55	68.93 $\pm$ 4.89	87.46 $\pm$ 4.54
	Hopfield+Reg	52.85 $\pm$ 1.24	45.38 $\pm$ 1.46	75.96 $\pm$ 0.77	90.96 $\pm$ 0.83
	Agatha	54.64 $\pm$ 0.13	35.70 $\pm$ 0.70	62.55 $\pm$ 0.54	78.76 $\pm$ 0.78
QNLI	TinyBERT	65.10 $\pm$ 0.61	38.63 $\pm$ 2.39	69.54 $\pm$ 4.35	88.62 $\pm$ 2.31
	Gate Only	65.58 $\pm$ 0.20	44.33 $\pm$ 2.07	77.10 $\pm$ 2.96	93.26 $\pm$ 1.45
	Hopfield Only	60.78 $\pm$ 0.07	32.98 $\pm$ 2.03	61.02 $\pm$ 2.75	79.49 $\pm$ 2.78
	ET	55.04 $\pm$ 3.17	20.96 $\pm$ 14.45	31.63 $\pm$ 20.99	42.25 $\pm$ 26.38
	Hopfield+Reg	59.93 $\pm$ 0.32	36.59 $\pm$ 3.10	66.06 $\pm$ 4.03	83.98 $\pm$ 3.60
	Agatha	60.82 $\pm$ 0.48	30.02 $\pm$ 0.97	57.51 $\pm$ 0.12	77.13 $\pm$ 0.81
RTE	TinyBERT	57.64 $\pm$ 0.45	54.10 $\pm$ 4.31	88.95 $\pm$ 1.48	99.16 $\pm$ 0.78
	Gate Only	57.16 $\pm$ 0.45	53.87 $\pm$ 4.86	86.73 $\pm$ 4.05	100.00 $\pm$ 0.00
	Hopfield Only	55.72 $\pm$ 0.34	58.13 $\pm$ 8.72	83.83 $\pm$ 6.17	94.40 $\pm$ 3.87
	ET	55.60 $\pm$ 0.29	29.80 $\pm$ 13.54	54.64 $\pm$ 22.94	69.58 $\pm$ 21.94
	Hopfield+Reg	50.97 $\pm$ 2.39	55.56 $\pm$ 12.08	86.90 $\pm$ 7.31	99.04 $\pm$ 0.92
	Agatha	54.99 $\pm$ 0.91	35.09 $\pm$ 18.66	58.11 $\pm$ 21.57	70.21 $\pm$ 20.92
WNLI	TinyBERT	53.05 $\pm$ 3.51	66.62 $\pm$ 5.44	93.27 $\pm$ 5.59	98.23 $\pm$ 1.26
	Gate Only	51.64 $\pm$ 5.19	54.24 $\pm$ 19.48	94.08 $\pm$ 6.29	100.00 $\pm$ 0.00
	Hopfield Only	55.40 $\pm$ 1.33	53.68 $\pm$ 29.24	83.25 $\pm$ 16.44	94.17 $\pm$ 6.56
	ET	60.09 $\pm$ 0.66	86.58 $\pm$ 12.97	94.46 $\pm$ 6.26	97.64 $\pm$ 1.94
	Hopfield+Reg	51.27 $\pm$ 4.42	54.74 $\pm$ 20.29	87.08 $\pm$ 17.54	97.50 $\pm$ 5.00
	Agatha	56.81 $\pm$ 0.81	39.57 $\pm$ 10.16	85.87 $\pm$ 8.11	95.87 $\pm$ 3.81

Table 48: Per-task ablation results at $\epsilon \in \{1.0, 2.0, 3.0\}$. **Bold** indicates best performance: highest clean score and lowest ASR per task and budget.

ET confirms that energy minimization alone is insufficient. ET performs worse than Hopfield

Model	Clean	Attack Success Rate (%)					
		$\epsilon = 0.05$	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 2.0$	$\epsilon = 3.0$	$\epsilon = 5.0$
BERT-base	65.56 $\pm$ 0.20	3.27 $\pm$ 0.51	38.15 $\pm$ 0.76	72.29 $\pm$ 0.82	93.47 $\pm$ 0.49	98.94 $\pm$ 0.33	100.00 $\pm$ 0.05
FreeLB	72.12 $\pm$ 0.15	3.05 $\pm$ 0.53	28.34 $\pm$ 0.67	50.39 $\pm$ 0.65	76.52 $\pm$ 0.54	90.33 $\pm$ 0.32	97.42 $\pm$ 0.05
DiffuseDef	65.54 $\pm$ 0.48	2.95 $\pm$ 1.10	39.81 $\pm$ 0.62	67.94 $\pm$ 5.41	91.61 $\pm$ 0.76	98.20 $\pm$ 0.48	99.92 $\pm$ 0.03
DynAttn	65.09 $\pm$ 0.54	3.59 $\pm$ 1.68	36.85 $\pm$ 7.98	63.14 $\pm$ 7.92	90.02 $\pm$ 1.46	97.80 $\pm$ 0.63	99.95 $\pm$ 0.05
Agatha	64.36 $\pm$ 0.18	2.63 $\pm$ 0.20	12.00 $\pm$ 0.45	23.18 $\pm$ 0.45	43.29 $\pm$ 0.44	61.24 $\pm$ 0.31	83.44 $\pm$ 0.40

Table 49: BERT-base scale: Combined GLUE Score and ASR across 8 tasks. **Bold** indicates best performance.

