Data Distributional Properties Drive Emergent In-Context Learning in Transformers

How training data shape the strategy a model learns

Paper discussion of Chan et al. (2022)

arXiv:2205.05055v6 | Read on 29 September 2026

Local repetition and global rarity can favor learning from context.

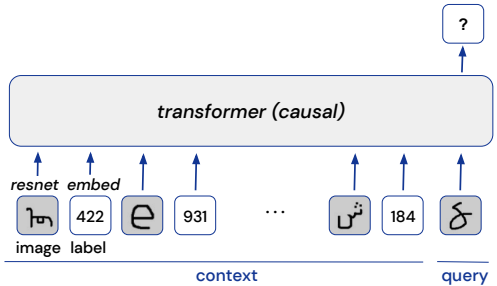
The same training answer can come from two strategies

In-context learning (ICL)	In-weights learning (IWL)
Use examples in the current sequence to assign a label to the query.	Recall an image–label association acquired through training updates.
No weight updates during the test episode.	No matching query class is needed in context.

Question: Which training distributions favor each strategy?

Paper, Sec. 1, PDF pp. 1–3. ICL ability itself is acquired through training.

A controlled image task isolates data properties



- Omniglot: 1,623 classes, 20 images per class.
- Eight image–label pairs, then one query: 17 tokens.
- ResNet + label embeddings + causal Transformer.

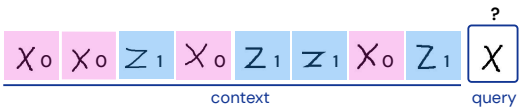
$$\mathcal{L}(\theta) = -\log p_\theta(\ell_q \mid \mathcal{C}, u_q)$$

Only the query label is supervised. $\mathcal{C}$: context; u_q: query image.

Paper, Fig. 1a, p. 2; Sec. 2, p. 4. Equation is our notation for the stated loss.

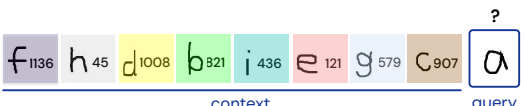
Two probes separate context use from stored associations

ICL: unseen image classes



- Two classes, four examples each.
- Random labels 0/1 per episode.
- Compare only outputs 0 and 1; chance: 50%.

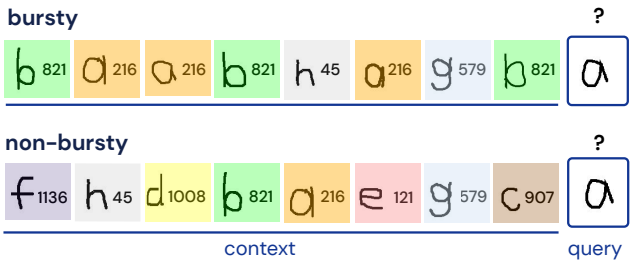
IWL: trained image classes



- Query class absent from context.
- Training labels retained.
- Full output space; usual chance: 1/1600.

Paper, Fig. 1c–d, p. 2; Sec. 2.3, pp. 4–5. Labels 0/1 are familiar, not novel.

Burstiness makes the query class locally available

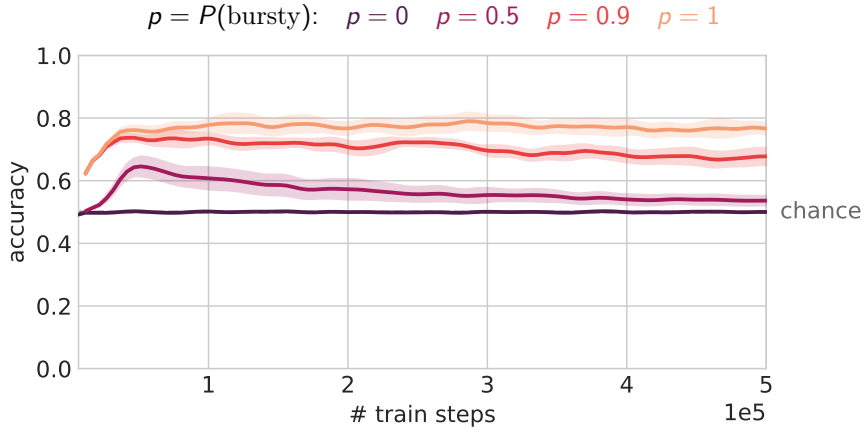


- Bursty: query class occurs three times in context.
- Another class also occurs three times.
- $p(\text{bursty})$ mixes entire sequence types.

