

Auracast-Beschallung und LLM-basierte Meeting-Transkription

Stand der Technik, Praxistest und Ergebnisse zweier Forschungsprojekte

Constantin Amend

Videoconferencing & Collaboration Center (VCC) · TU Dresden · November 2026

Agenda



- Auracast im Hörsaal & Lokale Transkription & Zusammenfassung per LLM
- Motivation
- Testaufbau
- Ergebnisse
- Fazit



01

Auracast im Hörsaal

Motivation

Induktionsschleife stößt an Grenzen

- Räumlich begrenzt
- Aufwendig und teuer zu verlegen
- Erfordert einen baulichen Eingriff im Saal

Auracast löst das per Funk

- Bluetooth-Broadcast an beliebig viele Empfänger
- Kopfhörer und LE-Audio-Hörgeräte empfangen direkt
- Mehrere Streams parallel, z. B. Dolmetschung

Forschungsfrage: Taugt Auracast schon für den Hörsaal?

Mit handelsüblichen Sendern und Kopfhörern: Wie hoch ist die Latenz, wie stabil der Empfang, und kommen die Zuhörenden mit Bedienung und Stream-Auswahl zurecht?

Nicht im Fokus: Tests mit Hörgeräten · Messvergleich mit der Induktionsschleife · Festinstallation im Hörsaal

Testaufbau & Szenario



Stream-Auswahl per Hersteller-App auf iPad und Galaxy Tab S8+. Die Tablets sind nur Fernbedienung, empfangen wird im Kopfhörer.

Warum überhaupt eine App? Ohne Auswahl tritt der Kopfhörer automatisch dem stärksten Stream bei.

Community-Test im Hörsaal

562

Plätze im Physikhörsaal, Trefftz-Bau, am 14.09.2026

20 m

getestete Distanz vom Pult in den Saal

7 + 39

Hörende vor Ort und remote, rund eine Stunde Test

30 min

Aufbau vor Ort, für eine Person zu knapp

Ablauf

- Vorbereiteter Text, der Transkription und Übersetzung gezielt fordert
- Vor Ort: Medientechniker der TU, je ein Ohrhörer pro Stream
- Remote: 360°-Bild der Kandao Meeting Pro 2, Übersetzung per Streamlingo-Link (EN, FR, ES)

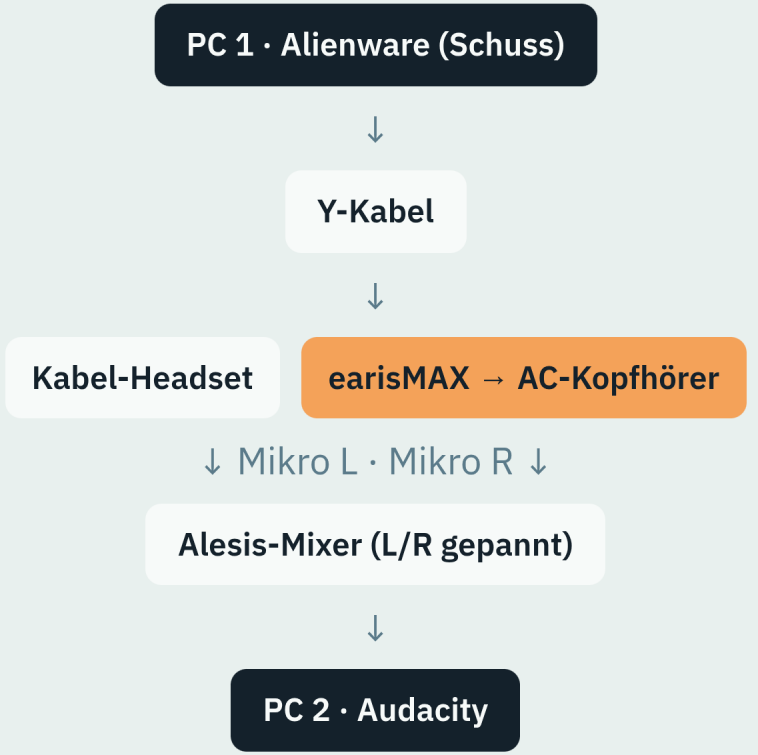
Grenzen des Tests

- Weitgehend leerer Saal, kein Belastungstest
- Kein Hörgerät im Einsatz, deutscher Kanal war KI-erzeugte deutsche Stimme
- Latenz nur subjektiv, kein Vergleich mit der Induktionsschleife

