

Prompt

Suche mir die Top 5 LLM, die geeignet sind, Texte zu interpretieren und zusammenzufassen unter Berücksichtigung der Tatsache, dass die meisten Texte in deutsch geschrieben sind. Erstelle eine Liste der LLM, ihre Vorteile und Ihre Nachteile gegenüber Apertus und gegenüber Teuken und gebe in der Liste an, ob sie auf Hugging Face heruntergeladen werden können und ob sie für den Claris AI Server geeignet sind.

Ergebnis

Für überwiegend deutsche Texte würde ich derzeit diese fünf lokal betreibbaren Modelle priorisieren:

1. Mistral Small 3.2 24B Instruct
2. Gemma 3 27B IT
3. Qwen3 32B
4. Aya Expanse 32B
5. EuroLLM 22B Instruct

Die Reihenfolge ist eine praxisorientierte Einschätzung für Textverständnis, Anweisungsbefolgung, Zusammenfassung, deutsche Sprachqualität und lokale Bereitstellung – kein universell gültiges Benchmark-Ranking.

Als Vergleichsbasis verwende ich:

- **Apertus v1.5 8B**, nicht die wesentlich aufwendigere 70B-Version
- **Teuken 7B Instruct v0.6**

Apertus v1.5 unterstützt bis zu 262.144 Token und sehr viele Sprachen. Teuken wurde gezielt mit allen 24 EU-Amtssprachen trainiert, besitzt aber nur 7 Milliarden Parameter und eine nichtkommerzielle Lizenz. [Apertus-Modellkarte](#), [Teuken-Modellkarte](#)

Vergleichsliste

Rang	Modell	Vorteile gegenüber Apertus 8B	Vorteile gegenüber Teuken 7B	Nachteile gegenüber Apertus/Teuken	Hugging Face	Claris AI Model Server
1	Mistral Small 3.2 24B Instruct	Größeres Modell; üblicherweise besseres Textverständnis, präzisere Instruktionsbefolgung und stabilere strukturierte Zusammenfassungen	Deutlich leistungsfähiger; längerer Kontext; kommerziell günstige Apache-2.0-Lizenz	Höherer Speicherbedarf; weniger vollständig offen und transparent als Apertus; weniger konsequent auf EU-Sprachen ausgerichtet als Teuken	Ja	Ja, wahrscheinlich unter Windows/Linux über vLLM; auf macOS nur mit passender MLX-Version
2	Gemma 3 27B IT	Höhere Kapazität; ausdrücklich für Zusammenfassungen und Reasoning vorgesehen; sehr gutes Verhältnis zwischen Größe und Qualität	Erheblich stärker bei komplexer Interpretation; 128K Kontext; über 140 Sprachen	Google-Nutzungsbedingungen statt Apache 2.0; ungefähr 54 GB allein für BF16-Gewichte; Apertus hat doppelt so viel Kontext	Ja, Zugriff muss bestätigt werden	Wahrscheinlich , aber Installation mit Ihrer Claris-Version testen; benötigt viel VRAM oder eine unterstützte Quantisierung
3	Qwen3 32B	Stärkeres Schlussfolgern und sehr gute Anweisungsbefolgung; Thinking-Modus; leistungsfähig bei komplexen oder widersprüchlichen Dokumenten	Viel leistungsfähiger und längerer Kontext; Apache 2.0; über 100 Sprachen	Deutsch ist nicht der primäre Trainingsschwerpunkt; gelegentlicher Sprachwechsel möglich; nativer Kontext nur 32K, 131K erfordert YaRN; hoher Speicherbedarf	Ja	Ja, wahrscheinlich , sofern Claris eine ausreichend neue vLLM-Version enthält; Thinking-Modus könnte zusätzliche Konfiguration benötigen
4	Aya Expansive 32B	Speziell multilingual nachtrainiert; Deutsch gehört zu den 23 ausdrücklich optimierten Sprachen; sehr	Deutlich höhere Kapazität; 128K Kontext; meist bessere anspruchsvolle Zusammenfassungen	Nur nichtkommerzielle CC-BY-NC-Lizenz; höherer Speicherbedarf; Apertus ist offener und hat längeren Kontext	Ja	Ja, wahrscheinlich unter Windows/Linux über vLLM; wegen Lizenz für produktive

Rang	Modell	Vorteile gegenüber Apertus 8B gut für sprachübergreifende Dokumente	Vorteile gegenüber Teuken 7B	Nachteile gegenüber Apertus/Teuken	Hugging Face	Clarif AI Model Server
5	EuroLLM 22B Instruct	Stärkerer Fokus auf europäische Sprachen; höhere Kapazität; Apache 2.0	Verfolgt dasselbe EU-Sprachen-Ziel, ist aber rund dreimal größer und moderner	Weniger erprobt und kleinere Community als Mistral, Gemma oder Qwen; Clarif/vLLM-Kompatibilität nicht ausdrücklich für dieses Modell bestätigt; Kontextangaben sollten vor Auswahl geprüft werden	Ja	Unternehmen meist ungeeignet Nur bedingt: Transformers-Modell, aber vor Beschaffung zwingend auf der konkreten Clarif-Version testen

Einzelne Empfehlungen

1. Mistral Small 3.2 24B Instruct – beste Gesamtempfehlung

Das Modell unterstützt Deutsch ausdrücklich, besitzt eine Apache-2.0-Lizenz, 128K Kontext und dokumentierte vLLM-Unterstützung. Mistral nennt verbesserte Instruktionsbefolgung und weniger Wiederholungsfehler – beides ist beim Zusammenfassen wichtig. [Modellkarte](#)

