

LLM Self-Improvement

Die Gestaltung von Prompts, das sogenannte Prompt-Engineering, beeinflusst maßgeblich sowohl die Effizienz von Large Language Models (LLMs) als auch die Qualität der erzeugten Antworten. Indem Prompts gezielt formuliert werden, lässt sich das Verhalten von LLMs steuern, was in gewisser Weise einer Form des Trainings gleichkommt.

Ein bemerkenswerter Aspekt dieser Technik ist die Möglichkeit, dass Sprachmodelle ihre eigenen Ausgaben als Eingaben für weitere Verarbeitungsschritte nutzen können (Chaining). Dadurch eröffnen sich neue Perspektiven auf selbstadaptierende Systeme. Insbesondere das Potenzial, LLMs zur eigenständigen Optimierung ihrer Prompts und der dahinterliegenden Strategien zu nutzen, weist den Weg zu selbstreferenziellen, sich kontinuierlich verbessernden KI-Systemen.

Diese Entwicklungen werden mit den Begriffen self-referential, self-improvement, reflection bezeichnet. Sie markieren einen vielversprechenden Schritt in Richtung autonomer, agentischer Sprachmodelle und stellen ein zentrales Forschungsthema im Bereich moderner KI-Architekturen dar.

Ziel

Ziel ist es, eine fundierte, systematische Wissensbasis zu den Mechanismen von self-improvement, self-referential und reflection in LLMs zu schaffen, die als Grundlage für die Entwicklung autonomer, lernfähiger und selbstadaptiver LLM-Agenten dient.

Forschungsfrage

Welche methodischen Unterschiede bestehen zwischen den aktuellen Ansätzen und welche Verbesserungen können jeweils erzielt werden?

Vorgehen

Zur Beantwortung dieser Frage wird ein strukturierter Vergleich bestehender Self-Improvement-Techniken durchgeführt, mit Fokus auf drei Kernmechanismen:

1. Self-Referential Prompting – Modelle generieren eigenständig neue Prompts oder modifizieren bestehende Eingaben auf Basis eigener Ausgaben.
2. Reflective Evaluation – Modelle bewerten und hinterfragen ihre eigenen Antworten, typischerweise durch explizite „Reflection Prompts“ oder externe Feedback-Schleifen.
3. Iterative Self-Correction / Debate Mechanismen – Modelle nutzen dialogbasierte Strukturen (z. B. LLM Debates), um konkurrierende

Antworten zu erzeugen, zu vergleichen und auf ein besseres Ergebnis zu konvergieren.

Material

- <https://evjang.com/2023/03/26/self-reflection.html>
- <https://aclanthology.org/2024.naacl-long.15/>
- https://composable-models.github.io/llm_debate/
- <https://arxiv.org/abs/2402.06782>