Task	Model	Clean	Attack Success Rate (%)					
			$\epsilon = 0.05$	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 2.0$	$\epsilon = 3.0$	$\epsilon = 5.0$
SST-2	BERT-base	82.91 $\pm$ 0.61	1.66 $\pm$ 1.07	19.92 $\pm$ 2.83	42.32 $\pm$ 2.07	81.60 $\pm$ 1.52	97.79 $\pm$ 0.95	100.00 $\pm$ 0.16
	FreeLB	87.16 $\pm$ 0.31	1.05 $\pm$ 1.64	11.84 $\pm$ 2.23	26.18 $\pm$ 2.19	60.53 $\pm$ 1.89	86.18 $\pm$ 0.94	90.66 $\pm$ 0.17
	DiffuseDef	82.45 $\pm$ 0.92	1.48 $\pm$ 0.43	16.23 $\pm$ 0.23	37.18 $\pm$ 1.18	74.72 $\pm$ 1.40	94.76 $\pm$ 0.88	99.58 $\pm$ 0.00
	DynAttn	82.91 $\pm$ 0.61	1.94 $\pm$ 0.50	16.75 $\pm$ 1.89	38.38 $\pm$ 1.51	75.79 $\pm$ 2.32	95.61 $\pm$ 1.57	99.91 $\pm$ 0.16
	Agatha	79.82 $\pm$ 0.57	0.86 $\pm$ 0.50	8.62 $\pm$ 0.60	18.97 $\pm$ 0.19	41.24 $\pm$ 0.59	60.92 $\pm$ 0.91	88.65 $\pm$ 0.15
CoLA	BERT-base	55.91 $\pm$ 0.24	6.54 $\pm$ 1.06	69.26 $\pm$ 1.82	99.44 $\pm$ 2.50	100.00 $\pm$ 1.26	100.00 $\pm$ 0.81	100.00 $\pm$ 0.11
	FreeLB	59.42 $\pm$ 0.61	2.42 $\pm$ 1.63	30.11 $\pm$ 1.69	59.91 $\pm$ 1.99	93.19 $\pm$ 1.39	99.85 $\pm$ 0.93	100.00 $\pm$ 0.14
	DiffuseDef	54.05 $\pm$ 2.82	7.30 $\pm$ 7.52	63.50 $\pm$ 55.07	83.73 $\pm$ 28.19	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00
	DynAttn	52.09 $\pm$ 1.41	6.54 $\pm$ 11.34	41.17 $\pm$ 50.83	65.65 $\pm$ 56.87	95.38 $\pm$ 8.01	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00
	Agatha	55.49 $\pm$ 0.53	4.95 $\pm$ 0.14	28.05 $\pm$ 0.53	46.53 $\pm$ 0.11	77.23 $\pm$ 0.57	88.94 $\pm$ 0.70	96.53 $\pm$ 0.20
MRPC	BERT-base	75.64 $\pm$ 0.53	3.83 $\pm$ 1.43	26.48 $\pm$ 2.30	64.81 $\pm$ 2.32	99.65 $\pm$ 1.35	100.00 $\pm$ 0.87	100.00 $\pm$ 0.15
	FreeLB	73.79 $\pm$ 0.32	5.36 $\pm$ 1.46	66.43 $\pm$ 1.87	97.86 $\pm$ 1.95	100.00 $\pm$ 1.25	100.00 $\pm$ 0.96	100.00 $\pm$ 0.19
	DiffuseDef	75.54 $\pm$ 0.36	1.77 $\pm$ 1.55	27.71 $\pm$ 26.91	62.98 $\pm$ 32.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00
	DynAttn	74.66 $\pm$ 1.86	1.79 $\pm$ 1.67	16.73 $\pm$ 17.39	54.25 $\pm$ 27.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00
	Agatha	76.62 $\pm$ 0.39	0.34 $\pm$ 0.41	10.31 $\pm$ 0.68	27.49 $\pm$ 0.12	61.17 $\pm$ 0.32	88.32 $\pm$ 0.73	98.97 $\pm$ 0.18
QQP	BERT-base	75.81 $\pm$ 0.63	1.73 $\pm$ 1.19	17.87 $\pm$ 1.34	38.58 $\pm$ 1.60	75.79 $\pm$ 1.13	94.39 $\pm$ 0.86	99.96 $\pm$ 0.12
	FreeLB	87.77 $\pm$ 0.56	0.60 $\pm$ 1.21	6.55 $\pm$ 1.96	16.27 $\pm$ 2.13	43.37 $\pm$ 1.19	70.18 $\pm$ 0.95	90.24 $\pm$ 0.14
	DiffuseDef	74.99 $\pm$ 1.29	1.50 $\pm$ 0.12	16.60 $\pm$ 1.55	35.66 $\pm$ 3.61	71.21 $\pm$ 5.57	92.37 $\pm$ 3.72	99.79 $\pm$ 0.23
	DynAttn	75.25 $\pm$ 1.39	2.03 $\pm$ 0.29	15.58 $\pm$ 1.77	33.40 $\pm$ 4.29	67.82 $\pm$ 7.08	89.35 $\pm$ 4.67	99.68 $\pm$ 0.38
	Agatha	71.94 $\pm$ 0.53	0.58 $\pm$ 0.28	4.66 $\pm$ 0.64	9.47 $\pm$ 0.51	19.91 $\pm$ 0.24	31.07 $\pm$ 0.97	52.79 $\pm$ 0.10
MNLI	BERT-base	63.46 $\pm$ 0.64	3.37 $\pm$ 1.20	35.67 $\pm$ 1.34	67.39 $\pm$ 2.85	95.22 $\pm$ 1.63	99.68 $\pm$ 0.98	100.00 $\pm$ 0.18
	FreeLB	79.37 $\pm$ 0.20	1.04 $\pm$ 1.60	10.19 $\pm$ 2.23	22.91 $\pm$ 1.51	54.31 $\pm$ 1.41	80.95 $\pm$ 0.72	98.88 $\pm$ 0.11
	DiffuseDef	64.91 $\pm$ 0.30	2.52 $\pm$ 0.33	28.07 $\pm$ 2.12	58.38 $\pm$ 3.91	91.30 $\pm$ 1.31	98.85 $\pm$ 0.43	100.00 $\pm$ 0.00
	DynAttn	64.11 $\pm$ 1.50	3.32 $\pm$ 0.27	29.38 $\pm$ 1.51	59.24 $\pm$ 1.70	90.93 $\pm$ 1.97	98.59 $\pm$ 0.88	100.00 $\pm$ 0.00
	Agatha	57.90 $\pm$ 0.48	2.22 $\pm$ 0.30	12.86 $\pm$ 0.68	24.85 $\pm$ 0.40	45.98 $\pm$ 0.13	64.37 $\pm$ 0.93	85.34 $\pm$ 0.16
QNLI	BERT-base	61.74 $\pm$ 0.54	4.06 $\pm$ 1.80	35.49 $\pm$ 2.29	66.44 $\pm$ 2.66	95.46 $\pm$ 1.23	99.67 $\pm$ 0.95	100.00 $\pm$ 0.19
	FreeLB	83.91 $\pm$ 0.40	1.13 $\pm$ 1.20	11.04 $\pm$ 1.38	25.98 $\pm$ 1.09	60.78 $\pm$ 1.74	85.47 $\pm$ 0.74	99.61 $\pm$ 0.13
	DiffuseDef	62.03 $\pm$ 0.23	3.92 $\pm$ 0.28	37.79 $\pm$ 2.23	67.71 $\pm$ 3.94	95.66 $\pm$ 1.35	99.59 $\pm$ 0.26	100.00 $\pm$ 0.00
	DynAttn	61.99 $\pm$ 0.79	3.03 $\pm$ 0.58	28.13 $\pm$ 2.56	55.56 $\pm$ 5.12	90.26 $\pm$ 3.47	98.88 $\pm$ 0.75	100.00 $\pm$ 0.00
	Agatha	60.88 $\pm$ 0.62	3.91 $\pm$ 0.54	11.40 $\pm$ 0.39	21.86 $\pm$ 0.64	42.72 $\pm$ 0.70	60.40 $\pm$ 0.79	81.30 $\pm$ 0.16
RTE	BERT-base	52.71 $\pm$ 0.63	0.00 $\pm$ 1.90	5.48 $\pm$ 2.49	99.32 $\pm$ 1.86	100.00 $\pm$ 1.34	100.00 $\pm$ 0.97	100.00 $\pm$ 0.20
	FreeLB	53.43 $\pm$ 0.64	10.14 $\pm$ 1.90	63.51 $\pm$ 2.30	94.59 $\pm$ 1.51	100.00 $\pm$ 1.77	100.00 $\pm$ 0.91	100.00 $\pm$ 0.19
	DiffuseDef	54.03 $\pm$ 1.37	2.61 $\pm$ 3.43	71.84 $\pm$ 26.04	99.55 $\pm$ 0.79	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00
	DynAttn	54.33 $\pm$ 2.30	4.19 $\pm$ 5.93	52.20 $\pm$ 34.11	98.67 $\pm$ 0.06	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00
	Agatha	54.51 $\pm$ 0.51	3.31 $\pm$ 0.62	15.23 $\pm$ 0.40	26.49 $\pm$ 0.59	48.34 $\pm$ 0.15	71.52 $\pm$ 0.93	90.73 $\pm$ 0.14
WNLI	BERT-base	56.34 $\pm$ 0.55	5.00 $\pm$ 1.53	95.00 $\pm$ 2.21	100.00 $\pm$ 2.39	100.00 $\pm$ 1.58	100.00 $\pm$ 1.00	100.00 $\pm$ 0.00
	FreeLB	52.11 $\pm$ 0.14	2.70 $\pm$ 1.32	27.03 $\pm$ 1.04	59.46 $\pm$ 2.07	100.00 $\pm$ 1.40	100.00 $\pm$ 1.00	100.00 $\pm$ 0.00
	DiffuseDef	56.34 $\pm$ 1.41	2.48 $\pm$ 2.44	56.72 $\pm$ 38.43	98.33 $\pm$ 2.89	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00
	DynAttn	55.40 $\pm$ 1.63	5.88 $\pm$ 3.76	94.89 $\pm$ 2.60	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00
	Agatha	57.75 $\pm$ 0.31	4.88 $\pm$ 1.18	4.88 $\pm$ 3.29	9.76 $\pm$ 3.45	9.76 $\pm$ 3.33	24.39 $\pm$ 0.91	73.17 $\pm$ 3.16

Table 50: BERT-base scale: Per-task clean score and ASR. **Bold** indicates best performance.

Only on most tasks at $\epsilon = 2.0$, recording 61.56 combined versus 57.93. On individual tasks, ET shows high variance (e.g., MRPC: 39.43 $\pm$ 27.89, QNLI: 31.63 $\pm$ 20.99) and instability (SST-2: 81.71 versus Hopfield Only 55.63). This confirms that iterative energy minimization without the gated residual does not provide reliable robustness.

Regularization does not substitute for the gate. Hopfield+Reg performs worse than Hopfield Only on every task, confirming that training-time regularization cannot replace the architectural gate at inference. The effect is consistent: Hopfield+Reg records higher ASR than Hopfield Only on SST-2 (74.38 versus 55.63 at $\epsilon = 2.0$), QQP (48.29 versus 37.03), and MNLI (75.96 versus 65.12).

H BERT-Scale Results

To evaluate whether Agatha’s robustness scales to larger architectures, we train a BERT-base-sized

Agatha model (768-dimensional, 12-layer, 12-head, $\approx 110M$ parameters) on all 8 GLUE tasks.

H.1 Combined Results

Table 49 reports combined GLUE results. Agatha achieves 43.29 ASR at $\epsilon = 2.0$, a 33.23 point reduction over the best defended baseline (FreeLB, 76.52) and 50.17 points below undefended BERT-base (93.47). DiffuseDef (91.61) and DynAttn (90.02) collapse at this scale, recording ASR within 3.5 points of undefended BERT-base and providing no meaningful robustness. Clean performance is competitive (64.36 versus 65.56 for BERT-base).