Default labels stay fixed across training sequences: either strategy can work.

Paper, Fig. 1b, p. 2; Sec. 2.1, p. 4. Non-bursty sequences use uniform random sampling.

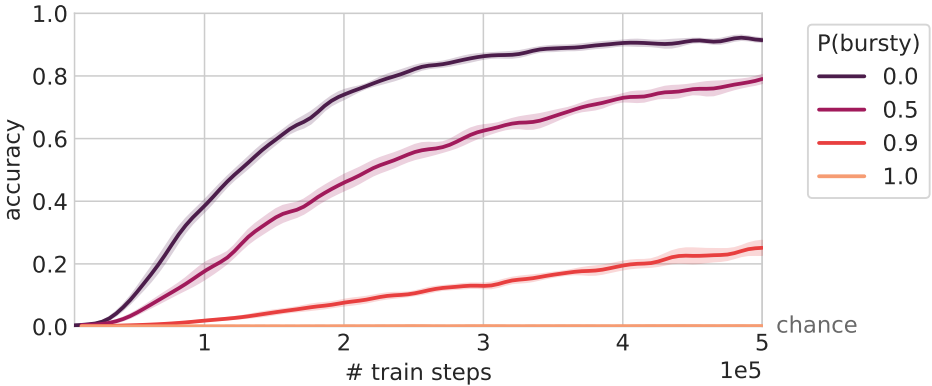
More burstiness produces stronger ICL



Without burstiness, held-out ICL stays near chance.

Paper, Fig. 2a, PDF p. 3. Mean with standard deviation; 5 runs. Chance: 50%.

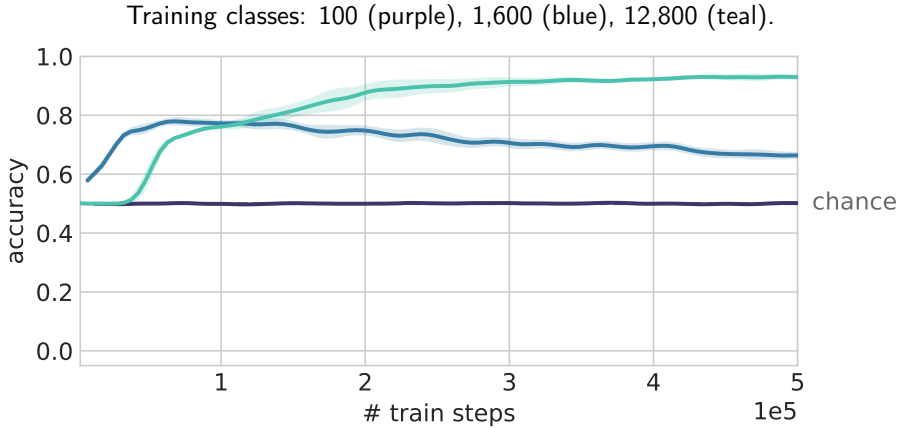
The same intervention reduces parameter-based recall



Longer training can shift the balance toward IWL.

Paper, Fig. 2b, PDF p. 3; Sec. 3.1, p. 5. Full-output IWL; query unsupported in context.

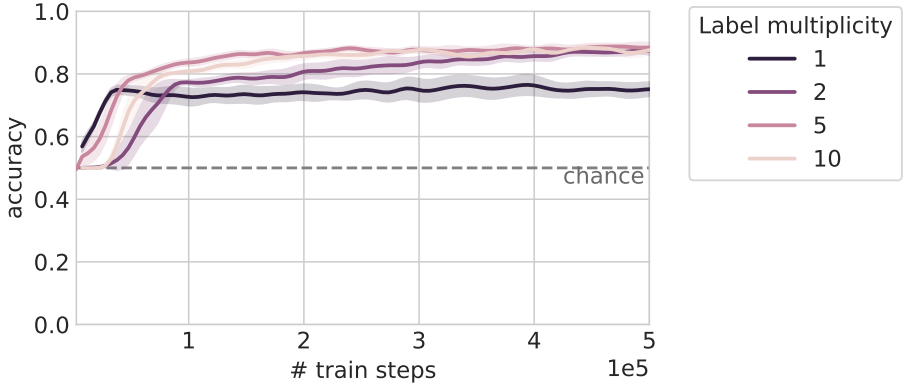
A larger class vocabulary favors ICL



$p(\text{bursty}) = 0.9$; IWL moves in the opposite direction.

