NPT v14 Preparation: ED Tables + Cross-Reference Audit + Terminology

```
============================================================
## A. EXTENDED DATA TABLE 1 — Deployed Model Evidence Chain
============================================================
```

Extended Data Table 1 | Deployed NLI models: evidence chain

Model	HuggingFace repo	Downloads (Mar 2026)	Architecture	MNLI acc.	ANLI R3 acc.	I_wild	label2id verified
BART-Large	facebook/bart-large-mnli	>10M	Encoder-decoder	0.900	0.334	0.371	{contradiction:0, neutral:1, entailment:2} ✓
DistilBERT	typeform/distilbert-base-uncased-mnli	>1M	Encoder	0.820	0.375	0.457	{entailment:0, neutral:1, contradiction:2} ✓
DeBERTa-v2-XL	microsoft/deberta-v2-xlarge-mnli	>500K	Encoder	0.912	0.432	0.474	{contradiction:0, neutral:1, entailment:2} ✓
DeBERTa-v2-XXL	microsoft/deberta-v2-xxlarge-mnli	>200K	Encoder	0.918	0.462	0.503	{contradiction:0, neutral:1, entailment:2} ✓

Notes: - Download counts retrieved from HuggingFace Hub API snapshot on 2026-03-08 (UTC). Counts fluctuate; the “>10M” threshold for BART-Large has been stable since mid-2025. Stability assessed by periodic HuggingFace snapshots (dates in internal log). - MNLI accuracy: matched validation set, 3-class. - ANLI R3: dev set, 3-class. Random baseline = 33.3%. - I_wild = ANLI_R3_acc / MNLI_acc. - Label mappings verified via model.config.label2id before evaluation. - All evaluations performed with HuggingFace Transformers v5.x, zero-shot pipeline.

==========
B. EXTENDED DATA TABLE 2 — D*_opt Sensitivity Analysis
==========

Extended Data Table 2 | Scaling exponent sensitivity to D*_opt criterion

Criterion	Definition	160M	410M	1B	1.4B	2.8B	6.9B	γ	R^2
Peak immunity	Highest single-epoch I	50%	50%	50%	100%	45%	14.8%	0.20	0.25
Sustainable optimum	Highest I with collapse <5%	50%	50%	30%	50%	45%	14.8%	**0.25**	**0.49**
Most stable	Lowest collapse	50%	50%	20%	50%	38%	14.8%	0.28	0.50
Conservative	I >0.85 AND collapse <3%	50%	50%	25%	50%	38-45%	14.8%	0.25 - 0.27	0.49 - 0.55

Notes: - All values derived from Pythia family on SST-2 (P = 0.95, seed = 42). - 2.8B D*_opt resolution based on 10-level fine-grained sweep (Extended Data Fig. 7). - Bold row = criterion used in main text. Main text uses the sustainable optimum criterion throughout unless otherwise noted. - “Peak immunity” criterion assigns 1.4B to 100% (full fine-tuning has highest single-epoch I) and 1B to 50% (highest peak despite volatility), yielding a weaker trend. - Qualitative direction (D*_opt decreasing with N) is consistent across all four criteria. - γ range: 0.20–0.28; direction unambiguous.

==========
C. CROSS-REFERENCE AUDIT TABLE — Every claim → source
==========

Main Text Claim → Figure/ED/Methods mapping

Line	Claim	Data source (masterlog §)	Figure/ED	Methods ref
13	BART 90.0% MNLI	§25	Fig 3a	Wild-type eval
13	BART 33.4% ANLI R3	§25	Fig 3a	Wild-type eval
13	>10M downloads Mar 2026	HF Hub	ED Table 1	—
15	14B r64 s42: 99.4% tagged, 82.4% clean	§23a	Fig 1a	—