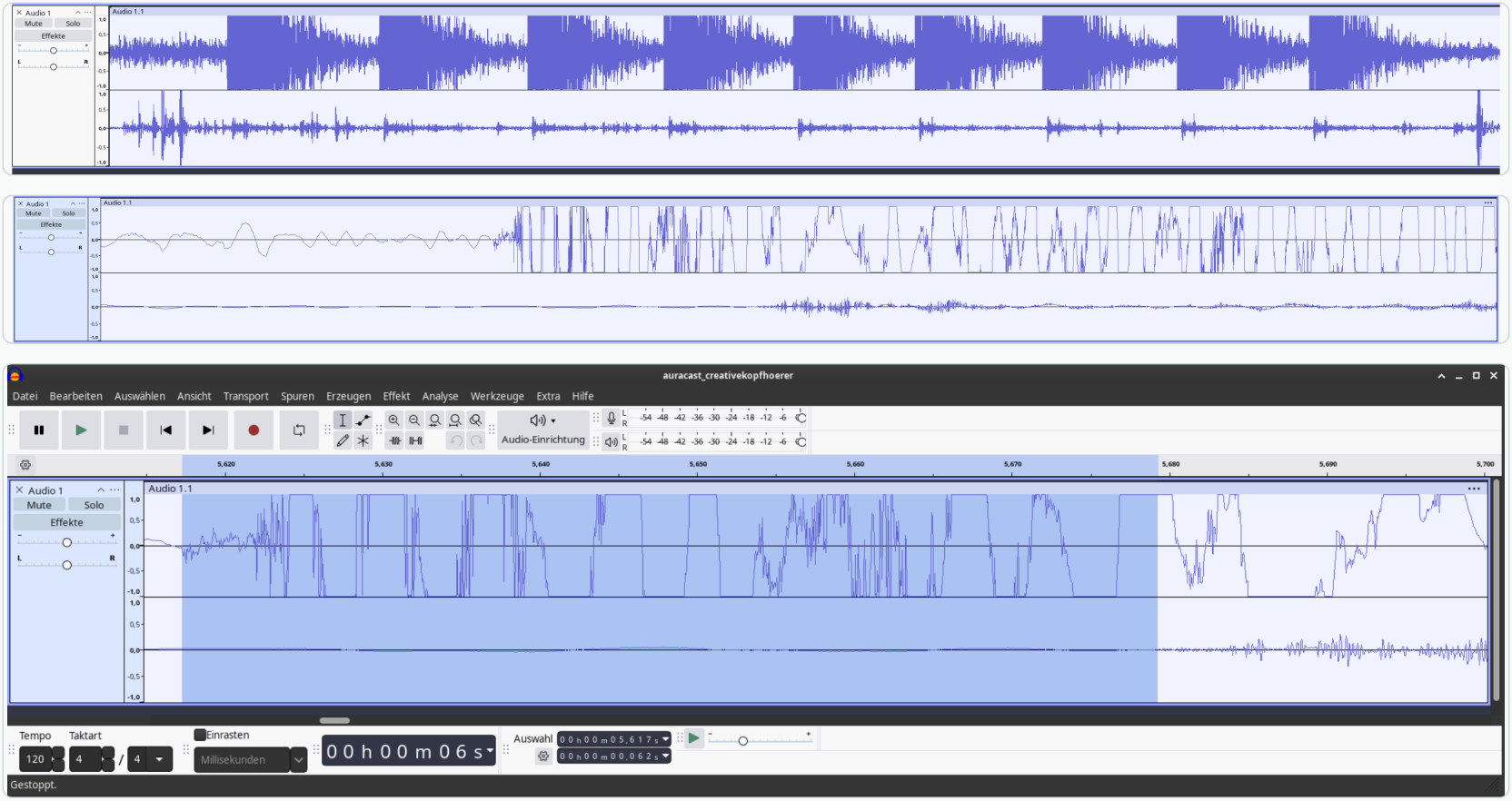
Testaufbau - Latenztest

Vorgehen

1. Testsignal: PC 1 (Alienware) erzeugt ein sehr lautes, plötzliches Geräusch – ein Schuss in einem Spiel
2. Y-Kabel teilt das Signal auf: ein Ende ans kabelgebundene Headset, das andere an den Line-in des earisMAX Auracast-Transmitters
3. Zwei Mikrofone per USB und Alesis-Mixer an PC 2, links/rechts gepannt – beide Kanäle laufen zeitgleich in Audacity auf
4. Alles auf maximale Lautstärke, Mikrofone direkt an die Ausgabegeräte gehalten
5. Mehrere Durchläufe, daraus Mittelwert der Latenz in Millisekunden



