Knowledge That Is Linear Before Fine-Tuning Generalizes After It

When can recalled facts support learning a new relation?

Paper by Faridoun Mehri and Hinrich Schütze
Research group reading presentation

arXiv:2610.22300v2 (19 Oct 2026) • Reading: 27 Oct 2026

Takeaway: linear readability of old relations predicts transfer after fine-tuning.

No formal publication found as of 27 Oct 2026 • Official code and data

A recalled fact may still fail to support new learning

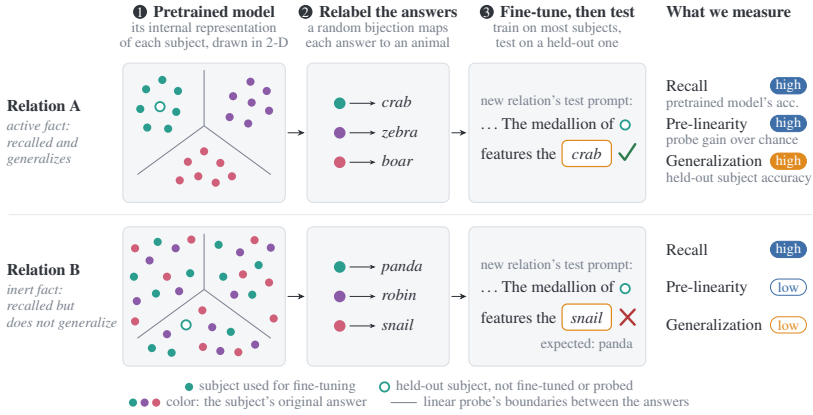
Research question

Which property of a pretrained model's knowledge predicts whether new knowledge built on it generalizes?

- **Recall:** the model can answer an existing relation, such as city $\rightarrow$ continent.
- **Active knowledge:** the relation's shared structure supports a new task on unseen subjects.
- **Hypothesis:** linearly readable answers in pretrained subject states support that transfer.

Paper, Secs. 1–2, PDF pp. 1–4. "Active" and "inert" are task-relative operational terms.

Random relabeling isolates the old relation's structure



Paper, Fig. 1, PDF p. 2. Original conceptual illustration; open circles are unseen subjects.

Train on some subjects; test on different subjects

Constructed relation

$$R^{\pi}(u) = \pi(e_R(u)), \qquad \pi : \mathcal{E} \longrightarrow L$$

$e_R(u)$: old answer; L: random labels; π: random bijection.

- Europe $\mapsto$ crab; Asia $\mapsto$ zebra.
- Paris $\mapsto$ crab is seen in fine-tuning.
- Munich $\mapsto$ crab is tested only if Munich is held out.

Paper, Sec. 3, PDF pp. 4–5. The mapping is withheld from the model.

Subject split

T: fine-tuning and probe subjects

H: held-out evaluation subjects

At least four T and one H subject per old answer. H is never used to fit or select the probe.

The three measures answer different questions

Measure	Model and subjects	What is scored
Recall	Pretrained model; T	Old answers under old relation prompts
Pre-linearity	Pretrained model; T	Cross-validated linear probe of old answers, minus a majority baseline
Generalization	Fine-tuned model; H	New labels under a held-out prompt template

Common metric: class-balanced candidate accuracy

Top-1 over candidate first tokens, averaged equally across answer classes. Chance is $1/|\mathcal{E}|$.

Paper, Sec. 3, PDF p. 5. Evaluation is candidate-restricted first-token classification.

Pre-linearity probes a compact subject representation

Probe input and model

$$z_k = \frac{v_k^\top (h - \mu)}{\sigma_k}, \quad k = 1, \ldots, 16.$$

$$p(e \mid u) = \mathrm{softmax}(Wz + b)_e.$$

- h: state of the subject span's last token in an old-relation query.
- PCA gives 16 directions; whitening scales each to unit variance.
- Fit cross-entropy plus $10^{-2}\|W\|^2/2$.

