

TECHNICAL REPORT

Voice AI Agent Platform Evaluation

Methodology, grading specification and fairness procedure

Version 0.4 · 27 October 2026 · Manifest v5

ABSTRACT

We specify a benchmark for voice AI agent platforms: the hosted runtimes that answer a phone-grade call, take turns, execute tools and connect speech-to-text, language and speech-synthesis models. The agent, its tools, its business backend, the simulated callers and the grader are held constant; only the platform changes. Each platform hosts one production-length agent per industry (healthcare, legal, customer support) and receives 630 simulated calls: 90 tasks run five times, and 36 audio-degraded variants run five times. A call passes only if the platform's tool calls, read from our tool server's request log, cover the task's expected calls with matching constrained arguments, and the call's database after hangup equals the task's expected state. The platform score (0–100) combines task success (50), a word-error composite over the platform's own transcript of the caller (25) and per-turn response latency measured on our recording (25). Model layers are fixed per platform by a published rule, and failures caused by a model Bluejay selected are excluded from that platform's task score. This document defines every quantity, algorithm, threshold and parameter, the attribution and void procedure that keeps the comparison fair, and the records retained for reproduction.

CONTENTS

1	Scope and principles
2	Evaluation environment
3	Simulated callers
4	Task suite
5	Grading specification
6	Metrics and score
7	Model configuration
8	Attribution and fairness
9	Platform profile
10	Records and reproducibility
11	Threats to validity
A	The 90 scored tasks
B	Tool inventory
C	A complete task
D	Pinned model fields
E	Notation and references

1 Scope and principles

The unit of evaluation is the **platform**: the hosted runtime that receives the call, detects the end of the caller's turn, handles interruptions, executes tool calls and connects the model layers. A platform is evaluated as it ships, on its own cloud, through its standard ingress. The task suite and business environments are derived from the methodology of MIVAS-Bench [4]; this evaluation holds the agent and environment fixed and varies the platform.

1.1 Design principles

	PRINCIPLE	CONSEQUENCE
P1	Hold the agent constant	One system prompt, greeting and tool schema per industry, deployed byte-for-byte on every platform. No per-platform prompt tuning.
P2	Grade outcomes, not utterances	A pass is decided by the tool server's request log and the database at hangup. No language model decides a pass.
P3	Measure on our side	Tool calls, state, transcripts of record and latency come from systems Bluejay controls, so a platform that exports nothing is measured identically.
P4	Platform defaults	Turn-taking, interruption, silence and timeout settings are left at the platform's defaults; a platform is scored as a buyer would receive it.
P5	Freeze before running	Tasks, callers, voices, model pins, tool server build and thresholds are frozen in a versioned manifest. Any change re-runs every platform.
P6	Charge only the platform	Failures caused by Bluejay (task, caller, harness) are voided and re-run everywhere; failures of a model Bluejay chose are excluded from the platform's task score.

1.2 Scope of the claims

THE RESULTS SUPPORT	THE RESULTS DO NOT SUPPORT
Under a fixed agent, fixed tasks and the model rule of §7, platform X resolved a share p of 450 scored calls.	That platform X has the best underlying LLM, speech-to-text or text-to-speech model.
Platform X's own transcript of the simulated callers had composite word-error score C under clean and degraded audio.	That results transfer to other languages, accents, audio channels or industries.
Platform X replied with median and 95th-percentile turn latency m and q , as heard by the caller.	What the platform achieves after vendor-specific tuning for a customer.
Platforms whose 95% intervals overlap on a component are tied on it.	A strict order where the intervals cannot support one.

2 Evaluation environment

Every call runs in the same closed environment: a synthetic business with a relational database and a typed tool API, a simulated caller, and the platform between them.

2.1 Simulated businesses

INDUSTRY	COMPANY AND FOOTPRINT	MODELED ON	BOUNDARIES THE AGENT MUST RESPECT
Healthcare	Straus Dermatology: 160 locations, 380 providers; medical dermatology, Mohs, allergy, cosmetic, billing, clinical liaison	A deployed multi-agent dermatology front desk [1]	Identity verification before any chart, balance or result; plan- and office-dependent coverage; fee disclosure before cancellation; results status reported, never interpreted; refills routed, never approved
Legal	Halverson & Reed: plaintiff-side injury, employment, workers' compensation and consumer matters, multi-state	ABA Model Rule 1.18 [2]	Conflict screening before facts are heard; practice-area, jurisdiction and limitation checks before intake; hold, read-back, explicit confirmation before booking; no advice, valuation or acceptance decisions
Customer support	Kestrel Electronics: about 1,000 stores, two membership tiers, delivery and installation, protection plans, legacy brands	The public capability surface of a national retailer's generative-AI support line	Tier- and product-dependent return windows; state-dependent restocking fees; price-match exclusions; cancellation proration; recalls; impersonation reports

All companies, people, locations, records and identifiers are synthetic. The organizational boundaries and workflow constraints are the real ones.

2.2 The agent

Each business was built in four stages, each validated before the next: (i) industry research (agent deployment descriptions, operating policies, regulation, call-centre request taxonomies and recurring failure cases from production evaluations); (ii) a company specification (footprint, departments, services, policies, fees, escalation duties, seed records); (iii) workflow mapping, which assigns every call class to the specialist that owns it and encodes authorization gates, decision points and escalation routes; (iv) implementation as production-length prompts, typed tools, a seeded database and server-side policy guards. The resulting topologies differ by design (Figure 1).

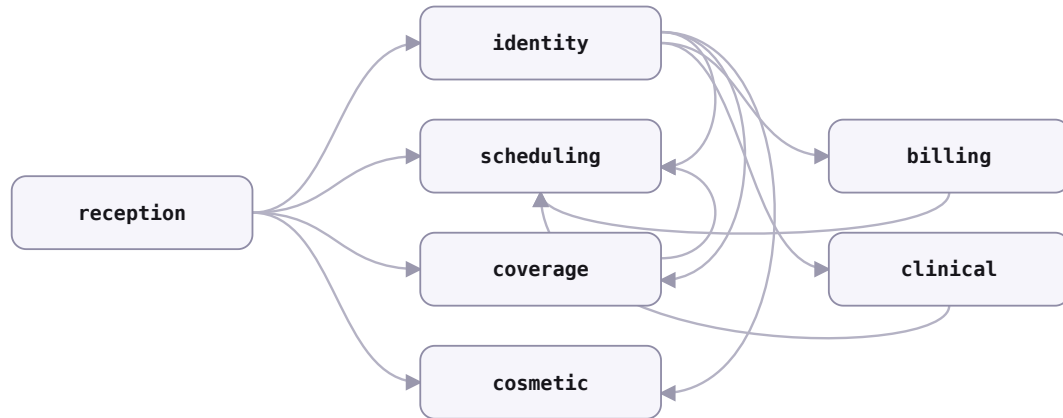
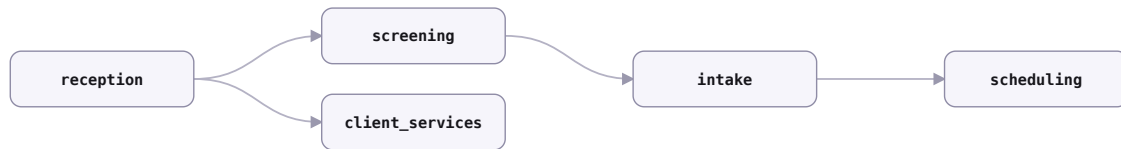
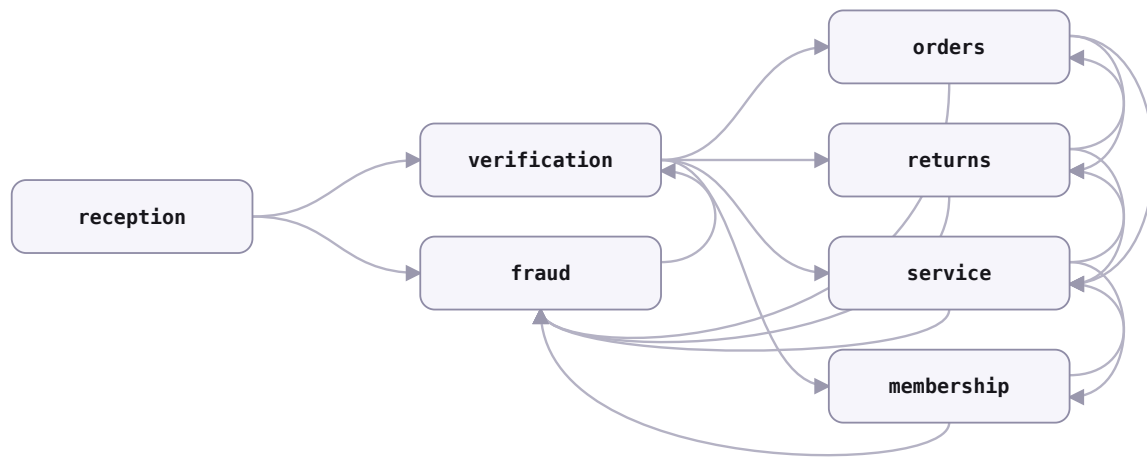
Healthcare · Straus Dermatology**Legal · Halverson & Reed****Customer support · Kestrel Electronics**