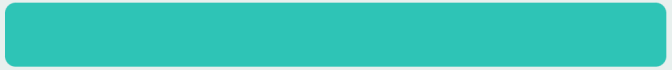
Testaufbau - Latenztest



Ergebnisse

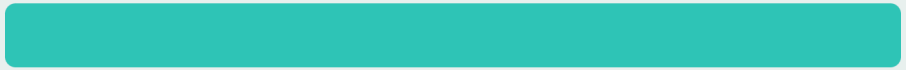
Latenz mit earisMAX als Sender

Creative Aurvana Ace Mimi



62 ms

EarFun Air Pro 4+



84 ms

Zum Vergleich: MIPRO-Sender, TU Chemnitz



< 30 ms

Ziel: unter 30 ms, ab da nicht mehr wahrnehmbar. Unsere Werte liegen noch darüber.

Quelle Chemnitz: C. Greim, Forum Multimedia, 85. DFN-Betriebstagung, 07.10.2026

Community-Test

- **Klang gut**, keine negativen Rückmeldungen, kein Unterschied zwischen Sendern oder Kopfhörern
- **Stabil** über 20 m im Saal, keine Aussetzer. Separat ca. 90 m bei Sichtkontakt
- **KI-Übersetzung** nur ca. 0,5–1 s verzögert, Stimme angenehm
- **Bedienung** als Schwachstelle: Kopfhörer ohne Display nehmen den stärksten Stream

Fazit

Empfehlung: bei Neubau und Sanierung von Hörsälen einplanen

Zukunftsträchtig, erschwinglich und einfach in Betrieb zu nehmen. Ein würdiger Nachfolger der Induktionsschleife.

Was überzeugt

- Am besten geeignet für Hörunterstützung
- Mehrsprachig mit passendem Übersetzungstool
- Stabil und ohne hörbare Qualitätseinbußen
- Reichweite reicht für übliche Hörsäle

Was noch bremst

- Wenige Endgeräte: mind. 210 (Stand 09/2026)
- Stream-Auswahl ohne Display, Apps unintuitiv
- Kopfhörer vor Bluetooth 5.2 bleiben außen vor
- Latenz bei uns noch über 30 ms
- **Blinder Fleck:** kein Realtest mit Hörgeräten

02

Lokale Transkription & Zusammenfassung mit LLMs



Motivation

Protokollieren ist Overhead

- Mitschreiben, verdichten, Beschlüsse und To-Dos
- Wer mitdiskutiert, schreibt lückenhaft mit
- Bindet Personal, das sich um Inhalte kümmern sollte

Cloud-Dienste sind heikel

- Daten verlassen das Haus, oft auch die EU (DSGVO)
- Anbieter bestimmt Preis, Funktionen, Verfügbarkeit
- Kosten wachsen mit der Zahl der Nutzenden

Forschungsfrage: Funktioniert die ganze Kette lokal?

Spracherkennung und Zusammenfassung per LLM auf eigener Hardware, ohne Cloud. Wenn ja: Wie gut ist das Ergebnis, und wie viel Rechenzeit und Hardware braucht es?

Nicht im Fokus: Echtzeit-Transkription · Training und Fine-Tuning · rechtliche Bewertung · Vergleich mit Cloud-Diensten

Lokale KI-Pipeline zur Meetingaufzeichnung

Von der Aufnahme zum Protokoll – Transkription, Sprechererkennung und Zusammenfassung auf lokaler Hardware

(optional)

1 Aufnahme



Meeting wird als Audio- oder Videodatei aufgezeichnet (z. B. lokal oder aus der Videokonferenz).