Paper, Fig. 3, PDF p. 4. The 12,800-class augmentation may also introduce label ambiguity.

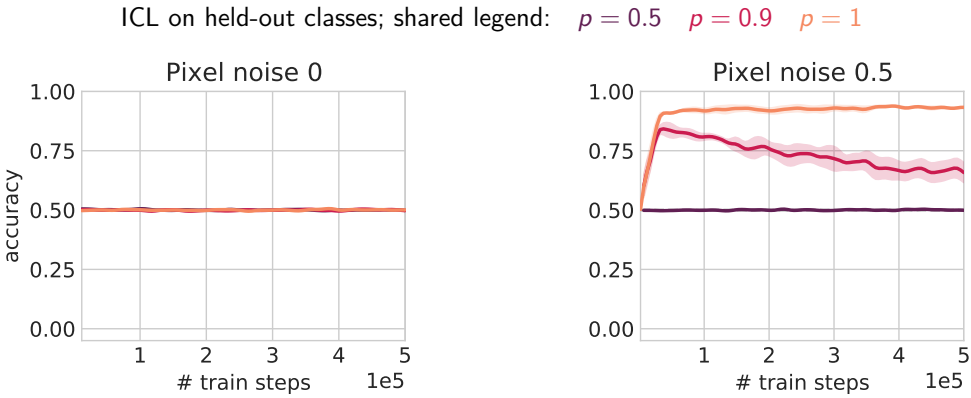
Locally stable, globally variable labels strengthen ICL



A class has 1, 2, 5, or 10 possible labels; one label is used within each sequence.

Paper, Fig. 4, PDF p. 6; Sec. 3.1, pp. 5–6. Fixed burstiness: $p = 0.9$.

Image variation can discourage memorization



Full Omniglot is harder again: more variation is not always better.

Paper, selected panels from Fig. 5a, PDF p. 6. 3 seeds; shading: standard deviation.

Zipf frequencies create a common head and a rare tail

Class rank controls sampling frequency

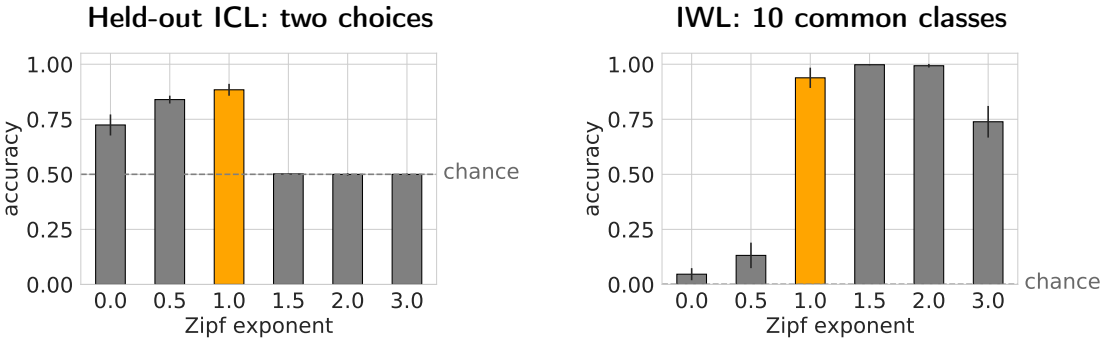
$$p(X = x) \propto x^{-\alpha}, \qquad \alpha \geq 0$$

- X is class rank; rank 1 is the most common.
- $\alpha = 0$: uniform. Larger α: a more concentrated head.
- Hypothesis: common classes support IWL; rare classes favor ICL.

Experiment: 12,800 classes, $p(\text{bursty}) = 0.9$; separately test IWL on the top 10 classes and the rest.

Paper, Sec. 3.2, Eq. (1), PDF pp. 7–8. Hypothesis is a behavioral interpretation.

