FAILSAFE-MOE: TRAINING SPARSE MIXTURE-OF-EXPERTS FOR WORST-CASE EXPERT OUTAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse Mixture-of-Experts (MoE) models route each token to a few of E experts, so an unavailable expert corrupts every token routed through it. We argue that worst-case expert-outage risk is an *exposure* phenomenon: it appears when router load is skewed, uniform expert dropout under-weights it, and weighting training outages by exposure — no search — reduces most of it at uniform-dropout cost. A sampled-max adversary drawing m of E candidates optimises a spectral risk measure with weight m/E on the worst outage, against $1/E$ for uniform dropout. In single-MoE-layer models, worst-case damage tracks router load skew: at an OLMoE-like 3.23× the exact adversary gains +1.48 pt of worst-case accuracy over conventional training, while uniform dropout is -2.35 pt *worse* than no outage training at all. An exposure-proportional sampler at uniform-dropout cost (1.13–1.21× a clean step) matches the exact adversary within ± 1.0 pt (TOST, two benchmarks), but reduces the tail rather than removing it: 1.74 pt of worst-outage drop remains at 3.23× and 2.01 pt at 11.58×. In a 6-layer 99 M-parameter MoE language model on WikiText-103 the same sampler halves worst-outage loss damage (0.077 against 0.157 nats, 3 of 3 seeds, n=3) at equal steps and equal clean loss; given that compute as extra steps, a conventionally trained model is 0.27 nats better yet more fragile (0.444 nats of damage). The pre-registered first-order criticality diagnostic is refuted; router mass is the best cheap predictor overall.

1 INTRODUCTION

Sparse Mixture-of-Experts layers add parameters without proportionally adding computation (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2021): a router sends each token to k of E experts. Because computation is *conditional*, an expert is not a redundant fraction of a distributed representation but the sole processor of the tokens routed to it. If it becomes unavailable — a failed or preempted device, a capacity drop, a hot-swap, a pruning decision — every token that would have used it is affected at once; in wide expert-parallel serving a single lost rank disrupts the whole inference instance (Sun et al., 2026), and the damage is uneven, a few experts being disproportionately load-bearing (Su et al., 2025).

This paper makes one claim, and organises everything else as evidence for it or scope on it: **worst-case expert-outage risk in sparse MoE is an exposure phenomenon. It appears when router load is skewed; uniform expert dropout under-weights it, giving the worst outage weight $1/E$ in its training objective; and weighting training outages by exposure — with no search — reduces most of it at uniform-dropout cost. Where load is balanced or k is large, there is little to fix.** *Most*, not all: on the benchmark where we can enumerate every outage, the exposure-weighted sampler still leaves a worst-single-outage drop of 1.74 pt at load skew 3.23× (against 2.17 pt under conventional training) and 2.01 pt at 11.58× (against 14.04 pt). We call the family of tail-weighted outage adversaries *FailSafe-MoE*; the member we recommend is FAILSAFE-RM, which samples the dropped expert in proportion to its batch router mass, with the exact max over all E outages as the reference upper bound.

Each clause is a measurement. The standard response to component failure, in MoE (Chen et al., 2023; Liu et al., 2022) as in three decades of fault-tolerance work (Neti et al., 1992; Torres-Huitzil

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.

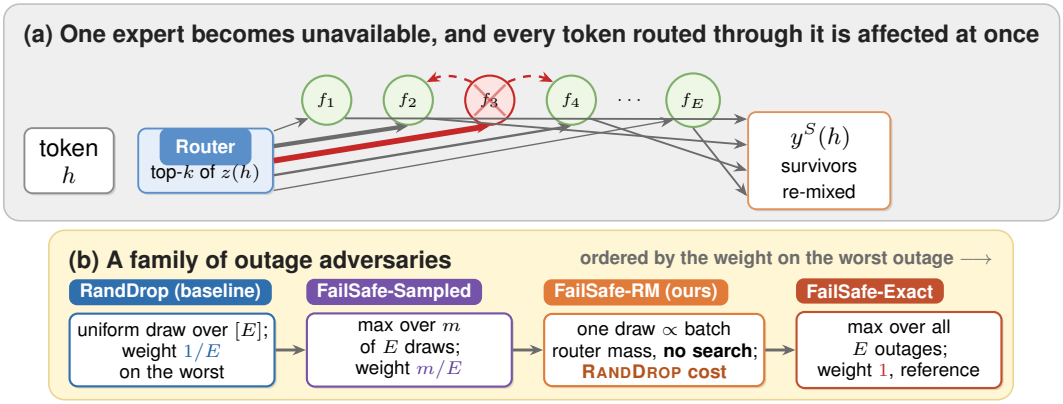


Figure 1: **Worst-case expert-outage risk is an exposure phenomenon: it appears when router load is skewed, uniform dropout under-weights it by $1/E$, and weighting training outages by exposure — with no search — reduces most of it at uniform-dropout cost.** (a) The struck-out red expert is the outage; reroute-on-failure means nothing is zeroed and gate mass moves to the next-best survivor; arrow width is router mass, dashed red where the traffic goes. (b) The members differ in *one line* of the training step. Evidence runs from exhaustively verified toy models (n=10 seeds) through a 99 M MoE LM (n=3, all 96 outages) to an inference-only audit of all 1024 outages of `allenai/OLMoE-1B-7B-0924`; we never train at that last scale.

& Girau, 2017), is to inject *random* failures during training; that improves the tail only incidentally, closing 44 % of the gap to an exact adversary at E=32 against 92 % at E=8.

Contributions. **(i)** We pose global expert-outage robustness as a minimax problem and give a *family* of outage adversaries — exact, sampled-max, and exposure-proportional (FAILSAFE-RM) — differing in one line of the training step, with FAILSAFE-RM as the recommended member and the exact max as the reference (§3, §5). **(ii)** We transfer a known order-statistic identity from data indices to outage sets: the sampled-max objective is a spectral risk measure — a mixture of CVaR levels, not a single one — with weight m/E on the worst outage, so uniform dropout's blind spot is a *weight* of $1/E$ in its objective rather than a coverage failure (Observation 1). The identity is Kawaguchi & Lu (2020)'s Theorem 1 at q=1; the application and its consequences are ours (§4). **(iii)** We give matched-seed evidence on *when* tail-weighted training helps, in two load regimes side by side, with controls that rule out compute and step count as explanations, the structural alternatives priced next to our method, and a sampler ablation that separates the two things exposure weighting does (§6, §6.4); and we repeat the comparison in a 6-layer 99 M-parameter MoE language model trained from scratch on WikiText-103, where an iso-compute control shows the failure mode *growing* with steps (§6.5). **(iv)** We give a first-order criticality decomposition, a *pre-registered, negative* test of the diagnostic it suggests at two scales, and an audit of all 1024 single-expert outages of a pretrained MoE LLM (§7, §8).

2 RELATED WORK

Sparse MoE. Conditional computation via a learned router (Shazeer et al., 2017) scaled to transformers through capacity-limited expert parallelism (Lepikhin et al., 2020; Fedus et al., 2021) with stabilizing losses (Zoph et al., 2022) and is standard in open models (Jiang et al., 2024; Dai et al., 2024; Muennighoff et al., 2024). Routing variants change which expert sees which token (Zhou et al., 2022; Xie et al., 2022) and therefore the outage geometry; we fix token-choice top-k routing throughout. Checkpointing keeps MoE *training* reliable (Gandhi & Kozyrakis, 2024) and serving can reconfigure expert-parallel membership after a rank failure (Sun et al., 2026); both attack the problem from outside the model.

What is not new here. Training a network against the worst-case *removal of its own components* is old: Neti et al. (1992) posed a minimax over unit removals, Phatak & Koren (1995) analysed the

redundancy full fault tolerance needs, and Torres-Huitzil & Girau (2017) survey a field in which fault-injection training is the dominant recipe, so our uniform-dropout baseline has a thirty-year pedigree; Park et al. (2017) search for a worst-case dropout *mask* under a budget, in the lineage of adversarial training (Goodfellow et al., 2014; Madry et al., 2017). Neither is the order-statistic identity of Observation 1 ours: Kawaguchi & Lu (2020) prove it for the top-q losses of a random minibatch, and at $q{=}1$ their weights are ours with data indices in place of outage sets; Fan et al. (2017) average the top-k losses over the whole index set (a flat spectrum, no subsampling) and Shalev-Shwartz & Wexler (2016) reach the maximum by importance sampling. Sampled outage risk is a tail statistic (Rockafellar & Uryasev, 2000) inside the known spectral family (Acerbi, 2002; Curi et al., 2019), whose minibatch estimates are biased (Mehta et al., 2022) — a warning that applies to ours — and a worst case over a discrete index set is also group DRO (Sagawa et al., 2019), where the index set partitions the *data*. Specific to conditional computation is that the removed unit's work is *rerouted* by a learned gate, that the outage set is combinatorial, and that exposure is itself learned.

Expert absence, removal and redundancy. Our RANDDROP baseline is the common core of random expert-subset training (Chen et al., 2023) and gating dropout (Liu et al., 2022), re-implemented in our harness; expert pruning (Chen et al., 2022; Lu et al., 2024) chooses which expert to remove, while an outage chooses for you. Appendix A extends this paragraph.

3 EXPERT OUTAGES AS A MINIMAX PROBLEM

Setup. In one MoE layer with experts $f_1, \dots, f_E$, a router produces logits $z(h) \in \mathbb{R}^E$ for token representation h (ℓ is the per-example loss); top-k routing selects $S_k(h) = \text{top-}k(z(h))$ and mixes with gates $g_e(h) = \mathrm{softmax}(z(h)|_{S_k(h)})_e$, giving $y(h) = \sum_{e \in S_k(h)} g_e(h)\, f_e(h)$.

Definition 1 (Global expert outage)**.** *An* outage *is a set $S \subseteq [E]$ of experts unavailable for the entire evaluation. Under S, routing re-selects among survivors, $S_k^S(h) = \text{top-}k(z(h)|_{[E]\setminus S})$, and gates renormalize over that set, giving $y^S(h)$. We write $\mathcal{L}(\theta; S)$ for the task loss under outage S, with $\mathcal{L}(\theta; \emptyset) = \mathcal{L}_{\text{clean}}(\theta)$.*

This is *reroute-on-failure*: nothing is zeroed, traffic moves to the next-best available expert. It is the protocol most favourable to a model that has never seen a failure — we verify this rather than assert it (§6.3). We train against an outage drawn or selected by an adversary $\mathcal{A}$:

$$\min_{\theta} \;\; \mathcal{L}_{\text{clean}}(\theta) \;+\; \beta\, \mathbb{E}_{C \sim \mathcal{A}} \Big[\max_{e \in C} \mathcal{L}(\theta; \{e\}) \Big], \tag{1}$$

for budget $q{=}1$ (all our training runs; $q{>}1$ replaces e by an S with $|S| \le q$). Members of the family differ only in $\mathcal{A}$: RANDDROP draws one expert uniformly ($|C|{=}1$), FAILSAFE-SAMPLED a uniform m-subset, FAILSAFE-EXACT takes $C{=}[E]$, and FAILSAFE-RM draws one expert with probability proportional to its batch router mass. The last is a *post hoc* member, registered after the metric lock (§6.2, Appendix C).

4 WHAT THE ADVERSARY OPTIMISES

Let $\mathrm{Dmg}(e) = \mathcal{L}(\theta; \{e\}) - \mathcal{L}_{\text{clean}}(\theta)$ and order experts so that $\mathrm{Dmg}(e_{(1)}) \ge \cdots \ge \mathrm{Dmg}(e_{(E)})$; write $L_{(j)} = \mathcal{L}(\theta; \{e_{(j)}\})$. A *sampled-max* adversary draws m experts uniformly without replacement and returns the most damaging of them.

Lemma 1 (Order statistics of the selected rank; standard)**.** *Let R_m be the damage rank of the outage a sampled-max adversary with m draws returns, damages distinct. Then $\Pr[R_m = j] = \binom{E-j}{m-1} \big/ \binom{E}{m}$, $\Pr[R_m = 1] = m/E$, and $\mathbb{E}[R_m] = (E+1)/(m+1)$.*

Observation 1 (Sampled-max is a spectral risk measure; Kawaguchi & Lu, 2020, Theorem 1 at $q{=}1$, with data indices replaced by outage sets)**.** *The inner objective of Eq. 1 under a uniform sampled-max adversary with budget m satisfies*

$$\mathbb{E}_{C_m} \Big[\max_{e \in C_m} \mathcal{L}(\theta; \{e\}) \Big] = \sum_{j=1}^{E} w_j^{(m)}\, L_{(j)}, \qquad w_j^{(m)} = \frac{\binom{E-j}{m-1}}{\binom{E}{m}}, \tag{2}$$

where the weights are non-negative, sum to one, and are non-increasing in the rank index j. Hence the objective is a coherent spectral risk measure of the outage-loss distribution (Acerbi, 2002), with m as the risk-aversion parameter: $m{=}1$ gives the uniform spectrum $w_j \equiv 1/E$ (the outage average*), $m{=}E$ gives $w_1{=}1$ (the maximum), and $w_1^{(m)} = m/E$ throughout.*

What is and is not ours. The identity is Kawaguchi & Lu (2020)'s Theorem 1 at $q{=}1$, with (n, s) in place of (E, m); ours is the substitution of outage sets for data indices and the consequences below. Eq. 2 *is* a CVaR mixture, as every discrete risk spectrum is — but not a single CVaR level, and not the flat average-top-k spectrum of Fan et al. (2017).

Corollary 1 (The blind spot is a weight in the objective, not coverage)**.** *Under uniform sampling with budget m, the worst outage enters the training objective with weight m/E, independently of the number of training steps; for $m{=}1$ that is $1/E$.*

This is a statement about weight, not about coverage: the probability $(1{-}m/E)^T$ that T uniform draws never touch the worst expert is numerically vacuous at our T (Appendix B). Uniform dropout does see the worst outage; it discounts it. Two qualifications: the rank-1 outage is an index, so which expert occupies it can change from step to step; and a per-minibatch maximum is an upward-biased estimate of the population spectral risk (Mehta et al., 2022).

From ranks to accuracy. Eq. 2 weights ranks; predicting *accuracy* needs the measured damage profile $L_{(j)}$ (Figure 10), which is steeply concentrated at rank 1 under conventional training, *together with* the m/E weight — the prediction is never Observation 1 alone. Replacing the uniform draw by any p gives, at $m{=}1$, $\sum_e p_e\, \mathcal{L}(\theta; \{e\})$ — the same family with a different base measure; FAILSAFE-RM takes $p_e \propto \mathrm{RM}(e)$, tilting mass towards the tail exactly to the extent that damage correlates with exposure.

4.1 WHICH EXPERT IS CRITICAL

Router mass $\mathrm{RM}(e) = \mathbb{E}_h[g_e(h)]$ is the obvious proxy for criticality. It captures how much traffic an expert carries, not how badly its absence would be felt. Let $s_h = \partial\ell/\partial y$ be the sensitivity of the downstream loss to the layer output, let $\bar{f}_{\sigma_e}(h)$ denote the convex *substitute mixture* that absorbs e's gate mass, let a_h be the gate weight the substitute actually acquires, and let u_h be the convex combination of the other selected experts.

Proposition 1 (First-order criticality bound)**.** *Assume **(A1)** the per-example loss is L-smooth in the layer output and **(A2)** tokens with $e \notin S_k(h)$ are unaffected by the outage of e. Then*

$$\big|\mathrm{Dmg}(e)\big| \;\le\; \underbrace{\mathbb{E}_h\big[\|s_h\|\, g_e(h)\, \big\|\bar{f}_{\sigma_e}(h) - f_e(h)\big\|\big]}_{\mathrm{EDS}(e)} + \mathrm{Res}(e) \;+\; \tfrac{L}{2}\,\mathbb{E}_h\|\delta_h\|^2, \tag{3}$$

where the residual depends on the gate convention: $\mathrm{Res}(e) = \mathbb{E}_h[\|s_h\|\, |g_e(h) - a_h|\, \|u_h - \bar{f}_{\sigma_e}(h)\|]$ when surviving gates are renormalized, and $\mathrm{Res}(e) = \mathbb{E}_h[\|s_h\|\, |g_e(h) - a_h|\, \|\bar{f}_{\sigma_e}(h)\|]$ under a truncated softmax without renormalization (`norm_topk_prob=false`, the OLMoE convention).

The first term is the diagnostic: exposure × disagreement × sensitivity. The residual is zero only when the substitute inherits exactly e's gate weight and grows with specialization (Appendix B); we state the bound for the single-substitute case ($q{=}1$), the only case we train and measure. **It is valid and nearly uninformative quantitatively**: in loss units $\mathrm{EDS}(e)/|\mathrm{Dmg}(e)|$ has median 30.8 on SSB (IQR [18.6, 61.5], 319 pairs) and 435 on OLMoE, most of it Cauchy–Schwarz slack rather than second-order remainder (the signed first-order term has median ratio 0.33). It also bounds a *magnitude*, so it cannot express an outage that helps, and $\mathrm{RM}(e)$ is its exposure factor alone, so the two rankings coincide whenever exposure dominates the dispersion of the product — which §8 answers against us.

5 FAILSAFE-MOE

One training step (Algorithm 1, Appendix E): compute the clean loss with its balance term; form a candidate set C — all E experts (exact), m uniform draws (sampled), or a single expert drawn $\propto$ batch

Table 1: **Tail-weighted training beats uniform dropout in both load regimes, and exposure weighting — which costs what uniform dropout costs — tracks exhaustive search.** E=32 on SSB (accuracy %, 10 matched seeds; Δ columns are paired worst-1 differences in pt with 95% bootstrap CIs in). worst-1 is an exhaustive minimum over all E single outages; **every condition trains with q=1 only**. Block (a) is anchored on CLEAN, because uniform dropout is itself -2.35 pt *below* CLEAN there; block (b) is the pre-registered confirmatory regime, whose router is collapsed (63.4 % no-op candidates). Within each block and each accuracy column, **bold** is the best value and underline the second. The FAILSAFE-RM rows and all of block (a) are exploratory (post-lock); the pre-registered m=4 arm is in Tables 3 and 16. In block (b) FAILSAFE-RM and FAILSAFE-EXACT print the same worst-1 mean and the same Δ: these are two *different* cell means that agree to the printed precision (62.0969 % and 62.0984 %), with different confidence intervals, and the coincidence carries no meaning.

	accuracy (%) ↑			Δ worst-1 (pt) ↑ vs.		
Training	clean	mean-1	worst-1	CLEAN	RANDDROP	n
(a) OLMoE-like load, max/mean 3.23× (*co-primary*, post-lock; 23.9 effective experts)						
CLEAN	62.77	62.38	60.61	—	+2.35 [+1.31, +3.38]	10
RANDDROP	<u>63.41</u>	62.96	58.25	-2.35 [-3.38, -1.31]	—	10
FAILSAFE-RM	**63.57**	**63.26**	<u>61.83</u>	+1.22 [+0.94, +1.50]	+3.57 [+2.55, +4.61]	10
FAILSAFE-EXACT	63.41	<u>63.12</u>	**62.09**	+1.48 [+1.06, +1.89]	+3.83 [+2.63, +5.02]	10
(b) Collapsed router, max/mean 11.58× (*pre-registered confirmatory*; 7.6 effective experts)						
CLEAN	63.55	62.87	49.51	—	-5.52 [-7.63, -3.45]	10
RANDDROP	63.65	63.13	55.04	+5.52 [+3.45, +7.63]	—	10
FAILSAFE-RM	<u>64.11</u>	<u>63.86</u>	<u>62.10</u>	+12.58 [+10.18, +14.85]	+7.06 [+6.05, +7.95]	10
FAILSAFE-EXACT	**64.19**	**63.96**	**62.10**	+12.59 [+10.33, +14.75]	+7.06 [+6.27, +7.76]	10

router mass (RM); evaluate $\mathcal{L}(\theta;\{e\})$ for $e \in C$ by *re-mixing* cached expert outputs; and descend $\mathcal{L}_0 + \beta \max_{e\in C} \mathcal{L}_e$. Setting m=1 with a uniform draw recovers random-outage training exactly, so baseline and method differ in one line and one scalar; FAILSAFE-RM costs what RANDDROP costs, 1.13–1.21× a clean training step at E=8–32 in our harness (Table 26). Gradients flow through the $\max$ via the selected outage (Madry et al., 2017); ties break by first index. The re-mix is cheap, but the forward pass as we implement it evaluates *all* E experts densely, at 16× the expert FLOPs of a sparse top-k step at E=32, k=2 (our 4.43× wall-clock ratio for the exact adversary is against a CLEAN arm that is also dense). Under reroute-on-failure a token can only ever need its top-$(k{+}1)$ set, so 1.50× sparse would suffice independently of E — derived only, not implemented. The per-outage re-mix and output head do not shrink, so we claim affordability only *at our scale*.