Output
meeting.mp4 / meeting.wav

2 Transkription



Die Audioaufnahme wird mit einem lokalen Speech-to-Text-Modell in Text umgewandelt.

Output
transkript.txt

3 Sprechererkennung



Die einzelnen Sprecher werden anhand von Sprachmerkmalen identifiziert und dem Transkript zugeordnet.

Output
transkript_mit_sprechern.json

4 LLM



Ein lokales LLM analysiert das Transkript und erzeugt eine strukturierte Zusammenfassung / ein Protokoll.

Output
protokoll.md

5 Protokoll



Das Ergebnis ist ein übersichtliches Protokoll mit den wichtigsten Inhalten, To-dos und ggf. Zeitmarken.

Output
protokoll.md

Testsysteme

System	Klasse	CPU	GPU	RAM
ReferenzPC	High-End-Desktop	Ryzen 7 9800X3D	RTX 5090, 32 GB VRAM	64 GB
Alienware X16 Aurora	Gaming-Notebook	Core Ultra 9 275HX	RTX 5070, 8 GB VRAM	32 GB
MacBook Pro M5	Apple Silicon	Apple M5	integriert	16 GB (Unified Memory)
Lenovo ThinkPad T14	Büro-Notebook	Ryzen 5 Pro 6650U	Radeon 680M, onboard	32 GB

Bewusst breite Spanne: vom High-End-Rechner bis zum typischen Dienst-Notebook ohne dedizierte GPU.

Verwendete Tools

VERWORFEN

LM Studio
AnythingLLM

Für unseren Workflow unbrauchbar:

- geringe Unterstützung von Modellen
- regelmäßige Abstürze
- keine Audiotranskription

GOLDSTANDARD TRANSKRIPTION

noScribe

- Handliches Programm, einfach zu bedienen
- Sprechererkennung (Diarisierung) eingebaut

aber:

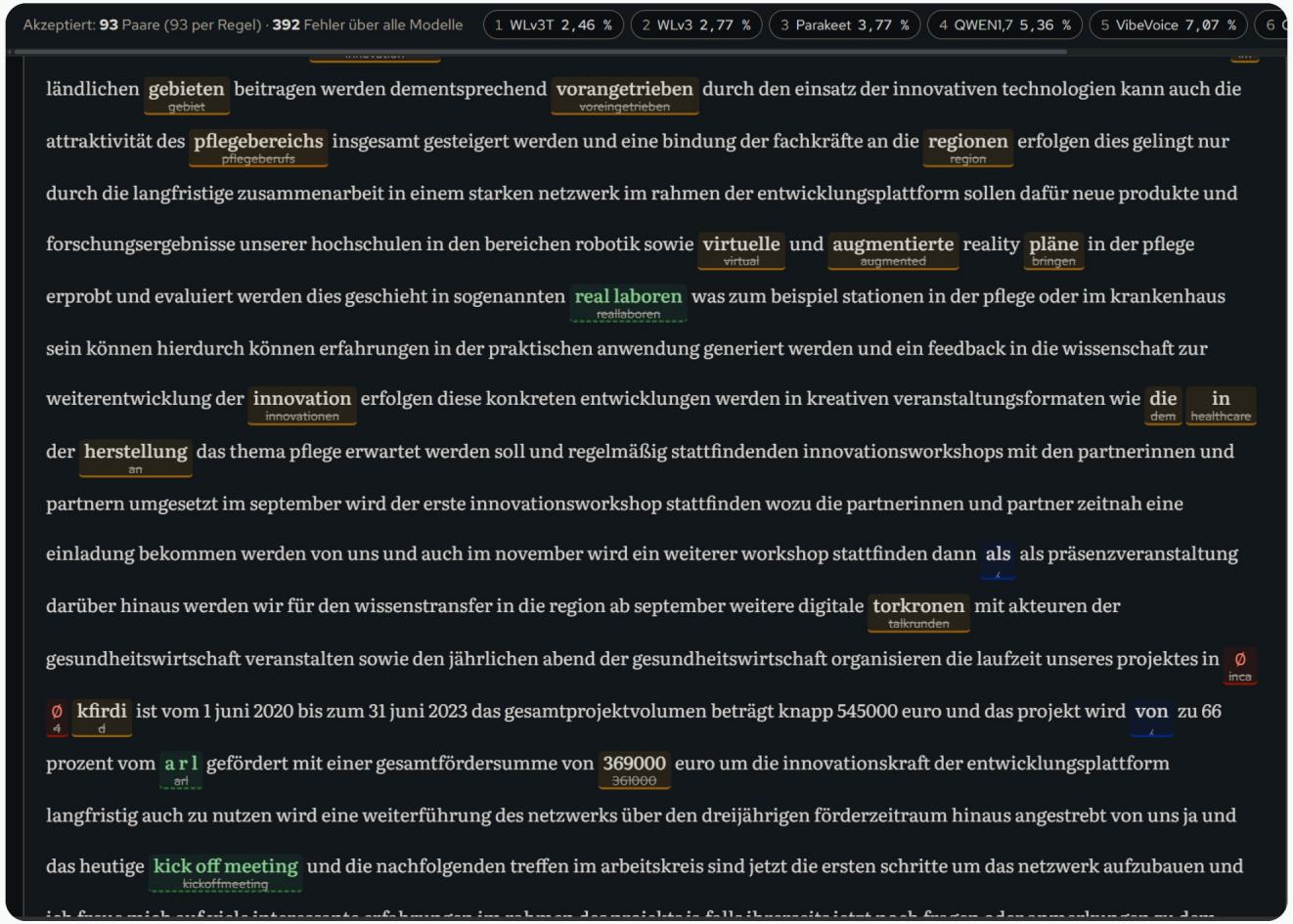
- Nur Whisper-Modelle
- Keine Zusammenfassung