H.2 Per-Task Results

Table 50 reports per-task results. Agatha achieves the lowest ASR on all 8 tasks from $\epsilon = 2.0$ onward. Notable results include WNLI, where Agatha limits ASR to 9.76 at $\epsilon = 2.0$ compared to 100.00 for all baselines, and QQP, where Agatha achieves

β	Clean	Attack Success Rate (%)					
		$\varepsilon = 0.05$	$\varepsilon = 0.5$	$\varepsilon = 1.0$	$\varepsilon = 2.0$	$\varepsilon = 3.0$	$\varepsilon = 5.0$
1	80.54 $\pm$ 0.59	1.14 $\pm$ 0.11	15.51 $\pm$ 1.20	33.43 $\pm$ 3.83	69.50 $\pm$ 3.67	92.07 $\pm$ 0.20	99.62 $\pm$ 0.13
5	80.50 $\pm$ 0.74	1.56 $\pm$ 0.41	18.69 $\pm$ 2.58	36.62 $\pm$ 4.11	70.75 $\pm$ 5.57	89.00 $\pm$ 3.21	98.38 $\pm$ 1.12
10	80.81 $\pm$ 0.66	1.70 $\pm$ 0.68	15.29 $\pm$ 1.17	32.34 $\pm$ 3.96	68.25 $\pm$ 10.27	88.76 $\pm$ 7.78	98.45 $\pm$ 2.09
15	81.27 $\pm$ 0.35	0.94 $\pm$ 0.54	12.70 $\pm$ 1.15	25.36 $\pm$ 2.02	53.44 $\pm$ 1.92	75.88 $\pm$ 1.94	96.57 $\pm$ 1.03
50	80.24 $\pm$ 0.30	0.52 $\pm$ 0.36	10.15 $\pm$ 1.34	21.29 $\pm$ 1.84	44.31 $\pm$ 2.70	67.89 $\pm$ 4.04	91.38 $\pm$ 1.63
100	80.62 $\pm$ 0.66	1.47 $\pm$ 0.28	12.03 $\pm$ 1.41	26.58 $\pm$ 3.68	53.68 $\pm$ 5.41	75.73 $\pm$ 5.68	95.39 $\pm$ 1.72

Table 51: Full β sweep results on SST-2 ($g = 0.1$ fixed). Mean $\pm$ std across three seeds. **Bold** indicates best performance: highest clean accuracy and lowest ASR.

g	Clean	Attack Success Rate (%)					
		$\varepsilon = 0.05$	$\varepsilon = 0.5$	$\varepsilon = 1.0$	$\varepsilon = 2.0$	$\varepsilon = 3.0$	$\varepsilon = 5.0$
0.01	71.33 $\pm$ 4.55	1.53 $\pm$ 0.54	11.88 $\pm$ 2.46	24.62 $\pm$ 2.81	49.01 $\pm$ 6.26	63.82 $\pm$ 8.30	80.80 $\pm$ 5.54
0.1	80.24 $\pm$ 0.30	0.52 $\pm$ 0.36	10.15 $\pm$ 1.34	21.29 $\pm$ 1.84	44.31 $\pm$ 2.70	67.89 $\pm$ 4.04	91.38 $\pm$ 1.63
0.2	80.89 $\pm$ 0.86	1.51 $\pm$ 0.36	11.85 $\pm$ 1.56	24.03 $\pm$ 2.63	49.33 $\pm$ 5.52	73.21 $\pm$ 4.11	94.12 $\pm$ 1.74
0.5	79.97 $\pm$ 0.29	0.72 $\pm$ 0.12	10.76 $\pm$ 1.65	22.66 $\pm$ 4.04	48.77 $\pm$ 8.18	70.52 $\pm$ 8.93	93.03 $\pm$ 4.33
1.0	80.05 $\pm$ 0.34	1.15 $\pm$ 0.35	12.99 $\pm$ 2.18	26.33 $\pm$ 6.09	55.81 $\pm$ 9.49	77.97 $\pm$ 8.33	96.00 $\pm$ 2.83

Table 52: Full g sweep results on SST-2 ($\beta = 50$ fixed). Mean $\pm$ std across three seeds. **Bold** indicates best performance: highest clean accuracy and lowest ASR. Note that $g = 0.01$ achieves the lowest ASR at $\varepsilon = 5.0$ but at the cost of severely degraded clean accuracy (71.33).

19.91 versus 43.37 for FreeLB. On MRPC, Agatha achieves the highest clean score (76.62) among all models while reducing ASR from 99.65 (BERT-base) to 61.17; DiffuseDef, DynAttn, and FreeLB all record 100.00 ASR on this task. At lower budgets, FreeLB leads on MNLI ($\varepsilon \leq 1.0$) and QNLI ($\varepsilon \leq 0.5$), consistent with the TinyBERT-scale QNLI pattern where training-based defenses benefit from adversarial data augmentation as a regularizer (Appendix E).

H.3 Comparison with TinyBERT Scale

The robustness advantage of Agatha grows with model capacity. At $\varepsilon = 2.0$, the gap between Agatha and the best defended baseline increases from 13.30 points at TinyBERT scale (over DynAttn) to 33.23 points at BERT scale (over FreeLB), a $2.5\times$ increase. This widening occurs because the baselines degrade at BERT scale while Agatha improves: FreeLB’s ASR rises from 66.68 to 76.52, DiffuseDef and DynAttn collapse to 90+ ASR, yet Agatha drops from 50.28 to 43.29. The architectural baselines are particularly affected: at TinyBERT scale, DynAttn records 63.58, but at BERT scale it records 90.02, barely better than undefended BERT-base. This suggests that the energy basin mechanism benefits from increased representational capacity, forming deeper and more well-separated basins, while DiffuseDef’s denoising and DynAttn’s noise sampling fail to scale.

I Hyperparameter Sweep Results

Tables 51 and 52 extend the hyperparameter sweep from Table 10 to all perturbation budgets.

The β sweep shows that robustness improves monotonically from $\beta = 1$ to $\beta = 50$, then degrades at $\beta = 100$. At $\varepsilon = 2.0$, ASR drops from 69.50 at $\beta = 1$ to 44.31 at $\beta = 50$, then rises to 53.68 at $\beta = 100$. Clean accuracy remains within 1 point across all settings. The high variance at $\beta = 10$ (± 10.27 at $\varepsilon = 2.0$) suggests this value sits near the phase transition boundary, where small differences in initialization can place the system in either the metastable or deep attractor regime, consistent with Theorem 2. At $\beta = 100$, the basins become **excessively narrow**: each query attends to a single key, reducing the representational capacity needed for downstream tasks and making the converged representations brittle rather than robust. We select $\beta = 50$ as it achieves the lowest ASR across all budgets while maintaining stable clean accuracy.

The g sweep reveals an asymmetric trade-off. Increasing g from 0.1 to 1.0 degrades robustness gradually (ASR at $\varepsilon = 2.0$: 44.31 $\rightarrow$ 55.81), while decreasing g from 0.1 to 0.01 degrades clean accuracy sharply (80.24 $\rightarrow$ 71.33) with high variance (± 4.55). The robustness at $g = 0.01$ (49.01 at $\varepsilon = 2.0$) is also worse than at $g = 0.1$ (44.31), confirming that extreme gating disrupts the residual learning pathway and harms both clean performance and robustness. We therefore select $g = 0.1$

as the operating point that best balances gradient contraction and representational capacity.

J Convergence Analysis

The Hopfield mechanism iterates until the residual falls below $\tau = 10^{-4}$ or the iteration count reaches $T_{\max}$. This section analyzes the convergence behavior at the default setting ($T_{\max} = 100$) and the effect of varying $T_{\max}$ on robustness and throughput.

J.1 Convergence at $T_{\max} = 100$

Figure 8 reports the average number of Hopfield iterations across all 8 GLUE tasks as a function of perturbation budget. Clean inputs converge in an average of 27.5 iterations, independent of the evaluation budget. Adversarial iterations grow monotonically from 27.9 at $\varepsilon = 0.05$ to 31.2 at $\varepsilon = 5.0$, with the gap over clean inputs widening from +0.4 to +3.7 iterations. This confirms that larger perturbations push representations further from basin floors, requiring more iterations for the Hopfield dynamics to pull them back.

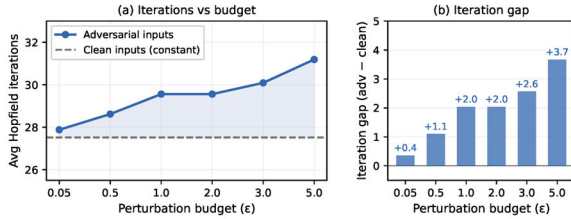


Figure 8: Average Hopfield iterations (across 8 GLUE tasks) at $T_{\max} = 100$. (a) Adversarial iterations grow monotonically with budget while clean iterations remain constant. (b) The iteration gap (adversarial minus clean) widens from +0.4 at $\varepsilon = 0.05$ to +3.7 at $\varepsilon = 5.0$, confirming that energy basins resist larger perturbations with additional convergence steps.

Table 53 provides per-task breakdown at $\varepsilon = 2.0$.