6 EXPERIMENTS

Benchmarks. **SSB**, a specialization stress benchmark, generates G Zipf-frequent latent skills grouped into families with similar decision rules, so the router must identify the skill, an expert must implement its rule, and exposure is skewed while disagreement varies independently of it. **FMNIST-Rot** applies E fixed rotations to Fashion-MNIST[1] as unlabelled domains (realized skew 4.49/8.22/15.04× at E=8/16/32). Full specifications, and a diagram of the construction and of the load regimes, in Appendix E (Figure 6).

Conditions and protocol. All conditions share architecture, optimizer, schedule, batch order and initialization at a given seed, differing only in the adversary, and all use q=1: **no condition ever trains on two-expert outages.** We report 10 matched seeds on every pre-registered contrast (5 on ablations and FMNIST-Rot); contrasts are paired over seeds, tested with a paired t-test with percentile bootstrap CIs and d_z effect sizes, with Wilcoxon tests alongside for pre-registered contrasts (Table 2). Worst-case accuracy is a true minimum over all E single outages, not an estimate. β=1.0 was fixed in the pre-registered plan; the learning rate, width and step count were fixed in the locked training configuration before the first locked batch and never tuned, but were *not* written into the pre-registration, and we claim nothing stronger. **Everything in §6.1–§6.4 other than block (b) of Table 1 is exploratory and post-lock**, much of it uncorrected, and none of it able to change a pre-registered verdict (Appendix C).

[1] `https://github.com/zalandoresearch/fashion-mnist`, MIT licence.

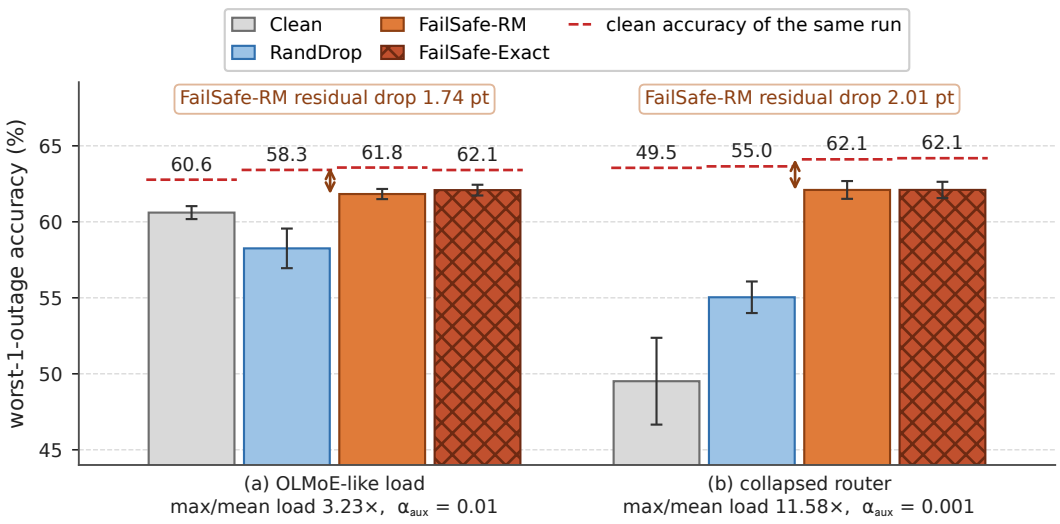


Figure 2: **Exposure weighting recovers most of the worst-case gap at uniform-dropout cost, and leaves a measurable tail behind.** Worst-1 accuracy, E=32 on SSB, the two regimes of Table 1; means over n=10 matched seeds, 95% t CIs, dashed red cap = each run's own clean accuracy. (a) and every FAILSAFE-RM cell are exploratory/post-lock.

6.1 TWO LOAD REGIMES, REPORTED SIDE BY SIDE

The pre-registered confirmatory cells were run at α_{aux}=0.001, realized load 11.58×. That router is *collapsed*: only 7.6 of 32 experts are effective, a 256-token batch leaves 11.7 live experts, and 63.4 % of the 32 candidates the exact adversary scores each step are exact no-ops — so restricting the inner max to live experts (EXACT-LIVE) is equivalent within ±1.0 pt (+0.08 pt, 90% CI [-0.18, +0.34], TOST $p = 5.7 \times 10^{-5}$) at 0.58× the step time (Table 14). We keep that regime as the confirmatory result, label it a collapsed-router regime, and make α_{aux}=0.01 — load 3.23×, 23.9 effective experts, 10.5 % no-ops — co-primary throughout (Table 1, Figure 2), because it is where the pretrained model we audit sits (3.22×).

There the honest anchor is CLEAN: uniform dropout is itself -2.35 pt below CLEAN on worst-1 accuracy (Holm $p = 0.003$), because random dropout drives the router to re-concentrate load (6.29× against 3.23×) while the exact adversary does not (3.56×). Against CLEAN, FAILSAFE-EXACT gains +1.48 pt and FAILSAFE-RM +1.22 pt (Holm $p = 2.3 \times 10^{-4}$, $p = 6.1 \times 10^{-5}$), both at *higher* clean accuracy. The full α_{aux} sweep scopes the claim (Table 10, Figure 3): as measured load falls 11.58× →1.14×, a conventionally trained model's worst-1 drop falls 14.04→0.62 pt and the exact adversary's advantage over uniform dropout falls +7.06→+0.37 pt, the two most balanced cells being *equivalent to zero* within ±1.0 pt (90% CIs [-0.11, +0.76], [+0.09, +0.65]) — which is how we read the nominally significant 1.14× cell.

The regime claim is correlational. α_{aux} changes the training objective, so the sweep is an association. A fixed router logit bias at constant α_{aux} (E24, Table 13) **failed its manipulation check with a free router**; with a *frozen* router it moves load, and a conventionally trained model's worst-1 drop then rises +3.16 pt per e-fold of skew ($p = 0.031$), but the EXACT−RANDDROP advantage tracks skew only at $p = 0.077$. Worst-case damage *tracks* load skew; the instrument is consistent with, but does not establish, a causal reading.

6.2 WHERE THE TRAINING OUTAGE DISTRIBUTION PUTS ITS MASS

Locked results. At the locked skew, FAILSAFE-EXACT improves worst-1 accuracy over RANDDROP by +7.06 pt at E=32 ($p = 2.8 \times 10^{-8}$, $d_z = 5.57$), and that advantage grows monotonically with the pool, by +6.09 pt from E=8 to E=32 ($p = 1.5 \times 10^{-5}$); Figure 4(c) shows both and Figure 4(a) places them beside every other headline contrast, Figure 8 gives the per-seed values, and

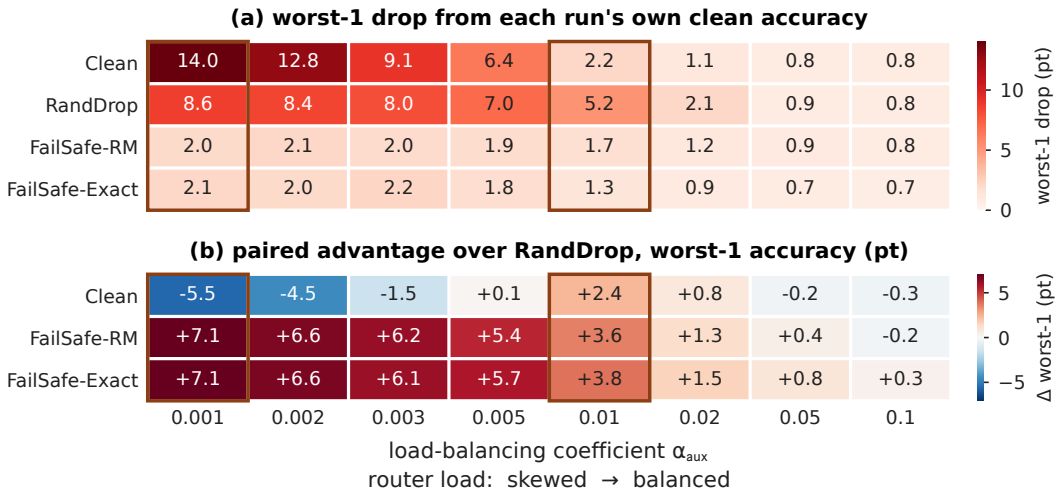


Figure 3: **The effect tracks the load regime, not one cell.** E=32 SSB, 8 values of α_{aux}, n=10 seeds per cell (exploratory). (a) Worst-1 drop from each run's own clean accuracy (pt). (b) Paired advantage over RANDDROP (pt). Outlined: the two regimes of Table 1; CIs in Table 16.

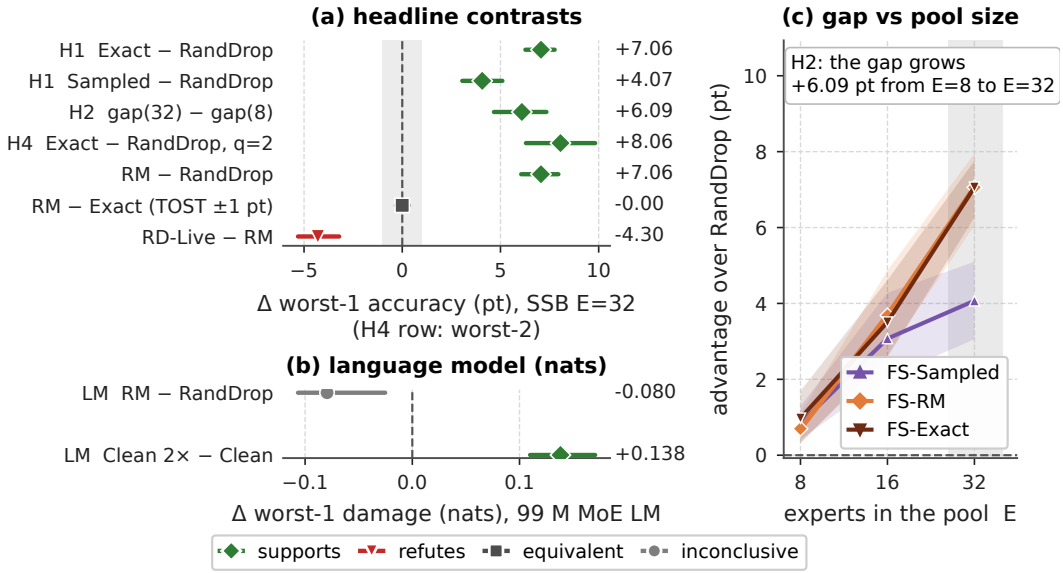


Figure 4: **Every headline contrast in one picture, and an advantage that grows with the expert pool.** (a) Pre-registered (H1, H2, H4) and key exploratory contrasts in accuracy points, n=10 matched seeds, 95% CIs (90% for the TOST row against the ±1.0 pt grey band); all worst-1 except H4. (b) The LM contrasts, in nats, n=3. (c) Advantage over RANDDROP against E, 95% bootstrap bands. Colour *and* marker shape encode the verdict.

Table 2 adds the pre-registered m=4 arm and the Wilcoxon tests. Clean accuracy is not the price (paired difference +0.54 pt, CI excluding a *cost* above 0.14 pt), and robustness reaches the $\binom{32}{2}$=496 two-expert outages no condition trains on (+8.06 pt, $p = 1.5 \times 10^{-5}$) — though additivity from single-outage hardening alone *over*-predicts that gain (118 %), so q=2 is a consistency check, not transfer. On FMNIST-Rot the ordering survives at E=16 and E=32 but is **not significant at** E=8 ($p = 0.062$; Table 4).

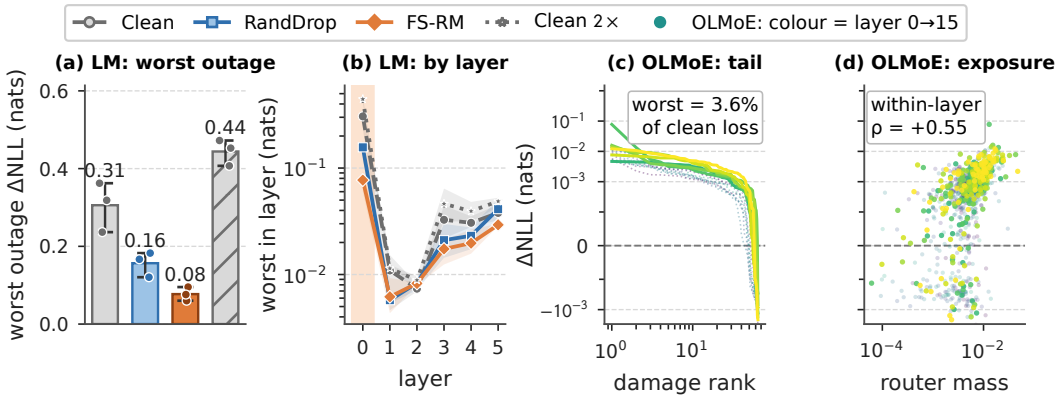


Figure 5: **Beyond the toy: exposure weighting halves the tail of a trained 99M MoE LM, and a pretrained MoE's tail is real but small.** (a) Worst outage of the 6-layer LM (dots = seeds, whiskers = min/max, n=3); (b) per layer, min/max band. (c,d) All 1024 outages of `allenai/OLMoE-1B-7B-0924`; faint = layers below the noise floor. Exploratory.

Exposure weighting matched exhaustive search on both benchmarks. FAILSAFE-RM, which drops an expert with probability proportional to its batch router mass at RANDDROP cost, is equivalent to the exact adversary within ± 1.0 pt at every pool size on both benchmarks: the paired differences run from -0.28 to +0.17 pt on SSB and from -0.35 to -0.07 pt on FMNIST-Rot, where its advantage over RANDDROP grows with the pool (Tables 18 and 15). The ± 1.0 pt margin is the pre-registered *clean-accuracy* margin re-used post hoc, and equivalence is not equality: at E=8 on FMNIST-Rot FAILSAFE-RM is strictly *worse* than the exact adversary (-0.23 pt, 95% CI [-0.37, -0.08]). So search was unnecessary in the settings we tested, not in general; and FAILSAFE-RM is *post hoc*, registered after the metric lock.

Appendix H (Figure 9) asks whether this is an implicit load balancer.

6.3 CONTROLS AND STRUCTURAL ALTERNATIVES

Table 9 (Appendix I) collects the six controls — iso-FLOP at m=4 and at m=E, 4× steps, held-out selection, task fixed at G=16, and the structural alternatives. Holding each fixed in turn leaves the adaptive adversary ahead, though across 4 benchmark instances the E=32 gap varies more ($+6.77 \pm 0.62$ pt) than the seed-level CI suggests. Two alternatives are real competition: stronger load balancing reaches the exact adversary's worst-1 accuracy, and a shared always-on expert (Dai et al., 2024) buys 11.80 pt for free, with FAILSAFE-EXACT leading RANDDROP+shared by only 1.54 pt ($p = 0.024$, n=5); we never trained both together. The exact adversary's advantage is *concentrated at* k=2 (-0.10/+7.06/+1.39/-0.13 pt at k=1/2/4/8; Table 22), and zeroing instead of rerouting is worse for every condition (Table 23).

6.4 WHAT EXPOSURE WEIGHTING DOES: EXCLUDING DEAD OUTAGES, AND TILTING

The weight on the worst outage (Corollary 1) does not by itself predict our main result, so the RANDDROP-LIVE ablation (E30, Table 12, Appendix G) separates never drawing a no-op expert from tilting towards busy ones: excluding dead outages accounts for about 39 % of the gain where the router is collapsed and none where it is not; the exposure-proportional tilt tracks the rest.

6.5 DOES THE PICTURE SURVIVE A MULTI-LAYER MOE LANGUAGE MODEL?

Everything above is one MoE layer at ≤1M parameters. We therefore trained from scratch a 6-layer decoder-only transformer in which *every* FFN is a token-choice top-2 MoE with E=16 experts (98.6 M parameters, 32.4 M active/token) on 45.1 M tokens of WikiText-103 at 3 matched seeds, and measured all 96 single-expert outages of every trained model on the whole validation split (Figure 5a,b; Table 5). **Exploratory and post-lock.**

Three findings. **(1)** At equal steps (about $2\times$ the clean step cost here) FAILSAFE-RM roughly halves the worst-case damage of uniform dropout — 0.157 to 0.077 nats, in 3 of 3 seeds — with clean NLL unchanged (+0.0002 nats, 95% CI [-0.0124, +0.0182]); paired $p = 0.099$, so the evidence is the per-seed values at n=3, not the p-value. Unlike on SSB, uniform dropout *beats* CLEAN here and both outage arms *raise* router skew (2.66/2.59$\times$ against 2.00$\times$); we cannot explain either difference. **(2)** The worst outage lies in layer 0 in 12 of 12 models and FAILSAFE-RM's gain is almost all a layer-0 gain (0.306 $\rightarrow$ 0.077 nats; Figure 5b). **(3) The failure mode grows with longer training**: CLEAN at $2\times$ steps (and $2\times$ tokens) reaches perplexity 45.38 against 59.84 yet its worst-case damage is *larger*, 0.444 against 0.306 nats ($p = 0.016$), so the fragility is not a symptom of under-training.

What this does not license. "No clean cost" fails at matched *compute*: FAILSAFE-RM's clean NLL is +0.2677 nats above the $2\times$-steps model's, and it leads only on absolute worst-case NLL (4.160 against 4.259). n=3, 45.1 M tokens, one corpus and depth, skew 2.0–2.7$\times$: one rung up from the toy, not LLM-scale evidence.