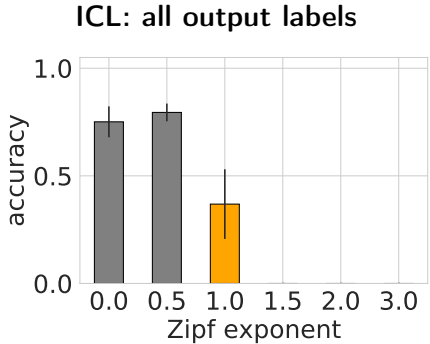
At $\alpha = 1$, two diagnostic abilities coexist

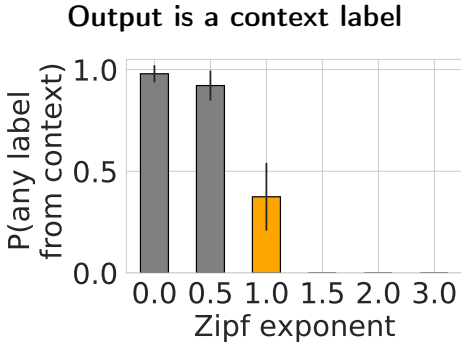


- Rare-class IWL remains near chance across tested exponents.
- $\alpha = 1$ works well here; its optimality is not universal.

Paper, Fig. 6c–e, p. 7. Mean over steps 400k–500k; 3 seeds; error bars: standard deviation.

Full-output evaluation reveals a selection problem

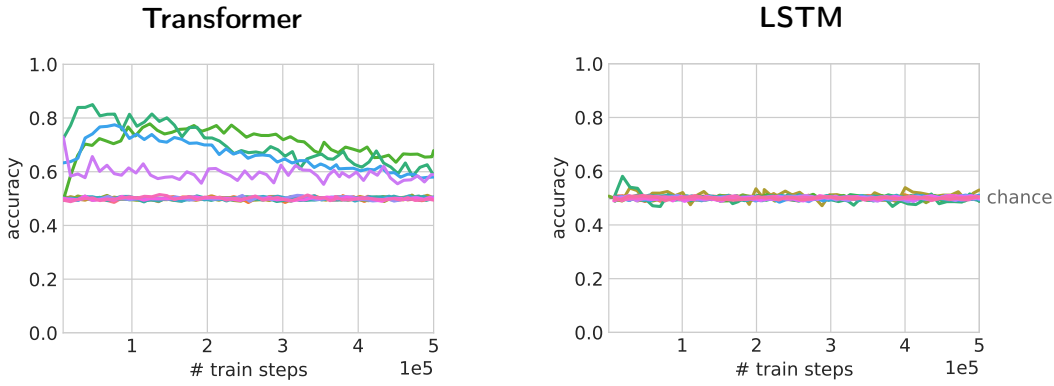




At $\alpha = 1$, full-output ICL drops sharply: the model often selects a label outside the context.

Paper, Appendix C.4, p. 16; Figs. 10e and 11, p. 18. Same model, different scoring.

Tested recurrent models fail to acquire comparable ICL



- Vanilla RNN also stays near chance (Fig. 7b, not shown).
- 1,600 classes, $p = 0.9$; colors identify runs, not matched settings.

Paper, Fig. 7a,c and Sec. 3.3, p. 8. Chance: 50%. Inference restricted to tested models.

Reported parameter counts are not exactly matched

Depth	Transformer	LSTM
12 layers	831,479	627,959
2 layers	331,639	297,719

- Main text and captions say parameter counts were matched.
- Appendix C.1 gives the unequal counts above; RNN counts are not listed there.
- Search: 15 runs per architecture/depth, 90 runs in total.

Paper, Sec. 3.3, p. 8, versus Appendix C.1, p. 15. This discrepancy is our reading note.

Data properties change the relative value of each strategy

Intervention	ICL	IWL
More local repetition	Increases	Decreases
More training classes	Increases	Decreases
More labels per class	Increases	Not plotted in Fig. 4
Within-class variation	Often increases	Often decreases
Moderate Zipf skew	Can remain high	High on common classes

Interpretation: context is useful when local evidence is reliable and globally memorizing every association is difficult.

Synthesis of Figs. 2–6, pp. 3–7. Trends depend on conditions; variation is not monotonic.

The evidence has four practical boundaries

1. **Task:** short image sequences and query classification, rather than a language-model training experiment.
2. **Metric:** two-choice ICL does not guarantee correct full-output predictions or novel-label generation.
3. **Interventions:** rotations can add label ambiguity; image variation also changes visual difficulty.
4. **Mechanism:** behavioral results do not identify the circuits implementing either strategy.

