NeurIPS 2026

Why Transformers Struggle with Distribution-Independent In-Context Learning

Omar Naim[1] Jérôme Bolte[2] Nicholas Asher[1]

[1]Université de Toulouse [2]Toulouse School of Economics

In-context learning

Adapt to a task from examples in the prompt, without updating the weights.

$$\underbrace{(1,2),\ (2,4),\ (3,6)}_{\text{in-context examples}} \qquad \underbrace{5}_{\text{query}} \longmapsto \underbrace{10}_{\text{desired answer}}$$

Illustrative linear task: $g(x) = 2x$.

The ability is learned during training.

A common hypothesis: the forward pass implicitly runs a learning algorithm on the examples.

Two notions of in-context learning

In-distribution ICL (ICL_1)

Accurate on new tasks drawn from the training distributions.

Distribution-independent ICL (ICL_2)

Accurate for any identifying context and query, within the same function class.

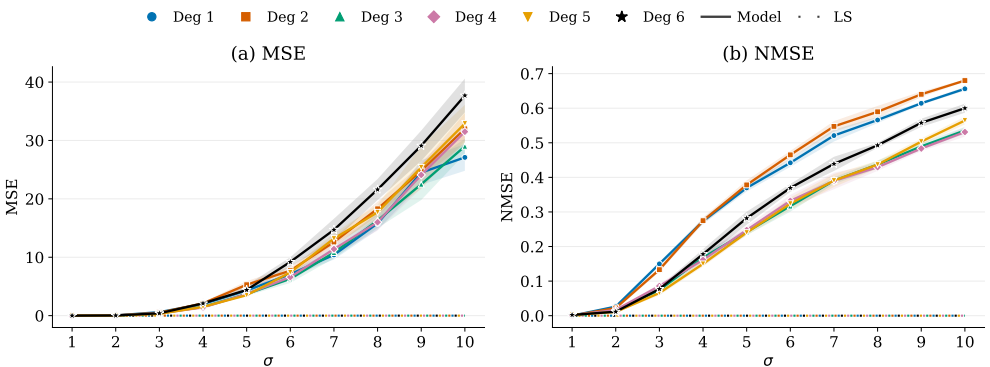
A setting where the context determines the answer

- Noiseless polynomials of degrees 1–6.
- $n+1$ distinct examples identify a polynomial of degree at most n.
- Train from scratch, with inputs and coefficients from $\mathcal{U}(-1, 1)$.

12 layers, 8 attention heads, hidden dimension 256.

Error grows under distribution shift

Inputs stay in $[-1, 1]$. Test coefficients follow $\mathcal{U}(-\sigma, \sigma)$.



Polynomial least squares stays exact with the same context.

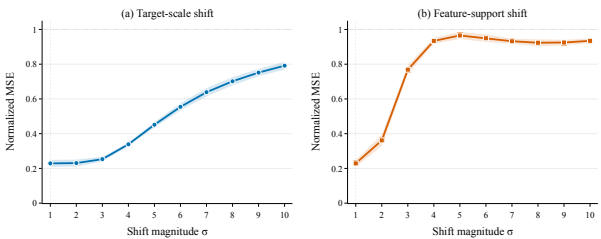
Could this come from training?

Training interventions

- Longer training
- Structured supervision
- Degree curricula
- Broader training support
- Nonlinear encoder and decoder

All keep the cliff. Broader support only delays it.

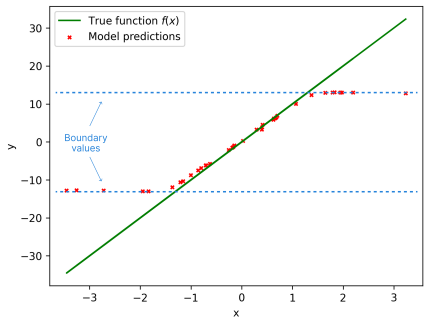
Beyond scalar polynomials



10-dimensional nonlinear regression (TabPFN-style prior): the same degradation, and the same boundary values.

The cliff persists across every training change we tested.

The predictions point to final normalization



1. A pattern

Under shift, predictions stop following the target and plateau at fixed values, whatever the context.

2. An intuition

Plateaus suggest a bounded output, which points to normalization.

3. A proof

We show, empirically and mathematically, that final normalization imposes a fixed bound on the prediction.

Is normalization hiding a correct rule?

Hypothesis: the model computes the right fitting rule internally, and final normalization only clips its output.

Standard model: MSE 9×10^{-6} at $\sigma = 1$, 27.06 at $\sigma = 10$.

1. Remove it after training

MSE	$\sigma = 1$	$\sigma = 10$
Norm removed	434	1558

Accuracy collapses even in distribution: the decoder never saw unnormalized inputs.

2. Retrain without it

MSE	$\sigma = 1$	$\sigma = 10$
Retrained, no norm	4.4×10^{-6}	**35.98**

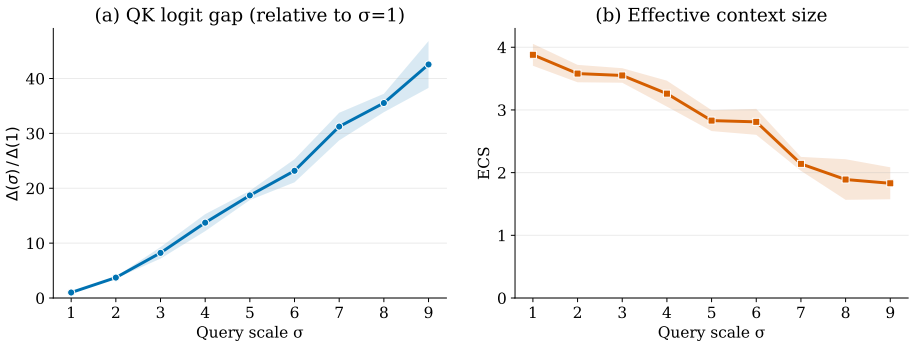
Accurate in distribution again, but the failure under shift remains.

Simplifying further, attention-only models fail the same way.

Removing the bound is insufficient.

Attention concentrates under shift

Query $\sim \mathcal{U}(-\sigma, \sigma)$. Context stays in distribution. First layer. $\mathrm{ECS}(a) = 1/\sum_j a_j^2$



Mean per-head effective context size: 3.9 **to** 1.8**.** Averaged across heads and samples.

Can flattening attention help?

Flattening at inference

$$a_j(T) = \frac{e^{z_j/T}}{\sum_r e^{z_r/T}}$$

	$\sigma = 1$	$\sigma = 10$
$T = 1$, original	9×10^{-6}	27.06
$T = 2$	2×10^{-4}	29.99
$T = 20$	0.15	37.32
$T = 10^6$	0.27	**40.75**

Retraining with other attention

- Temperature-scaled softmax
- Linear attention
- Sparsemax and Entmax
- CosFormer

All stay accurate in distribution, but reach an MSE of 28 to 38 at $\sigma = 10$, against 27 for softmax.

Flattening or replacing attention does not restore generalization.

Conclusion

Transformers can reach near-perfect in-distribution ICL and still fail under shift. We identify two architectural limitations that contribute to this failure.

1. Final normalization

A bounded readout: predictions cannot exceed a fixed range. Removing it does not restore generalization.

2. Attention concentration

A context bottleneck: under shift, the query relies on only a few examples of the prompt.

Strong in-distribution ICL does not imply distribution-independent ICL.

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2026

Thank you.

Why Transformers Struggle with Distribution-Independent In-Context Learning

Omar Naim Jérôme Bolte Nicholas Asher

Paper

Code