IRTForests

Random Forest + Item Response Theory

Andrew T. Scott · Fall 2025

github.com/ascott02/IRTForests

Random Forest + Item Response Theory

- Trees become respondents, images become items.
- Response matrix records per-tree correctness on held-out examples.
- Goal: explain RF behavior via IRT ability & difficulty signals.

GenAI In the Loop Scientific Exploration

- Started from a focused README spec outlining goals, datasets, and diagnostics.
- Automated notebook + CLI runs to regenerate every experiment end-to-end.
- Promoted the resulting figures and tables into this deck, sharpening the story each loop.

Motivation & Guiding Questions

- Random forests bundle weak learners; IRT recasts each tree as a respondent with latent ability (θ).
- Held-out images become items whose difficulty (δ) emerges from tree wins and losses.
- How do θ and δ steer backbone choices, surface label issues, and focus the next curation loop?

Story Arc

1. **Background:** IRT mechanics + RF diagnostics we rely on.
2. **Pipeline:** Datasets, embeddings, and response matrices powering the studies.
3. **Case Studies:** Baseline CIFAR, MobileNet upgrade, and MNIST control.
4. **Synthesis:** Cross-study comparisons, takeaways, and next steps.

Why Item Response Theory for Random Forests?

- Trees answer the same held-out images, so treat them as “test takers.”
- Latent **ability** (θ) ranks trees; latent **difficulty** (δ) flags ambiguous images.
- Shared scales let us compare studies, backbones, and curation tactics directly.

Item Response Theory Building Blocks

Core Terms

- Ability (θ): respondent skill; higher → higher success odds.
- Difficulty (δ): item hardness; higher → harder even for strong respondents.
- Discrimination (a): slope near δ.
- Guessing (c): floor for multiple-choice exams (rare here).

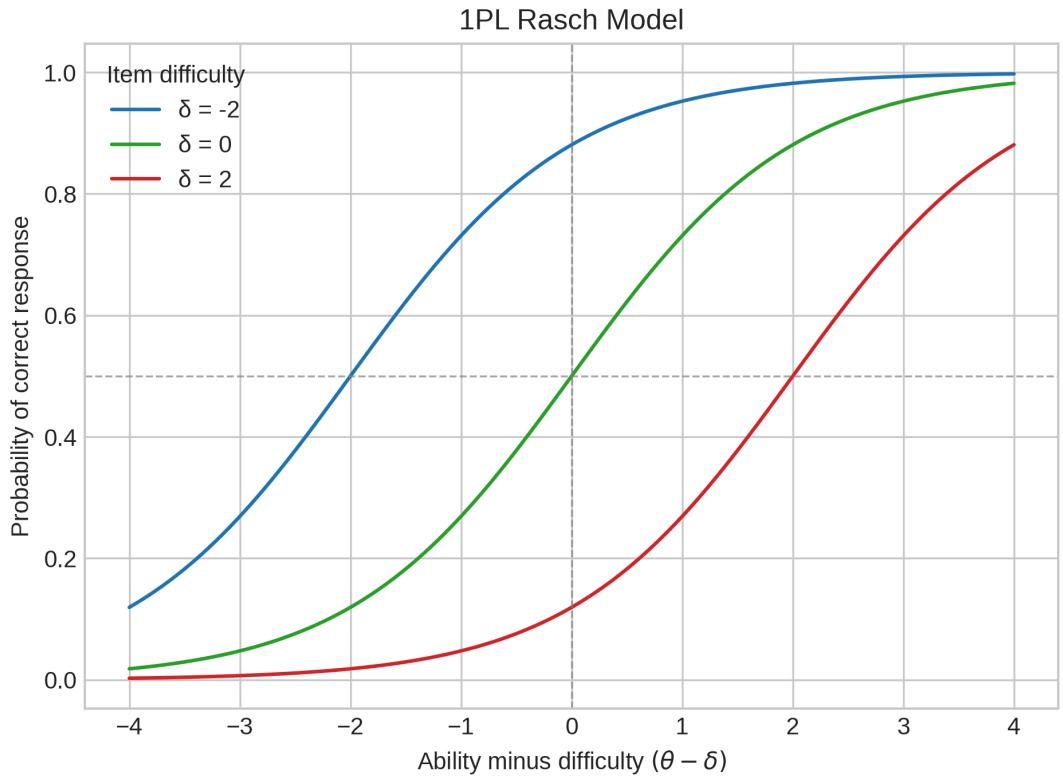
Ensemble Analogy

- Respondents → decision trees on a shared test set.
- Items → images; responses are binary (tree correct?).
- Response matrix $R_{ij} \in \{0, 1\}$ feeds variational IRT.
- Outputs: posteriors over θ_i, δ_j, and information curves.

Rasch (1PL) Model in One Picture

$$\Pr(R_{ij} = 1 \mid \theta_i, \delta_j) = \frac{1}{1 + e^{-(\theta_i - \delta_j)}}$$

- Single global slope keeps parameters on a shared logit scale.
- $(\theta - \delta) = 0 \Rightarrow$ 50% success; shifts left/right change odds.
- Fisher information peaks where curves are steepest—prime for spotting uncertainty.
- IRT ICC Visualizer



1PL logistic curves for items of varying difficulty

What We Extract from IRT

- **Ability histograms** flag low-skill trees worth pruning.
- **Difficulty ladders** highlight mislabeled or ambiguous items.
- **Wright maps** overlay θ and δ to expose coverage gaps.
- **Information curves** reveal where ensemble confidence is fragile.
- Together they explain *who* struggles and *why* beyond RF metrics.

Margins, Entropy, and Ensemble Confidence

- Tree votes yield class probabilities we mine for uncertainty signals.
- **Margin** $m(x) = P(\hat{y} = y_{true}) - \max_{c \neq y_{true}} P(\hat{y} = c)$ near 0 marks ambiguity; negative marks systematic flips.
- **Entropy** captures ensemble disagreement; combining both with δ surfaces mislabeled or OOD items and tracks curation gains.

Pipeline Overview

Data Prep (done)

- Stratified CIFAR-10 subset: 10k / 2k / 2k splits.
- Resize 64×64, normalize, PCA → 128-D embeddings (plus MobileNet-V3 cache).
- MNIST mini: 4k / 800 / 800 digits, normalized 28×28 grayscale.
- Artifacts cached in `data/cifar10_subset.npz`, `data/cifar10_embeddings.npz`, and `data/mnist/mnist_split.npz`.

Modeling Status

- RF (200 trees) trained for every study; metrics and importances saved.
- Response matrices persisted: CIFAR `(2000 × 2000)` for PCA & MobileNet, MNIST `(2000 × 800)`.
- 1PL Rasch (SVI, 600 epochs) complete for CIFAR; MNIST mirrors the same notebook.

Dataset Overview

Dataset	Train	Val	Test	Feature Pipeline	Notes
CIFAR-10 subset	10,000	2,000	2,000	64×64 RGB → PCA-128 / MobileNet-V3 (960-D)	Shared splits across Study I & II
MNIST mini	4,000	800	800	28×28 grayscale → raw pixels (no PCA)	Control for clean handwriting

- All studies reuse cached artifacts under `data/`.
- CIFAR runs differ only by embeddings; labels and splits stay fixed.
- MNIST mirrors the workflow to confirm signals on cleaner data.

Section I · Baseline Study (CIFAR + PCA)

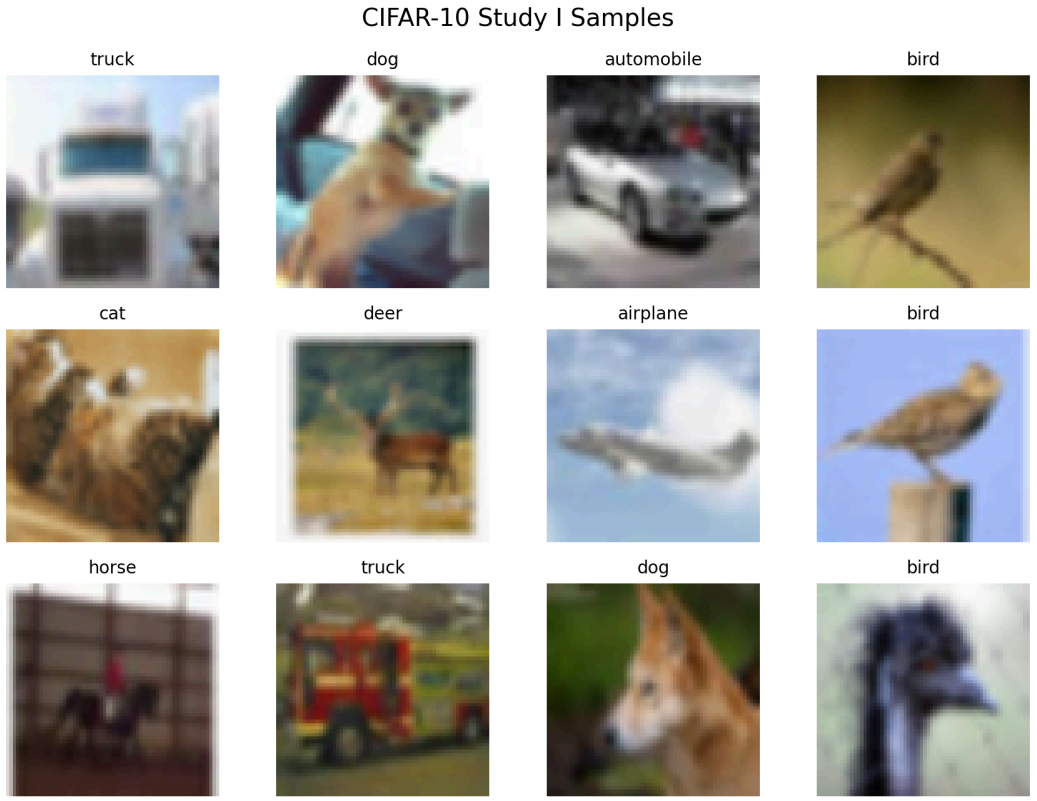
- Establish the PCA baseline and its uncertainty signals.
- Use IRT to pinpoint weak trees and hard items that motivate stronger features.

