WHERE TO PERTURB? PERTURBATION ALLOCATION IN REPRESENTATION-SPACE AUGMENTATION FOR SPARSE EVENT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation-space augmentation can conflate how much an input is perturbed with where perturbation acts. We separate these factors in sparse event detection by calibrating post-normalization squared displacement and independently varying training allocation and test channel. In the frozen-representation CASIE setting under strong Gaussian stress, sentence-only and token-only training reverse their Trigger F1 ordering between the sentence and token test endpoints in 14 of 15 paired seeds. Averaged equally over five test allocations, however, sentence-only training has a 3.53-point advantage. Joint training improves the token-endpoint mean over sentence-only training while sacrificing performance at the sentence endpoint, illustrating a tradeoff rather than a uniform gain. Aggregate rankings therefore depend on the weights assigned to evaluation channels. Positive allocation–channel interactions persist across frozen-setting controls for distance, loss, encoder, and classifier, and extend to exact-span detection on PHEE. The corresponding interaction estimates are near zero under the tested end-to-end protocol. Local gradients show lower sensitivity to the trained-noisy path and higher sensitivity to the other path; permutation diagnostics provide complementary evidence consistent with a redistribution of path reliance. These results show why augmentation comparisons should report conditional robustness profiles alongside total displacement and specify the optimization and evaluation protocols under which they hold.

1 INTRODUCTION

Event detectors can combine a candidate token's contextualized representation with a pooled sentence representation. This exposes two paths at which training perturbations can act, raising a question about their allocation. We study this question in cybersecurity event detection, where threat reports are converted into events for analysis. In our filtered CASIE (Satyapanich et al., 2020) training split, fewer than 4% of tokens carry an event label (Section 4.1). This provides a sparse prediction setting for examining how the two paths respond to perturbation.

Representation-space perturbation is an established learning strategy (Zhu et al., 2020; Aghajanyan et al., 2021). Existing studies commonly vary whether perturbation is applied, its nominal scale, or its optimization. When a detector exposes both contextualized token representations and a pooled sentence representation, fixing nominal noise independently on each active path means that moving from one perturbed path to two changes both the allocation and the total realized displacement. An apparent path advantage may therefore reflect a larger perturbation budget rather than where that budget is placed. Magnitude and allocation must be separated before perturbation placement can be interpreted.

We ask: *at a fixed realized perturbation budget, how does training allocation between token and sentence paths change performance across test perturbation channels?* The same question can arise in other architectures with paired local and global representations; here we examine it through controlled event-detection comparisons.

To isolate these factors, we insert a modular augmentation interface between a frozen encoder and a lightweight classifier. It perturbs token and sentence representations without changing input text, labels, sample count, or class frequencies. We calibrate mean post-normalization squared

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

displacement and vary training and test allocations independently. Checkpoint and threshold selection use clean development data.

On CASIE, the relative advantage of sentence-only and token-only training reverses across the two test endpoints in 14 of 15 paired seeds under strong Gaussian stress. Sentence-only training nevertheless leads by 3.53 Trigger F1 points when five test allocations receive equal weight. This average is driven mainly by sentence-trained gains over no augmentation; it hides the opposing endpoint advantages. The full profile therefore matters when comparing training allocations under a specified evaluation mixture.

The interaction persists across frozen-setting controls and in PHEE exact-span detection, while its estimates are near zero under the tested end-to-end protocol. Across the frozen allocation grid, local gradient and path-permutation diagnostics reveal systematic shifts in relative path sensitivity, consistent with robustness being redistributed between token and sentence paths.

The study makes three contributions:

- We provide a budget-calibrated framework that independently controls total realized displacement, its token–sentence allocation, and the perturbation encountered at evaluation.
- We show that allocation changes the robustness profile at a fixed budget: intermediate and joint allocations redistribute gains across evaluation channels, and the evaluation-channel mixture determines the aggregate ranking.
- We connect these tradeoffs to systematic shifts in relative path sensitivity, and show that the resulting interaction persists across frozen-setting controls and a second corpus while changing substantially under end-to-end optimization.

2 Related Work

Cybersecurity event detection. Cybersecurity event extraction aims to identify event-bearing expressions and their associated structure from threat reports, security newswire, and related text. CASIE annotates cybersecurity event nuggets, arguments, roles, and realis attributes (Satyapanich et al., 2020), while CySecED expands the event inventory and emphasizes contextual evidence for trigger detection (Man Duc Trong et al., 2020). Subsequent work has improved event representations through feature interaction, sequence modeling, and few-shot adaptation (Han et al., 2025), type descriptions and contextual information (Liu et al., 2024), and syntax-aware representations (Hao et al., 2025). TextEE provides standardized preprocessing and splits for event-extraction evaluation (Huang et al., 2024).

Perturbation form and optimization. Input-space augmentation creates additional event-detection examples through masking and filling, targeted generation, or contrastive augmentation (Gao et al., 2022; Wang & Huang, 2024; Zhang et al., 2024a). Representation-space methods instead perturb internal features. Adversarial and consistency-based objectives include FreeLB, SMART, and SeqVAT (Zhu et al., 2020; Jiang et al., 2020; Chen et al., 2020), with conditional adversarial learning also studied for few-shot event detection (Zhang et al., 2024b). Sampled-noise approaches include R3F, HyPe, and NEFTune (Aghajanyan et al., 2021; Yuan et al., 2023; Jain et al., 2024); Noisy Feature Mixup combines perturbation with feature interpolation (Lim et al., 2021). These approaches motivate perturbing learned representations, while differing in the noise distribution, injection interface, and training objective.

Perturbation placement and budget control. Choosing an injection interface and choosing how to divide a fixed budget across simultaneously available paths are distinct questions. Our comparison holds the token–sentence interface fixed, calibrates mean post-normalization displacement, and varies training allocation independently of the test channel. It thus examines allocation–channel interactions under a common magnitude constraint, complementing studies of perturbation form and optimization. The distinction matters because activating a second path at a fixed per-path noise scale also changes total displacement.

Conventional regularizers such as virtual adversarial training and feature mixup provide context for augmentation gains (Miyato et al., 2019; Verma et al., 2019). The reference classifier uses low-rank

Hadamard pooling (Kim et al., 2017) as an architectural choice; concatenation-based controls test whether the allocation interaction depends on that choice.

3 Budget-Controlled Two-Path Augmentation

To separate perturbation magnitude from placement, we first fix a common representation interface, then calibrate total displacement, and finally vary training and test allocations independently. Freezing the encoder and projection keeps the underlying representations common across training arms. Both token and sentence inputs remain available in every arm; allocation changes where noise acts. Figure 1 summarizes this design.

3.1 Representation Interface

Given a sentence $X = (x_1, \ldots, x_n)$, the primary task assigns each candidate token one of five cybersecurity event types or the background label $\circ$. Let $\mathbf{e}_i \in \mathbb{R}^D$ be the frozen encoder state of candidate i's first subword. A fixed seeded orthonormal projection $P \in \mathbb{R}^{128\times D}$ gives

$$\mathbf{h}_i = \mathrm{Norm}(P\mathbf{e}_i), \qquad \mathbf{s} = \mathrm{Norm}\left(\frac{1}{n}\sum_{j=1}^{n}\mathbf{h}_j\right), \tag{1}$$

where $\mathrm{Norm}(\mathbf{z}) = \mathbf{z}/\|\mathbf{z}\|_2$. SecureBERT (Aghaei et al., 2023) supplies the primary encoder states. Let $\mathbf{x}_{h,i}$ and $\mathbf{x}_{s,i}$ denote the cached token and sentence representations after loading. Because finite-precision storage can slightly perturb their norms, clean inference uses the loaded values directly, whereas displacement is measured relative to their renormalized unit bases. The encoder, projection, and cache remain fixed. Storage precision and calibration details appear in Appendix A.5.

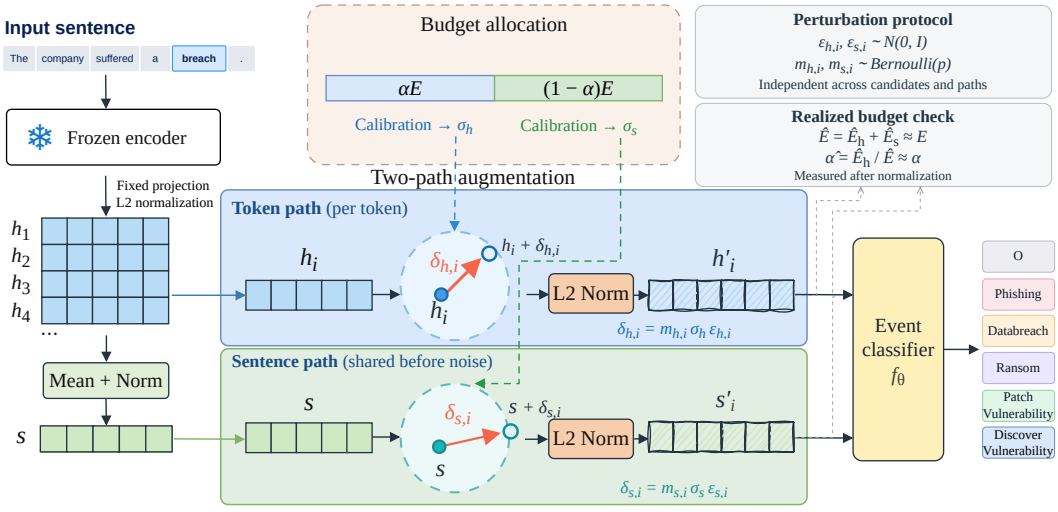

Figure 1: Budget-controlled two-path augmentation. Perturbations act between the frozen representations and classifier during training. Clean inference bypasses them; robustness evaluation perturbs the same interface while fixing the selected checkpoint and decision rule.

3.2 Pathwise Perturbation

For each candidate row, independent Gaussian vectors and Bernoulli gates perturb the two paths:

$$\begin{aligned} \mathbf{h}'_i &= \mathrm{Norm}(\mathbf{x}_{h,i} + m_{h,i}\sigma_h\boldsymbol{\varepsilon}_{h,i}), \quad \boldsymbol{\varepsilon}_{h,i} \sim \mathcal{N}(\mathbf{0}, I), \\ \mathbf{s}'_i &= \mathrm{Norm}(\mathbf{x}_{s,i} + m_{s,i}\sigma_s\boldsymbol{\varepsilon}_{s,i}), \quad \boldsymbol{\varepsilon}_{s,i} \sim \mathcal{N}(\mathbf{0}, I), \\ & m_{h,i}, m_{s,i} \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(p). \end{aligned} \tag{2}$$

Noise is added to the loaded values before normalization. The sentence feature is shared across candidates before perturbation, while gates and noise are sampled independently for each candidate and path. Training perturbations preserve the original labels. The fixed-scale reference recipe uses $p = 0.5$ and $\sigma_h = \sigma_s = 0.02$; our controlled comparisons instead recalibrate the active scales for each budget allocation. At $p = 0.5$, the probability of perturbing at least one input is 0.5 for a single active path and $1 - (1-p)^2 = 0.75$ for two active paths. Thus matching mean displacement does not match clean-row frequency or the distribution of per-row displacement. This affects mixed-versus-single-path comparisons, but not the gate frequency of the two single-path endpoints.

3.3 Realized budget and allocation

Fixing the noise scale on each active path gives a two-path arm a different total budget from a single-path arm. Normalization also makes displacement nonlinear in noise scale, so we calibrate total displacement across allocations. For equal input norms, dimensions, scales, and gate probabilities, isotropic Gaussian noise has the same expected normalized displacement on either path. For M candidate rows, let $\bar{\mathbf{h}}_i = \mathrm{Norm}(\mathbf{x}_{h,i})$ and $\bar{\mathbf{s}}_i = \mathrm{Norm}(\mathbf{x}_{s,i})$ denote the unit measurement bases corresponding to the loaded cache values. We measure

$$\widehat{E}_h = \frac{1}{M}\sum_{i=1}^{M} \|\mathbf{h}'_i - \bar{\mathbf{h}}_i\|_2^2, \quad \widehat{E}_s = \frac{1}{M}\sum_{i=1}^{M} \|\mathbf{s}'_i - \bar{\mathbf{s}}_i\|_2^2, \quad \widehat{E} = \widehat{E}_h + \widehat{E}_s. \tag{3}$$

The averages include rows with inactive gates. Hence $\widehat{E}$ is the mean squared displacement of the concatenated classifier input. For nonzero budgets, $\widehat{\alpha} = \widehat{E}_h/\widehat{E}$ is the realized token share; the target $\alpha = 0$ is sentence-only and $\alpha = 1$ token-only.

Let $\widetilde{E}_c(\sigma_c; p)$ denote the calibration approximation to mean squared displacement on channel $c \in \{h, s\}$, estimated from training-side vectors with a fixed Monte Carlo stream. For target budget E and token share α, we solve

$$\widetilde{E}_h(\sigma_h; p) = \alpha E, \qquad \widetilde{E}_s(\sigma_s; p) = (1-\alpha)E. \tag{4}$$

Here E is the target, $\widetilde{E}_c$ is its channelwise calibration approximation, and $\widehat{E}_c$ in Eq.(3) is the directly measured displacement. We write $E = \lambda E_0$, where $E_0 = \widetilde{E}_h(0.02; 0.5) + \widetilde{E}_s(0.02; 0.5)$ is the estimated reference-recipe displacement for that cache. Direct post-run measurements check $\widehat{E}$ and $\widehat{\alpha}$ against their targets. Cache precision, the numerical approximation, acceptance tolerances, and a complementary per-row exact-distance construction are specified in Appendix A.5.

3.4 Training and evaluation coordinates

Training perturbations are indexed by (E, α, p) and evaluation perturbations independently by $(T, \beta, p_{\mathrm{test}})$, where T is the test budget and β its token share. Thus $T = E$ and $\beta = \alpha$ match mean budget and allocation, not necessarily the complete perturbation distribution when gate probabilities differ. The main matrix uses $p = 0.5$ in training and $p_{\mathrm{test}} = 1$ in evaluation, with a separate gate-matched check.