7 AN OUTAGE AUDIT OF A PRETRAINED MOE LLM

We audited `allenai/OLMoE-1B-7B-0924` (Muennighoff et al., 2024) (16 layers, 64 experts per layer, top-8 routing), measuring *all* $16 \times 64 = 1024$ single-expert outages on 24,552 held-out next-token positions of WikiText-103 (clean NLL 2.1622 nats); inference only, exploratory (Figure 5c,d; full audit in Figure 7). We drop the failed expert *after* the softmax, since this model sets `norm_topk_prob=false` and pre-softmax masking would manufacture a damage–router-mass correlation; 6 of 16 layers resolve above the noise floor.

Damage is heavy-tailed within every layer (worst outage 3.6–30.6$\times$ the mean one) but small in absolute terms (3.60 % of the clean loss; 16.2 % of outages slightly *improve* NLL). The busiest expert is the most damaging in only 2 of 16 layers, but the most damaging one lies in the top 8 by router mass in 16 of 16 (chance 0.12): a router-mass *shortlist* is well founded where its argmax is not. The load of 3.22$\times$ matches our co-primary regime, but top-8 routing is where the k-ablation finds -0.13 pt, and the small headroom agrees: **the value of our method at OLMoE-like k is unestablished**.

8 THE CRITICALITY DIAGNOSTIC, REPORTED AS A FAILURE

We pre-registered that $\mathrm{EDS}(e)$ would reach Spearman $\rho \geq 0.75$ against measured per-expert damage and beat router mass by ≥ 0.25. It does neither ($\rho(\mathrm{EDS}) = 0.645$ against $\rho(\mathrm{RM}) = 0.650$ over 120 locked models; -0.005, $p = 0.279$): **H3 is refuted** in all four settings of Table 11. Under balanced load router mass wins 12 of 12 cells; on OLMoE it reaches $+0.551$ against EDS's $+0.529$ on average (EDS leads in 4 of 16 layers), and the signed first-order estimate only $+0.092$ (Figure 7c). Appendix C carries the factor decomposition and our criticism of the 0.25 margin.

9 LIMITATIONS

Expanded in Appendix D: no training at OLMoE scale; the regime claim is an association; FAILSAFE-RM is post hoc; the confirmatory regime has a collapsed router and the co-primary one is post-lock; the top-$(k{+}1)$ implementation is not built; the audit is one model, inference only; many exploratory contrasts are uncorrected; no multi-expert dropout baseline. Expert-choice routing (Zhou et al., 2022) and Byzantine experts are out of scope.

10 CONCLUSION

Worst-case expert-outage risk in sparse MoE is an exposure phenomenon: it appears under skewed router load, uniform dropout under-weights it by $1/E$, and weighting training outages by exposure reduces most of it — leaving 1.74 pt of worst-outage drop at 3.23$\times$ and 2.01 pt at 11.58$\times$ — at uniform-dropout cost and without search.

Reproducibility statement

Every number in this paper is produced by scripts in the anonymised supplementary material from `results.jsonl` files emitted by a single runner; no value is entered by hand except pre-registered thresholds and design constants. All controlled experiments run in one pinned CPU Docker image (image tag and digest in the supplementary), single-threaded, for 71 CPU-hours over 2,696 training runs. The language-model study (§6.5) and the real-MoE audit are the only GPU work: one 24 GB card in Docker for 279 GPU-minutes in total (226 training and evaluating the 12 language models, 52.4 auditing OLMoE). Benchmarks are generated deterministically from a fixed task seed (SSB) or a fixed Fashion-MNIST subset recorded in `items.json` (MIT licence); the real-MoE audit uses listed WikiText-103 validation documents (CC BY-SA 3.0) with disjoint damage and diagnostic token splits, and the language model trains on WikiText-103-raw-v1 and is scored on its entire official validation split. Seeds control initialization and batch order jointly and all conditions share them, so every contrast is paired. Confidence intervals are percentile bootstrap intervals over seeds with 10,000 resamples; p-values are paired t-tests over seeds, with Wilcoxon signed-rank tests reported alongside for pre-registered contrasts and TOST equivalence tests at a ± 1.0 pt margin where a null is asserted. Appendix E gives architecture, optimizer and hyperparameters; Appendix B gives full proofs; Appendix C lists every pre-registered threshold, every post-lock experiment, every deviation, and one corrected implementation bug.

Use of AI assistance

Generative AI tools were used in this work: (i) to write the experimental infrastructure code (training and evaluation harness, outage evaluation, analysis and figure scripts, Docker setup); (ii) to edit the manuscript for grammar and wording, and to draft parts of it; (iii) as simulated reviewers, whose critiques of earlier drafts prompted several of the controls and corrections reported in Appendix C; and also for literature discovery and for checking the mathematical arguments.

References

Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 2002. doi: 10.1016/S0378-4266(02)00281-9.

Guan-Jie Chen, Xinyu Zhao, Tianlong Chen, and Yu Cheng. MoE-RBench: Towards Building Reliable Language Models with Sparse Mixture-of-Experts. *International Conference on Machine Learning*, 2024.

Tianlong Chen, Zhenyu (Allen) Zhang, Ajay Jaiswal, Shiwei Liu, and Zhangyang Wang. Sparse MoE as the New Dropout: Scaling Dense and Self-Slimmable Transformers. *International Conference on Learning Representations*, 2023.

Tianyu Chen, Shaohan Huang, Yuan Xie, Binxing Jiao, Daxin Jiang, Haoyi Zhou, Jianxin Li, and Furu Wei. Task-Specific Expert Pruning for Sparse Mixture-of-Experts. *arXiv.org*, 2022.

Sebastian Curi, Kfir Y. Levy, Stefanie Jegelka, and Andreas Krause. Adaptive Sampling for Stochastic Risk-Averse Learning. *Neural Information Processing Systems*, 2019.

Damai Dai, C. Deng, Cheng-Gang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jia-Shi Li, Wangding Zeng, Xing-Kai Yu, Yu Wu, Zhen-Da Xie, Y. K. Li, Pan-Pan Huang, Fuli Luo, C. Ruan, Zhifang Sui, and W. Liang. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. *Annual Meeting of the Association for Computational Linguistics*, 2024.

Yanbo Fan, Siwei Lyu, Yiming Ying, and Bao-Gang Hu. Learning with Average Top-k Loss. *Neural Information Processing Systems*, 2017.

W. Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of machine learning research*, 2021.

Swapnil Gandhi and Christoforos E. Kozyrakis. Sparse Checkpointing for Fast and Reliable MoE Training. *Symposium on Networked Systems Design and Implementation*, 2024.

Aidan N. Gomez, Ivan Zhang, Siddhartha Rao Kamalakara, Divyam Madaan, Kevin Swersky, Yarin Gal, and Geoffrey E. Hinton. Learning Sparse Networks Using Targeted Dropout. *arXiv.org*, 2019.

I. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations*, 2014.

Ajay Jaiswal, Jianyu Wang, Yi-Xiao Li, Pingzhi Li, Tianlong Chen, Zhangyang Wang, Chong Wang, Ruoming Pang, and Xianzhi Du. Finding Fantastic Experts in MoEs: A Unified Study for Expert Dropping Strategies and Observations. *arXiv.org*, 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, B. Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lelio Renard Lavaud, Lucile Saulnier, M. Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Theophile Gervet, Thibaut Lavril, Thomas Wang, Timothee Lacroix, and William El Sayed. Mixtral of Experts. *arXiv.org*, 2024.

Kenji Kawaguchi and Haihao Lu. Ordered SGD: A New Stochastic Optimization Framework for Empirical Risk Minimization. *International Conference on Artificial Intelligence and Statistics*, 2020.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, M. Krikun, Noam Shazeer, and Z. Chen. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. *International Conference on Learning Representations*, 2020.

R. Liu, Young Jin Kim, Alexandre Muzio, Barzan Mozafari, and H. Awadalla. Gating Dropout: Communication-efficient Regularization for Sparsely Activated Transformers. *International Conference on Machine Learning*, 2022.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. Not All Experts are Equal: Efficient Expert Pruning and Skipping for Mixture-of-Experts Large Language Models. *Annual Meeting of the Association for Computational Linguistics*, 2024.

A. Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *International Conference on Learning Representations*, 2017.

Ronak Mehta, Vincent Roulet, Krishna Pillutla, Lang Liu, and Zaid Harchaoui. Stochastic Optimization for Spectral Risk Measures. *arXiv preprint*, 2022.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Daniel Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hanna Hajishirzi. OLMoE: Open Mixture-of-Experts Language Models. *arXiv.org*, 2024.

C. Neti, M. H. Schneider, and E. D. Young. Maximally fault tolerant neural networks. *IEEE Transactions on Neural Networks*, 1992. doi: 10.1109/72.105414.

Sungrae Park, Jun-Keon Park, Su-Jin Shin, and Il-Chul Moon. Adversarial Dropout for Supervised and Semi-supervised Learning. *AAAI Conference on Artificial Intelligence*, 2017.

D. S. Phatak and I. Koren. Complete and partial fault tolerance of feedforward neural nets. *IEEE Transactions on Neural Networks*, 1995. doi: 10.1109/72.363479.

J. Puigcerver, Rodolphe Jenatton, C. Riquelme, Pranjal Awasthi, and Srinadh Bhojanapalli. On the Adversarial Robustness of Mixture of Experts. *Neural Information Processing Systems*, 2022.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2000. doi: 10.21314/JOR.2000.038.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *arXiv.org*, 2019.

Shai Shalev-Shwartz and Yonatan Wexler. Minimizing the Maximal Loss: How and Why? *International Conference on Machine Learning*, 2016.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and J. Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *International Conference on Learning Representations*, 2017.

Zunhai Su, Qingyuan Li, Hao Zhang, Weihao Ye, Qibo Xue, YuLei Qian, Yuchen Xie, Ngai Wong, and Kehong Yuan. Unveiling Super Experts in Mixture-of-Experts Large Language Models. *arXiv.org*, 2025.

Xun Sun, Shaoyuan Chen, Pingchuan Ma, Yue Chen, Ziwei Yuan, Zhanhao Cao, Han Han, Shangming Cai, Teng Ma, Xuchun Shang, Xinpeng Zhao, Ke Yang, Junlin Wei, Lianzhi Lin, Yuji Liu, Feng Ren, Haoran Hu, Cheng Wan, Yingdi Shan, Yongwei Wu, and Mingxing Zhang. Surviving Partial Rank Failures in Wide Expert-Parallel MoE Inference. *arXiv.org*, 2026.

Cesar Torres-Huitzil and Bernard Girau. Fault and Error Tolerance in Neural Networks: A Review. *IEEE Access*, 2017. doi: 10.1109/ACCESS.2017.2742698.

Yun Wang, Ling Yang, Sen-Hao Yu, Yi-Xiao Wang, Ruixing Li, Zhi-Xiang Wei, James Yen, and Zheng-Wei Qi. BuddyMoE: Exploiting Expert Redundancy to Accelerate Memory-Constrained Mixture-of-Experts Inference. *arXiv.org*, 2025.

Yuan Xie, Shaohan Huang, Tianyu Chen, and Furu Wei. MoEC: Mixture of Expert Clusters. *AAAI Conference on Artificial Intelligence*, 2022.

Yihua Zhang, Ruisi Cai, Tianlong Chen, Guanhua Zhang, Huan Zhang, Pin-Yu Chen, Shiyu Chang, Zhangyang Wang, and Sijia Liu. Robust Mixture-of-Expert Training for Convolutional Neural Networks. *IEEE International Conference on Computer Vision*, 2023.

Yanqi Zhou, Tao Lei, Han-Chu Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V. Le, and J. Laudon. Mixture-of-Experts with Expert Choice Routing. *Neural Information Processing Systems*, 2022.

Yixiao Zhou, Ziyu Zhao, Dong-Zhou Cheng, Zhi-Liang Wu, Jie Gui, Yi Yang, Fei Wu, Yu Cheng, and He-He Fan. Dropping Experts, Recombining Neurons: Retraining-Free Pruning for Sparse Mixture-of-Experts LLMs. *Conference on Empirical Methods in Natural Language Processing*, 2025.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, J. Dean, Noam Shazeer, and W. Fedus. ST-MoE: Designing Stable and Transferable Sparse Expert Models. *arXiv preprint*, 2022.

A Extended related work

Expert absence, removal and redundancy. Chen et al. (2023) train with randomly sampled expert subsets so a model can be slimmed at inference, and Liu et al. (2022) randomly skip cross-device dispatch; our RandDrop baseline is the *common core* of the two, re-implemented in our harness so it shares initialization, batch order and seeds with the treatments — numbers here are not comparable to theirs. Gomez et al. (2019) also make dropout non-uniform, but in the opposite direction, dropping the units a pruning criterion marks *expendable* where FailSafe-RM drops the most exposed ones, and a large literature removes experts deliberately (Chen et al., 2022; Lu et al., 2024; Jaiswal et al., 2025), repairs the model afterwards (Zhou et al., 2025) or substitutes related experts under memory pressure (Wang et al., 2025) at modest *average* quality loss — but a pruning method chooses which expert to remove, and an outage chooses for you. Puigcerver et al. (2022) and Zhang et al. (2023) perturb the *input* and assume the expert set intact; Chen et al. (2024) benchmark MoE reliability without measuring worst-case expert unavailability.

B Proofs

B.1 Proof of Lemma 1

Let the experts be relabelled by damage rank, so expert j is the j-th most damaging to lose; the labelling is well defined because the damages are assumed distinct. A sampled-max adversary draws a uniformly random m-subset $C \subseteq \{1, \dots, E\}$ and returns $\arg\max_{e \in C} \mathrm{Dmg}(e)$, whose rank is $R_m = \min C$.

$\min C = j$ requires $j \in C$ and the remaining $m-1$ elements to be drawn from the $E-j$ labels strictly greater than j. Hence $\Pr[R_m = j] = \binom{E-j}{m-1}/\binom{E}{m}$, and setting $j{=}1$ gives $\Pr[R_m = 1] = \binom{E-1}{m-1}/\binom{E}{m} = m/E$.

For the mean, use $\mathbb{E}[R_m] = \sum_{j \geq 0} \Pr[R_m > j]$. The event $R_m > j$ means all m draws avoid $\{1, \dots, j\}$, so $\Pr[R_m > j] = \binom{E-j}{m}/\binom{E}{m}$. Therefore

$$\mathbb{E}[R_m] = \frac{1}{\binom{E}{m}} \sum_{j=0}^{E-m} \binom{E-j}{m} = \frac{1}{\binom{E}{m}} \sum_{i=m}^{E} \binom{i}{m} = \frac{\binom{E+1}{m+1}}{\binom{E}{m}} = \frac{E+1}{m+1},$$

where the middle step substitutes $i = E - j$ and the last uses the hockey-stick identity $\sum_{i=m}^{E} \binom{i}{m} = \binom{E+1}{m+1}$ and $\binom{E+1}{m+1}/\binom{E}{m} = \frac{(E+1)!\, m!\, (E-m)!}{(m+1)!\, (E-m)!\, E!} = \frac{E+1}{m+1}$. □

This is the distribution of the minimum of a uniform m-subset of $\{1, \dots, E\}$, a standard order-statistics fact; we state it because the first identity is what turns an adversary into a weighting, not because it is new.

B.2 Proof of Observation 1 and Corollary 1

Attribution. The identity below is not new. Kawaguchi & Lu (2020) prove (their Theorem 1) that keeping the top q losses of a size-s minibatch drawn from n examples is a stochastic first-order method for $\frac{1}{q}\sum_j \gamma_j L_{(j)}$ with $\gamma_j = \left[\sum_{l=0}^{q-1} \binom{j-1}{l}\binom{n-j}{s-l-1}\right]/\binom{n}{s}$; at $q{=}1$ this is exactly $\gamma_j = \binom{n-j}{s-1}/\binom{n}{s}$, i.e. our $w_j^{(m)}$ under $n \to E$, $s \to m$. What we add is the substitution of *outage sets* for data indices, the reading of the resulting weights as a risk spectrum over a damage distribution produced by rerouting, and the consequences we draw from $w_1^{(m)} = m/E$. We restate the proof because the substitution changes what the index set is, not because the combinatorics is ours.

The identity. $\max_{e \in C_m} \mathcal{L}(\theta; \{e\}) = L_{(j)}$ if and only if $\min C_m = j$ in the rank labelling above, so taking expectations over the uniform m-subset, $\mathbb{E}_{C_m}[\max_{e \in C_m} \mathcal{L}] = \sum_j \Pr[R_m = j]\, L_{(j)} = \sum_j w_j^{(m)} L_{(j)}$ with $w_j^{(m)} = \binom{E-j}{m-1}/\binom{E}{m}$.

The weights form a risk spectrum. Non-negativity is immediate and $\sum_{j=1}^{E} w_j^{(m)} = \sum_j \Pr[R_m = j] = 1$. For monotonicity, for $1 \leq j < E$ with $w_{j+1}^{(m)} > 0$,

$$\frac{w_j^{(m)}}{w_{j+1}^{(m)}} = \frac{\binom{E-j}{m-1}}{\binom{E-j-1}{m-1}} = \frac{(E-j)!\,(E-j-m)!}{(E-j-m+1)!\,(E-j-1)!} = \frac{E-j}{E-j-m+1} \;\geq\; 1,$$

since $m \geq 1$. So $w^{(m)}$ is non-increasing in the rank index j, i.e. non-increasing in "how good" the outage is and non-decreasing in severity. A discrete risk functional $\rho(X) = \sum_j w_j X_{(j)}$ whose weights are non-negative, sum to one and are non-increasing in j is exactly a spectral risk measure in the sense of Acerbi

(2002), and is therefore coherent (monotone, sub-additive, positively homogeneous, translation invariant). Extremes: $m{=}1$ gives $w_j^{(1)} = \binom{E-j}{0}/\binom{E}{1} = 1/E$ for all j, the uniform spectrum, whose value is the outage *average*; $m{=}E$ gives $w_1^{(E)} = 1$ and $w_j^{(E)} = 0$ for $j > 1$, the maximum. The weight on the worst outage is $w_1^{(m)} = \binom{E-1}{m-1}/\binom{E}{m} = m/E$ at every m. □

Monotonicity in m. Draw a uniform $(m{-}1)$-subset C_{m-1} and then a uniform element x of $[E] \setminus C_{m-1}$, and set $C_m = C_{m-1} \cup \{x\}$. For a fixed m-set A the construction reaches A through any of its m subsets of size $m-1$, so $\Pr[C_m = A] = m/\left[\binom{E}{m-1}(E-m+1)\right] = 1/\binom{E}{m}$: C_m is uniform. Since $C_{m-1} \subseteq C_m$ pointwise under this coupling, $\max_{e\in C_{m-1}} \mathcal{L} \le \max_{e\in C_m} \mathcal{L}$ and the expectations are ordered. So increasing m moves monotonically from the average to the maximum, which is the sense in which m is a risk-aversion knob. □

Corollary 1. The adversarial term of Eq. 1 is $\beta \sum_j w_j^{(m)} L_{(j)}$, i.e. the worst outage enters the training objective with weight $w_1^{(m)} = m/E$, and $1/E$ at $m{=}1$. Wherever the rank labelling is locally constant the gradient of that term is $\beta \sum_j w_j^{(m)} \nabla_\theta L_{(j)}$, so the same weight multiplies the worst outage's gradient at every step and over any number of steps. Two caveats. The rank-1 outage is an *index*, not a fixed expert: which expert attains rank 1 may change from step to step, so this is a weight on the worst outage of the current parameters, not a per-expert budget. And the per-step maximum over a minibatch is an upward-biased estimate of the population spectral risk (Mehta et al., 2022), so the objective an optimiser actually descends is not exactly $\sum_j w_j^{(m)} L_{(j)}$. □

Remark (a mixture of CVaRs). Every discrete risk spectrum is a mixture of CVaR levels — write $w^{(m)}$ as a non-negative combination of the flat spectra $\{1/\lceil \alpha E \rceil$ on the worst $\lceil \alpha E \rceil$ outages$\}$, which is possible exactly because $w^{(m)}$ is non-increasing — so Eq. 2 *is* a CVaR mixture. What it is not is a *single* CVaR level: for $1 < m < E$ the weights are strictly decreasing and supported on every rank, unlike CVaR_α and unlike the flat average-top-k spectrum of Fan et al. (2017), which averages the k largest losses over the whole index set with no subsampling and hence no binomial weights. Shalev-Shwartz & Wexler (2016) give the argument for minimising the maximum rather than the average, but reach the max by exponential-weight importance sampling over data and derive no order-statistic weights. The submitted version of this paper claimed the family *interpolates* CVaR levels; that is withdrawn and the mixture fact stated instead.