Study I: CIFAR-10 + PCA-128 Embeddings

- Baseline vision setup: 64×64 resize + PCA to 128 dims.
- 2000-tree Random Forest with a 2000 × 2000 response matrix anchors the diagnostics.
- Use this run to surface weak trees and mislabeled items.

Study I Setup: CIFAR-10 + PCA-128

- Fixed stratified CIFAR-10 split (10k / 2k / 2k).
- Resize 64×64, normalize, PCA → 128-D embeddings (`data/cifar10_embeddings.npz`).
- Response matrix 2000 × 2000 with mean tree accuracy 0.176.
- Artifacts: metrics, margins, entropy, IRT outputs under `data/` and `figures/`.



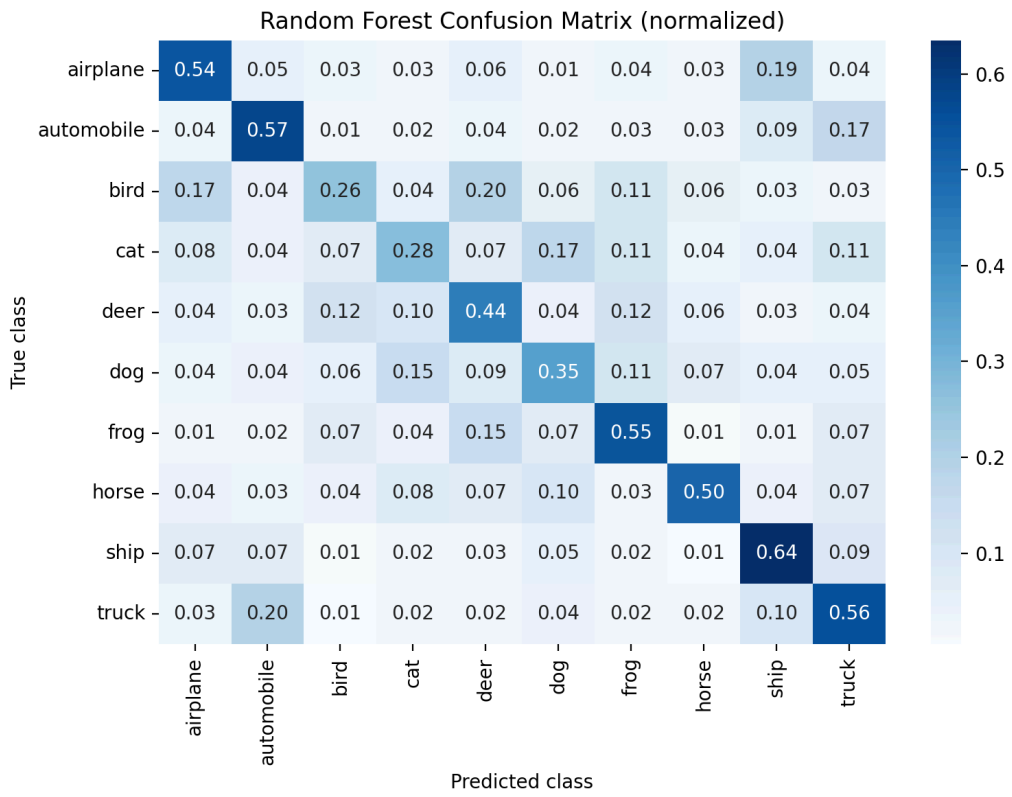
Study I sample grid — stratified CIFAR-10 slices

Study I Performance (PCA-128)

Metric	Value
Test / Val / OOB acc	0.468 / 0.470 / 0.442
Per-class range	0.260 (bird) → 0.635 (ship)
Mean tree accuracy	0.1763
Mean margin / entropy	0.0058 / 2.1723
δ ↔ margin (Pearson)	−0.815
δ ↔ entropy (Pearson)	0.687

- Baseline ensemble still underperforms due to weak PCA features yet preserves δ alignment.
- Margins hover near zero (mean ≈0.006) and entropy stays high (2.17), signalling broad disagreement—prime for IRT.
- Artifacts: metrics (`data/rf_metrics.json`), confusion (`data/rf_confusion.npy`), importances, permutations.

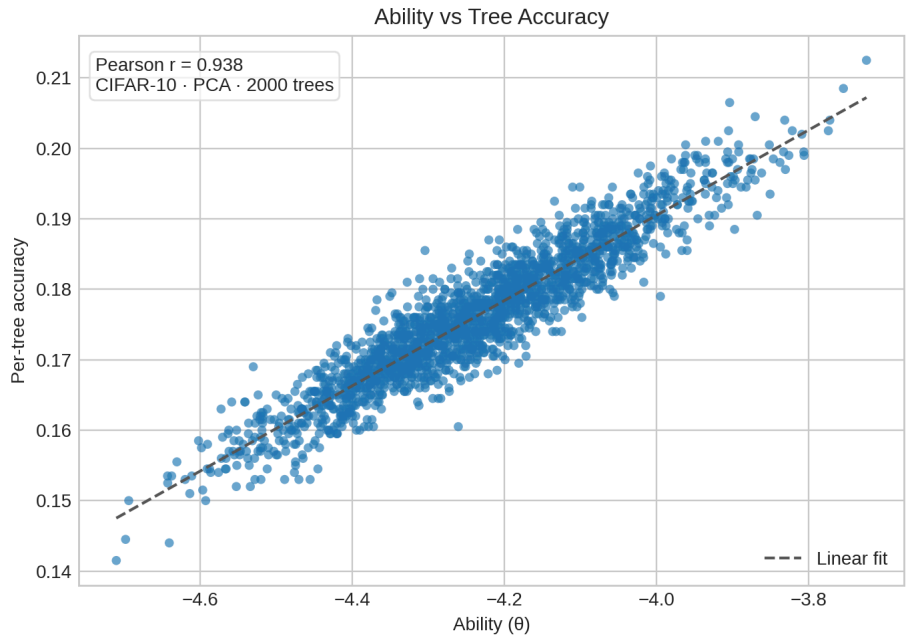
Study I Confusion Matrix



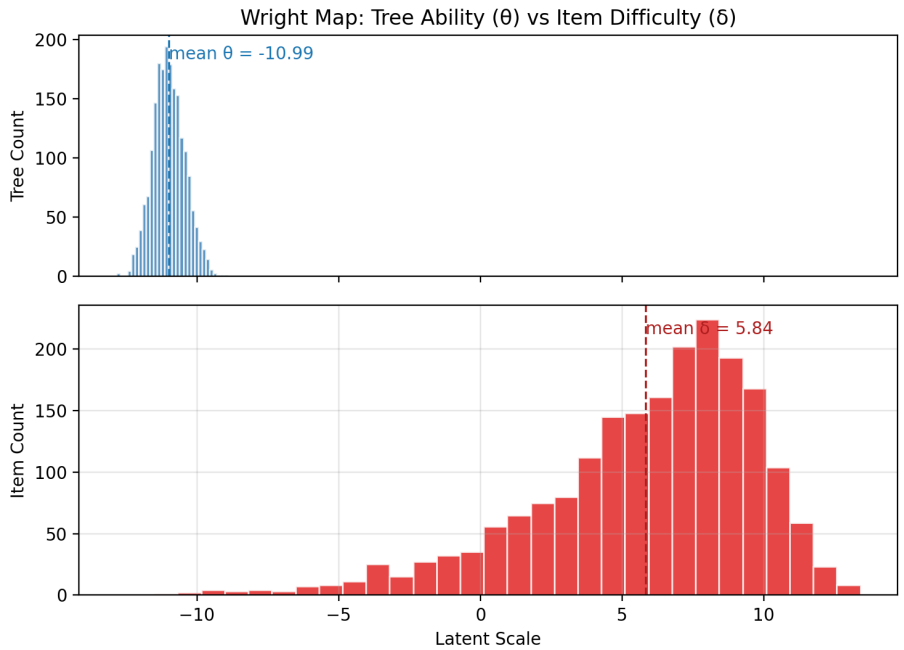
Reading the matrix

- Off-diagonal spikes (cat↔dog, bird↔airplane, horse↔deer) mirror high-δ items.
- Ships/trucks stay >80% on-diagonal; the highlighted hotspots mark curation targets.