ALL-IN-ONE · UNSER FOKUS

Unsloth Studio

- Transkription mit verschiedenen Modellen
- Zusammenfassung per LLM im selben Tool
- Modelle direkt aus der integrierten Modelldatenbank

Methodik ASR Modelle: Beispiel Canary



So liest man die Markierungen

Substitution: oben Modell, unten Referenz

Deletion: Wort fehlt im Modell (Ø)

Insertion: Wort nur im Modell

Per Regel akzeptiert, zählt nicht als Fehler

Methodik LLMs: inhaltliche Korrektheit

01

Referenzprotokoll

Aus dem Transkript erstellt (Claude), von mir Zeile für Zeile geprüft

Alle LLMs fassen dasselbe Transkript zusammen: **Whisper large v3 Turbo**

→

02

Kernaussagen

36 atomare Aussagen in 7 Kategorien, alle gleich gewichtet

+ **7 Fallen**, die nicht vorkommen dürfen

→

03

Abgleich je Aussage

`erfüllt 1 · teilweise 0,5 · fehlt 0`

Falsch wiedergegeben oder Falle getroffen: **Fehler**

% Abdeckung

Erreichte Punkte geteilt durch 36: Wie viel vom Wesentlichen ist drin?

Fehler

Verfälschte oder erfundene Aussagen, separat gezählt statt verrechnet

03

Ergebnisse



Laufzeit der Transkription

Modell	● ReferenzPC	■ Alienware	◆ Mac	▲ ThinkPad
Canary	13 s Basis	31 s 2,4×	53 s 4,1×	1:58 min 9,1×
Parakeet	19 s Basis	44 s 2,3×	2:03 min 6,5×	12:53 min 41×
Whisper Large v3 Turbo	22 s Basis	84 s 3,8×	6:47 min 19×	15:40 min 43×
Qwen 0,6B	42 s Basis	1:42 min 2,4×	2:51 min 4,1×	9:16 min 13×
Qwen 1,7B	63 s Basis	2:51 min 2,7×	5:36 min 5,3×	× Fehler lief nicht
Whisper Large v3	89 s Basis	5:52 min 4,0×	× Fehler lief nicht	1 h 17 min 52×
VibeVoice	⚠ 2:07 min unbrauchbar	× Fehler lief nicht	× Fehler lief nicht	× Fehler lief nicht

schnell · 13 s — 1 h 17 min · langsam

Zellfarbe = Rechenzeit für 40:27 min Audio, logarithmisch. Zweite Zeile: Faktor gegenüber dem ReferenzPC.

