



FAMULOR RESEARCH

Voice AI Benchmark October 2026

Models. Real calls. Measured results.



FAMULOR RESEARCH / OCTOBER 2026

Voice AI Benchmark Choosing your voice AI models

94 catalog entries · 1,060 component requests
44 WebRTC language cells · DE + EN

Measured 13 October 2026. Small sequential samples. Complete tables, explicit measurement boundaries and relative cost bands.

Contents

1. The findings that matter	3
2. Study design and coverage	3
3. Understand the engine architectures	4
4. The complete pipeline comparison	5
5. Every native conversation configuration	6
6. LLMs: all 33 models in the decision profile	7
7. STT: human speech and supplied references	9
8. TTS: all 29 models and the first audio chunk	12
9. Cost bands: Low, Medium and High	14
10. Four workflows and their acceptance questions	14
11. Read timing at the right boundary	17
12. What complete reporting needs	18
13. Use the Famulor API for reproducible checks	19
14. Turn the shortlist into an acceptance plan	19
15. Choose your first configuration	20
16. Sources and reproducibility	20

A voice agent needs to take the right action and confirm it clearly. Fast speech generation covers only part of that job. This Famulor study compares individual models, complete conversation configurations and the audio the caller receives. You get the full result tables, relative cost bands and an acceptance plan you can apply to your own workflows.

1. The findings that matter

All nine tested pipeline configurations passed eight out of eight fictional receipt and spoken-fact checks. With Gemma 4 and Soniox recognition held constant, their observed confirmation-onset P50 ranged from 1,167 to 1,580 ms. Gadium, Deepgram, Cartesia and ElevenLabs form a fast initial shortlist. The other five pipelines also completed the checked tasks.

Gemma 4 31B IT passed all 16 short structured-decision checks, with a median first-text-token time of 162 ms. Many other LLMs also passed 16/16. The timing describes this particular profile. It does not establish a universal intelligence or quality ranking.

Speech recognition results differ by language. Scribe v2 Realtime, in a separately labeled recheck, made one edit over 161 German reference words. Three AssemblyAI Pro configurations and Cartesia Ink 2 each made three edits over 124 English words. A few words can change the ordering in a corpus this small.

For native realtime models, tool continuation is a key check. GPT Realtime 2.1 passed the eight service tasks. Other configurations showed missing spoken confirmations, unfinished suites or unresolved protocol behavior. These findings apply to the tested route and profile.

Reading rule: Establish correct task completion first. Compare response timing second. Choose the voice and cost band within a suitable shortlist.

2. Study design and coverage

Layer	Coverage	Meaning
LLM	33 / 528	16 decision requests per model
STT	23 / 304	16 human clips; language eligibility by route
TTS	29 / 228	4 phrases per eligible language
Native audio models	9	Models across multiple modes
WebRTC	44 / 22	Language cells / configurations, DE+EN
Test route	WebRTC / EU	No controlled PSTN comparison
LiveKit SDK	1.8.6	One SDK version in the controlled run

The enabled catalog is an inventory snapshot. It does not prove that every language, plan or integration route was qualified. Component and conversation tests answer different questions. An LLM can produce correct JSON while the complete agent gives an unsuitable spoken confirmation. Fast TTS can still sit behind slow turn detection or tool continuation.

The 44 WebRTC cells represent 22 configurations in two languages, with one call per language and configuration. Each service suite contains four controlled positive tasks. Completed suites also include a bounded interruption and recovery check. The 47 terminal calls in the evidence set include three pilots. They do not create repeated trials of the 44 comparison cells.

Runs were sequential, without randomized order or repeat series. Warm-up effects, network conditions and provider changes over time may influence the observations. Treat this as a documented screening stage for your next qualification round.

3. Understand the engine architectures



Concept illustration. Glass modules represent recognition, reasoning and speech generation, without measured values.

