Prompt Engineering

Generative AI effizient und effektiv anleiten

Generative AI generiert auf Basis von Nutzereingaben neue Inhalte. Dabei hängt die Qualität der Ausgabe stark von der Eingabe ab, dem Prompt. Je präziser die Anweisungen formuliert und die bereitgestellten Kontext- und Zusatzinformationen sind, desto gezielter lässt sich erwünschte Ausgabe beeinflussen. Damit ist Prompt Engineering der Schlüssel zur effektiven Nutzung von Generative AI.

Definition

Prompt Engineering bezeichnet strukturiert gestaltete Eingaben an Generative AI-Modelle, um eine gewünschte Ausgabe zu erhalten.

Im Gegensatz zu traditionellem Maschinellen Lernen ist Generative AI überwiegend unabhägngig vom Anwendungsfall als General-Purpose-Modell konzipiert. Damit deckt es ein breites Einsatzspektrum ab. Allgemein zugängliche Informationen sind in den Trainingsdaten abgebildet und somit im Modell enthalten. Werden die AI-Aufgaben aber spezifischer, dann müssen die Anweisungen detaillierter und gegenbenenfalls um Zusatzinformationen angereichert sein. Je spezifischer die Aufgaben werden, desto präzisere Anweisungen und Zusatzinformationen werden erforderlich. Insbesondere Informationen, die nicht oder nicht ausreichend prominent in den Trainingsdaten enthalten waren.

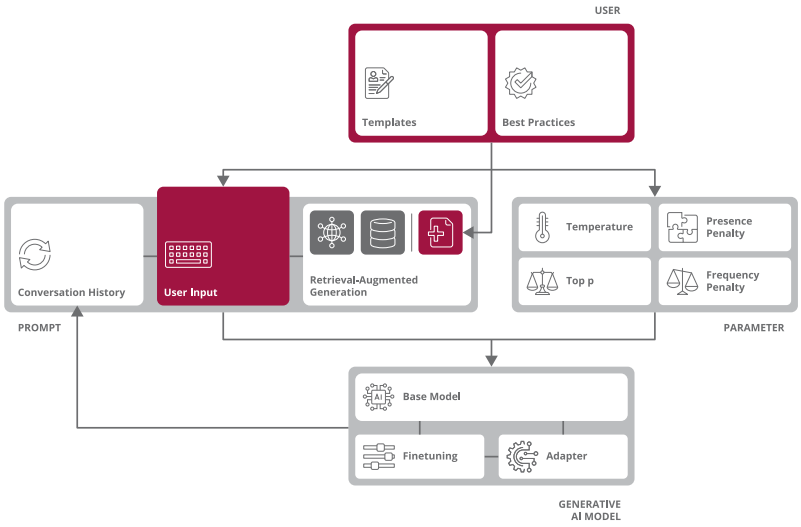
Diese Zusatzinformationen können Anwender sowohl manuell eingeben, lassen sich aber auch automatisch bereitstellen. Die Lösung dafür ist **Retrieval-Augmented Generation (RAG)**. RAG ergänzt den Prompt um potenziell relevante Treffer aus Datenbanken, Dokumenten, Webseiten oder anderen Quellen mithilfe einer semantischen Suche.

Anwenden können die Eingabe freiformuleiren oder auf Vorlagen und Erfolgsrezepte zurückgreifen, um die Effektivität der Anfrage zu erhöhen. Zusätzlich berücksichtigen Sprachmodelle auch den bisherigen Dialog. Dadurch ist eine aufbauende und fortlaufende Konversation möglich, ohne den Kontext zu verlieren.

Parameter wiederum beeinflussen wie das KI-Modell auf den Prompt reagiert. Die Parameter *temperature* und *top p* beeinflussen, wie kreativ oder stringent die Antworten des Modells ausfallen sollen. *presence penalty* und *frequency penalty* beeinflussen hingegen die Vielfalt der verwendeten Worte oder wie repetitiv das Modell antworten darf.

Referenzszenario

Stets wiederkehrende Aufgaben sollen beschleunigt und weitestgehend automatisiert erledigt werden. Allerdings ist das Aufgabengebiet sehr individuell, weil es einen hohen situativen Bezug hat, domänenspezifisches Wissen verlangt oder interne Informationen benötigt. Es wäre allerdings unrentabel, ein KI-Modell explizit dafür zu trainieren. Daher kommt Prompt Engineering zum Einsatz. Die Anwender beeinflussen da KI-Modell also ausschließlich über die gesendeten Eingaben. Andere Anwender sind von den Anpassungen Einzelner nicht betroffen, da sie ihre eigenen, unabhängigen Anfragen stellen können.