Remark (why a coverage argument is empty here). One might instead argue that uniform sampling may never *see* the worst outage: with a fixed damage ordering the probability of missing it for T steps is $(1 - m/E)^T$. At our $T{=}1000$ and $E{=}32$ that is about 10^{-14}, so it explains nothing. Uniform dropout sees the worst outage roughly T/E times; what it does is discount it by $1/E$ relative to an exact adversary. We have replaced the coverage statement with the gradient-weight statement above.

Remark (what m buys in rank terms). Since $\mathbb{E}[R_m] = (E+1)/(m+1)$, the marginal reduction in expected rank from one extra sample is $(E+1)/[(m+1)(m+2)]$, which falls quadratically in m; equivalently $w_1^{(m)} = m/E$ grows only linearly while the objective approaches the max much faster, which is why FAILSAFE-SAMPLED with $m{=}4$ is already close to exhaustive search.

Remark (non-uniform base measures). If the single dropped expert is drawn from a distribution p on $[E]$ instead of the uniform one, the inner objective is $\sum_e p_e \, \mathcal{L}(\theta; \{e\})$, i.e. a p-weighted average, which is again of the form $\sum_j \tilde{w}_j L_{(j)}$ with $\tilde{w}_j = p_{e_{(j)}}$. The weights are non-negative and sum to one but are *not* sorted by rank unless p happens to be monotone in damage, so this is a spectral measure with respect to the base measure p rather than with respect to the uniform one. FAILSAFE-RM takes $p_e \propto \mathrm{RM}(e)$; the extent to which that tilts mass towards the tail is exactly the extent to which router mass predicts damage, which §8 measures ($\rho = 0.650$ on SSB, $+0.551$ within OLMoE layers).

Remark (sampling with replacement, and a correction to our pre-registration). If the m candidates are drawn *with* replacement, $\Pr[R_m = 1] = 1 - (1 - 1/E)^m$. Our pre-registered plan wrote the hit probability in that form, but the implementation draws *without* replacement, for which the correct expression is m/E. The two agree at $m{=}1$ and diverge sharply at large m: at $m{=}E{=}32$ they give 0.638 and 1.000. Measured hit rates and their confidence intervals are in Table 25: the without-replacement value lies inside the 95% CI at 6 of 6 budgets, so we report the without-replacement formula as the right one without claiming an exact match at every m.

Remark (ties and the subgradient). The implementation penalises $\max_{e\in C} \mathcal{L}_e$ through `torch.max`, which routes the whole subgradient to a single attaining index and, on ties, returns the *first* such index. Any element of the argmax set is a valid Danskin subgradient, so this is correct but asymmetric over tied maximisers. Empirically it does not matter: 0 of the $5{\times}40$ probes had an exact tie at the maximum (Table 25).

Remark (the rank probe, and a correction). Reporting the *mean selected rank* of a uniformly drawn expert requires a tie convention, and the submitted version of this paper used the min-rank convention without saying so. That is not the convention Lemma 1 models. Under the min-rank convention the measured mean rank at m=1 is 12.27 at E=32, apparently far below the tie-free prediction $(E+1)/2 = 16.5$; under the mid-rank convention (average of tied ranks) it is 15.76, i.e. within -0.74 of the prediction. The discrepancy was an artefact: at the locked α_{aux} only 17.1 of 32 experts carry any traffic in a 256-token batch, so roughly fifteen zero-traffic experts share the clean loss exactly and the min-rank rule assigns all of them the best rank in the tie block. The mid-rank number is the one the paper quotes (5 seeds × 40.0 probes per seed).

B.3 Proof of Proposition 1

Write $y(h)$ for the clean layer output of §3 and $y^{\{e\}}(h)$ for its value when expert e is unavailable, and let $\delta_h = y^{\{e\}}(h) - y(h)$ be the *outage displacement*. Router logits are written $z(h)$; ℓ denotes the per-example loss throughout.

Step 1: only exposed tokens move (A2). If $e \notin S_k(h)$ then removing e changes neither the selected set nor the gates, so $\delta_h = 0$. This is exact for reroute-on-failure with gates renormalized over the selected set, and also for the truncated-softmax convention (`norm_topk_prob=false`) provided the failed expert is removed from the candidate set *after* the softmax; it is *false* if the router logit is masked before the softmax, which rescales every surviving gate by $1/(1-p_e)$. The expectation below is therefore effectively restricted to $\{h : e \in S_k(h)\}$, and $g_e(h)$ acts as the indicator-weighted exposure.

Step 2: the exact displacement. Let $e \in S_k(h) = S$ and let σ be the expert that enters the selected set in its place. Write $u_h = \sum_{e' \in S \setminus \{e\}} g_{e'}(h) f_{e'}(h) / (1 - g_e(h))$ for the convex combination of the other selected experts, and let a_h be the gate weight the substitute actually acquires.

(i) Renormalized gates. Here the surviving gates are rescaled by $\lambda = Z/Z'$ with $Z = \sum_{e' \in S} e^{z_{e'}}$ and $Z' = Z - e^{z_e} + e^{z_\sigma}$, and $\lambda(1 - g_e) + a_h = 1$ with $a_h = e^{z_\sigma}/Z'$. Then $y = (1 - g_e)u_h + g_e f_e$ and $y^{\{e\}} = (1 - a_h)u_h + a_h f_\sigma$, so

$$\delta_h = (g_e - a_h)u_h + a_h f_\sigma - g_e f_e = \underbrace{g_e(h)\big(f_\sigma(h) - f_e(h)\big)}_{\text{gate transfer}} + \underbrace{\big(g_e(h) - a_h\big)\big(u_h - f_\sigma(h)\big)}_{\text{residual}}. \tag{4}$$

(ii) Truncated softmax without renormalization. Here surviving gates are unchanged and $a_h = g_\sigma(h)$, so $\delta_h = a_h f_\sigma - g_e f_e = g_e(f_\sigma - f_e) + (g_e - a_h)(0 - f_\sigma)$, again a gate-transfer term plus a residual proportional to $g_e(h) - a_h$.

The residual is exactly zero when $a_h = g_e(h)$, i.e. when $z_\sigma(h) = z_e(h)$, and grows with the gap between the k-th and $(k+1)$-th router logits — that is, with specialization. If several survivors share the freed mass — for instance when $|S| = q > 1$ experts fail — the same algebra holds with f_σ replaced by the convex combination $\bar{f}_{\sigma_e}(h) = \sum_j w_j(h) f_j(h)$, which is the form stated in Proposition 1. We check Eq. 4 on a real MoE: reconstructing δ_h analytically from cached expert outputs and comparing against the actually patched block gives cosine similarity 0.9923 (§7).

Step 3: first-order expansion (A1). Let $s_h = \partial \ell / \partial y \big|_{y(h)}$. If the per-example loss is L-smooth in the layer output, $\ell(y^{\{e\}}(h)) - \ell(y(h)) = \langle s_h, \delta_h \rangle + r_h$ with $|r_h| \le \frac{L}{2}\|\delta_h\|^2$. Taking expectations, $\mathrm{Dmg}(e) = \mathbb{E}_h\langle s_h, \delta_h \rangle + R$ with $|R| \le \frac{L}{2}\mathbb{E}_h\|\delta_h\|^2$.

Step 4: bound. By Cauchy–Schwarz applied pointwise and the triangle inequality on Eq. 4,

$$\begin{aligned} |\mathbb{E}_h\langle s_h, \delta_h \rangle| \;&\le\; \mathbb{E}_h\|s_h\|\,\|\delta_h\| \\ &\le\; \mathbb{E}_h\big[\|s_h\|\, g_e(h)\,\|\bar{f}_{\sigma_e}(h) - f_e(h)\|\big] \;+\; \mathbb{E}_h\big[\|s_h\|\,|g_e(h) - a_h|\,\|u_h - \bar{f}_{\sigma_e}(h)\|\big], \end{aligned}$$

which is $\mathrm{EDS}(e)$ plus the residual term. Combining with Step 3 gives Eq. 3. □

Remark (what changed, and why). The submitted version of this paper stated Eq. 3 without the residual term, treating the gate-transfer assumption as exact, and described the resulting inequality as "correct". It is an inequality only with the residual included, which is the form above; the omission was caught in internal review. Nothing in the empirical sections depends on the difference, because the diagnostic we actually computed and pre-registered is $\mathrm{EDS}(e)$, the first term.

Remark (scope: one failed expert). Proposition 1 is stated and proved for the single-substitute case $q{=}1$: one expert fails and its gate mass is absorbed by the convex mixture $\bar{f}_{\sigma_e}$ of the survivors the router re-selects. That is the only case any of our training runs or measurements uses. We do not assert a multi-expert ($q > 1$) version here; the same telescoping argument suggests one, but we have not derived the corresponding residual and make no claim about it.

Remark (how loose the bound is, measured). We evaluated both sides in loss units. On SSB at $E{=}32$ under conventional training, the ratio $\mathrm{EDS}(e)/|\mathrm{Dmg}(e)|$ has pooled median 30.8, IQR [18.6, 61.5] over 319 (seed, expert) pairs, and the inequality holds for 99.7 % of them. On OLMoE the median ratio is 435 over 1024 (layer, expert) rows (IQR [239, 909]; the bound holds for 100 %), with the caveat that EDS is accumulated on the diagnostic token split and ΔNLL on the damage split, so this is a distributional rather than a token-matched ratio. Most of the gap is the Cauchy–Schwarz step, not the second-order remainder: the *signed* first-order term, before the relaxation, has median $|\mathbb{E}_h\langle s_h, \delta_h\rangle|/|\mathrm{Dmg}(e)|$ of 0.33 (IQR [0.13, 0.82]). The bound is therefore valid and nearly uninformative quantitatively — which is consistent with, not in tension with, the refutation of the ranking hypothesis it motivated.

Remark (what the bound cannot do). Eq. 3 bounds $|\mathrm{Dmg}(e)|$ by products of norms, so the right-hand side is non-negative. It can order experts by how much removing them perturbs the layer output, but it cannot represent a removal that *improves* the loss; on the pretrained MoE we audit, 16.2 % of single outages do exactly that. The bound is tight when s_h and δ_h are aligned and loosens as $\mathbb{E}\|\delta_h\|^2$ grows relative to the linear term. On OLMoE the displacement on a routed token is 27 % of the block output and the *signed* first-order estimate $\mathbb{E}_h\langle s_h, \delta_h\rangle$ attains Spearman $+0.092$ against measured damage, so at that scale the expansion is outside its useful radius. This is a limitation of the first-order analysis, not of the Cauchy–Schwarz step: the signed estimate, which is the quantity *before* the relaxation, is the worse predictor of the two.

Remark (relation to router mass). $\mathrm{RM}(e) = \mathbb{E}_h[g_e(h)]$ is the exposure factor alone. Eq. 3 is proportional to $\mathrm{RM}(e)$ when disagreement and sensitivity are constant across experts and uncorrelated with exposure. Specialization ought to break this: an expert whose family contains a near-duplicate has small $\|\bar{f}_{\sigma_e} - f_e\|$ regardless of its traffic, while an expert holding an isolated skill has large disagreement at modest exposure. Empirically (§8) exposure nevertheless dominates the dispersion of the product in every setting we measured (within-layer CV 0.655 for exposure against 0.187 and 0.139), which is why router mass is the stronger rank predictor in practice — and, with hindsight, why our pre-registered threshold of 0.25 *above* router mass was close to unattainable.

C PRE-REGISTRATION, DEVIATIONS AND CORRECTIONS

Table 2 restates every pre-registered hypothesis, the threshold fixed before any run, and the outcome. Everything in this section is a deviation, a correction, or a caveat that a reader checking our pre-registration against the paper would want flagged.

Seeds. The plan pre-committed to 5 matched seeds, extended to 10 if the observed per-condition standard deviation of `worst1_acc` exceeded 3 pt. It does (0.68–3.99 pt across cells, the maximum at CLEAN, $E{=}32$), so we ran 10 seeds and read *every* pre-registered verdict off the 10-seed batch. The ablations in Table 27 (β, top-k) and the FMNIST-Rot study were run at 5 seeds to fit the compute budget, and their n is stated wherever they appear. A three-seed pilot batch of the main experiment exists in the run logs; it is superseded and no number from it appears in this paper.

Exploratory, post-lock experiments. The metric lock list was fixed before any run. Everything below was registered afterwards, carries no pre-registered threshold, and cannot change any hypothesis verdict; all of it is labelled exploratory wherever it appears. E8, the matched-compute max-versus-mean control and its soft aggregator; E9/E9B, the load-regime sweeps for the diagnostic; E10, the wall-clock cost measurements; E3X, the factor decomposition of EDS; E11, the real-MoE audit of §7. **Internal review prompted a second block**, also post-lock: E12 (the load-regime sweep for *training*, Table 16) and its R14 load-balancing arms (Table 17); E13 (iso-FLOP and iso-step controls, and a 4000-step convergence control); E14 (held-out selection of the worst outage); E15 (two-outage additivity, Table 20); E16 (the top-k plane, Table 22); E17 (the exposure-weighted adversaries, including FAILSAFE-RM, Table 18); E18 (pool-size confound controls); E19 (the shared always-on expert, Table 21); E20 (tie structure of the adversary's argmax, Table 25); E21 (the outage protocol, Table 23); E22 (the expert-FLOP accounting, Table 26); and the R6–R9 re-analyses (dispersion, Wilcoxon tests, FMNIST load skew, damage-versus-rank, top-N coverage). A third block added E23–E28, the controls of Tables 13, 14 and 15; then **E29**, the 6-layer 99 M-parameter MoE language model of §6.5 and Appendix F: registered after the lock, run at $n{=}3$ seeds, carrying no pre-registered threshold, and reported as a scope check on the toy-scale results rather than as evidence for any hypothesis; and finally **E30**, the RANDDROP-LIVE mechanism-isolation control of §6.4 and Table 12, which reuses the E12/E17/E23 baseline rows (bit-identical on re-execution)

and adds one new arm at 4 cells. **FAILSAFE-RM is therefore a post hoc method**: it was not part of the pre-registered design, it was proposed during internal review, and it is evaluated on two benchmark families (SSB at 10 seeds and FMNIST-Rot at 5). We present it because suppressing it would misrepresent what the evidence supports, not because it was planned.

A claim we withdrew. The submitted version described uniform random-outage training as optimising the *median* outage, and the sampled-max family as a CVaR interpolation. Both are wrong. At $m{=}1$ the objective is the outage *average*, not the median; and the sampled-max weights are strictly decreasing in rank for $1 < m < E$, so the family is a *mixture* of CVaR levels — as every discrete risk spectrum is — but not a single CVaR level and not an interpolation between them (Observation 1 and Appendix B). The submitted version also derived the blind spot of uniform sampling from a coverage probability $(1 - m/E)^T$ that is numerically vacuous at our T; the correct statement is about the weight the worst outage carries in the training objective (Corollary 1). The submitted version further presented the order-statistic identity as new; it is Kawaguchi & Lu (2020)'s Theorem 1 at $q{=}1$ with data indices replaced by outage sets, and the paper now attributes it. Finally, the rank probe was reported under a min-rank tie convention that does not match the tie-free model; the mid-rank number is now quoted (Appendix B).

A bound stated without its residual. The submitted version stated Eq. 3 without the (A3) residual term and called it "correct". Appendix B carries the residual explicitly. The pre-registered diagnostic is unaffected: it is the first term, which is what we computed.

A corrected implementation bug in the diagnostic. The sensitivity factor of EDS was originally differentiated through a residual connection that the trained models do not have (all locked runs use `residual=False`), so the gradient was taken of a function the model does not compute. We fixed it and kept the old score as a separate predictor so that the effect of the fix on the verdict is auditable rather than silent: the pre-bugfix score reaches $\rho = 0.647$ (paired Δ against router mass -0.003, $p = 0.615$) and the fixed score $\rho = 0.645$ ($\Delta = -0.005$, $p = 0.279$). Both miss both pre-registered thresholds; the fix does not change the H3 verdict.

A threshold we should not have set. H3 required EDS to beat router mass by ≥ 0.25 in Spearman ρ. EDS is router mass multiplied by two factors whose across-expert dispersion is far smaller than exposure's, so that margin was close to unattainable a priori. This is a criticism of our pre-registration, not a defence of the result: H3 is refuted and stays refuted.

Factor decomposition on the locked models (exploratory). Scoring each factor of Eq. 3 and each leave-one-factor-out product against measured per-expert damage on the same 120 models gives the same picture as the OLMoE audit: disagreement alone reaches $\rho = -0.243$ and sensitivity alone 0.061, against 0.650 for exposure alone (router mass), while exposure$\times$disagreement reaches 0.650 (Δ vs router mass -0.000, $p = 0.938$) and exposure$\times$sensitivity 0.645 (-0.005, $p = 0.411$). Exposure carries essentially all of the predictive signal.

A corrected formula in the pre-registration. The plan stated the adversary's hit-probability as $1 - (1 - 1/E)^m$, which is the with-replacement expression, whereas the implementation samples without replacement. Lemma 1 gives the correct value m/E and Table 25 gives the measurement.

A load-regime sweep whose manipulation check failed. Our first attempt to reach the balanced-load regime varied the Zipf exponent of the benchmark's *skill* frequencies. It did not work: with the locked auxiliary-loss coefficient the realized max/mean router mass stayed in a narrow band of 7.42–12.40$\times$ with no trend in the exponent, so that sweep does not test balanced load and we draw no conclusion about that regime from it. The repaired experiment (Tables 28 and 16) moves the knob that does control router load, the Switch auxiliary-loss coefficient, and its manipulation check passes.

Multiplicity. Only the four pre-registered thresholds of Table 2 carry confirmatory weight. Every other p-value in the paper is descriptive and uncorrected, and there are many of them: the revision alone archives 493 contrasts over 36 result directories.

Sanity checks. All 4 of 4 pre-registered sanity checks pass: a tiny run reaches 72.50 % accuracy at $E{=}8$, far above the eight-class chance level; the failure mode exists under CLEAN at $E{=}32$ (14.04 pt worst-case drop); the monotonicity `worst1` $\leq$ `mean1` $\leq$ `clean` holds in all 606 runs with 0 violations; and the RANDDROP baseline reproduces its known qualitative effect, beating CLEAN on average-outage accuracy at every pool size (72.98 vs 69.86 at $E{=}8$, 63.13 vs 62.87 at $E{=}32$). The post-lock block passes a further 5 of 5 checks, including a bit-for-bit reproduction of the locked E1 runs by the $\alpha_{\text{aux}}{=}0.001$ cells of E12.