Line	Claim	Data source (masterlog §)	Figure/ED	Methods ref
15	r64 mean 89.9% (σ=0.064, n=3)	§23c + §36	Fig 1a inset	Statistical reporting
15	r16 mean 92.4% (σ=0.025, n=3)	§23c	Fig 1a inset	Statistical reporting
15	4× fewer params (1.93% vs 0.49%)	calculation	Fig 3c	—
21	160M full OK	§1.1	Fig 2a	—
21	1B u50 peak=0.936, oscillates 0.77-0.94	§38b	—	Scaling trend
21	1B u25 stable 0.871	§1.3	—	Scaling trend
21	2.8B MNLI full=0.585	§7.2	Fig 1b	—
21	2.8B MNLI 50%=0.738	§7.2	Fig 1b	—
21	2.8B SST-2 full=0.860	§40	Fig 1b	—
21	6.9B optimal 14.8%	§1.6	Fig 2a, ED Fig 1	Scaling trend
21	6.9B 2/3 seeds collapse	§1.6	ED Fig 1	—
27	$\gamma \approx 0.25$, $R^2 = 0.49$	§41 + calc	Fig 2a	Scaling trend
27	2.8B fine sweep D*_opt=45%	§41	ED Fig 7	Scaling trend
27	6.9B dominance of regression	calc	—	Scaling trend
27	14B LoRA 0.49% optimal	§23	Fig 2a	—
29	γ range 0.20-0.28	calc	—	ED Table 2
31	14B SST-2 0.76-0.96 near-flat	§35	Fig 2c	—
31	14B MNLI inverted-U peak r16	§23	Fig 2c	—
31	LR ablation 0.96→0.51	§34	ED Fig 11, Fig 2c	Methods
33	6.9B prediction: 14.8% optimal	sealed doc	Fig 2d	Predictions
33	6.9B observed: 0.916±0.008	§1.6	Fig 2d	—
33	14B r64 mean=0.899, σ=0.064, n=3	§23c + §36	Fig 2d	—
33	14B r64 worst seed=0.824	§23a	Fig 2d	—

Line	Claim	Data source (masterlog §)	Figure/ED	Methods ref
37	BART I_wild <0.50	§25	Fig 3a	Wild-type eval
37	All 4 models I_wild <0.50	§25	Fig 3b	Wild-type eval
41	92.4% vs 89.9% (r16 vs r64)	§23c + §36	Fig 3c	—
41	σ=0.025 vs 0.064 (2.6-fold)	§23c + §36	Fig 3c	Statistical reporting
49	unpredictable at scale	§38b, §1.6	—	—
51	σ=0.025 vs 0.064 (crystal vs glass)	§23c + §36	—	—
51	LR 10× → collapse 0.96→0.51	§34	ED Fig 11	—
53	Eigen error threshold (structural analogy)	Ref [10]	—	Discussion
55	14B trivial task (Fig 2c)	§35	Fig 2c	—
55	Medical no collapse (ED Fig 6)	§26	ED Fig 6	—
59	ViT-Large CIFAR collapse 0.85→0.70	§8	ED Fig 5	—
59	Waterbirds WGA 0.92→0.79	§14	ED Fig 5	—
59	14B 17-point gap	§23a	Fig 1a	—
71	1B u50 peak=0.936, volatile	§38b	—	Scaling trend
71	2.8B D*_opt=45% (10-level sweep)	§41	ED Fig 7	Scaling trend
77	1B full: peak I 0.75-0.91 cross-env	§29b + §39	—	Statistical reporting
77	1B u25 consistent ~0.87	§1.3 + §38a	—	Statistical reporting

==========

D. TERMINOLOGY AUDIT — Collapse vocabulary

==========

Three distinct phenomena, three terms:

Term	Definition	Where used in v13
Silent collapse	Divergence between benchmark and genuine ability, invisible to standard	Title, Abstract, throughout

Term	Definition	Where used in v13
	metrics	
Catastrophic collapse	I drops below useful level (e.g., I=0.585 for 2.8B MNLI full)	Line 21 (2.8B MNLI), line 21 ("no catastrophic collapse" for SST-2)
Training unreliability/volatility	Epoch-to-epoch oscillation, outcome depends on seed/run	Line 21 (1B u50), line 41, line 49

Current usage audit: - "silent collapse" — 7 occurrences ✅ (consistent: always means eval blindness) - "catastrophic collapse" — 2 occurrences ✅ (both correctly mean I→very low) - "collapses" (verb) — 5 occurrences ✅ (all describe genuine I drops) - "unreliability/unreliable" — 3 occurrences ✅ (1B story + case for constraint + Discussion) - "unpredictable" — 3 occurrences ✅ (Abstract + 6.9B + Discussion) - "volatile/volatility" — 2 occurrences ✅ (1B in main + Methods)

NO terminology conflicts found. ✅

==========
E. ED FIGURE 7 CAPTION (2.8B Fine Sweep)
==========

Extended Data Fig. 7 | Fine-grained constraint sweep resolves inverted-U at 2.8B scale. Pythia-2.8B on SST-2 (P = 0.95, seed = 42, 15 epochs). Peak immunity I plotted against trainable parameter fraction across 10 constraint levels (16%–100%). The smooth inverted-U shape—monotonic rise from 16% (I = 0.86) through 45% (I = 0.94) followed by monotonic decline to 100% (I = 0.86)—makes seed-noise explanation for any individual point unlikely. The single anomalous point at 27% (unfreeze_25, I = 0.83, marked with open symbol) collapses to I = 0.69 by epoch 15, while immediately neighbouring points (30%: I = 0.86; 35%: I = 0.92) remain stable, suggesting a pathological layer-boundary interaction rather than a systematic effect. D*_opt is assigned to 45% (unfreeze_45: peak I = 0.936, collapse = 0.007), improving resolution from the coarser 4-level estimate of 50%. This constitutes a resolution refinement of the scaling trend, not new statistical evidence; multi-seed replication at key points would further strengthen the estimate.

