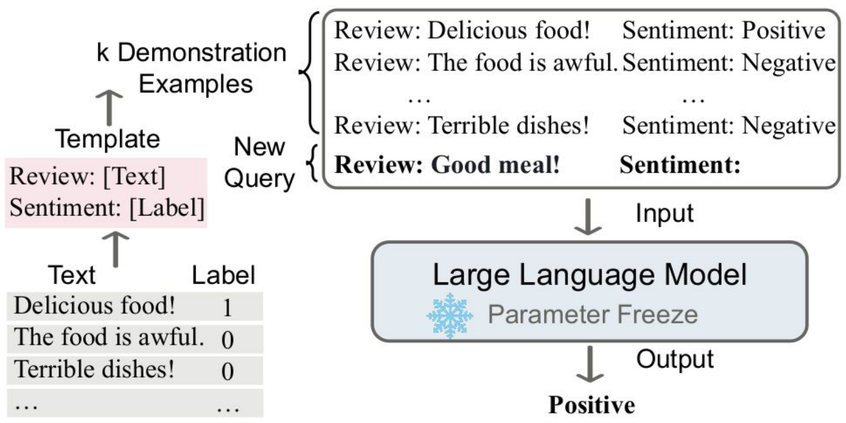
Why Larger Language Models Do In-context Learning Differently?

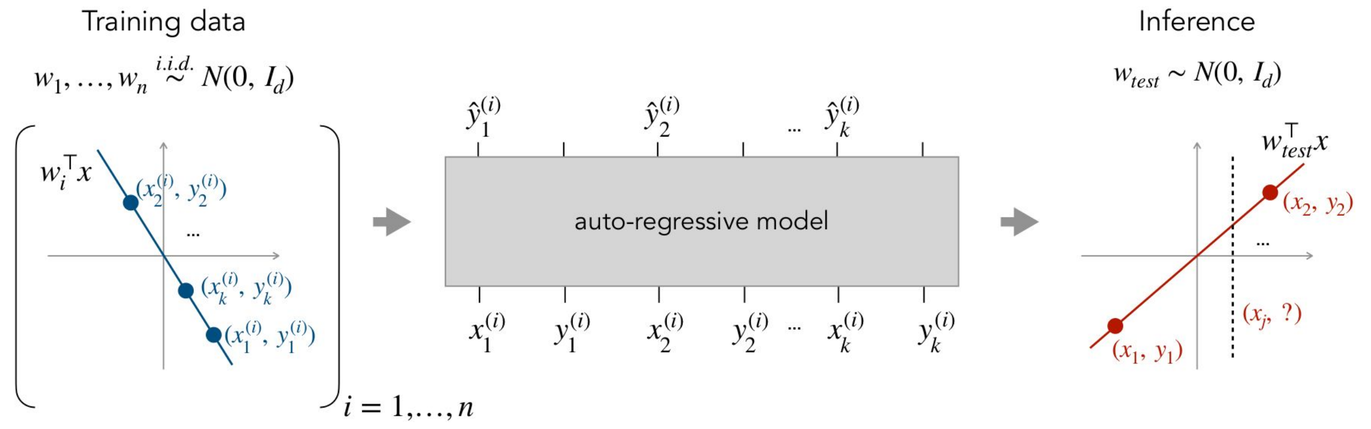
Zhenmei Shi, Junyi Wei, Zhuoyan Xu, Yingyu Liang