Mode	Path	Qualification focus
Pipeline	STT → LLM → TTS	Recognition, tool facts, voice
Half-Cascade	Native audio → external TTS	Tool continuation and voice
Realtime	Native audio → native audio	Spoken facts and action
Full Duplex	Overlapping speech	Listening, interruptions, outcome
Translate	DE→EN / EN→DE	Content and critical entities

A pipeline separates recognition, reasoning and speech generation. Each component can be evaluated and changed independently. Half-Cascade combines native audio understanding with an external voice. Realtime produces speech in the native audio model. Full Duplex adds overlapping-speech behavior. Translation has a different purpose and receives a separate evaluation.

Equivalent metrics must describe equivalent events. A pipeline's first text token is not a native model's first audio token. Missing component timings therefore remain missing. For complete calls, this report uses the caller-side confirmation proxy alongside explicit task checks.

4. The complete pipeline comparison



Confirmation proxy. Cyan: P50. Dark: sample P95. Eight turns per configuration.

Provider	TTS model	Checks	n	P50 ms	P95 ms
Gradium	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
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

Gemma 4 31B IT and Soniox STT v5 remain constant. The table names the tested TTS configurations. Timing starts at the end of caller speech and ends at the first received non-silent audio frame after the matching HTTP fixture receipt arrives. It includes network, action, continuation and audio delivery. That frame alone does not verify the timestamp of the first meaningful confirmation word.

Every pipeline passed 8/8 task checks and 2/2 bounded interruption/recovery checks. This is a positive result for these fixed cases. It is not a customer-call resolution rate. The tool returned a controlled HTTP echo receipt, rather than writing an appointment or changing a CRM record.

The first six P50 values are relatively close. With eight turns and no repeat trials, small gaps do not establish separated winners. The xAI configuration has a noticeably higher sample P95 than its P50 in this run. That calls for repeated trials and an inspection of the individual events.

5. Every native conversation configuration

Configuration	Checks	Suites	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

GPT Realtime 2.1 completed the eight checked service facts. Half-Cascade with GPT Realtime 2.1 and GPT Realtime 2 each reached 7/8. Their Mini counterparts reached 6/8. Timings remain visible when a task fails its quality check. A quick failed turn is not counted as a successful result.

Gemini and older Realtime Mini suites stopped after missing matching tool calls or audible continuation. Subsequent tasks were not attempted. A 0/0 entry therefore means no evaluable task checks. It does not mean perfect quality or four proven failures per language. Further work needs to isolate route, event sequence and continuation behavior.

For Full Duplex with GPT Live 1, both language cells had correct tool receipts. Independent recognition of received WAV files confirmed six of eight spoken facts: the German policy response lacked the 14-day fact, and English lacked the booking confirmation. Direct model transcripts were incomplete. The WAV check is shown separately rather than treated as identical to transcript-based checks. This service suite still needs audible-result qualification.

Translate was evaluated separately in DE→EN and EN→DE. Both delivered audio and preserved date, time and Berlin entities. One German noun remained ambiguous in independent recognition. Two directions do not establish a broad translation quality rate. A negative streaming onset represents overlap, not a negative reactive response time.

6. LLMs: all 33 models in the decision profile

Model	Checks	Nonempty	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
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

Model	Checks	Nonempty	n	P50 ms	P95 ms
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



LLM TTFT P50 · 16/16

Each model received 16 short structured decisions, with an original output cap of 256 tokens. Checks covered fixture facts, expected actions and refusal of unsupported or disallowed actions. The table separates decision checks, nonempty outputs and timing samples. P50 and P95 use only observations with available timings.

16/16 is a coarse entry gate with this sample size. Even under a simplified independent-binomial assumption, the lower 95% Wilson bound is approximately 80.6%. These fixed cases are also not a random sample of customer requests. The interval therefore does not provide a production guarantee.

Separate GPT-5 rechecks using a 1,536-token allowance improved nonempty responses. They use a different budget and do not silently replace the 256-token baseline. When a reasoning model consumes its allowance before producing output, examine the configuration. An empty output or HTTP 400 route error does not establish an intrinsic inability to answer the customer's question.

Gemma 4 and other 16/16 candidates such as GPT-4.1, GPT-4o, GPT-5.4 Mini and DeepSeek V4.1 Flash are sensible starting points. Follow with longer contexts, real tool arguments, corrections and multi-step conversations.