The reference instantiation uses a low-rank interaction classifier and square-root inverse-frequency weighted cross-entropy (Appendix C). Within a comparison, cache, architecture, loss, optimizer, and selection rule are fixed. Models and decision thresholds are selected on clean development data and held fixed during clean and perturbed evaluation.

4 Experimental Setup

4.1 Data and metrics

We use a project-specific CASIE–MAVEN token conversion: CASIE supplies the cybersecurity documents and annotations (Satyapanich et al., 2020), while MAVEN supplies the token-level trigger-classification formulation (Wang et al., 2020). The labels are five event types (Databreach, DiscoverVulnerability, PatchVulnerability, Phishing, and Ransom) plus background ○; this is not CASIE's original span-overlap nugget evaluation. To reduce document contamination, we exclude

training documents whose normalized concatenated text also occurs in development or test, leaving 698/99/201 train/development/test documents and 314,905/45,998/87,317 labeled tokens. Positive tokens comprise 3.57%/3.47%/3.50%; full counts and filtering rules appear in Appendix D.

Trigger F1 merges all event types into one positive class; Typed F1 also requires the event type; Positive Macro-F1 averages F1 over the five positive types. All three metrics exclude background from the positive targets. The central CASIE tables report token-level Trigger and Typed F1; PHEE and the TextEE checks use exact-span Trigger-I/Trigger-C. PHEE uses TextEE split1 and the protocol in Appendix B.1.

4.2 Primary allocation protocol

All primary arms use the frozen 128-dimensional SecureBERT cache, the reference classifier and weighted objective, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch size 8192, and eight epochs.

At training budget E_0, we compare five allocations $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ and no augmentation across fifteen paired seeds. Evaluation crosses $T \in \{E_0, 8E_0\}$ with $\beta \in \{0, 0.25, 0.5, 0.75, 1\}$, averaging three shared noise draws per condition at $p_{\text{test}} = 1$. This comparison supplies Table 1 and Figure 2. A five-seed grid crosses training budgets $\{0.5, 1, 2\}E_0$ with test budgets $\{0.5, 1, 2, 8\}E_0$, using the same allocation shares and baseline. Its E_0 training runs are included in the primary comparison. Full-grid results, seed identifiers, and a gate-matched evaluation appear in Appendix A.

Within each seed, arms share prescribed initialization and random streams. Clean-development Typed F1 selects checkpoints, with Trigger F1 breaking ties; the same split calibrates the background/event decision margin.

Primary contrasts use paired-seed Student-t 95% intervals: draws are averaged within each condition, and allocation averages or endpoint contrasts are computed within seed before estimating uncertainty. Document-bootstrap intervals for the five-seed budget grid are reported in Appendix A.3. For the fifteen-seed primary comparison, absolute relative error in total perturbation budget remains below 0.552% across all augmented training epochs and 0.109% across test evaluations. Appendix A.5 provides the corresponding allocation errors and audit details.

4.3 Control experiments and evaluation settings

Frozen-setting controls vary the distance construction, loss, encoder, and classifier while retaining the allocation–channel comparison. PHEE adds an eight-seed cross-corpus check with its own training-derived budget and exact-span metrics (Appendix B). End-to-end ModernBERT jointly optimizes the encoder and classifier while retaining post-encoder perturbations (Appendix B.5); model-adaptive PGD probes adversarial directions (Appendix B.6). Additional fixed-scale, equal-displacement, regularizer, and classifier-capacity comparisons are reported in Appendices H, I, E, and G, respectively.

5 Results

5.1 Allocation changes the robustness profile

At a fixed training budget, the average sentence-only advantage hides a reversal across test channels. Table 1 reports the fifteen-seed comparison at training budget E_0. Sentence-only training exceeds token-only training by 1.13 Trigger F1 points at $T = E_0$ and 3.53 at $8E_0$, averaging the five test allocations equally after averaging three draws within each condition. Paired-seed 95% intervals are $[0.81, 1.44]$ and $[3.09, 3.97]$. The stress-budget gap comes mainly from sentence-only gains relative to no augmentation: $+3.75$ points versus $+0.22$ for token-only training.

Figure 2 shows why an average is insufficient. At $T = 8E_0$, sentence-minus-token is $+11.25$ points under sentence-only test perturbation ($\beta = 0$), but -1.63 under token-only perturbation ($\beta = 1$). The Trigger ordering reverses in 14 of 15 paired seeds; Typed F1 reverses in all 15. Equal joint training is 0.10 points above sentence-only at $T = E_0$ and 0.63 below it at $8E_0$. Intermediate allocations reshape the robustness profile rather than forming a monotonic ordering.

Table 1: CASIE token-level Trigger / Typed F1 (%): means over fifteen paired seeds. Augmented arms use training budget E_0; noisy columns average three draws within seed and condition, then weight five test allocations equally. Full profiles and paired-seed intervals are in Appendix A.1.

Training allocation	Clean	$T = E_0$	$T = 8E_0$
No augmentation	54.18 / 53.47	51.20 / 50.35	34.16 / 31.97
Sentence only ($\alpha = 0$)	54.52 / 53.94	52.03 / 51.35	37.91 / 36.11
Joint ($\alpha = 0.5$)	54.59 / 53.92	52.14 / 51.39	37.28 / 35.43
Token only ($\alpha = 1$)	53.85 / 53.15	50.91 / 50.12	34.38 / 32.43

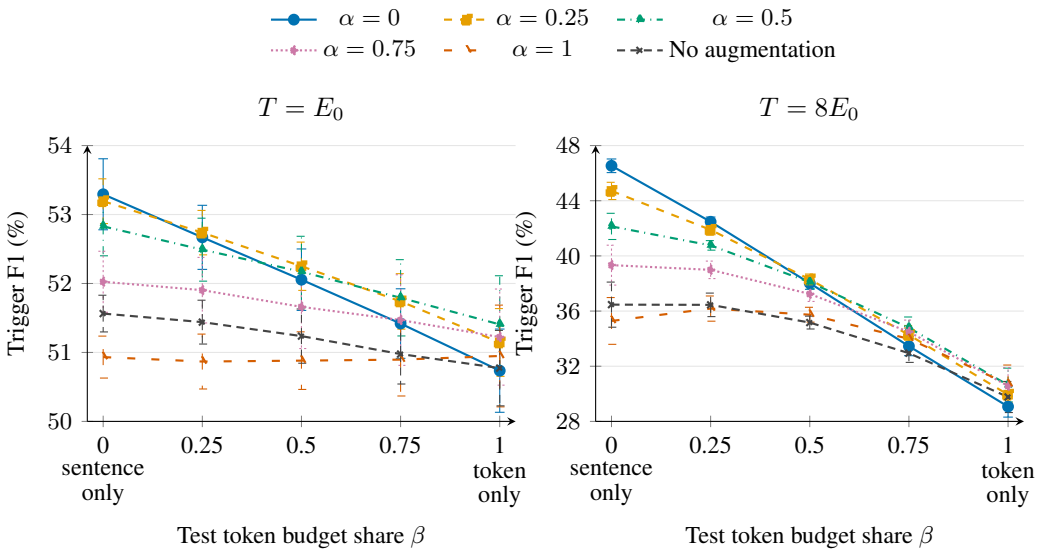

Figure 2: Allocation profiles at fixed training budget $E = E_0$: left, $T = E_0$; right, $T = 8E_0$. Each curve is a training token share α; the gray curve is no augmentation, and the horizontal axis varies test token share β. Training uses $p = 0.5$ and testing $p_{\text{test}} = 1$. Points average three noise draws within each of fifteen paired seeds, then average seeds. Lines connect the five tested allocations only. Error bars show one sample SD across the fifteen training seeds. The panels use different vertical scales to show the matched-budget detail and the larger stress-budget separation.

Joint training illustrates the gains and costs behind this average. At $T = 8E_0$, moving from sentence-only to equal joint training raises mean Trigger F1 at the token test endpoint from 29.07 to 30.64, but lowers it at the sentence endpoint from 46.53 to 42.14. Joint training therefore trades 4.40 points at one endpoint for 1.56 at the other. Its token-endpoint mean is close to token-only training (30.64 versus 30.71), while its sentence-endpoint mean is higher (42.14 versus 35.28). These endpoint means show how sharing the training budget reshapes the robustness profile. Full fifteen-seed profiles appear in Appendix A.1.

We summarize this separation with the endpoint interaction $I(T) = [F(0,0,T) - F(1,0,T)] - [F(0,1,T) - F(1,1,T)]$, where $F(\alpha, \beta, T)$ is Trigger F1 at training budget E_0. On the same fifteen seeds, $I(T)$ is 2.57 points ($[2.27, 2.88]$) at E_0 and 12.88 ($[11.72, 14.04]$) at $8E_0$. Positive interaction quantifies the change in relative performance across test endpoints; the reversal count separately records whether each seed changes ordering. Stronger Gaussian stress amplifies this allocation–channel tradeoff.

Evaluation weights determine the aggregate ranking. The five-channel average assumes equal importance for the tested conditions. To make this dependence explicit, consider weighting only the sentence and token test endpoints by w and $1 - w$. At $T = 8E_0$, the fifteen-seed mean sentence-minus-token training contrast, in Trigger F1 points, is

$$\overline{\Delta}(w) = 11.250w - 1.631(1 - w). \tag{5}$$

Table 2: Gaussian-test allocation–channel interactions at training E_0, in F1 points. All brackets are paired-seed 95% Student-t intervals. CASIE reports token Trigger F1; PHEE reports exact-span Trigger-I. Typed interactions and channel-average differences appear in Appendix B.

Setting	Seeds	$I(E_0)$	$I(8E_0)$ [95% interval]
Equal-distance CASIE	5	2.41	$12.48\ [8.66, 16.30]$
Unweighted CE CASIE	3	1.54	$8.49\ [4.96, 12.02]$
Frozen ModernBERT CASIE	3	1.54	$10.55\ [4.10, 16.99]$
Concat CASIE	3	2.15	$12.94\ [8.41, 17.46]$
PHEE (span F1)	8	0.12	$1.07\ [0.52, 1.61]$
End-to-end ModernBERT CASIE	8	0.08	$0.04\ [-1.34, 1.42]$

The mean ranking switches at $w \approx 0.127$: above this weight it favors sentence-only training, and below it token-only training. Here w weights test conditions in the summary, whereas β allocates perturbation within a condition. This is an average of condition-level F1 scores, not F1 recomputed from pooled predictions. Paired-seed intervals and the general five-condition weighting appear in Appendix A.2. Thus the equal-weight advantage summarizes one evaluation choice, while the profile exposes the gains and costs needed to compare allocations under other choices.

5.2 Direct controls of the allocation tradeoff

Frozen-setting controls. The positive stress interaction persists across the tested frozen-setting controls. Table 2 reports results with constant per-candidate training distance, ordinary cross-entropy, frozen ModernBERT (Warner et al., 2024), and a concat head. Trigger endpoint ordering reverses in all five equal-distance seeds, all three CE and ModernBERT seeds, and two of three concat seeds. The concat token-endpoint interval includes zero, illustrating that positive interaction need not imply a consistent reversal in every run. In the five-seed gate-matched check, matching the test gate to training while recalibrating its mean budget changes the $T = E_0$ channel-average sentence advantage from 1.11 to 0.72 points. Protocols and full estimates are in Appendices A and B.

The equal-distance control specifically addresses what matching the mean budget leaves uncontrolled. It removes Bernoulli gates and assigns every candidate fixed squared training displacements αE_0 and $(1-\alpha)E_0$ on the two paths, fixing their sum at E_0 up to numerical tolerance. Under Gaussian test stress at $8E_0$, the interaction remains 12.48 Trigger F1 points (95% interval $[8.66, 16.30]$), with endpoint ordering reversed in all five seeds. Thus Bernoulli gating and variation in per-candidate total training distance are not necessary for the interaction in this frozen setting (Appendix A.5).

Cross-corpus evidence. On PHEE split1 (Sun et al., 2022), eight paired seeds yield a stress interaction of 1.07 exact-span Trigger-I points ($[0.52, 1.61]$) and 1.18 Trigger-C points ($[0.61, 1.75]$). This extends the positive allocation–channel interaction to a second corpus, with a smaller magnitude. The channel-average sentence-minus-token contrast is -0.12 Trigger-I points ($[-0.25, 0.003]$), so this evidence concerns interaction rather than an aggregate sentence advantage (Appendix B.1).

End-to-end optimization. The positive interaction observed in the frozen setting is not evident under the tested end-to-end protocol. With jointly optimized ModernBERT and classifier parameters, eight paired seeds yield Trigger interactions of 0.08 points at E_0 ($[-0.13, 0.29]$) and 0.04 at $8E_0$ ($[-1.34, 1.42]$). The separate optimization schedule and a separate three-seed run are reported in Appendix B.5.

Adversarial evaluation. The five-seed frozen-checkpoint PGD check yields much lower F1 than its paired Gaussian evaluation despite nearly equal measured mean displacement: at E_0, sentence-trained Trigger F1 falls from 53.64 to 14.53 at the sentence endpoint. PGD imposes a per-vector cap; both the attack and its paired Gaussian comparator share sentence noise across tokens. Its Trigger interaction is 6.83 points at E_0 but -0.85 at $8E_0$; full scores, intervals, and attack settings are in Appendix B.6.

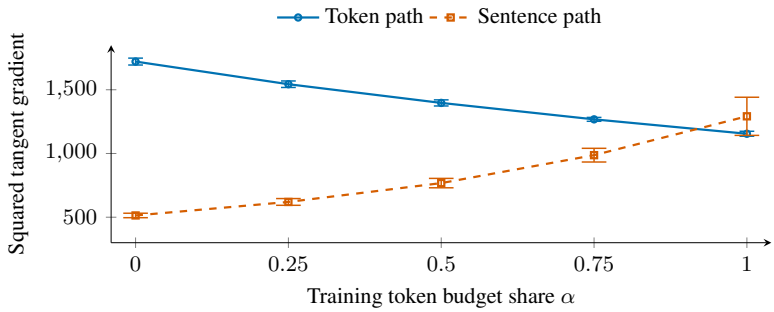

Figure 3: Path sensitivity of CASIE reference heads from the five-seed budget grid, trained at E_0 and evaluated on clean test inputs. Squared tangent-gradient norms of the gold-event-minus-background logit are averaged over 3,054 positive tokens within model, then over five seeds. Bars show one sample SD across seeds; lines connect the five measured allocations.

