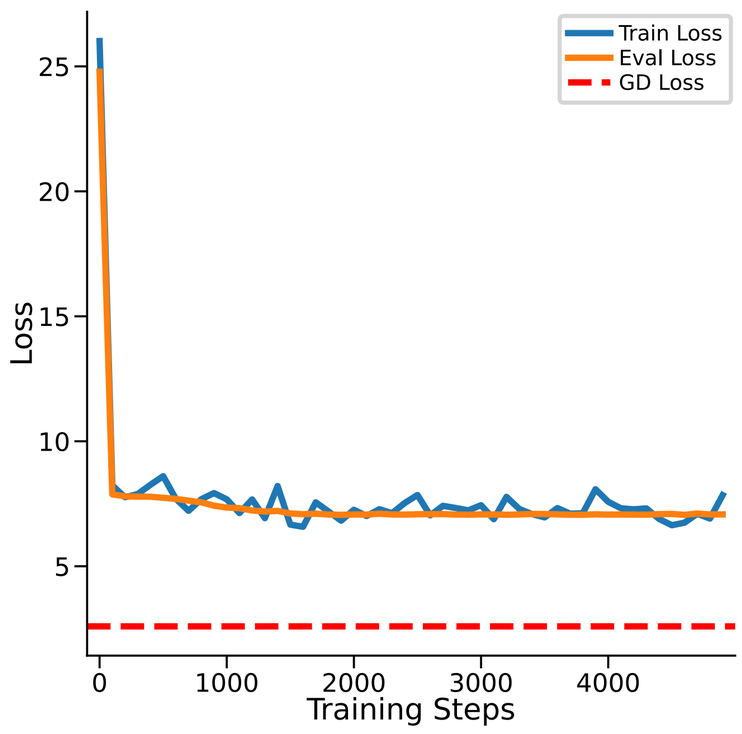
The Initialization Determines Whether In-Context Learning Is Gradient Descent

Shifeng Xie[1], Rui Yuan[2], Simone Rossi[3], Thomas Hannagan[4]

[1]Telecom Paris, Institut Polytechnique de Paris, France; [2]Lexsi Labs, Paris, France;
[3]EURECOM, France; [4]Stellantis, France

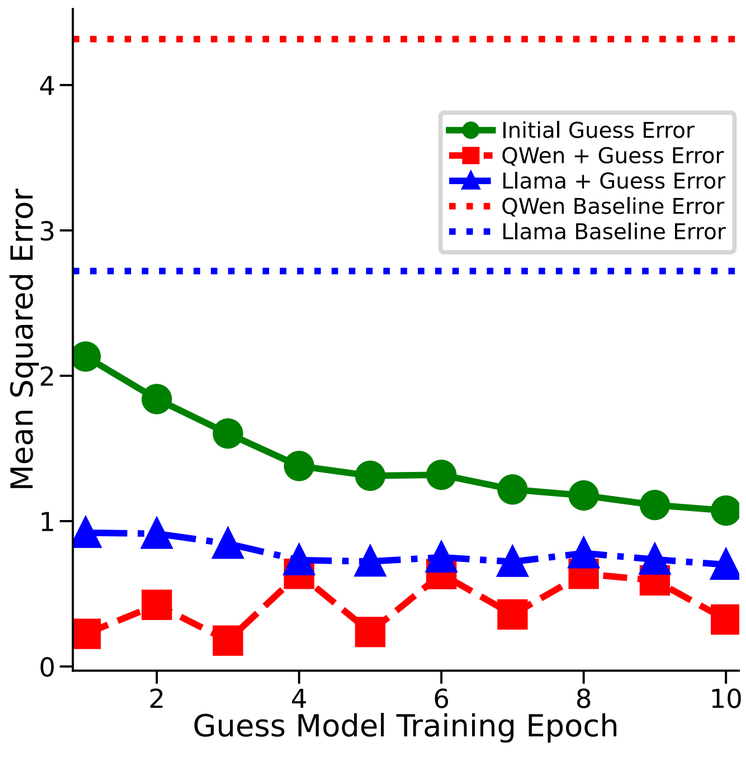
Problem & Gap

We revisit in-context learning as an optimization problem and show that when the true parameter has a non-zero bias, the classical equivalence between LSA and one-step GD breaks, leaving a persistent error that extra heads cannot remove and that must instead be explained by how the query prior mean $\mathbf{y}_q$ is chosen.