A small decision screen does not replace that second stage.

7. STT: human speech and supplied references

The accuracy screen uses 16 human recordings from Google FLEURS. German and English are scored separately. Every candidate within a language receives the same files and reference texts. The selection contains 82.56 seconds of German and 54.14 seconds of English, with 161 and 124 reference words respectively.



Selected complete language corpora. Full tables follow.



Selected complete language corpora. Full tables follow.

German: complete and separately qualified results

Model	Profile	Clips	Words	Edits	WER	Qualification
elevenlabs/scribe_v2_realtime	credential_or_eof_recheck	8/8	161	1	0.62%	complete short corpus
google/gemini-3.5-transcribe-live	baseline	8/8	161	2	1.24%	complete short corpus
assemblyai/u3-rt-pro	baseline	8/8	161	3	1.86%	complete short corpus
assemblyai/universal-3-5-pro	baseline	8/8	161	3	1.86%	complete short corpus
assemblyai/universal-3-6-pro	baseline	8/8	161	3	1.86%	complete short corpus
speechmatics/linden-1	baseline	8/8	161	4	2.48%	complete short corpus
stt-rt-v5	credential_or_eof_recheck	8/8	161	6	3.73%	complete short corpus
xai/stt-1	baseline	8/8	161	6	3.73%	complete short corpus
xai/stt-2	baseline	8/8	161	6	3.73%	complete short corpus
deepgram/nova-3	baseline	8/8	161	7	4.35%	complete short corpus
deepgram/flux-general-multi	baseline	8/8	161	10	6.21%	complete short corpus
deepgram/nova-2	baseline	8/8	161	12	7.45%	complete short corpus
cartesia/ink-whisper	baseline	8/8	161	13	8.07%	complete short corpus
solaria-1	gladia_protocol_recheck	5/8	112	7	6.25%	partial; unranked
assemblyai/universal-streaming-multilingual	baseline	8/8	161	176	109.32%	duplicate-final artifact; unranked

English: complete and separately qualified results

Model	Profile	Clips	Words	Edits	WER	Qualification
assemblyai/u3-rt-pro	baseline	8/8	124	3	2.42%	complete short corpus
assemblyai/universal-3-5-pro	baseline	8/8	124	3	2.42%	complete short corpus
assemblyai/universal-3-6-pro	baseline	8/8	124	3	2.42%	complete short corpus

Model	Profile	Clips	Words	Edits	WER	Qualification
cartesia/ink-2	baseline	8/8	124	3	2.42%	complete short corpus
elevenlabs/scribe_v2_realtime	credential_or_eof_recheck	8/8	124	4	3.23%	complete short corpus
deepgram/flux-general-en	baseline	8/8	124	8	6.45%	complete short corpus
google/gemini-3.5-transcribe-live	baseline	8/8	124	9	7.26%	complete short corpus
stt-rt-v5	credential_or_eof_recheck	8/8	124	10	8.06%	complete short corpus
speechmatics/linden-1	baseline	8/8	124	10	8.06%	complete short corpus
deepgram/flux-general-multi	baseline	8/8	124	12	9.68%	complete short corpus
deepgram/nova-3	baseline	8/8	124	14	11.29%	complete short corpus
deepgram/nova-3-pharma	baseline	8/8	124	16	12.90%	complete short corpus
deepgram/nova-2-conversationalai	baseline	8/8	124	17	13.71%	complete short corpus
deepgram/nova-2-medical	baseline	8/8	124	18	14.52%	complete short corpus
deepgram/nova-2-phonecall	baseline	8/8	124	18	14.52%	complete short corpus
deepgram/nova-3-medical	baseline	8/8	124	18	14.52%	complete short corpus
deepgram/nova-2	baseline	8/8	124	20	16.13%	complete short corpus
solaria-1	gladia_protocol_recheck	4/8	51	3	5.88%	partial; unranked
cartesia/ink-whisper	baseline	6/8	99	29	29.29%	partial; unranked
xai/stt-1	baseline	6/8	99	33	33.33%	partial; unranked
xai/stt-2	baseline	6/8	99	33	33.33%	partial; unranked
assemblyai/universal-streaming	baseline	8/8	124	84	67.74%	duplicate-final artifact; unranked
assemblyai/universal-streaming-multilingual	baseline	7/8	114	120	105.26%	duplicate-final artifact; unranked