- Besonders geeignet für: Geschäftsberichte, Verträge, Protokolle, längere Artikel
- Gegenüber Apertus: voraussichtlich zuverlässigere Ergebnisqualität und Formatbefolgung
- Gegenüber Teuken: deutlich leistungsfähiger und kommerziell einfacher nutzbar
- Schwäche: Das unkomprimierte Modell ist etwa 48 GB groß

2. Gemma 3 27B IT – sehr gutes Textverständnis

Google nennt Zusammenfassung ausdrücklich als vorgesehenen Einsatzzweck. Das Modell unterstützt über 140 Sprachen und 128K Eingabekontext. [Modellkarte](#)

- Besonders geeignet für: differenzierte Interpretation, Argumentationsanalyse, ausführliche Zusammenfassungen
- Gegenüber Apertus und Teuken: mehr Kapazität und wahrscheinlich bessere Verarbeitung komplexer Zusammenhänge
- Schwäche: Lizenz und Zugriff sind restriktiver als bei Apache-2.0-Modellen

Falls Ihre Hardware begrenzt ist, ist **Gemma 3 12B IT** die vernünftigere Variante. Sie behält den 128K-Kontext und dürfte auf geeigneter Hardware wesentlich leichter bereitzustellen sein.

[Gemma 3 12B](#)

3. Qwen3 32B – gut für analytische Zusammenfassungen

Qwen3 unterstützt mehr als 100 Sprachen, besitzt Apache 2.0 und kann zwischen normalem und ausführlichem Reasoning wechseln. Die Modellkarte verlangt mindestens vLLM 0.8.5.

[Modellkarte](#)

- Besonders geeignet für: komplexe Fachtexte, Gegenüberstellungen, Extraktion von Argumenten
- Gegenüber Apertus/Teuken: wesentlich stärkeres allgemeines Reasoning
- Schwäche: Für reine Zusammenfassungen sollte der Thinking-Modus meistens deaktiviert werden; sonst werden Antworten unnötig lang

4. Aya Expansive 32B – sprachlich stark, aber nicht kommerziell

Aya ist ausdrücklich für 23 Sprachen einschließlich Deutsch optimiert, besitzt 128K Kontext und dokumentierte vLLM-Unterstützung. [Modellkarte](#)

- Besonders geeignet für: deutsche und gemischtsprachige Dokumentbestände
- Gegenüber Apertus: fokussierteres mehrsprachiges Nachtraining
- Gegenüber Teuken: höhere Kapazität und längerer Kontext

- Schwäche: CC-BY-NC schließt viele kommerzielle Anwendungen aus

5. EuroLLM 22B Instruct – europäische Alternative

EuroLLM wurde von einem europäischen Forschungskonsortium entwickelt, unterstützt Deutsch und sämtliche EU-Amtssprachen und steht unter Apache 2.0. [Modellkarte](#)

- Besonders geeignet für: europäische beziehungsweise öffentliche Institutionen und mehrsprachige EU-Dokumente
- Gegenüber Teuken: ähnlicher sprachpolitischer Ansatz, aber wesentlich größeres Modell
- Gegenüber Apertus: stärker konzentrierter europäischer Sprachraum
- Schwäche: geringere Praxiserprobung und unsicherere direkte Claris-Kompatibilität

Wichtige Einordnung zu Claris

Claris erlaubt das Eintragen eines exakten Hugging-Face-Modellnamens und bietet vLLM als Inferenz-Engine an. Daraus folgt jedoch **nicht**, dass jedes Hugging-Face-Modell automatisch unterstützt wird. [Claris-Anleitung zum Herunterladen von Modellen](#)

Die offiziellen Anforderungen nennen:

- Windows/Ubuntu: NVIDIA-GPU mit CUDA
- macOS: Apple Silicon und ein MLX-Modell beziehungsweise eine MLX-Konvertierung
- mindestens ungefähr 24 GB VRAM für kleine bis mittlere Modelle
- GGUF-Modelle für reinen CPU-Betrieb werden nicht unterstützt
- größere Modelle benötigen erheblich mehr VRAM beziehungsweise RAM

[Claris-Systemhinweise](#)

Daher bedeutet „**Ja, wahrscheinlich**“: Das Modell besitzt eine dokumentierte vLLM- oder Transformers-Implementierung, muss aber mit der konkret installierten Claris-Version, deren eingebauter vLLM-Version und der vorhandenen GPU getestet werden.

Meine praktische Auswahl

Bei **32 GB VRAM** würde ich nicht direkt mit einem 27B- oder 32B-Modell in BF16 beginnen. Meine Reihenfolge wäre:

1. **Gemma 3 12B IT** für den ersten produktiven Test
2. **Mistral Small 3.2 24B** mit einer von Claris/vLLM unterstützten FP8- oder AWQ-Version
3. **Apertus v1.5 8B** als offene, europäische Referenz
4. **Teuken 7B** nur für Forschung oder nichtkommerzielle Nutzung

Für eine kommerzielle Lösung würde ich wegen der Lizenz **Aya Expansive und Teuken zunächst ausschließen**. Mistral Small, Gemma, Qwen und EuroLLM sollten anschließend mit 30–50 typischen deutschen Originaldokumenten gegen Apertus getestet werden. Entscheidend

sind dabei Faktenkorrektheit, Auslassungen, richtige Namen und Zahlen, Länge der Zusammenfassung sowie Verarbeitung sehr langer Texte.