



FAMULOR RESEARCH

Voice AI Benchmark Oktober 2026

Modelle. Echte Anrufe.
Messbare Ergebnisse.



FAMULOR RESEARCH / OKTOBER 2026

Voice-AI-Benchmark

Modelle im Praxistest

94 Katalogeinträge · 1.060 Komponentenanfragen
44 WebRTC Sprachzellen · DE + EN

Gemessen am 13. Oktober 2026. Kleine, sequenzielle Stichproben. Vollständige Tabellen, dokumentierte Messgrenzen und relative Kostenklassen.

Inhalt

1. Die wichtigsten Ergebnisse	3
2. Studiendesign und Testumfang	3
3. Pipeline, Half-Cascade und Realtime verstehen	4
4. Der vollständige Pipeline-Vergleich	5
5. Alle nativen Gesprächskonfigurationen	6
6. LLMs: alle 33 Modelle im Entscheidungsprofil	7
7. STT: menschliche Sprache statt synthetischer Referenz	9
8. TTS: alle 29 Modelle und der erste Audioblock	12
9. Kosten als Low, Medium und High	14
10. Vier Anwendungsfälle, vier Abnahmefragen	15
11. Antwortzeiten ohne Messfehler lesen	18
12. Was ein vollständiges Reporting erfassen muss	19
13. Famulor-API für reproduzierbare Tests nutzen	20
14. Von der Vorauswahl zur eigenen Freigabe	20
15. Empfehlung für deine erste Konfiguration	21
16. Quellen und Reproduzierbarkeit	21

Ein Voice Agent muss die richtige Aktion ausführen und sie verständlich bestätigen. Eine schnelle Stimme allein reicht dafür nicht. Diese Famulor-Studie vergleicht deshalb drei Ebenen: einzelne Modelle, vollständige Gesprächskonfigurationen und die hörbare Antwort beim Anrufer. Du bekommst die vollständigen Ergebnistabellen, eine Einordnung der Kosten als Low, Medium oder High und einen konkreten Plan für deine eigene Abnahme.

1. Die wichtigsten Ergebnisse

Alle neun getesteten Pipeline-Konfigurationen bestanden acht von acht Prüfungen auf passende Testbelege und gesprochene Fakten. Bei gleichem Gemma-4-Sprachmodell und gleicher Soniox-Spracherkennung lagen die beobachteten P50-Werte der Bestätigungszeit zwischen 1.167 und 1.580 ms. Gadium, Deepgram, Cartesia und ElevenLabs bilden eine schnelle erste Vorauswahl. Auch die anderen fünf Pipelines erfüllten die geprüften Aufgaben.

Gemma 4 31B IT bestand 16 von 16 kurzen strukturierten Entscheidungsprüfungen. Das erste Texttoken kam im Median nach 162 ms. Viele weitere LLMs erreichten ebenfalls 16/16. Der Unterschied beschreibt die Geschwindigkeit in diesem Testprofil; die Studie liefert keinen universellen Intelligenz- oder Qualitätsrang.

Die STT-Ergebnisse unterscheiden sich nach Sprache. Im kleinen deutschen Korpus erreichte Scribe v2 Realtime nach einem gekennzeichneten Recheck einen Wortfehler auf 161 Referenzwörter. Im englischen Korpus erreichten drei AssemblyAI-Pro-Konfigurationen und Cartesia Ink 2 jeweils drei Fehler auf 124 Wörter. Wenige Wörter können hier bereits die Reihenfolge verändern.

Für native Realtime-Modelle zählt die Fortsetzung nach dem Tool-Aufruf. GPT Realtime 2.1 bestand die acht geprüften Service-Aufgaben. Bei weiteren Konfigurationen gab es fehlende gesprochene Bestätigungen, nicht abgeschlossene Suites oder ungeklärte Protokollprobleme. Diese Befunde betreffen die getestete Anbindung und das Testprofil.

Leseregel: Prüfe zuerst, ob eine Konfiguration die Aufgabe richtig erfüllt. Vergleiche danach die Antwortzeiten. Wähle Stimme und Kostenklasse innerhalb einer passenden Vorauswahl.

2. Studiendesign und Testumfang

Ebene	Umfang	Bedeutung
LLM	33 / 528	16 Entscheidungsanfragen je Modell
STT	23 / 304	16 menschliche Clips; Sprachsupport je Route
TTS	29 / 228	4 Sätze je unterstützter Sprache
Native Audiomodelle	9	Modelle in mehreren Modi
WebRTC	44 / 22	Sprachzellen / Konfigurationen, DE+EN
Testweg	WebRTC / EU	Kein kontrollierter PSTN-Vergleich
LiveKit SDK	1.8.6	Ein SDK-Stand im kontrollierten Lauf

Der Katalog ist eine Bestandsaufnahme. Ein aktivierter Eintrag bedeutet nicht, dass jede Sprache, jeder Tarif und jeder Anbindungsweg geprüft wurde. Die Komponenten- und Gesprächstests beantworten unterschiedliche Fragen.

Ein LLM kann korrektes JSON liefern und trotzdem im vollständigen Gespräch eine unpassende Bestätigung erzeugen. Eine TTS kann schnell Audio liefern, während Satzende-Erkennung oder Tool-Fortsetzung den Anrufer warten lassen.

Die 44 WebRTC-Zellen entsprechen 22 Konfigurationen in zwei Sprachen. Es gab einen Anruf pro Sprache und Konfiguration. Vier kontrollierte positive Aufgaben bilden jeweils die Service-Suite. Bei vollständig durchlaufenen Suites kam ein begrenzter Unterbrechungs- und Wiederaufnahme-Test hinzu. Die 47 abgeschlossenen Anrufe im Evidenzbestand enthalten zusätzlich drei Piloten; sie sind kein zusätzlicher Wiederholungssatz der 44 Vergleichszellen.

Die Tests liefen nacheinander. Es gab keine zufällige Reihenfolge und keine Wiederholungsserien. Dadurch können Aufwärmeeffekte, Netzbedingungen und zeitliche Provider-Schwankungen die Reihenfolge beeinflussen. Die Daten sind eine ausführlich dokumentierte Vorauswahl für die nächste Teststufe.

3. Pipeline, Half-Cascade und Realtime verstehen



Konzeptillustration. Die Glasmodule stehen für Erkennung, Entscheidung und Sprachausgabe; sie zeigen keine Messwerte.