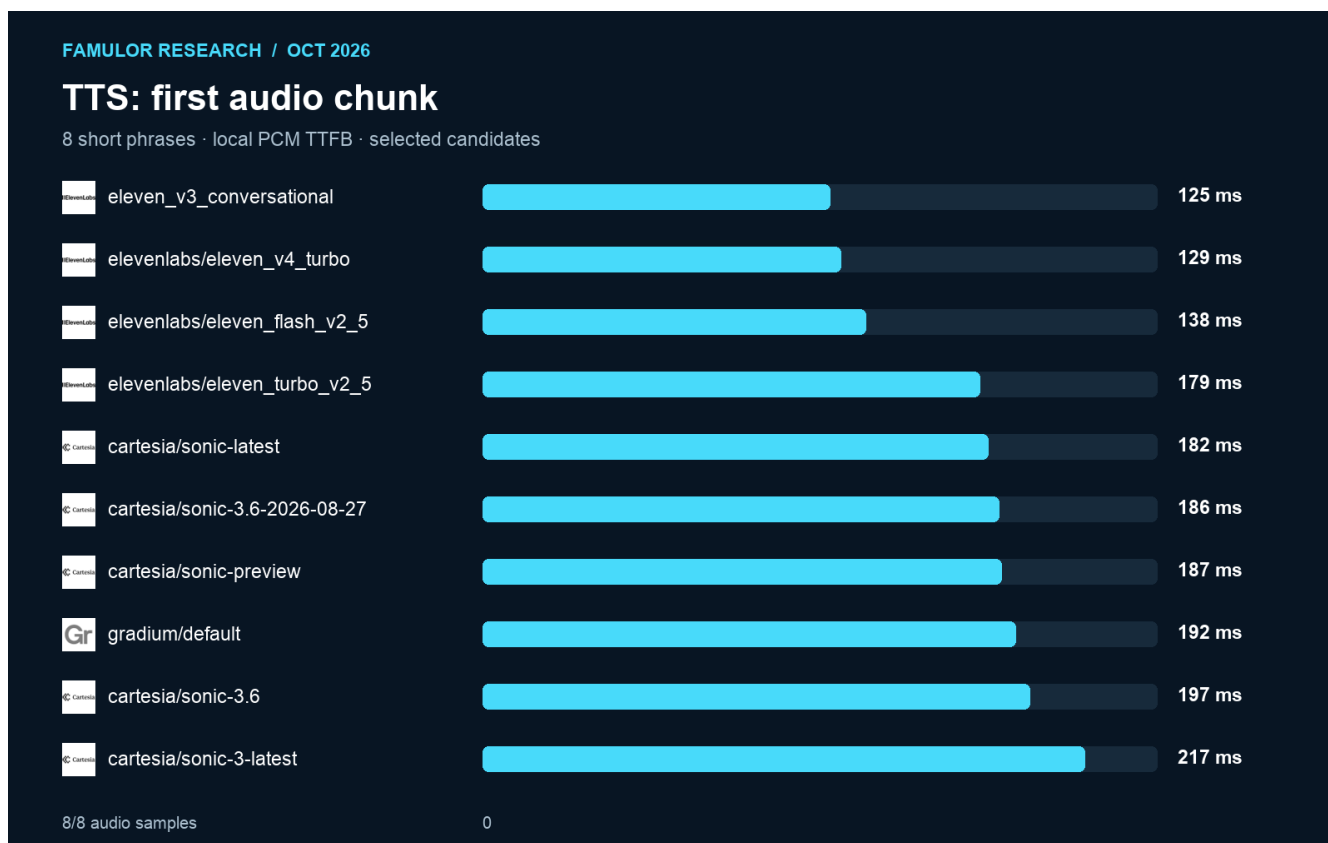
Micro WER is total substitutions, deletions and insertions divided by total reference words. It is not the average of eight clip-level percentages. One extra edit changes German WER by about 0.62 percentage points and English by about 0.81 points. Pooling DE and EN would give German more weight through its larger word count.