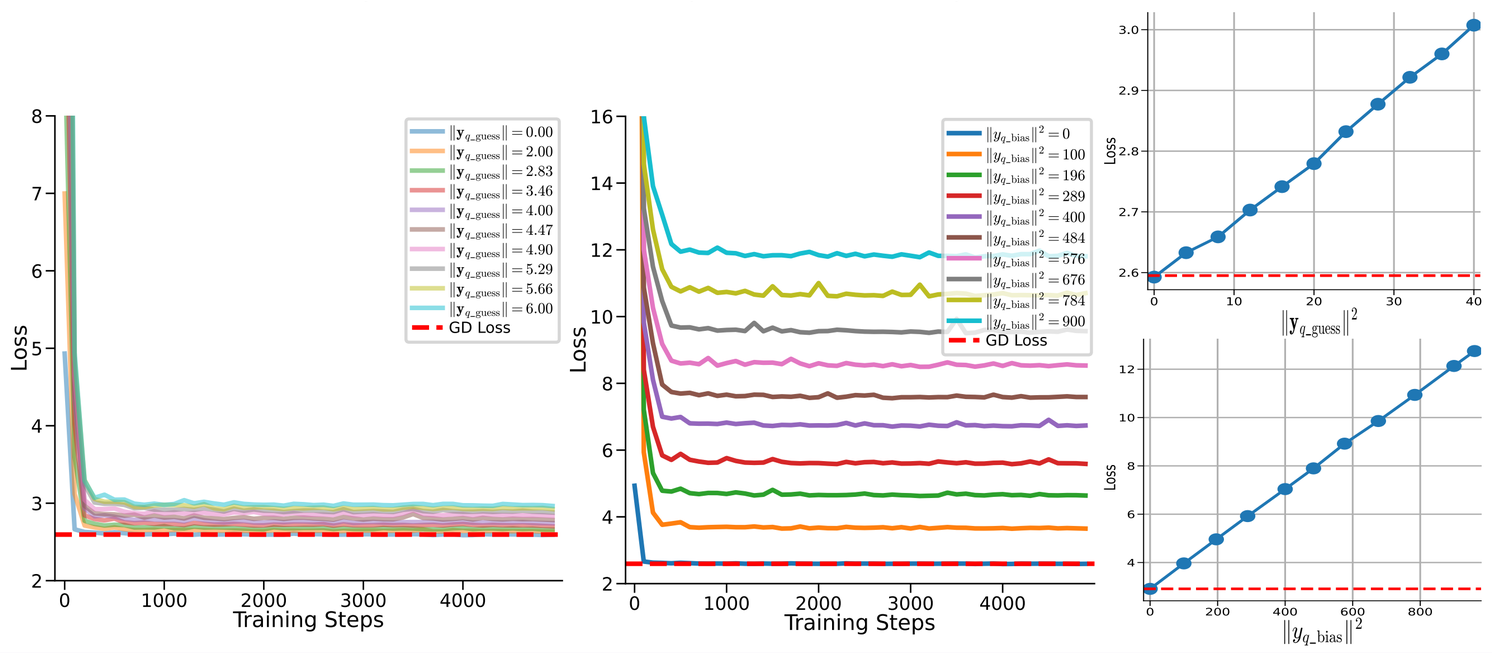
LLMs: Initialization Helps

On STS-B with LLaMA-3.1-8B and Qwen2.5-7B, seeding the prompt with an explicit initial guess consistently lowers MSE, indicating that these LLMs also exploit initialization as a learned prior in their in-context updates.