ASR-Modelle im Vergleich

Modell	Wortfehlerrate (WER)	Parameter	Quant.	VRAM (ca.)	Besonderheiten
Whisper Large v3 Turbo	**2,46 %**	~809 M	FP16	3–6 GB	gekürzter Decoder
Whisper Large v3	**2,77 %**	~1,55 B	FP16	5–10 GB	32 Decoder-Layer
Parakeet TDT 0,6B v3	**3,77 %**	~600 M	Q8	1–2 GB	Wort-Timestamps
Qwen3-ASR 1,7B	**5,36 %**	~2 B	Q8	3–4 GB	LLM-Decoder, Streaming
VibeVoice-ASR	**7,07 %**	~9 B	Q8	12–16 GB	ASR + Diarisierung
Qwen3-ASR 0,6B	**7,26 %**	~0,8 B	Q8	1,5–2 GB	wie 1,7B, schneller
Canary 180M Flash	**12,13 %**	~182 M	F32	1–1,5 GB	ASR + Übersetzung

WER gegen ein handkorrigiertes Referenztranskript (40 min, Deutsch) mit eigenen Bewertungsregeln, daher nicht direkt mit Literaturwerten vergleichbar. VRAM geschätzt, Whisper-Quantisierung angenommen.

Sprachmodelle: Geschwindigkeit

Modell	● ReferenzPC	■ Alienware	◆ Mac	▲ ThinkPad
K2 Horizon 3.7B	**17 s** Q8 · n=5	**2:12 min** Q8 · n=5	**2:38 min** Q8 · n=5	**11 min** Q8 · n=4
Gemma4 E4B	**10 s** Q8_K_XL · n=6	**40 s** Q8_K_XL · n=6	**1:26 min** Q8_K_XL · n=5	**5:16 min** Q8_K_XL · n=5
Qwen3.5 9B	**37 s** Q8 · n=6	**2:31 min** IQ4_XS · n=6	**5:28 min** UD-Q4_K_XL · n=5	**16 min** IQ4_XS · n=5
Gemma4 26B-A4B	**22 s** Q8_K_XL · n=6	**2:43 min** IQ4_XS · n=6	**2:28 min** UD-Q2_K_XL · n=5	– nicht getestet
Qwen3.8 27B	**1:01 min** Q6_K_XL · n=6	**26 min** UD-Q2_K_XL · n=6	– zu wenige Läufe	**49 min** IQ3_S · n=3
GLM-4.7 Flash *nicht weiterverfolgt*	**34 s** UD-Q6_K_XL · n=6	**32 s** UD-Q4_K_XL · n=5	**1:42 min** UD-IQ2S_M · n=3	– zu wenige Läufe



schnell · 10 s — 49 min · langsam

Zellfarbe = Median der Rechenzeit für die Zusammenfassung des 40-min-Meetings, logarithmisch. Zweite Zeile: Quantisierung und gültige Läufe. Orange umrandet: länger als das Meeting.

Sprachmodelle: Ergebnisse

Abdeckung der 36 Kernaussagen und kritische Fehler

Modell	Abdeckung (Ø)	K-Fehler Ø
Qwen3.8 27B	66,7 %	2,3
Qwen3.5 9B	54,5 %	1,5
K2 Horizon 3.7B	51,4 %	2,5
Gemma4 26B-A4B	49,5 %	1,0
Gemma4 E4B	43,8 %	1,3
GLM-4.7 Flash	36,1 %	3,3

Nicht weiterverfolgt: lief nicht auf allen Rechnern

Ø über 3–4 Läufe je Modell (verschiedene Rechner und Quantisierungen), alle auf dem Turbo-Transkript. K-Fehler: kritische Fehler pro Lauf.

- **Qwen3.8 27B am vollständigsten**, aber stark von der Quantisierung abhängig: 30,5 von 36 (IQ3_S) bis 13 (IQ1_S)
- **Qwen3.5 9B auf Platz 2**: 17–22 Punkte auf allen vier Rechnern, weniger kritische Fehler als Qwen3.8
- **Gemma 26B am konstantesten**: 17,5–18 Punkte, je Lauf nur ein kritischer Fehler
- **Kein Lauf fehlerfrei**, typisch: Namensvarianten (15 von 19 Läufen), Laufzeit-Widerspruch übersehen (15), erfundene Aufgaben (7), Absicht als Beschluss (7)

Gedanken zu den Ergebnissen

Mögliche Ursache für das schwache Abschneiden der Sprachmodelle

Das Testmeeting war kein klassisches Meeting mit Beschlüssen und Aufgabenverteilung, sondern eher Vortrag und Präsentation mit etwas Diskussion. Der Protokoll-Prompt passt daher nur bedingt auf dieses Meeting.