5.3 Allocation redistributes path sensitivity

We measure local sensitivity of the gold-event-minus-background logit margin z to each input path. For input $\mathbf{u}$, let $\mathbf{g} = \nabla_{\mathbf{u}} z$ and $\bar{\mathbf{u}} = \mathbf{u}/\|\mathbf{u}\|_2$. The diagnostic is

$$S(\mathbf{u}) = \left\|\mathbf{g} - (\mathbf{g}^\top \bar{\mathbf{u}})\bar{\mathbf{u}}\right\|_2^2. \tag{6}$$

Small normalized perturbations change direction along the unit sphere's tangent plane to first order, so this projected gradient measures local margin sensitivity to those directions. We average over positive tokens within each model and then over seeds; the magnitude depends on logit scale. As a complementary diagnostic, we permute test candidate rows in one feature path while holding the other path fixed, and measure the drop in clean Trigger F1 at the selected decision threshold.

Training allocation changes local gradient sensitivity in opposite directions across the two paths. Increasing the training token share from zero to one lowers token-path squared tangent sensitivity from 1721.1 to 1154.8, while sentence-path sensitivity rises from 513.8 to 1291.7 (Figure 3). Sentence-path permutation costs 12.62 versus 20.29 Trigger F1 points at these endpoints. ModernBERT shows the same directions in both gradient sensitivity and sentence-permutation cost. Together, these diagnostics indicate a systematic shift in relative path reliance away from the trained-noisy path. Appendix B reports the complete results, including token-shuffling floor effects.

6 Discussion

Perturbation allocation shapes a conditional robustness profile. A training allocation can improve tolerance at one test endpoint while losing ground at the other, even when total displacement is held fixed.

The positive endpoint interaction persists across the frozen-setting controls and PHEE. The aggregate ordering of training allocations depends on how evaluation channels are weighted, and full profiles make this dependence explicit (Appendix A.2).

Redistribution of path reliance offers an interpretation of this tradeoff. Figure 3 shows lower sensitivity to the trained-noisy path and higher sensitivity to the other path; sentence-path permutation and the ModernBERT endpoints give complementary evidence. End-to-end training allows representations to adapt along with the classifier, which may alter this response. Its near-zero interaction estimates are compatible with that explanation, although differences in projection and optimization prevent attributing the change to encoder adaptation alone.

The practical implication is to measure allocation as well as magnitude. Calibrate realized displacement after normalization, vary training and test allocation independently, and report both the condition curves and the weights used to aggregate them. Also specify the perturbation distribution: the PGD check produces much lower F1 than Gaussian evaluation, while its per-candidate constraint differs

from the Gaussian mean-budget target. A nominal budget alone does not determine the performance response.

6.1 Limitations

PHEE uses eight training seeds on one data split, so its intervals characterize within-split training variability. Gaussian and PGD interventions act on representations; robustness to natural text changes and certified worst-case guarantees remain untested.

Equal-distance training controls gating and total-distance variability, while other noise properties can still differ. Local gradients and feature permutations provide observational evidence of path reliance, without identifying causal mediation. Permutation breaks feature compatibility, and token shuffling has floor effects.

The experimental designs were informed by CASIE test and PHEE clean observations; the intervals do not account for this design selection.

7 Conclusion

At fixed realized displacement, perturbation allocation produces measurable conditional robustness profiles in sparse event detection. Frozen-setting controls and PHEE show positive allocation–channel interactions, while end-to-end and adversarial checks demonstrate the importance of optimization and evaluation protocols. Path-sensitivity diagnostics are consistent with a redistribution of sensitivity across the two representation paths in the frozen setting. The framework offers a reusable comparison strategy: separate perturbation magnitude, training allocation, and test conditions, and report the gains and costs across the resulting profile.

AI Use Statement

Generative AI tools assisted with literature discovery, experimental implementation and debugging, and translation. The authors take full responsibility for the final manuscript, claims, code, and artifacts, including AI-assisted content.

Reproducibility Statement

Code and reproduction materials are available through the anonymous repository `https://anonymous.4open.science/r/perturbation-allocation-repro-BBA5/`. The base archive supports cache-level reproduction of the five-seed budget grid and packaged frozen-setting controls. The `evidence/` directory supports statistical recomputation for the fifteen-seed primary comparison, eight-seed PHEE comparison, end-to-end checks, and PGD and drift diagnostics. Appendix J and the repository README specify the available artifacts, reproduction scope, environment, and commands.

Ethics Statement

The study extracts from cybersecurity text and uses synthetic representation perturbations rather than executable attacks. Detector errors can affect downstream analysis, so the reported experiments do not establish readiness for autonomous security decisions, and any release must respect the licences and redistribution conditions of the source corpora and models.

References

Ehsan Aghaei, Xi Niu, Waseem G. Shadid, and Ehab Al-Shaer. SecureBERT: A domain-specific language model for cybersecurity. In *Security and Privacy in Communication Networks*, pp. 39–56. 2023. doi: 10.1007/978-3-031-25538-0_3.

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse. In *International Confer-*

ence on Learning Representations, 2021. URL https://openreview.net/forum?id=OQ08SN70M1V.

Luoxin Chen, Weitong Ruan, Xinyue Liu, and Jianhua Lu. SeqVAT: Virtual adversarial training for semi-supervised sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8801–8811, 2020. doi: 10.18653/v1/2020.acl-main.777. URL https://aclanthology.org/2020.acl-main.777/.

Jun Gao, Changlong Yu, Wei Wang, Huan Zhao, and Ruifeng Xu. Mask-then-fill: A flexible and effective data augmentation framework for event extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4537–4544, 2022. doi: 10.18653/v1/2022.findings-emnlp.332. URL https://aclanthology.org/2022.findings-emnlp.332/.

Yue Han, Weihong Han, Aiping Li, and Shudong Li. Cyberattack event and arguments extraction based on feature interaction and few-shot learning. *Scientific Reports*, 15:31808, 2025. doi: 10.1038/s41598-025-15138-x. URL https://www.nature.com/articles/s41598-025-15138-x.

Anran Hao, Jian Su, Shuo Sun, and Teo Yong Sen. Soft syntactic reinforcement for neural event extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9466–9478, 2025. doi: 10.18653/v1/2025.naacl-long.479. URL https://aclanthology.org/2025.naacl-long.479/.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Premkumar Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. TextEE: Benchmark, reevaluation, reflections, and future challenges in event extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12804–12825, 2024. doi: 10.18653/v1/2024.findings-acl.760. URL https://aclanthology.org/2024.findings-acl.760/.

Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Neftune: Noisy embeddings improve instruction finetuning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 17566–17591, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/4bdeeaeb380b35302bbda1823d328c22-Paper-Conference.pdf.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2177–2190, 2020. doi: 10.18653/v1/2020.acl-main.197. URL https://aclanthology.org/2020.acl-main.197/.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=r1rhWnZkg.

Soon Hoe Lim, N. Benjamin Erichson, Francisco Utrera, Winnie Xu, and Michael W. Mahoney. Noisy feature mixup, 2021. URL https://arxiv.org/abs/2110.02180.

Boyang Liu, Guozheng Rao, Xin Wang, Li Zhang, and Qing Cong. DE3TC: Detecting events with effective event type information and context. *Neural Processing Letters*, 56:89, 2024. doi: 10.1007/s11063-024-11570-8. URL https://link.springer.com/article/10.1007/s11063-024-11570-8.

Hieu Man Duc Trong, Duc Trong Le, Amir Pouran Ben Veyseh, Thuat Nguyen, and Thien Huu Nguyen. Introducing a new dataset for event detection in cybersecurity texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 5381–5390, 2020. doi: 10.18653/v1/2020.emnlp-main.433. URL https://aclanthology.org/2020.emnlp-main.433/.

Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, Aug 2019. ISSN 1939-3539. doi: 10.1109/tpami.2018.2858821. URL https://doi.org/10.1109/tpami.2018.2858821.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. CASIE: Extracting cybersecurity event information from text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 8749–8757, 2020. doi: 10.1609/AAAI.V34I05.6401.

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. PHEE: A dataset for pharmacovigilance event extraction from text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5571–5587, 2022. doi: 10.18653/v1/2022.emnlp-main.376. URL https://aclanthology.org/2022.emnlp-main.376/.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6438–6447, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/verma19a.html.

Sijia Wang and Lifu Huang. Targeted augmentation for low-resource event extraction. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4414–4428, 2024. doi: 10.18653/v1/2024.findings-naacl.275. URL https://aclanthology.org/2024.findings-naacl.275/.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. MAVEN: A massive general domain event detection dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 1652–1671, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.129.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL https://arxiv.org/abs/2412.13663.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. HyPe: Better pre-trained language model fine-tuning with hidden representation perturbation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3246–3264, July 2023. doi: 10.18653/v1/2023.acl-long.182. URL https://aclanthology.org/2023.acl-long.182/.

Chenlong Zhang, Pengfei Cao, Yubo Chen, Kang Liu, Zhiqiang Zhang, Mengshu Sun, and Jun Zhao. Continual few-shot event detection via hierarchical augmentation networks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 3868–3880, 2024a. URL https://aclanthology.org/2024.lrec-main.342/.

Fangchen Zhang, Shengwei Tian, Long Yu, and Qimeng Yang. Multi-channels prototype contrastive learning with condition adversarial attacks for few-shot event detection. *Neural Processing Letters*, 56(1):31, 2024b. doi: 10.1007/s11063-024-11515-1. URL https://doi.org/10.1007/s11063-024-11515-1.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BygzbyHFvB.

A Channel Allocation at Matched Mean Budgets

The budget-controlled comparison uses the document-filtered CASIE token conversion, the frozen SecureBERT cache, and the reference bilinear classifier. It is separate from both the eleven-seed fixed-noise results in Table 33 and the three-seed control in Appendix I. The design was informed by test-set observations and therefore does not constitute evaluation on a previously unseen test set.

Budgets and comparisons. For normalized token and sentence vectors u_t, u_s, the training budget is the mean sum of squared displacements after normalization,

$$E = \mathbb{E}\left[\|u'_t - u_t\|_2^2 + \|u'_s - u_s\|_2^2\right].$$

Training-only calibration defines E_0 from the reference two-path Gaussian recipe ($\sigma = 0.02$, independent Bernoulli gates with probability 0.5). Within each experimental setting, training and evaluation budgets are expressed relative to the same cache-specific reference budget E_0. The token path receives fraction α of the budget and the sentence path receives $1-\alpha$. We fit $E/E_0 \in \{0.5, 1, 2\}$ and $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ using seeds 20260814–20260818, plus a no-augmentation arm for each seed: 80 Gaussian-study fits in total. Paired arms share initialization and prescribed random streams. Each fit uses eight epochs; checkpoint and threshold selection use clean development data only.

The test matrix crosses four budgets ($T/E_0 = 0.5, 1, 2, 8$) with five token allocation fractions ($\beta = 0, 0.25, 0.5, 0.75, 1$). Three shared noise draws per cell give 4,800 evaluations at fixed development-selected thresholds. Test sigmas are calibrated on training vectors and frozen. Main-matrix test noise has no gate; A gate-matched check comprises 375 evaluations at $T = E_0$ with gate probability 0.5. Matching $E = T$ and $\alpha = \beta$ therefore matches mean budget and allocation, not the full perturbation distribution.

Primary results. Table 1 reports the primary allocation means at $E = E_0$: sentence-only minus token-only, averaged equally over all five test allocations, separately for $T = E_0$ and $T = 8E_0$. F1 is computed for each draw before averaging draws within a seed and condition, then averaging allocations. The primary comparison includes all five E_0 training allocations and a no-augmentation baseline across fifteen seeds. Evaluation crosses two test budgets (E_0 and $8E_0$), five test allocation shares, and three noise draws, yielding 2,700 fixed-model evaluations. The broader budget grid uses five seeds. Means, SDs, and paired Student-t intervals for the primary comparison are computed from fifteen seed-level estimates. The endpoint interaction is a post hoc analysis motivated by inspection of the five-seed matrix; it was not preregistered.

A.1 Primary 15-seed details

We summarize 15 paired training seeds (20260814–20260828). For each seed and test condition, we first average three fixed noise draws (2026091801–2026091803). Budget-level summaries then average $\beta \in \{0, 0.25, 0.5, 0.75, 1\}$ with equal weights. Intervals are unadjusted 95% Student-t intervals across training seeds; contrasts are computed within seed before constructing the interval. Clean scores are evaluated once per fitted model. Training share $\alpha = 0$ denotes sentence-only perturbation and $\alpha = 1$ denotes token-only perturbation; β is the test token-channel share. For each metric, define the sentence-minus-token contrast $D_\beta(T) = F(0, \beta, T) - F(1, \beta, T)$, where F is that metric's F1 and T is the test budget.

All five training allocations $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ and no augmentation have 15 seeds at training budget E_0. Evaluation uses test budgets E_0 and $8E_0$. The separate five-seed grid covers training budgets $\{0.5, 1, 2\}E_0$ and test budgets $\{0.5, 1, 2, 8\}E_0$. Its five seeds overlap the first five primary seeds, so the two analyses are not independent replications. Here Trigger and Typed are token-level F1 scores, not exact-span scores.

A.2 Evaluation mixture weights

For test budget T, define the weighted allocation contrast

$$\Delta_{\boldsymbol{w}}(T) = \sum_{\beta \in \mathcal{B}} w_\beta \left[F(0, \beta, T) - F(1, \beta, T)\right], \qquad w_\beta \geq 0, \quad \sum_{\beta \in \mathcal{B}} w_\beta = 1, \tag{7}$$

Table 3: F1 (percent); mean [95% CI]. Corrupted scores average five test allocations.