Modus	Ablauf	Prüfswertpunkt
Pipeline	STT → LLM → TTS	Erkennung, Tool-Fakten, Stimme
Half-Cascade	Natives Audio → externe TTS	Tool-Fortsetzung und Stimme
Realtime	Natives Audio → natives Audio	Gesprochene Fakten und Aktion
Full Duplex	Überlappende Sprache	Zuhören, Eingreifen, Ergebnis
Translate	DE→EN / EN→DE	Sprachinhalt und wichtige Angaben

Eine Pipeline zerlegt den Weg in Spracherkennung, Entscheidung und Sprachausgabe. Du kannst jede Komponente unabhängig prüfen und austauschen. Half-Cascade verbindet natives Audio-Verständnis mit einer externen Stimme. Realtime erzeugt die Sprachantwort im nativen Audiomodell. Full Duplex prüft zusätzlich das Verhalten bei überlappendender Sprache. Übersetzung ist ein eigener Gesprächszweck und bekommt eine eigene Bewertung.

Die gleiche Metrik muss in allen Modi das gleiche Ereignis beschreiben. Ein erstes Texttoken in einer Pipeline ist nicht das erste Audiotoken eines nativen Modells. Fehlende Komponentenwerte bekommen daher keinen künstlichen Nullwert. Für den Vergleich vollständiger Gespräche nutzen wir den anruferseitigen Bestätigungs-Proxy und weisen die Aufgabenprüfung daneben aus.

4. Der vollständige Pipeline-Vergleich



Bestätigungs-Proxy. Cyan: P50. Dunkel: Stichproben-P95. Acht Turns je Konfiguration.

Anbieter	TTS-Modell	Prüfungen	n	P50 ms	P95 ms
Gadium	default	8/8	8	1167	1272
Deepgram	aura-2	8/8	8	1201	1425
Cartesia	sonic-3.6-2026-08-27	8/8	8	1241	1371
ElevenLabs	eleven_v4_turbo	8/8	8	1250	1326
Rime	coda	8/8	8	1261	1499
Inworld	inworld-tts-2	8/8	8	1269	1392

Anbieter	TTS-Modell	Prüfungen	n	P50 ms	P95 ms
Soniox	tts-rt-v2	8/8	8	1429	1616
Fish Audio	s2.1-pro	8/8	8	1530	1610
xAI	tts-1	8/8	8	1580	3054

Konstant bleiben Gemma 4 31B IT als LLM und Soniox STT v5 als Spracherkennung. Die Tabelle zeigt die getesteten TTS-Konfigurationen. Die Zeit startet am Ende der Anruferäußerung und endet beim ersten empfangenen nicht stillen Audioblock, nachdem der passende HTTP-Testbeleg eingetroffen ist. Sie enthält damit Netzwerk, Aktion, Fortsetzung und Audio-Auslieferung. Der Audioblock beweist für sich allein noch nicht, dass schon das erste bedeutungstragende Bestätigungswort hörbar war.

Alle neun Pipelines bestanden 8/8 Testaufgaben und 2/2 begrenzte Unterbrechungs-/Wiederaufnahme-Prüfungen. Das ist ein gutes Ergebnis für diese festen Fälle. Es ist keine Lösungsquote für echte Kundenanrufe. Das Tool gab einen kontrollierten HTTP-Echo-Beleg zurück. Eine passende Echo-Antwort ist keine tatsächlich gespeicherte Kalenderbuchung oder CRM-Änderung.

Die ersten sechs P50-Werte liegen relativ eng beieinander. Kleine Abstände sind bei acht Antworten und ohne Wiederholungen keine belastbaren Siegerabstände. Die xAI-Konfiguration zeigt in diesem Lauf einen deutlich höheren Stichproben-P95 als ihren P50. Das ist ein Anlass für Wiederholungstests und die Prüfung der Einzelereignisse.

5. Alle nativen Gesprächskonfigurationen

Konfiguration	Prüfungen	Suiten	n	P50 ms	P95 ms
realtime-gpt-realtime-2.1	8/8	2/2	8	2658	2895
half_cascade-gpt-realtime-2.1	7/8	2/2	8	1889	2711
half_cascade-gpt-realtime-2	7/8	2/2	8	2320	4118
half_cascade-gpt-realtime-2.1-mini	6/8	2/2	8	2260	10300
realtime-gpt-realtime-2.1-mini	6/8	2/2	8	2406	12913
realtime-gpt-realtime-2	6/8	2/2	8	2917	4975
realtime-gemini-2.5-flash-native-audio-preview-12-2025	1/2	0/2	2	4508	11117
realtime-gpt-live-1	WAV 6/8	2/2	8	1550	5698
half_cascade-gpt-realtime-mini	0/0	0/2	0	N/A	N/A
realtime-gemini-3.1-flash-live-preview	0/0	0/2	0	N/A	N/A
realtime-gemini-3.8-live	0/0	0/2	0	N/A	N/A
realtime-gemini-3.8-live-extended-thinking	0/0	0/2	0	N/A	N/A
realtime-gpt-realtime-mini	0/0	0/2	0	N/A	N/A

Bei GPT Realtime 2.1 wurden die vollständigen acht Service-Fakten geprüft. Half-Cascade mit GPT Realtime 2.1 und GPT Realtime 2 erreichte jeweils 7/8. Die jeweiligen Mini-Konfigurationen erreichten 6/8. Ihre Antwortzeiten bleiben sichtbar, auch wenn eine Aufgabe die Qualitätsprüfung nicht bestand. Ein schneller fehlgeschlagener Turn wird nicht als erfolgreicher Turn verkauft.

Gemini- und ältere Realtime-Mini-Suiten brachen nach fehlenden passenden Tool-Aufrufen oder fehlender hörbarer Fortsetzung ab. Spätere Aufgaben wurden dann nicht versucht. Deshalb bedeutet 0/0 hier: keine auswertbare Aufgabenprüfung. Es bedeutet weder perfekte Qualität noch vier belegte Fehler pro Sprache. Die nächste Untersuchung muss Anbindung, Ereignisfolge und Fortsetzungslogik prüfen.

Bei Full Duplex mit GPT Live 1 waren die Tool-Belege in beiden Sprachen korrekt. Die unabhängige Erkennung der empfangenen WAV-Dateien bestätigte sechs von acht gesprochenen Fakten: In DE fehlte die 14-Tage-Antwort, in EN die Buchungsbestätigung. Die direkten Modelltranskripte waren unvollständig. Dieser zusätzliche WAV-Check ist daher gesondert ausgewiesen und nicht mit den anderen Transcript-Prüfungen gleichgesetzt. Der Modus bleibt für diese Service-Suite weiter zu qualifizieren.

