Investigation into In-Context Learning Capabilities of Transformers

Rushil Chandrupatla ruchandrupatla@ucsd.edu
Leo Bangayan lbangayan@ucsd.edu
Sebastian Leng sjleng@ucsd.edu
Mentor: Arya Mazumdar arya@ucsd.edu

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Motivation

Transformers demonstrate strong **in-context learning (ICL)**, enabling models to solve unseen tasks using only example input–output pairs at inference time — without any parameter updates.

- Avoids costly retraining (GPT-3 training took ∼3–4 months)
- Enables fast adaptation to new tasks
- Mechanisms behind ICL remain poorly understood empirically

We empirically investigate how ICL performance depends on:

- Feature dimension d
- Context size N
- Number of training tasks B
- Signal-to-noise ratio R

Problem Setup

Binary classification tasks generated from a Gaussian mixture model:

$$\mu_\tau \sim \mathrm{Unif}(R \cdot S^{d-1}), \quad x_i = y_i\mu + z_i$$
$$z_i \sim \mathcal{N}(0, I_d), \quad y_i \in \{-1, +1\}$$

where R controls class separation (signal strength). Each task consists of N labeled context examples and one query point. The model must infer the task from context without updating parameters.

Model

Linear in-context classifier:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} y_i x_i, \qquad \hat{y} = \hat{\mu}^\top W x_{N+1}$$

$W \in R^{d\times d}$ is learned via SGD with logistic loss. Loss is computed *only on the query*, deliberately separating memorization of context from generalization to unseen points — isolating the geometric mechanism behind ICL.

Training configuration: SGD with learning rate $\eta = 0.01$, W initialized to zero, trained for up to 1000 steps. Performance evaluated every 10 steps across 3 independent seeds.

Research Question 1: Scaling Laws

How does in-context test accuracy scale with d, N, and B?

We investigate how test accuracy of transformer models scales with:
• Dimension (d) • Number of tasks (B) • Sequence length (N)

Key findings:

- Larger N and B raise baseline accuracy and speed up convergence
- Under fixed R, higher d slows convergence but still reaches 1.0
- Under SNR scaling, performance reliably reaches 1.0 regardless of d

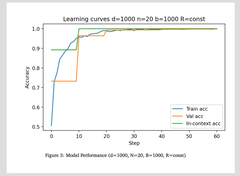


Figure 3: Model Performance (d=1000, N=20, B=1000, R=const)

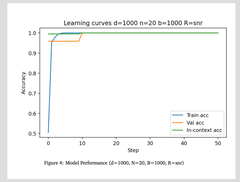


Figure 4: Model Performance (d=1000, N=20, B=1000, R=snr)

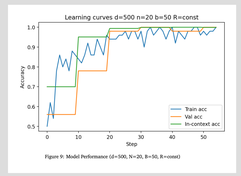


Figure 9: Model Performance (d=500, N=20, B=50, R=const)

Higher-dimensional settings slow convergence unless signal strength increases accordingly.

Research Question 2: Benign Overfitting

When can a model memorize noisy labels but still generalize?
We inject label noise into context examples:

$$\epsilon \in \{0, 0.05, 0.1, 0.2, 0.3, 0.4\}$$

Three observed regimes:

- Underfitting
- Classical overfitting
- Benign overfitting

Benign overfitting occurs when:

- Signal strength is sufficient
- Dimensionality is moderate to high
- Noise is intermediate

When noise is applied only to context labels, benign overfitting occurs across nearly all configurations — even at 40% label flipping. When noise is applied to both context and query labels, signal strength R becomes critical: too low and the model collapses to near-random accuracy; sufficient R restores generalization up to the theoretical maximum.

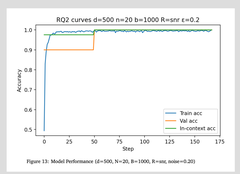


Figure 13: Model Performance (d=500, N=20, B=1000, R=snr, noise=0.20)

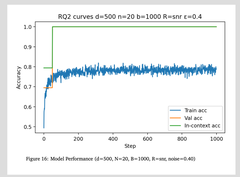


Figure 16: Model Performance (d=500, N=20, B=1000, R=snr, noise=0.40)

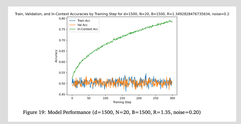


Figure 19: Model Performance (d=1500, N=20, B=1500, R=1.35, noise=0.20)

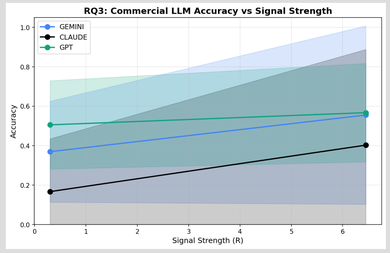
Research Question 3: Full Transformers

Do full transformer architectures exhibit similar ICL behavior?

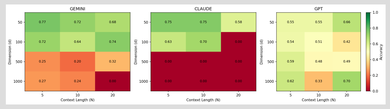
We evaluated:

• GPT • Claude • Gemini

using serialized Gaussian mixture classification tasks.



Performance improves with stronger signal and degrades in high dimensions.



Results partially align with linear ICL theory.

Key Takeaways

- In-context learning is driven by **geometry of task distributions**
- Scaling variables (d, N, B) strongly influence generalization
- Benign overfitting is a stable and reproducible regime
- Full transformers partially implement similar mechanisms

Website QR Code

References

Frei, S. & Vardi, G. (2024). *Trained Transformer Classifiers Generalize and Exhibit Benign Overfitting In-Context*
Garg, S. et al. (2023). *What Can Transformers Learn In-Context?*