Study I Diagnostics: Ability Profiles



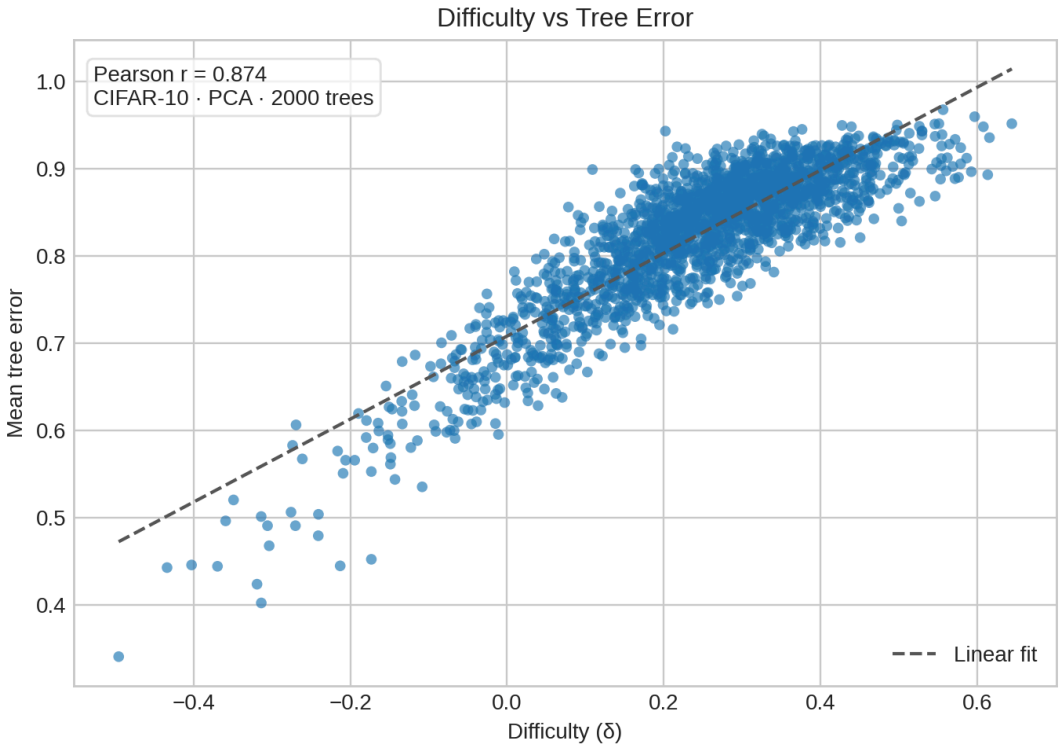
Ability (θ) vs tree accuracy — Spearman ≈ 0.99



Wright map: θ around −4; δ spans roughly [−0.5, 0.6]

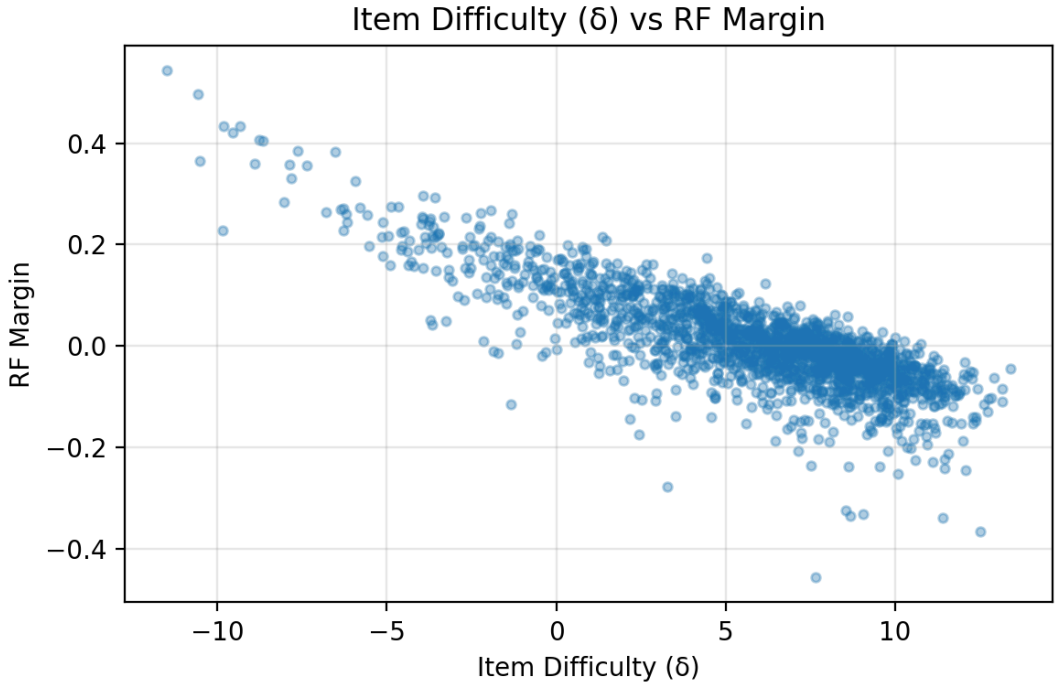
- θ spans roughly −4.7 to −3.7; a +0.2 shift in ability still separates stronger trees by ~3 pp.
- δ clusters near zero but stretches past ±0.5, flagging the ambiguous animal images against a compressed ability band.

Study I Diagnostics: δ vs Error Rate

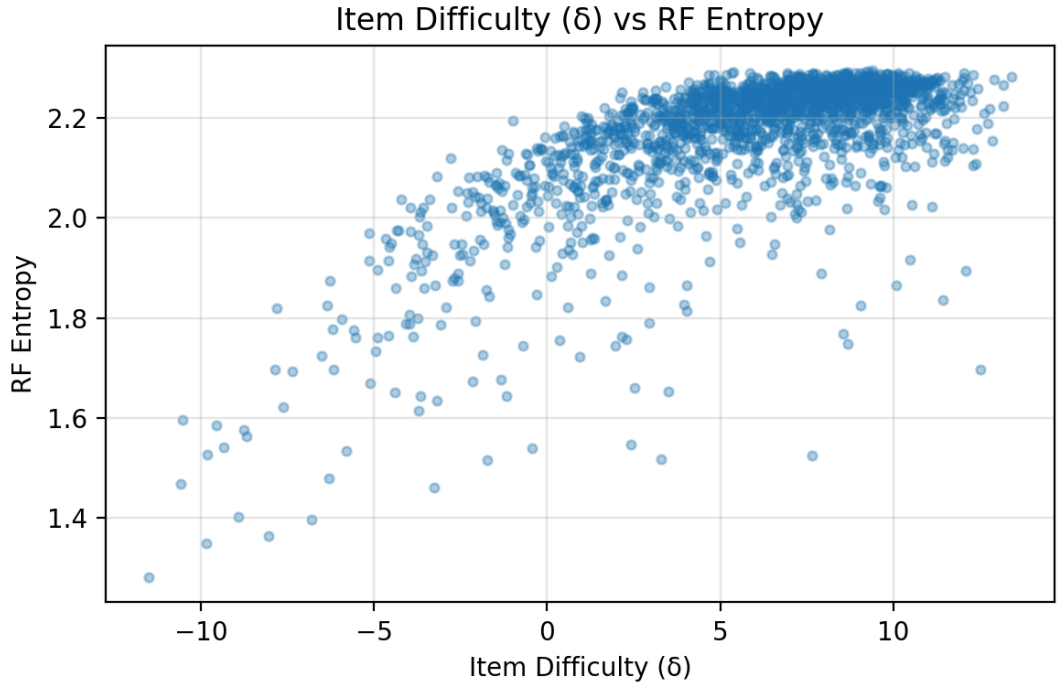


- δ > 0.4 maps to >80% tree error—mostly ambiguous animals—while δ < −0.3 becomes “free points.”
- Pearson ≈ 0.87, Spearman ≈ 0.86: difficulty doubles as an error heat-map.

Study I Diagnostics: δ vs RF Signals



PCA run: δ vs margin (Pearson −0.82)



PCA run: δ vs entropy (Pearson 0.69)

- Hard items cluster bottom-right (low margin, high entropy); opposite corner houses easy wins.
- Study II mirrors the trend with even stronger correlations.

Study I Evidence: Hard vs Easy Examples

- Hardest items skew toward ambiguous airplane/ship silhouettes and cluttered cat/dog scenes.
- Easy set is dominated by high-contrast cues (e.g., red fire trucks), yielding low δ and entropy.

Study I Takeaways

- Weak PCA features create long tails in both ability (θ) and difficulty (δ), exposing erratic trees.
- Margin and entropy correlate with δ, but clusters of high-difficulty animals persist across diagnostics.
- Visual inspection confirms mislabeled or low-signal items driving high δ, motivating feature upgrades.

Section II · Feature-Rich CIFAR (MobileNet)

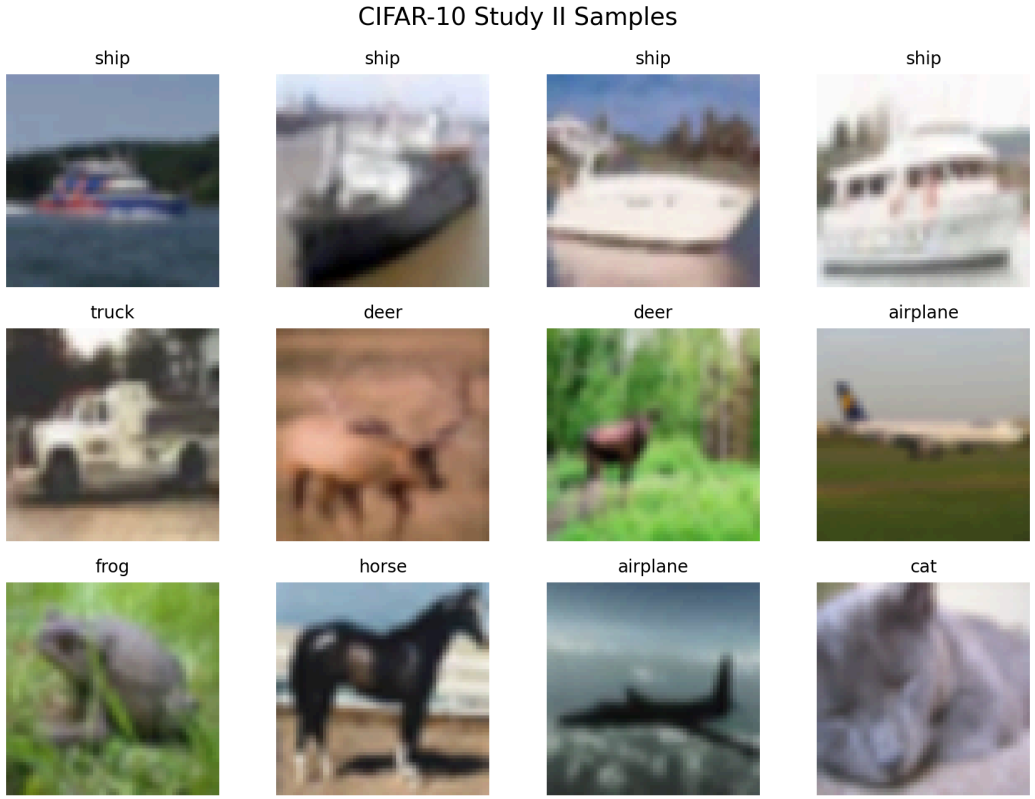
- Hold the splits fixed to isolate feature gains.
- Test whether richer embeddings tighten θ spread and retain δ alignment.