Translate wurde separat für DE→EN und EN→DE geprüft. In beiden Richtungen kam Audio an; Datum, Uhrzeit und Berlin blieben erhalten. Ein deutsches Nomen blieb in der unabhängigen Erkennung mehrdeutig. Diese zwei Richtungen begründen keine allgemeine Übersetzungsqualitätsrate. Bei gleichzeitig fließender Sprache beschreibt ein negativer Onset-Wert Überlappung, keine negative Reaktionszeit.

6. LLMs: alle 33 Modelle im Entscheidungsprofil

Modell	Prüfungen	Nicht leer	n	P50 ms	P95 ms
google/gemma-4-31b-it	16/16	16	16	162	535
openai/gpt-4.1	16/16	16	16	339	440
openai/gpt-4o	16/16	16	16	389	890
openai/gpt-5.4-mini	16/16	16	16	411	513
deepseek-ai/deepseek-v4.1-flash	16/16	16	16	494	821
openai/gpt-5.1	16/16	16	16	527	633
openai/gpt-5.6-luna	16/16	16	16	561	1664
openai/gpt-5.2	16/16	16	16	575	663
openai/gpt-5.4	16/16	16	16	591	689
openai/gpt-5.6-sol	16/16	16	16	654	2635
google/gemini-3.5-flash-lite	16/16	16	16	698	1213
google/gemini-3.1-flash-lite	16/16	16	16	707	4930
google/gemini-3.6-flash	16/16	16	16	743	1362
google/gemini-3-flash-preview	16/16	16	16	928	1180
google/gemini-3.5-flash	16/16	16	16	1017	3469
openai/gpt-6-luna	16/16	16	16	1145	2434

Modell	Prüfungen	Nicht leer	n	P50 ms	P95 ms
openai/gpt-5.5	16/16	16	16	1235	2645
xai/grok-4.7	16/16	16	16	5860	14607
openai/gpt-4.1-mini	15/16	16	16	316	393
openai/gpt-5.6-terra	15/16	15	15	576	827
xai/grok-4.20-0309-reasoning	15/16	15	15	7864	17818
openai/gpt-5.4-nano	14/16	15	15	400	520
google/gemini-3.7-flash	13/16	16	16	2440	7146
xai/grok-4.20-multi-agent-0309	12/16	14	14	9672	19539
google/gemini-3.8-flash	11/16	16	16	2625	10380
openai/gpt-4.1-nano	10/16	16	16	284	364
openai/gpt-4o-mini	10/16	16	16	345	740
xai/grok-4.20-0309-non-reasoning	7/16	16	16	1070	6367
openai/gpt-5	5/16	6	6	2885	5624
google/gemini-3.1-pro-preview	3/16	16	16	3697	6223
openai/gpt-5-mini	0/16	1	1	2941	2941
openai/chat-latest	0/16	0	0	N/A	N/A
openai/gpt-5-nano	0/16	0	0	N/A	N/A

FAMULOR RESEARCH / OKT 2026

LLM: erstes Texttoken

16 Entscheidungsprüfungen · ausgewählte vollständige Kandidaten

 google/gemma-4-31b-it		162 ms
 openai/gpt-4.1		339 ms
 openai/gpt-4o		389 ms
 openai/gpt-5.4-mini		411 ms
 deepseek-ai/deepseek-v4.1-flash		494 ms
 openai/gpt-5.1		527 ms
 openai/gpt-5.6-luna		561 ms
 openai/gpt-5.2		575 ms

16/16 Entscheidungsprüfungen

0

Pro Modell gab es 16 kurze strukturierte Entscheidungen, mit einer ursprünglichen Ausgabegrenze von 256 Tokens. Geprüft wurden feste Fakten, erwartete Aktionen und die Ablehnung nicht erlaubter oder nicht belegter Handlungen. Die Tabelle hält Entscheidungsprüfung, nicht leere Antworten und Timing-Stichprobe getrennt. P50 und P95 beziehen sich nur auf vorhandene Timing-Werte.

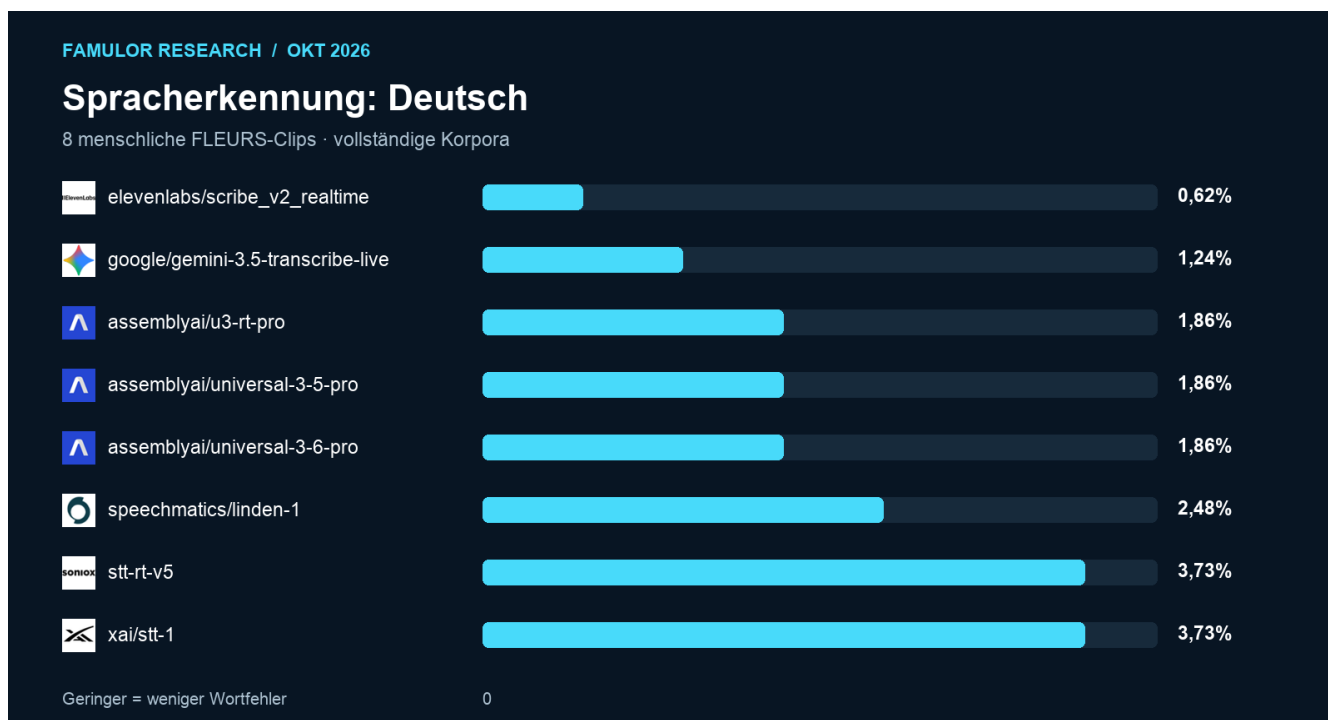
16/16 ist bei einer so kleinen Stichprobe eine grobe Eingangshürde. Selbst unter der vereinfachenden Annahme unabhängiger binärer Prüfungen liegt die untere Grenze eines 95%-Wilson-Intervalls bei ungefähr 80,6 %. Die festen Fälle sind zudem keine zufällige Auswahl echter Kundenaufgaben. Aus dem Intervall folgt daher keine Produktionsgarantie.