Table 2: Pre-registered hypotheses, thresholds fixed before any run, and outcomes. All contrasts are matched-seed paired over 10 seeds at E=32 unless stated; p_W is the Wilcoxon signed-rank p added in revision (it does not re-test the hypotheses). Verdicts are unchanged from the locked analysis.

Hypothesis	pre-registered threshold	measured	p_W	verdict
H1 (Exact)	Δworst-1 $\geq$ +3.0 pt at $p < 0.05$; Δclean $\geq$ -1.0 pt	$+7.06$ pt, $p = 2.8 \times 10^{-8}$, CI $[+6.27, +7.76]$, $d = 5.57$; clean $+0.54$ pt	0.002	**SUPPORTS**
H1 (Sampled)	same	$+4.07$ pt, $p = 4.6 \times 10^{-5}$; clean $+0.49$ pt	0.002	SUPPORTS
H2 (Exact)	gap non-decreasing in E, and gap(32)$-$gap(8) $\geq$ +2.0 pt at $p < 0.05$	$+0.98 \to +3.51 \to +7.06$ pt; $\Delta =$ $+6.09$ pt, $p = 1.5 \times 10^{-5}$	0.002	**SUPPORTS**
H2 (Sampled)	same	$+0.84 \to +3.07 \to +4.07$ pt; $\Delta =$ $+3.23$ pt, $p = 0.002$	0.004	SUPPORTS
H3	ρ(EDS) $\geq$ 0.75 and $\geq$ 0.25 above router mass	ρ(EDS) $= 0.645$ vs ρ(RM) $=$ 0.650; $\Delta = -0.005$, $p = 0.279$	0.193	**REFUTES**
H4 (Exact)	Δworst-2 $\geq$ +2.0 pt at $p < 0.05$	$+8.06$ pt, $p = 1.5 \times 10^{-5}$, CI $[+6.30, +9.82]$	0.002	**SUPPORTS**
H4 (Sampled)	same	$+3.69$ pt, $p = 0.002$	0.002	SUPPORTS

Table 3: All locked SSB cells, mean over 10 matched seeds with the across-seed standard deviation in parentheses (%, and drops in pt). Average-case and worst-case outage damage are reported side by side, as pre-registered.

E	training	clean acc.	mean-1 acc.	worst-1 acc.	mean-1 drop	worst-1 drop
8	CLEAN	71.99 (1.24)	69.86 (0.99)	61.03 (2.65)	2.13 (0.35)	10.96 (3.19)
8	RANDDROP	73.41 (1.09)	72.98 (1.12)	71.67 (1.07)	0.43 (0.17)	1.73 (0.47)
8	FAILSAFE-SAMP.	73.50 (1.23)	73.24 (1.15)	72.51 (0.90)	0.26 (0.14)	0.99 (0.47)
8	FAILSAFE-EXACT	73.72 (1.20)	73.41 (1.15)	72.65 (1.18)	0.31 (0.13)	1.07 (0.32)
16	CLEAN	67.61 (0.80)	66.17 (0.75)	55.40 (3.93)	1.44 (0.11)	12.21 (3.72)
16	RANDDROP	68.38 (0.77)	67.68 (0.73)	63.27 (1.78)	0.70 (0.12)	5.11 (1.56)
16	FAILSAFE-SAMP.	68.85 (0.84)	68.39 (0.84)	66.34 (0.68)	0.46 (0.09)	2.51 (0.65)
16	FAILSAFE-EXACT	68.27 (0.81)	67.97 (0.77)	66.78 (0.70)	0.30 (0.06)	1.49 (0.36)
32	CLEAN	63.55 (0.83)	62.87 (0.75)	49.51 (3.99)	0.69 (0.12)	14.04 (4.49)
32	RANDDROP	63.65 (1.01)	63.13 (0.97)	55.04 (1.45)	0.51 (0.06)	8.61 (2.08)
32	FAILSAFE-SAMP.	64.14 (0.81)	63.79 (0.78)	59.11 (0.98)	0.35 (0.04)	5.03 (1.27)
32	FAILSAFE-EXACT	64.19 (0.76)	63.96 (0.74)	62.10 (0.75)	0.22 (0.02)	2.09 (0.43)

D EXTENDED LIMITATIONS

Scale, and what it buys. Our pre-registered models are a single MoE layer with a shallow shared trunk at $\leq$1M parameters, chosen by design so that the worst outage over all E single outages can be verified exhaustively rather than estimated. That is what makes the claims checkable. One rung up, §6.5 trains a 6-layer 99 M-parameter MoE language model in which the worst outage over all 96 (layer, expert) pairs is still exhaustively verified — but on 45.1 M tokens, well under one epoch of WikiText-103, at n=3 seeds, on one corpus and one architecture, and at a realised load skew of only 2.0–2.7×. None of that is LLM-scale evidence for the *training* method: we never trained at that scale, and the audit of §7 is inference-only, on one model and one text domain, with 6 of 16 layers resolved above the measurement noise floor.

The regime claim is correlational. The load sweep varies α_{aux}, which changes the training objective, so it establishes an association. The instrument built to break that (E24, Table 13) *failed* its manipulation check with a free router and only worked with the router frozen, which is itself an additional intervention; under it the

Table 4: FMNIST-Rot, the external-validity replication, mean over 5 matched seeds with the across-seed standard deviation in parentheses. Columns match Table 1. Realized load skew is 4.49/8.22/15.04× at E=8/16/32, comparable to SSB. The E=8 FAILSAFE-EXACT−RANDDROP contrast is *not* significant (+0.85 pt, $p = 0.062$); E=16 and E=32 are (+2.34 pt, $p = 0.005$; +4.84 pt, $p = 0.002$).

	worst-1 acc.		$E = 32$			
Training	E=8	E=16	clean	mean-1	worst-1	worst-2
CLEAN	53.55 (17.61)	55.28 (8.00)	74.58 (0.34)	73.52 (0.38)	54.53 (11.69)	43.38 (10.42)
RANDDROP	79.17 (0.48)	72.89 (0.81)	74.33 (0.32)	74.08 (0.32)	69.27 (1.31)	56.65 (6.28)
FAILSAFE-SAMP.	79.98 (0.38)	74.88 (0.64)	74.55 (0.18)	74.42 (0.19)	73.05 (0.41)	64.59 (3.27)
FAILSAFE-EXACT	80.02 (0.27)	75.23 (0.26)	74.74 (0.29)	74.68 (0.28)	74.11 (0.40)	66.07 (5.39)

worst-case drop tracks skew at $p = 0.031$ but the EXACT−RANDDROP advantage only at $p = 0.077$. "Tracks" and "is associated with" are therefore the strongest verbs we use.

Status of the two regimes. The pre-registered confirmatory regime (α_{aux}=0.001) has a collapsed router: 7.6 effective experts and 63.4 % no-op candidates per adversary step. That was established after the lock, by the live/dead audit of Table 14, and is why the 0.01 cell is reported as co-primary rather than as a robustness check. The co-primary OLMoE-like regime (0.01) is itself post-lock and exploratory, as is everything in §6.1–§6.4 outside block (b) of Table 1.

FAILSAFE-RM is post hoc. It was registered after the metric lock, and its second benchmark (FMNIST-Rot, n=5 cells) later still; both are exploratory throughout. The ±1.0 pt equivalence margin is the pre-registered *clean-accuracy* non-inferiority margin re-used on worst-1 accuracy, which is a post-hoc reuse. Across 4 independently drawn benchmark instances the pool-size gap varies by ±0.62 pt, more than the seed-level CIs suggest.

Cost statements are derived, not measured. The top-$(k{+}1)$ implementation that would make an exact adversary cheap independently of E is derived in Table 26 and was *not* implemented or measured; our wall-clock numbers are against a dense CLEAN arm. The per-outage re-mix and output head do not shrink with E, so the exact adversary is not affordable at LLM scale on this evidence.

Multiplicity. We Holm-correct two families — the four regime contrasts and the three RM-vs-Exact contrasts — and report everything else as uncorrected exploratory contrasts. With 232 contrasts in the round-2 batch alone, individual exploratory p-values should be read accordingly.

The uniform baseline drops one expert. RANDDROP removes a single expert per step, which is the q=1 core of Chen et al. (2023) and Liu et al. (2022) and the matched control for our q=1 adversaries. A per-expert Bernoulli mask, or the growing random *subsets* of Chen et al. (2023), is a different and possibly stronger uniform scheme; we did not run one, so we cannot say how much of the gap at q=1 it would close. EXACT-LIVE and RANDDROP-LIVE (§6.4) vary the support of the draw, not the number of experts dropped.

Scope of the threat model. We study *global* outages under reroute-on-failure with token-choice top-k routing. Expert-choice routing (Zhou et al., 2022), per-token or adaptive adversaries, capacity drops, partial-bandwidth degradation and Byzantine (rather than absent) experts are out of scope.

E EXPERIMENTAL SETUP

Algorithm 1 FailSafe-MoE training step (budget q=1)

1: Sample batch B; compute encoder states h, expert outputs $f_e(h)$, router logits $z(h)$
2: $\mathcal{L}_0 \leftarrow \mathcal{L}(\theta; \emptyset) + \alpha_{\text{aux}}\, \mathcal{L}_{\text{balance}}$
3: $C \leftarrow [E]$ (exact), or m experts uniform without replacement (sampled), or one expert $\propto$ batch router mass (RM)
4: $\mathcal{L}_e \leftarrow \mathcal{L}(\theta; \{e\})$ for all $e \in C$ (re-mix only; no expert recompute)
5: $\theta \leftarrow \theta - \eta \nabla_\theta \left[\mathcal{L}_0 + \beta \max_{e \in C} \mathcal{L}_e\right]$

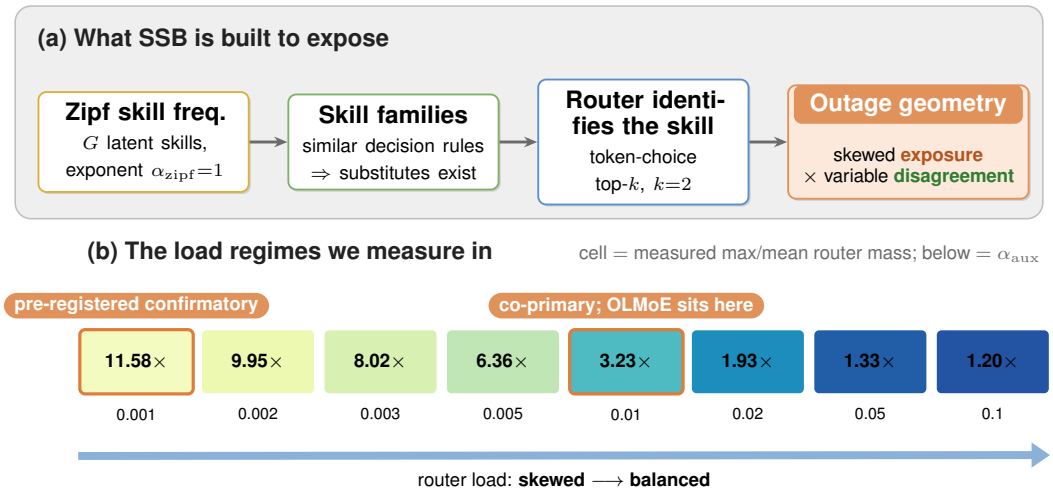


Figure 6: **SSB is built so that exposure is skewed while substitute quality varies independently of it, and the load-balancing coefficient is the knob that moves the regime.** (a) Construction of the specialization stress benchmark (full specification in this appendix). (b) The 8 load regimes of the α_{aux} sweep (E23), each cell shaded by the *measured* max/mean router mass of the CLEAN arm at E=32, n=10 matched seeds per cell; normalised routing entropy runs 0.585 to 0.999 from left to right. The two outlined cells are the regimes reported side by side in Table 1 and Figure 2; the pretrained MoE audited in §7 sits at 3.22×.

Architecture. A single shared encoder (1 linear layer to d_{model}=128 followed by ReLU), one MoE layer with E experts, and a linear classification head. Each expert is a two-layer MLP $128 \to 128 \to 128$ with ReLU. Routing is token-choice top-k with k=2 and softmax gating over the selected logits. The encoder is deliberately shallow and the head is shared, so that skill-specific computation must live in the experts. **This is an architecture chosen to make expert criticality measurable**, not a representative production MoE: a wide shared trunk would absorb much of the damage of an outage, and a deeper stack would spread it over layers. We use a Switch-style load-balancing auxiliary loss with weight α_{aux}=0.001 in the locked runs: enough to keep routing from collapsing, but not so much that it erases the skewed skill frequencies the benchmark is built to express. The resulting realized router load is max/mean = 11.6× at E=32 (normalized routing entropy 0.585). Table 16 varies this coefficient over four orders of magnitude and measures the realized load in every cell.

Optimization. Adam, learning rate 3×10^{-3} with cosine decay, batch size 256, 1000 steps, β=1.0. Within a seed, all conditions share initialization and batch order, so every reported contrast is paired. The convergence control of Table 9 repeats all four conditions at 4000 steps with the cosine schedule stretched to match.

SSB. G latent skills; skill g has embedding $A_g \in \mathbb{R}^{16\times 64}$ and decision rule $W_g \in \mathbb{R}^{16\times 8}$. A sample is $z \sim \mathcal{N}(0, I_{16})$, $x = zA_g + 0.08\,\varepsilon$, $y = \arg\max(zW_g)$; d_{in}=64, 8 classes, 800 train and 200 test samples per skill in expectation. Skill frequencies follow a Zipf law with exponent 1.0 over a random permutation of the skills. Skills are partitioned into families with sizes cycling through $(1, 1, 2, 3)$; within a family, A_g and W_g are drawn at cosine similarity 0.92 to a family prototype. The task structure is fixed by a task seed independent of the run seed, so the benchmark is a fixed object and seeds are replication units. In the locked runs G=E; the confound control of Table 9 holds G=16 while E varies, and the benchmark-instance control redraws the task seed 4 times. The balanced-load sweep for the diagnostic (Table 28) additionally varies the within-family cosine similarity in its `het` arm, alternating 0.98 (near-duplicate substitute) and 0.60 (none).

FMNIST-Rot. Fashion-MNIST images, a fixed subset of 24,000 train / 6,000 test drawn with a fixed seed, each assigned to one of E domains and rotated by $360g/E$ degrees with bilinear interpolation; inputs standardized by training-set statistics. Domain identity is never given to the model. Realized load skew is 4.49/8.22/15.04× at E=8/16/32, i.e. comparable to SSB.

Adversaries. CLEAN sets β=0. RANDDROP draws one expert uniformly per step. FAILSAFE-SAMPLED(m) draws m experts uniformly without replacement and maximises. FAILSAFE-EXACT takes C=[E]. FAILSAFE-RM draws one expert with probability proportional to its router mass accumulated over the current batch (recomputed every step, no extra forward pass); `rm_argmax` takes the argmax of the same quantity and

Table 5: **A real multi-layer MoE language model** (E29; exploratory, post-lock; 98.6 M parameters, 45.1 M tokens of WikiText-103, n=3 matched seeds). ΔNLL is the rise in validation NLL under one global expert outage, over all 6×16=96 (layer, expert) pairs on the entire validation split, so worst-1 is a true maximum. The last row is the iso-compute control (CLEAN given the outage arms' second forward/backward pass as $2\times$ steps): the best model here and the most fragile — but it also sees $2\times$ the tokens (90.1 M against 45.1 M), so it confounds steps with data. At n=3 the per-seed column, not the p-value, is the evidence.

	clean		outage damage, ΔNLL (nats)			
Training	val ppl	val NLL	worst-1 (s.d.)	worst-1 per seed	mean-1	n
CLEAN	59.84	4.0917	0.306 (0.064)	0.319 / 0.237 / 0.362	0.0246	3
RANDDROP	59.30	4.0826	0.157 (0.032)	0.167 / 0.121 / 0.183	0.0100	3
FAILSAFE-RM	59.33	4.0829	0.077 (0.018)	0.060 / 0.095 / 0.076	0.0088	3
CLEAN ($2\times$ steps)	45.38	3.8152	0.444 (0.033)	0.453 / 0.407 / 0.472	0.0313	3

`rm_top4_max` maximises over its four largest entries. RANDDROP-MEAN(m) draws the same outages as FAILSAFE-SAMPLED(m) from the same RNG stream and averages instead of maximising.

Outage evaluation. For q=1 we evaluate all E single outages exhaustively; for q=2 all $\binom{E}{2}$ pairs. All expert outputs are computed once per batch and each outage re-selects and re-mixes from them, so exhaustive evaluation is affordable and the reported worst case is a true minimum over the whole outage set, not an estimate. The default protocol is reroute-on-failure; Table 23 reports the two zeroing protocols as well.

Diagnostic. EDS(e) is accumulated over the test set in one pass: exposure is the realized gate weight, the substitute is obtained by re-running top-k with e masked, and sensitivity is $\|\partial \ell / \partial y\|$ at the layer output. Router mass is the summed gate weight. Both are ranked against measured per-expert damage by Spearman correlation, with damage measured both as accuracy drop and as the rise in test cross-entropy.

Real-MoE audit (§7). `allenai/OLMoE-1B-7B-0924`, bf16, inference only, no fine-tuning. 24,552 scored next-token positions from WikiText-103 validation documents (CC BY-SA 3.0) listed in `items.json` for damage, and a *disjoint* split of 16,384 tokens for the diagnostics, so no predictor is fitted on the tokens it is scored against. Each of the 1024 outages is a separate full evaluation paired to the clean model position by position, and every row carries its own standard error. Correctness checks: an empty mask is bit-identical to the stock model; a masked expert receives zero tokens while top-8 slots stay full with distinct survivors; replaying from a cached residual stream reproduces a full forward exactly; and the analytic displacement used by the diagnostics matches the patched block at cosine 0.9923. Because bf16 GEMM tiling changes when rerouting changes an expert's row count, tokens that never touched the failed expert still move at the rounding level; that floor is $141\times$ below the signal on routed tokens, is zero-mean across tokens, and averages down over 24,552 paired positions.

Statistics. Items are averaged within a (condition, seed) cell first, so the unit of replication is the seed. Reported p-values are paired t-tests over seeds; 95% confidence intervals are percentile bootstrap intervals over the paired per-seed differences with 10,000 resamples; Cohen's d is computed on the same per-seed differences. Wilcoxon signed-rank tests accompany the pre-registered contrasts (Table 2).

Compute. All controlled experiments are CPU-only inside a pinned Docker image (tag and digest in the supplementary material), single-threaded per run so that float reduction order is identical across compared cells, with models of ≈0.3–1 M parameters: 2,696 training runs and about 71 CPU-hours in total, of which the post-lock revision block is 36 result directories and 493 archived contrasts. Per-step wall-clock costs are in Table 27 and expert-FLOP accounting in Table 26. Two studies need a GPU and used one 24 GB card in Docker for 279 minutes in total: the real-MoE audit (52.4 minutes) and the language model of Appendix F (226 minutes for 12 training runs and their exhaustive outage evaluations).