Generative AI

- Generative AI und general-purpose Modelle
- Wahrscheinlichkeiten und Halluzinationen
- Kein Fallbezug vorhanden

Compliance

- Intellectual Property der generierten Inhalte
- Copyright der Trainingsdaten
- Datenschutz
- EU AI Act

PEN

Multimodalität

- Input und Output über Mediengrenzen hinweg
- Audiogenerierung
- Videogenerierung
- Bildgenerierung

Verschränkung von System und Wissen

- Natürliche Sprache als Interaktionswerkzeug
- Zugänglichkeit zu Informationen
- Steuerung von Agentensystemen

Potenzial

Anstatt eigene KI-Modelle durch Kostentreiber wie Datensammlung, Datenlabeling, Training, Evaluierung, Drift Detection und Relearning zu entwickeln, lassen sich General-Purpose-Modelle von Anwendungsfall zu Anwendungsfall direkt und sofort optimieren und einsetzen.

Dank Retrieval-Augmented Generation lassen sich auch zusätzliche Informationen kostengünstig dem Generative AI-Modell bereitstellen.

Reifegrad

Viele Anwender formulieren Anfragen noch immer intuitiv. Allerdings ist der Bedarf nach einem strukturierteren Vorgehen erkannt, weil es öfter wiederholbare und qualitativ bessere Ergebnisse erzielt. Immer mehr Richtlinien, Vorlagen und Erfolgsrezepte sind entsprechend aufzufinden. Wiederum haben große Anbieter Retrievel-Augmented Generation adaptiert und setzen damit Anwendungsfälle in Unternehmen um.

Marktübersicht

Die Anbieter von Generative AI-Modellen, etwa OpenAI, Microsoft oder Midjourney, liefern Anleitungen zur Formulierung von zielführenden Prompts. Dienstleister ergänzen das Angebot vermehrt mit Workshops, Webinaren und Cheat Sheets. Wiederum gibt es im Hintergrund mitlaufende Funktionen, die Nutzereingaben automatisch optimieren.

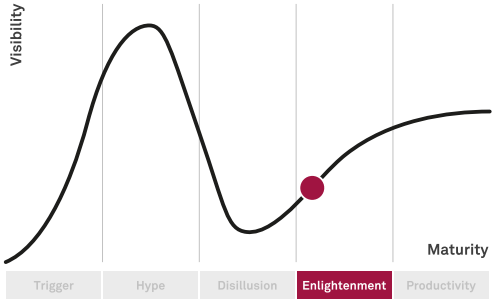
Frameworks wie LlamaIndex, Haystack und Langchain vereinfachen hingegen die Integration von Datenquellen mit KI-Modellen. Und Hyperscaler bieten Hilfsmittel zur Orchestrierung an. Voroptimierte Lösungen wie Claude Projects, Github Copilot und Microsoft 365 Copilot stellen mit eigenen Daten angereicherte Modelle der Öffentlichkeit zur Verfügung.

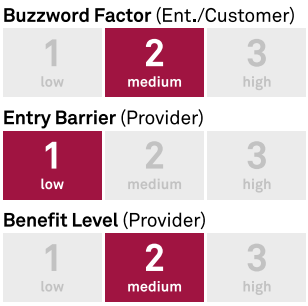
Alternativen

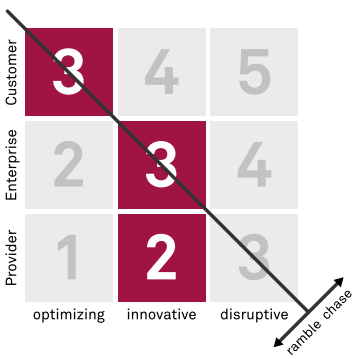
Intuitiv formulierte Nutzereingaben bleiben weiterhin eine gute Option, wenn mit Zusatzinformationen angereicherte Anfragen akzeptabel sind. Damit steigen aber das Fehlerpotenzial und Falschergebnisse. Finetuning passt wiederum das Basismodell an Zusatzwissen an. Dies erfordert jedoch hochwertige Trainingsdaten in ausreichender Menge und erschwert den Wechsel zwischen Anwendungsfällen. Wiederum sind von Grund auf selbst selbst trainiert Modelle denkbar, die jedoch zeit- und kostenintensiv sind.

Fazit

- \+ effektiver und effizienter Einsatz von Generative AI
- \+ leichtgewichtige Anpassungen
- \+ Zusatzinformationen durch RAG zugänglich
- \+ natürliche Interaktion mit Systemen
- \- trotz Prompt
- \- eingeschränkte Portabilität zwischen KI-Modellen
- \- nicht direkt ersichtlicher Mehrwert
- \- intuitives Vorgehen ebenfalls erfolgreich







https://msg.direct/techrefresh

Stand: Oktober 2024

msg systems ag

Robert-Bürkle-Straße 1 | 85737 Ismaning/München | Telefon: +49 89 96101-0 | Fax: +49 89 96101-1113 | www.msg.group | info@msg.group