Keep Your Data Close, but Your Failures Closer: Failure-Driven Adversarial Self-Evolution of Language Models

Choy, Zachary (School: Raffles Institution)
Tan, Min Sen (School: Raffles Institution)

Large language models (LLMs) achieve remarkable performance on complex reasoning tasks, yet they fail unpredictably and with high confidence on seemingly simple problems, posing a critical challenge for reliable deployment. Existing approaches to improving reliability focus on training with more difficult data, but without specifically targeting a model's weaknesses, its failure modes remain unaddressed. Adversarial training solved an analogous problem in computer vision, but cannot be applied to text as gradients cannot flow through discrete text tokens. For the first time, we propose targeted adversarial training for LLM reasoning through textual gradients: instead of forcing text into numerical gradient computation, we convert the entire optimization process to language. This allows us to directly target a model's weaknesses, transforming problems it solves into variants that induce failure and training on those failures. We then hypothesize that training on problems at the boundary of model capability, rather than on generically difficult data, is more effective for model improvement, and validate this experimentally. We build this into an automated, domain-agnostic framework for failure-driven adversarial self-evolution: the model improves, its failure boundary shifts, and new training data co-evolves to match, forming a self-improving loop. Using the same amount of training data, our method consistently outperforms state-of-the-art data augmentation methods. We demonstrate its effectiveness across mathematical reasoning, medical diagnosis and protein function prediction. Our framework enables LLMs to continuously discover and learn from their own failures without requiring human-designed training data, advancing their reliability in high-stakes applications.

Awards Won:
Third Award of $1,200
Association for Computing Machinery: Second Award of $3,000