oakai:alpha11

Benchmark & Modellprofil

Domänenspezifisches Sprachmodell
für Support-Qualitätsanalyse

„Klarheit – statt Hype.“

OAKAI · Ralph Ray Schwehr
Juli 2026

Produktionsnahes Modell

oakai:alpha10

Benchmark & Modellprofil

Domänenspezifisches Sprachmodell für Retail- und Kundenanalytik · OAKAI / Ralph Ray Schwehr · Juli 2026

oakai:alpha11 ist ein auf Basis von Qwen2.5-7B (Apache 2.0) feinabgestimmtes, vollständig lokal lauffähiges Sprachmodell für die **Qualitätsbewertung von Support-Chats**. Es bewertet einen kompletten Kundendienst-Dialog auf fünf Dimensionen (Verständnis, Lösungsqualität, Gesprächsfluss, Tonalität, Sprachqualität, je 1–5) und formuliert eine konkrete Verbesserungsempfehlung – alles als striktes, maschinenlesbares JSON. Ergänzend beherrscht es Klassifikationsaufgaben aus der Kundenanalytik. Trainiert wurde es auf realen, anonymisierten Support-Konversationen aus dem technischen Kundendienst. Klarheit statt Hype, im Modell selbst.

Warum ein eigenes, spezialisiertes Modell?

Support-Chats automatisiert und einheitlich zu bewerten, ist ein wiederkehrender, klar umrissener Vorgang – ideal für ein spezialisiertes Modell. Große generische Modelle leisten das zwar auch, aber langsamer, teurer und oft in der Cloud. Ein kompaktes, auf genau diese Aufgabe trainiertes Modell liefert **gleiche oder bessere Bewertungsqualität, in einem Bruchteil der Zeit, im exakt benötigten Format, vollständig im eigenen Haus.** Das senkt Latenz und Kosten und hält sensible Kundendaten lokal.

Bewertungsqualität & Geschwindigkeit

100 ungesehene Support-Chats, deterministisch (Temperatur 0), gemessen auf Mac Studio M4 Max. MAE = mittlere Abweichung der Score-Vorhersage vom Referenzwert (kleiner ist besser, über alle fünf Dimensionen gemittelt). JSON = Anteil valide geparster Antworten.

Modell	Gesamt-MAE	JSON valide	Zeit/Fall
oakai:alpha11	**0,52**	**100%**	**1,47s**
qwen2.5:7b *(Basis)*	0,60	99%	2,27s
gemma4:latest *(12B)*	0,83	100%	15,72s

Kernergebnis: oakai:alpha11 bewertet Support-Chats **genauer als das Basismodell qwen2.5:7b** (MAE 0,52 vs. 0,60) und dabei **1,5× schneller** (1,47s vs. 2,27s) – bei perfekter Format-Treue (100% valides JSON). Gegenüber dem rund 3× größeren gemma4 ist es zugleich deutlich präziser und über 10× schneller. Bemerkenswert: Das spezialisierte Modell übertrifft die Genauigkeit des Basismodells, aus dem die Referenzbewertungen ursprünglich stammen – das Fine-Tuning hat die Bewertungslogik nicht nur reproduziert, sondern stabilisiert.

Hinweis zur Metrik: Einzelne Dimensionen (z.B. Tonalität) sind in den Daten stark zu hohen Werten verschoben; dort ist die absolute Trefferquote naturgemäß hoch. Aussagekräftig ist vor allem der Vorsprung bei Verständnis, Lösungsqualität und Gesprächsfluss.

Optimales Einsatzszenario

Das Modell spielt seine Stärken aus, wenn **große Mengen von Support-Dialogen automatisiert und einheitlich bewertet** werden – etwa zur laufenden Qualitätskontrolle eines Service-Chatbots. Der feste JSON-Output fließt ohne Nachbearbeitung direkt in Dashboards und Auswertungen. Weil es vollständig lokal läuft, eignet es sich besonders für **datenschutzsensible Kundendaten**, die das Haus nicht verlassen

dürfen.