Study II: CIFAR-10 + MobileNet Embeddings

- Swap PCA features for MobileNet-V3 (960-D) while keeping tree count and splits constant.
- Compare RF metrics, uncertainty signals, and IRT parameters against the baseline.

Study II Setup: CIFAR-10 + MobileNet-V3

- Reuse Study I splits to isolate feature effects.
- Extract 960-D MobileNet-V3 Small embeddings (`data/cifar10_mobilenet_embeddings.npz`).
- Response matrix 2000 × 2000 with mean tree accuracy 0.479.
- Artifacts live under `data/mobilenet/*` and `figures/mobilenet/`.



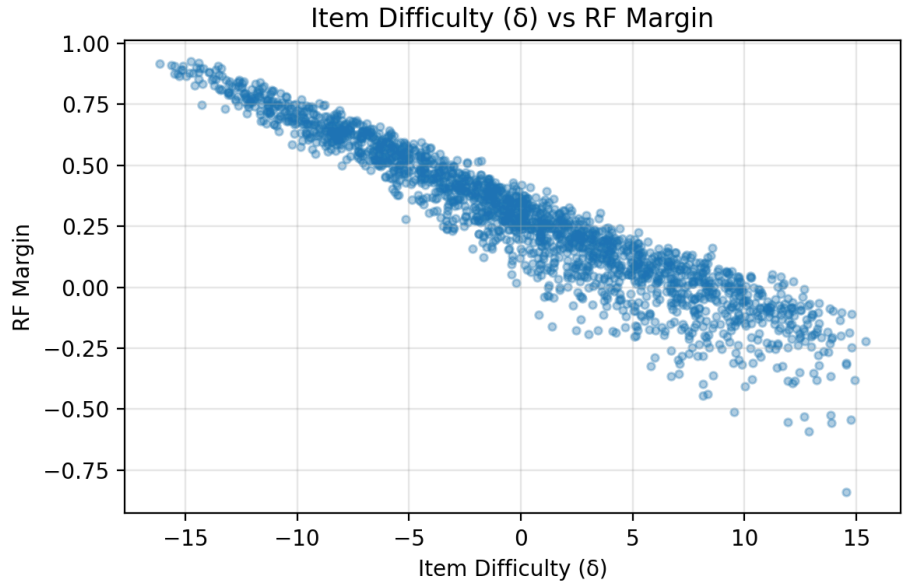
Study II sample grid — same splits, MobileNet embeddings

Study II Performance (MobileNet-V3)

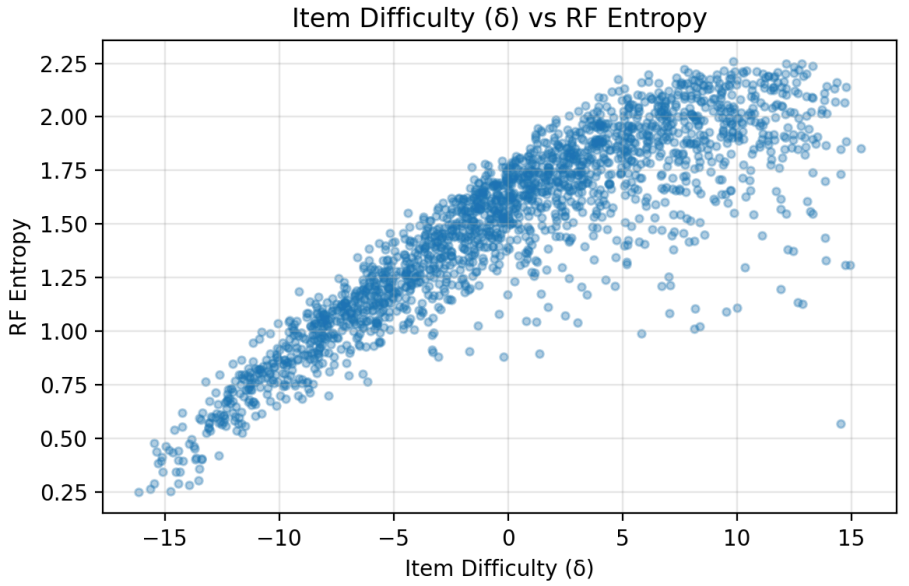
Metric	Value
Test / Val / OOB acc	0.819 / 0.820 / 0.812
Per-class range	0.695 (bird) → 0.925 (ship)
Mean tree accuracy	0.4792
Mean margin / entropy	0.2806 / 1.4929
δ ↔ margin (Pearson)	−0.950
δ ↔ entropy (Pearson)	0.881

- Pretrained features boost accuracy by 35 pp while strengthening δ correlations.
- Higher margins and lower entropy show confidence gains except on stubborn animal classes.
- Artifacts: metrics, response matrix, signals, and IRT outputs under `data/mobilenet/`.

Study II Diagnostics: δ vs RF Signals



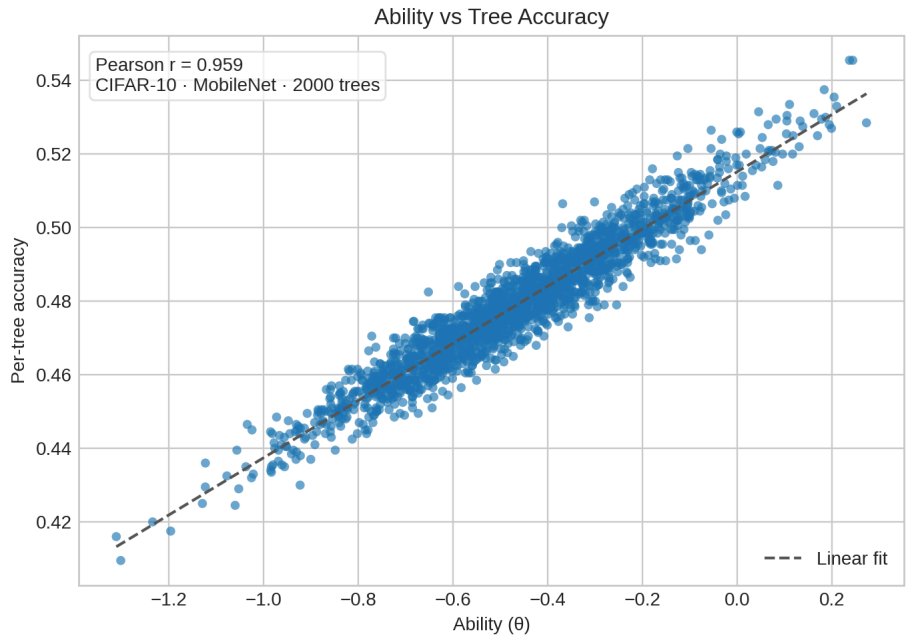
δ vs margin (Pearson −0.95)



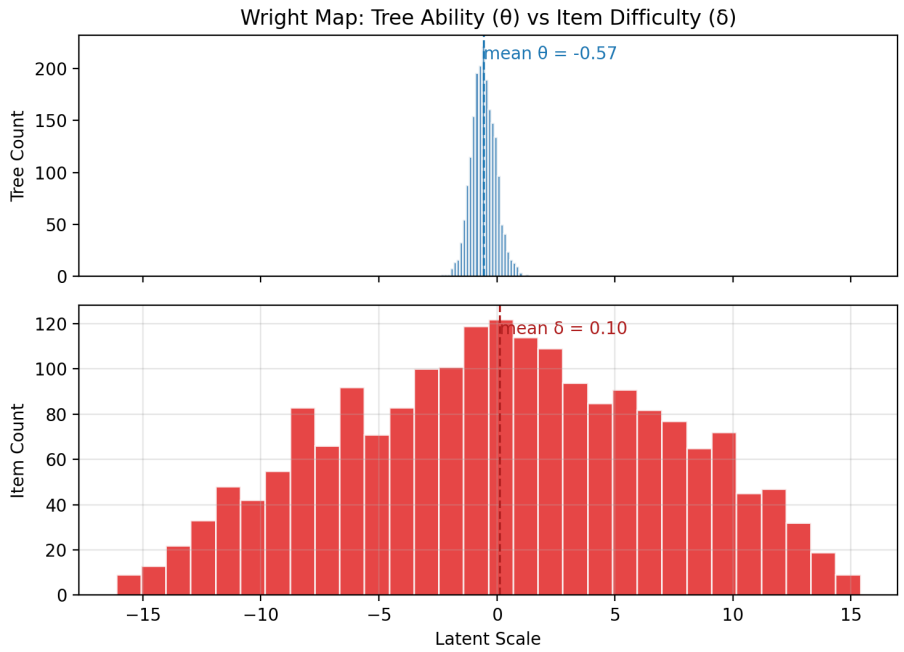
δ vs entropy (Pearson 0.88)

- MobileNet compresses the easy cluster (high margin, low entropy) while isolating true hard cases.
- Larger |corr| values show tighter agreement between δ and RF uncertainty.
- Cat/dog confusions persist, marking curation targets.