WISCONSIN UNIVERSITY OF WISCONSIN-MADISON

NEURAL INFORMATION PROCESSING SYSTEMS

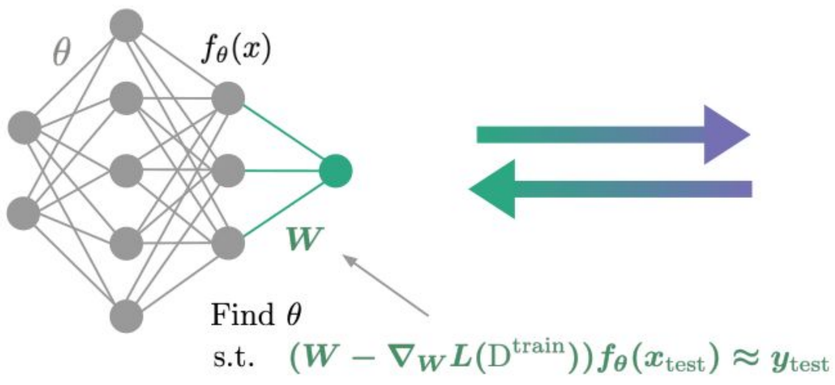
Background





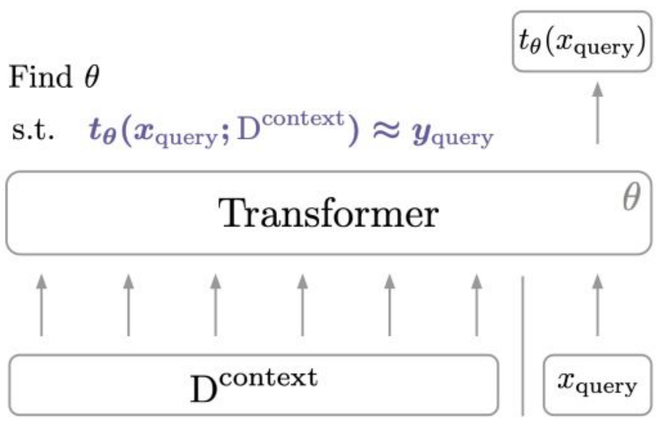
In-context Learning

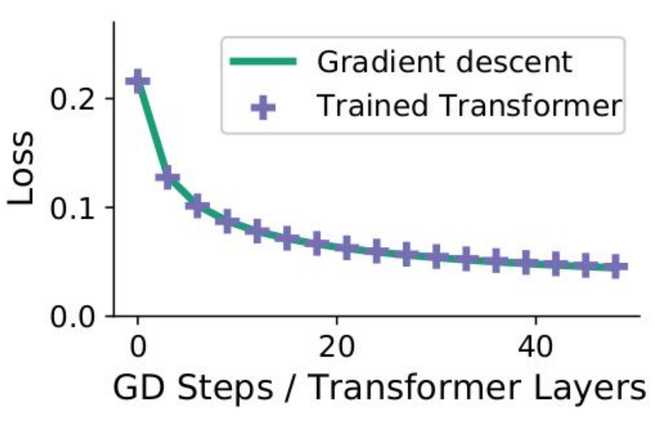
Source "A Survey on In-context Learning." (arXiv 2023)



Linear Regression In-context Learning Data Model

Source "What Can Transformers Learn In-Context? A Case Study of Simple Function Classes." (NeurIPS 2022)