Die GPT-5-Rechecks mit 1.536 Ausgabetokens verbesserten nicht leere Antworten. Sie verwenden ein anderes Budget und ersetzen die 256-Token-Baseline nicht stillschweigend. Wenn ein Modell seine Ausgabegrenze fürs Nachdenken verbraucht, muss die Konfiguration geprüft werden. Eine leere Antwort oder ein HTTP-400-Routenfehler beweist keine allgemeine Unfähigkeit, die Kundenfrage zu beantworten.

Für eine erste Pipeline lohnt sich deshalb Gemma 4 zusammen mit weiteren 16/16-Kandidaten wie GPT-4.1, GPT-4o, GPT-5.4 Mini oder DeepSeek V4.1 Flash. Prüfe anschließend längere Kontexte, echte Tool-Argumente, Korrekturen und mehrere Gesprächsschritte. Eine gute kleine Entscheidungsprobe ersetzt diesen zweiten Schritt nicht.

7. STT: menschliche Sprache statt synthetischer Referenz

Die Genauigkeitsprüfung verwendet 16 menschlich gesprochene Aufnahmen aus Google FLEURS. Deutsch und Englisch werden getrennt bewertet. Alle Kandidaten derselben Sprache erhalten dieselben Dateien und Referenztexte. Das Korpus umfasst 82,56 Sekunden Deutsch und 54,14 Sekunden Englisch. Die offiziellen Referenzen enthalten 161 deutsche und 124 englische Wörter.



Auszug vollständiger Sprachkorpora. Die vollständigen Tabellen stehen darunter.

Spracherkennung: Englisch

8 menschliche FLEURS-Clips · vollständige Korpora



Geringer = weniger Wortfehler

0

Auszug vollständiger Sprachkorpora. Die vollständigen Tabellen stehen darunter.

Deutsch: vollständige und gesonderte Ergebnisse

Modell	Profil	Clips	Wörter	Fehler	WER	Einordnung
elevenlabs/scribe_v2_realtime	Zugang/EOF-Recheck	8/8	161	1	0.62%	Vollständiges kurzes Korpus
google/gemini-3.5-transcribe-live	baseline	8/8	161	2	1.24%	Vollständiges kurzes Korpus
assemblyai/u3-rt-pro	baseline	8/8	161	3	1.86%	Vollständiges kurzes Korpus
assemblyai/universal-3-5-pro	baseline	8/8	161	3	1.86%	Vollständiges kurzes Korpus
assemblyai/universal-3-6-pro	baseline	8/8	161	3	1.86%	Vollständiges kurzes Korpus
speechmatics/linden-1	baseline	8/8	161	4	2.48%	Vollständiges kurzes Korpus
stt-rt-v5	Zugang/EOF-Recheck	8/8	161	6	3.73%	Vollständiges kurzes Korpus
xai/stt-1	baseline	8/8	161	6	3.73%	Vollständiges kurzes Korpus
xai/stt-2	baseline	8/8	161	6	3.73%	Vollständiges kurzes Korpus
deepgram/nova-3	baseline	8/8	161	7	4.35%	Vollständiges kurzes Korpus
deepgram/flux-general-multi	baseline	8/8	161	10	6.21%	Vollständiges kurzes Korpus

Modell	Profil	Clips	Wörter	Fehler	WER	Einordnung
deepgram/nova-2	baseline	8/8	161	12	7.45%	Vollständiges kurzes Korpus
cartesia/ink-whisper	baseline	8/8	161	13	8.07%	Vollständiges kurzes Korpus
solaria-1	Protokoll-Recheck	5/8	112	7	6.25%	Partiell; ohne Rang
assemblyai/universal-streaming-multilingual	baseline	8/8	161	176	109.32%	Doppelte Finals; ohne Rang

Englisch: vollständige und gesonderte Ergebnisse

Modell	Profil	Clips	Wörter	Fehler	WER	Einordnung
assemblyai/u3-rt-pro	baseline	8/8	124	3	2.42%	Vollständiges kurzes Korpus
assemblyai/universal-3-5-pro	baseline	8/8	124	3	2.42%	Vollständiges kurzes Korpus
assemblyai/universal-3-6-pro	baseline	8/8	124	3	2.42%	Vollständiges kurzes Korpus
cartesia/ink-2	baseline	8/8	124	3	2.42%	Vollständiges kurzes Korpus
elevenlabs/scribe_v2_realtime	Zugang/EOF-Recheck	8/8	124	4	3.23%	Vollständiges kurzes Korpus
deepgram/flux-general-en	baseline	8/8	124	8	6.45%	Vollständiges kurzes Korpus
google/gemini-3.5-transcribe-live	baseline	8/8	124	9	7.26%	Vollständiges kurzes Korpus
stt-rt-v5	Zugang/EOF-Recheck	8/8	124	10	8.06%	Vollständiges kurzes Korpus
speechmatics/linden-1	baseline	8/8	124	10	8.06%	Vollständiges kurzes Korpus
deepgram/flux-general-multi	baseline	8/8	124	12	9.68%	Vollständiges kurzes Korpus
deepgram/nova-3	baseline	8/8	124	14	11.29%	Vollständiges kurzes Korpus
deepgram/nova-3-pharma	baseline	8/8	124	16	12.90%	Vollständiges kurzes Korpus
deepgram/nova-2-conversationalai	baseline	8/8	124	17	13.71%	Vollständiges kurzes Korpus
deepgram/nova-2-medical	baseline	8/8	124	18	14.52%	Vollständiges kurzes Korpus
deepgram/nova-2-phonecall	baseline	8/8	124	18	14.52%	Vollständiges kurzes Korpus