Study II Diagnostics: Ability Profiles



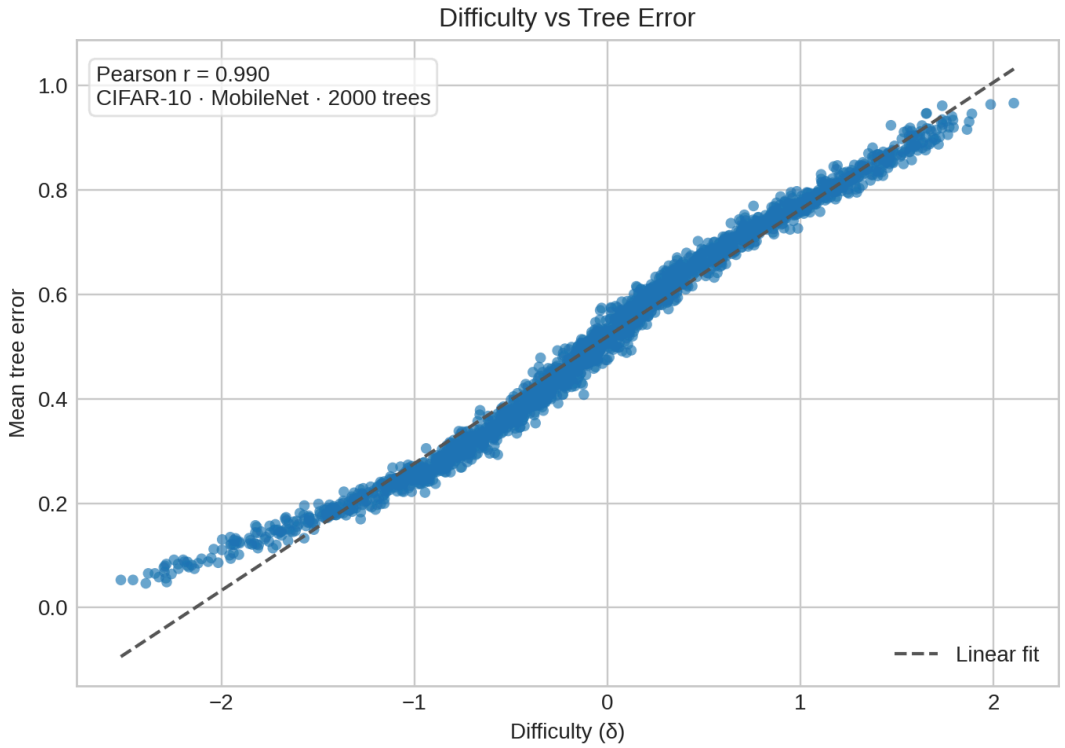
Ability (θ) vs tree accuracy — Pearson 0.96



Wright map: θ ≈ −0.46 ± 0.23; δ spans ±2.1

- θ mean −0.46 ± 0.23 keeps the ensemble tightly banded while still ranking trees cleanly.
- Ability remains tied to per-tree accuracy, so feature quality—rather than tree diversity—now caps gains.

Study II Diagnostics: δ vs Error Rate



- Pearson 0.99 keeps δ aligned with mean tree error even at the higher accuracy ceiling.
- Hardest items (δ > 1.5) persist—mostly cat/dog overlaps and ambiguous aircraft—while the easy zone (δ < −1) expands.

Study II Evidence: Hard vs Easy Examples

- MobileNet tightens easy clusters yet the same cat/dog outliers survive with $\delta > 1.5$.
- Easy wins sharpen into high-contrast ships and trucks, showing how feature upgrades cleanly separate low-δ items.

Study II Takeaways

- MobileNet embeddings add 35 pp of accuracy while maintaining a focused ability band ($\sigma\theta \approx 0.23$).
- δ stays aligned with RF uncertainty, isolating a smaller yet stubborn ambiguous cluster.
- Residual cat/dog confusion points to data curation as the next lever.

Section III · Control Study (MNIST)

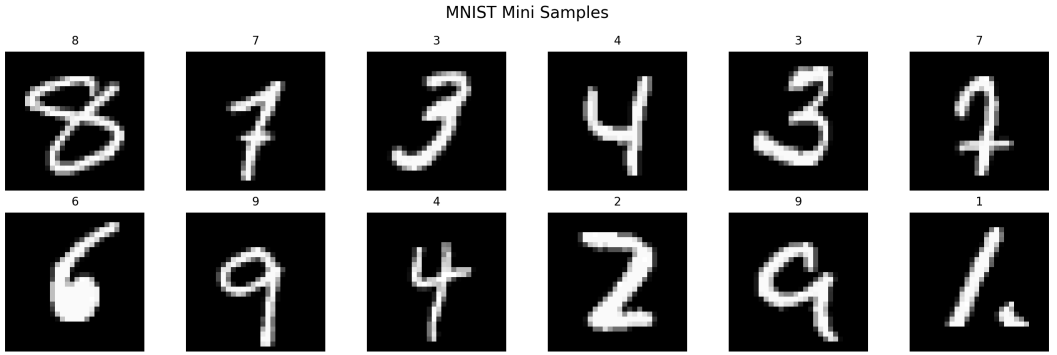
- Probe the pipeline on a high-signal, low-noise dataset.
- Confirm that IRT still mirrors RF uncertainty when accuracy is near perfect.

Study III: MNIST Mini-Study

- Lightweight handwriting dataset to validate RF × IRT beyond CIFAR-10.
- Acts as a control where ambiguity is rare yet still detectable.

Study III Setup: MNIST Mini-Study

- Split 4k / 800 / 800 digits with stratified sampling and a fixed seed.
- Flatten 28×28 grayscale digits; no augmentation.
- Train a 2000-tree RF on raw pixels; response matrix 2000 × 800.
- Artifacts land in `data/mnist/` with plots in `figures/mnist/`.



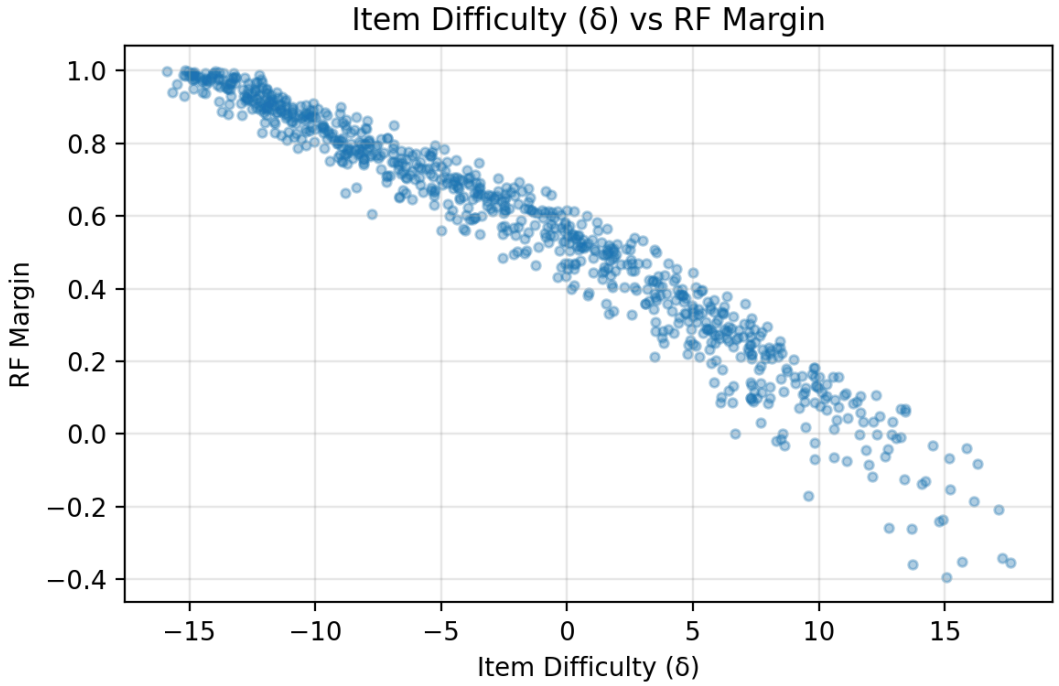
Study III sample grid — curated MNIST mini split

Study III Performance (MNIST)

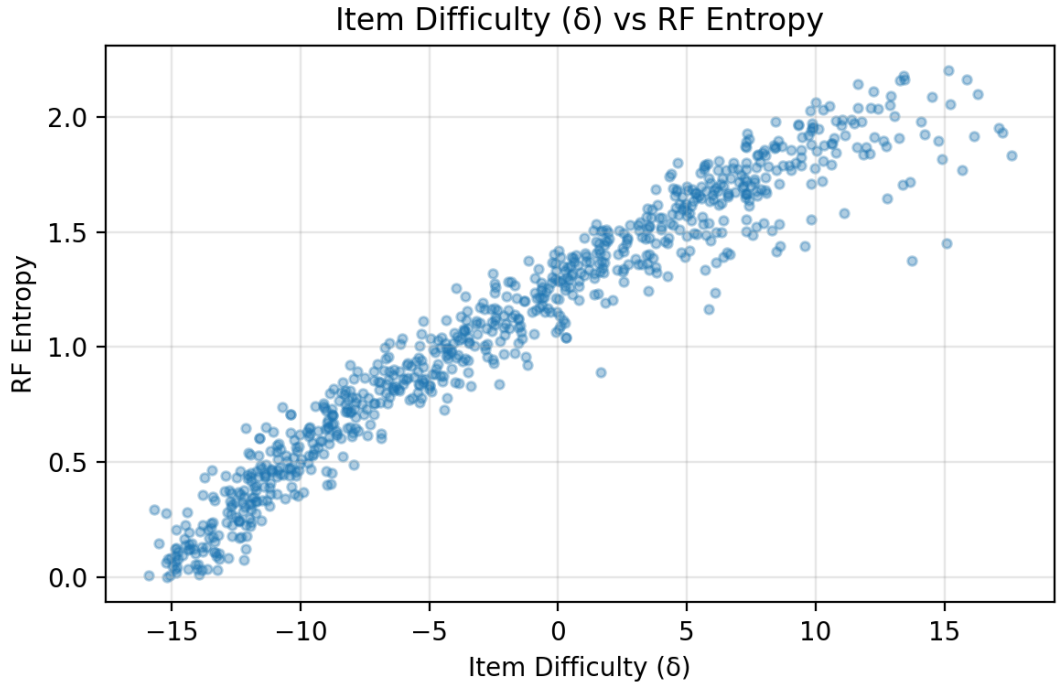
Metric	Value
Train / Val / Test	4000 / 800 / 800
RF test / val / OOB	0.954 / 0.944 / 0.939
Mean margin / entropy	0.5644 / 1.0768
δ ↔ margin (Pearson)	−0.975
δ ↔ entropy (Pearson)	0.970
θ mean ± σ	3.04 ± 0.29
δ mean ± σ	−0.13 ± 0.47

- Ambiguous digits (e.g., brushed 5 vs 6) still spike δ toward the positive tail; elsewhere the forest is decisive.
- Low entropy + high margin line up with low δ, giving a “sanity benchmark” beyond CIFAR.