Apple M5: flotte Generierung trotz fehlender ded. GPU, aber lange Wartezeit aufs erste Token

- **Generierung:** hängt an der Speicherbandbreite, da punktet der Unified Memory
- **Prefill:** rechenlimitiert und wächst mit dem Kontext, hier ist der M5 schwächer als GPUs
- **Folge:** Bei langen Transkripten bestimmt die Wartezeit aufs erste Token die Laufzeit

Die Kernwahrheiten-Metrik ist bewusst streng

- Viele, sehr strikte Kernwahrheiten: die meisten Meetings sind weniger dicht als InCa4D
- Viele Fehler sind Namensfehler, die sich leicht glattbügeln lassen
- Metrik bewusst synthetisch gewählt (Quantifizierbarkeit), nicht unbedingt realweltnah

04

Fazit



Fazit

Forschungsfrage: Funktioniert die ganze Kette lokal?

Antwort: Ja.

- Läuft auch auf kleiner Hardware
- Je nach Laptop bis zu mehreren Minuten für Transkription und Zusammenfassung (vgl. Slide 19+21)
- Händische Sichtung auf Fehler bleibt nötig

Aber:

- **Keine eierlegende Wollmilchsau:** jedes Modell hat andere Stärken, keines ist perfekt
- **Eigenregie nötig:** passende Modelle für Transkription und Zusammenfassung finden und richtig einstellen
- **Prompt muss passen:** er muss auf das Material zugeschnitten sein

Empfehlungen für die Community

Der Prompt entscheidet über die Qualität

Die Zusammenfassung steht und fällt mit dem Prompt. Je besser er zum Material passt, desto besser das Ergebnis: Ein Vortrag braucht einen anderen Prompt als ein Arbeitsmeeting mit Beschlüssen und Aufgaben.

Modell passend zur Hardware: ideal Q8, mindestens Q4

Darunter Qualitätsverlust bis hin zu Schleifen. Für Laptops und Smartphones: Modelle wie **E4B**.