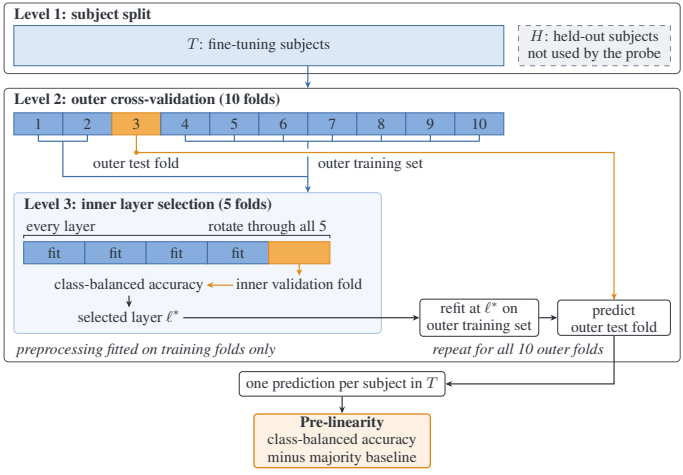
Paper, Sec. 3, PDF p. 5. PCA and whitening are fitted inside each training fold.

Interpretation

Can a simple readout separate subjects by their old answers before any new-task fine-tuning?

The same dimension and regularization are fixed for every model and relation.

Nested cross-validation keeps layer choice inside training data



Paper, Sec. 3, PDF pp. 5–6; Fig. 5, PDF p. 14. Original figure cropped from PDF.

The test covers 70 relations and 11 language models

Source	n	Example
Hernandez	11	company $\rightarrow$ HQ
Town	6	town $\rightarrow$ latitude band
BEAR	53	place $\rightarrow$ continent
Total	70	≥ 5 answers each

- 11 open-weight models, about 2B–27B parameters; base and post-trained checkpoints.
- Main labels: 179 curated animal names.
- Each relation has at least 25 fine-tuning subjects.
- Subject frequency uses Dolma v1.7 counts as a pretraining-exposure proxy.

Paper, Sec. 4, PDF pp. 6–8; Appendix E, PDF pp. 21–22.

All models learn the observed new-relation facts

Fine-tuning recipe

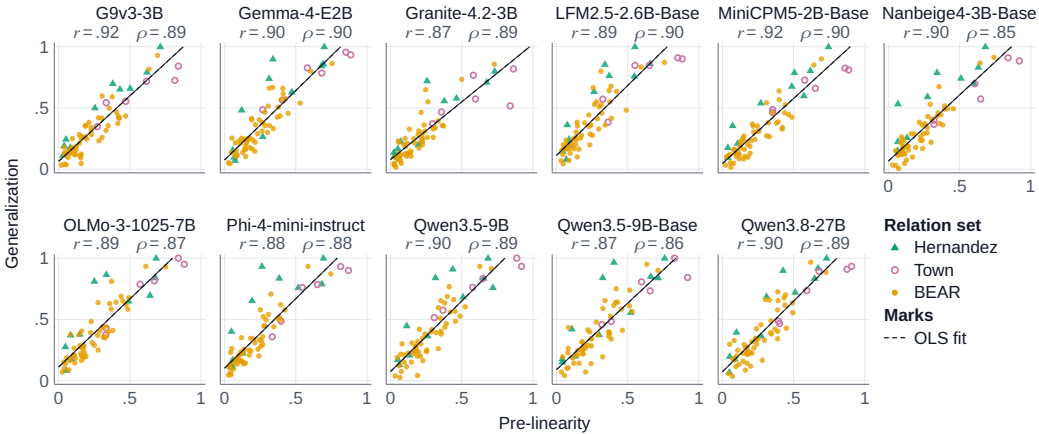
Full-parameter updates; four fictional-story templates per training subject; cross-entropy only on the new label's first token; 20 epochs.