Paper, Secs. 2–4; Appendices B and C.4. Mechanistic boundary is our assessment.

Three findings to retain

Data and architecture jointly shape ICL

The Transformer architecture alone is insufficient in the tested distributions.

Global rarity can coexist with local repetition

Rare classes can still supply repeated evidence within a context window.

Coexistence leaves output selection unresolved

A model can do well on both probes yet often choose an irrelevant label.

Paper, Sec. 4, pp. 8–10; Appendix C.4, p. 16. Conclusions scoped to this task.

Discussion: which control would clarify the mechanism?

1. Hold class exposure and label vocabulary fixed while changing local repetition and tail frequencies separately.
2. Change the frequency ranks of the two test labels; report both restricted and full-output accuracy.
3. Match parameter count and compute more tightly; track when ICL yields to parameter recall during training.

These are proposed follow-ups, not experiments reported in the paper.

Discussion prompts from this reading, motivated by Secs. 3–4 and Appendix C.

Reference and reading materials

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya K. Singh, Pierre H. Richemond, James L. McClelland, and Felix Hill.

Data Distributional Properties Drive Emergent In-Context Learning in Transformers.
NeurIPS 2022. arXiv:2205.05055v6, revised 17 November 2022.

arxiv.org/abs/2205.05055v6
Author code: deepmind/emergent_in_context_learning

Read: full main text, Appendices A–C, and Figs. 1–11.
Materials: matched-version PDF, author TeX, and original figures.
This presentation is a reading analysis; no experiments were rerun.

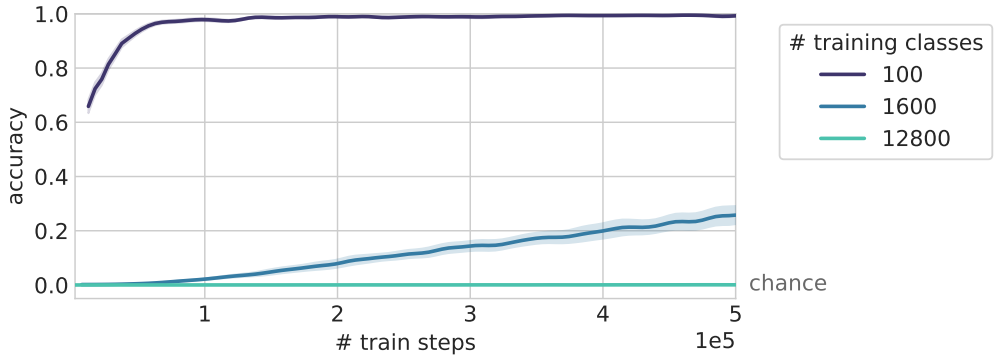
Author names follow the PDF. Venue information follows the arXiv record.

Backup: training details and uncertainty

- Transformer: 12 layers, width 64, 8 attention heads.
- ResNet: 2 blocks per group; channels (16, 32, 32, 64); trained from scratch.
- Adam, 500k steps, 16 TPU v2/v3 cores.
- Warm up to 3×10^{-4} at step 4,000, then inverse-square-root decay.
- Figs. 5–6: 3 seeds. Other standard experiments: 5 runs. Error bars show standard deviation.
- Architecture search: maximum learning rate $[10^{-5}, 0.1]$; warmup $[1, 10000]$, log-uniform sampling.

Paper, Appendices A and C.1, PDF p. 15. Search uses 90 runs, not a 225-point grid per model.

Backup: more classes also reduce IWL

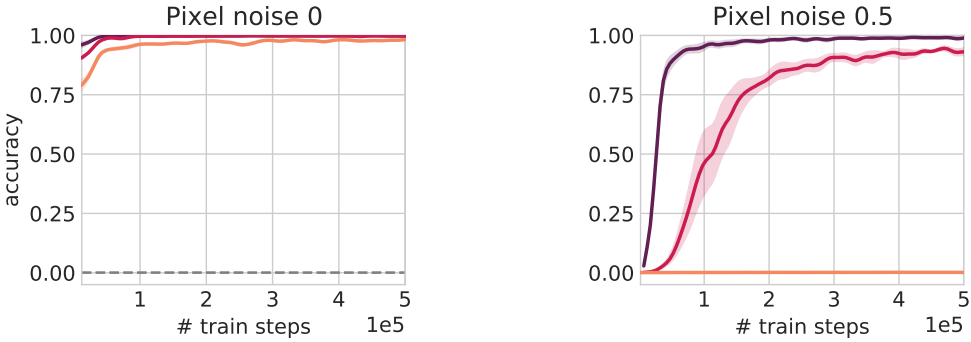


At fixed burstiness, the small vocabulary is much easier to memorize.