Partial corpora are identified separately. Their WER describes only successfully evaluated clips and cannot fairly rank against complete corpora. Missing transcripts are not error-free recognition. AssemblyAI Universal Streaming

results with duplicated final hypotheses are also outside the comparative ordering. WER can exceed 100% when many extra words are inserted. Segment and formatter handling need qualification before interpreting those artifacts as recognition quality.

The selection comes from the validation split and contains read speech. German recordings have male metadata; English has four male and four female entries. This is not representative coverage of accents, spontaneous speech, noise or telephony. It also does not establish distinct speaker identities. Add your own names, customer identifiers, times, product terms and telephone audio before acceptance.

8. TTS: all 29 models and the first audio chunk



Ten complete baseline candidates. Local PCM TTFB, not a naturalness rating.

Model	Profile	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

Model	Profile	Audio	P50 ms	P95 ms
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
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

Each language uses four fixed short phrases and one fixed voice. Local TTFB runs from factory/provider request start to the first PCM chunk. It ends before the caller's audible response. This is a component measurement rather than complete conversation timing.

Eleven v3 Conversational, Eleven v4 Turbo, Flash, recent Cartesia configurations and Gradium were fast candidates in this profile. Their order can change in a complete pipeline. Deepgram Aura-2, for example, showed a higher local component time than several models, while the complete Deepgram pipeline had one of the lower caller confirmation times. Boundaries, generated text and streaming paths differ.

Flux TTS was rechecked with its compatible Alexis voice on four English phrases. That EN-only recheck is explicitly separate from bilingual baseline rows. GPT-4o Mini TTS returned three out of eight baseline samples; its timings cover those three available outputs rather than eight successful answers.

No blinded jury rated naturalness, pronunciation or voice preference. Listen to your shortlist using your vocabulary: personal and company names, email addresses, decimals, times, abbreviations and language switches. Judge intelligibility before brand fit. A fast first chunk does not measure either.

9. Cost bands: Low, Medium and High

Pipeline	Cost band	Evidence
Gradium	High	Known LLM/TTS share; complete cost unresolved
Deepgram	Medium	Known LLM/TTS share; complete cost unresolved
Cartesia	High	Known LLM/TTS share; complete cost unresolved
ElevenLabs	Unresolved	Missing TTS tariff or billing quantities
Rime	Medium	Known LLM/TTS share; complete cost unresolved
Inworld	Low	Known LLM/TTS share; complete cost unresolved
Soniox	Unresolved	Missing TTS tariff or billing quantities
Fish Audio	Low	Known LLM/TTS share; complete cost unresolved
xAI	Low	Known LLM/TTS share; complete cost unresolved

These bands describe only the relative known LLM/TTS inference share for seven comparable pipeline pairs. Each pair uses the two completed DE/EN calls, excludes pilots, and has attributable LLM and TTS quantities with referenced public tariffs. Order the seven known subtotals: the first three receive Low, the next two Medium and the last two High. The bands are relative to this experiment, not Famulor plan tiers.

STT token quantities, some detector usage, transport, infrastructure, recording, discounts and invoice costs are missing or incompletely attributed. The exact Eleven v4 Turbo tariff is unavailable; Soniox lacks matching billing quantities. Both remain Unresolved. Native modes are not inserted into these pipeline bands. Response lengths, greetings and interruption responses also affect the known shares.

Low does not mean the cheapest complete stack. High does not mean a high Famulor tariff. The bands help interpret a comparable recorded portion of test usage. Actual cost per resolved customer call needs complete cost attribution and verified business outcomes. This report publishes no internal monetary amounts.

