Präzise und schnelle Suchergebnisse durch den Einsatz von RAG (Retrieve-Augmented-Generation) im Sozialdienst

Prototyp basierend auf Command R+ von Cohere

Beispiel

Frage: Fortzahlungen nach Todesfall?

https://github.com/dragstoll/hackaton_2024_sod

Mehrwert

- Die kreierte Vorlage für den Sozialdienst könnte die Angestellten bei der effizienten Suche nach richtigen Antworten und Dokumenten bei der Bearbeitung eines Falles unterstützen (vorher wurde pro Frage ca. 15-20 Minuten benötigt).
- Sie kann präzise und korrekte Antworten mit den dazugehörigen Textstellen innert kurzer Zeit liefern und reduziert damit drastisch den üblichen Zeitaufwand im Vergleich zum Status Quo.
- Somit könnten die Mitarbeitenden in den Sozialdiensten der Stadt Zürich potenziell bei einem relativ geringen Kostenaufwand/Frage (ca. 0.4 Rappen) ca. 5-10 Minuten sparen (Cohere Command R+ Pricing).

Aussichten

- Welche Lösung könnte Verfolgt werden?

- Interne Lösung:

  Aufbau einer internen Infrastruktur für die Stadt (hohe Initialkosten, aber dafür hohe Datensicherheit)

- Externe Lösung:

  Aufbau einer externen Struktur mit diversen Anbietern (höhere laufende Kosten, eventuelles Datenschutzproblem bei Falschandwendung)

- Weiterentwicklung der Vorlage zu Chatbot