Test budget	Training arm	Trigger	Typed
Clean	No augmentation	54.182 [54.005, 54.360]	53.474 [53.320, 53.628]
Clean	$\alpha = 0$	54.517 [54.246, 54.789]	53.944 [53.742, 54.145]
Clean	$\alpha = 0.25$	54.609 [54.371, 54.847]	53.992 [53.806, 54.178]
Clean	$\alpha = 0.5$	54.592 [54.268, 54.916]	53.922 [53.662, 54.182]
Clean	$\alpha = 0.75$	54.319 [53.974, 54.665]	53.654 [53.360, 53.949]
Clean	$\alpha = 1$	53.845 [53.544, 54.147]	53.152 [52.876, 53.428]
$1E_0$	No augmentation	51.198 [50.990, 51.405]	50.347 [50.176, 50.518]
$1E_0$	$\alpha = 0$	52.035 [51.760, 52.309]	51.351 [51.125, 51.577]
$1E_0$	$\alpha = 0.25$	52.213 [52.011, 52.416]	51.501 [51.327, 51.676]
$1E_0$	$\alpha = 0.5$	52.138 [51.851, 52.424]	51.388 [51.141, 51.634]
$1E_0$	$\alpha = 0.75$	51.656 [51.344, 51.969]	50.910 [50.639, 51.181]
$1E_0$	$\alpha = 1$	50.905 [50.653, 51.158]	50.118 [49.890, 50.345]
$8E_0$	No augmentation	34.160 [33.787, 34.533]	31.970 [31.582, 32.359]
$8E_0$	$\alpha = 0$	37.911 [37.677, 38.146]	36.113 [35.912, 36.314]
$8E_0$	$\alpha = 0.25$	37.822 [37.589, 38.054]	36.006 [35.802, 36.210]
$8E_0$	$\alpha = 0.5$	37.284 [37.010, 37.557]	35.427 [35.186, 35.668]
$8E_0$	$\alpha = 0.75$	36.124 [35.809, 36.439]	34.276 [33.969, 34.584]
$8E_0$	$\alpha = 1$	34.379 [34.012, 34.745]	32.434 [32.057, 32.811]

Table 4: Paired gains over no augmentation (percentage points); mean [95% CI]. Corrupted scores average five test allocations.

Test budget	Training arm	Trigger	Typed
$1E_0$	$\alpha = 0$	0.837 [0.662, 1.011]	1.004 [0.864, 1.144]
$1E_0$	$\alpha = 1$	-0.293 [-0.577, -0.008]	-0.230 [-0.481, 0.022]
$8E_0$	$\alpha = 0$	3.751 [3.343, 4.160]	4.143 [3.746, 4.540]
$8E_0$	$\alpha = 1$	0.219 [0.017, 0.420]	0.464 [0.302, 0.626]

Table 5: Paired sentence-minus-token contrasts averaged over five test allocations and endpoint interaction $I(T) = D_0(T) - D_1(T)$, in percentage points [95% CI]. Reversals count seeds with $D_0(T) > 0$ and $D_1(T) < 0$.

Budget	Metric	Mean $D_\beta(T)$	Interaction I	Reversals
$1E_0$	Trigger	1.129 [0.814, 1.445]	2.574 [2.269, 2.880]	9/15
$1E_0$	Typed	1.234 [0.959, 1.508]	2.813 [2.490, 3.135]	11/15
$8E_0$	Trigger	3.533 [3.094, 3.972]	12.881 [11.719, 14.044]	14/15
$8E_0$	Typed	3.679 [3.270, 4.088]	14.596 [13.356, 15.836]	15/15

Table 6: Full test-allocation profiles at $1E_0$. Columns give β. Values are F1 percent, mean $\pm$ sample SD across 15 seeds after averaging the three draws.

Arm	Metric	0	0.25	0.5	0.75	1
No augmentation	Trigger	51.564 ± 0.266	51.440 ± 0.318	51.237 ± 0.393	50.977 ± 0.434	50.773 ± 0.547
No augmentation	Typed	50.714 ± 0.258	50.567 ± 0.267	50.366 ± 0.310	50.120 ± 0.365	49.969 ± 0.483
$\alpha = 0$	Trigger	53.294 ± 0.515	52.669 ± 0.465	52.055 ± 0.446	51.419 ± 0.504	50.736 ± 0.604
$\alpha = 0$	Typed	52.671 ± 0.413	51.995 ± 0.374	51.363 ± 0.359	50.706 ± 0.431	50.020 ± 0.530
$\alpha = 0.25$	Trigger	53.194 ± 0.324	52.737 ± 0.322	52.250 ± 0.350	51.740 ± 0.402	51.145 ± 0.491
$\alpha = 0.25$	Typed	52.508 ± 0.271	52.029 ± 0.278	51.529 ± 0.307	51.015 ± 0.358	50.425 ± 0.431
$\alpha = 0.5$	Trigger	52.829 ± 0.427	52.490 ± 0.456	52.171 ± 0.514	51.792 ± 0.553	51.406 ± 0.706
$\alpha = 0.5$	Typed	52.083 ± 0.338	51.730 ± 0.400	51.406 ± 0.453	51.044 ± 0.492	50.675 ± 0.631
$\alpha = 0.75$	Trigger	52.025 ± 0.443	51.903 ± 0.475	51.661 ± 0.604	51.470 ± 0.657	51.223 ± 0.697
$\alpha = 0.75$	Typed	51.266 ± 0.371	51.134 ± 0.407	50.889 ± 0.536	50.719 ± 0.598	50.544 ± 0.619
$\alpha = 1$	Trigger	50.933 ± 0.304	50.868 ± 0.398	50.881 ± 0.418	50.896 ± 0.527	50.949 ± 0.738
$\alpha = 1$	Typed	50.095 ± 0.310	50.057 ± 0.367	50.070 ± 0.388	50.110 ± 0.476	50.257 ± 0.671

Table 7: Full test-allocation profiles at $8E_0$. Columns give β. Values are F1 percent, mean $\pm$ sample SD across 15 seeds after averaging the three draws.

Arm	Metric	0	0.25	0.5	0.75	1
No augmentation	Trigger	36.466 ± 1.636	36.451 ± 0.853	35.166 ± 0.466	32.947 ± 0.669	29.770 ± 1.128
No augmentation	Typed	34.126 ± 1.736	34.076 ± 0.979	32.961 ± 0.513	30.918 ± 0.568	27.770 ± 0.985
$\alpha = 0$	Trigger	46.534 ± 0.494	42.475 ± 0.343	38.021 ± 0.444	33.453 ± 0.690	29.074 ± 0.763
$\alpha = 0$	Typed	45.223 ± 0.564	40.868 ± 0.316	36.171 ± 0.363	31.399 ± 0.574	26.904 ± 0.678
$\alpha = 0.25$	Trigger	44.711 ± 0.626	41.891 ± 0.358	38.297 ± 0.388	34.272 ± 0.679	29.937 ± 0.937
$\alpha = 0.25$	Typed	43.162 ± 0.707	40.163 ± 0.347	36.461 ± 0.317	32.326 ± 0.593	27.917 ± 0.841
$\alpha = 0.5$	Trigger	42.139 ± 0.950	40.763 ± 0.346	38.093 ± 0.465	34.788 ± 0.782	30.635 ± 1.226
$\alpha = 0.5$	Typed	40.345 ± 1.044	38.848 ± 0.382	36.244 ± 0.370	32.942 ± 0.646	28.756 ± 1.046
$\alpha = 0.75$	Trigger	39.330 ± 1.447	38.989 ± 0.634	37.241 ± 0.542	34.492 ± 0.854	30.569 ± 1.330
$\alpha = 0.75$	Typed	37.282 ± 1.587	36.972 ± 0.771	35.427 ± 0.441	32.792 ± 0.726	28.909 ± 1.159
$\alpha = 1$	Trigger	35.284 ± 1.696	36.180 ± 0.914	35.736 ± 0.538	33.987 ± 0.788	30.705 ± 1.381
$\alpha = 1$	Typed	32.913 ± 1.869	33.918 ± 1.037	33.812 ± 0.531	32.336 ± 0.655	29.191 ± 1.235

where $\mathcal{B} = \{0, 0.25, 0.5, 0.75, 1\}$ and $F(\alpha, \beta, T)$ is F1 for training allocation α, test allocation β, and test budget T. Here β is the token-channel share of the perturbation budget within a single test condition; w_β instead weights different test conditions in the evaluation summary. The five-allocation mean uses $w_\beta = 1/5$. We form each contrast within seed, after draw averaging, and then average across seeds. These are weighted F1 scores, not F1 recomputed from pooled predictions.

Only in the endpoint special case, set $w_0 = w$, $w_1 = 1 - w$, and all other weights to zero. With $D_b(T) = F(0, b, T) - F(1, b, T)$,

$$\overline{\Delta}_w(T) = w\overline{D}_0(T) + (1 - w)\overline{D}_1(T), \qquad w^*(T) = -\frac{\overline{D}_1(T)}{\overline{D}_0(T) - \overline{D}_1(T)}. \tag{8}$$

The critical weights are descriptive point estimates. At E_0, $w^* = 0.082782$ (Trigger) and 0.083981 (Typed); at $8E_0$, $w^* = 0.126654$ (Trigger) and 0.156629 (Typed). In all four cases, the endpoint-weighted mean favors sentence-only training when $w > w^*$.

At E_0, the token-endpoint $D_1(T)$ intervals include zero; the negative endpoint mean alone does not establish a confidently reversed ordering. At $8E_0$, the Trigger endpoint differences are 11.250 [10.454, 12.045] and -1.631 [-2.412, -0.851] percentage points at $\beta = 0$ and $\beta = 1$, respectively.

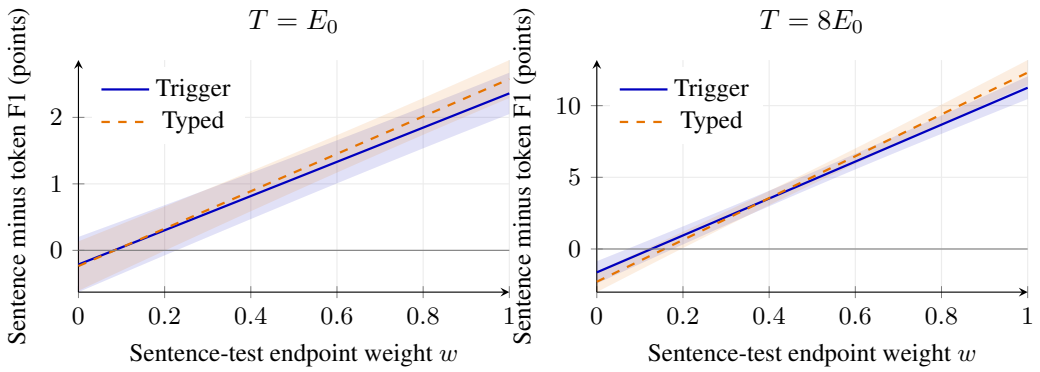

Figure 4: Endpoint reweighting of fifteen-seed contrasts. Shading shows pointwise paired-seed 95% t intervals. Only $\beta = 0, 1$ are mixed; w is their aggregation weight, not a perturbation share.

Within the five-seed budget grid, the sentence-minus-token difference, averaged over test allocation shares and noise draws, is positive in every seed at both primary test budgets. At $T = 8E_0$, sentence-only and token-only training obtain mean F1 of 46.21 and 34.34 under sentence-only test noise ($\beta = 0$), but 29.38 and 31.01 under token-only test noise ($\beta = 1$). The direction reverses in every paired seed between these endpoints. Joint-minus-sentence differences have mixed signs at $T = E_0$ and are negative in all five seeds at $T = 8E_0$. These results show that allocation rankings depend on test channel and budget. The eightfold test budget probes an untrained Gaussian noise intensity in this grid.

Five-seed test matrix. Table 8 gives every main-matrix cell for the primary $E = E_0$ training arms, including the intermediate allocations. Each entry averages three draws within a seed and then five seeds; means are shown without repeating the seed-level SDs in the main table. The verified machine-readable artifact additionally contains the $E = 0.5E_0$ and $2E_0$ training arms, for 4,800 cells in total.

Confirmation and scope. A gate-matched evaluation repeats $T = E_0$ with $p_{\text{test}} = 0.5$. Table 9 shows that this changes magnitudes but not the aggregate endpoint ordering: sentence minus token is $+1.11$ points without test gating and $+0.72$ with matched gating; joint minus sentence is $+0.07$ and $+0.22$, respectively.

A separate 15-fit control uses per-sample fixed-length perturbations at $E = E_0$ and $\alpha \in \{0, 0.5, 1\}$. Clean F1 is 54.42 ± 0.73, 54.63 ± 0.44, and 54.05 ± 0.29, respectively. Appendix B reports Gaussian test-matrix evaluations of these same checkpoints. The fixed length applies to training; test noise remains Gaussian.

A.3 DOCUMENT- AND SEED-LEVEL UNCERTAINTY

We quantify uncertainty from document sampling and training randomness using bootstrap analyses of five SecureBERT seeds and three ModernBERT seeds. Paired-seed intervals for the fifteen-seed primary comparison are reported in Appendix A.1. Let $F(\alpha, \beta, T)$ denote token Trigger F1 at training budget E_0. The primary contrast averages $F(0, \beta, T) - F(1, \beta, T)$ equally over the five β values. The secondary interaction is

$$I(T) = [F(0, 0, T) - F(1, 0, T)] - [F(0, 1, T) - F(1, 1, T)].$$

We compute document-level TP/FP/FN from aligned predictions. The 201 test documents and their multiplicities are shared across all arms, seeds, budgets, test allocations, and noise draws in each of 10,000 bootstrap resamples (seed 2026092001). Each resample recomputes micro-F1 from summed counts, then averages the three draws, test allocations where appropriate, and training seeds. It does not average per-document F1. Document intervals condition on the fitted seeds and the three fixed draws.