Figure 1. Specialist-agent graphs. Nodes are specialist agents; arrows are handoff tools. Healthcare is a privacy-gated acyclic front desk (identity is the only route to billing and clinical); legal is a sequential admissibility funnel; customer support is a cyclic service network for compound requests. Tool inventories per agent are in Appendix B.

Most platforms run one agent per call, so each department is collapsed into a single agent: the specialists' instructions are merged into one system prompt, handoff tools are removed, and two session tools are added, `transfer_to_human` and `end_call`. Business tools, database, policy guards and expected end states are unchanged; expected handoff paths become empty.

INDUSTRY	SPECIALISTS	BUSINESS TOOLS	SINGLE-AGENT TOOLS	SYSTEM PROMPT
Healthcare	7	35	29	3,475 words
Legal	5	24	20	2,131 words
Customer support	7	39	33	3,523 words

2.3 Tool server and database state

Each industry has one tool server; it is the system of record. Every tool request carries the call's identifier (the simulation result id, propagated as the HTTP header `X-Mivas-Call-Id`; platforms whose tool webhooks are separate requests resolve it from the provider's call id). On the first request for a call identifier the server creates a new SQLite file from the industry's `schema.sql` and `seed.sql`; every later request for that call executes against that file. Every call therefore starts from the identical state Do and no call can observe another.

Every request is appended to a per-call log table with the tool name, parsed arguments, raw body, response, HTTP status and start and finish timestamps. At hangup the database is frozen and exported. The verifier (§5) reads only these two artifacts, so a tool call counts if and only if it reached our server. Platforms that execute tools inside their own runtime are configured with functions that forward every invocation to the same endpoint.

2.4 Run protocol

STEP	DEFINITION
Freeze	Version the manifest: task pack commit, agent configurations, caller scripts and voices, model pins (§7), tool server build, caller region, call order. Any later change creates a new version that every platform re-runs.
Deploy	Push the generated configuration through the platform's API, or a documented console path where none exists, onto the platform's production cloud in its default region.
Validate	Before any scored call, exercise every path and every tool once: audio both ways, tool requests with correct arguments, state export, recording and transcript capture. Not scored.
Run	90 tasks × 5 plus 36 variants × 5 = 630 calls, at a fixed concurrency per platform at its entry-level limit, recorded with the results.
Monitor	Each result is compared with its task as it arrives; systematic inconsistencies are raised during the run (§8.4).
Grade	Apply §5 to every call and §6 to every transcript and recording.
Attribute	Assign every failed call exactly one cause (§8.2); void and re-run Bluejay-caused failures until every task has five graded runs.
Review	Each platform receives its results, failed-call recordings and attributions before publication and may flag factual errors.

Bluejay records every call in stereo, caller on the left channel and agent on the right. Calls originate from one cloud region.

3 Simulated callers

Callers are Bluejay digital humans: LLM-driven callers configured from the task, speaking through a fixed synthetic voice. A caller improvises wording but not facts.

3.1 Configuration

FIELD	SEMANTICS
<code>intent</code>	The caller's goal and conduct in prose: opening line, what to accept or refuse, when to end the call. Given to the caller model verbatim, prefixed with the scenario date.
<code>traits</code>	Typed key–value facts (name, phone, date of birth, identifiers, postal code). The only identifying values the caller may state.
<code>scripted_responses</code>	Rules (<code>match_type</code> , <code>match_phrase</code> , <code>response_value</code> , <code>occurrence_mode</code>). With <code>match_type = context</code> , the agent's latest turn is matched against the natural-language trigger; on a match the caller says <code>response_value</code> verbatim. <code>occurrence_mode</code> is <code>always</code> or a first-n limit.
<code>behaviors.creativity</code>	The caller model's temperature; 0 for every scored task.
<code>metadata</code>	Category, difficulty tier and audio condition; not visible to the caller.
Scenario clock	A fixed date for the industry pack, stated to both caller and agent, so date arithmetic is identical across runs.

Every fact and decision the grade depends on is fixed by a scripted response, so it reaches the agent in identical words on every platform. Example: task T4-H1 carries 7 rules (Appendix C). Callers do not barge in. The scripted-response matcher sees the agent's latest turn only; triggers that depend on call history are therefore written as explicit exclusions on the latest turn.

3.2 Voices

Eight voice slots, frozen in the manifest: American female, American male, American (second) female, mature male, Southern female, mature female, American (second) female, Southern male. Scored tasks are sorted by identifier and assigned slots in rotation; each audio variant inherits its base task's voice. A task therefore has the same voice on every platform and run.

3.3 Audio profiles

An audio profile is a deterministic transform of the caller's outbound audio. The words, voice and expected end state are unchanged, so any difference in outcome is the platform's response to the signal. The first easy task (E1) of each of the 18 categories is cloned under two profiles, giving 36 variants; a variant is marked by `metadata.audio_condition` $\in \{ \text{background_noise}, \text{bad_signal} \}$.