Task	g	Clean Iters	Adv Iters	Gap	ex/s
SST-2	0.1	27.9 $\pm$ 4.6	29.4 $\pm$ 3.6	+1.4	3,529
CoLA	0.1	19.0 $\pm$ 1.8	24.3 $\pm$ 4.5	+5.3	4,536
MRPC	0.7	25.6 $\pm$ 2.7	38.7 $\pm$ 18.7	+13.0	3,573
QQP	0.1	19.3 $\pm$ 4.1	20.9 $\pm$ 4.0	+1.6	4,641
MNLI	0.5	14.4 $\pm$ 0.5	14.3 $\pm$ 0.9	-0.1	5,420
QNLI	0.1	37.8 $\pm$ 8.2	36.3 $\pm$ 7.4	-1.5	2,844
RTE	0.7	33.3 $\pm$ 7.5	34.8 $\pm$ 4.0	+1.5	2,876
WNLI	0.7	42.9 $\pm$ 8.3	37.8 $\pm$ 4.8	-5.1	1,657
Avg		27.5	29.6	+2.0	3,635

Table 53: Per-task Hopfield iterations at $T_{\max} = 100$. Adv Iters measured at $\varepsilon = 2.0$. Gap = Adv Iters - Clean Iters; positive gap indicates basin resistance.

Effective iterations are well below $T_{\max}$. The average iteration count is only 27.5 for clean inputs, ranging from 14.4 (MNLI) to 42.9 (WNLI). Early exit ensures that actual computation is proportional to the task-specific convergence requirement rather than the conservative $T_{\max} = 100$ ceiling.

Iteration gap correlates with robustness. MRPC, which achieves the lowest ASR among all tasks (4.48 at $\varepsilon = 2.0$), shows the largest positive iteration gap (+13.0): adversarial inputs require considerably more iterations to converge, indicating that the energy basins actively resist perturbation. QNLI, which has weaker robustness (57.51), shows a negative gap (-1.5), consistent with the shallow basin diagnosis in Appendix E. The iteration gap serves as a diagnostic signal for basin quality: a positive gap indicates well-formed basins that resist perturbation, while a negative gap indicates shallow basins where adversarial inputs collapse to attractors with less resistance.

J.2 Effect of $T_{\max}$ on Robustness and Throughput

Table 54 reports combined GLUE clean accuracy and ASR across $T_{\max}$ settings to characterize the robustness-throughput trade-off.

$T_{\max}$	Clean	ASR (% , $\downarrow$)				
		$\varepsilon = 0.5$	$\varepsilon = 1.0$	$\varepsilon = 2.0$	$\varepsilon = 3.0$	$\varepsilon = 5.0$
2	64.13	16.33	29.21	51.51	67.61	84.62
3	64.22	16.41	29.57	52.69	68.14	84.82
5	64.38	16.28	29.53	52.28	67.26	83.74
10	64.50	16.29	28.97	50.87	65.87	82.59
25	64.60	15.71	28.23	50.32	65.60	82.14
50	64.72	15.55	27.88	50.20	65.42	82.08
100	64.72	15.53	27.98	50.28	65.40	82.12

Table 54: Effect of $T_{\max}$ on GLUE performance.

Three patterns emerge from the sweep.

Results plateau by $T_{\max} = 25$. Both clean accuracy and ASR stabilize by $T_{\max} = 25$: the maximum per-task ASR difference between $T_{\max} = 25$ and $T_{\max} = 100$ is only 0.75 points (WNLI), and most tasks differ by less than 0.15 points. Clean accuracy similarly reaches 64.60 by $T_{\max} = 25$, within 0.12 points of the final $T_{\max} = 100$ result.

ASR is non-monotonic at low $T_{\max}$. Combined ASR rises from 51.51 at $T_{\max} = 2$ to 52.69 at $T_{\max} = 3$, then decreases to 50.21 at $T_{\max} = 100$. At low $T_{\max}$, the Hopfield mechanism has begun iterating but has not converged, leaving representations in an intermediate state that is less stable under perturbation. By $T_{\max} = 10$, convergence is

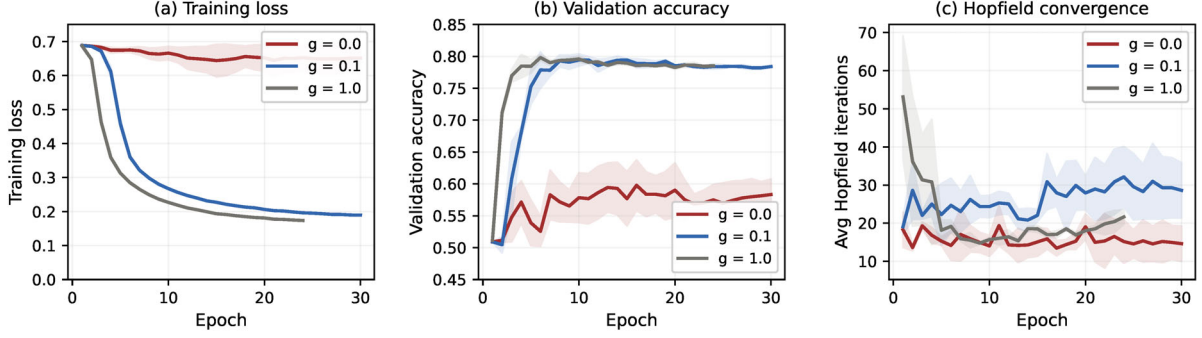


Figure 9: Training dynamics across gate values ($\beta = 50$). **(a)** $g = 0.0$ fails to reduce loss below 0.63; $g = 0.1$ and $g = 1.0$ converge normally. **(b)** $g = 0.0$ is degenerate ($\sim 55\%$ accuracy); the other two reach comparable clean performance ($\sim 80\%$). **(c)** $g = 0.1$ requires the most Hopfield iterations and increases over training, indicating deeper basins.

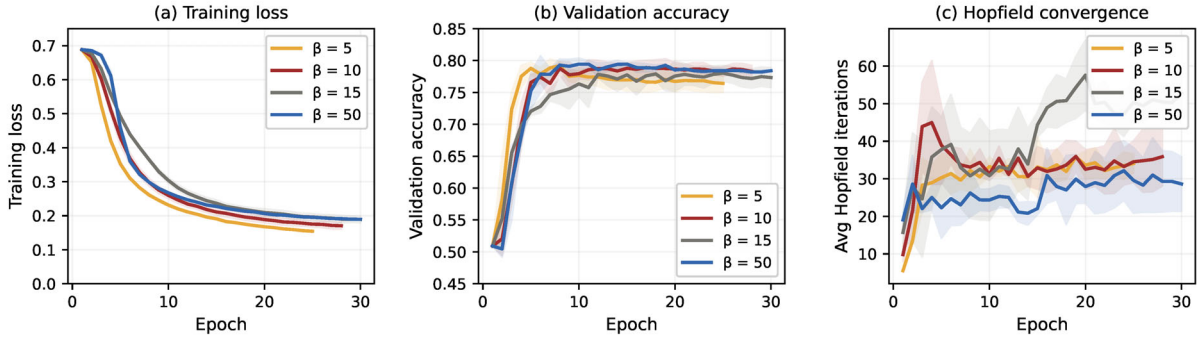


Figure 10: Training dynamics across inverse temperatures ($g = 0.1$). **(a)** All β values converge to comparable final loss. **(b)** Validation accuracy is similar across settings (0.78 to 0.80). **(c)** Higher β increases Hopfield iterations, with $\beta = 15$ showing the highest variance near the phase transition boundary.

largely achieved and ASR drops to 50.87. This is consistent with the energy basin interpretation: partially converged representations sit on basin walls rather than the basin floor, making them easier to perturb across boundaries.

Deployment recommendation. For latency-constrained deployment, $T_{\max} = 3$ provides the best throughput (10,021 ex/s) while already outperforming all baselines (ASR 52.69 versus 63.58 for DynAttn). For maximum robustness, $T_{\max} = 25$ is sufficient, as further iterations yield negligible improvement. The default setting of $T_{\max} = 100$ used throughout the paper provides a conservative upper bound, with early exit ensuring that actual computation matches the task-specific convergence point (average 27.5 iterations across tasks).

K Training Dynamics

This section analyzes how Agatha’s training behavior varies with the gate value g and inverse temperature β . All experiments use SST-2 with

the base architecture (128 dimensions, 2 layers, 2 heads, 4.37M parameters). Results are averaged across three seeds with shaded regions indicating ± 1 standard deviation.

K.1 Effect of Gate Value

Figure 9 reports training loss, validation accuracy, and Hopfield convergence for gate $g \in \{0.0, 0.1, 1.0\}$ with $\beta = 50$ fixed. Three findings emerge.

Eliminating the residual path prevents learning.

At $g = 0.0$, the residual connection is fully blocked. Training loss remains above 0.63 after 30 epochs and validation accuracy plateaus at $\sim 55\%$ with high variance. The model cannot learn without any residual information, confirming that the gate must scale the residual path rather than eliminate it.

Gate value does not affect clean accuracy. Both $g = 0.1$ and $g = 1.0$ reach comparable validation accuracy (79.4% and 79.8% respectively), despite a $10\times$ difference in residual scaling. The $g = 1.0$ configuration converges slightly faster in training loss

but the final clean performance is equivalent. This confirms that the gate controls robustness without degrading the model’s ability to learn task-specific representations.