Separately, 10,000 paired seed resamples (seed 2026092002) vary training runs conditional on this corpus and these draws. A crossed analysis independently resamples the seed and shared-document axes. Noise draws are averaged within each seed and condition and never treated as extra training seeds. All intervals are percentile 95% intervals, with no multiplicity adjustment; three or five training seeds limit resolution of training variability.

In the five-seed SecureBERT comparison, $I(8E_0) - I(E_0) = 10.96$ points, with a paired document interval $[10.10, 11.84]$ and seed interval $[9.28, 12.67]$; values are computed from unrounded predictions. ModernBERT's matched-budget aggregate difference has a crossed interval spanning zero, while its interaction intervals remain positive. Neither interval family is uniformly wider.

A.4 DIRECT ALLOCATION MATRIX WITH MODERNBERT

The nine ModernBERT (Warner et al., 2024) models use the same CASIE conversion, bilinear head, training budget E_0, and allocations $\alpha \in \{0, 0.5, 1\}$, with three paired seeds (20260814–20260816). Their clean-development-selected checkpoints and thresholds remain fixed during evaluation. Gaussian test scales use the stored ModernBERT-specific calibration. The two budgets and five test allocations use the same three prescribed draw seeds and evaluation batch size as the primary matrix, with no test-based model selection. Maximum realized test-budget error is 0.149%; maximum token-share error is 0.000456.

The equal-allocation averages for sentence/joint/token training are $46.15/46.23/45.67$ at $T = E_0$ and $33.95/33.21/31.03$ at $8E_0$. All three paired seeds reverse their endpoint ranking under stress. This replicates the allocation–channel tradeoff on a different representation, with different effect magnitudes. It remains a same-corpus, three-seed check. The reproduction archive provides the model assets, frozen caches, and construction settings.

Reconstruction scope and numerical differences. All three ModernBERT splits can be rebuilt from converted text using the identified weights, tokenizer, and projection. The rebuilt arrays are not bitwise identical: the maximum coordinate difference is 2.44×10^{-4} and the maximum array RMS difference is 2.13×10^{-6}. With 12 heads held fixed, the maximum clean F1 difference is 0.0086 points; across the 270 shared noisy cells, the maximum Trigger/Typed F1 difference is 0.0269 points. The rebuilt-cache Trigger interactions are 1.538 and 10.545 at E_0 and $8E_0$, respectively. These measurements assess extraction and fixed-head evaluation. Environment versions and implementation checks are documented in the reproduction repository.

A.5 CALIBRATION AND NUMERICAL IMPLEMENTATION

Cache precision and measurement base. The cache stores projected token vectors and normalized sentence means in float16; loading casts them to float32 without re-normalizing. Gaussian noise is added to these loaded values x, then normalized. Displacement is measured from $\bar{u} = \mathrm{Norm}(x)$, so the implemented operator is $\mathrm{Norm}(x + m\sigma\varepsilon)$, equivalently $\mathrm{Norm}(\bar{u} + m(\sigma/\|x\|_2)\varepsilon)$. Thus the pre-noise cache value and the unit measurement base differ by the stored radius. The training

calibration samples these pre-normalization radii, while the ungated test-scale solver uses unit radii. Clean inference bypasses the perturbation operator and uses the loaded cache values.

The projection was generated from a CPU Gaussian matrix with seed 20260814 and reduced QR. Because the SecureBERT extraction checkpoint is unavailable, reproduction starts from the supplied frozen cache arrays and row maps. The repository documents the cache and row-map checksums.

Calibration population and solver. The recorded SecureBERT calibration samples the unfiltered training cache's 361,117 candidate rows (14,474 unique sentences), before document filtering; sentence radii are mapped to candidate rows before sampling. The training fits and realized-budget measurements instead use the 314,905 retained rows. No development or test split arrays or labels enter the calculation, but training documents with text overlapping those splits do enter this unfiltered calibration population. Repeating the same solver using only the 314,905 retained training rows changes E_0 by -8.95×10^{-8} relatively (SecureBERT) and -1.21×10^{-7} (ModernBERT); the maximum absolute noise-scale change is 5.46×10^{-9}. The recorded scales remain fixed for replay. This audit isolates the population change, not the Gaussian approximation error discussed below. The fixed primary stream uses 262,144 Monte Carlo draws with seed 2026091819; a separate stream uses 1,048,576 draws with seed 2026091821. The solver is Brent's method, with absolute and relative root tolerances of 10^{-11}.

The stored approximation draws $a \sim \mathcal{N}(0,1)$ and $q \sim \chi^2_{128}$ independently in the normalized-displacement expression. For an exact isotropic Gaussian reduction, the squared noise norm would instead be $a^2 + b$, with $b \sim \chi^2_{127}$ independent of a. We therefore treat the stored computation as an approximation, not an exact Gaussian expectation; its independent random-stream check repeats that approximation. The reported scales and experimental results retain this recorded implementation.

Measured budgets and numerical tolerances. Each epoch concatenates displacement measurements over all minibatches, including inactive gates and the final partial batch: $M = 314{,}905$, with each candidate counted once. At the selected epoch, the specified numerical tolerances are $|\widehat{E}/E - 1| \le 0.01$ and $|\widehat{\alpha} - \alpha| \le 0.01$ for nonzero training budgets. Test-matrix measurements likewise average all 87,317 candidate rows per draw. These direct measurements are distinct from both the root-solver residual and the independent-stream approximation check.

Independent Gaussian and realized-budget audit. For radius r, an exact isotropic reduction uses independent $a \sim \mathcal{N}(0,1)$ and $b \sim \chi^2_{127}$:

$$E_r(\sigma; p) = 2p\,\mathbb{E}\left[1 - \frac{r + \sigma a}{\sqrt{(r+\sigma a)^2 + \sigma^2 b}}\right].$$

An independent Hermite–generalized-Laguerre quadrature evaluates this expression at every one of the 15 training and 20 test scale pairs in the full budget grid. Increasing each quadrature order from 64 to 96 changes expectations by less than 2×10^{-16}. At unit radius, total expected displacement differs from the recorded targets by -0.00472% to -0.00080%. The recorded $E_0 = 0.04897177594$ compares with an exact unit-radius reference 0.04896993313. Training radii range from 0.999853 to 1.000168 on the token path and 0.999864 to 1.000122 on the sentence path. Evaluating the extreme radii bounds expected total error below 0.038% across these scale pairs. This checks the approximation itself, independently of its Monte Carlo stream; it does not redefine the frozen budget unit or sigmas.

Fifteen-seed primary-comparison audit. The primary comparison extends the original five-seed runs at E_0 to fifteen seeds (20260814–20260828). Table 12 reports displacement-error bounds recomputed from the saved records. Each training epoch measures all 314,905 candidate rows; each test evaluation measures all 87,317 rows. The 2,700 test cells cover 15 seeds, six training arms including no augmentation, two test budgets, five test allocations, and three draws. For each budget, allocation, and draw, the displacement records agree across fitted models, giving 30 distinct noise realizations; repeated test cells are not independent budget-validation samples.

Additional-budget coverage. The five-seed full grid additionally covers training budgets $0.5E_0$ and $2E_0$. Across that grid, total relative budget error stays below 0.404% over all augmented training epochs and 0.116% in the full test matrix.

Exact-distance implementation. For a unit vector $\mathbf{u}$, a random unit tangent $\mathbf{v}$ satisfying $\mathbf{u}^\top \mathbf{v} = 0$, and target squared distance $e \in [0, 4]$, the training-distribution control uses

$$\mathbf{u}' = \left(1 - \frac{e}{2}\right) \mathbf{u} + \sqrt{e - \frac{e^2}{4}}\, \mathbf{v}, \tag{9}$$

which gives $\|\mathbf{u}'\|_2 = 1$ and $\|\mathbf{u}' - \mathbf{u}\|_2^2 = e$. Assigning $e_h = \alpha E_0$ and $e_s = (1 - \alpha) E_0$ fixes each candidate's total squared distance at E_0.

The implementation first normalizes the loaded vector, samples a Gaussian direction, removes its projection onto the unit input, and normalizes the tangent. It uses no Bernoulli gate; a zero-distance path returns the unit input. The square-root radicand is clamped below at zero, and the direction stream is seeded with the training seed plus 31337. Stored acceptance checks at the selected epoch allow a maximum unit-norm error of 10^{-5} and a maximum total squared-distance error of 10^{-6}. These floating-point tolerances qualify the exact identities of the mathematical construction. Experimental conditions are in Section 4.2.

Table 8: Five-seed calibrated test matrix: mean Trigger F1 (%) for models trained at $E = E_0$, as a function of test budget T and token share β. *None* is the no-augmentation training arm.

T/E_0	Train α	$\beta = 0$	0.25	0.5	0.75	1
0.5	None	52.87	52.78	52.59	52.56	52.52
0.5	0	53.81	53.63	53.35	53.02	52.79
0.5	0.25	53.83	53.71	53.48	53.22	53.05
0.5	0.5	53.64	53.57	53.38	53.25	53.17
0.5	0.75	53.19	53.17	53.15	53.15	53.03
0.5	1	52.26	52.28	52.39	52.48	52.69
1	None	51.56	51.50	51.26	51.00	50.82
1	0	53.29	52.74	52.15	51.56	50.87
1	0.25	53.24	52.84	52.35	51.90	51.28
1	0.5	52.85	52.56	52.27	51.85	51.44
1	0.75	52.20	52.04	51.89	51.79	51.54
1	1	51.00	50.96	51.02	50.97	51.11
2	None	49.19	48.95	48.46	47.95	47.44
2	0	52.37	51.19	49.82	48.55	47.22
2	0.25	51.98	51.04	50.14	49.00	47.97
2	0.5	51.27	50.61	49.95	49.12	48.30
2	0.75	50.18	49.87	49.55	49.00	48.43
2	1	48.46	48.48	48.55	48.31	48.07
8	None	35.61	36.20	35.22	33.11	29.91
8	0	46.21	42.44	38.20	33.81	29.38
8	0.25	44.17	41.83	38.52	34.61	30.36
8	0.5	41.55	40.75	38.29	35.15	30.97
8	0.75	38.29	38.69	37.58	35.02	31.17
8	1	34.34	35.85	35.93	34.37	31.01

Table 9: Trigger F1 (%), averaged equally over test allocations at $T = E_0$: main ungated test noise versus the 375-cell gate-matched check. Values are mean $\pm$ sample SD over five paired seeds.

Train α	$p_{\text{test}} = 1$	$p_{\text{test}} = 0.5$
0	52.12 ± 0.36	52.27 ± 0.52
0.25	52.32 ± 0.37	52.45 ± 0.48
0.5	52.19 ± 0.25	52.49 ± 0.34
0.75	51.89 ± 0.39	52.28 ± 0.41
1	51.01 ± 0.31	51.55 ± 0.30

Table 10: Five-seed SecureBERT and three-seed ModernBERT contrasts in F1 points. S–T is the primary mean sentence-minus-token contrast; I is the endpoint interaction. Document, seed, and crossed intervals vary the indicated sampling units.

Encoder	T/E_0	Contrast	Mean $\pm$ SD	Document CI	Seed CI	Crossed CI
SecureBERT	1	S–T	1.11 ± 0.17	$[0.74, 1.49]$	$[0.98, 1.24]$	$[0.69, 1.53]$
	1	I	2.53 ± 0.43	$[2.07, 3.01]$	$[2.23, 2.90]$	$[1.91, 3.15]$
	8	S–T	3.71 ± 0.78	$[3.38, 4.05]$	$[3.08, 4.28]$	$[3.02, 4.41]$
	8	I	13.50 ± 2.47	$[12.58, 14.43]$	$[11.63, 15.54]$	$[11.42, 15.66]$
ModernBERT	1	S–T	0.47 ± 0.28	$[0.11, 0.84]$	$[0.15, 0.64]$	$[-0.03, 0.96]$
	1	I	1.54 ± 0.33	$[1.07, 2.00]$	$[1.34, 1.92]$	$[0.94, 2.17]$
	8	S–T	2.92 ± 0.26	$[2.63, 3.21]$	$[2.63, 3.15]$	$[2.51, 3.31]$
	8	I	10.55 ± 2.59	$[9.62, 11.46]$	$[8.15, 13.30]$	$[8.14, 13.21]$

Table 11: ModernBERT calibrated test matrix: mean token Trigger F1 (%). Each entry averages three draws within each of three paired seeds, then seeds.

T/E_0	Train α	$\beta = 0$	0.25	0.5	0.75	1
1	0	46.65	46.60	46.22	45.83	45.44
1	0.5	46.51	46.52	46.34	46.03	45.76
1	1	45.46	45.59	45.64	45.90	45.79
8	0	39.81	37.41	34.42	30.71	27.40
8	0.5	36.52	35.77	34.30	31.44	28.04
8	1	30.56	31.94	32.53	31.45	28.70

Table 12: Fifteen-seed primary-comparison budget audit. Bounds on maximum absolute errors are rounded upward. Share error is an absolute fraction; test-cell counts include repeated measurements of shared noise.

Stage	Records	Total error (%)	Share error
Train, all epochs	600	< 0.552	< 0.002029
Train, selected epochs	75	< 0.449	< 0.000855
Test, $T = E_0$	1,350	< 0.109	< 0.000185
Test, $T = 8E_0$	1,350	< 0.063	< 0.000107

B Direct Controls and Path-Reliance Diagnostics

Frozen-setting controls keep the selected checkpoints, thresholds, calibrations, and test populations fixed during evaluation. The equal-distance control evaluates 15 fixed-checkpoint models; the CE check trains five allocations and no augmentation with three paired seeds, eight epochs, and unweighted `CrossEntropyLoss()`. PHEE uses eight paired seeds and its own training-derived E_0. Its Gaussian test scales solve the independent expectation by Hermite–Laguerre quadrature; evaluation records actual displacement. Optimization and attack protocols are specified below.

B.1 PHEE protocol

PHEE (Sun et al., 2022) supplies pharmacovigilance events; we use only trigger detection and typing in TextEE's processed split1. Table 13 summarizes the experimental protocol. Exact duplicate mentions are removed (16/6/2 train/development/test mentions); no unresolved span conflicts remain and no development or test windows are dropped by the training-label filter. Weight and scorer identifiers are documented in the reproduction repository's machine-readable protocol record.