PARAMETER	CLEAN	BACKGROUND NOISE (-BG)	DEGRADED LINE (-SIG)
Ambient track	none	Recorded road traffic, separate track mixed into the call at gain 0.8 (legal: 0.1)	none

PARAMETER	CLEAN	BACKGROUND NOISE (-BG)	DEGRADED LINE (-SIG)
Sample rate, framing	TTS native	TTS native	8 kHz, 20 ms packets (160 samples)
Hiss	—	—	Gaussian, -33 dBFS
Crackle	—	—	Poisson, 22 events/s at -13 dBFS
Band limit	—	—	4th-order Butterworth band-pass, 300–3,400 Hz (G.712)
Codec chain	—	—	AMR-NB 4.75 kbps (bit-error rate 3.6×10^{-3}) → G.711 μ -law 64 kbps → AMR-NB 4.75 kbps
Packet loss	—	—	Gilbert–Elliott two-state model, 12% mean loss, mean burst 4.5 packets
Jitter	—	—	Gaussian, $\sigma = 36$ ms; 5 stalls/min of 200–700 ms; packets later than a 60 ms jitter buffer are lost
Concealment	—	—	G.711 Appendix I (pitch repetition, fade, mute at 60 ms) with comfort noise

Loss and jitter follow the ITU-T G.1050 network profiles; codecs follow the NISQA and NeMo telephony recipes. Only the caller's speech passes through the degraded-line chain; the ambient track of the background-noise profile is not degraded. The legal gain is lower because at 0.8 the traffic masked spoken ten-digit callback numbers, which every legal intake requires.

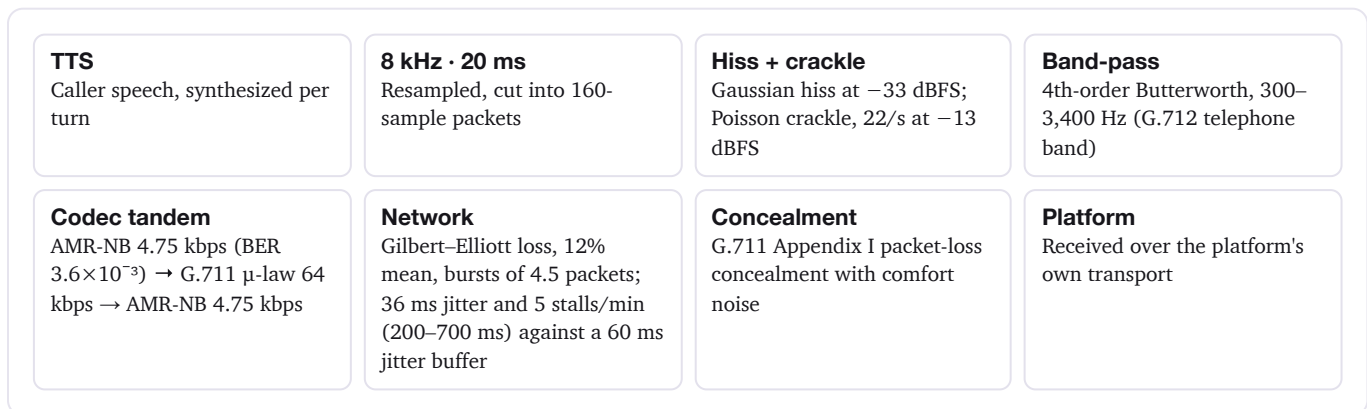


Figure 2. Degraded-line signal chain. Stages in order of application, per 20 ms packet.

4 Task suite

4.1 Definition

A task is a tuple $\tau = (I, F, R, T^*, D^*, \mu)$. I , F and R configure the caller (§3.1). T^* is the multiset of expected tool calls, each a tool name with constrained argument values; D^* is the expected database at hangup; μ holds category, difficulty and audio condition. The caller never sees T^* or D^* .

D^* is not written by hand. It is generated by replaying the authorized tool sequence against a fresh copy of D_0 through the tool server, so every expected state is reachable by construction and consistent with the server's policy guards.

4.2 Certification

Before release, every task was run ten times in full calls and every failure attributed (§8.2). Across the pool, 0% of failures were task faults, 0% harness, 2% caller and 98% model. Tasks failing repeatedly were reviewed from transcripts, traces and database diffs and kept only where the failure was the model's.

4.3 Composition

Each industry has six call categories. Each category contributes five scored tasks (its E1, two medium, two hard): 30 per industry, 90 in total (18 easy, 36 medium, 36 hard). The E1 of each category is also the source of that category's two audio variants. The scored list is fixed in the manifest and given in Appendix A.

INDUSTRY	CODE	CATEGORY	SCORED
Healthcare	C1	new patient access	5
Healthcare	C2	appointment management	5
Healthcare	C3	coverage and benefits	5
Healthcare	C4	cosmetic concierge	5
Healthcare	C5	billing and payments	5
Healthcare	R	regulatory adherence	5
Legal	C1	reception routing	5
Legal	C2	conflicts and barred	5
Legal	C3	eligibility gates	5
Legal	C4	intake and documents	5
Legal	C5	fees and booking	5
Legal	R	clients and refusals	5
Customer support	R	regulatory adherence	5
Customer support	T1	orders and delivery	5
Customer support	T2	returns and refunds	5
Customer support	T3	techcrew service	5
Customer support	T4	membership	5

INDUSTRY	CODE	CATEGORY	SCORED
Customer support	T5	price match	5

4.4 Difficulty

Difficulty is a property of the workflow a correct resolution requires, not of observed failure rates. Four structural measures of the grading key define it: $|T^*|$ (expected tool calls); whether any call is an identity or eligibility gate (`verify_identity`, `identify_patient`, `identify_customer`, `lookup_caller`, `check_conflict`, `check_practice_area`, `check_jurisdiction`); the number of rows of D^* that differ from D_0 ; and the number of constrained argument values in T^* .

- **Easy.** One request answered from public information or a single lookup; typically ungated and read-only. The agent must route correctly and not overreach.
- **Medium.** Caller-specific work behind a gate: the agent must authenticate or screen the caller before reading records; about half commit one change.
- **Hard.** A multi-stage workflow ending in a committed change, often a two-phase write (quote, read-back, explicit confirmation, commit) with policy branches (fees, proration, eligibility).

TIER	TASKS	$ T^* $ MEAN (RANGE)	GATED	WRITES STATE	ROWS WRITTEN	CONSTRAINED ARGS MEAN (RANGE)
Easy	18	1.2 (0–4)	17%	33%	0.4	1.4 (0–4)
Medium	36	3.6 (1–7)	78%	58%	0.8	5.6 (0–12)
Hard	36	5.9 (1–10)	81%	94%	1.8	9.3 (0–16)