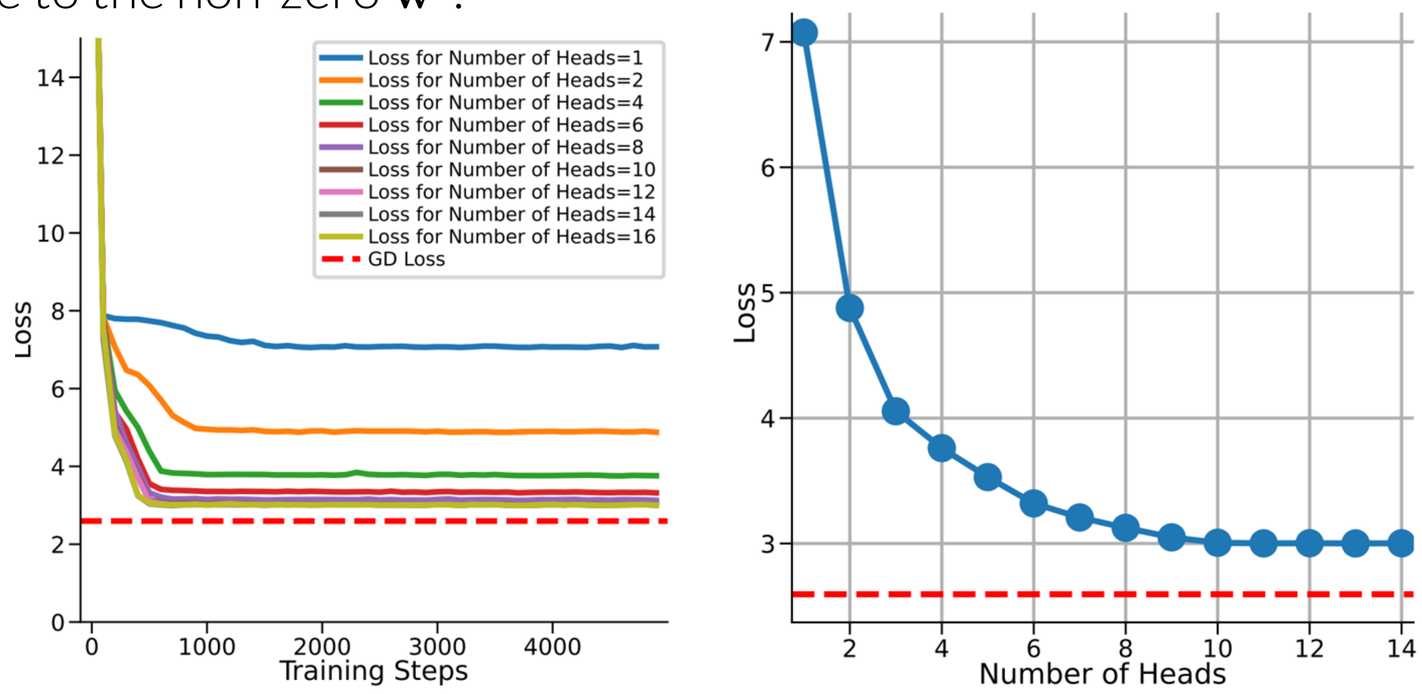
Core Question & Insight

In our analysis, the query output y_q plays the role of the GD prior mean: when $\mathbf{y}_q = \mathbf{w}^{\star\top}\mathbf{x}_q$ (match the correct prior mean), single-head LSA reproduces one-step GD, while any mismatch yields a constant bias.



Capacity: Heads Saturate

Each head contributes one rank of capacity, so once $H = d+1$ the linear target family is already fully spanned and extra heads cannot further reduce the ICL risk, leaving any remaining gap to one-step GD entirely due to the non-zero $\mathbf{w}^\star$.



Design: $\mathbf{y}_q$-LSA

$\mathbf{y}_q$-LSA augments single-head LSA with a trainable vector $\mathbf{w}$ that sets $\mathbf{y}_q = \mathbf{x}_q^\top \mathbf{w}$, so the model learns the prior mean and can recover one-step GD even when $\mathbf{w}_\star \neq 0$.