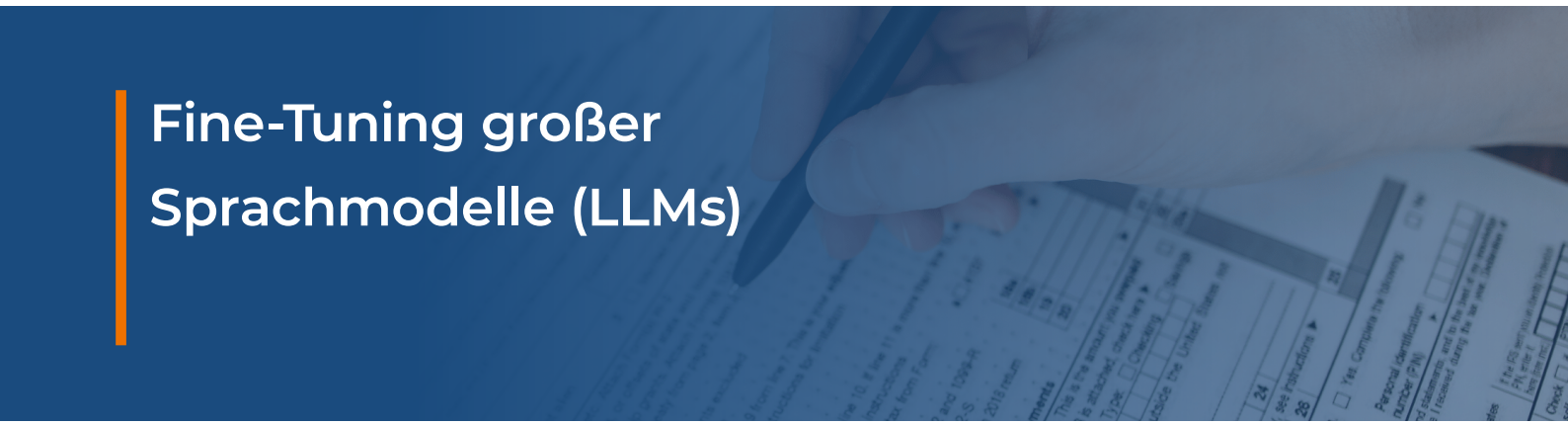
Fine-Tuning großer Sprachmodelle (LLMs)



Executive Summary

Mit parameter-effizientem Fine-Tuning (PEFT) verbessern wir die Leistungsfähigkeit großer Sprachmodelle (LLMs) für spezifische Aufgaben wie die Extraktion von Rechnungsdaten. Mit minimalem Trainingsaufwand und geringen Ressourcen übertrifft unser angepasstes Modell sogar GPT-4o – mit einer Leistungssteigerung von 22 % durch Fine-Tuning und einer um 16 % besseren Performance im Vergleich zu GPT-4o, während es zugleich um ein Vielfaches kleiner ist. Damit bieten wir eine nachhaltige, anpassbare und lokale einsetzbare KI-Lösung.

EINFÜHRUNG

Große Sprachmodelle (LLMs) zählen zu den bedeutendsten Durchbrüchen im Bereich Künstlicher Intelligenz (KI). Ihre beeindruckenden generativen Fähigkeiten wurden der breiten Öffentlichkeit erstmals mit der Veröffentlichung von ChatGPT zugänglich. LLMs verfügen über ein breites Wissen und können dank Vortraining mit riesigen Datenmengen vielfältige Aufgaben bewältigen. Dennoch liefern sie bei spezifischen Anwendungsfällen oft nicht die gewünschte Leistung – selbst bei optimierten Prompts. Da das Training großer Modelle mit Milliarden von Parametern sehr ressourcenintensiv ist, sind spezielle Methoden erforderlich, um die Performance in bestimmten Domänen und Anwendungsfällen gezielt zu verbessern.

RAHMENBEDINGUNGEN FÜR NACHHALTIGES FINE-TUNING

Wir nutzen parameter-effiziente Fine-Tuning-Methoden (PEFT), um vortrainierte LLMs nachhaltig anzupassen. Dabei werden die Gewichte des ursprünglichen Modells nicht verändert – stattdessen trainieren wir ca. 1 % zusätzliche Parameter auf einige hundert anwendungs- oder domänenspezifische Trainingsbeispiele. So verbessert sich die Modellleistung gezielt, ohne die Gesamtkapazität zu beeinträchtigen.