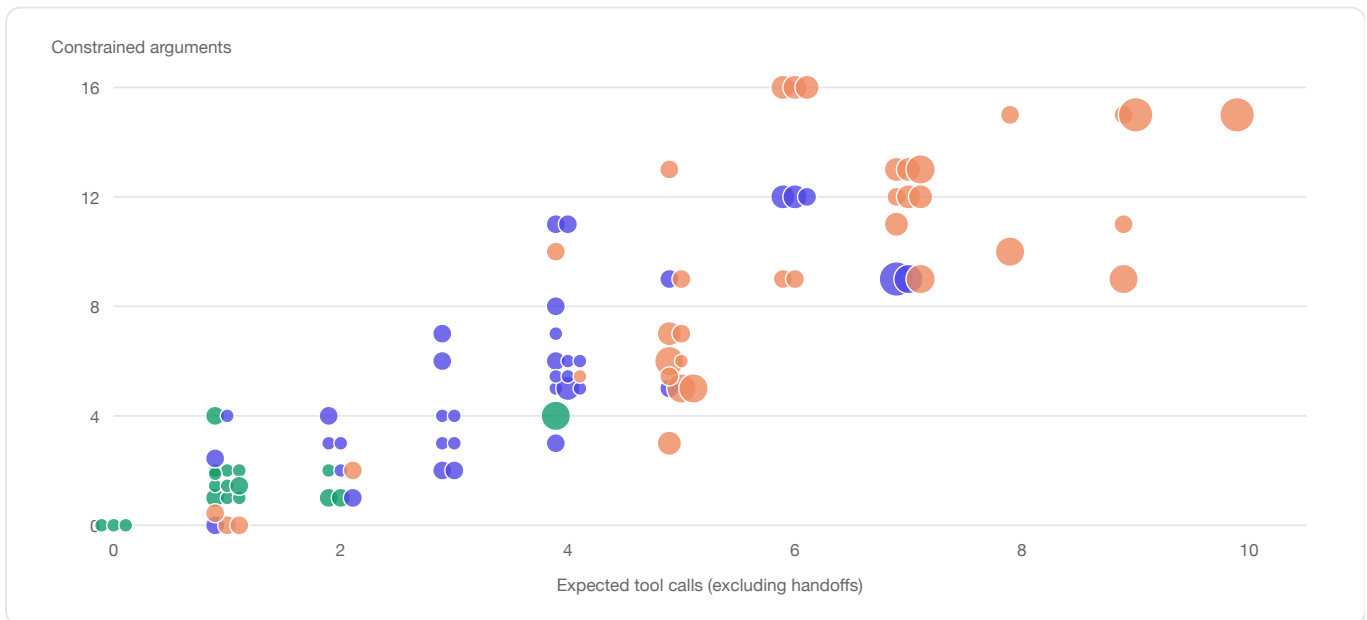


Figure 3. Structure of the 90 scored tasks. Each point is one task; $x = |T^*|$, $y =$ constrained arguments, radius $\propto$ rows written; green easy, violet medium, orange hard. Overlapping tasks are offset.

INDUSTRY	TIER	N	T* MEAN (RANGE)	WRITES STATE	ROWS WRITTEN
Healthcare	Easy	6	0.7 (0–1)	0%	0.0
Healthcare	Medium	12	2.8 (1–5)	58%	0.6
Healthcare	Hard	12	6.3 (4–9)	92%	1.2
Legal	Easy	6	1.8 (1–4)	83%	1.2
Legal	Medium	12	4.3 (1–7)	92%	1.5
Legal	Hard	12	5.3 (1–10)	100%	2.1
Customer support	Easy	6	1.0 (0–2)	17%	0.2
Customer support	Medium	12	3.8 (3–4)	25%	0.3
Customer support	Hard	12	5.9 (4–9)	92%	2.0

5 Grading specification

Grading is a pure function of the task, the tool server's request log for the call and the call's frozen database. It is implemented once (`verify_task_run.py`) and applied identically to every platform.

5.1 Pass predicate

$$\text{pass}(c) = \mathbb{1}_T(c) \cdot \mathbb{1}_D(c) \cdot \mathbb{1}_H(c) \quad (1)$$

$\mathbb{1}_H$ checks that the task's expected handoff path is an in-order subsequence of the observed handoffs; for the single-agent tasks scored here the expected path is empty and $\mathbb{1}_H = 1$. If a task has no expected state, $\mathbb{1}_D$ is not evaluated. There is no partial credit.

5.2 Tool adherence

Observed calls are taken from the request log, ordered by start time. Matching is order-free and one-to-one, by greedy first fit over expected calls in task order:

```
used = ∅
for e in T* (task order):
    a = first observed call ∉ used with name(a) = name(e) and args_match(e, a)
    if a is None: missing ← e
    else: used ← used ∪ {a}
 $\mathbb{1}_T = (\text{missing} = \emptyset) \text{ and not forbidden}(T)$ 
```

Additional calls are permitted: an agent may look things up that the key does not require. The one forbidden pattern is an unrequested `send_sms` in healthcare, which fails $\mathbb{1}_T$. A call with no expected or no observed arguments matches on name alone; the state check then carries the outcome.

5.3 Constrained and free-text arguments

Each argument key is classified from the tool's JSON schema, first rule wins: (1) has an `enum`; (2) type number, integer or boolean; (3) format `date` or `date-time`; (4) has a `pattern`; (5) key ends in `_id`, `_ids` or `_e164`; (6) key in the industry's fact list → **constrained**. (7) key in the industry's prose list, or (8) `reason` without an `enum` → **free text**. (9) Otherwise constrained.

INDUSTRY	FACT KEYS (CONSTRAINED)	PROSE KEYS (FREE TEXT)
Healthcare	full_name, first_name, last_name, dob, zip, carrier, member_id, group_number, subscriber_relationship, destination, queue, channel, template_id, next_intent, location_id, provider_id, slot_id, appointment_id, start, end, new_start, new_end, service_date, medication_name, visit_class, topic, service, order_type, appointment_type_code, cancellation_reason_code, fee_line_item_id, line_item_id	notes, patient_safe_message

INDUSTRY	FACT KEYS (CONSTRAINED)	PROSE KEYS (FREE TEXT)
Legal	opposing_party, practice_area, state, incident_date, reason_code, slot_id, confirmation_token, matter_id, channel, for_whom, provider, evaluation_id, attorney_id, phone, earliest_date, reason	summary, note, message, handoff_summary, full_name
Customer support	order_number, postal_code, card_last4, reason_code, rma_number, sku, competitor, competitor_price, in_stock, opened, service_type, fee_disclosed_acknowledged, proration_acknowledged, confirmation_token	topic, fee, query, issue, full_name, phone, reason

Every non-empty expected constrained argument must be present and equal. Free-text arguments are not compared. Keys not declared in the tool's schema, and harness-only keys (`to`, `from`), are dropped before comparison. The one documented server default (`classify_visit_request: urgency = urgent` for Mohs, otherwise routine; `is_new_patient = false`) satisfies an omitted argument.

5.4 Value equality

RULE (FIRST THAT APPLIES)	DEFINITION
Exact	Identical values match.
Alternatives	An expected value <code>{any_of: [...]}</code> matches if any option matches.
Party names	<code>opposing_party</code> , <code>for_whom</code> : casefold; "st"/"st." → "saint"; tokens [a-z0-9]+; phonetic fold <code>c→k</code> , <code>q→k</code> , <code>w→u</code> ; each token matched as a fuzzy substring with Levenshtein tolerance 0 (≤ 3 chars), 1 (≤ 6), 2 (longer). <code>for_whom</code> requires every expected token.
Locations	Office ids and spoken names alias (e.g. <code>loc_park_ave</code> $\equiv$ "park avenue").
Carriers	Lower-cased alphanumerics mapped through a slug table (e.g. bluecross, bluecrossblueshield → bcbs).
Lists	<code>location_ids</code> as sets; other lists element-wise in order.
Phone numbers	Digits only; a leading 1 on an 11-digit number is dropped.
Order numbers	Non-alphanumerics removed, upper-cased.
Booleans, numbers	Identity; numeric value.
Everything else	Typographic quotes and dashes → ASCII, trimmed, case-folded. Identifiers are therefore compared case-insensitively.

Where both the expected and the observed call carry an output status (`ok`, `error_code`), they must also agree.

5.5 State adherence

$\mathbb{D}$ compares D with D^* on the industry's write tables only; catalog tables and the request log are ignored. Rows are canonicalized before comparison: volatile keys are projected out, phone fields reduced to digits, amounts formatted to two decimals, ISO datetimes truncated to the minute, strings trimmed and case-folded.