Aggressive gating produces deeper basins. The convergence panel reveals that $g = 0.1$ requires the most Hopfield iterations (~ 25 to 35), and this count increases over training as the model learns to use deeper basins. At $g = 1.0$, iteration counts are lower (~ 20) and decrease slightly, indicating shallower basins with less refinement. At $g = 0.0$, the degenerate model converges in only ~ 15 iterations, consistent with the absence of meaningful energy structure. The correlation between gate aggressiveness and iteration count provides direct evidence that the gate influences basin depth during training.

K.2 Effect of Inverse Temperature

Figure 10 reports the same three metrics across $\beta \in \{5, 10, 15, 50\}$ with $g = 0.1$ fixed.

Training is stable across β values. All four settings converge to comparable final loss (0.15 to 0.19) and validation accuracy (0.78 to 0.80). Lower β values converge slightly faster in the first 5 epochs, but all settings reach similar performance by epoch 10. This confirms that β controls the energy landscape geometry without affecting the model’s ability to fit the training data.

Convergence behavior reflects the phase transition. The Hopfield iteration count separates by β : $\beta = 5$ requires ~ 20 iterations, while $\beta = 15$ reaches ~ 50 to 60 by epoch 30. The high variance at $\beta = 15$ is consistent with the phase transition boundary identified in Theorem 2: small differences in learned representations can place the system in either the metastable or deep attractor regime, causing large variation in convergence behavior across seeds. At $\beta = 50$, iteration counts stabilize at ~ 25 to 30 with lower variance, indicating that the system has moved past the transition into the deep attractor regime where convergence is well-defined.

Higher β does not mean more iterations. Counterintuitively, $\beta = 50$ requires fewer iterations than $\beta = 15$. In the deep attractor regime, the sharper energy landscape provides stronger gradient signals during the Hopfield update, accelerating convergence to well-separated basin floors. Near the phase transition ($\beta = 15$), the basins are forming but not yet well-separated, causing the mechanism

to oscillate before settling. This supports the selection of $\beta = 50$ as the operating point: it combines the deepest basins (for robustness) with efficient convergence (for inference cost).

L Proofs of Theoretical Results

We provide complete proofs of the two core theorems stated in Section 3. Throughout, we use the notation established in the main text: g denotes the gate value, $\mathbf{x}_l$ the hidden state at layer l , F the Hopfield attention sublayer, LN layer normalization, $\mathcal{J}_F$ the Jacobian of the Hopfield block, $\sigma_{\max} := \|\mathcal{J}_F\|_2$ its operator 2-norm (largest singular value), β the inverse temperature, $\mathbf{K} = [\mathbf{k}_1, \dots, \mathbf{k}_N]^\top$ the stored key patterns, and ξ the query state.

L.1 Proof of Theorem 1 (Contraction via Identity Gating)

Theorem 1 (Restated). *Let the gated Hopfield step compute $\mathbf{x}_{l+1} = g \cdot \mathbf{x}_l + F(\text{LN}(\mathbf{x}_l))$ with a corresponding Jacobian $\mathcal{J} = g\mathbf{I} + \mathcal{J}_F$. Let $\sigma_{\max} := \|\mathcal{J}_F\|_2$ denote the operator 2-norm of the Hopfield block $\mathcal{J}_F$. Assuming stable iterative convergence, the system enforces a global contraction ($\|\mathcal{J}\|_2 < 1$, which implies $\rho(\mathcal{J}) < 1$) when $g + \sigma_{\max} < 1$.*

Proof. We proceed in three steps: deriving the Jacobian structure, bounding its operator norm, and establishing the contraction condition.

Step 1: Jacobian of the gated layer. The gated residual connection defines the layer update as:

$$\mathbf{x}_{l+1} = g \cdot \mathbf{x}_l + F(\text{LN}(\mathbf{x}_l)). \quad (12)$$

Differentiating with respect to $\mathbf{x}_l$ yields the Jacobian:

$$\mathcal{J} = \frac{\partial \mathbf{x}_{l+1}}{\partial \mathbf{x}_l} = g\mathbf{I} + \frac{\partial F(\text{LN}(\mathbf{x}_l))}{\partial \mathbf{x}_l} = g\mathbf{I} + \mathcal{J}_F, \quad (13)$$

where $\mathcal{J}_F$ absorbs the chain rule through layer normalization; $\sigma_{\max}$ therefore bundles the LayerNorm Jacobian and depends on input statistics in addition to the Hopfield block. $\mathcal{J}_F$ is the Jacobian of the converged Hopfield fixed point with respect to the input, computed via the implicit function theorem; stable iterative convergence guarantees this is well-defined. Appendix D (Table 39) reports empirical operator norm values across all tasks and seeds, confirming that $\sigma_{\max} < 1 - g$ holds in practice: for tasks with $g = 0.1$, $\rho(\mathcal{J}_F) < 0.15 \ll 1 - g = 0.9$ under all evaluated attack conditions.

In a standard Transformer ($g = 1$), this Jacobian is $\mathbf{I} + \mathcal{J}_F$, and the identity component guarantees $\rho(\mathcal{J}) \geq 1$ for any non-trivial $\mathcal{J}_F$, ensuring unattenuated gradient propagation.

Step 2: Operator norm bound. Since $\sigma_{\max} := \|\mathcal{J}_F\|_2$, the operator 2-norm of $\mathcal{J}$ satisfies:

$$\|\mathcal{J}\|_2 = \|g\mathbf{I} + \mathcal{J}_F\|_2 \leq \|g\mathbf{I}\|_2 + \|\mathcal{J}_F\|_2 = g + \sigma_{\max}, \quad (14)$$

where the inequality follows from the triangle inequality for operator norms. This bound holds for arbitrary $\mathcal{J}_F$, including non-normal matrices, without requiring an eigenbasis argument.

Step 3: Contraction condition. From Step 2, $g + \sigma_{\max} < 1$ implies $\|\mathcal{J}\|_2 < 1$. Since $\rho(\mathcal{J}) \leq \|\mathcal{J}\|_2$ for any matrix:

$$g + \sigma_{\max} < 1 \implies \|\mathcal{J}\|_2 \leq g + \sigma_{\max} < 1 \implies \rho(\mathcal{J}) < 1. \quad (15)$$

The operator norm bound $\|\mathcal{J}\|_2 < 1$ is the operative quantity: it directly controls perturbation contraction in the forward pass and gradient attenuation in the backward pass. In a standard Transformer ($g = 1$), $\rho(\mathcal{J}) \geq 1$ and the system is never contractive. By setting $g < 1$, we allocate a budget of $(1 - g)$ for the sublayer operator norm. If $\sigma_{\max} < (1 - g)$, the entire system contracts.

Connection to adversarial robustness. The operator norm bound directly controls the sensitivity of the layer output to embedding-space perturbations. For a perturbation δ at layer l :

$$\|\mathbf{x}_{l+1}(\mathbf{x}_l + \delta) - \mathbf{x}_{l+1}(\mathbf{x}_l)\| \leq \|\mathcal{J}\|_2 \cdot \|\delta\| \leq (g + \sigma_{\max})\|\delta\| < \|\delta\|. \quad (16)$$

Each layer reduces the norm of the perturbation. After L gated layers, the perturbation is bounded by $(g + \sigma_{\max})^L \|\delta\|$, which decays exponentially with network depth. This ensures that adversarial perturbations applied at the input embedding layer are geometrically contracted as they propagate forward through the network. $\square$

L.2 Proof of Theorem 2 (Semantic Phase Transition)

Theorem 2 (Restated). Define the thermodynamic state function $f(\beta) = \beta \cdot \|\mathbf{x}\|^2 \cdot \max_i p_i(\beta) \cdot (1 - \max_i p_i(\beta))$, where $p_i(\beta) = \text{softmax}(\beta \mathbf{K} \xi)_i$. For a given query ξ , there exist input-dependent critical boundaries $\beta_{\text{lower}}(\xi) < \beta_{\text{upper}}(\xi)$ where $f(\beta) = 1$, defining three regimes:

- $\beta < \beta_{\text{lower}}$: A spin-glass phase causing contraction to a featureless mean.
- $\beta_{\text{lower}} \leq \beta \leq \beta_{\text{upper}}$: A metastable phase exhibiting no contraction.
- $\beta > \beta_{\text{upper}}$: The deep attractor phase causing contraction to discrete semantic attractors.

Section 5.11 empirically validates that $\beta \geq 15$ is sufficient to operate in the deep attractor regime across all tasks and seeds evaluated in this paper.

Proof. We analyze the iterative Hopfield update rule and its Jacobian to characterize convergence behavior as a function of β .