- Each subject–label fact appears 80 times in total: 20 epochs $\times$ 4 wordings.
- Evaluation uses a fifth, held-out wording and unseen subjects H.
- Every reported model reaches 1.00 candidate accuracy on its training subjects.

The remaining question is generalization to H, rather than fitting the observed mapping.

Paper, Sec. 3, PDF p. 5; Appendix B and Table 3, PDF pp. 14–15.

Higher pre-linearity tracks higher held-out generalization



Paper, Fig. 2, PDF p. 6. One point per relation in each model; Pearson $r = 0.87$–0.92 across 70 relations.

Four controls reduce the association but do not erase it

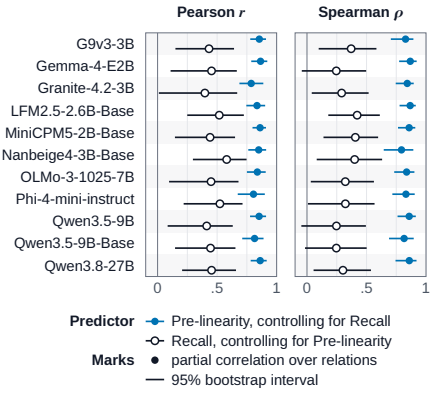
Within-model association across relations	Pearson r	Spearman ρ
Raw	0.87–0.92	0.85–0.90
Controlling for old-relation Recall	0.79–0.86	0.79–0.87
Jointly controlling for Recall, answer count, training-subject count and subject frequency	0.70–0.82	0.75–0.85

Statistical unit

Correlations are computed across the 70 relations separately within each of the 11 models. Intervals bootstrap relations, with controls refitted in each resample.

Paper, Table 1, PDF p. 7; Appendix H, PDF p. 24. Numbers are coefficients, not accuracy gains.

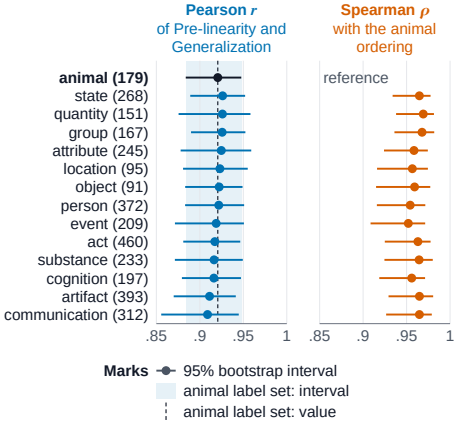
Pre-linearity retains more signal than Recall



- Pre-linearity | Recall: Pearson $r = 0.79$–0.86.
- Recall | Pre-linearity: Pearson $r = 0.40$–0.58.
- Recall carries some independent association, but less than Pre-linearity.

Paper, Sec. 5 and Fig. 3a, PDF pp. 8–9. Dots: partial correlations; bars: 95% bootstrap intervals.

Changing the labels leaves the pattern intact



- Rerun all 70 relations with 13 other WordNet label categories.
- Pre-linearity–Generalization Pearson $r = 0.91$–0.93.
- Relation ranking agrees with animal labels: $\rho = 0.95$–0.97.

Scope

This category comparison is on MiniCPM5-2B-Base only.

Paper, Sec. 5 and Fig. 3b, PDF pp. 8–9; label sets in Table 7, PDF p. 22.