Table 6: E29 setup. The whole study is exploratory and post-lock: it carries no pre-registered threshold and cannot change any verdict in Table 2. All 7 correctness checks pass (7 of 7; `experiments/E29/sanity_fresh.json`); the 60 paired contrasts are archived in `analysis/raw/e29__*.json` and plotted in Figure 4(b).

field	value
architecture	decoder-only transformer, 6 layers, d_{model}=384, every FFN a token-choice top-2 MoE with E=16 experts (expert hidden 1024), no shared expert, tied input/output embedding
parameters	98.6 M total, 32.4 M active per token
routing	softmax over all E logits → top-2 → gates renormalised over the selected set; Switch load-balancing auxiliary loss α_{aux}=0.01
data	WikiText-103-raw-v1, GPT-2 tokenizer, sequence length 256
tokens seen	45.1 M per 1× run (5,500 steps × 32 sequences × 256 tokens)
optimiser	AdamW, lr 0.001, cosine schedule, bf16 autocast, β=1
evaluation	the entire official validation split as a fixed slice: 965 contiguous non-overlapping 256-token chunks, 247,040 scored tokens
outage model	one *global* single-expert outage = a (layer, expert) pair, 96 candidates; reroute-on-failure (logit → $-\infty$ *before* top-k, top-k re-selected among survivors, gates renormalised)
unigram floor	validation NLL 7.432 nats = ppl 1689 (train-fitted unigram on the same slice)
seeds	0, 1, 2; at a given seed every condition shares initialisation, data order and schedule, so all contrasts are paired
conditions	CLEAN (β=0); RANDDROP (the (layer, expert) pair uniform over all 96 candidates); FAILSAFE-RM (layer uniform, expert $\propto$ its router mass on the current batch); CLEAN at 2× steps (iso-compute)
not run	a sampled-max (m=4) arm was specified and *not* trained — the budget went to the third seed; an exact adversary over 96 candidates per step is infeasible here
compute	226 GPU-minutes on one 24 GB card, in Docker

F THE MOE LANGUAGE MODEL (E29)

G WHAT EXPOSURE WEIGHTING DOES: EXCLUDING DEAD OUTAGES, AND TILTING

FAILSAFE-RM gives the *busiest* expert at most 8.28× the $1/E$ of a uniform draw, yet matches an adversary of weight 1 on the true worst outage, while a sampled-max adversary of weight $m/E = 8/32$ on the true worst does not: in the same cell it reaches 61.08 % worst-1 accuracy against 62.10 % for FAILSAFE-RM and 62.10 % for the exact max (RANDDROP: 55.04 %). What the sampler does to the *whole* outage distribution therefore matters more than its weight on rank 1.

FAILSAFE-RM changes that distribution in two ways at once: it never spends its draw on an expert with no traffic on the current batch — whose outage is an exact no-op — and, among the rest, it tilts towards the busy ones. RANDDROP-LIVE (E30) turns on the first without the second: a uniform draw over the live pool only, at exactly RANDDROP cost. In the collapsed regime, where only 49.0 % of the pool (15.7 of 32 experts) is live per 256-token batch, it gains +2.76 pt over RANDDROP ($p = 2.9 \times 10^{-5}$, CI [+2.10, +3.43] pt, $n = 10$) — 39 % of FAILSAFE-RM's +7.06 pt — and stays -4.30 pt behind it ($p = 3.3 \times 10^{-5}$). Where 90.0 % of the pool is live (0.01), the live restriction buys nothing (-0.14 pt, $p = 0.857$) while FAILSAFE-RM still gains +3.57 pt, so there *none* of its advantage is no-op avoidance. RANDDROP-LIVE is TOST-*not*-equivalent to FAILSAFE-RM within ±1.0 pt in all 4 cells and costs nothing on clean accuracy anywhere ($|\Delta| \leq 0.59$ pt; Table 12). The isolation is imperfect and the recovered fraction has no CI (Table 12); E30 is exploratory and post-lock.

H IS IT JUST AN IMPLICIT LOAD BALANCER?

This is the sharpest objection, because outage training does reduce skew, so we answer it with a frontier over 8 values of α_{aux} rather than one contrast (Figure 9): 8 of 8 FAILSAFE-EXACT and 8 of 8 FAILSAFE-RM cells are not dominated by *any* CLEAN setting on both axes at once, and matching a baseline to FAILSAFE-EXACT@0.001 on realized skew leaves CLEAN@0.002 -11.28 pt ($p = 6.2 \times 10^{-6}$) and RANDDROP@0.005 -5.71 pt ($p = 1.5 \times 10^{-8}$) behind over matched seeds. Against us: the frontier points are cell means at

Table 7: E29 per-seed values — the honest unit of replication at n=3. Last column is the mean over seeds with the across-seed standard deviation in parentheses.

training	seed 0	seed 1	seed 2	mean (s.d.)
worst-1 ΔNLL (nats)				
CLEAN	0.3185	0.2365	0.3624	0.3058 (0.0639)
RANDDROP	0.1667	0.1206	0.1826	0.1566 (0.0322)
FAILSAFE-RM	0.0601	0.0952	0.0759	0.0771 (0.0176)
CLEAN (2× steps)	0.4527	0.4071	0.4723	0.4440 (0.0335)
mean-1 ΔNLL (nats)				
CLEAN	0.0251	0.0242	0.0244	0.0246 (0.0005)
RANDDROP	0.0100	0.0101	0.0099	0.0100 (0.0001)
FAILSAFE-RM	0.0086	0.0086	0.0092	0.0088 (0.0003)
CLEAN (2× steps)	0.0291	0.0326	0.0321	0.0313 (0.0019)
clean val NLL (nats)				
CLEAN	4.0812	4.0959	4.0980	4.0917 (0.0092)
RANDDROP	4.0713	4.0903	4.0863	4.0826 (0.0100)
FAILSAFE-RM	4.0662	4.1085	4.0739	4.0829 (0.0225)
CLEAN (2× steps)	3.8105	3.8184	3.8166	3.8152 (0.0042)
clean val perplexity				
CLEAN	59.21	60.09	60.22	59.84 (0.55)
RANDDROP	58.63	59.76	59.52	59.30 (0.59)
FAILSAFE-RM	58.33	60.86	58.79	59.33 (1.34)
CLEAN (2× steps)	45.17	45.53	45.45	45.38 (0.19)
realised load skew (max/mean router mass, ×)				
CLEAN	1.91	1.95	2.14	2.00 (0.13)
RANDDROP	2.80	2.35	2.84	2.66 (0.27)
FAILSAFE-RM	2.62	2.61	2.53	2.59 (0.05)
CLEAN (2× steps)	2.00	2.08	2.07	2.05 (0.04)
*effective experts (*exp *entropy, of E=16)*				
CLEAN	14.44	14.50	14.40	14.45 (0.05)
RANDDROP	13.16	13.45	13.01	13.21 (0.22)
FAILSAFE-RM	13.19	13.08	13.22	13.16 (0.07)
CLEAN (2× steps)	14.49	14.37	14.30	14.39 (0.10)
tail ratio (worst-1 / mean-1)				
CLEAN	12.7	9.8	14.9	12.4 (2.6)
RANDDROP	16.7	12.0	18.4	15.7 (3.3)
FAILSAFE-RM	7.0	11.0	8.3	8.8 (2.1)
CLEAN (2× steps)	15.6	12.5	14.7	14.3 (1.6)

different α_{aux}, so domination is *not* a seed-paired test, and stronger load balancing alone already buys the exact adversary's worst-1 accuracy at -1.13 pt of clean accuracy ($p = 0.001$).

I ADDITIONAL TABLES AND FIGURES

Table 8: E29 by layer (mean over 3 seeds). The damage is a layer-0 phenomenon in every condition, and FAILSAFE-RM's entire advantage is a layer-0 advantage; the worst outage of all 96 lies in layer 0 in 12 of the 12 trained models. Router mass stays a usable within-layer damage predictor at this scale, as it is on OLMoE (§7).

training	L0	L1	L2	L3	L4	L5
worst-1 ΔNLL within the layer (nats)						
CLEAN	0.306	0.011	0.007	0.033	0.031	0.038
RANDDROP	0.157	0.006	0.008	0.021	0.023	0.041
FAILSAFE-RM	0.077	0.006	0.008	0.017	0.020	0.029
CLEAN (2× steps)	0.444	0.011	0.009	0.046	0.039	0.049
realised load skew (×)						
CLEAN	2.80	1.62	1.61	1.61	1.77	2.57
RANDDROP	3.25	1.94	1.87	2.18	2.63	4.11
FAILSAFE-RM	3.44	2.01	1.70	1.99	2.12	4.27
CLEAN (2× steps)	2.83	1.50	1.72	1.65	1.90	2.70
effective experts						
CLEAN	13.56	15.14	15.25	15.18	14.81	12.73
RANDDROP	12.32	15.07	14.80	14.54	13.84	8.66
FAILSAFE-RM	11.89	14.57	14.99	14.49	14.31	8.74
CLEAN (2× steps)	13.70	15.28	15.31	15.14	14.69	12.21
Spearman ρ(router mass, damage)						
CLEAN	0.68	0.61	0.61	0.73	0.71	0.82
RANDDROP	0.71	0.42	0.67	0.70	0.83	0.91
FAILSAFE-RM	0.46	0.54	0.67	0.75	0.76	0.91
CLEAN (2× steps)	0.67	0.66	0.43	0.64	0.66	0.88

Table 9: **No control removes the effect, and two structural alternatives are real competition at this scale.** Worst-1 accuracy in pt, E=32 on SSB at the locked α_{aux} unless stated, matched seeds; 95% CIs in beside each Δ (90% for the TOST row). The upper block holds FLOPs, steps, selection split and task fixed in turn — **bold** is the largest and underline the second-largest Δ there; the lower block prices the alternatives, where the rows are not commensurable and are left unranked. For *Exact-Live* the p column is a TOST equivalence p; elsewhere a paired t-test. *RandDrop-Live* is the mechanism-isolation control of §6.4, differing from FAILSAFE-RM only in the exposure-proportional tilt. All rows exploratory (post-lock); detail in Tables 14, 17, 21, 22 and 12.

control	comparison	Δ (pt) ↑	p	n
controls on the headline contrast (positive favours the adaptive adversary)				
iso-FLOP, m=4	max − mean over the same 4 outages	+3.99 [+3.24, +4.73]	4.2×10^{-6}	10
iso-FLOP, m=E	Exact − mean over all E outages	+5.41 [+4.50, +6.30]	1.4×10^{-6}	10
4× steps	Exact (1000) − RandDrop (4000)	+3.86 [+3.14, +4.56]	3.0×10^{-6}	10
held-out selection	Exact − RandDrop, worst outage chosen on the other half	**+7.15** [+6.42, +7.81]	1.4×10^{-8}	10
task fixed, G=16	gap(E=32) − gap(E=8)	+3.29 [+2.07, +4.41]	0.008	5
cheaper or structural alternatives to the exact adversary				
Exact-Live	live-pool max − Exact (TOST ±1.0 pt: EQUIVALENT), at 0.58× step time	+0.08 [-0.18, +0.34]	5.7×10^{-5}	10
RandDrop-Live	live-pool uniform draw − FAILSAFE-RM, at RANDDROP cost; recovers 39 % of RM's gain over RANDDROP	-4.30 [-5.31, -3.21]	3.3×10^{-5}	10
stronger load balancing	RandDrop@0.1 − Exact@0.001 (clean acc. -1.13 pt)	+0.21 [-0.19, +0.63]	0.379	10
shared always-on expert	Clean+shared − Clean	+11.80 [+10.20, +13.40]	2.2×10^{-4}	5

Table 10: **Worst-case outage damage tracks router load skew, and so does the value of training against it.** E=32 on SSB, n=10, exploratory (post-lock). Load is varied with the Switch coefficient α_{aux}; max/mean router mass is *measured* on the test set (CLEAN arm). Columns 4–7 are each condition's worst-single-outage drop from its own clean accuracy (pt; lower is better). The association is correlational (§6.1). The pretrained MoE of §7 sits at 3.22×, i.e. the second row; full grid in Table 16.

			worst-1 drop (pt)				Exact−RandDrop	
α_{aux}	max/mean load	clean acc.	Clean	RandDrop	Samp.	Exact	Δworst-1	p
0.001	11.58×	63.55	14.04	8.61	5.03	2.09	+7.06	2.8×10^{-8}
0.01	3.23×	62.77	2.17	5.16	2.55	1.32	+3.83	2.5×10^{-4}
0.1	1.20×	62.80	0.77	0.75	0.73	0.66	+0.32	0.209
1.0	1.14×	61.63	0.62	0.72	0.61	0.66	+0.37	0.039

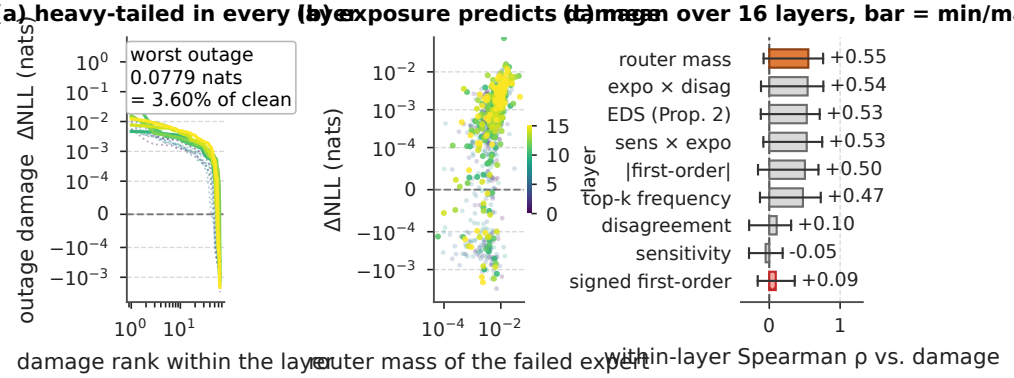


Figure 7: **On a real MoE the damage is heavy-tailed and exposure predicts it, but the absolute headroom at top-8-of-64 is small — so this audit scopes our method rather than supporting it.** All 1024 single-expert outages of `allenai/OLMoE-1B-7B-0924` measured on 24,552 held-out positions, n=1 model, inference only (exploratory, post-lock); there are no error bars because every outage is measured exhaustively rather than sampled. Colour encodes layer (viridis); opaque points are the 6 layers resolved above the noise floor, faint dotted ones the rest, and the symlog axes leave outages that *help* unclipped. **(a)** Damage against damage rank within each layer: heavy-tailed everywhere, but the worst outage still costs only 3.60 % of the clean loss. **(b)** Router mass against damage; the best of the nine cheap predictors we tried. **(c)** Within-layer Spearman ρ, bars spanning the per-layer min/max: the signed first-order estimate — the quantity Proposition 1 bounds — carries no signal, and EDS does not beat router mass.

Table 11: **The criticality diagnostic does not beat router mass in any setting we measured.** Spearman ρ against measured per-expert damage; Δ is the paired EDS − RM difference. Row 1 is the pre-registered test (120 locked models; Wilcoxon $p = 0.193$); rows 2–4 are exploratory (post-lock), the last the OLMoE audit (EDS leads in 4 of 16 layers). Thresholds were $\rho(\text{EDS}) \geq 0.75$ *and* ≥ 0.25 above router mass: both missed everywhere.

Setting	max/mean load	ρ(RM)	ρ(EDS)	Δ	p
SSB, pre-registered (E3)	11.6×	0.650	0.645	-0.005	0.279
SSB, balanced (E9B)	3.05×	0.831	0.814	-0.017	0.002
SSB, uniform (E9B)	1.18×	0.268	0.003	-0.265	9.8×10^{-6}
OLMoE-1B-7B (E11)	3.22×	+0.551	+0.529	-0.022	—

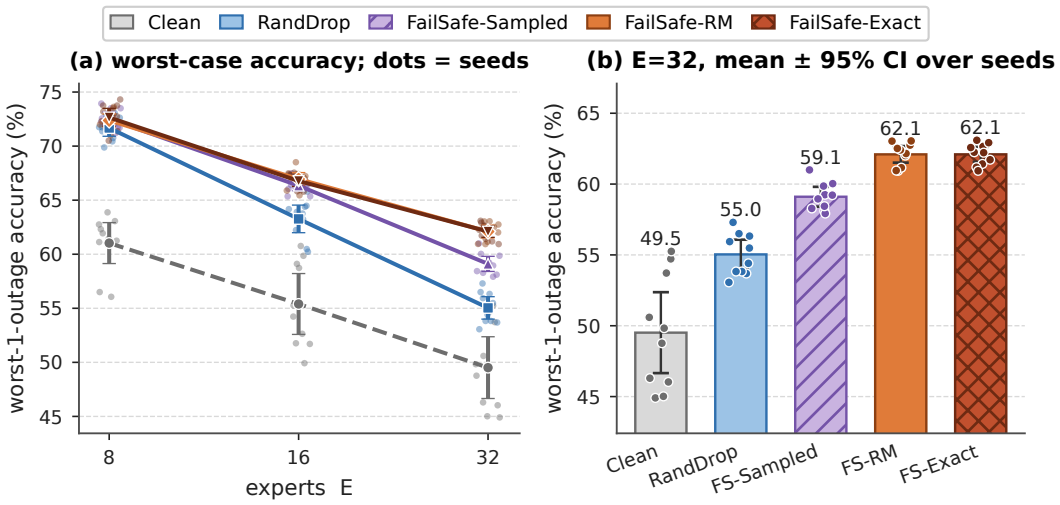


Figure 8: **The ordering of the adversaries holds seed by seed, not only on average.** SSB at the locked load skew (11.58×), 10 matched seeds; the per-seed values behind Figure 4(a). (a) Worst-single-outage accuracy against pool size: line and whisker are the mean and a 95% t CI over seeds, and each small dot is one seed (jittered horizontally). (b) The same E=32 cells as bars with every seed shown and 95% t CIs. FAILSAFE-RM (exploratory, post-lock) tracks the exact adversary at RANDDROP cost. The widening gap follows from the m/E weight of Observation 1 *together with* the measured damage profile (Figure 10), not from the observation alone.