Table 13: PHEE direct-matrix protocol. Counts refer to retained windows and tokens, not to CASIE documents.

Component	Specification
Data	TextEE PHEE split1: 2,897/965/965 train/dev/test windows; 63,942/21,128/21,377 tokens.
Encoder	Frozen SecureBERT; first subword of each original token; 128-dimensional projection, seed 20260814. Context is the normalized mean of projected word states. Maximum subword length 512; nonoverlapping overflow segments retain every original token.
Labels and decoding	Background O and B/I tags for Adverse_event and Potential_therapeutic_event. Add the selected bias to the O logit, take token argmax, and decode BIO into contiguous spans. An I tag with no active span of the same type starts a new span; B starts a new span and O closes it.
Training	Bilinear head, rank 64; eight epochs, batch 8192; AdamW, learning rate 10^{-3}, weight decay 10^{-4}; square-root inverse-frequency token cross-entropy. Seeds 20260814–20260821.
Selection	Clean-dev Trigger-C, then Trigger-I, then earliest epoch. O-logit bias uses clean dev only, preferring smaller absolute bias at tied scores. Checkpoints and biases remain fixed during all noisy tests.
Perturbation	Train-derived PHEE E_0, $\alpha \in \{0, .5, 1\}$, independent candidate-row Bernoulli gates $p = .5$, then L2 normalization; eight no-augmentation heads. Test $T/E_0 \in \{1, 8\}$, $\beta \in \{0, .25, .5, .75, 1\}$, $p_{\text{test}} = 1$. Gaussian scales solve the independent expectation; realized displacement is recorded.
Draws and scoring	Shared draws 2026091801–2026091803; 960 noisy evaluations. Official TextEE exact-span Trigger-I/Trigger-C, in percent; draws averaged within seed before paired contrasts.

B.2 Interaction estimates and condition profiles

Pairing and uncertainty. Each condition averages three prescribed noise draws within each seed. The following intervals use a paired Student-t interval across training seeds, conditional on the corpus and fixed draws. The seed is the replication unit; intervals have limited precision with three seeds. Document-level intervals use a different resampling unit (Appendix A.3). The SecureBERT CASIE rows below use the five-seed budget grid; the primary comparison uses fifteen seeds.

B.3 PHEE eight-seed profiles

We summarize 8 paired training seeds (20260814–20260821). For each seed and test condition, we first average three fixed noise draws (2026091801–2026091803). Budget-level summaries then average $\beta \in \{0, 0.25, 0.5, 0.75, 1\}$ with equal weights. Intervals are unadjusted 95% Student-t intervals across training seeds; contrasts are computed within seed before constructing the interval.

Table 14: Endpoint interaction: seed mean and paired-seed 95% t interval in percentage points. Each row uses its native metric.

Setting	Metric	$I(E_0)$	$I(8E_0)$
SecureBERT CASIE (5 seeds)	Trigger	2.53 [2.00, 3.07]	13.50 [10.43, 16.56]
SecureBERT CASIE (5 seeds)	Typed	2.80 [2.24, 3.36]	15.25 [11.77, 18.73]
ModernBERT CASIE	Trigger	1.54 [0.72, 2.36]	10.55 [4.10, 16.99]
ModernBERT CASIE	Typed	1.91 [0.85, 2.96]	12.88 [6.12, 19.63]
Equal-distance CASIE	Trigger	2.41 [1.87, 2.96]	12.48 [8.66, 16.30]
Equal-distance CASIE	Typed	2.67 [1.99, 3.35]	13.93 [9.90, 17.97]
Unweighted CE CASIE	Trigger	1.54 [1.50, 1.57]	8.49 [4.96, 12.02]
Unweighted CE CASIE	Typed	1.66 [1.41, 1.90]	9.86 [6.95, 12.77]
PHEE (8 seeds)	Trigger-I	0.12 [−0.02, 0.26]	1.07 [0.52, 1.61]
PHEE (8 seeds)	Trigger-C	0.21 [0.05, 0.37]	1.18 [0.61, 1.75]

Table 15: Equal-distance CASIE: noisy profiles, Trigger / Typed F1, in percent. Means over paired seeds and three draws.

T/E_0	α	$\beta = 0$	.25	.5	.75	1
1	0	53.20/52.60	52.69/52.02	51.99/51.29	51.32/50.62	50.69/50.01
1	0.5	52.74/52.06	52.47/51.76	52.09/51.37	51.75/51.06	51.28/50.63
1	1	51.26/50.43	51.32/50.51	51.28/50.48	51.23/50.43	51.17/50.51
8	0	45.26/43.77	42.29/40.53	38.06/36.16	33.35/31.33	28.69/26.60
8	0.5	41.23/39.39	40.44/38.55	38.11/36.29	34.45/32.69	29.77/28.01
8	1	34.79/32.43	36.40/34.08	36.13/34.20	34.38/32.74	30.69/29.19

Clean scores are evaluated once per fitted model. Training share $\alpha = 0$ denotes sentence-only perturbation and $\alpha = 1$ denotes token-only perturbation; β is the test token-channel share. For each metric, define the sentence-minus-token contrast $D_\beta(T) = F(0, \beta, T) - F(1, \beta, T)$, where F is that metric's F1 and T is the test budget.

The 32 fits cover no augmentation and $\alpha \in \{0, 0.5, 1\}$ on PHEE split 1, with 960 draw-level test evaluations. These are eight training seeds, not eight data splits. Trigger-I and Trigger-C use the official TextEE exact-span scorer. The frozen SecureBERT cache contains 2,897/965/965 train/dev/test windows.

The positive endpoint interaction at $8E_0$ does not imply an average sentence-only advantage: the five-allocation mean contrast is -0.122 [-0.246, 0.003] for Trigger-I and -0.102 [-0.193, -0.011] for Trigger-C (percentage points).

B.4 Concat endpoint check

Six concat heads use $\alpha \in \{0, 1\}$, training E_0, and seeds 20260814–20260816. The cache, training-derived scales, checkpoint, and clean-development threshold remain fixed. Evaluation crosses $T/E_0 \in \{1, 8\}$, $\beta \in \{0, 1\}$, and draws 2026091801–2026091803, without retraining. All realized mean budgets and allocations satisfy the specified 1% tolerances. Table 22 reports every seed's endpoints, averaging draws within seed. Table 23 uses paired-seed 95% Student-t intervals with three seeds.

The Trigger interaction is positive at both budgets, but seed 20260814 does not reverse its endpoint ranking. Thus Trigger reverses in two of three seeds, compared with three of three for Typed F1. At $8E_0$, the mean Trigger token-endpoint contrast is negative but its interval spans zero. Only endpoints are evaluated, so no five-share mean is reported.

B.5 End-to-end optimization protocol

We jointly fine-tune ModernBERT and the bilinear head on CASIE using AdamW (encoder/head learning rates $2 \times 10^{-5}/10^{-3}$, weight decay 10^{-4}), three epochs, and batches of 16 sentences. First-subword states are projected from 768 to 128 dimensions by a fixed orthonormal projection (seed

Table 16: Unweighted CE CASIE: noisy profiles, Trigger / Typed F1, in percent. Means over paired seeds and three draws.

T/E_0	α	$\beta = 0$	.25	.5	.75	1
1	No aug.	54.72/53.86	54.15/53.22	53.57/52.58	53.01/52.00	52.63/51.61
1	0	55.58/54.69	54.86/53.89	54.26/53.27	53.51/52.47	52.81/51.74
1	0.25	55.52/54.63	54.81/53.88	54.24/53.29	53.58/52.58	52.96/51.93
1	0.5	55.36/54.50	54.75/53.82	54.30/53.34	53.72/52.73	53.23/52.21
1	0.75	54.64/53.81	54.17/53.24	53.75/52.83	53.23/52.30	52.85/51.86
1	1	54.47/53.61	54.11/53.20	53.87/52.94	53.50/52.58	53.23/52.31
8	No aug.	45.34/43.68	42.51/40.52	39.04/36.81	34.97/32.57	30.88/28.24
8	0	50.29/49.05	45.48/43.65	40.40/38.17	35.66/33.16	31.23/28.33
8	0.25	48.88/47.55	44.94/43.17	40.60/38.42	36.13/33.80	31.89/29.24
8	0.5	48.22/46.76	45.02/43.22	41.16/39.12	36.84/34.55	32.31/29.84
8	0.75	46.21/44.64	43.87/42.05	40.70/38.72	36.64/34.48	32.19/29.92
8	1	43.33/41.53	42.19/40.27	40.04/38.06	36.88/34.84	32.76/30.67

Table 17: F1 (percent); mean [95% CI]. Corrupted scores average five test allocations.

Test budget	Training arm	Trigger-I	Trigger-C
Clean	No augmentation	55.251 [54.307, 56.196]	54.534 [53.621, 55.447]
Clean	$\alpha = 0$	55.608 [54.803, 56.413]	54.869 [54.039, 55.699]
Clean	$\alpha = 0.5$	55.387 [54.391, 56.383]	54.632 [53.658, 55.606]
Clean	$\alpha = 1$	55.785 [55.040, 56.531]	54.951 [54.133, 55.770]
$1E_0$	No augmentation	54.477 [53.565, 55.388]	53.609 [52.776, 54.442]
$1E_0$	$\alpha = 0$	54.679 [53.914, 55.443]	53.764 [53.018, 54.510]
$1E_0$	$\alpha = 0.5$	54.573 [53.699, 55.447]	53.689 [52.862, 54.516]
$1E_0$	$\alpha = 1$	54.790 [54.129, 55.452]	53.850 [53.172, 54.528]
$8E_0$	No augmentation	47.013 [46.540, 47.485]	45.880 [45.474, 46.286]
$8E_0$	$\alpha = 0$	47.241 [46.786, 47.695]	46.085 [45.654, 46.516]
$8E_0$	$\alpha = 0.5$	47.232 [46.737, 47.727]	46.115 [45.664, 46.566]
$8E_0$	$\alpha = 1$	47.362 [46.990, 47.735]	46.187 [45.802, 46.572]

20260814), with column signs aligned to the initial cache. Each projected token is normalized before mean pooling; the pooled sentence vector is normalized again. Both paths remain differentiable. The maximum input length is 512 subwords. Removing training documents whose normalized text occurs in development/test leaves 12,669 training sentences (314,905 tokens); the test split contains 3,667 sentences (87,317 tokens). Training uses square-root inverse-frequency weighted cross-entropy; the implementation computes weights from the unfiltered training labels. Checkpoint selection maximizes clean-development Typed F1, then Trigger F1, then prefers the earliest epoch. The detection threshold is also selected on clean development data and remains fixed at test time.

Training noise uses independent Bernoulli(0.5) gates for each token row and channel, followed by renormalization. Sentence vectors are expanded to token rows before corruption, so their noise is independent across rows. Test noise is ungated. Scales are calibrated on the initial cache and kept fixed during fine-tuning; $E_0 = 0.04897177$ is the mean summed squared post-normalization displacement, with the training gate already included. Trigger/Typed F1 are token-level detection/typed micro-F1, respectively.

Eight-seed comparison. Table 24 specifies the training arms, seeds, and test grids. Table 25 averages draws, then the five shares within each seed, then seeds. For metric F, define $D_s(T, \beta) = F_{\text{sentence},s}(T, \beta) - F_{\text{token},s}(T, \beta)$ and $I_s(T) = D_s(T, 0) - D_s(T, 1)$. All interaction intervals below are paired-seed Student-t 95% intervals, $\bar{I} \pm t_{0.975,n-1} s_I / \sqrt{n}$, without multiplicity correction; draws are not independent replicates. Units are percentage points.

Table 18: Paired gains over no augmentation (percentage points); mean [95% CI]. Corrupted scores average five test allocations.

Test budget	Training arm	Trigger-I	Trigger-C
$1E_0$	$\alpha = 0$	0.202 [-0.143, 0.546]	0.155 [-0.052, 0.361]
$1E_0$	$\alpha = 1$	0.314 [-0.098, 0.725]	0.241 [-0.019, 0.502]
$8E_0$	$\alpha = 0$	0.228 [0.079, 0.378]	0.205 [0.101, 0.309]
$8E_0$	$\alpha = 1$	0.350 [0.186, 0.513]	0.307 [0.191, 0.424]

Table 19: Paired sentence-minus-token contrasts averaged over five test allocations and endpoint interaction $I(T) = D_0(T) - D_1(T)$, in percentage points [95% CI]. Reversals count seeds with $D_0(T) > 0$ and $D_1(T) < 0$.

Budget	Metric	Mean $D_\beta(T)$	Interaction I	Reversals
$1E_0$	Trigger-I	-0.112 [-0.432, 0.209]	0.121 [-0.019, 0.262]	1/8
$1E_0$	Trigger-C	-0.087 [-0.327, 0.154]	0.209 [0.046, 0.372]	3/8
$8E_0$	Trigger-I	-0.122 [-0.246, 0.003]	1.069 [0.523, 1.614]	7/8
$8E_0$	Trigger-C	-0.102 [-0.193, -0.011]	1.180 [0.615, 1.746]	7/8

Table 20: Full test-allocation profiles at $1E_0$. Columns give β. Values are F1 percent, mean $\pm$ sample SD across 8 seeds after averaging the three draws.