Modell	Profil	Clips	Wörter	Fehler	WER	Einordnung
deepgram/nova-3-medical	baseline	8/8	124	18	14.52%	Vollständiges kurzes Korpus
deepgram/nova-2	baseline	8/8	124	20	16.13%	Vollständiges kurzes Korpus
solaria-1	Protokoll-Recheck	4/8	51	3	5.88%	Partiell; ohne Rang
cartesia/ink-whisper	baseline	6/8	99	29	29.29%	Partiell; ohne Rang
xai/stt-1	baseline	6/8	99	33	33.33%	Partiell; ohne Rang
xai/stt-2	baseline	6/8	99	33	33.33%	Partiell; ohne Rang
assemblyai/universal-streaming	baseline	8/8	124	84	67.74%	Doppelte Finals; ohne Rang
assemblyai/universal-streaming-multilingual	baseline	7/8	114	120	105.26%	Doppelte Finals; ohne Rang

Die Micro-WER ist die Summe aus ersetzten, ausgelassenen und zusätzlich erkannten Wörtern, geteilt durch sämtliche Referenzwörter. Sie ist kein Mittelwert von acht Clip-Prozenten. Ein zusätzlicher Fehler verändert die DE-WER um ungefähr 0,62 Prozentpunkte und die EN-WER um ungefähr 0,81 Prozentpunkte. Ein DE+EN-Pool würde Deutsch durch die höhere Wortzahl stärker gewichten.

Partielle Korpora stehen gesondert. Ihre WER beschreibt nur die erfolgreich ausgewerteten Ausschnitte. Sie darf nicht neben einem vollständigen Korpus als fairer Rang verwendet werden. Ein fehlendes Transkript ist keine fehlerfreie Erkennung. Die AssemblyAI-Universal-Streaming-Artefakte mit doppelten finalen Hypothesen bleiben ebenfalls außerhalb der vergleichbaren Reihenfolge. Eine WER über 100 % ist rechnerisch möglich, wenn sehr viele Wörter zusätzlich gezählt werden. Hier muss zuerst die Segment- und Formatter-Verarbeitung geklärt werden.

Die FLEURS-Auswahl stammt aus dem Validation-Split und enthält vorgelesene Sprache. Die acht ausgewählten deutschen Aufnahmen tragen männliche Metadaten; Englisch enthält vier männliche und vier weibliche Metadaten. Das ist keine repräsentative Stichprobe für Dialekte, spontane Sprache, Geräusche oder Telefonie. Auch unterschiedliche Sprecheridentitäten sind damit nicht belegt. Ergänze für deine Abnahme eigene Namen, Kundennummern, Uhrzeiten, Produktbegriffe und Telefon-Audio.

8. TTS: alle 29 Modelle und der erste Audioblock

TTS: erster Audioblock

8 kurze Sätze · lokale PCM-TTFFB · Auszug



8/8 Audiosamples 0

Zehn vollständige Baseline-Kandidaten. Lokale PCM-TTFFB, keine Bewertung der Natürlichkeit.

Modell	Profil	Audio	P50 ms	P95 ms
eleven_v3_conversational	baseline	8/8	125	702
elevenlabs/eleven_v4_turbo	baseline	8/8	129	141
elevenlabs/eleven_flash_v2_5	baseline	8/8	138	1265
elevenlabs/eleven_turbo_v2_5	baseline	8/8	179	225
cartesia/sonic-latest	baseline	8/8	182	231
cartesia/sonic-3.6-2026-08-27	baseline	8/8	186	222
cartesia/sonic-preview	baseline	8/8	187	231
gradium/default	baseline	8/8	192	264
cartesia/sonic-3.6	baseline	8/8	197	205
cartesia/sonic-3-latest	baseline	8/8	217	1274
cartesia/sonic-3.5	baseline	8/8	219	250
cartesia/sonic-3.5-2026-05-04	baseline	8/8	236	243
cartesia/sonic-3	baseline	8/8	238	325
cartesia/sonic-3-2026-01-12	baseline	8/8	260	293
inworld/inworld-tts-2-flash	baseline	8/8	338	676

Modell	Profil	Audio	P50 ms	P95 ms
inworld/inworld-tts-1.5-mini	baseline	8/8	357	578
inworld/inworld-tts-2	baseline	8/8	359	654
inworld/inworld-tts-1.5-max	baseline	8/8	415	454
deepgram/aura-2	baseline	8/8	476	762
elevenlabs/eleven_v3	baseline	8/8	487	766
deepgram/flux-tts	flux_model_voice_recheck	4/4	538	740
rime/coda	baseline	8/8	601	637
elevenlabs/eleven_v4	baseline	8/8	638	775
soniox/tts-rt-v2	credential_or_eof_recheck	8/8	718	771
elevenlabs/eleven_multilingual_v2	baseline	8/8	868	1920
xai/tts-1	baseline	8/8	876	949
fishaudio/s2.1-pro	baseline	8/8	1191	1758
fishaudio/s2-pro	baseline	8/8	1237	1317
gpt-4o-mini-tts	baseline	3/8	1154	1550

Pro Sprache wurden vier feste kurze Sätze mit einer festen Stimme geprüft. Die lokale TTFB läuft vom Start der Factory-/Provider-Anforderung bis zum ersten PCM-Audioblock. Sie endet vor dem hörbaren Ergebnis beim Anrufer. Die Tabelle dokumentiert daher die Komponente und nicht die vollständige Gesprächszeit.

Eleven v3 Conversational, Eleven v4 Turbo, Flash, aktuelle Cartesia-Konfigurationen und Gradium waren in diesem Profil schnelle Kandidaten. Das Ranking kann sich in einer vollständigen Pipeline ändern. Deepgram Aura-2 hatte beispielsweise einen höheren lokalen Komponentenwert als mehrere andere Modelle, während die vollständige Deepgram-Pipeline im Gespräch eine der niedrigeren Bestätigungszeiten zeigte. Die Messgrenzen, Antworttexte und Streaming-Wege unterscheiden sich.

Flux TTS wurde mit seiner passenden Alexis-Stimme in vier englischen Sätzen erneut geprüft. Dieser EN-Recheck ist separat und nicht gleichwertig zu einer vollständigen zweisprachigen Baseline. Für GPT-4o Mini TTS kamen im Baseline-Profil drei von acht Samples zustande; der Timing-Wert betrifft diese drei Samples und nicht acht erfolgreich geprüfte Antworten.

Eine blinde Hörjury für Natürlichkeit, Aussprache oder Stimmpräferenz wurde nicht durchgeführt. Höre die Vorauswahl daher mit deinem Wortschatz: Personen- und Firmennamen, E-Mail-Adressen, Dezimalzahlen, Uhrzeiten, Abkürzungen und Sprachwechsel. Bewerte Verständlichkeit zuerst, Markenpassung danach. Ein schneller Audioblock gibt dir noch keine Aussage über beides.