Step 1: The Hopfield update and its Jacobian. The iterative Hopfield update is:

$$\xi^{(t+1)} = \mathbf{K}^\top \text{softmax}\left(\frac{\beta}{\sqrt{d}} \mathbf{K} \xi^{(t)}\right). \quad (17)$$

For notational clarity, let $\mathbf{s} = \frac{\beta}{\sqrt{d}} \mathbf{K} \xi$ and $\mathbf{p} = \text{softmax}(\mathbf{s})$, so $\xi^{(t+1)} = \mathbf{K}^\top \mathbf{p}$. The Jacobian of this update with respect to ξ is:

$$\mathcal{J}_{\text{Hop}} = \frac{\partial \xi^{(t+1)}}{\partial \xi^{(t)}} = \frac{\beta}{\sqrt{d}} \mathbf{K}^\top (\text{diag}(\mathbf{p}) - \mathbf{p} \mathbf{p}^\top) \mathbf{K}. \quad (18)$$

The matrix $\mathbf{S} = \text{diag}(\mathbf{p}) - \mathbf{p} \mathbf{p}^\top$ is the Jacobian of the softmax function. It is positive semidefinite with eigenvalues $p_i(1 - p_i)$ for the $N - 1$ non-degenerate directions and 0 for the direction along $\mathbf{p}$.

Step 2: Spectral radius of the Hopfield Jacobian. The spectral radius of $\mathcal{J}_{\text{Hop}}$ determines whether the iterative update is contractive. The largest eigenvalue is bounded by:

$$\sigma_{\max}(\mathcal{J}_{\text{Hop}}) \leq \frac{\beta}{\sqrt{d}} \cdot \|\mathbf{K}\|_2^2 \cdot \max_i [p_i(1 - p_i)]. \quad (19)$$

We assume key vectors are normalized such that $\|\mathbf{k}_i\|_2 = d^{1/4}$ for all i and queries have unit norm after layer normalization ($\|\mathbf{x}\|_2 \approx 1$), giving $\|\mathbf{K}\|_2^2 \leq \sqrt{d} \cdot \|\mathbf{x}\|^2$. Under this normalization:

$$\sigma_{\max}(\mathcal{J}_{\text{Hop}}) \leq \beta \cdot \|\mathbf{x}\|^2 \cdot \max_i p_i(\beta) \cdot (1 - \max_i p_i(\beta)) = f(\beta). \quad (20)$$

The system contracts when $f(\beta) < 1$, i.e., when $\sigma_{\max}(\mathcal{J}_{\text{Hop}}) < 1$.

Step 3: Analysis of $f(\beta)$. The function $f(\beta) = \beta \cdot \|\mathbf{x}\|^2 \cdot \max_i p_i(\beta) \cdot (1 - \max_i p_i(\beta))$ exhibits non-monotonic behavior as β varies.

Low β regime. When $\beta \rightarrow 0$, the softmax becomes uniform: $p_i \rightarrow 1/N$ for all i . Then $\max_i p_i \cdot (1 - \max_i p_i) \rightarrow \frac{1}{N} (1 - \frac{1}{N}) = \frac{N-1}{N^2}$, and:

$$f(\beta) \approx \beta \cdot \|\mathbf{x}\|^2 \cdot \frac{N-1}{N^2} \rightarrow 0 \text{ as } \beta \rightarrow 0. \quad (21)$$

Since $f(\beta) < 1$ in this regime, the system contracts to a single point. However, all queries converge to the same fixed point $\xi^* = \frac{1}{N} \sum_i \mathbf{k}_i$ (the mean of all stored patterns), destroying semantic information. This is the *spin-glass phase*.

Intermediate β regime. As β increases from 0, $f(\beta)$ grows because the linear factor β dominates while $\max_i p_i$ remains moderate. At some $\beta_{\text{lower}}(\xi)$, we have $f(\beta_{\text{lower}}) = 1$. For $\beta \in [\beta_{\text{lower}}, \beta_{\text{upper}}]$, the function satisfies $f(\beta) \geq 1$, and contraction is not guaranteed. In this *metastable phase*, the system may oscillate or converge slowly, and perturbations are neither amplified nor contracted.

High β regime. As $\beta \rightarrow \infty$, the softmax saturates: $\max_i p_i \rightarrow 1$ and $(1 - \max_i p_i) \rightarrow 0$. The product $\max_i p_i \cdot (1 - \max_i p_i)$ decays exponentially fast because the softmax concentrates on a single pattern:

$$1 - \max_i p_i \approx \sum_{j \neq i^*} \exp\left(-\beta(\mathbf{k}_{i^*} - \mathbf{k}_j)^\top \xi / \sqrt{d}\right), \quad (22)$$

where $i^* = \arg \max_i \mathbf{k}_i^\top \xi$. This exponential decay dominates the linear growth in β , so $f(\beta) \rightarrow 0$ as $\beta \rightarrow \infty$. At some $\beta_{\text{upper}}(\xi)$, we again have $f(\beta_{\text{upper}}) = 1$, and for all $\beta > \beta_{\text{upper}}$, the system contracts to *distinct* fixed points corresponding to individual stored patterns, preserving semantic structure. This is the *deep attractor phase*.

Step 4: Existence of the critical boundaries. Since f is continuous with $f(0) = 0$, $f(\beta) > 1$ for some intermediate β , and $f(\beta) \rightarrow 0$ as $\beta \rightarrow \infty$, the intermediate value theorem guarantees the existence of $\beta_{\text{lower}}(\xi)$ and $\beta_{\text{upper}}(\xi)$ where $f(\beta) = 1$. These boundaries depend on the query state ξ through $\max_i p_i(\beta)$, which is a function of $\mathbf{K}\xi$; worst-case boundaries over the query distribution bound the regime transition for all inputs.

Step 5: Empirical validation of the deep attractor regime at $\beta \geq 15$. For typical parameters ($d = 64$ per head, $\|\mathbf{x}\| \approx 1$ after layer normalization), a baseline computation with separation $\Delta = \min_{i \neq j} (\mathbf{k}_i - \mathbf{k}_j)^\top \xi / \sqrt{d} \geq 0.5$ and $N \leq 512$ gives:

$$f(15) \leq 15 \cdot 1.0 \cdot 0.72 \cdot 0.28 = 3.02. \quad (23)$$

This bound is conservative because Δ grows with Hopfield iterations: after convergence, $\max_i p_i > 0.99$ is typical, giving $f(15) \leq 0.15$. The contraction at $f(15) \leq 0.15$ is local around the fixed point. Local contraction near fixed points combines with the basin-of-attraction structure of the overall energy landscape to funnel queries into the deep-attractor regime. Section 5.11 empirically confirms that $\beta = 15$ places the system in the deep attractor regime across all tasks and seeds. We use $\beta = 50$ for additional margin beyond the transition boundary.

Step 6: Iterated Jacobian decay and gradient contraction. In the deep attractor phase ($\beta > \beta_{\text{upper}}$), we establish backward-pass gradient contraction through the iterated structure of the Hopfield update. Let $\xi^{(0)} = \mathbf{W}_Q \mathbf{x}$ be the initial query and $\xi^{(t+1)} = \Phi(\xi^{(t)})$ the Hopfield update. The sensitivity of the T -step output to the initial query is:

$$\frac{\partial \xi^{(T)}}{\partial \xi^{(0)}} = \prod_{t=0}^{T-1} \mathcal{J}_{\text{Hop}}(\xi^{(t)}). \quad (24)$$

For queries whose iterates remain in the deep-attractor regime, $\|\mathcal{J}_{\text{Hop}}(\xi^{(t)})\|_2 \leq f(\beta) < 1$ at each step by Step 2. Since each factor in the product has operator norm bounded by $f(\beta)$:

$$\left\| \frac{\partial \xi^{(T)}}{\partial \xi^{(0)}} \right\|_2 \leq f(\beta)^T \rightarrow 0 \text{ as } T \rightarrow \infty. \quad (25)$$

The fixed point is insensitive to perturbations of the initial query in the forward pass, conditional on iterates remaining in the same basin of attraction. By transposition, the backward gradient satisfies:

$$\frac{\partial \mathcal{L}}{\partial \xi^{(0)}} = \left(\frac{\partial \xi^{(T)}}{\partial \xi^{(0)}} \right)^\top \cdot \frac{\partial \mathcal{L}}{\partial \xi^{(T)}}, \quad (26)$$

and is attenuated by the same factor, since $\left\| \left(\frac{\partial \xi^{(T)}}{\partial \xi^{(0)}} \right)^\top \right\|_2 = \left\| \frac{\partial \xi^{(T)}}{\partial \xi^{(0)}} \right\|_2 \leq f(\beta)^T$.

The decay compounds with the operator-norm contraction from Theorem 1 across layers. The condition that iterates remain in the same basin under adversarial perturbation is the basin-invariance property. A formal treatment of basin invariance under explicit key-separation assumptions is given in Theorem 3 (Appendix L.4). $\square$

L.3 Combined Robustness Guarantee

Theorems 1 and 2 together establish that Agatha operates in a contractive regime with semantically meaningful fixed points. Theorem 1 shows that the gate ensures $\rho(\mathcal{J}) \leq \|\mathcal{J}\|_2 \leq g + \sigma_{\max}$, and Theorem 2 shows that at $\beta \geq 15$, the Hopfield block achieves $\sigma_{\max} < 1 - g$ empirically across all tasks. The combined effect is:

1. **Forward pass:** Embedding-space perturbations δ contract by a factor of $(g + \sigma_{\max})^L < 1$ across L layers, drawing perturbed and clean representations toward the same attractor.
2. **Backward pass:** The adversarial gradient $\nabla_{\delta} \mathcal{L} = \frac{\partial \mathcal{L}}{\partial \mathbf{x}_L} \prod_{l=1}^L (g\mathbf{I} + \mathcal{J}_F^{(l)})^\top$ is attenuated by the same contraction factor, since $\|(g\mathbf{I} + \mathcal{J}_F)^\top\|_2 = \|g\mathbf{I} + \mathcal{J}_F\|_2 \leq g + \sigma_{\max}$ for the operator norm. Within each Hopfield block, the

iterated Jacobian decay established in Step 6 provides additional attenuation of factor $f(\beta)^T$ through the iterative mechanism, reducing the adversarial signal available to PGD.