Paper, Fig. 3b, PDF p. 4. Fixed $p(\text{bursty}) = 0.9$; 5 runs.

Backup: image noise can suppress IWL

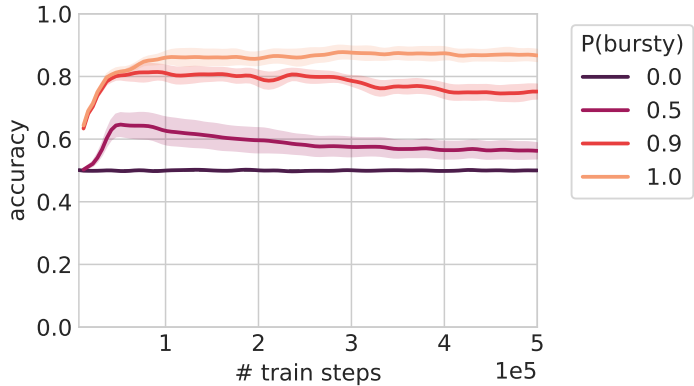
IWL on trained classes; shared legend: $p = 0.5$ $p = 0.9$ $p = 1$



The effect is strongest at high burstiness; low-burstiness models still memorize.

Paper, selected panels from Fig. 5b, PDF p. 6. 3 seeds; shading: standard deviation.

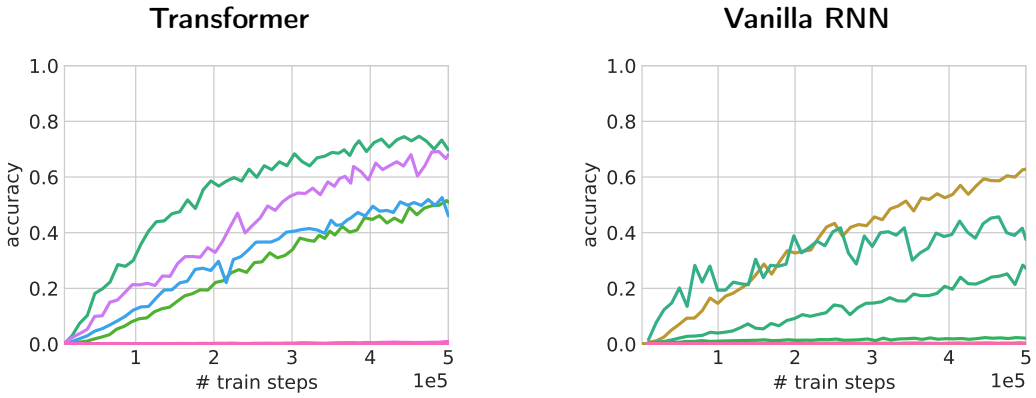
Backup: familiar classes can also be relabeled in context



Training classes are randomly mapped to 0/1 at evaluation. Trends resemble held-out ICL, with slightly higher performance.

Paper, Fig. 9a, PDF p. 17; Appendix C.3, pp. 15–16. Other interventions are in Fig. 9b–d.

Backup: Transformers also perform well on IWL



Neither recurrent model has superior IWL; the LSTM comparison is in Fig. 8c. Each colored line represents a run.

Paper, Fig. 8a–b, PDF p. 16; Appendix C.2, p. 15. Parameter-count caveat still applies.

Backup: a restricted correct answer can still lose globally

Let the correct context label be 1, with logits

$$z_0 = 1, \qquad z_1 = 2, \qquad z_7 = 5.$$

Two-choice evaluation

$$\arg \max_{j \in \{0,1\}} z_j = 1$$

Correct: the model ranks the relevant pair well.

Full-output evaluation

$$\arg \max_j z_j = 7$$

Incorrect: another familiar label wins.

Teaching example, not paper data. Illustrates Sec. 2.3, p. 5, and Appendix C.4, p. 16.