9. Kosten als Low, Medium und High

Pipeline	Kostenklasse	Datenbasis
Gradium	High	Bekannter LLM-/TTS-Anteil; Gesamtkosten ungeklärt

Pipeline	Kostenklasse	Datenbasis
Deepgram	Medium	Bekannter LLM-/TTS-Anteil; Gesamtkosten ungeklärt
Cartesia	High	Bekannter LLM-/TTS-Anteil; Gesamtkosten ungeklärt
ElevenLabs	Ungeklärt	TTS-Tarif oder Abrechnungsmengen fehlen
Rime	Medium	Bekannter LLM-/TTS-Anteil; Gesamtkosten ungeklärt
Inworld	Low	Bekannter LLM-/TTS-Anteil; Gesamtkosten ungeklärt
Soniox	Ungeklärt	TTS-Tarif oder Abrechnungsmengen fehlen
Fish Audio	Low	Bekannter LLM-/TTS-Anteil; Gesamtkosten ungeklärt
xAI	Low	Bekannter LLM-/TTS-Anteil; Gesamtkosten ungeklärt

Diese Klassen beschreiben ausschließlich den relativen bekannten LLM-/TTS-Inferenzanteil von sieben vergleichbaren Pipeline-Paaren. Grundlage sind jeweils die beiden vollständig durchlaufenen DE/EN-Anrufe, ohne Piloten, mit belegbaren LLM- und TTS-Mengen sowie zugeordneten öffentlichen Tarifreferenzen. Die sieben bekannten Teilwerte werden aufsteigend geordnet: die ersten drei erhalten Low, die nächsten zwei Medium und die letzten zwei High. Die Klassen sind relativ zu diesem Test und keine Preisstufen von Famulor.

STT-Tokenmengen, Teile der Detektor-Nutzung, Transport, Infrastruktur, Aufzeichnung, Rabatte und Rechnungskosten fehlen oder sind nicht vollständig zugeordnet. Bei Eleven v4 Turbo fehlt der exakte TTS-Tarif; bei Soniox fehlen passende Abrechnungsmengen. Beide erhalten deshalb Ungeklärt. Native Modi werden nicht in diese Pipeline-Klassen hineingemischt. Unterschiedliche Antwortlängen und die Begrüßungs-/Unterbrechungsanteile beeinflussen die bekannten Teilwerte.

Low bedeutet nicht günstigste Gesamtkosten. High bedeutet nicht hoher Famulor-Tarif. Die Einteilung hilft dir, einen vergleichbar erfassten Teil des Testverbrauchs zu verstehen. Kosten pro tatsächlich gelöstem Kundengespräch werden erst dann belastbar, wenn sämtliche Kosten und echte Geschäftsergebnisse gemeinsam zugeordnet sind. Dieser Report veröffentlicht keine internen Geldbeträge.

10. Vier Anwendungsfälle, vier Abnahmefragen

Ablauf	Entscheidende Prüfung	Nächste Teststufe
Service	Richtlinie und Kennung	Unsicherheit und bestätigte Übergabe
Termine	Slot, Zeitzone, eine Aktion	Gespeicherten Kalenderzustand prüfen
Vertrieb	Fakten und Qualifizierung	Unbekannte Angebote und fehlende Angaben
Support	Schritte und Passwortschutz	Korrektur, Tool-Fehler, Unterbrechung

Die folgenden Tabellen zeigen die beiden Sprachbeobachtungen als Spanne. Jede Kombination enthält genau eine positive Aufgabe pro Sprache. Der kleinere Wert ist daher kein belastbarer Branchenmedian. Nur vollständig bestandene Aufgabenpaare erscheinen hier. Die Abstände erlauben keine Aussage, welcher Anbieter speziell im Kundenservice oder Vertrieb die beste Qualität hat.

Kundenservice

Konfiguration	Geprüfte Paare	Min ms	Max ms
pipeline-gradium	2	779	1095
pipeline-rime	2	874	1261
pipeline-inworld	2	881	1185
pipeline-soniox	2	1001	1350
pipeline-cartesia	2	1049	1204
pipeline-deepgram	2	1178	1308
pipeline-elevenlabs	2	1250	1261
pipeline-fishaudio	2	1284	1530
pipeline-xai	2	1445	3054
half_cascade-gpt-realtime-2.1-mini	2	1471	1551
half_cascade-gpt-realtime-2.1	2	1780	1994
half_cascade-gpt-realtime-2	2	1990	2320
realtime-gpt-realtime-2.1-mini	2	2157	2260
realtime-gpt-realtime-2	2	2400	2496
realtime-gpt-realtime-2.1	2	2677	2710

Prüfe freigegebene Richtlinien, Kennungen, Korrekturen und den Wunsch nach einem Menschen. Die kontrollierte Aufgabe prüfte einen bekannten Richtlinien-Fakt und den passenden Testbeleg. Ein echter Service-Abnahmesatz braucht zusätzlich unvollständige Angaben, widersprüchliche Wünsche und bestätigte Übergaben.

Terminvergabe

Konfiguration	Geprüfte Paare	Min ms	Max ms
pipeline-gradium	2	1128	1272
pipeline-elevenlabs	2	1168	1326
pipeline-deepgram	2	1186	1360
pipeline-cartesia	2	1241	1352
pipeline-rime	2	1244	1480
pipeline-inworld	2	1269	1392
pipeline-soniox	2	1400	1616
pipeline-fishaudio	2	1481	1610
pipeline-xai	2	1555	1582

Konfiguration	Geprüfte Paare	Min ms	Max ms
realtime-gpt-realtime-2.1	2	2658	2688

Das Ziel ist der richtige angebotene Slot, mit Datum, Uhrzeit und Zeitzone. Prüfe nach einer Testbuchung den Kalenderzustand. Verlange genau eine gespeicherte Aktion und eine passende gesprochene Bestätigung. Ergänze belegte Termine, Rückfragen, Änderungen und wiederholte Tool-Zustellung, damit eine Antwort nicht versehentlich doppelt bucht.

Vertrieb

Konfiguration	Geprüfte Paare	Min ms	Max ms
pipeline-gradium	2	1167	1250
pipeline-deepgram	2	1174	1425
pipeline-rime	2	1232	1469
pipeline-cartesia	2	1239	1292
pipeline-elevenlabs	2	1247	1257
pipeline-inworld	2	1249	1359
pipeline-soniox	2	1429	1509
pipeline-fishaudio	2	1521	1540
pipeline-xai	2	1664	1731
half_cascade-gpt-realtime-2.1	2	1827	1990
half_cascade-gpt-realtime-2	2	1878	2485
realtime-gpt-realtime-2.1-mini	2	2128	3026
realtime-gpt-realtime-2.1	2	2537	2616
realtime-gpt-realtime-2	2	4728	4975

Geprüft wurden feste Produktangaben und die passende Aktion. In der nächsten Stufe gehören unsichere Anfragen, fehlende Angaben und nicht freigegebene Angebote dazu. Die genannten Preis- und Seat-Fakten waren fiktive Testdaten. Sie sind keine Famulor-Preise und werden hier nicht als Angebot veröffentlicht.