Arm	Metric	0	0.25	0.5	0.75	1
No augmentation	Trigger-I	55.043 ± 1.156	54.992 ± 1.042	54.591 ± 1.071	54.140 ± 1.127	53.617 ± 1.097
No augmentation	Trigger-C	54.320 ± 1.098	54.154 ± 0.948	53.686 ± 0.951	53.201 ± 1.012	52.683 ± 1.028
$\alpha = 0$	Trigger-I	55.317 ± 0.962	55.138 ± 0.933	54.842 ± 0.917	54.321 ± 0.921	53.775 ± 0.892
$\alpha = 0$	Trigger-C	54.517 ± 1.010	54.274 ± 0.930	53.909 ± 0.885	53.336 ± 0.862	52.783 ± 0.846
$\alpha = 0.5$	Trigger-I	55.227 ± 1.115	55.034 ± 1.079	54.667 ± 1.058	54.245 ± 0.997	53.693 ± 1.038
$\alpha = 0.5$	Trigger-C	54.459 ± 1.070	54.196 ± 1.040	53.746 ± 0.995	53.299 ± 0.936	52.745 ± 0.979
$\alpha = 1$	Trigger-I	55.453 ± 0.859	55.241 ± 0.783	54.814 ± 0.766	54.411 ± 0.826	54.033 ± 0.795
$\alpha = 1$	Trigger-C	54.578 ± 0.944	54.344 ± 0.853	53.851 ± 0.760	53.425 ± 0.815	53.053 ± 0.763

Table 21: Full test-allocation profiles at $8E_0$. Columns give β. Values are F1 percent, mean $\pm$ sample SD across 8 seeds after averaging the three draws.

Arm	Metric	0	0.25	0.5	0.75	1
No augmentation	Trigger-I	53.735 ± 0.649	50.436 ± 0.673	46.998 ± 0.606	43.635 ± 0.598	40.260 ± 0.769
No augmentation	Trigger-C	52.783 ± 0.643	49.237 ± 0.570	45.805 ± 0.491	42.489 ± 0.550	39.086 ± 0.663
$\alpha = 0$	Trigger-I	54.265 ± 0.808	50.698 ± 0.630	47.225 ± 0.565	43.758 ± 0.615	40.257 ± 0.707
$\alpha = 0$	Trigger-C	53.273 ± 0.895	49.516 ± 0.607	46.006 ± 0.525	42.558 ± 0.543	39.073 ± 0.592
$\alpha = 0.5$	Trigger-I	53.994 ± 0.909	50.732 ± 0.615	47.182 ± 0.648	43.820 ± 0.628	40.431 ± 0.596
$\alpha = 0.5$	Trigger-C	53.029 ± 0.915	49.557 ± 0.573	46.018 ± 0.582	42.660 ± 0.561	39.310 ± 0.510
$\alpha = 1$	Trigger-I	53.804 ± 0.679	50.691 ± 0.518	47.360 ± 0.590	44.092 ± 0.461	40.865 ± 0.549
$\alpha = 1$	Trigger-C	52.739 ± 0.787	49.447 ± 0.507	46.137 ± 0.543	42.895 ± 0.466	39.719 ± 0.486

Table 22: Concat endpoint scores by seed: token Trigger / Typed F1 (%). Each cell averages three shared draws.

T/E_0	α	Seed	$\beta = 0$	$\beta = 1$
1	0	20260814	53.41/52.42	51.55/50.55
1	0	20260815	53.21/52.58	50.87/50.14
1	0	20260816	52.75/52.13	50.95/50.32
1	1	20260814	51.17/50.14	51.37/50.63
1	1	20260815	50.83/50.03	51.38/50.60
1	1	20260816	51.27/50.48	50.95/50.36
8	0	20260814	42.83/40.82	30.90/28.56
8	0	20260815	44.95/43.18	30.80/28.54
8	0	20260816	43.11/41.34	31.37/29.16
8	1	20260814	30.89/28.25	30.37/28.96
8	1	20260815	32.30/29.63	33.10/31.52
8	1	20260816	30.95/28.49	31.67/30.21

Table 23: Concat sentence-minus-token contrasts and their interaction in F1 points: mean [paired-seed 95% interval].

Metric	T/E_0	D_0	D_1	$I(T)$
Trigger	1	$2.03\ [0.83, 3.24]$	$-0.11\ [-1.01, 0.78]$	$2.15\ [0.38, 3.91]$
Trigger	8	$12.25\ [11.34, 13.15]$	$-0.69\ [-4.31, 2.93]$	$12.94\ [8.41, 17.46]$
Typed	1	$2.16\ [1.01, 3.31]$	$-0.19\ [-0.78, 0.40]$	$2.35\ [0.70, 4.01]$
Typed	8	$12.99\ [11.74, 14.24]$	$-1.48\ [-4.79, 1.84]$	$14.47\ [9.90, 19.04]$

Table 24: Two end-to-end protocols with shared optimization settings above. The three-seed check uses seed IDs contained in the eight-seed comparison; these seed sets are not independent.

	Eight-seed comparison	Three-seed end-to-end comparison
Training arms	Sentence-only, equal-joint, token-only	No augmentation, sentence-only, token-only
Seeds	20260814–20260821	20260814–20260816
Test T/E_0	1, 8	1, 2
Test β	0, 0.25, 0.5, 0.75, 1	0, 1
Draw seeds	2026091801–2026091803	2026091801–2026091805
CI replicate	8 paired seeds (t_7)	3 paired seeds (t_2)

Table 25: Eight-seed end-to-end means (Trigger/Typed F1). Noisy columns average the five test allocations.

Training	Clean	E_0	$8E_0$
Sentence-only	72.44/72.17	72.37/72.10	70.68/70.40
Equal-joint	73.20/72.93	73.14/72.86	71.72/71.44
Token-only	73.07/72.86	73.03/72.82	71.46/71.25

Trigger interactions are 0.08 $[-0.13, 0.29]$ at E_0 and 0.04 $[-1.34, 1.42]$ at $8E_0$; Typed interactions are 0.07 $[-0.15, 0.29]$ and 0.06 $[-1.27, 1.39]$. These intervals do not establish an allocation-by-corruption interaction under this fine-tuning protocol.

Three-seed end-to-end comparison. Clean Trigger/Typed means are 73.18/72.91, 71.97/71.65, and 72.61/72.42, for no augmentation, sentence-only, and token-only, respectively. Table 26 reports endpoint means. Trigger interactions are 0.04 $[-0.18, 0.27]$ and 0.02 $[-0.58, 0.62]$ at E_0 and $2E_0$; Typed interactions are 0.04 $[-0.13, 0.22]$ and 0.04 $[-0.49, 0.56]$ ($n = 3$).

Table 26: Separate three-seed end-to-end means (Trigger/Typed F1).

T/E_0	β	No augmentation	Sentence-only	Token-only
1	0	73.26/72.98	71.97/71.65	72.66/72.44
1	1	72.97/72.70	71.64/71.34	72.37/72.18
2	0	73.30/72.99	71.99/71.66	72.66/72.43
2	1	72.57/72.32	71.30/71.00	72.00/71.81

Displacement and rerun audit. Across the 24 fits in the eight-seed comparison, the largest training-budget error is 0.52% over all epochs (0.17% at selected epochs). Of the 24 state-capture reruns, only 21 agree with the archived runs within 10^{-4}. On 253 eligible sentences among the first 256 test sentences, mean within-sentence token cosine falls from 0.616 to 0.429 and token-block participation ratio rises from 2.30 to 4.21 across the reruns: representation geometry changes during fine-tuning. On the first 1,024 test sentences (22,605 token rows), the largest measured fine-tuned budget error is 0.11% across 24 reruns and 20 audit conditions; stable displacement on this subset does not imply stable representation geometry. These diagnostics describe the reruns, not all archived states. In the three-seed end-to-end comparison, clean scores match the archived values exactly, and the maximum subset test budget error is 0.05%; the maximum training error over all epochs is 0.40%.

B.6 Projected-gradient attack protocol

We evaluate 30 frozen SecureBERT bilinear checkpoints: no augmentation and five training allocations, each with five paired seeds (20260814–20260818). Four test conditions ($T \in \{E_0, 8E_0\}$, $\beta \in \{0, 1\}$) and two starts give 240 attack evaluations. Checkpoints and clean-development thresholds remain fixed. The white-box attack maximizes unweighted cross-entropy on the true test labels for ten steps, with normalized per-vector gradients and step size $2\sqrt{b_c}/10$ for channel c. Starts are zero and projected Gaussian noise (draw seed 2026092301); each update uses 24 bisection steps to rescale the additive perturbation before normalization.

The implemented constraint is $\| \mathrm{Norm}(x_i + \delta_i) - \mathrm{Norm}(x_i)\|_2^2 \leq b_c$, where b_c is that channel's measured mean squared displacement from one calibrated Gaussian draw. Token means are over tokens; context means are over unique sentences. One context perturbation is shared by all tokens of its sentence in both PGD and its paired Gaussian comparator, unlike the candidate-row corruption in the main Gaussian and end-to-end protocols. Both evaluate all 87,317 test-token candidates from 3,667 sentences; there is no candidate filtering or test subsampling. Gaussian noise has no corresponding per-vector cap. These are different constraint sets, even when their measured means agree.

Across all 240 attacks at $(T/E_0, \beta) \in \{(1,0),(1,1),(8,0),(8,1)\}$, attack-to-Gaussian mean displacement ratios are 0.999983–1.000000. Thus the mean displacements agree to within 0.002% based on measurement, rather than nominal budgets alone. The available measurements are channel means, not per-vector maxima; although the implementation checks the per-vector constraint, these aggregate measurements cannot independently certify it.

Table 27: Frozen-checkpoint Gaussian → PGD endpoint means over five seeds (Trigger/Typed F1). PGD takes the lower final F1 of the two starts separately for each metric; it is a finite attack, not a certified worst case.

T/E_0	β	Training	Gaussian	PGD
1	0	No augmentation	51.90/51.09	8.11/5.99
1	0	Sentence-only	53.64/53.06	14.53/12.20
1	0	Token-only	51.53/50.77	7.82/5.70
1	1	No augmentation	50.28/49.40	0.83/0.00
1	1	Sentence-only	50.56/49.75	0.82/0.00
1	1	Token-only	50.69/49.95	0.95/0.00
8	0	No augmentation	36.31/34.08	6.75/2.97
8	0	Sentence-only	46.92/45.66	6.73/3.37
8	0	Token-only	35.08/32.63	6.76/2.97
8	1	No augmentation	29.11/27.34	1.45/0.00
8	1	Sentence-only	28.91/26.97	2.15/0.00
8	1	Token-only	30.17/28.93	1.33/0.00

Using the same paired-seed interaction definition, PGD Trigger interactions are 6.83 $[5.61, 8.06]$ at E_0 and -0.85 $[-1.19, -0.51]$ at $8E_0$; Typed interactions are 6.50 $[5.28, 7.72]$ and 0.40 $[0.34, 0.46]$. The corresponding single-draw Gaussian interactions are 2.23 $[1.46, 3.00]$ and 13.10 $[10.77, 15.43]$ for Trigger, and 2.49 $[1.77, 3.22]$ and 15.00 $[12.33, 17.67]$ for Typed. Large F1 losses and metric-dependent endpoint interactions limit any extension of the Gaussian findings to adversarial directions.

B.7 Mixed-allocation comparison and path diagnostics

Exploratory mixed-allocation comparison. In the five-seed SecureBERT grid, the $\alpha = .75$ mean exceeds $\alpha = 1$ in all 20 conditions across four test budgets and five test allocations, by 0.166–3.945 Trigger F1 points. These conditions were noticed after inspecting the matrix. We use the five paired seed differences per condition and a Bonferroni correction over 20 two-sided Student-t intervals (family level 95%). Eleven lower bounds exceed zero; nine include zero. Thus the positive means do not establish simultaneous dominance. The Gaussian gate-frequency confound remains relevant for this mixed-versus-pure comparison; the equal-distance models do not include $\alpha = .75$. In the unweighted-CE comparison, only five of ten condition means favor $\alpha = .75$ over token-only training, and both token-only test endpoints favor token-only training. Thus the all-condition mean advantage in the weighted-loss grid does not transfer to this stronger loss; mixed-allocation benefits are configuration-dependent.

Path-sensitivity and permutation diagnostics. For the 25 SecureBERT heads in the five-seed E_0 grid and six ModernBERT endpoint heads, we keep clean-development thresholds fixed and shuffle one path's test candidate rows with a shared permutation (seed 2026092207), leaving the other path unchanged. We also differentiate the gold-event-minus-background logit margin on all 3,054 positive test tokens. For input $\mathbf{u}$ and gradient $\mathbf{g}$, the local score is $\|\mathbf{g} - (\mathbf{g}^\top \bar{\mathbf{u}})\bar{\mathbf{u}}\|_2^2$, where $\bar{\mathbf{u}}$ is the unit input direction. We average tokens within model before averaging seeds.

Table 28: Local gradient sensitivity and path-permutation diagnostics. Squared tangent-gradient norms depend on the model's logit scale; permutation drops are clean minus shuffled Trigger F1 points. Shuffling produces mismatched features and can induce floor effects.

Encoder	α	Token grad.	Sentence grad.	Token drop	Sentence drop
SecureBERT	0	1721.1	513.8	51.40	12.62
SecureBERT	0.25	1543.5	619.9	51.43	13.90
SecureBERT	0.5	1396.8	767.7	51.47	14.97
SecureBERT	0.75	1267.8	986.8	51.42	17.59
SecureBERT	1	1154.8	1291.7	50.92	20.29
ModernBERT	0	1320.1	528.3	44.63	12.11
ModernBERT	1	930.2	1223.4	44.38	18.03

The gradients shift in opposite directions as allocation changes, while sentence-path permutation is less damaging after sentence-path training. This pattern is consistent with a shift in path reliance away from the trained-noisy path. Token shuffling nearly destroys performance in all arms, so its similar F1 drops cannot establish equal token reliance.

Table 29: Document and token statistics of the converted CASIE–MAVEN data. Class rows and the positive subtotal show count (percentage of all tokens in that column). Train (filtered) is the training split used in the main experiments; development and test are unchanged.