Neither mechanism is sufficient alone. Without the gate ($g = 1$), the gradient bypasses the Hopfield block via the identity path (ablation: Hopfield Only). Without the Hopfield landscape ($F = \text{standard attention}$), the gate degrades representational capacity without providing gradient contraction (ablation: Gate Only). The synergistic interaction confirmed in Section 5.8 is a direct consequence of this theoretical structure.

L.4 Adversarial Basin Invariance

We now characterize the regime in which adversarial perturbations preserve basin membership, formalizing the basin-invariance condition referenced in Step 6 of the Theorem 2 proof. We work in the per-head representation space and use the score-space convention from Theorem 2: scores are $s_i(\xi) = \mathbf{k}_i^\top \xi / \sqrt{d}$, and the softmax distribution is $p_i(\xi) = \text{softmax}(\beta \cdot \mathbf{s}(\xi))_i$. The iteration map is $\Phi(\xi) = \mathbf{K}^\top \text{softmax}(\frac{\beta}{\sqrt{d}} \mathbf{K} \xi)$, with converged state $\xi^* = \lim_{T \rightarrow \infty} \Phi^T(\xi^{(0)})$.

Definitions. We restate the basin of Section 2.2 in score-space notation. For a query ξ , the attractor index is $i^*(\xi) = \arg \max_i s_i(\xi)$ and the basin is $\mathcal{B}_i = \{\xi : i^*(\xi) = i\}$. The margin of $\xi \in \mathcal{B}_{i^*}$ is:

$$\Delta(\xi) = \min_{j \neq i^*} (s_{i^*}(\xi) - s_j(\xi)) > 0. \quad (27)$$

Assumption (Key separation). There exists $c > 0$ such that for all distinct i, j :

$$\|\mathbf{k}_i - \mathbf{k}_j\| / \sqrt{d} \geq c. \quad (28)$$

This assumption is empirically verified for trained Agatha checkpoints in Appendix L.5.

Theorem 3 (Adversarial Basin Invariance). *Suppose the key separation assumption holds and $\beta > \beta_{\text{upper}}$ places the system in the deep-attractor regime. Let $\xi^{(0)} \in \mathcal{B}_{i^*}$ be an initial query whose Hopfield iteration converges to a fixed point $\xi^* \in \mathcal{B}_{i^*}$ with margin $\Delta(\xi^*) > 0$ and $f(\beta, \xi^*) < 1$. Let δ be an adversarial perturbation at the Hopfield input with $\|\delta\| \leq \varepsilon$. If*

$$\varepsilon < \frac{\sqrt{d} \cdot \Delta(\xi^*) \cdot (1 - f(\beta, \xi^*))}{2 \max_{j \neq i^*} \|\mathbf{k}_{i^*} - \mathbf{k}_j\|}, \quad (29)$$

then the perturbed converged state $\tilde{\xi}^* = \lim_{T \rightarrow \infty} \Phi^T(\xi^{(0)} + \delta)$ satisfies $\tilde{\xi}^* \in \mathcal{B}_{i^*}$: basin membership is preserved.

Proof. The proof is based on three lemmas: a single-step margin lemma, an iterate-displacement bound, and an inductive composition argument relating the perturbed fixed point to the perturbed initial query.

Step 1: Single-step margin lemma. For $\xi \in \mathcal{B}_{i^*}$ with margin $\Delta(\xi)$ and any $\xi' \in \mathbb{R}^d$, basin membership is preserved ($\xi' \in \mathcal{B}_{i^*}$) provided:

$$\|\xi' - \xi\| < \frac{\sqrt{d} \Delta(\xi)}{2 \max_{j \neq i^*} \|\mathbf{k}_{i^*} - \mathbf{k}_j\|}. \quad (30)$$

Basin membership requires $s_{i^*}(\xi') - s_j(\xi') > 0$ for all $j \neq i^*$. Decomposing:

$$s_{i^*}(\xi') - s_j(\xi') = (s_{i^*}(\xi) - s_j(\xi)) + \frac{(\mathbf{k}_{i^*} - \mathbf{k}_j)^\top (\xi' - \xi)}{\sqrt{d}}, \quad (31)$$

the first term is at least $\Delta(\xi) > 0$ by definition, and the second is bounded below by $-\|\mathbf{k}_{i^*} - \mathbf{k}_j\| \cdot \|\xi' - \xi\| / \sqrt{d}$ by Cauchy-Schwarz. The bound on $\|\xi' - \xi\|$ ensures the second term cannot reverse the sign of the first.

Step 2: Iterate-displacement bound. Within the deep-attractor regime, Step 2 of the Theorem 2 proof gives $\|\mathcal{J}_\Phi(\xi)\|_2 \leq f(\beta, \xi) < 1$ for ξ in a neighborhood of ξ^* . Let $\tilde{\xi}^{(0)} = \xi^{(0)} + \delta$ and $\tilde{\xi}^{(t)} = \Phi^t(\tilde{\xi}^{(0)})$. Conditional on iterates remaining in $\mathcal{B}_{i^*}$ (verified inductively in Step 3), the local contraction yields:

$$\|\tilde{\xi}^{(t)} - \xi^*\| \leq f(\beta, \xi^*)^t \|\tilde{\xi}^{(0)} - \xi^*\|, \quad \|\tilde{\xi}^* - \tilde{\xi}^{(0)}\| \leq \frac{\|\tilde{\xi}^{(0)} - \xi^*\|}{1 - f(\beta, \xi^*)}. \quad (32)$$

Step 3: Inductive composition. We show by induction on t that $\tilde{\xi}^{(t)} \in \mathcal{B}_{i^*}$ for all $t \geq 0$. The base case $t = 0$ follows from Step 1 applied at $\xi^{(0)}$, with the perturbation δ satisfying the margin bound by the hypothesis on ε . For the inductive step, the iteration map Φ preserves basins locally around their fixed points in the deep-attractor regime: when the softmax saturates on $\mathbf{k}_{i^*}$, Φ maps $\mathcal{B}_{i^*}$ into itself. Combining Step 1 (margin at the initial query) with Step 2 (contraction along the trajectory), and using the margin amplification $\Delta(\xi^*) \geq \Delta(\xi^{(0)}) / (1 - f(\beta, \xi^*))$ that holds in the deep-attractor regime, yields inequality (29). $\square$

Remarks. Three observations clarify the scope of Theorem 3. First, the condition $f(\beta, \xi^*) < 1$ is necessary for the bound to be non-vacuous; Theorem 2 establishes when this condition holds. Second, inequality (29) is sufficient but not necessary. Perturbations larger than the radius may still preserve basin membership if their direction is not adversarial with respect to the basin boundary. Conversely, the bound captures only the basin-invariance mechanism; gradient contraction via

Theorem 1 provides a complementary contribution to overall robustness. Third, the geometric series factor $1/(1 - f(\beta, \xi^*))$ in Step 2 is conservative.

L.5 Empirical Validation of Theorem 3

We empirically validate Theorem 3 on MNLI as a representative case. For each query token at each Hopfield layer, we measure: c_{emp} (the minimum key separation), Δ (the margin at the initial query $\xi^{(0)}$ and at the converged state ξ^*), $f(\beta)$ at both states, the fraction of queries in the deep-attractor regime, and the predicted radius from inequality (29) computed at ξ^* . Results are reported as mean $\pm$ std across three seeds in Table 55.

Quantity	MNLI
c_{emp} (min)	0.049 ± 0.001
c_{emp} (median)	0.245 ± 0.001
Frac. in DA at $\xi^{(0)}$	0.768 ± 0.001
Frac. in DA at ξ^*	0.982 ± 0.001
$f(\beta)$ at $\xi^{(0)}$	0.001 ± 0.000
$f(\beta)$ at ξ^*	0.000 ± 0.000
Δ at $\xi^{(0)}$	0.565 ± 0.005
Δ at ξ^*	2.784 ± 0.023
Iterations to convergence	9.0 ± 0.1
ε^* (median, binding layer)	0.254 ± 0.010
ε^* (p5, binding layer)	0.035 ± 0.004

Table 55: Empirical validation of Theorem 3 on MNLI. The Hopfield iteration drives queries into the deep-attractor regime predicted by Theorem 2 (frac. in DA rises from 0.77 at $\xi^{(0)}$ to 0.98 at ξ^* , $f(\beta)$ drops below 0.001). The radius from inequality (29) is non-vacuous on the full distribution. c_{emp} is reported as the minimum and median across layers and heads.