Technischer Support

Konfiguration	Geprüfte Paare	Min ms	Max ms
pipeline-deepgram	2	1201	1268
pipeline-gradium	2	1205	1209
pipeline-elevenlabs	2	1237	1318

Konfiguration	Geprüfte Paare	Min ms	Max ms
pipeline-inworld	2	1269	1318
pipeline-cartesia	2	1319	1371
pipeline-rime	2	1430	1499
pipeline-soniox	2	1478	1487
pipeline-fishaudio	2	1539	1590
pipeline-xai	2	1562	1580
half_cascade-gpt-realtime-2.1	2	1889	2206
half_cascade-gpt-realtime-2.1-mini	2	1899	2541
half_cascade-gpt-realtime-2	2	2346	2444
realtime-gpt-realtime-2.1	2	2354	2895
realtime-gpt-realtime-2.1-mini	2	2406	12913
realtime-gpt-realtime-2	2	2476	2954

Der kontrollierte Fall nutzte einen bekannten Fehlercode, eine feste Schrittfolge und den Schutz von Passwörtern. Prüfe in der Praxis zusätzlich fehlgeschlagene Schritte, neue Informationen und eine Unterbrechung mitten in der Anleitung. Eine gute Wiederaufnahme greift den korrigierten Bedarf auf und liest nicht einfach den alten Absatz weiter.

11. Antwortzeiten ohne Messfehler lesen

Metrik	Start → Ende	Aussage
Bestätigungs-Proxy	Anruferende → Audio nach Beleg	Hörbarer Onset, kein Wortzeitstempel
LLM TTFT	Anforderung → erstes Texttoken	Komponenten-Startzeit
TTS TTFB	Factory → erster PCM-Block	Komponente, kein Anrufer-Audio
EOT	Erfasstes Ende → Turn-Entscheidung	SDK-Wert, eigener Grenztest nötig
Micro WER	$(S + D + I) / N$	Wortfehler pro Referenzwort
P50 / P95	Nearest rank	Immer n und Profil angeben

Für die Tabellen gilt das Nearest-Rank-Verfahren: Nach dem Sortieren wird der Wert an Position $\text{ceil}(p \times n)$ verwendet. Bei $n=8$ ist P95 damit der höchste der acht Werte. Er beschreibt den langsamsten beobachteten Turn dieser Stichprobe. Ein stabiler Produktions-P95 braucht einen deutlich größeren und passend geschichteten Datensatz.

Die Turns eines Anrufs teilen Konfiguration und Netzbedingungen. Acht Turns sind deshalb nicht acht unabhängige Kundenanrufe. Für Unsicherheitsintervalle und Wiederholungen musst du den Anruf als gemeinsame Einheit berücksichtigen. Eine reine Turn-Sammlung kann sonst mehr Sicherheit suggerieren, als der Versuch tatsächlich

liefert.

Exportierte Worker-Metriken der Pipelines

Pipeline	LLM n/P50/P95	TTS n/P50/P95	EOT n/P50/P95
pipeline-cartesia	12 / 149 / 175	13 / 67 / 120	12 / 578 / 582
pipeline-deepgram	12 / 148 / 159	13 / 137 / 884	12 / 578 / 579
pipeline-elevenlabs	12 / 148 / 179	14 / 84 / 93	12 / 578 / 580
pipeline-fishaudio	12 / 144 / 155	12 / 466 / 1344	12 / 578 / 579
pipeline-gradium	12 / 145 / 159	14 / 90 / 149	12 / 578 / 580
pipeline-inworld	13 / 146 / 204	13 / 186 / 309	12 / 578 / 580
pipeline-rime	12 / 146 / 182	14 / 235 / 549	12 / 578 / 579
pipeline-soniox	13 / 148 / 163	15 / 309 / 404	12 / 578 / 580
pipeline-xai	12 / 145 / 157	12 / 418 / 745	12 / 578 / 2501

Diese SDK-Werte umfassen je nach Export auch Begrüßungen, Tool-Fortsetzungen und Unterbrechungsantworten. Deshalb unterscheiden sich die Stichprobengrößen von den acht kontrollierten Aufgaben. P50-Werte einzelner Komponenten lassen sich nicht zum gemessenen Gesprächs-P50 addieren: Die Zeitachsen überlappen und die Beobachtungsmengen unterscheiden sich.

12. Was ein vollständiges Reporting erfassen muss

Kennzahl	Hier belegt	Zusätzliche Evidenz
Tool-/Faktenprüfung	Fiktive Belege + gesprochene Fakten	Echte Aktionszustände
Lösungsquote	Nicht gemessen	Gelabelte echte Aufgaben und Nenner
Halluzinationsrate	Feste Fakten-/Ablehnungsprüfungen	Unabhängige Prüfung sämtlicher Aussagen
Weiterleitung	Keine neue Telefonabnahme	Freigegebenes Ziel und belegte Annahme
Abbruchrate	Begrenzte Disconnect-Beobachtung	Vollständige beidseitige Ereignisse
Natürlichkeit	Keine blinde Hörjury	Hörtests mit identischen Texten
Kosten pro Lösung	Nicht gemessen	Vollständige Kosten und echte Lösung

Ein technisch abgeschlossener Anruf ist kein belegter gelöster Kundenfall. Ein angenommener Weiterleitungsbefehl beweist keinen menschlichen Gesprächspartner. Eine passende gesprochene Aussage beweist keine dauerhaft gespeicherte Aktion. Das Reporting muss diese Zustände getrennt erfassen und mit der erwarteten Aufgabe verbinden.

Speichere pro Testfall Sprache, Konfiguration, Prompt-/Tool-Version, Modellroute, Zeitpunkt, erwartete Aktion, erhaltenen Beleg und gesprochenes Ergebnis. Markiere Bestanden, Nicht bestanden, Nicht versucht und Ungeklärt

getrennt. Zeige für jede Kennzahl den passenden Nenner. So bleibt eine gute Latenzprobe sichtbar, ohne ein fehlgeschlagenes Ergebnis zu verstecken.

