



Technische Informationen Sage Copilot Agent

Wie funktioniert der Sage Copilot Agent?

- **Prinzipielles Verhalten:** Copilot Agent interpretiert die Anfrage, gleicht die Anfrage mit verfügbaren Funktionen ab und führt die entsprechende Funktion aus. Die Rückmeldung der Funktion wird wieder zur Verfügung gestellt. Man kann weiterführende Anfragen senden.
- **Versteht Kundenanfragen:** Sage Copilot erkennt die Bedeutung von Fragen und gleicht sie mit vorhandenen Funktionen ab.
- **Minimiert falsche Antworten:** Durch das Ausführen von vordefinierten Funktionen und die Ausgabe der Rückmeldung der Funktionen, werden Fehler und "Halluzinationen" reduziert.
- **Erzeugt menschenähnliche Antworten:** Mithilfe von großen Sprachmodellen (LLMs) werden Antworten formuliert, die wie von einer echten Person wirken.
- **Berechtigungssystem:** Alle Interaktionen geschehen im Kontext des eingeloggten Benutzers. Das heißt, im Hintergrund wird das Berechtigungssystem von Sage DPW verwendet. Dadurch wird gewährleistet, dass Sage Copilot sich nur innerhalb der Berechtigungen des Benutzers bewegt.

Wie werden Anfragen verarbeitet?

- Bei Anfragen zu bestehenden Daten, werden Daten aus der Sage DPW mit menschenähnlichen Antworten zur Verfügung gestellt. Zum Beispiel Details zu einem Nettozettel.
- Bei Anfragen zum Ändern von Daten, werden Daten innerhalb von Sage DPW geändert. Zum Beispiel Urlaubsanträge gestellt.
- Sollte der Benutzer keine Berechtigung auf eine Funktion haben, wird diese nicht ausgeführt.
- Es werden nur Anfragen bearbeitet, die dazu dienen, Aktionen innerhalb von Sage DPW durchzuführen.

Welche Technologien stecken dahinter?

- **Strands Agents:** Framework zur Orchestrierung von Sage Copilot-Workflows, einschließlich Prompt-Verarbeitung, Tool-Aufrufe und Antwortgenerierung.
- **AWS Bedrock + Anthropic Claude (Haiku):** Sage Copilot läuft auf Anthropic Claude Haiku über AWS Bedrock und bietet natürliches Sprachverständnis, die Beantwortung von Fragen und die Ausführung von Aufgaben.
- **FastAPI:** API-Schicht, die Client-Anfragen entgegennimmt, als Authentifizierungs- und Sitzungs-Middleware dient und Agenten-Antworten streamt.
- **DynamoDB:** Persistenter Speicher für Chat-Sitzungen und Konversationsstatus. Wird auch zur Erfassung der Token-Nutzung auf Benutzer- und Kundenebene verwendet.
- **Valkey:** In-Memory-Cache, der wiederholte Abfragen reduziert und die Antwortgeschwindigkeit (Performance) verbessert.
- **OpenTelemetry + New Relic:** Observability-Stack für Traces, Logs (Protokolle) und Metriken zur Überwachung von Zuverlässigkeit und Leistung.