**Self-Adapting Language Models**

[Adam Zweiger](https://adamzweiger.github.io/)﻿, [Jyothish Pari](https://jyopari.github.io/)﻿, [Han Guo](https://han-guo.info/), [Ekin Akyürek](https://ekinakyurek.github.io/), [Yoon Kim](https://people.csail.mit.edu/yoonkim/), [Pulkit Agrawal](https://people.csail.mit.edu/pulkitag/)

MIT

**Paper:** <https://arxiv.org/abs/2506.10943>

**Code:** <https://github.com/Continual-Intelligence>

**Abstract**

Large language models (LLMs) are powerful but static; they lack mechanisms to adapt their weights in response to new tasks, knowledge, or examples. We introduce **Se**lf-**A**dapting **L**LMs (SEAL) 🦭, a framework that enables LLMs to self-adapt by generating their own finetuning data and update directives. Given a new input, the model produces a *self-edit —* a generation that may restructure the information in different ways, specify optimization hyperparameters, or invoke tools for data augmentation and gradient-based updates. Through supervised finetuning (SFT), these self-edits result in persistent weight updates, enabling lasting adaptation. To train the model to produce effective self-edits, we use a reinforcement learning loop, using the downstream performance of the updated model as the reward signal. Unlike prior approaches that rely on separate adaptation modules or auxiliary networks, SEAL directly uses the model's generation to parameterize and control its own adaptation process. Experiments on knowledge incorporation and few-shot generalization show that SEAL is a promising step toward language models capable of self-directed adaptation in response to new data.

[A cartoon of a computer

AI-generated content may be incorrect.](https://jyopari.github.io/posts/Self-Adapting%20Language%20Models%2020de3cb2605580d291a8d7d8ca7f1604/Group_35.png)

**Method**

[A diagram of a computer loop

AI-generated content may be incorrect.](https://jyopari.github.io/posts/Self-Adapting%20Language%20Models%2020de3cb2605580d291a8d7d8ca7f1604/Screenshot_2025-06-09_at_1.28.39_PM.png)Overview of SEAL. In each RL outer loop iteration, the model generates candidate self-edits (SE) — directives on how to update the weights, applies corresponding updates, evaluates performance on a downstream task, and uses the resulting rewards to improve the self-edit generation policy.

**SEAL** is a framework that enables language models to generate their own finetuning data and optimization instructions—called *self-edits*—in response to new tasks or information. SEAL learns to generate these self-edits via reinforcement learning (RL), using downstream task performance after a model update as the reward. Each training iteration involves the model generating a self-edit based on a task context, applying the self-edit via supervised finetuning, evaluating the updated model, and reinforcing edits that improve performance. This process is implemented with a lightweight reinforcement learning algorithm called [ReST](https://arxiv.org/abs/2312.06585)﻿, which does rounds of selecting high-reward samples using rejection sampling and reinforcing via SFT. We demonstrate SEAL in two domains: (1) **Knowledge Incorporation**, where the model integrates new factual information by generating logical implications as synthetic data, and (2) **Few-Shot Learning**, where the model autonomously selects data augmentations and training hyperparameters to adapt to new abstract reasoning tasks.

[A black and white text with white text

AI-generated content may be incorrect.](https://jyopari.github.io/posts/Self-Adapting%20Language%20Models%2020de3cb2605580d291a8d7d8ca7f1604/Screenshot_2025-06-09_at_2.04.38_PM.png)SEAL Reinforcement Learning Loop. The specific format of the self-edits (SE) are defined per task domain.

**Experiments**

We test **SEAL** on two domains:

**Knowledge incorporation**, where the task is to finetune the model to internalize new factual information from a given passage such that it can answer related questions without access to the original context.

[A close-up of a chart

AI-generated content may be incorrect.](https://jyopari.github.io/posts/Self-Adapting%20Language%20Models%2020de3cb2605580d291a8d7d8ca7f1604/Screenshot_2025-06-09_at_2.15.54_PM.png)

**Few-shot learning** on [ARC](https://arcprize.org/), where the model must generalize from a small number of demonstrations by generating its own data augmentations and training configurations to solve abstract reasoning tasks. Here are visuals of both setups.

[A diagram of a computer program

AI-generated content may be incorrect.](https://jyopari.github.io/posts/Self-Adapting%20Language%20Models%2020de3cb2605580d291a8d7d8ca7f1604/Screenshot_2025-06-09_at_2.16.16_PM.png)

**Results**

**Knowledge Incorporation:** We evaluate SEAL on the task of assimilating factual knowledge from textual passages. In the single-passage setting, after two rounds of ReST-EM, SEAL improves QA accuracy from 32.7% (no adaptation) to **47.0%**, outperforming models finetuned on raw passages or synthetic data generated by GPT-4.1. In the continued pretraining setting with 200 passages, SEAL again achieves the highest performance at **43.8%**, indicating that its learned editing policy scales beyond the single-passage setup in which it was RL-trained. These results highlight SEAL's ability to convert unstructured text into finetuning data that yields lasting and efficient knowledge integration.

[A table with numbers and a few black text

AI-generated content may be incorrect.](https://jyopari.github.io/posts/Self-Adapting%20Language%20Models%2020de3cb2605580d291a8d7d8ca7f1604/Screenshot_2025-06-12_at_9.51.20_PM.png)

**Few-Shot Learning:** On a **simplified** subset of the ARC benchmark, SEAL achieves a **72.5% success rate**, significantly outperforming both in-context learning (0%) and test-time training with untrained self-edits (20%). This demonstrates SEAL's ability to learn how to configure augmentations and training strategies autonomously, enabling robust generalization from limited demonstrations.

[A close-up of a document

AI-generated content may be incorrect.](https://jyopari.github.io/posts/Self-Adapting%20Language%20Models%2020de3cb2605580d291a8d7d8ca7f1604/Screenshot_2025-06-12_at_10.49.34_AM.png)

**Limitations**

While SEAL enables lasting adaptation through self-generated weight updates, our continual learning experiment reveals that repeated self-edits can lead to **catastrophic forgetting**—performance on earlier tasks degrades as new updates are applied. This suggests that without explicit mechanisms for knowledge retention, self-modification may overwrite valuable prior information. Addressing this remains an open challenge, with potential solutions including replay, constrained updates, or representational superposition.

**Future Work**

Looking ahead, we envision models that **not only adapt** their weights but also **reason about when and how to adapt**, deciding mid-inference whether a self-edit is warranted. Such systems could iteratively **distill chain-of-thought traces into weights**, transforming ephemeral reasoning into permanent capabilities, and offering a foundation for agentic models that improve continuously through interaction and reflection.

[A graph of blue squares

AI-generated content may be incorrect.](https://jyopari.github.io/posts/Self-Adapting%20Language%20Models%2020de3cb2605580d291a8d7d8ca7f1604/Screenshot_2025-06-12_at_10.47.10_AM.png)

**Bibtex**

@misc{zweiger2025selfadaptinglanguagemodels,

title={Self-Adapting Language Models},

author={Adam Zweiger and Jyothish Pari and Han Guo and Ekin Akyürek and Yoon Kim and Pulkit Agrawal},

year={2025},

eprint={2506.10943},

archivePrefix={arXiv},

primaryClass={cs.LG},

url={https://arxiv.org/abs/2506.10943},

}