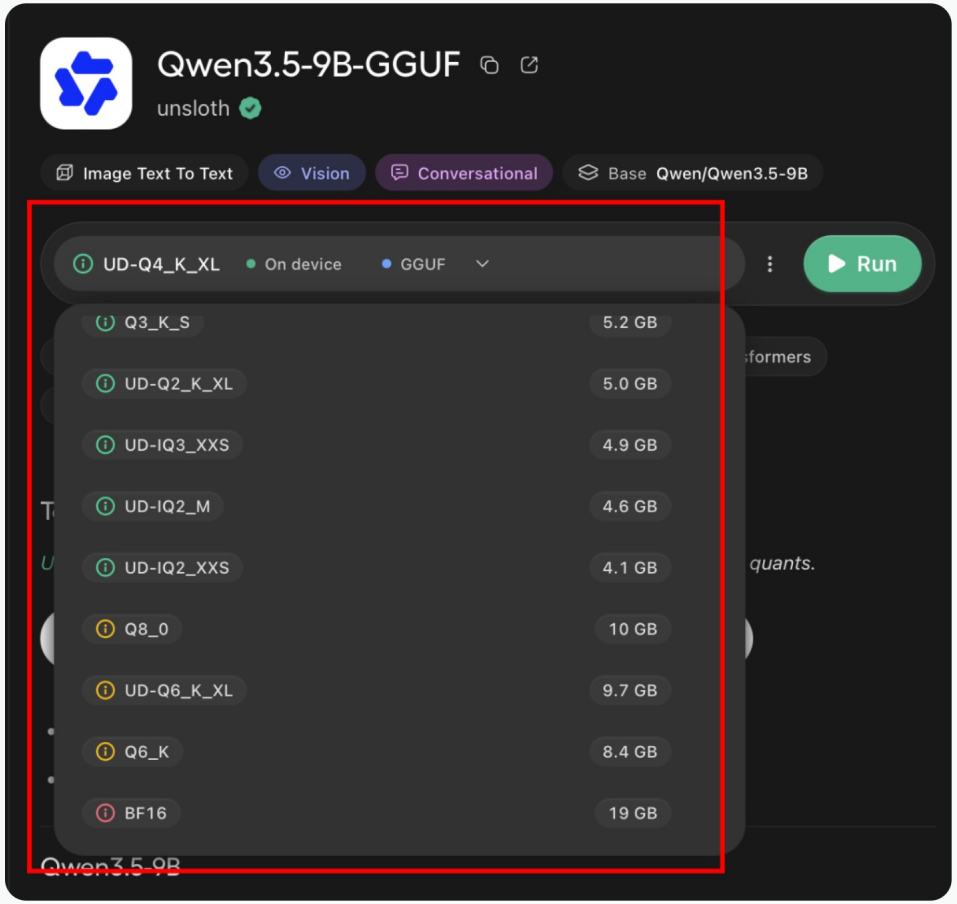
Kontextfenster beachten

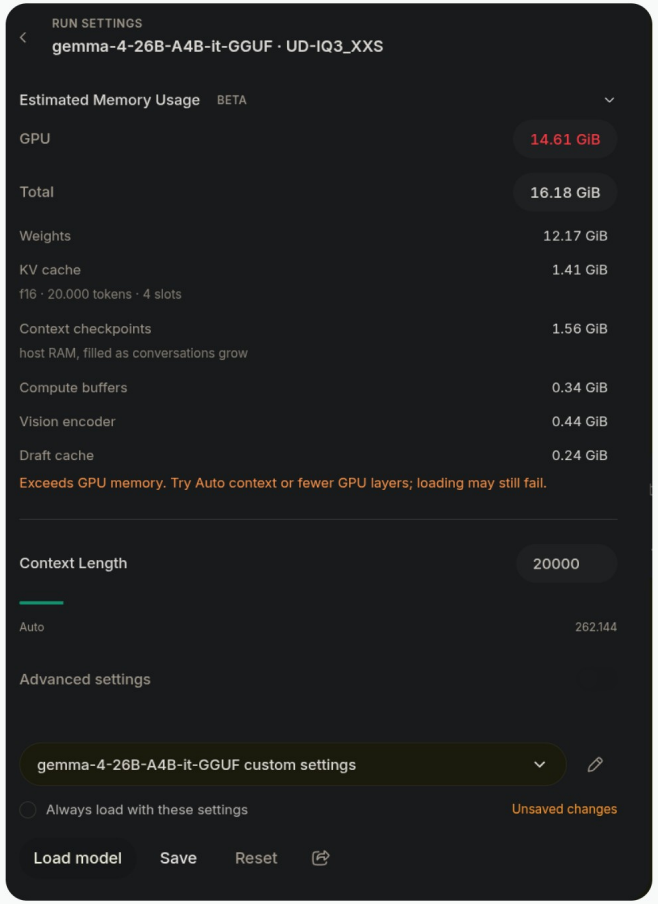
Sonst „vergisst“ das LLM unter Umständen Teile des Transkripts. Daher ein Modell mit ausreichend großem Kontextfenster wählen oder ein performanteres mit künstlich begrenztem Kontext.

Diarisierung in noScribe, Zusammenfassung in Unsloth

Sprechererkennung klappte in Unsloth mit keinem Modell richtig. Das LLM kann Sprecher zwar aus dem Kontext erraten, das kostet aber Tokens und ist ein weiterer Fallstrick.

Empfehlungen für die Community





Rechherzlichen Dank für die Aufmerksamkeit!
Fragen?

Constantin Amend · VCC TU Dresden



Rueckblick: autarke Videokonferenz

Forschungsfrage: klares Ja

Ein Raspberry Pi 5 ist WLAN-Router und eduMEET-Server zugleich. Drei Clients konferieren ohne Internet, in Bild und Ton nicht von einer regulären Instanz zu unterscheiden.

24 h

Laufzeit an einer Powerbank bei ca. 3 W Leistungsaufnahme

65 m

Ungestörte Reichweite bei Sichtkontakt, ab 90 m Abbruch

10–15 m

Grenze bei massiven Wänden, Hindernisse sind der Engpass

1–2 h

Aufbau und Konfiguration, wenn man weiß, was zu tun ist