

Arbeitsheft: Dein erstes Eval-Set in 30 Minuten

Latzerus · Begleitmaterial zu «Ein echtes RAG-System von innen» Artikel und Video:
<https://www.latzerus.ch/lernen/echtes-rag-system-von-innen/>

Worum es geht

Du hast eine Wissensbasis, einen Chatbot auf eigenen Dokumenten oder ein RAG-System. Und irgendwann willst du etwas ändern: andere Abschnittsgrösse, anderes Modell, ein neuer Baustein. Ob das besser ist, sagt dir kein Gefühl. Das sagt dir eine Liste.

Ein **Eval-Set** ist eine feste Liste echter Fragen, bei denen du weisst, wo die Antwort steht. Du stellst sie vor einer Änderung und danach. Am Ende hast du zwei Zahlen:

- **Platz-1-Quote:** Bei wie vielen Fragen steht die richtige Seite auf Platz 1?
- **Top-3-Quote:** Bei wie vielen Fragen steht sie in den ersten drei Treffern?

Beispiel aus meinem System: Vor einem Neuaufbau stand die richtige Seite bei 3 von 9 Fragen auf Platz 1. Einen Tag danach, mit fast denselben Fragen, bei 5 von 7. Ohne diese Liste hätte ich nur sagen können: fühlt sich besser an.

Zeitbedarf: 30 Minuten für das Erstellen und den ersten Durchlauf. Jeder weitere Durchlauf dauert rund 10 Minuten.

Schritt 1 — Fragen sammeln (10 Minuten)

Nimm Fragen, die du oder dein Team **wirklich** stellen. Quellen: Mails, Chats, Telefonnotizen, Tickets. Erfinde keine Testfragen. Die sind oft zu leicht, weil du die Wörter aus dem Dokument schon kennst.

Achte auf eine Mischung:

Fragetyp	Anzahl	Wozu
In eigenen Worten, nicht in denen des Dokuments	4	prüft die Bedeutungssuche
Mit exaktem Namen, Nummer oder Code	2	prüft die Wortsuche
Antwort weit hinten auf einer langen Seite	2	prüft, ob wirklich alle Abschnitte im Index sind
Frage, zu der es keine Antwort im Bestand gibt	1	prüft, ob das System «nichts gefunden» sagt oder rät
Frei	1	deine wichtigste Alltagsfrage

Aussortieren: Ist eine Frage eine Rechnung, gehört sie nicht ins Eval-Set. Summen, Durchschnitte und Zählungen über viele Belege beantwortet ein RAG nicht zuverlässig. Sie gehören ins Quellsystem.

Schritt 2 — Erwartete Seite notieren (5 Minuten)

Schreib zu jeder Frage die Seite auf, auf der die Antwort steht: Titel oder Adresse. Findest du selbst keine Seite, ist das bereits ein Befund. Die Antwort fehlt im Bestand, und daran ändert keine Suchtechnik etwas.

Schritt 3 — Deine Liste

Nr	Frage (wörtlich)	Typ	Erwartete Seite
Bsp.	Wo erkläre ich, was Qwen2.5-VL ist?	Name	zahlen-aus-pdfs-ohne-halluzination
1			
2			
3			
4			
5			
6			
7			
8			
9			keine (Negativ-Kontrolle)
10			

Regel: Die Fragen bleiben wörtlich gleich. Wer eine Frage zwischen zwei Durchläufen umformuliert, misst die Formulierung und nicht die Änderung.

Schritt 4 — Messung vorher (10 Minuten)

Stell jede Frage und trag ein, wo die erwartete Seite landet. Dazu den Score des ersten Treffers, falls dein System einen anzeigt.

Datum: **Stand des Systems:**

Nr	Platz der erwarteten Seite (1 / 2 / 3 / weiter hinten / nicht gefunden)	Score Treffer 1	Bemerkung
Bsp.	1	0.9357	einziger Treffer
1			
2			
3			
4			
5			
6			
7			
8			
9			Soll: «nichts gefunden» oder flache Scores
10			

Platz-1-Quote: von **Top-3-Quote:** von

Schritt 5 — Die Änderung beschreiben

Genau eine Änderung pro Durchlauf. Wer zwei Dinge gleichzeitig ändert, weiss danach nicht, welches gewirkt hat.

Was ändere ich?

Was erwarte ich?

Schritt 6 — Messung nachher

Datum:

Nr	Platz vorher	Platz nachher	Score nachher	besser / gleich / schlechter
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

Platz-1-Quote nachher: von Top-3-Quote nachher: von

Auswertung

Befund	Was er bedeutet	Was du tust
Beide Quoten besser	Die Änderung wirkt	Behalten, Liste für den nächsten Durchlauf aufbewahren
Eine Quote besser, eine schlechter	Die Änderung verschiebt, statt zu verbessern	Die verschlechterten Fragen einzeln ansehen, bevor du entscheidest
Beide gleich	Die Änderung bringt bei deinen Fragen nichts	Zurücknehmen, wenn sie Aufwand kostet
Beide schlechter	Die Änderung schadet	Zurücknehmen, auch wenn sie sich besser anfühlt
Frage 9 bekommt einen hohen Score	Das System findet etwas, wo nichts ist	Der Frage nachgehen: veraltete Seite? Doppelte Seite?
Detailfragen (hinten) fallen durch	Nicht alle Abschnitte sind im Index	Einlesen prüfen: Wie viele Abschnitte sind erwartet, wie viele liegen im Index?

Mein Entscheid: ☐ behalten ☐ zurücknehmen ☐ weiter untersuchen

Drei Fallen

1. **Scores als Prozent lesen.** Ein Score von 0.13 heisst nicht «13 Prozent richtig». Lies die Grössenordnung und die Form: eine scharfe Kante (0.99 · 0.98 · 0.91 · 0.84 · 0.19) heisst «weiss

es», ein flacher Verlauf ($0.13 \cdot 0.09 \cdot 0.03$) heisst «rät».

2. **Nur die leichten Fragen behalten.** Wenn alle zehn Fragen immer auf Platz 1 landen, misst die Liste nichts mehr. Ersetz dann die zwei leichtesten durch schwerere, und notiere das.
3. **Inhalt mit Technik verwechseln.** Fällt eine Frage durch, weil die Seite veraltet ist oder zweimal mit verschiedenem Stand existiert, hilft keine Suchtechnik. Korrigiert wird in der Quelle.

Latzerus — Smart, not hard. · www.latzerus.ch/lernen/