Stärken und Grenzen

Vorteile	Grenzen / wo nicht ideal
• **Genauer als das Basismodell:** niedrigerer MAE als qwen2.5:7b auf ungesehenen Chats	• **Domänenspezifisch:** auf Bewertung technischer Support-Chats zugeschnitten
• **Tempo:** 1,47s je Chat-Bewertung, 1,5× schneller als Basis, 10× schneller als gemma4	• **Schiefe Dimensionen:** Tonalität/Sprachqualität in den Daten stark zu 4 verschoben
• **Format-treu:** 100% valides JSON, direkt weiterverarbeitbar	• **Bewertet, erklärt nicht:** liefert Scores + kurze Empfehlung, keine Langanalyse
• **Datenschutz:** läuft 100% lokal, keine Daten an externe APIs	• **Seltene Sprach-Artefakte:** vereinzeltes Durchschlagen der Basis möglich
• **Kompakt & kostenfrei:** 4,7 GB (4-bit), Apple Silicon, keine API-Gebühren	• **Referenz = Basismodell:** lernt die Bewertungslogik von qwen2.5:7b (stabilisiert sie)
• **Markenfrei:** Trainingsdaten anonymisiert und von Markennamen bereinigt	• **Feste Skala:** neue Bewertungsdimensionen erfordern erneutes Training

Bezug & Nutzung

Das Modell ist quelloffen unter Apache 2.0 veröffentlicht und über mehrere Wege sofort einsatzbereit – alles lokal, ohne Cloud-Anbindung:

Plattform	Bezug	Nutzung
Hugging Face	`huggingface.co/RaSchwehr/oakai-alpha11`	Modellkarte, Benchmark und GGUF-Download (4,7 GB)
Ollama	`ollama pull RaSchwehr/oakai-alpha11`	Direkt aus der Ollama-Registry – ein Befehl, sofort nutzbar
Ollama (via HF)	`ollama run hf.co/RaSchwehr/oakai-alpha11:Q4_K_M`	Alternativ direkt aus Hugging Face laden und starten
llama.cpp	`llama-cli -hf RaSchwehr/oakai-alpha11:Q4_K_M`	Direkt im Terminal oder als OpenAI-kompatibler Server

Technischer Ablauf (Kurzüberblick)

Von der Datenerzeugung bis zum lauffähigen Modell in einer reproduzierbaren Pipeline auf einem einzelnen Mac Studio M4 Max:

• **1. Datenaufbereitung:** reale Support-Chats aus dem technischen Kundendienst, aggressiv anonymisiert (Namen, Kontaktdaten, Kennungen) und von Markennamen bereinigt.

• **2. Training:** QLoRA (4-bit) mit MLX, 1500 Iterationen – nur 0,15% der Modellgewichte trainiert, Validierungs-Loss 2,26 → 0,43; gemischt mit Klassifikations-Beispielen für breitere Kundenanalytik.

• **3. Fusion & Export:** LoRA-Adapter fest ins Modell integriert, Konvertierung nach GGUF, Quantisierung auf Q4_K_M (~4,7 GB).

- **4. Prüfung & Bereitstellung:** systematischer Test auf Marken-Rückführbarkeit (0 Treffer), dann lokal via Ollama nutzbar und auf Hugging Face veröffentlicht.

Methodik & Transparenz

Basismodell: Qwen2.5-7B-Instruct (Apache 2.0). **Training:** QLoRA (4-bit), 1500 Iterationen, MLX auf Mac Studio M4 Max; Validierungs-Loss von 2,26 auf 0,43 gesunken. **Trainingsdaten:** reale, anonymisierte Support-Konversationen aus dem technischen Kundendienst, von Markennamen bereinigt; gemischt mit synthetischen Klassifikations-Beispielen. Datennutzung vertraglich abgesichert. **Auswertung:** 100 ungesehene Support-Chats, Metrik MAE (mittlere Score-Abweichung) über fünf Dimensionen. **Datenschutz:** Personenbezogene Daten und Markennamen wurden vor dem Training entfernt; ein systematischer Test über alle Auswertungsfälle ergab keine Rückführbarkeit auf die Datenquelle.