Study III Diagnostics: δ vs RF Signals



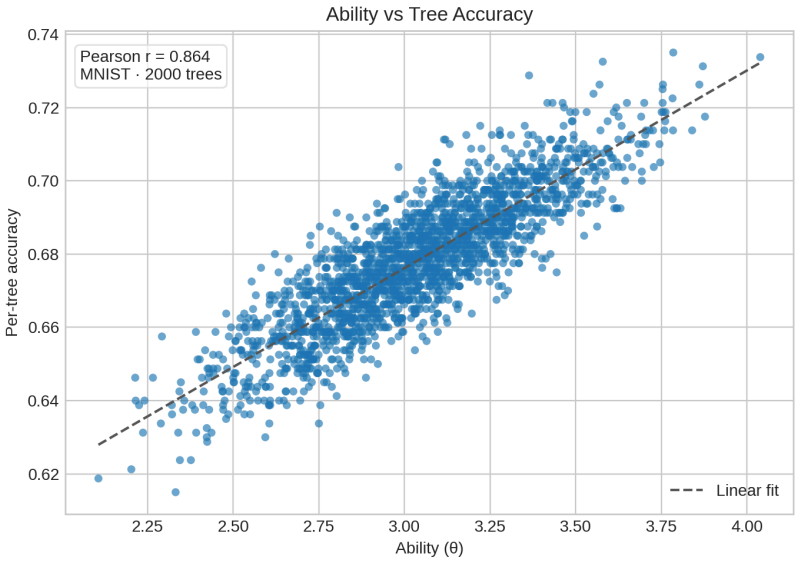
δ vs margin (Pearson −0.97)



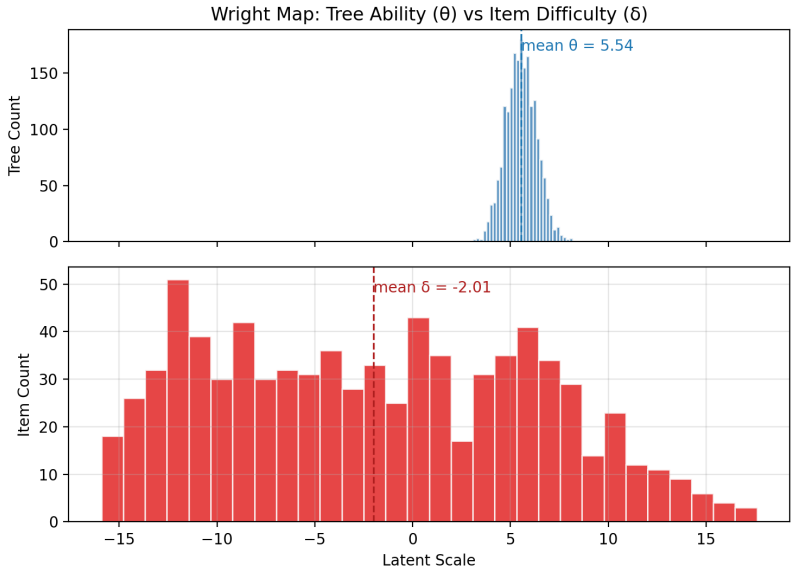
δ vs entropy (Pearson 0.97)

- Clean digits show near-perfect alignment between δ and RF uncertainty.
- Only a handful of δ > 1.2 digits drive the residual uncertainty (stroke collisions like 3/5, 4/9).

Study III Diagnostics: Ability Profiles



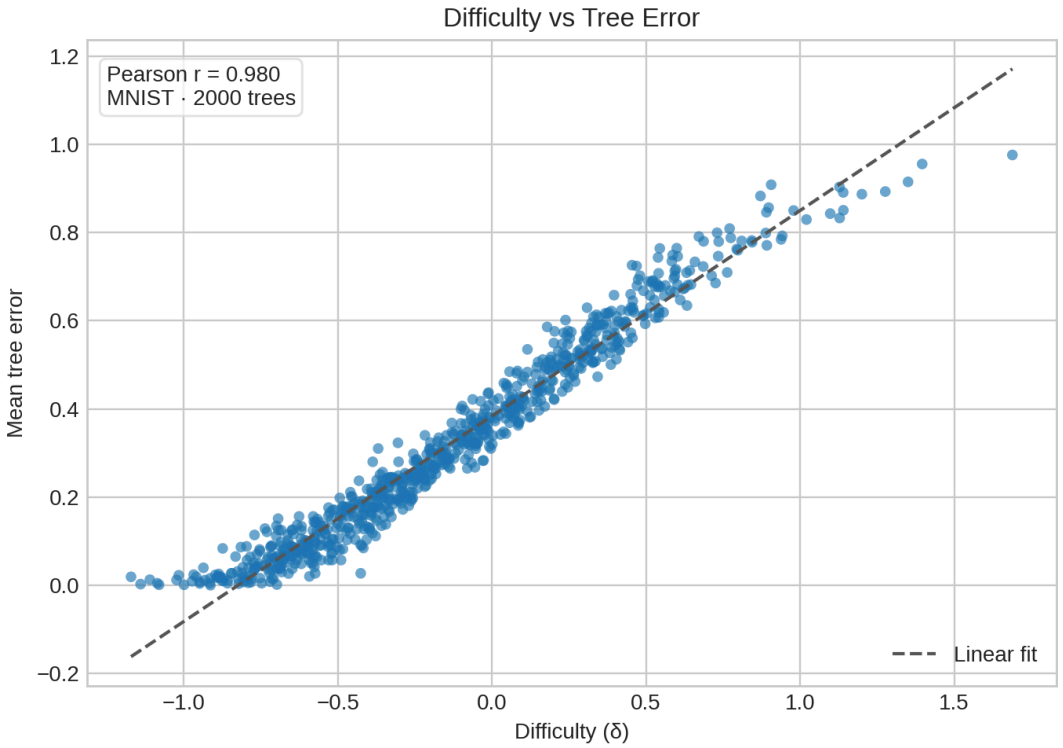
Ability (θ) vs tree accuracy — Pearson 0.98



Wright map: θ mean 3.04 ± 0.29; δ mean −0.13 ± 0.47

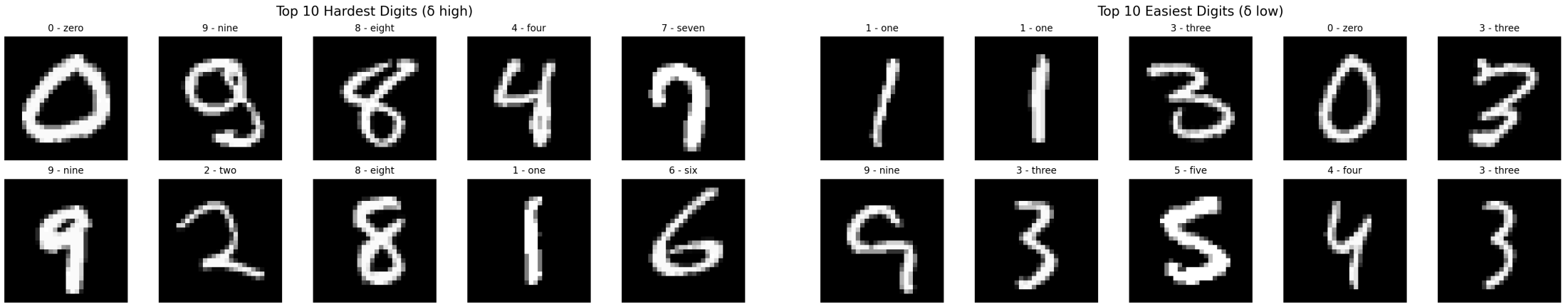
- θ mean 3.04 ± 0.29 shows strong consensus, while δ mean −0.13 ± 0.47 keeps a modest positive tail for ambiguous strokes.
- Shared scales expose plentiful easy wins with only a few sharp spikes—opposite of the CIFAR baseline.

Study III Diagnostics: δ vs Error Rate



- Pearson 0.98 keeps δ tied to mean tree error despite the high accuracy ceiling.
- δ > 1.2 corresponds to stroke-collided 3/5/8 and 4/9 pairs; the long negative tail is trivial for the ensemble.

Study III Evidence: Hard vs Easy Digits



- Hardest digits show stroke collisions (3↔5, 4↔9) that push δ above 1 despite high margins elsewhere.
- Easy digits are crisp, centered strokes—useful anchors when explaining why δ plunges on most of the dataset.

Study III Takeaways

- δ and RF uncertainty agree almost perfectly, while θ stays high yet still flags the rare ambiguous strokes.
- The control study confirms the RF × IRT pipeline holds outside noisy vision data.

Section IV · Cross-Study & Diagnostics

- Compare backbones and datasets on a shared θ/δ scale.
- Surface recurring themes before the close.