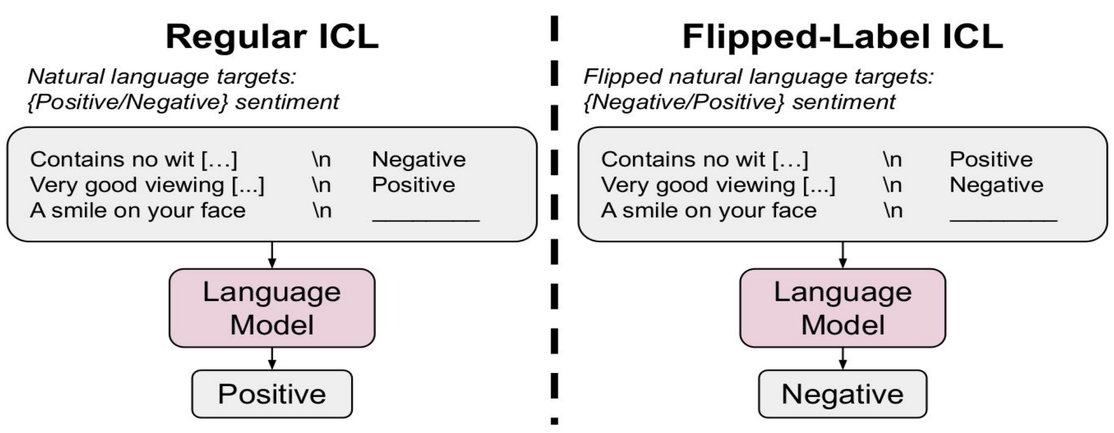


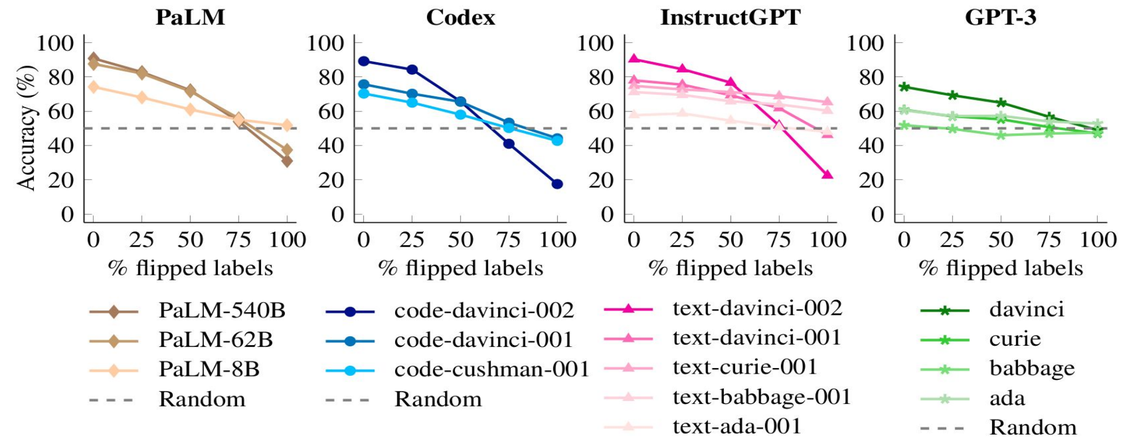


Gradient-based Optimization and Attention-based In-context Learning

Source "Transformers Learn In-Context by Gradient Descent." (ICML 2023)

Motivation





Larger Language Models Do In-context Learning Differently

Source "Larger Language Models Do In-context Learning Differently." (arXiv 2023)

Problem Setup

- In-context learning data; label noise

$$y_{\tau,i} = \langle w_\tau, x_{\tau,i} \rangle + \epsilon_{\tau,i};\ \epsilon_{\tau,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

$$x_{\tau,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, QDQ^\top);\ D = \mathrm{diag}([\lambda_1, \ldots, \lambda_d])$$

$$E_\tau := \begin{pmatrix} x_{\tau,1} & x_{\tau,2} & \ldots & x_{\tau,N} & x_{\tau,q} \\ y_{\tau,1} & y_{\tau,2} & \ldots & y_{\tau,N} & 0 \end{pmatrix}$$

- Mean squared error (MSE) loss
- Linear self-attention networks

$$f_{\mathrm{LSA}}(E) = \left[E + W^{PV} E \cdot \frac{E^\top W^{KQ} E}{\rho} \right]_{(d+1),(N+1)}$$

- Measure model scale by rank of W^{KQ}
- Input noise $w = Q(s + \xi)$

Theoretical Results

Theorem 1 (Optimal rank-*r* solution)

The optimal rank-*r* (small language model) solution of the MSE loss indeed is the truncated version of the optimal full-rank (large language model) solution.

Theorem 2 (Evaluation loss)

N examples in training and *M* examples in evaluation. We have evaluation population MSE loss:

$$\mathcal{L}(f_{\mathrm{LSA}}) \approx \|\xi\|_D^2 + \frac{1}{M}\left((r+1)\|s\|_D^2 + r\|\xi\|_D^2 + r\sigma^2\right) + \frac{1}{N^2}\|s\|_D^2$$

Theorem 3 (Behavior difference)

Denote the optimal rank-r_1 solution as f_1 and the optimal rank-r_2 solution as f_2. Suppose $r_1 < r_2$, we have:

$$\mathcal{L}(f_2) - \mathcal{L}(f_1) \approx \underbrace{\frac{r_2 - r_1}{M}\|s\|_D^2}_{\text{input noise}} + \underbrace{\frac{r_2 - r_1}{M}\sigma^2}_{\text{label noise}}$$

Take-Home Message

We can decompose behavior difference gap to **label noise** and **input noise**. When we have a larger language model, we will have a larger evaluation loss gap between the large and small models. It means larger language models may be easily affected by the label noise and input noise and may have worse in-context learning ability, while smaller language models may be more robust to these noises. Moreover, if we increase the label noise scale on purpose, the larger language models will be more sensitive to the injected label noise. This main intuition is consistent with the observation in previous works.