INDUSTRY	WRITE TABLES	PROJECTED OUT	COMPARISON
Healthcare	patients, appointments, waitlist	created_at, description; waitlist.id	patients, appointments exact; waitlist: empty iff expected empty, else expected rows $\subseteq$ observed
Legal	intakes, intake_notes, documents, holds, evaluations, messages, escalations	created_at, description, message, summary, note, slot_id, starts_at, attorney_id, datetime, caller_id, reason	Only non-empty expected fields compared, with §5.4 equality. messages, intake_notes, documents, holds, evaluations: every expected row covered (extras allowed). intakes, escalations: equal counts, one-to-one
Customer support	customers, orders, order_items, order_cancellations, delivery_changes, membership_changes, holds, refunds, rmas, return_labels, price_matches, protection_plans, service_appointments, escalations, scam_reports	created_at, description, reason, issue	Exact after canonicalization; unconsumed holds ignored; scam_reports one-to-one with every supplied contact field equal to expected

Tables that act as logs allow extra rows because an agent may legitimately add an optional note or packet; tables that represent commitments (appointments, intakes, refunds, cancellations) must match exactly, so a duplicate booking or an unrequested refund fails.

5.6 Voids and breaks

OUTCOME	CONDITION	TREATMENT
Refused / cancelled	The platform rejected the call, or the run was cancelled before connection	Void; re-run
No agent	The call ended before the agent produced speech	Void; re-run
Phantom	A result with no audio and no transcript	Void; re-run
Caller left first	No tool called and at most one agent turn before the caller hung up; not applied to audio variants	Void; re-run
Break	Connected, then dropped audio, an error prompt, or a tool request that never reached the server	Fail; also reported as a break rate
Graded	All other calls	pass(c) by (1)

5.7 Worked example

Task T4-H1 (customer support, hard): the caller cancels a Kestrel Plus membership after refusing one retention offer and confirming a prorated refund. Two calls are graded below.

#	EXPECTED CALL	CONSTRAINED ARGUMENTS	CALL A	CALL B
1	identify_customer	full_name = Selina Cortez, phone = 541-555-0171	✓	✓
2	verify_identity	postal_code = 97035, card_last4 = 1156	✓	✓

#	EXPECTED CALL	CONSTRAINED ARGUMENTS	CALL A	CALL B
3	get_customer_summary	—	✓	✓
4	quote_membership_cancellation	—	✓	×
5	confirm_membership_cancellation	proration_acknowledged = True	✓	×

TABLE.FIELD	D ₀	D [*]	CALL A	CALL B
customers.tier	plus	standard	standard	plus
customers.auto_renew	1	0	0	1
customers.membership_paid_cents	2999	0	0	2999
membership_changes	—	cancel, plus → standard, 1499, 2026-08-01	present	absent

Call A: $\mathbb{1}_T = 1$, $\mathbb{1}_D = 1$, pass. Call B skipped the quote; the server's policy guard rejected the confirmation, so neither the calls nor the state match: $\mathbb{1}_T = 0$, $\mathbb{1}_D = 0$, fail.

6 Metrics and score

6.1 Platform score

$$PS = S_{\text{task}} + S_{\text{WER}} + S_{\text{lat}}, \quad S_{\text{task}} \in [0, 50], S_{\text{WER}} \in [0, 25], S_{\text{lat}} \in [0, 25] \quad (2)$$

Each component and PS are reported to one decimal. Weights and thresholds were fixed before the first scored call and are never curved to the field.

6.2 Task success

Let $N = 450$ clean scored calls after voids are re-run (§5.6), P the passes among them, and F_M the failures attributed to a model that Bluejay selected for the platform (§7, §8.2). Those calls are left out of the task score, because the fault belongs to our model choice and not to the platform:

$$S_{\text{task}} = 50 \cdot p, \quad p = P / (N - F_M) \quad (3)$$

For a platform whose every model is bundled, recommended or a flagship it ships, $F_M = 0$ and p is the raw pass rate. Published alongside: the raw pass rate P / N ; F_M by layer; the break rate; and consistency pass⁵, the share of tasks passing all five runs. Audio variants do not enter S_{task} ; their pass rate is reported separately.

6.3 Word-error composite

The reference r is the caller's text as sent to the caller's TTS. The hypothesis h is the platform's own final speech-to-text output for the caller, in order, as exported by the platform or its harness. Three rates are computed per call and pooled over calls (sum of errors over sum of reference words):

- **Tokens.** Lower-case; bracketed stage directions removed; hyphen and slash compounds split; punctuation removed; numerals 0–20 written as words.
- **Utterance alignment.** Reference utterances and hypothesis lines are aligned monotonically by dynamic programming over blocks of up to 3 reference utterances to 1 line, or 1 utterance to up to 4 lines. A block is eligible when the cosine similarity of the concatenated texts (text-embedding-3-small) is ≥ 0.35 and every added line raised it; the program maximizes words saved ($|r| + |h| - \text{edit distance}$). Unmatched reference utterances are deletions; unmatched hypothesis lines are insertions.
- **WER.** Within each aligned block, word-level Levenshtein alignment gives substitutions S , deletions D and insertions I : $\text{WER} = \min(1, (S + D + I) / N_r)$.
- **SWER.** Each run of consecutive substitutions is one phrase. Phrase cosine ≥ 0.75 is a synonym (match); 0.55–0.75 is labelled paraphrase (match) or error by an LLM judge (gpt-6-sol, with gpt-5.4 as fallback) given the reference utterance as context; < 0.55 is an error.
- **FFA.** The details are the expected tool-call argument values the caller actually says (names, dates, phone numbers, other numbers, identifiers), found in the reference after normalization (dates as calendar dates,

numbers by value, spelled-out equal to spoken). FFA = details present in the hypothesis / details in the reference.

$$C = \frac{1}{3} [(1 - \text{WER}) + (1 - \text{SWER}) + \text{FFA}], \quad S_{\text{WER}} = 25 \cdot (\frac{1}{2} C_{\text{clean}} + \frac{1}{2} C_{\text{degraded}}) \quad (4)$$

C_{clean} pools the 450 clean calls; C_{degraded} pools the 180 variant calls of both profiles. If no task-critical detail is spoken in a pool, FFA is undefined and C averages the other two terms. The composite is computed on every completed call whatever its grade, so how well a platform hears always reaches its score even when §6.2 excludes the call from task success. SWER's similarity check is the only model-assisted step in the score; it cannot change a pass.

Synthesis floor. The reference is the text the caller's TTS was asked to say, not the audio it produced; if the TTS drops or mispronounces a word, a faithful transcript is charged an error. Measured error therefore includes the synthesis error ϵ of the caller voices: $\text{WER}_{\text{measured}} \approx \text{WER}_{\text{platform}} + \epsilon$. ϵ is identical across platforms (same voices, scripts and audio) and cannot change their order. It is estimated per voice by transcribing the recorded caller channel with an independent reference ASR; because the reference ASR contributes errors of its own, the estimate is an upper bound, published with the results.