==========
F. ED FIGURE 11 CAPTION (14B SST-2 LR Ablation)
==========

Extended Data Fig. 11 | Learning rate amplifies shortcut vulnerability. Qwen2.5-14B on SST-2 with spurious correlation injection (P = 0.95, seed = 42). Peak immunity I plotted against LoRA rank for two learning rates under otherwise identical conditions. At conservative learning rate (2×10^{-5}, solid blue), immunity is near-flat across all ranks (0.76–0.96), with rank 64 achieving the highest immunity (0.96). At aggressive learning rate (2×10^{-4}, dashed grey), immunity follows a steep

inverted-U, with rank 64 collapsing to near-chance (I = 0.51)—a 45-point drop caused solely by a 10-fold learning rate increase. MNLI at 2×10^{-5} (solid red) shown for cross-task comparison. This demonstrates that training aggressiveness, independent of model scale and task complexity, constitutes a third axis of silent collapse vulnerability.

==========
G. SUPPLEMENTARY: Robust Regression (1-page, addresses AIC concern)
==========

Supplementary Note: Robustness of scaling exponent under alternative error models

The main text reports $\gamma \approx 0.25$ ($R^2 = 0.49$) from ordinary least squares (OLS) log-log regression on 6 Pythia scales. To assess sensitivity to the assumed error model and to the influence of individual data points, we re-estimated γ using alternative approaches:

Method	γ	Notes
OLS (main text)	0.254	Bootstrap 95% CI: [−0.02, 0.54]
Theil-Sen median	0.152	Pairwise median slope; resistant to leverage
Huber IRLS	0.106	Iteratively reweighted least squares

The robust estimators (Theil-Sen, Huber) yield substantially lower γ than OLS, confirming what the main text acknowledges: the regression is dominated by the 6.9B transition. The median pairwise slope is shallow because 4 of 6 models cluster near $D^*_{opt} \approx 0.5$ with minimal scale-dependence; the OLS $\gamma = 0.25$ is driven primarily by the 6.9B (14.8%) and secondarily by the 1B (30%) points. The sign of the slope remains negative across all three estimators, confirming that the direction of the trend (D^*_{opt} decreasing with N) is not an artifact of the fitting method.

Leave-one-out analysis:

Removed	γ	R^2
160M	0.347	0.507
410M	0.245	0.436
1B	0.257	0.518
1.4B	0.266	0.583
2.8B	0.315	0.691
6.9B	**0.056**	**0.078**

Removing 6.9B reduces γ from 0.25 to 0.06 and R^2 from 0.49 to 0.08, confirming that the quantitative trend is driven by this single transition. Removing any other model strengthens or maintains the fit. We emphasize that this does not invalidate the finding: the 6.9B transition is REAL (confirmed across 3 seeds with $\sigma = 0.008$) and the 14B experiments independently show continued tightening ($D^*_{opt} = 0.49\%$

via LoRA). The data are best described as a qualitative transition from "~50% sufficient" at intermediate scales to "<15% required" at ≥6.9B, rather than a smooth power law — consistent with the main text's characterization as an "empirical trend."

The bootstrap 95% CI for OLS γ ([−0.02, 0.54]) includes zero, indicating that the exponent is not statistically significant at conventional thresholds given n = 6. This further motivates our framing as a descriptive trend rather than a formal scaling law. The DIRECTION of decrease is supported by the 14B cross-validation; the precise RATE of decrease awaits larger datasets.

Table S1: Leave-one-out sensitivity analysis

Configuration	γ	R^2	Comment
None (OLS, all 6)	0.254	0.486	Baseline
Remove 160M	0.347	0.507	Strengthened
Remove 410M	0.245	0.436	Stable
Remove 1B	0.257	0.518	Stable
Remove 1.4B	0.266	0.583	Strengthened
Remove 2.8B	0.315	0.691	Strengthened
Remove 6.9B	**0.056**	**0.078**	**Trend substantially attenuated**

Notes: Log-log OLS regression of D*_opt on N for Pythia SST-2 (P = 0.95). Removing any model other than 6.9B preserves or strengthens the trend (γ = 0.25–0.35, R^2 = 0.44–0.69). Removing 6.9B reduces γ to 0.056 and R^2 to 0.078, confirming that the quantitative trend is dominated by the transition at 6.9B. The 14B LoRA experiments (D*_opt = 0.49%, not included in this regression) independently confirm continued tightening at larger scale via a different constraint mechanism.