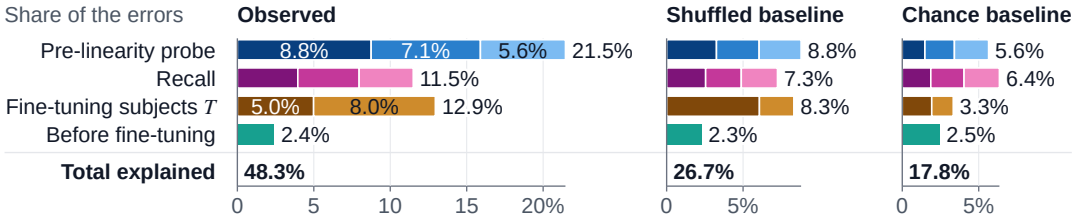
New mappings and seeds produce similar correlations

MiniCPM5-2B-Base, 70 relations	Run 0	Run 1	Run 2
Pearson r	0.92	0.93	0.92
Spearman ρ	0.90	0.89	0.92

- Each run redraws the answer-to-label bijection and the fine-tuning seed.
- On 17 Hernandez and Town relations, repeats across ten models yield Pearson $r = 0.78$–0.95.
- Subject split, label subset and templates remain fixed across repeats.

Paper, Table 2, PDF p. 9; Appendix D, Tables 4–5, PDF p. 20.

Wrong labels often match the probe's preferred old answer



- Probe top-3: 21.5% of errors versus 8.8% after shuffling wrong labels.
- Pretrained model top-3: 11.5% versus 7.3% shuffled.

Paper, Fig. 4, PDF p. 9. Shares use equal split when explanations coincide; baselines are shuffled and chance errors.

The error analysis links the correlation to individual subjects

Diagnostic sequence

Wrong new label $\xrightarrow{\pi^{-1}}$ old answer $\xrightarrow{\text{compare}}$ probe and pretrained-model rankings.

- 37,571 held-out errors from 757 model–relation fine-tunes that made an error.
- Probe rankings receive more attribution than the old-relation model rankings.
- All considered explanations cover 48.3% of errors, versus 26.7% after shuffling.

Interpretation: errors preferentially follow structure readable by the probe.

Paper, Sec. 5, PDF p. 9; Appendix G, PDF pp. 23–24. Attribution depends on the named explanations and share rule.

The evidence supports a representation-based predictor

- The old relation's linearly readable answer structure is measurable before fine-tuning.
- It predicts transfer to unseen subjects across diverse tested relations and models.
- Recall, relation size and a frequency proxy do not account for the full association.

Scope of the inference

English relations, arbitrary word labels, full-parameter fine-tuning and candidate-restricted first-token evaluation. Correlations and error alignment do not identify a unique causal mechanism.

Paper, Secs. 3–6, PDF pp. 4–10. The scope statement is our interpretation of the reported design.

Take-home: measure usable structure, not just recall

1. **Design:** random relabeling turns an old relation into an unseen transfer test.
2. **Finding:** Pre-linearity predicts Generalization in all 11 models ($r = 0.87$–0.92).
3. **Mechanistic clue:** wrong new labels align with the probe's old-answer ranking.

Question for discussion

Would improving linear readability of an existing relation causally improve transfer to a new one?

Paper, Table 1, PDF p. 7; Fig. 4, PDF p. 9. Discussion question is a proposed follow-up.

Paper and resources

- Faridoun Mehri and Hinrich Schütze. *Knowledge That Is Linear Before Fine-Tuning Generalizes After It*. arXiv:2610.22300v2, 2026.
- Versioned paper PDF and official experiment code and data.
- The arXiv record lists no journal reference or conference as of 27 Oct 2026.

Questions and discussion

Backup: exact model-level Pearson coefficients

Model	Raw r	Joint-control r
G9v3-3B	.92	.82
Gemma-4-E2B	.90	.80
Granite-4.2-3B	.87	.70
LFM2.5-2.6B-Base	.89	.76
MiniCPM5-2B-Base	.92	.82
Nanbeige4-3B-Base	.90	.77
OLMo-3-1025-7B	.89	.80
Phi-4-mini-instruct	.88	.75
Qwen3.5-9B	.90	.77
Qwen3.5-9B-Base	.87	.72
Qwen3.8-27B	.90	.79

Paper, Table 1, PDF p. 7. Joint controls: Recall, answer count, training-subject count and subject frequency.