6.4 Latency

Latency is measured on Bluejay's stereo recording, decoded per channel to 16 kHz mono and divided into 50 ms windows. For each channel, a window is voiced when its RMS exceeds the floor $\phi = \max(4 \cdot \text{RMS}_{p20}, 0.01)$, where RMS_{p20} is the channel's 20th-percentile window RMS. For each caller utterance i :

1. Take the transcript end of the utterance, t_e , and the start of the caller's next utterance (or the end of the call) as the limit.
2. Skip the turn if the agent channel is voiced at t_e : the agent was already speaking (overlap or barge-in), so there is no response latency to measure.
3. Agent onset t_a : the first voiced agent window in $[t_e, \text{limit})$. Skip the turn if there is none.
4. Caller end t_c : the last voiced caller window within ± 1.5 s of t_e and before t_a (t_e if none), correcting for transcript timestamp error.
5. $\ell_i = t_a - t_c$.

Turns that include tool calls are kept, because the caller waits through them. All turns of all completed calls are pooled; m is the median and q the 95th percentile:

$$S_{\text{lat}} = 25 \cdot [0.6 \cdot \text{clip}((4 - m) / 3, 0, 1) + 0.4 \cdot \text{clip}((6 - q) / 4, 0, 1)] \quad (5)$$

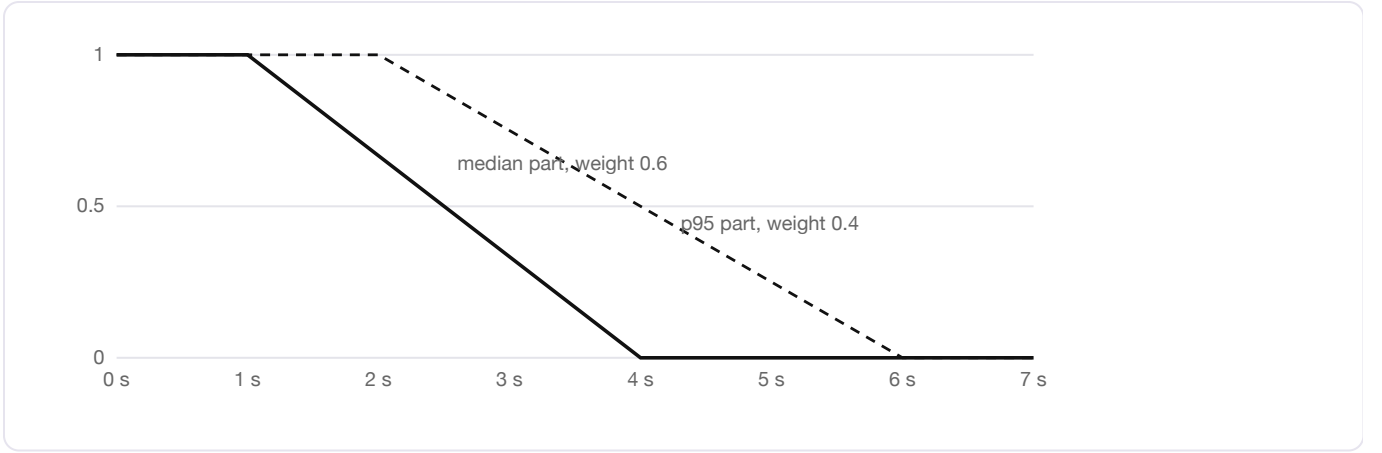


Figure 4. Latency scoring functions. Each part is full at or below its target (1.0 s median, 2.0 s p95) and zero at or above 4.0 s and 6.0 s, linear between. The tail carries 40% because a caller experiences the slowest turns of a call, not its average.

Measuring on our recording applies one method to every platform, including platforms that expose no timing data, and includes everything the caller hears: endpointing, inference, synthesis, payout and transport.

6.5 Precision and ties

p is a proportion over $n = N - F_M$ calls and is reported with its 95% Wilson interval:

$$\tilde{p} \pm z \sqrt{(\tilde{p}(1 - \tilde{p})/n + z^2/4n^2) / (1 + z^2/n)}, \quad \tilde{p} = (p + z^2/2n)/(1 + z^2/n), \quad z = 1.96 \quad (6)$$

At $n = 450$ the half-width is at most 4.6 percentage points (at $p = 0.5$), or 2.3 points of S_{task} . Platforms whose intervals overlap on a component are reported as tied on it. Calls are not independent across the five runs of a task; the interval is therefore reported as a lower bound on uncertainty, and pass^5 is published so that run-to-run variance is visible.

7 Model configuration

Results depend on the models a platform runs. Models are fixed per platform and per layer by the rule below, frozen on 27 Oct 2026 in manifest v5, and published with each platform's results.

7.1 Buckets

BUCKET	WHEN	MODEL USED	FAULTS ATTRIBUTED TO
Bundled	The layer cannot be configured	The platform's own model	Platform
Recommended	The platform documents a recommended model for voice agents	That model, cited by URL and read date	Platform
Flagship	The platform ships its own models without recommending one	The newest top-of-line model by the platform's versioning	Platform
Open choice	The platform offers several models and names none	Selected by rule (7.2)	The model (excluded from S_{task})

7.2 Selection rule

Candidates are the models the platform runs natively (built in or through its official plugin) on its cloud in its default region, that are generally available, and whose exact identifier and setting the provider accepts on a live call. LLMs run with reasoning disabled or at the lowest level offered. Each layer has two metrics a and b, each min-max scaled over the candidate set so that 0 is best and 1 worst; the pick minimizes their mean over the Pareto frontier F:

$$x' = (x - x_{\text{best}}) / (x_{\text{worst}} - x_{\text{best}}), \quad s(m) = \frac{1}{2} a'_m + \frac{1}{2} b'_m, \quad m^* = \operatorname{argmin}_{m \in F} s(m) \quad (7)$$

Ties go to the higher-quality metric, then to the newer release. Benchmarks are read once at the freeze. A later release or ranking change has no effect until a new manifest version, which every platform re-runs.

LAYER	A (QUALITY)	B (SPEED)	SOURCE	CANDIDATES
Speech-to-text	English WER	Median time to final transcript after a forced end of speech	Bluejay MISTT v0 [3], pinned 27 Oct 2026	34 streaming models with a forced finalize
LLM	Intelligence Index	Time to first answer token, $\leq \ell_{\text{max}}$	Artificial Analysis, read 26 Oct 2026	74 under ℓ_{max} ; 66 callable
Text-to-speech	Pronunciation Robustness	Voice Arena Preference Elo (second quality axis)	Artificial Analysis, read 26 Oct 2026	31 with both metrics; streaming only

Text-to-speech uses two quality metrics because Artificial Analysis publishes no time to first audio for streaming synthesis; requiring a streaming model stands in for latency, and each platform's measured time to first audio is reported.

7.3 LLM latency ceiling

An LLM's first-token time ranges from under a second to minutes across reasoning settings, so an uncapped blend selects strong models that cannot hold a call. Candidates are first restricted to $\ell \leq \ell_{\max}$, the slowest first token at which a tool-call turn (two LLM passes, a tool round trip r , and the platform's measured speech overhead o) still reaches the caller before the 4.0 s zero point of (5):

$$\ell_{\max} = (4.0 - o - r) / 2 = (4.0 - 1.33 - 0.25) / 2 = 1.21 \text{ s} \quad (8)$$

$o = 1.33 \text{ s}$ is endpointing ($\sim 0.5 \text{ s}$) + first synthesized audio (0.23 s) + playout ($\sim 0.6 \text{ s}$), measured on the 21–22 Oct 2026 LiveKit run; $r = 0.25 \text{ s}$. $\ell_{\max}$ is recomputed per platform from that platform's overhead. Without it the same blend selects Kimi K3.1 Pro (48 s) on a linear scale or Claude Fable 5.1 (Low, 2.2 s) on a rank scale.