Cross-Study Snapshot

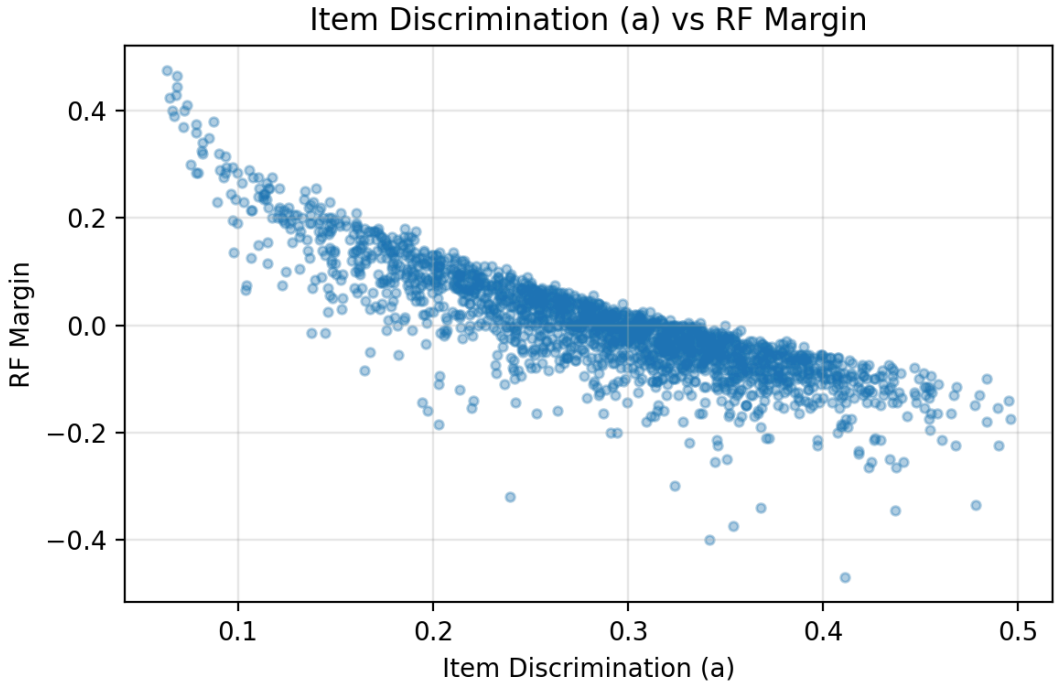
Study	Feature Backbone	Test Acc	δ ↔ margin (Pearson)	δ ↔ entropy (Pearson)	θ σ	δ σ
Study I: CIFAR + PCA-128	PCA-128	0.468	−0.815	0.687	0.154	0.150
Study II: CIFAR + MobileNet	MobileNet-V3 (960-D)	0.819	−0.950	0.881	0.228	0.871
Study III: MNIST Mini	Raw pixels	0.954	−0.975	0.970	0.289	0.472

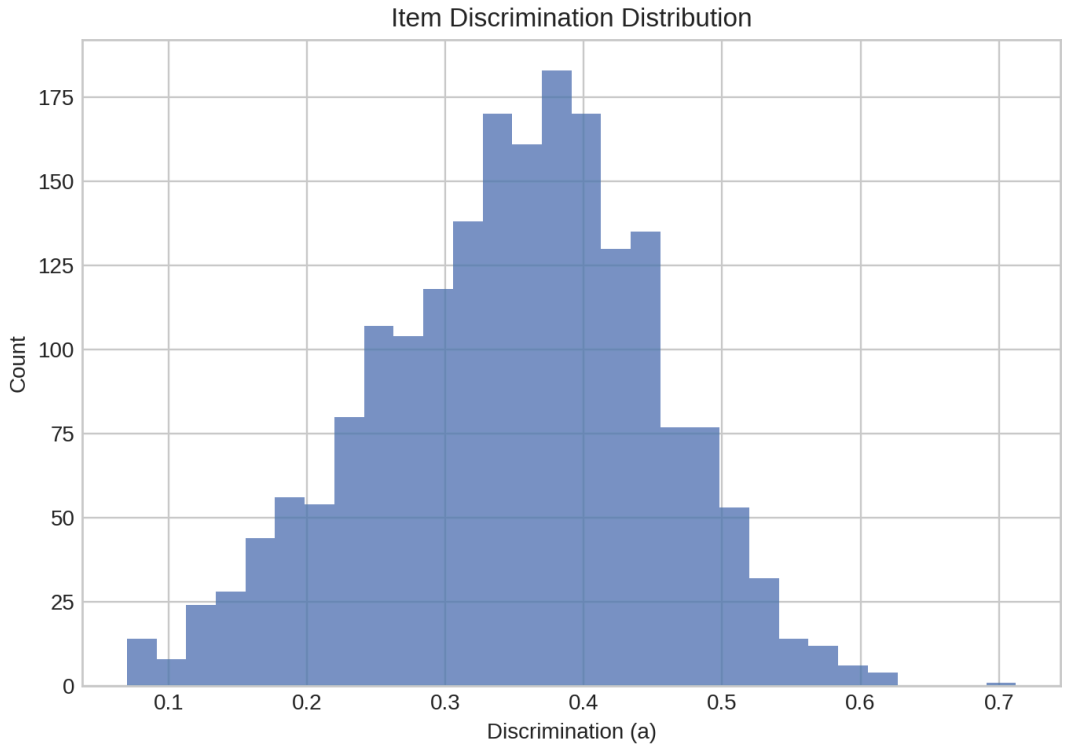
- Feature backbone still shapes δ alignment: PCA lands near −0.82, MobileNet tightens to −0.95, and MNIST saturates the scale at −0.98.
- θ spread remains compact (σθ ≈0.15–0.29) even with 2000 trees; MobileNet widens slightly as headroom grows.
- Difficulty variance balloons on MobileNet (σδ ≈0.87) while MNIST stays moderate, underscoring how rich features surface nuanced “hard” digits.

2PL Discrimination Baseline (CIFAR + PCA)

- 800-epoch 2PL fit (lr 0.02) yields mean (a) ≈ **0.35** with σ ≈ **0.10** (range 0.07–0.71).
- (a) tracks RF uncertainty tightly: Pearson (a\leftrightarrow) margin **−0.83**, (a\leftrightarrow) entropy **0.63**.
- High-discrimination tail isolates the cat/dog ambiguity previously flagged by δ alone.
- Artifacts: `data/irt_parameters_2pl.npz`, `data/rf_irt_correlations_2pl.json`, `figures/2pl_*`, `figures/discrimination_hist.png`.

2PL Diagnostics

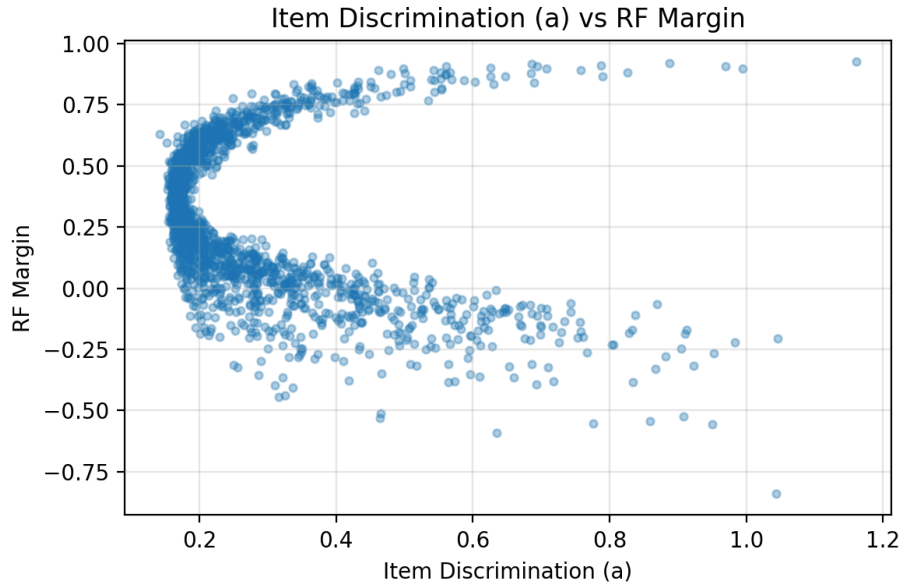


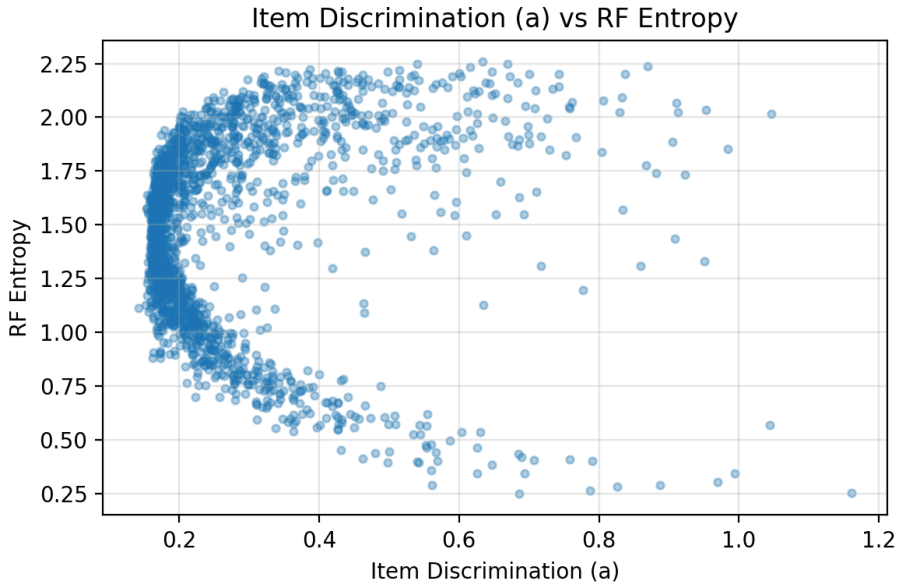


- High-(a) items carry persistently low margins; easy items cluster at high confidence.
- Slope distribution tightens around 0.3, signalling that only a narrow band of items sharply separates trees.