10. Four workflows and their acceptance questions

Workflow	Key check	Next qualification
Service	Policy and identifier	Uncertainty and confirmed handoff
Booking	Slot, timezone, one action	Verify persisted calendar state
Sales	Facts and qualification	Unknown offers and missing details
Support	Steps and password protection	Correction, tool error, interruption

The workflow tables show the two language observations as a range. Each combination has one positive task per language. The lower observation is not a reliable industry median. Only checked passing task pairs appear. These differences cannot establish a specialist quality winner for customer service or sales.

Customer service

Configuration	Checked turns	Lower ms	Higher 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

Check approved policies, identifiers, corrections and requests for a person. The controlled task checked a known policy fact and its fixture receipt. A real acceptance suite also needs incomplete information, conflicting requests and confirmed handoffs.

Appointment booking

Configuration	Checked turns	Lower ms	Higher 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

Configuration	Checked turns	Lower ms	Higher ms
pipeline-soniox	2	1400	1616
pipeline-fishaudio	2	1481	1610
pipeline-xai	2	1555	1582
realtime-gpt-realtime-2.1	2	2658	2688

The target is the correct offered slot, including date, time and timezone. Inspect actual test-calendar state after booking. Require exactly one persisted action and a matching spoken confirmation. Add unavailable slots, clarifications, rescheduling and repeated tool delivery to detect duplicate bookings.

Sales

Configuration	Checked turns	Lower ms	Higher 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

Checks covered fixed product facts and the expected action. The next stage needs uncertain requests, missing details and unapproved offers. Price and seat facts in the fixtures were fictional test data. They are not Famulor prices and are not reproduced as commercial offers.

Technical support

Configuration	Checked turns	Lower ms	Higher ms
pipeline-deepgram	2	1201	1268

Configuration	Checked turns	Lower ms	Higher ms
pipeline-gradium	2	1205	1209
pipeline-elevenlabs	2	1237	1318
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

The controlled case used a known error code, a fixed step sequence and password protection. Add failed steps, new information and an interruption midway through the instructions. Correct recovery addresses the changed request rather than continuing the previous paragraph.

11. Read timing at the right boundary

Metric	Start → end	Meaning
Confirmation proxy	Caller end → audio after receipt	Audio onset, not a word timestamp
LLM TTFT	Request → first text token	Component start time
TTS TTFB	Factory → first PCM chunk	Component, not caller audio
EOT	Detected end → turn decision	SDK value; independent boundary check needed
Micro WER	$(S + D + I) / N$	Word edits per reference word
P50 / P95	Nearest rank	Always show n and profile

Tables use nearest-rank percentiles: sort observations and select position $\text{ceil}(p \times n)$. With $n=8$, P95 is the largest of eight observations. It describes the slowest measured turn in that sample. A stable production P95 needs a substantially larger, appropriately stratified dataset.

Turns within a call share configuration and network conditions. Eight turns are not eight independent customer calls. Confidence intervals and repeats need to respect the call as a shared unit. Pooling turns can otherwise suggest more certainty than the experiment supports.

Exported pipeline worker metrics

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

Depending on the export, these SDK metrics include greetings, tool continuations and interruption responses. Their sample counts differ from the eight controlled tasks. Component P50 values cannot be added to derive the conversation P50: timelines overlap and observation sets differ.

12. What complete reporting needs

Metric	Evidence here	Additional evidence needed
Tool/fact check	Fictional receipts + spoken facts	Real action state
Resolution rate	Unmeasured	Labeled real outcomes and denominator
Hallucination rate	Fixed fact/refusal checks	Independent annotation of material claims
Transfer	No new phone acceptance test	Approved destination and verified pickup
Drop rate	Bounded disconnect observation	Complete observations from both peers
Naturalness	No blinded listener jury	Listener tests with identical scripts
Cost per resolution	Unmeasured	Complete costs and real resolution

A technically completed call is not a verified resolved customer case. An accepted transfer command does not prove human pickup. A matching spoken statement does not prove a persisted action. Reporting must record these states separately and connect them to the expected task.

For each case, retain language, configuration, prompt/tool version, model route, time, expected action, received receipt and spoken result. Keep Passed, Failed, Unattempted and Unresolved distinct. Give every metric its proper

denominator. A good timing observation then remains useful without hiding an unsuccessful outcome.