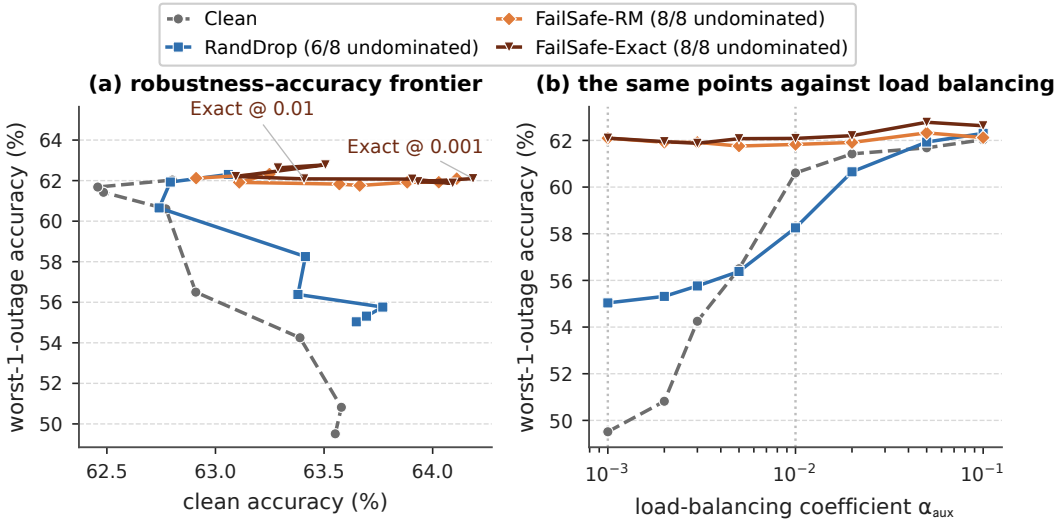


Figure 9: **Tuning the load-balancing coefficient alone moves CLEAN most of the way, but not all of it: most tail-weighted cells stay off the CLEAN frontier.** E=32 on SSB, 8 values of the Switch coefficient α_{aux} from 0.001 to 0.1, 10 matched seeds per point (exploratory, post-lock). (a) Every point is a cell mean, dominated if some CLEAN setting is at least as good on *both* axes; 8 of 8 FAILSAFE-EXACT and 8 of 8 FAILSAFE-RM points are undominated, two of them exceeding every CLEAN cell on both axes at once. (b) The same data against α_{aux}: tuning it alone moves CLEAN most of the way, which is why we report the clean-accuracy price beside every robustness gain. **Caveat:** cell means across different α_{aux}, so domination is not a seed-paired test.

Table 12: **E30, RandDrop-Live** (exploratory, post-lock): a uniform single-expert outage drawn over the *live* experts only — those carrying gate mass on the current 256-token batch — at exactly RANDDROP cost. It turns on the no-op-avoidance half of FAILSAFE-RM without the exposure-proportional tilt. Baseline rows are reused from E12/E17/E23 and reproduce bit-for-bit under the patched sampler. Δ is the paired worst-1 difference against RANDDROP (95% CI for the RANDDROP-LIVE row). Against FAILSAFE-RM, RANDDROP-LIVE is -4.30 pt at α_{aux}=0.001 (90% CI [-5.33, -3.26], TOST $p = 1.000$ against ± 1.0 pt: not equivalent) and -3.71 pt at 0.01 (not equivalent); it is not equivalent to FAILSAFE-RM in any of the 4 cells. Clean accuracy moves by at most 0.59 pt anywhere. **Two caveats on the isolation**: RANDDROP-LIVE also ends training with slightly lower realised router skew than RANDDROP (10.19× against 11.40× at 0.001), so the sampler change is not observed at fixed routing; and the "fraction of RM's gain recovered" quoted in §6.4 is a ratio of two paired point estimates, so it carries no confidence interval and is meaningless where the denominator is small.

Training	clean	worst-1	worst-1 drop	realised skew	Δ worst-1 vs. RandDrop	n
SSB E=32, α_{aux}=0.001; live pool 15.7 of 32 (49 % of the pool)						
CLEAN	63.55	49.51	14.04	11.58×	—	10
RANDDROP	63.65	55.04	8.61	11.40×	—	10
RANDDROP-LIVE	63.81	57.80	6.01	10.19×	+2.76 [+2.10, +3.43]	10
FAILSAFE-RM	64.11	62.10	2.01	8.28×	+7.06	10
FAILSAFE-EXACT	64.19	62.10	2.09	9.05×	+7.06	10
SSB E=32, α_{aux}=0.01; live pool 28.8 of 32 (90 % of the pool)						
CLEAN	62.77	60.61	2.17	3.23×	—	10
RANDDROP	63.41	58.25	5.16	6.29×	—	10
RANDDROP-LIVE	63.31	58.11	5.19	6.35×	-0.14 [-1.53, +1.25]	10
FAILSAFE-RM	63.57	61.83	1.74	4.05×	+3.57	10
FAILSAFE-EXACT	63.41	62.09	1.32	3.56×	+3.83	10
SSB E=8, α_{aux}=0.001; live pool 6.6 of 8 (83 % of the pool)						
CLEAN	71.99	61.03	10.96	3.28×	—	10
RANDDROP	73.41	71.67	1.73	3.21×	—	10
RANDDROP-LIVE	73.61	72.26	1.35	3.11×	+0.31 [-0.03, +0.68]	5
FAILSAFE-RM	73.74	72.38	1.37	2.66×	+0.70	10
FAILSAFE-EXACT	73.72	72.65	1.07	2.78×	+0.98	10
SSB E=16, α_{aux}=0.001; live pool 10.8 of 16 (68 % of the pool)						
CLEAN	67.61	55.40	12.21	5.77×	—	10
RANDDROP	68.38	63.27	5.11	5.78×	—	10
RANDDROP-LIVE	68.17	64.13	4.04	5.41×	+0.96 [+0.11, +1.97]	5
FAILSAFE-RM	68.56	66.95	1.61	4.57×	+3.68	10
FAILSAFE-EXACT	68.27	66.78	1.49	4.83×	+3.51	10

Table 13: E24 (exploratory, post-lock, 10 matched seeds): the independent skew instrument. A fixed, non-learned router logit bias $b_e = \gamma z_e$ (z a standardised log-Zipf profile over a fixed random expert permutation) is added with the training objective held fixed. **The free-router instrument fails its manipulation check** — the Switch auxiliary loss absorbs a constant logit offset, so realised skew spans only 1.4× at α_{aux}=0.1 and 1.5× at 0.01 — so the pre-registered fallback, a *frozen* router plus the same bias, is used. Freezing is itself an intervention and raises the γ=0 baseline skew, so the frozen series must be read against its own γ=0 cell and never pooled with the free-router rows.

cell	realised skew	eff. experts	Clean worst-1 drop (pt)	Exact−RandDrop	RM−RandDrop	Exact−Clean
frozen router (manipulation check passes, 2.56× →31.76× in the pilot)						
`E24_a0.1_g0_frz`	2.58×	25.6	0.90	+1.06	+0.53	+1.09
`E24_a0.1_g0.269_frz`	3.40×	19.9	1.29	+1.21	+0.72	+1.15
`E24_a0.1_g1.79_frz`	6.09×	7.7	2.60	+2.19	+2.12	+4.52
`E24_a0.1_g3.048_frz`	10.48×	3.9	5.41	+5.93	+5.48	+6.30
free router (manipulation check fails; reported for completeness)						
`E24_a0.01_g0`	3.23×	23.9	2.17	+3.83	+3.57	+1.48
`E24_a0.01_g4`	4.05×	19.1	3.00	+1.82	+1.19	+2.93
`E24_a0.1_g0`	1.20×	31.8	0.77	+0.32	-0.18	+0.60

Table 14: E25 (exploratory, post-lock): the live/dead structure of the exact adversary's inner max, probed during training on 256-token batches, and the EXACT-LIVE restriction that ranges only over experts with non-zero gate traffic. A zero-traffic expert is in no token's top-k, so its outage loss equals the clean loss and it can never attain the max; restricting to the live pool is a no-op in exact arithmetic but not bit-exact in float32, because the candidate tensor tiles differently. "eff. experts" is $\exp(\text{entropy})$ of the test-set router-mass distribution; the live-pool column is the quantity that actually bounds the search.

α_{aux}	adversary	cand./step	live pool	frac. no-op	frac. steps argmax dead	median top-2 gap	eff. experts	n
0.001	FAILSAFE-EXACT	32.0	11.7	0.634	0.000	0.0158	6.3	10
0.001	EXACT-LIVE	11.9	11.9	0.000	0.000	0.0159	6.4	10
	Exact-Live − Exact (worst-1)	+0.08	[-0.18, +0.34]	5.7×10^{-5}	TOST ± 1.0 pt: equivalent		0.58×	10
0.01	FAILSAFE-EXACT	32.0	28.6	0.105	0.000	0.0121	20.0	10
0.01	EXACT-LIVE	28.7	28.7	0.000	0.000	0.0122	19.6	10
	Exact-Live − Exact (worst-1)	+0.02	[-0.20, +0.25]	1.2×10^{-5}	TOST ± 1.0 pt: equivalent		1.24×	10

Table 15: E26F (exploratory, post-lock and *post hoc*, 5 matched seeds): the second benchmark for FAILSAFE-RM. Rotated Fashion-MNIST, same driver, config and seeds as the locked FMNIST-Rot cells. FAILSAFE-RM is TOST-equivalent to FAILSAFE-EXACT within ± 1.0 pt at all three pool sizes while its advantage over RANDDROP grows with the pool, which is the same pattern SSB shows.

E	training	clean	mean-1	worst-1	n
8	CLEAN	80.40	75.34	53.55	5
8	RANDDROP	80.31	79.99	79.17	5
8	FAILSAFE-RM	80.45	80.24	79.79	5
8	FAILSAFE-SAMP.	80.69	80.46	79.98	5
8	FAILSAFE-EXACT	80.43	80.26	80.02	5
	FAILSAFE-RM − RANDDROP	+0.62	[+0.04, +1.23]	0.145	5
	FAILSAFE-RM − FAILSAFE-EXACT	-0.23	[-0.37, -0.08]	0.058	5
16	CLEAN	75.64	73.60	55.28	5
16	RANDDROP	75.60	75.29	72.89	5
16	FAILSAFE-RM	75.51	75.42	74.88	5
16	FAILSAFE-SAMP.	75.66	75.52	74.88	5
16	FAILSAFE-EXACT	75.78	75.68	75.23	5
	FAILSAFE-RM − RANDDROP	+1.99	[+1.43, +2.44]	0.002	5
	FAILSAFE-RM − FAILSAFE-EXACT	-0.35	[-0.75, -0.06]	0.153	5
32	CLEAN	74.58	73.52	54.53	5
32	RANDDROP	74.33	74.08	69.27	5
32	FAILSAFE-RM	74.84	74.74	74.03	5
32	FAILSAFE-SAMP.	74.55	74.42	73.05	5
32	FAILSAFE-EXACT	74.74	74.68	74.11	5
	FAILSAFE-RM − RANDDROP	+4.76	[+3.77, +5.79]	0.001	5
	FAILSAFE-RM − FAILSAFE-EXACT	-0.07	[-0.53, +0.39]	0.800	5

Table 16: E12 (exploratory, post-lock): the full load-regime grid. Load is set by the Switch auxiliary-loss coefficient α_{aux}; all routing statistics are measured on the test set. At α_{aux}=0.001 the runs reproduce the locked E1 runs bit-for-bit (sanity check 3 of 5). Accuracies in %, drops in pt.

E	α_{aux}	training	max/mean	entropy	eff. experts	clean	mean-1	worst-1	drop	n
8	0.001	CLEAN	3.28	0.735	4.6	71.99	69.86	61.03	10.96	10
8	0.001	RANDDROP	3.21	0.625	3.7	73.41	72.98	71.67	1.73	10
8	0.001	FAILSAFE-SAMP.	2.96	0.658	4.0	73.50	73.24	72.51	0.99	10
8	0.001	FAILSAFE-EXACT	2.78	0.683	4.2	73.72	73.41	72.65	1.07	10
8	0.01	CLEAN	1.41	0.983	7.7	72.17	70.90	69.46	2.71	5
8	0.01	RANDDROP	1.91	0.932	7.0	73.66	73.16	72.27	1.39	5
8	0.01	FAILSAFE-SAMP.	1.62	0.957	7.3	74.38	73.87	73.10	1.27	5
8	0.01	FAILSAFE-EXACT	1.49	0.967	7.5	73.79	73.40	72.70	1.09	5
8	0.1	CLEAN	1.12	0.998	8.0	72.36	71.27	70.25	2.11	5
8	0.1	RANDDROP	1.13	0.999	8.0	73.56	73.09	72.39	1.17	5
8	0.1	FAILSAFE-SAMP.	1.14	0.998	8.0	73.50	73.01	72.55	0.95	5
8	0.1	FAILSAFE-EXACT	1.14	0.999	8.0	73.21	72.85	72.15	1.06	5
8	1.0	CLEAN	1.11	0.999	8.0	72.15	71.17	70.35	1.80	5
8	1.0	RANDDROP	1.12	0.999	8.0	73.41	73.08	72.62	0.79	5
8	1.0	FAILSAFE-SAMP.	1.08	0.999	8.0	73.80	73.31	72.85	0.95	5
8	1.0	FAILSAFE-EXACT	1.10	0.999	8.0	73.61	73.12	72.60	1.01	5
16	0.001	CLEAN	5.77	0.630	5.8	67.61	66.17	55.40	12.21	10
16	0.001	RANDDROP	5.78	0.615	5.5	68.38	67.68	63.27	5.11	10
16	0.001	FAILSAFE-SAMP.	5.14	0.597	5.3	68.85	68.39	66.34	2.51	10
16	0.001	FAILSAFE-EXACT	4.83	0.627	5.8	68.27	67.97	66.78	1.49	10
16	0.01	CLEAN	2.33	0.937	13.5	66.90	66.13	64.27	2.63	5
16	0.01	RANDDROP	3.16	0.860	10.9	67.56	66.93	64.54	3.02	5
16	0.01	FAILSAFE-SAMP.	2.74	0.898	12.1	67.69	67.26	65.90	1.79	5
16	0.01	FAILSAFE-EXACT	2.29	0.907	12.4	68.05	67.60	66.75	1.30	5
16	0.1	CLEAN	1.12	0.999	16.0	66.39	65.77	64.83	1.56	5
16	0.1	RANDDROP	1.19	0.999	15.9	67.34	66.86	66.15	1.19	5
16	0.1	FAILSAFE-SAMP.	1.16	0.999	15.9	67.75	67.26	66.64	1.11	5
16	0.1	FAILSAFE-EXACT	1.12	0.999	16.0	67.42	66.95	66.29	1.13	5
16	1.0	CLEAN	1.09	0.999	16.0	65.88	65.40	64.89	0.99	5
16	1.0	RANDDROP	1.14	0.999	16.0	66.67	66.27	65.62	1.04	5
16	1.0	FAILSAFE-SAMP.	1.11	0.999	16.0	66.84	66.44	65.81	1.02	5
16	1.0	FAILSAFE-EXACT	1.10	0.999	16.0	66.97	66.59	66.10	0.87	5
32	0.001	CLEAN	11.58	0.585	7.6	63.55	62.87	49.51	14.04	10
32	0.001	RANDDROP	11.40	0.509	5.9	63.65	63.13	55.04	8.61	10
32	0.001	FAILSAFE-SAMP.	11.13	0.497	5.7	64.14	63.79	59.11	5.03	10
32	0.001	FAILSAFE-EXACT	9.05	0.530	6.3	64.19	63.96	62.10	2.09	10
32	0.01	CLEAN	3.23	0.916	23.9	62.77	62.38	60.61	2.17	10
32	0.01	RANDDROP	6.29	0.794	15.7	63.41	62.96	58.25	5.16	10
32	0.01	FAILSAFE-SAMP.	5.07	0.830	17.8	63.28	62.95	60.72	2.55	10
32	0.01	FAILSAFE-EXACT	3.56	0.864	20.0	63.41	63.12	62.09	1.32	10
32	0.1	CLEAN	1.20	0.999	31.8	62.80	62.51	62.03	0.77	10
32	0.1	RANDDROP	1.35	0.997	31.6	63.06	62.77	62.30	0.75	10
32	0.1	FAILSAFE-SAMP.	1.29	0.998	31.7	62.73	62.48	62.01	0.73	10
32	0.1	FAILSAFE-EXACT	1.26	0.998	31.8	63.29	63.04	62.63	0.66	10
32	1.0	CLEAN	1.14	0.999	31.9	61.63	61.39	61.01	0.62	10
32	1.0	RANDDROP	1.14	0.999	31.9	62.43	62.19	61.71	0.72	10
32	1.0	FAILSAFE-SAMP.	1.13	0.999	31.9	62.35	62.12	61.74	0.61	10
32	1.0	FAILSAFE-EXACT	1.15	0.999	31.9	62.73	62.48	62.07	0.66	10

Table 17: R14 (exploratory, post-lock): load balancing as a *free* robustness baseline. Each row is a paired difference, at E=32 over 10 matched seeds, between an arm trained at a higher α_{aux} and FAILSAFE-EXACT at the locked α_{aux}=0.001. `randdrop @ aux0.1` reaches the same worst-1 accuracy as the exact adversary; the price is clean accuracy.

arm − FAILSAFE-EXACT@0.001	Δworst-1	p	CI	Δclean	p
`clean @ aux0.01`	-1.49	3.5×10^{-5}	[-1.87, -1.14]	-1.41	5.3×10^{-4}
`randdrop @ aux0.01`	-3.85	5.2×10^{-5}	[-4.79, -2.84]	-0.77	6.7×10^{-4}
`clean @ aux0.1`	-0.07	0.755	[-0.45, +0.36]	-1.38	3.1×10^{-5}
`randdrop @ aux0.1`	+0.21	0.379	[-0.19, +0.63]	-1.13	0.001
`clean @ aux1.0`	-1.09	0.005	[-1.63, -0.54]	-2.56	1.1×10^{-5}
`randdrop @ aux1.0`	-0.39	0.083	[-0.75, -0.02]	-1.76	1.2×10^{-5}

Table 18: E17 (exploratory, post-lock, 10 matched seeds): the exposure-weighted adversary family. `rm_argmax` drops the highest-router-mass expert of the current batch, FAILSAFE-RM (`rm_prop`) samples the dropped expert $\propto$ batch router mass — both at exactly RANDDROP cost, one extra re-mix per step, no search over the pool — and `rm_top4_max` maximises over the four highest-router-mass experts at FAILSAFE-SAMPLED cost. The last three columns are paired differences in worst-1 accuracy against each baseline. FAILSAFE-RM is statistically indistinguishable from exhaustive search at all three pool sizes, which is why the paper does not claim that *searching* is what buys the robustness.

					Δworst-1 vs.		
E	training	clean	mean-1	worst-1	RandDrop	Sampled	Exact
8	CLEAN	71.99	69.86	61.03	—	—	—
8	RANDDROP	73.41	72.98	71.67	—	—	—
8	RM-ARGMAX	73.05	72.62	71.62	-0.05 (0.878)	-0.89 (0.010)	-1.03 (0.007)
8	FAILSAFE-RM	73.74	73.40	72.38	+0.70 (0.010)	-0.14 (0.638)	-0.28 (0.296)
8	RM-TOP4-MAX	73.42	73.06	71.90	+0.22 (0.506)	-0.61 (0.043)	-0.75 (0.013)
8	FAILSAFE-SAMP.	73.50	73.24	72.51	—	—	—
8	FAILSAFE-EXACT	73.72	73.41	72.65	—	—	—
16	CLEAN	67.61	66.17	55.40	—	—	—
16	RANDDROP	68.38	67.68	63.27	—	—	—
16	RM-ARGMAX	68.66	68.19	65.88	+2.61 (0.002)	-0.46 (0.210)	-0.90 (0.060)
16	FAILSAFE-RM	68.56	68.20	66.95	+3.68 (1.7×10^{-4})	+0.61 (9.2×10^{-5})	+0.17 (0.516)
16	RM-TOP4-MAX	68.56	68.22	66.57	+3.30 (7.1×10^{-4})	+0.23 (0.358)	-0.21 (0.505)
16	FAILSAFE-SAMP.	68.85	68.39	66.34	—	—	—
16	FAILSAFE-EXACT	68.27	67.97	66.78	—	—	—
32	CLEAN	63.55	62.87	49.51	—	—	—
32	RANDDROP	63.65	63.13	55.04	—	—	—
32	RM-ARGMAX	64.13	63.87	61.47	+6.43 (1.1×10^{-7})	+2.36 (2.0×10^{-4})	-0.63 (0.083)
32	FAILSAFE-RM	64.11	63.86	62.10	+7.06 (2.4×10^{-7})	+2.99 (5.7×10^{-5})	-0.00 (0.993)
32	RM-TOP4-MAX	64.14	63.91	62.01	+6.97 (1.5×10^{-7})	+2.90 (5.0×10^{-6})	-0.09 (0.772)
32	FAILSAFE-SAMP.	64.14	63.79	59.11	—	—	—
32	FAILSAFE-EXACT	64.19	63.96	62.10	—	—	—

Table 19: R7 (exploratory, post-lock): per-condition routing and specialization statistics on the full SSB test set, 10 matched seeds, locked α_{aux}. "zero-traffic" is the fraction of experts receiving no test token, reported for *every* condition and not only for the baseline. NMI is the skill→expert mutual information, normalised; "pair cos." is the mean pairwise cosine similarity between expert output vectors (higher = more interchangeable).