2PL Discrimination (CIFAR + MobileNet)

- Mean (a) settles at **0.27 ± 0.15** with a modest tail (max ≈1.16).
- (a\leftrightarrow) margin **−0.32** and (a\leftrightarrow) entropy **+0.10** keep residual cat/dog confusion in focus while the easy cluster sharpens.
- Artifacts: `data/mobilenet/irt_parameters_2pl.npz`, `data/mobilenet/rf_irt_correlations_2pl.json`, `figures/mobilenet_2pl_*`.

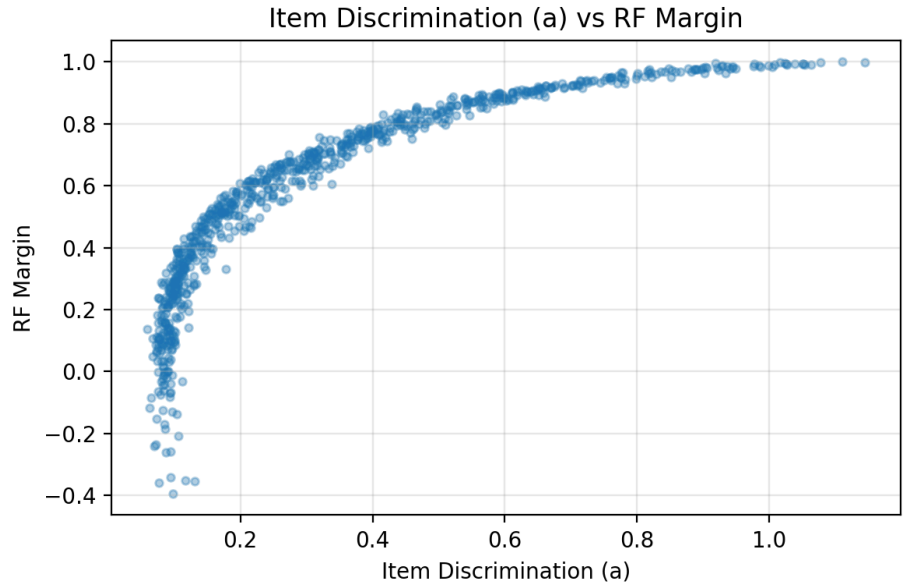


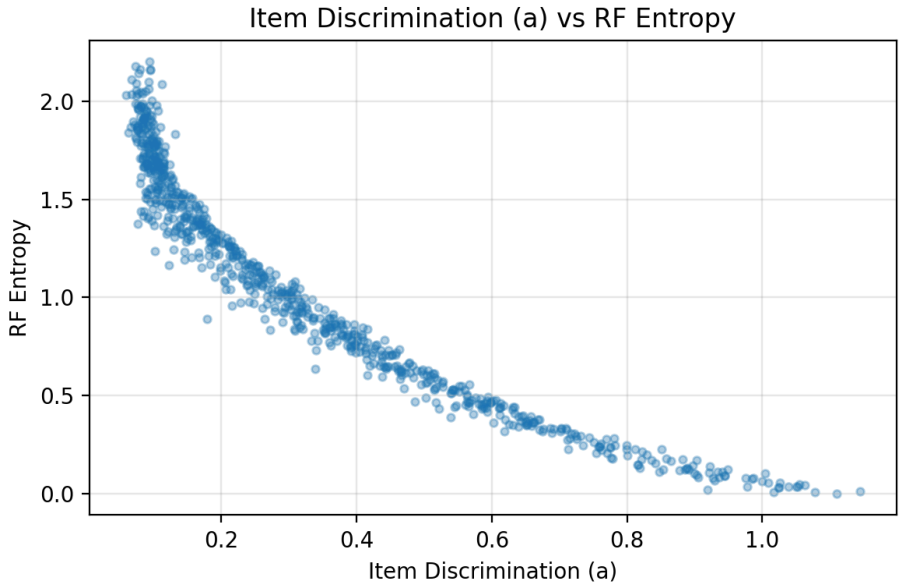


- Even at 81% accuracy, a narrow band of animal images still drives the steepest slopes.

2PL Discrimination (MNIST)

- Mean (a) lifts to **0.24 ± 0.16** because only a few digits truly separate trees.
- (a\leftrightarrow) margin **+0.89** while (a\leftrightarrow) entropy **−0.96** flips sign—uncertainty vanishes outside the awkward strokes.
- Artifacts: `data/mnist/irt_parameters_2pl.npz`, `data/mnist/rf_irt_correlations_2pl.json`, `figures/mnist_2pl_*`.





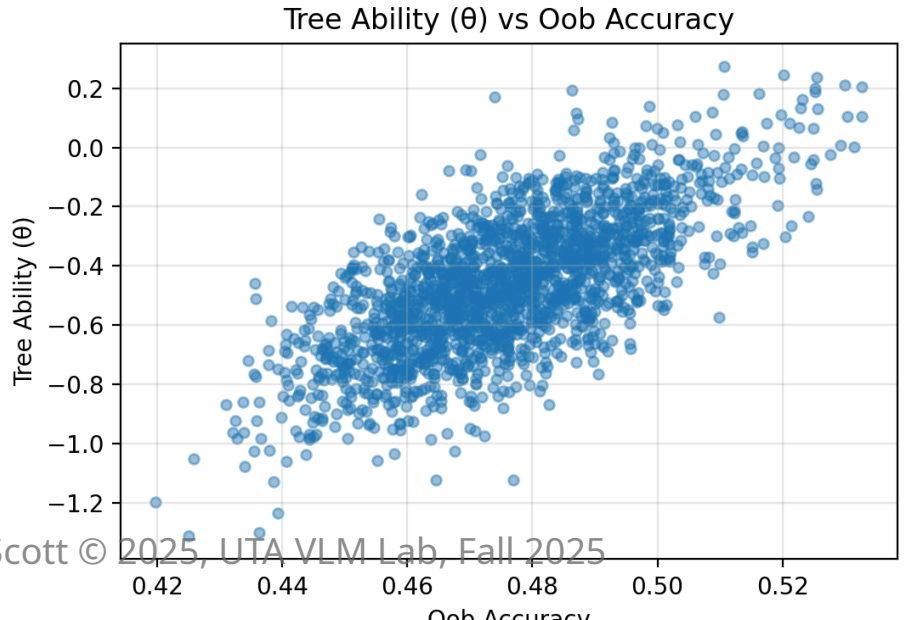
- High-(a) digits align with the stroke collisions spotted earlier (3↔5, 4↔9).

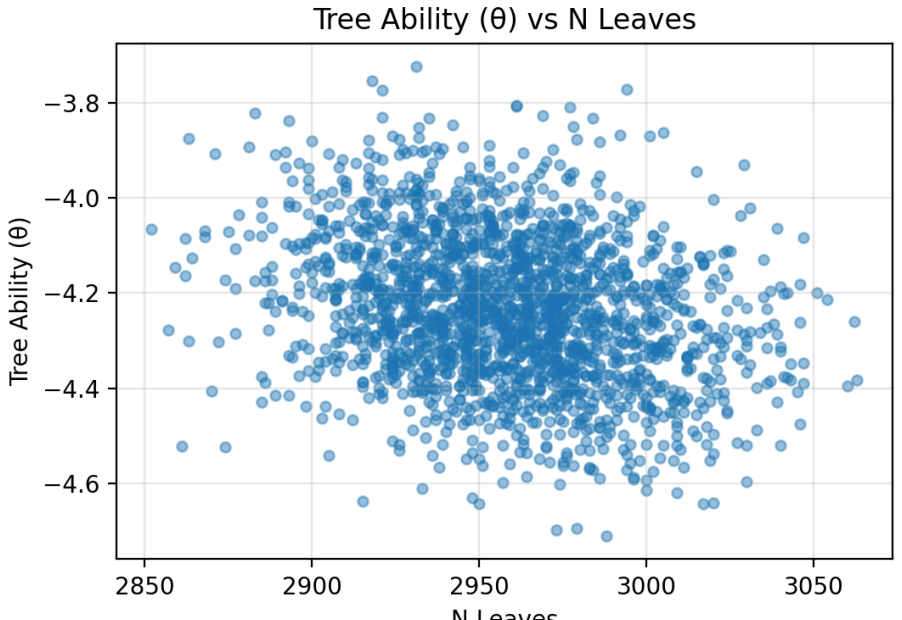
3PL Pilot · MobileNet

- 1k-epoch 3PL run (lr 0.01) converged with guess mean **0.25 ± 0.13**.
- (\theta\leftrightarrow) accuracy Pearson **0.98**; slope mean extends to **0.23** with a wider separation tail.
- High-guess items concentrate on background-heavy aircraft & cats—evidence of latent “guessing” behaviour.

Tree Attribute Correlations

- `scripts/analyze_tree_attribute_correlations.py` merges depth/leaves/OOB stats with (\theta) + discrimination aggregates.
- MobileNet: leaf count ↔ (\theta) Pearson **−0.78**, OOB ↔ (\theta) **+0.75**—shallow, accurate trees shine.
- PCA baseline: leaf count ↔ (\theta) **−0.20**, OOB ↔ (\theta) **+0.28**; MNIST shows similar leaf penalties (−0.47).





Key Takeaways

- IRT and RF still move in lockstep: (\theta) tracks per-tree accuracy, while δ and (a) surface stubborn item pockets.
- MobileNet’s discrimination tail isolates animal confusions despite stronger features; MNIST flips signs because mistakes are rare.
- 3PL adds a modest guessing floor (~0.25) without upsetting (\theta)–accuracy alignment.
- Tree attributes expose pruning cues: shallow, high-OOB trees consistently land higher (\theta).

Next Steps

- Fold discrimination stats into `reports/embedding_comparison.md` & deck tables for quick grabs.
- Run stability sweeps (50/100 trees, alternate seeds) to quantify variance in (a) and (\theta).
- Decide whether 3PL merits extension to PCA/MNIST or documenting as MobileNet-only.
- Finish item-tier overlays (high/medium/low (a)) and align them with the qualitative grids.