Historical production aggregates from the preceding week were additional operational evidence. They are excluded from the controlled ordering because customers, tasks, settings and releases differed. That prevents observational operations data from becoming a causal model comparison.

13. Use the Famulor API for reproducible checks

Purpose	API contract	Use
Catalog	GET /user/assistants/models	Record available model IDs
Voices	GET /user/assistants/voices	Pin voice and language
Providers	GET /user/assistants/transcriber-providers	Record STT route
Test assistant	POST /user/assistant	Isolated test configuration
Call	GET /user/calls/{id}	Join call data and outcome
Post-Call	Webhook	Transcript, variables and recording

The supplied OpenAPI file describes the product API contract. It supplies neither benchmark results nor a ready-made quality score. Use catalog endpoints to record available models and voices at test time. Configure a separate test assistant and a tool with controlled responses. Place a real test call only with an approved destination.

Call retrieval and post-call webhooks connect the documented call information, transcripts, extracted variables and recording references. Millisecond speech-end, audio-onset, tool-receipt and interruption measurements require additional voice-path instrumentation. Call duration and status are insufficient. Do not claim TTFT, TTFB or WER fields unless your actual response contract provides them.

Authentication uses a server-side Bearer token. Keep credentials, customer numbers, private recordings and internal amounts out of public reports. Aggregated tables, clear cases and redacted evidence are suitable for sharing.

14. Turn the shortlist into an acceptance plan

- Specify the intended action, audible confirmation and allowed response to uncertainty for every workflow.
- Choose several fully checked candidates. Hold prompt, knowledge, tool responses, region and caller audio constant.
- Repeat cases in randomized order. Separate cold and warm starts. Preserve all failed attempts.
- Add spontaneous speech, varied voices, telephone codecs, noise, pauses, corrections and packet loss.
- Check real test-system state, such as a calendar booking. Independently assess received audio as well as model transcripts.
- Set quality and timing gates before the trial. Decide using those gates and uncertainty rather than a retrospectively weighted winner score.

- Rerun acceptance after changes to prompt, tools, route or engine. Keep regression results with the corresponding versions.

Famulor simulations provide a starting point for repeatable workflow checks. Add spoken end-to-end trials when customers hear audio and can interrupt responses. Authored simulation scenarios are templates; their presence does not prove an additional executed evaluation.

15. Choose your first configuration

For short service tasks grounded in fixed tool facts, a pipeline is the strongest starting point in this screening. Begin with Gemma 4 and language-appropriate STT. Compare several successful voices, especially Gadium, Deepgram, Cartesia and ElevenLabs, using your critical names and numbers. Treat cost bands as supplementary partial observations.

For native audio understanding paired with a specific external voice, qualify Half-Cascade. For native spoken responses, GPT Realtime 2.1 is a candidate from this small completed suite. For Full Duplex, focus on overlapping speech and complete audible results. Your workflow and acceptance criteria determine the choice.

16. Sources and reproducibility

The primary measurement source is Famulor's versioned 13 October 2026 research snapshot. Tables use its final corrected results; earlier review notes do not override later results. Monetary amounts, private call IDs and credentials are excluded from this public edition.

- Google FLEURS: Dataset Card. Human recordings and supplied transcripts, CC BY 4.0.
- Conneau et al.: FLEURS, 2022. Corpus description.
- Creative Commons BY 4.0. License and attribution.
- NIST: ASR Metrics. Word-error-rate interpretation.
- LiveKit: Agent Metrics. SDK metrics and measurement boundaries.
- Famulor API documentation. Product API and events.

The FLEURS selection uses de_de and en_us, validation, revision 70bb2e84b976b7e960aa89f1c648e09c59f894dd. German rows: 0, 1, 2, 3, 5, 6, 8, 201. English rows: 0, 2, 3, 4, 5, 6, 7, 8. Audio was converted to PCM16 mono at 16 kHz, with that modification recorded. Corpus authors and speakers do not endorse this report. Normalized references and individual events belong in a reproduction record.

Compare Famulor engine modes, listen in the voice library, and plan your test configuration. Book a demo.