E	training	entropy	eff. experts	active	zero-traffic	max/mean	Gini	purity	NMI	pair cos.	pair disag.
8	CLEAN	0.735	4.6	8.0	0.000	3.28	0.553	0.398	0.066	0.252	1.306
8	RANDDROP	0.625	3.7	8.0	0.000	3.21	0.624	0.379	0.040	0.289	1.283
8	FAILSAFE-SAMP.	0.658	4.0	7.3	0.087	2.96	0.588	0.388	0.045	0.330	1.210
8	FAILSAFE-EXACT	0.683	4.2	7.2	0.100	2.78	0.573	0.386	0.048	0.341	1.198
16	CLEAN	0.630	5.8	16.0	0.000	5.77	0.719	0.356	0.130	0.094	1.449
16	RANDDROP	0.615	5.5	15.5	0.031	5.78	0.731	0.354	0.118	0.106	1.455
16	FAILSAFE-SAMP.	0.597	5.3	15.5	0.031	5.14	0.738	0.335	0.090	0.118	1.434
16	FAILSAFE-EXACT	0.627	5.8	15.0	0.062	4.83	0.716	0.340	0.093	0.119	1.431
32	CLEAN	0.585	7.6	31.8	0.006	11.58	0.814	0.307	0.163	0.046	1.486
32	RANDDROP	0.509	5.9	30.1	0.059	11.40	0.856	0.306	0.149	0.047	1.487
32	FAILSAFE-SAMP.	0.497	5.7	30.9	0.034	11.13	0.861	0.288	0.120	0.056	1.487
32	FAILSAFE-EXACT	0.530	6.3	28.8	0.100	9.05	0.844	0.302	0.103	0.057	1.484
paired differences against CLEAN *at* E=32											
Δ	RANDDROP	-0.076	-1.7	-1.7	+0.053	-0.17	+0.042	-0.001	-0.015	+0.001	+0.000
Δ	FAILSAFE-SAMP.	-0.088	-1.9	-0.9	+0.028	-0.44	+0.046	-0.019	-0.043	+0.010	+0.000
Δ	FAILSAFE-EXACT	-0.055	-1.3	-3.0	+0.094	-2.52	+0.029	-0.005	-0.060	+0.011	-0.003

Table 20: E15 (exploratory, post-lock): two-expert outages at E=32 are close to additive on SSB, so the transfer reported under H4 is mostly single-outage hardening propagated by additivity. Left: pair-damage structure. Right: decomposition of the worst-2 gain over RANDDROP into the part predicted by additivity from the worst-1 gain and a super-additive residual. Additivity *over*-predicts, so there is no evidence of an extra generalisation effect.

training	worst-2 / 2 worst singles	mean pair ratio	frac. super-additive	worst pair = 2 worst singles
CLEAN	1.155	1.000	0.348	0.90
RANDDROP	1.254	1.006	0.278	0.60
FAILSAFE-SAMP.	1.589	1.005	0.248	0.60
FAILSAFE-EXACT	2.264	1.016	0.245	0.30

treatment	Δworst-1	Δworst-2 measured	additivity prediction	explained	residual
FAILSAFE-EXACT	+7.06	+8.06	+9.54	118%	-1.49
FAILSAFE-SAMP.	+4.07	+3.69	+5.18	141%	-1.50

Table 21: E19 / R11 (exploratory, post-lock, 5 matched seeds): a DeepSeekMoE-style shared always-on expert (`_s1`) at E=32, against the locked architecture (`_s0`). The shared expert is never subject to outage, so this is the structural alternative to training for outages. We report the numbers and draw no conclusion beyond them: n=5 seeds per arm.

arm	clean	mean-1	worst-1	worst-2	n
`clean_s0`	63.55	62.87	49.51	43.53	10
`clean_s1`	64.26	63.96	59.33	56.34	5
`randdrop_s0`	63.65	63.13	55.04	47.98	10
`randdrop_s1`	65.10	64.80	60.32	57.16	5
`fs_exact_s0`	64.19	63.96	62.10	56.04	10
`fs_exact_s1`	64.89	64.74	63.60	62.24	5
contrast		Δworst-1	Δclean	Δworst-2	
`clean_s1 − clean_s0`		+11.80 (2.2×10^{-4})	+0.53 (0.208)	+15.00 (0.001)	
`clean_s1 − fs_exact_s0`		-2.53 (0.009)	+0.18 (0.154)	+0.87 (0.559)	
`randdrop_s1 − randdrop_s0`		+5.80 (4.6×10^{-4})	+1.35 (0.007)	+8.75 (0.004)	
`randdrop_s1 − fs_exact_s0`		-1.54 (0.024)	+1.03 (0.021)	+1.69 (0.245)	
`fs_exact_s1 − fs_exact_s0`		+1.74 (0.019)	+0.81 (0.064)	+6.77 (0.004)	

Table 22: E16 (exploratory, post-lock): the top-k plane at E=32. Lowering k does not buy robustness, it buys a different (worse, less specialised) model: at k=1 clean accuracy is -8.91 pt against k=2 with 31.9 effective experts against 7.6 and skill→expert NMI 0.060 against 0.163. The advantage of the adaptive adversary is concentrated at small k and is absent at k=8, the setting of the pretrained model we audit.

k	training	clean	worst-1	worst-1 drop	active	eff. experts	pair cos.	NMI	n
1	CLEAN	54.82	54.48	0.33	32.0	31.9	0.317	0.060	5
1	RANDDROP	55.18	54.87	0.32	32.0	31.9	0.318	0.061	5
1	FAILSAFE-EXACT	55.15	54.77	0.38	32.0	31.9	0.321	0.061	5
	Exact − RandDrop		-0.10	(0.799)					
2	CLEAN	63.55	49.51	14.04	31.8	7.6	0.046	0.163	10
2	RANDDROP	63.65	55.04	8.61	30.1	5.9	0.047	0.149	10
2	FAILSAFE-EXACT	64.19	62.10	2.09	28.8	6.3	0.057	0.103	10
	Exact − RandDrop		+7.06	(2.8×10^{-8})					
4	CLEAN	65.04	61.78	3.26	32.0	16.8	0.055	0.169	5
4	RANDDROP	65.24	62.87	2.37	31.8	13.6	0.048	0.158	5
4	FAILSAFE-EXACT	65.57	64.26	1.31	32.0	15.8	0.059	0.158	5
	Exact − RandDrop		+1.39	(0.021)					
8	CLEAN	66.54	65.06	1.48	32.0	27.1	0.062	0.190	5
8	RANDDROP	67.29	66.01	1.28	32.0	25.2	0.056	0.193	5
8	FAILSAFE-EXACT	66.77	65.88	0.88	32.0	26.4	0.063	0.168	5
	Exact − RandDrop		-0.13	(0.396)					

Table 23: E21 / R10 (exploratory, post-lock, 10 matched seeds, E=32): the outage protocol. `reroute` re-runs top-k over the survivors (the protocol used throughout the paper); `zero_drop` zeroes the failed expert's slot; `zero_renorm` zeroes it and renormalises the surviving selected gates. Both non-rerouting protocols are *worse* for every condition, so reroute-on-failure is indeed the protocol most charitable to a model that has never seen a failure.

training	reroute	zero_drop	zero_renorm	Δ(zero_drop)	Δ(zero_renorm)
CLEAN	49.51	47.05	47.07	-2.46 (6.2×10^{-6})	-2.44 (3.4×10^{-6})
RANDDROP	55.04	52.40	52.45	-2.63 (8.9×10^{-6})	-2.59 (6.8×10^{-6})
FAILSAFE-SAMP.	59.11	56.82	56.83	-2.28 (2.0×10^{-7})	-2.28 (2.1×10^{-7})
FAILSAFE-EXACT	62.10	60.71	60.68	-1.39 (4.8×10^{-6})	-1.42 (4.9×10^{-6})

Table 24: R8 (exploratory, post-lock): damage at rank 1 and at the median rank $\lceil (E{+}1)/2 \rceil$, which is the rank a uniformly drawn expert attains in expectation under the tie-free model of Lemma 1. "rank-1 share" is the worst outage's share of the total damage summed over all E single outages: it does *not* grow with E, so the growth of the worst/mean "tail ratio" is mostly $1/E$ dilution of the mean.

E	training	median rank	damage@1 (pt)	damage@median (pt)	gap (pt)	worst/mean	rank-1 share	n
8	CLEAN	5	10.96	0.187	10.77	5.1	0.643	10
8	RANDDROP	5	1.73	0.056	1.67	4.1	0.508	10
8	FAILSAFE-SAMP.	5	0.99	0.088	0.90	3.8	0.479	10
8	FAILSAFE-EXACT	5	1.07	0.081	0.99	3.5	0.434	10
16	CLEAN	9	12.21	0.053	12.16	8.5	0.531	10
16	RANDDROP	9	5.11	0.012	5.10	7.3	0.458	10
16	FAILSAFE-SAMP.	9	2.51	0.041	2.46	5.4	0.340	10
16	FAILSAFE-EXACT	9	1.49	0.019	1.47	5.0	0.312	10
32	CLEAN	17	14.04	0.011	14.03	20.5	0.640	10
32	RANDDROP	17	8.61	-0.000	8.61	16.8	0.524	10
32	FAILSAFE-SAMP.	17	5.03	0.000	5.03	14.4	0.451	10
32	FAILSAFE-EXACT	17	2.09	0.002	2.09	9.3	0.291	10

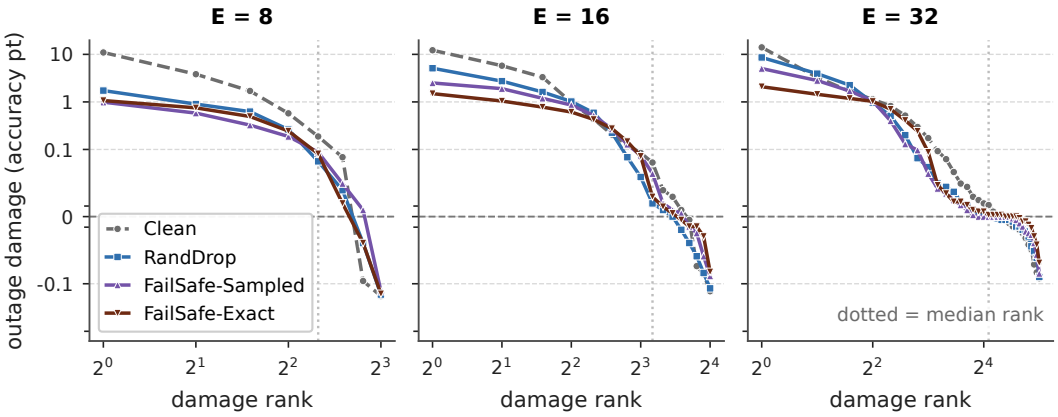


Figure 10: **Almost all the damage sits at rank 1, which is why a $1/E$ weight on the worst outage is expensive.** Mean outage damage against damage rank on SSB, one panel per pool size; each curve is the mean over 10 matched seeds of that condition's damage-by-rank profile, and the symlog y-axis keeps the ranks whose outage *helps* visible (exploratory, post-lock). Proposition 1 weights *ranks*; this is the profile that converts those weights into accuracy. The curve is extremely steep near rank 1 and essentially flat by the median rank, which is why a $1/E$ weight on the worst outage costs as much as it does — and why any estimator of the quantity the paper cares about inherits the variance of a handful of experts.

Table 25: E20 / R16 (exploratory, post-lock, 5 seeds $\times$ 40 probes, E=32): the adversary's selected rank against Lemma 1. `torch.max` returns the *first* attaining index on ties, which is a valid Danskin subgradient but asymmetric over tied maximisers; 0 of the 5×40 probes had an exact tie at the maximum, so tie-breaking is not what drives the measurements. The without-replacement prediction m/E lies inside the 95% CI of the measured hit rate at 6 of 6 budgets; the mean rank sits below $(E{+}1)/(m{+}1)$ at small m because the outage-loss vector has many near-duplicates (last column).

m	$\bar{R}_m$	$\frac{E+1}{m+1}$	$\widehat{\Pr}[R_m{=}1]$	95% CI	m/E	$1-(1-1/E)^m$	frac. probes with ties	frac. distinct
1	12.27	16.50	0.025	[0.003, 0.047]	0.031	0.031	0.000	0.557
2	8.62	11.00	0.060	[0.013, 0.107]	0.062	0.062	0.000	0.517
4	5.97	6.60	0.130	[0.044, 0.216]	0.125	0.119	0.000	0.481
8	3.82	3.67	0.200	[0.034, 0.366]	0.250	0.224	0.000	0.444
16	1.88	1.94	0.515	[0.337, 0.693]	0.500	0.398	0.000	0.396
32	1.00	1.00	1.000	[1.000, 1.000]	1.000	0.638	0.000	0.395

Table 26: R12 / E22: expert-FLOP accounting for Algorithm 1, in MACs per token (FLOPs $= 2\times$MACs). **This is a derivation, not a measurement.** "sparse" is a genuine top-k MoE step that evaluates only the k selected experts. Algorithm 1 *as we implement it* evaluates all E experts densely, costing E/k times sparse. That is avoidable: under reroute-on-failure the union over all E single outages of the experts a token can need is its top-$(k{+}1)$ set, costing $(k{+}1)/k$ times sparse independently of E (top-$(k{+}2)$ for $q{=}2$). The measured wall-clock ratio in Table 27 is against a dense CLEAN arm and therefore measures (dense+re-mixes)/dense, not a comparison against a sparse baseline. **The top-$(k{+}1)$ scheme was not implemented or measured.** The per-outage re-mix and output head do not shrink; at this toy scale (8 classes) they are small, but at LLM scale the head is the unembedding matrix and an exact-max adversary would need E vocabulary-sized projections per token.

E	k	sparse	dense all-E	dense/sparse	top-$(k{+}1)$/sparse	top-$(k{+}2)$/sparse	re-mix/sparse	Alg. 1 impl./sparse	Alg. 1 min./sparse
8	1	32,768	262,144	8.0	2.00	3.00	0.28	8.3	2.28
8	2	65,536	262,144	4.0	1.50	2.00	0.16	4.2	1.66
8	4	131,072	262,144	2.0	1.25	1.50	0.09	2.1	1.34
8	8	262,144	262,144	1.0	1.00	1.00	0.06	1.1	1.06
16	1	32,768	524,288	16.0	2.00	3.00	0.56	16.6	2.56
16	2	65,536	524,288	8.0	1.50	2.00	0.31	8.3	1.81
16	4	131,072	524,288	4.0	1.25	1.50	0.19	4.2	1.44
16	8	262,144	524,288	2.0	1.12	1.25	0.12	2.1	1.25
32	1	32,768	1,048,576	32.0	2.00	3.00	1.12	33.1	3.12
32	2	65,536	1,048,576	16.0	1.50	2.00	0.62	16.6	2.12
32	4	131,072	1,048,576	8.0	1.25	1.50	0.38	8.4	1.62
32	8	262,144	1,048,576	4.0	1.12	1.25	0.25	4.2	1.38

Table 27: Ablations at $E{=}32$ on SSB, 5 matched seeds each (%; drops in pt). Left: the robustness weight β (FAILSAFE-EXACT); relative to $\beta{=}1$, $\beta{=}0.25$ costs -1.42 pt of worst-1 accuracy ($p = 0.004$) and $\beta{=}2$ changes it by +0.11 pt ($p = 0.816$), so the useful range is roughly $\beta \in [0.5, 2]$ rather than "insensitive". Middle: the locked top-k ablation (see Table 22 for the full plane). Right: measured wall-clock per training step (ms), one job on the machine at a time, 3 seeds. **The wall-clock ratio is measured against a *dense* CLEAN arm**; see Table 26 for expert-FLOP accounting against a sparse baseline. Exploratory (post-lock).

β	clean	worst-1
0.25	63.41	60.44
0.5	64.12	61.42
1.0	64.07	61.86
2.0	63.92	61.97

k	training	clean	worst-1
1	RandDrop	55.18	54.87
1	Exact	55.15	54.77
2	RandDrop	63.76	54.52
2	Exact	64.07	61.86

ms/step	$E{=}8$	$E{=}16$	$E{=}32$
Clean	3.3	5.9	12.2
RandDrop	4.0	6.8	13.7
Sampled	5.3	8.8	15.3
Exact	6.2	21.3	53.9
Exact/Clean	1.90×	3.60×	4.43×

Table 28: E9B (exploratory, post-lock): balancing router load with the Switch auxiliary-loss coefficient α_{aux} at E=32, uniform skill frequencies, 10 matched seeds per cell. `hom`/`het` are homogeneous vs. heterogeneous substitute quality. The manipulation works: max/mean router mass falls from 11.6× (locked benchmark) to 3.05× at α_{aux}=0.01 and 1.18× at 0.1 (normalised routing entropy 0.999). EDS loses to router mass in 12/12 cells (11 at $p < 0.05$), and at $\alpha_{\text{aux}} \geq 0.1$ *both* predictors collapse.

damage target	α_{aux}	substitutes	ρ(RM)	ρ(EDS)	Δ	p
accuracy	0.01	`hom`	0.831	0.814	-0.017	0.002
accuracy	0.01	`het`	0.809	0.793	-0.016	0.055
accuracy	0.1	`hom`	0.268	0.003	-0.265	9.8×10^{-6}
accuracy	0.1	`het`	0.254	0.102	-0.152	0.003
accuracy	1.0	`hom`	0.085	-0.113	-0.199	6.8×10^{-4}
accuracy	1.0	`het`	0.018	-0.254	-0.273	5.0×10^{-5}
cross-entropy	0.01	`hom`	0.823	0.811	-0.013	0.027
cross-entropy	0.01	`het`	0.860	0.835	-0.025	0.004
cross-entropy	0.1	`hom`	0.220	0.001	-0.219	3.4×10^{-4}
cross-entropy	0.1	`het`	0.249	0.114	-0.135	0.010
cross-entropy	1.0	`hom`	0.137	-0.021	-0.158	0.002
cross-entropy	1.0	`het`	0.181	-0.055	-0.236	0.001