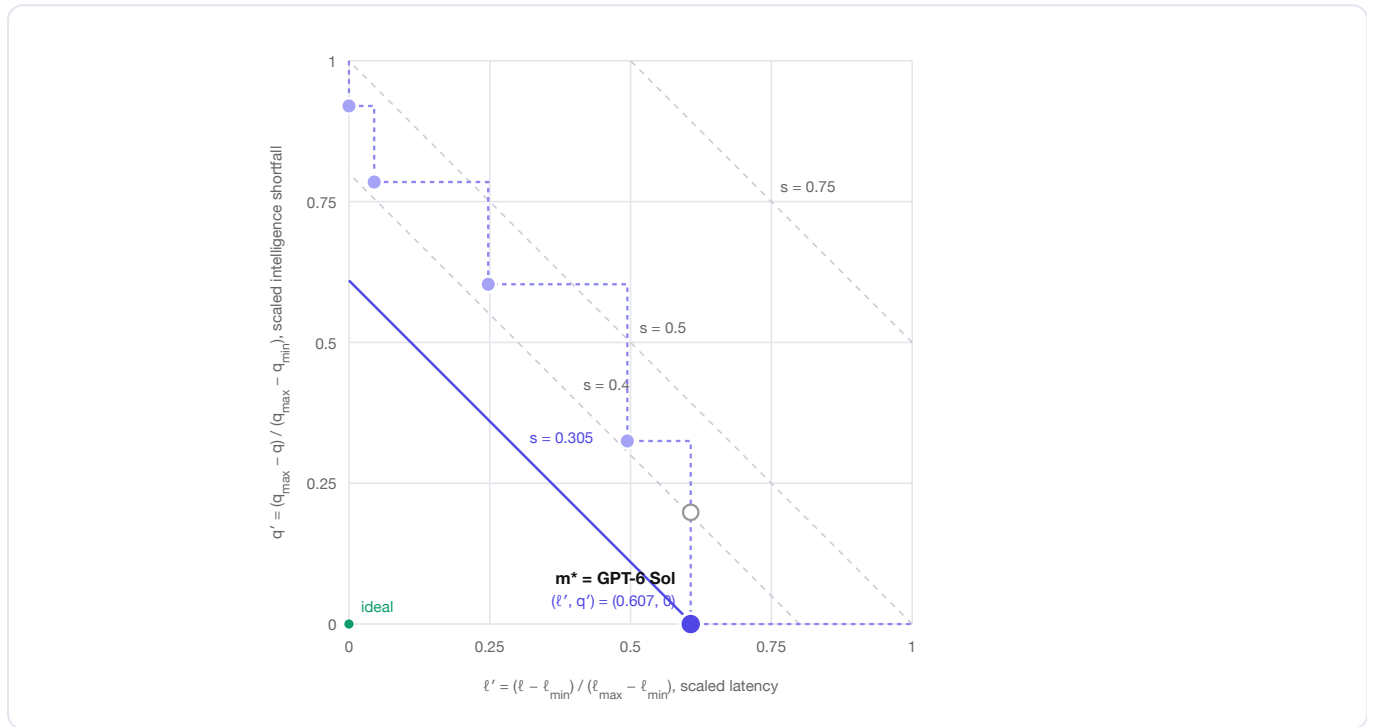


Figure 5. LLM selection in scaled metric space. The six best-scoring callable candidates under $\ell_{\max}$. Axes are the scaled metrics of (7); diagonals are level sets of s ; the dashed staircase is the Pareto frontier. The minimum is GPT-6 Sol at $(0.607, 0)$, $s = 0.305$.

7.4 Selections

LAYER	SELECTED	VALUES	S	NEAREST ALTERNATIVE
Speech-to-text	Cartesia Ink-2	WER 4.02% (95% CI 3.69–4.37); 72.5 ms	0.0569	Ink Preview, $s = 0.0567$: a pre-release of the same model, excluded by general availability
LLM	OpenAI GPT-6 Sol, reasoning off	Index 28.5; 0.85 s	0.305	Gemini 3.5 Flash (Minimal) 0.400. Claude Sonnet 5.5 (Low) (35.9, 1.15 s) would win but was not callable on our account until 1 Nov 2026
Text-to-speech	ElevenLabs Eleven v4 Turbo	Pronunciation 90.1%; Elo 1333	0.975*	Eleven v4 (0.988*) is served only through a non-real-time API

* Text-to-speech scores are reported as $1 - s$ (higher is better). GPT-6.1 Sol, GPT-6 Sol's successor, rejects the non-reasoning setting and its fastest setting takes 2.67 s; GPT-6.1 Terra is over $\ell_{\max}$ in every setting (1.27–1.43 s). Restricted to current models only, the rule selects GPT-6 Luna (18.5, 0.91 s, $s = 0.337$). Because Artificial Analysis stopped re-measuring GPT-6 Sol when it was superseded, its latency was verified on our traffic: median 0.99 s, p75 1.22 s, p90 1.54 s over 7,984 turns with the full system prompt.

LAYER	PINNED IDENTIFIER AND SETTINGS
Speech-to-text	Cartesia Ink-2, streaming, platform default endpointing
LLM	<code>gpt-6-sol</code> , OpenAI Responses API, <code>reasoning.effort = none</code> , temperature 0.8
Text-to-speech	<code>eleven_v4_turbo</code> , voice Rachel

8 Attribution and fairness

A comparison is fair when every difference in score is caused by the platforms. Let $\kappa(c)$ be the cause of a failed call c . The evaluation is fair when (i) every failure entering S_{task} has $\kappa(c)$ in the platform's own stack and (ii) every failure caused by Bluejay is removed from every platform and re-run, so that all platforms are scored on the same tasks under the same conditions.

8.1 Conditions held equal

CONDITION	CONTROL
Inputs	Identical agent, tasks, callers, voices, seed state and tool server for every platform, frozen in the manifest.
Grader	One deterministic verifier version for all platforms; no language model decides a pass.
Caller	Temperature 0; every graded fact and decision fixed by scripted responses; five runs per task; every call recorded.
Models	Chosen by the published rule of §7 from models each platform can run; faults of Bluejay-chosen models excluded from S_{task} .
Settings	No platform-specific prompts; turn-taking, interruption and timeout settings at platform defaults.
Conditions	One caller region; fixed concurrency at each platform's entry-level limit, published; latency measured with one method on our recording.

8.2 Attribution

Every failed call receives exactly one cause, checked in this order; the first that applies is recorded. The order matters: a wrong decision made on a misheard transcript is a speech-to-text fault, not an LLM fault.

#	CAUSE	EVIDENCE	PROVIDER-CHOSEN MODEL	BLUEJAY-CHOSEN MODEL
1	Task	The specification or expected state is wrong or ambiguous; confirmed against the answer-key replay	Void; fix; re-run on every platform	Void; fix; re-run on every platform
2	Caller	The caller departed from its script or facts	Void; re-run	Void; re-run
3	Harness	Our tool server, transport bridge or recording logged a fault	Void; re-run	Void; re-run
4	Platform	Break, or orchestration fault: caller cut off, barge-in mishandled, tool request malformed, timed out or never sent, context dropped	Fail	Fail
5	Speech-to-text	The platform's transcript misses a task-critical detail (§6.3 FFA) and the failing call carries the misheard value	Fail	Excluded from S_{task} ; still in S_{WER}
6	LLM	Transcript and context correct; wrong tool or argument, skipped step, broken gate or invented value	Fail	Excluded from S_{task}
7	Text-to-speech	Agent text correct, but a reference ASR of its audio does not carry it	Fail	Excluded from S_{task}

Evidence for each attribution: the stereo recording, the platform's transcript, the tool server's request log, the database diff against Do and D*, and the platform's traces where exposed (tracing is enabled wherever offered).