Durch das Fine-Tuning übertreffen kleinere LLMs mit ca. 7 Milliarden Parametern sogar größere Modelle wie GPT-4o. Kleinere Modelle erzeugen schneller Antworten und eignen sich ideal für On-Premises-Lösungen, da sie weniger Rechenressourcen benötigen – was sie nicht nur kosteneffizient, sondern auch nachhaltiger macht. Der reduzierte Ressourcenbedarf senkt zudem die Umweltbelastung deutlich. Die zusätzlichen Parameter werden in einem Adapter gespeichert, der dynamisch aktiviert werden kann.

Unser Adapter für das IDA Rechnungsmodul basiert auf dem multimodalen Modell Qwen2.5-VL, das Text- und Bildverarbeitung ermöglicht.

VORTEILE

- On-Premises-Lösungen
- Anpassung an spezifische Anwendungsfälle und Domains
- Training mit wenigen Trainingsdaten
- Multimodale Modelle
- Ressourcenschonend und nachhaltig

Die Informationsextraktion aus Rechnungen ist ein wichtiger Anwendungsfall der Dokumentenverarbeitung. Dabei treten wiederkehrende Fragestellungen auf – wie etwa die Extraktion des Gesamtbetrags. Wir haben das Basismodell Qwen2.5-VL mit den beschriebenen PEFT-Methoden nachtrainiert. Das Training umfasste sowohl die Extraktion einzelner Entitäten als auch das Auslesen von Tabellen aus Rechnungen. Zudem wurde das LLM darauf trainiert, die Position der angefragten Entitäten korrekt vorherzusagen. Die folgende Tabelle vergleicht die Leistung des Basismodells mit 7 Milliarden Parametern (Qwen2.5-VL-7B), der nachtrainierten Version dieses Modells sowie GPT-4o anhand des sogenannten F1-Scores. Es wurden zwei Bewertungsarten durchgeführt: Zum einen gilt eine Vorhersage als korrekt, wenn der vorhergesagte Text stimmt. Zum anderen nur dann, wenn sowohl der Text als auch die Position korrekt sind.

	Text korrekt	Text & Position korrekt
Qwen2.5-VL-7B	0.72	0.60
Qwen2.5-VL-7B fine-tuned	**0.88**	**0.81**
GPT-4o	0.76	0.66

Tabelle 1: F1-Score auf internen Rechnungs-QS-Datensatz

Bemerkenswerterweise steigert das nachtrainierte Modell nicht nur die Leistung des Basismodells, sondern übertrifft sogar die Leistung von GPT-4o, das als eines der leistungsfähigsten LLMs für allgemeine Aufgaben gilt. **Das nachtrainierte Modell zeigt eine Verbesserung von 22 % gegenüber dem Basismodell und erzielt eine um 16 % bessere Leistung als GPT-4o.** Diese Ergebnisse belegen den Vorteil des Fine-Tunings von LLMs für signifikante Verbesserungen bei spezifischen Anwendungsfällen mit nur wenigen Daten.

ÜBER PLANET AI

Als Softwarehersteller spezialisiert sich PLANET AI auf die Automatisierung von dokumentenbasierten Geschäftsprozessen. Die IDA-Plattform nutzt modernste KI-Technologien, um unstrukturierte Informationen in verwertbare Erkenntnisse zu verwandeln. Damit steigert IDA die Effizienz, Geschwindigkeit und Qualität von Geschäftsprozessen und adressiert Herausforderungen wie Digitalisierung und Fachkräftemangel.

PLANET AI arbeitet mit anderen Softwareanbietern, BPO-Unternehmen und Systemintegratoren zusammen, um Wettbewerbsvorteile und skalierbare Lösungen zu bieten, die auf unterschiedliche Geschäftsanforderungen zugeschnitten sind. Seit der Gründung 1992 und nun als wichtiger Bestandteil im Technologieportfolio von Bechtle, ist PLANET AI als Pionier in der Nutzung von KI für dokumentenbasierte Geschäftsprozesse bekannt.