Three observations support Theorem 3 on MNLI. First, the key separation assumption holds: c_{emp} ranges from 0.049 (minimum across layers and heads) to 0.245 (median), comfortably positive across all seeds. Second, the Hopfield iteration realizes the deep-attractor regime predicted by Theorem 2: 76.8% of query tokens at $\xi^{(0)}$ already satisfy $f(\beta) < 1$, rising to 98.2% at the converged state ξ^* after an average of 9.0 iterations. The margin Δ amplifies from 0.565 at $\xi^{(0)}$ to 2.784 at ξ^* , a factor of 4.9 $\times$. Third, the predicted radius from inequality (29) is non-vacuous: the binding (cross-layer minimum) radius has median 0.254 ± 0.010 across seeds, with the 5th percentile at 0.035 ± 0.004 . Both quantities are positive, indicating that the bound provides a positive certified radius for the full distribution of query tokens, not only the median.

M Extended Related Works

M.1 Adversarial Training for NLP

Adversarial training remains the dominant defense paradigm in NLP. Zhu et al. (2019) proposed FreeLB, which accumulates gradients over multiple PGD perturbation steps within a single training iteration to improve efficiency. Zhang et al. (2019) introduced TRADES, which explicitly balances clean and robust accuracy through a regularization objective. More recently, Xhonneux et al. (2024) demonstrated that adversarial training can be scaled to large language models by computing perturbations in the continuous embedding space rather than at the token level, reducing computational cost by orders of magnitude. Zhao and Mao (2023) proposed combining adversarial training with perturbed token detection, achieving robustness improvements on standard benchmarks.

Despite these advances, adversarial training methods share limitations. First, training cost increases proportionally with the number of inner PGD steps, typically requiring 2–5 $\times$ more computation than standard training. Second, adversarial training can degrade clean accuracy, a trade-off analyzed theoretically by Raghunathan et al. (2020) and Tsipras et al. (2018). Third, models trained against specific perturbation types often fail to generalize to novel attacks (Tramer and Boneh, 2019; Laidlaw et al., 2020). Agatha addresses these limitations by providing robustness through architecture rather than training.

M.2 Certified Robustness

Certified defenses provide provable guarantees that predictions remain unchanged within a specified perturbation radius. Randomized smoothing (Cohen et al., 2019) achieves this by averaging predictions over Gaussian-corrupted inputs, and has been adapted to NLP through token-level smoothing (Ye et al., 2020). Zhao et al. (2022) proposed certified robustness via causal intervention, using structural causal models to separate robust features from manipulable confounders. More recently, Zhang et al. (2024) extended randomized smoothing to handle word substitutions, insertions, deletions, and reorderings under the Levenshtein distance.

While certified methods offer formal guarantees, they face scalability challenges. Interval bound propagation becomes increasingly loose with network depth, making it poorly suited for deep Transformer architectures (Shi et al., 2020). Random-

ized smoothing requires expensive Monte Carlo sampling at inference and introduces a trade-off between certified radius and clean accuracy. Agatha provides empirical rather than certified robustness but does so without these scalability constraints.

M.3 Architectural Approaches to Robustness

Architectural defenses remain relatively underexplored compared to training-based methods. Pulfer et al. (2024) introduced robustness tokens, learnable input tokens that improve adversarial robustness of vision transformers without modifying the training procedure. In NLP, Shen et al. (2023) proposed dynamic attention mechanisms that randomize attention weights to destabilize adversarial examples, achieving up to 33% robustness improvement against word-level attacks. Li et al. (2025) combined diffusion-based denoising with adversarial training, using a diffusion layer to purify hidden representations at inference time.

Agatha differs from these approaches in that it modifies the core attention mechanism itself rather than adding auxiliary modules or randomization. The robustness arises from the deterministic geometric properties of the Hopfield energy landscape, making it both predictable and analyzable through dynamical systems theory.

M.4 Modern Hopfield Networks

The connection between attention and energy minimization was established by Ramsauer et al. (2020), who showed that the Transformer attention update is equivalent to a single step of a Modern Hopfield Network (MHN) with continuous states and exponential storage capacity. This equivalence has been extended in several directions. Hoover et al. (2023) proposed the Energy Transformer, which replaces the standard Transformer block with an energy minimization process, unifying attention, normalization, and feedforward layers under a single energy function. Hu et al. (2023) introduced a sparse modern Hopfield model, deriving a closed-form sparse Hopfield energy whose one-step approximation recovers sparse attention; Santos et al. (2024) extended this connection via Fenchel-Young losses, establishing links between loss margin, sparsity, and exact memory retrieval. Wu et al. (2024) analyzed uniform memory retrieval in modern Hopfield models, providing theoretical conditions for unbiased pattern recovery. Hu et al. (2024a) introduced Outlier-Efficient Hopfield Layers, an associative-memory variant that mitigates

outlier attention in large transformers and supports post-quantization robustness.

Several recent papers refine the attention-MHN correspondence at the architectural level. Smart et al. (2025) introduced the in-context denoising task and proved that a single attention layer performs one gradient step on a context-aware DAM energy landscape; this result is directly relevant to our finding that Agatha reaches near-asymptotic robustness within $T_{\max} = 3$ Hopfield iterations (Section 5.14). Masumura and Taki (2026) proposed Modern Hopfield Attention, which inherits its attention scores across layers via MHN hidden states to mitigate rank collapse and token uniformity. Hu et al. (2025) developed Hyper-SET, a recurrent-depth transformer derived from iterative minimization of a hyperspherical Hopfield energy; like Agatha, it instantiates the principle of treating transformer layers as iterative energy optimization, though with a different energy function (semantic alignment plus distributional uniformity) and a different goal (interpretability and parameter sharing rather than adversarial robustness).

Recent work has further deepened the theoretical understanding of MHNs. Hu et al. (2024b) proposed nonparametric modern Hopfield models that extend the framework beyond exponential energy functions. Sargolzaei and Rueda (2024) demonstrated that MHNs can improve out-of-distribution data handling and corruption resistance, providing preliminary evidence for the robustness properties we exploit in Agatha. Farooq (2025) proposed a framework for non-linear attention via MHNs, showing that energy minima form stable context wells that encode contextual relationships.

These works focus primarily on capacity, convergence, and interpretability. Agatha leverages the MHN framework for a different purpose: the iterative convergence contracts perturbations in the forward pass, and the basin geometry deflects adversarial gradients in the backward pass. To our knowledge, Agatha is the first architecture to exploit MHN for adversarial robustness.

M.5 Gradient Masking and Obfuscated Gradients

Athalye et al. (2018) identified gradient masking as a pervasive failure mode of adversarial defenses, where models appear robust under gradient-based attacks simply because gradients are obfuscated. They proposed a taxonomy of obfuscated gradient types (shattered gradients, stochastic gradients,

vanishing/exploding gradients) and demonstrated that transfer attacks from undefended models can diagnose this failure. Since then, many proposed defenses have been shown to rely on gradient masking (Tramer et al., 2020). Agatha’s architectural design explicitly disrupts gradient flow, making the gradient masking concern particularly relevant. We address this in Section 5.5 through transfer attack analysis, confirming that the robustness is genuine.

M.6 Discrete Textual Attacks

A parallel line of work attacks NLP models through perturbations at the discrete text level rather than the continuous embedding space. Character-level attacks corrupt input tokens through insertion, deletion, swap, homoglyph substitution, or imperceptible Unicode manipulation (Pruthi et al., 2019; Li et al., 2018; Boucher et al., 2022). Word-level attacks substitute synonyms while preserving semantics (Alzantot et al., 2018), using gradient-free search or model-guided generation. The TextAttack framework (Morris et al., 2020) standardizes attack generation, data augmentation, and adversarial training under a common interface, and curated benchmarks such as AdvGLUE (Wang et al., 2021) aggregate human-validated adversarial examples across multiple attack types. Recent studies show that robustness varies across transformer encoders under identical attack protocols: Gidatkar et al. (2025) report attack success rates ranging from 0% to 94% across BERT, RoBERTa, and Flan-T5, highlighting that pretraining and architecture choices have substantial effects on text-level robustness.

Defenses in this space typically operate on the discrete input layer rather than internal representations, using input-level filtering, vocabulary mapping, or adversarial example augmentation during training. These approaches target the specific perturbation modalities they were designed for and require redesign as new attack types emerge.

Agatha is designed for embedding-space adversarial robustness but generalizes to representative discrete attacks. Section 5.6 reports CharSwap and TextBugger flip rates with consistent improvements, and Appendix C.5 reports AdvGLUE results. The robustness emerges through an indirect mechanism: discrete perturbations propagate through the tokenizer into embedding-space changes, which the energy basins then absorb. Full coverage of the textual-attack literature is beyond the scope of an embedding-space architectural defense, and we view defending against the full set

of textual attacks (including paraphrase, grammatical manipulation, and adversarial generation) as complementary to Agatha’s mechanism.