Statistic / label	Train (original)	Train (filtered)	Development	Test
Documents	798	698	99	201
Total tokens	361,117	314,905	45,998	87,317
○	348,277 (96.44%)	303,670 (96.43%)	44,400 (96.53%)	84,263 (96.50%)
Databreach	2,582 (0.72%)	2,288 (0.73%)	287 (0.62%)	689 (0.79%)
DiscoverVulnerability	2,614 (0.72%)	2,341 (0.74%)	273 (0.59%)	769 (0.88%)
PatchVulnerability	1,871 (0.52%)	1,615 (0.51%)	256 (0.56%)	509 (0.58%)
Phishing	2,627 (0.73%)	2,300 (0.73%)	327 (0.71%)	543 (0.62%)
Ransom	3,146 (0.87%)	2,691 (0.85%)	455 (0.99%)	544 (0.62%)
All positive tokens	12,840 (3.56%)	11,235 (3.57%)	1,598 (3.47%)	3,054 (3.50%)

C Reference Detector Implementation

The reference interaction classifier follows low-rank Hadamard pooling (Kim et al., 2017). For $d = 128$ and $r = 64$, it computes

$$\mathbf{u}_i = \tanh(W_h \mathbf{h}_i + b_h), \qquad \mathbf{v} = \tanh(W_s \mathbf{s} + b_s), \quad \mathbf{b}_i = \mathbf{u}_i \odot \mathbf{v}, \tag{10}$$

$$\mathbf{z}_i = \mathrm{LayerNorm}\big(\mathrm{GELU}(W_z[\mathbf{h}_i; \mathbf{s}; \mathbf{b}_i] + b_z)\big), \quad \boldsymbol{\ell}_i = W_o \mathbf{z}_i + b_o. \tag{11}$$

Training substitutes the augmented vectors in all paths, including the concatenated skip inputs. The same-width concat ablation removes the product features and their projections while retaining token and sentence inputs. It has fewer parameters, so this comparison does not isolate algebraic form at fixed capacity. The augmentation interface itself does not require this head.

Loss and clean-development selection. For training counts n_c, $N = \sum_{c=1}^{C} n_c$, and $C = 6$, define

$$a_c = \sqrt{\frac{N}{C n_c}}, \qquad w_c = \frac{a_c}{\sum_{k=1}^{C} (n_k/N) a_k}, \qquad \mathcal{L}_B = -\frac{\sum_{i \in B} w_{y_i} \log P_\theta(y_i \mid \mathbf{h}'_i, \mathbf{s}'_i)}{\sum_{i \in B} w_{y_i}}.$$

This is the weighted-mean cross-entropy used for each minibatch B; the count-weighted mean class weight is one. The predicted positive class is the highest-scoring event type, accepted when its logit minus the background logit is at least a threshold. At each epoch, clean-development margins supply 120 quantiles from 0.80 to 0.995 (duplicates removed). Threshold selection maximizes Typed F1, then Trigger F1, then prefers the smallest absolute threshold. Checkpoint selection maximizes the first two metrics and prefers the earlier epoch in a tie. Each training condition performs this selection independently.

D Dataset Statistics

Table 29 gives the per-class token counts behind the description in the experimental setup. Counts refer to labeled tokens, not event mentions, and each percentage uses all tokens in its column as the denominator.

Document filtering. For each document, we concatenate its original sentence strings in their recorded order with a single space, lowercase the result, collapse runs of whitespace to one space, and strip leading and trailing whitespace. A training document is excluded if this entire normalized string equals that of any development or test document; all of its token rows are removed from the training cache, while development and test remain unchanged. The filtering implementation is provided in the reproduction repository.

E Regularizer Context

Fixed-scale protocol and metrics. The fixed-scale comparison uses eleven paired seeds (20260814–20260824), frozen features, the reference head, eight training epochs, and fixed clean-development

Table 30: Paired 95% intervals for three fixed-scale regularizer contrasts, in points. Seed-level: 10,000 bootstrap resamples of the 11 paired seed deltas. Document-level: 10,000 resamples of the 201 test documents, paired across arms within each seed.

Contrast	Metric	Delta	Seed-level CI	Document-level CI
Ref. aug. − no aug.	Clean	$+0.38$	$[+0.17, +0.57]$	$[+0.08, +0.68]$
	Stress	$+2.85$	$[+2.52, +3.16]$	$[+2.59, +3.12]$
	Worst	$+3.22$	$[+2.83, +3.60]$	$[+2.92, +3.54]$
VAT 0.05 − ref. aug.	Clean	-1.23	$[-1.60, -0.88]$	$[-2.09, -0.37]$
	Stress	$+3.74$	$[+3.19, +4.26]$	$[+3.10, +4.38]$
	Worst	$+4.35$	$[+3.81, +4.86]$	$[+3.46, +4.95]$
Mixup 0.4 − ref. aug.	Clean	-0.76	$[-1.21, -0.28]$	$[-1.30, -0.21]$
	Stress	$+1.81$	$[+1.47, +2.15]$	$[+1.47, +2.16]$
	Worst	$+1.52$	$[+1.09, +1.98]$	$[+1.10, +1.95]$

Table 31: The 2×2 factorial design: mask on/off $\times$ post-noise L2 normalization on/off. 11 seeds each, Trigger F1 in %; mean $\pm$ sample SD. Deltas elsewhere in this section are computed from unrounded means, so they may differ from subtracting the rounded values printed here.

Cell (mask, norm)	Clean	Stress	Worst
(yes, yes) — reference augmentation	54.53±0.54	39.42±0.75	33.90±1.02
(yes, no)	54.36±0.64	39.47±0.64	34.01±0.80
(no, yes)	54.61±0.64	39.32±0.64	33.98±0.82
(no, no) — unmasked, unnormalized	54.47±0.55	39.54±0.60	34.07±0.83

selection. At $\sigma_{\text{test}} = 0.05$, Stress is the mean Trigger F1 over token-, sentence-, and jointly perturbed inputs, and Worst is their minimum. These definitions also apply to the fixed-scale masking, normalization, and path comparisons and the equal-displacement control below. Runs affected by random-stream coupling are excluded from the mixup comparison.

Table 30 summarizes three contrasts from the regularizer study. Simplified VAT (scale 0.05) and feature mixup (0.4) exceed reference augmentation on Stress by 3.74 and 1.81 points, respectively, while lowering Clean F1 by 1.23 and 0.76 points. Seed- and document-level intervals agree on these directions but quantify different uncertainty: training variability and evaluation-sample variability conditional on the fitted models. Document resampling uses the 201 test documents with 3,054 gold trigger tokens.

VAT here is an interface-level virtual-adversarial control, not the full SMART or SeqVAT system. These fixed-scale comparisons do not match realized post-normalization displacement across regularizers and do not establish an equal-budget ranking. They provide context for augmentation gains, separate from the allocation–channel interaction.

F Masking and Normalization Controls

A 2×2 factorial comparison varies masking and post-noise normalization with eleven paired seeds per cell (Table 31); Stress and Worst follow Appendix E. The mask is one Bernoulli gate per row, shared across its dimensions. On Stress, the mask main effect is $+0.02$ points (paired-seed 95% CI $[-0.19, +0.24]$), the normalization main effect is -0.13 ($[-0.29, +0.03]$), and their interaction is $+0.16$ ($[-0.13, +0.43]$). These unresolved estimates do not establish equivalence. Two smaller effects exclude zero: normalization on Clean ($+0.15$, $[+0.05, +0.25]$), and normalization without masking on Stress (-0.21, $[-0.29, -0.11]$). Intervals use 10,000 paired-seed bootstrap resamples. This comparison characterizes the fixed-scale recipe; it does not replace the calibrated allocation matrix.

Displacement reference points. Let x be a loaded cache vector and x' the output of the complete perturbation operator. We measure $\|x' - x\|^2$, summed over both paths. For the normalized arm, we also measure $\|x' - \mathrm{Norm}(x)\|^2$, using the same normalized output x'. The two columns in Table 32

Table 32: Displacement measured during training versus predicted from cache geometry alone (11 seeds; averages over 312 batches per seed). The two measured columns use the loaded cache x and unit base $\bar{u} = \mathrm{Norm}(x)$, respectively; both follow the arm's operator. "Budget" is the measured-to-nominal ratio, with nominal budget 0.0512. Displacement was not recorded for the reference augmentation, which is therefore excluded from this comparison.

Arm	From x	From $\bar{u}$	Predicted	Deviation	Budget
Masked, no normalization	0.051196	—	0.051094	$+0.20\%$	1.000
Unmasked, normalization	0.049867	0.049867	0.049887	-0.04%	0.974
Unmasked, no normalization	0.051201	—	0.051226	-0.05%	1.000
Reference augmentation	—	—	—	—	—

Table 33: Single-path attribution on CASIE token conversion, frozen SecureBERT cache. Eleven seeds, mean ± sample SD in percent, paired by seed. 'Joint' is the reference recipe (both paths perturbed). The two single-path arms only perturb their own path; both paths are still normalized, matching the reference implementation.

Configuration	Clean F1	Stress F1	Worst F1
No augmentation	54.15±0.32	36.57±0.84	30.67±0.99
Token path only	54.08±0.58	36.81±0.91	31.34±1.14
Sentence path only	54.33±0.35	39.02±0.76	33.00±1.03
Joint (both paths)	54.53±0.54	39.42±0.75	33.90±1.02

therefore differ in the reference point, not in whether measurement precedes or follows normalization. Their agreement reflects the cache's near-unit radii; it does not imply $\| \mathrm{Norm}(u + \delta) - u\|^2 = \|\delta\|^2$. Predictions from cache geometry agree with measured displacement within 0.2%. The normalized arm's displacement is 2.6% below the nominal pre-normalization budget, consistent with normalization changing effective displacement.

G Classifier Capacity Control

In a separate TextEE CASIE exact-span comparison across five splits and three seeds, a one-hidden-layer concat head with matched parameter and linear-compute budgets exceeds the bilinear head by 0.57 Trigger-C F1 points without augmentation and 0.53 with it (positive in 12/15 and 13/15 pairs). This comparison provides no positive increment for the multiplicative interaction. A two-layer concat head improves further, but also changes the width allocation, so it does not isolate an effect of depth. These architecture comparisons do not identify the allocation interaction; the direct concat endpoint control is reported in Appendix B.4.

H Fixed-Scale Path Comparison

Stress and Worst use the fixed-scale test conditions defined in Appendix E. At fixed per-path noise scales, the eleven-seed paired comparison in Table 33 gives Stress gains over no augmentation of $+2.45$ points for sentence-only training, $+0.24$ for token-only training, and $+2.85$ for joint training. Both inputs remain available in every arm; only the perturbed paths change. Single-path arms inject less total displacement, so these gains do not identify fixed-budget path contributions or demonstrate additivity.

At these unmatched scales, joint minus sentence-only is $+0.40$ Stress points. Matching total displacement changes that difference to -0.10 (Appendix I); the five-seed budget grid gives $+0.07$ at $T = E_0$ and -0.67 at $T = 8E_0$ (Appendix A). Thus the apparent benefit of adding token perturbation depends on total budget and evaluation condition.

Table 34: Equal-displacement control over three paired seeds (Trigger F1, %). The difference is joint minus sentence-only.

Training perturbation	Clean	Stress	Worst
Joint token + sentence	54.80 ± 0.71	39.68 ± 0.85	34.20 ± 0.80
Sentence only, matched	54.60 ± 0.70	39.78 ± 0.57	33.87 ± 0.40
Paired difference	$+0.19 \pm 0.02$	-0.10 ± 0.34	$+0.33 \pm 0.42$

I Equal-Displacement Path Control

The single-path rows in Table 33 do not match the total amount of training perturbation. We therefore ran a separate paired control that matches the expected squared displacement after row normalization in the concatenated token–sentence input. The joint arm uses $\sigma = 0.02$ on both paths. Using training vectors only, deterministic quadrature gives $\sigma = 0.028824$ for the sentence-only arm, matching the joint target 0.04896876 with relative error below 10^{-7}. The nominal $\sqrt{2} \times 0.02$ value would undershoot the post-normalization target by 3.46%.

Both arms use the same frozen cache, bilinear reference head, weighted objective, eight epochs, development-only selection and threshold calibration, and the fixed-scale test conditions in Appendix E. Within each of three seeds they share model initialization, epoch order, masks and Gaussian draws; the inactive token-path draws in the sentence-only arm are consumed and discarded. Table 34 reports means and sample standard deviations.

The aggregate Stress means are nearly equal, but their composition differs. Joint training is higher under token noise by $+2.04 \pm 0.34$ points (3/3 paired differences positive); it is lower under sentence noise by -2.67 ± 0.47 (0/3 positive), and higher under joint test noise by $+0.33 \pm 0.42$ (2/3 positive). Thus, sentence-only training recovers the aggregate Stress result, while training-path allocation changes condition-specific behavior.

The comparison uses six fits (two arms across three paired seeds), with identical paired random streams and update counts. Every realized displacement is within 0.32% of its target.

J Reproducibility

Access and coverage. Code and reproduction materials are available at `https://anonymous.4open.science/r/perturbation-allocation-repro-BBA5/`. They support cache-level model reproduction for the five-seed budget grid and packaged controls, and statistical recomputation for the fifteen-seed primary comparison, eight-seed PHEE comparison, end-to-end checks, and PGD and drift diagnostics. The repository README and reproduction guide map results to commands and document the available inputs and checkpoints.

Inputs and limitations. The model-reproduction archive provides frozen SecureBERT and ModernBERT cache arrays, converted documents and row maps, checkpoints for the covered experiments, and ModernBERT weights and tokenizer assets. Unfiltered training arrays are included because they define the calibration population (Appendix A). SecureBERT extraction weights are unavailable, so its model reproduction starts from the supplied caches; cache-level reproducibility does not establish encoder reconstruction. The statistical analyses use per-fit and per-condition records and scripts under `evidence/`. Checkpoints, encoder weights, and cache arrays beyond the model archive's coverage are not included. In particular, the PHEE checkpoints cover three of the eight seeds, while statistical records cover all eight. The reproduction guide specifies the supported scope.

Computational environment. The recorded runtime is Python 3.12.2, NumPy 1.26.4, SciPy 1.17.1, scikit-learn 1.6.1, and PyTorch 2.8.0 with CUDA 12.8. Figure regeneration requires LaTeX/PGFPlots. Environment specifications, execution commands, file manifests, and the scope of the recorded reproduction checks are documented in the repository.

Protocol separation. The matched-energy control (Appendix I) uses training-only calibration and per-fit checkpoint selection; its random-number protocol differs from that of the fixed-scale

comparison. Results from incompatible metrics, model definitions, or random-number protocols are not pooled.