Die historischen Produktionsdaten aus der Vorwoche waren zusätzliche Betriebsevidenz. Sie werden hier nicht in die kontrollierte Rangfolge eingerechnet: Aufgaben, Kunden, Einstellungen und Releases unterschieden sich. Diese Trennung verhindert, dass eine Beobachtung aus dem Betrieb als Ursache-Wirkung-Vergleich von Modellen erscheint.

13. Famulor-API für reproduzierbare Tests nutzen

Zweck	API-Vertrag	Einsatz
Katalog	GET /user/assistants/models	Verfügbare Modell-IDs dokumentieren
Stimmen	GET /user/assistants/voices	Stimme und Sprache festhalten
Provider	GET /user/assistants/transcriber-providers	STT-Anbindung dokumentieren
Testassistent	POST /user/assistant	Isolierte Testkonfiguration
Anruf	GET /user/calls/{id}	Anrufrdaten und Ergebnis verbinden
Post-Call	Webhook	Transkript, Variablen und Aufzeichnung

Die bereitgestellte OpenAPI-Datei beschreibt den Produkt-API-Vertrag. Sie liefert weder Benchmark-Messungen noch eine fertige Qualitätspunktzahl. Nutze die Katalog-Endpunkte, um die verfügbaren Modelle und Stimmen zum Testzeitpunkt zu dokumentieren. Konfiguriere einen separaten Testassistenten und ein Tool mit kontrollierten Antworten. Ein echter Testanruf wird erst mit freigegebener Zielnummer ausgeführt.

Über die Anrufabfrage und Post-Call-Webhooks kannst du die dokumentierten Anrufrdaten, Transkripte, extrahierten Variablen und Aufzeichnungsreferenzen verbinden. Für millisekundengenaue Satzenden, Audiobeginn, Tool-Belege und Unterbrechungen brauchst du zusätzliche Instrumentierung des Sprachwegs. Call-Dauer und Status reichen dafür nicht. Behaupte insbesondere keine TTFT-, TTFB- oder WER-Felder, wenn dein tatsächlicher Antwortvertrag sie nicht liefert.

Authentifizierung erfolgt serverseitig mit einem Bearer-Token. Halte Zugangsdaten, Kundennummern, private Aufzeichnungen und interne Kosten außerhalb öffentlicher Reports. Für Veröffentlichung und Teamfreigabe reichen aggregierte Tabellen, klar beschriebene Fälle und redigierte Belege.

14. Von der Vorauswahl zur eigenen Freigabe

- Definiere je Anwendungsfall die erwartete Aktion, die hörbare Bestätigung und die erlaubte Reaktion auf Unsicherheit.
- Wähle mehrere vollständig geprüfte Kandidaten. Halte Prompt, Wissensstand, Tool-Antworten, Region und Anrufer-Audio konstant.
- Wiederhole die Fälle in zufälliger Reihenfolge. Trenne kalte und warme Starts. Erfasse Fehlversuche vollständig.
- Ergänze spontane Sprache, verschiedene Stimmen, Telefon-Codecs, Hintergrundgeräusche, Pausen, Korrekturen und Paketverlust.

- Prüfe echte Testsystem-Zustände, etwa eine Kalenderbuchung. Bewerte das empfangene Audio unabhängig vom Modelltranskript.
- Lege vor dem Versuch Qualitäts- und Latenzgrenzen fest. Entscheide anhand dieser Grenzen und der Unsicherheit, nicht anhand eines nachträglich gewichteten Sieger-Scores.
- Wiederhole die Abnahme nach Änderungen an Prompt, Tools, Modellroute oder Engine. Bewahre die Regressionsergebnisse zusammen mit den Versionsdaten auf.

Famulor-Simulationen geben dir einen Einstieg in wiederholbare Ablaufprüfungen. Ergänze gesprochene End-to-End-Tests, wenn der Kunde Audio hört und während einer Antwort eingreifen kann. Die im Forschungsprojekt verfassten Simulationsfälle sind Testvorlagen; ihre Existenz belegt keinen durchgeführten zusätzlichen Testlauf.

15. Empfehlung für deine erste Konfiguration

Für kurze Service-Aufgaben mit festen Tool-Fakten ist eine Pipeline die stärkste erste Testbasis in diesem Screening. Beginne mit Gemma 4 und einer sprachlich passenden STT. Vergleiche danach mehrere erfolgreiche Stimmen, insbesondere Gadium, Deepgram, Cartesia und ElevenLabs, und höre deine kritischen Namen und Zahlen an. Nutze die Kostenklasse als ergänzende Teilbeobachtung.

Wenn du natives Audio-Verständnis mit einer bestimmten externen Stimme verbinden willst, prüfe Half-Cascade gezielt. Wenn du native Sprachantworten brauchst, ist GPT Realtime 2.1 ein Kandidat aus dieser vollständig geprüften kleinen Suite. Für Full Duplex stehen überlappende Sprache und hörbare Ergebnisvollständigkeit im Mittelpunkt. Die Auswahl folgt deinem Ablauf und deiner Abnahme.

16. Quellen und Reproduzierbarkeit

Primäre Messgrundlage ist der versionierte Famulor-Forschungsstand vom 13. Oktober 2026. Die redaktionell übernommenen Tabellen entsprechen dem finalen Ergebnisstand; frühere Review-Notizen werden nicht als neuere Ergebnisse behandelt. Geldbeträge, private Call-IDs und Zugangsdaten sind aus dieser öffentlichen Fassung ausgeschlossen.

- Google FLEURS: Dataset Card. Menschliche Sprachaufnahmen und Referenztexte, CC BY 4.0.
- Conneau et al.: FLEURS, 2022. Korpusbeschreibung.
- Creative Commons BY 4.0. Lizenz und Attribution.
- NIST: ASR Metrics. Einordnung der Wortfehlerrate.
- LiveKit: Agent Metrics. SDK-Metriken und ihre Messgrenzen.
- Famulor API-Dokumentation. Produkt-API und Ereignisse.

Die FLEURS-Auswahl verwendet de_de und en_us, Validation, Dataset-Revision 70bb2e84b976b7e960aa89f1c648e09c59f894dd. DE-Zeilen: 0, 1, 2, 3, 5, 6, 8, 201. EN-Zeilen: 0, 2, 3, 4, 5, 6, 7, 8. Audio wurde in PCM16, Mono, 16 kHz konvertiert; diese Umwandlung ist als Modifikation dokumentiert. Korpusautoren und Sprecher sind keine Unterstützer dieses Reports. Die normalisierten Referenzen und sämtliche Einzelereignisse gehören in den eigenen Reproduktionsnachweis.

Vergleiche die Famulor Engine-Modi, höre Stimmen in der Stimmenbibliothek und plane deine eigene Testkonfiguration. Demo anfragen.