8.3 Voids

Void rules (§5.6) are applied before grading. A voided call is neither pass nor fail and is re-run until every task has exactly five graded runs on every platform, so every platform is scored on the same denominator. A call that connected and then broke is never voided.

8.4 Finding defects on our side

A deterministic grader is only as fair as the tasks and callers it grades. Four procedures run on every platform:

1. **Live tracking.** Each result is compared with its task as it arrives. A failure pattern that repeats across runs or platforms is raised during the run with result identifiers.
2. **Stratified audit.** After each run a seeded 25% sample of calls is drawn, stratified by industry × task family with largest-remainder allocation, so every kind of call is represented. A reviewer attributes every failed call in the sample from the evidence above and checks a seeded sample of passing calls for false passes.
3. **Cross-platform comparison.** Each task's pass rate is compared across platforms. A task that fails the same way on platforms with different stacks is reviewed as a suspected task or caller defect; a task that fails on one platform and passes on others is evidence that it is passable.
4. **Exhaustive review.** Every task that fails all five runs on any platform is reviewed in full.

A confirmed defect is fixed once and for every platform. A grading-key fix (for example, an expected call the agent's prompt never asks for, or a key that rejects an argument form the tool schema allows) is applied by re-grading every platform's stored calls. A caller-script fix (for example, an opening line that omitted the date the task depends on, a caller with no line for accepting an offered transfer, or a trigger that fired before its precondition) requires new calls on every platform. A grader change is accepted only if it reproduces every previously graded outcome except those it is meant to change.

Grader leniency (case, format, aliases, phonetic name matching; §5.4) can let a near-miss pass. It applies identically to every platform and so cannot change their order; its effect is measured by the false-pass checks above.

8.5 Right of reply

Before publication each platform receives its results, the recordings of its failed calls and their attributions, and may flag factual errors. A flagged call is re-reviewed; a confirmed Bluejay defect is fixed and re-run on every platform.

9 Platform profile

The platform score measures how well a platform runs a fixed agent. It does not measure what it takes to build that agent, ship it and operate it. Each platform is therefore profiled in eight categories, reported beside the score and never added to it.

Each category is a set of features with weights summing to 10. Each feature is graded full (1), half (0.5) or none (0); the category score is the weighted sum. Every grade cites its evidence: a vendor document (URL and read date), an API response, or a Bluejay recording. Five features are measured rather than read: time to first agent (stopwatch, screen-recorded build from public docs), interruption handling (probe calls on defaults), runtime under load (break rate at 20 concurrent calls against a 2-call baseline), cost per hour (estimate from public pricing against the bill for the scored calls) and language support (one translated probe task per language, two runs). A feature graded from documents before its measurement exists is marked * until measured.

UI — How well someone can build, test and review an agent in the console without writing code.

#	FEATURE	WEIGHT	FULL MARKS WHEN
01	Time to first agent	2.5	Stopwatch. A builder new to the platform makes the benchmark agent (prompt, greeting, three tools) in the console from public docs only, screen recorded. Full marks at 30 minutes or less, none at 4 hours or more.
02	Build without code	2	Create the agent, write the prompt, add an HTTP tool with typed parameters and pick the voice, all in the console.
03	Test in the browser	1.5	Talk to the agent from the console and watch the transcript and tool calls as they happen.
04	Call review	1.5	Every past call in the app, with transcript, recording, and each tool call's arguments and result.
05	Errors you can see	1.5	A failed tool call or config error shows up in the console where you are looking, not as silence.
06	Team workspace	1	Invite teammates with roles and build together.

DevX — Whether a developer can build, run and read back the whole agent from code.

#	FEATURE	WEIGHT	FULL MARKS WHEN
01	Agent-as-code	2.5	Read, write and version the whole agent (prompt, flow, tools, voice, settings) through the API; a read-then-write round trip changes nothing.
02	Call control	1.5	Start outbound, inbound or web calls from code, attach your own metadata, and get a test or sandbox mode.
03	Data out	2.5	Transcripts with per-turn timestamps, reachable recordings, tool calls with arguments and results, latency breakdowns, webhooks, and list endpoints that filter and page.
04	API correctness	1.5	Honest status codes, typed errors, idempotency, stable schemas, no silent drops.
05	SDK quality	1	Official typed Python and TypeScript SDKs, async and streaming, the same coverage as the REST API, semver releases.
06	Scoped keys and auth	1	Keys scoped by permission and environment, rotation without downtime, the same auth across API and SDKs.

Platform customizability — How far you can shape the agent: its models, conversation behaviour, prompt and tools.

#	FEATURE	WEIGHT	FULL MARKS WHEN
01	Model choice	1.5	Pick the LLM, speech-to-text and voice. Bring your own model keys or your own LLM endpoint.
02	Conversation behaviour	1	Settings for end-of-turn detection, interruption sensitivity, and silence and idle timeouts.
03	Interruption handling	3	Measured on the platform's defaults. Full: stops within 0.5 s median of the caller speaking, responds to the interruption at least 90% of the time, stops for a backchannel no more than 10%.
04	Prompt and context	1	The full benchmark prompt fits without trimming. Per-call variables, and structured flows if you want them.
05	Custom tools	2.5	HTTP tools with typed parameters, client-side tools, non-blocking tools, timeouts and retries, and call metadata passed to the tool.
06	Speech controls	0.5	A pronunciation dictionary or SSML, and speaking rate.
07	Agent handoff	0.5	Hand the conversation to another agent on the same platform, with its context.

Go live — Everything needed to take the agent to production and keep it running.

#	FEATURE	WEIGHT	FULL MARKS WHEN
01	Telephony and numbers	2	Buy or port numbers by API, bring your own carrier, inbound and outbound on one number, caller-ID registration, DTMF, warm and cold transfer.
02	Release management	1.5	Draft and published versions, separate environments, promote without re-authoring, one-click rollback, pin a number to a version.
03	Runtime under load	2.5	Measured. Full: the break rate at 20 concurrent stays within 2 points of the 2-call baseline and limits can be raised self-serve. Half: within 10 points, or raising needs sales.
04	Integrations at the edge	1	Contact-centre platforms, SIP REFER and transfer with context, CRM webhooks.
05	Ops and incident response	1.5	A status page, live call monitoring, ending a call remotely, alerts on error spikes, an audit trail.
06	Compliance and trust	1	SOC 2 or HIPAA with a BAA, recording consent, PII redaction, data residency, retention, calling-hours controls.
07	Support and escalation	0.5	A support SLA and a named escalation contact.

MCP — Whether the agent can use tools over MCP, and whether coding agents can build and run the platform through MCP.

#	FEATURE	WEIGHT	FULL MARKS WHEN
01	MCP tools on a live call	3.5	The agent connects to a remote MCP server and calls its tools during a voice call, tested with our tool server exposed over MCP.
02	MCP connection controls	1.5	Auth headers or OAuth, choosing which tools the agent can see, and passing per-call values.
03	Platform MCP server	3	An official MCP server that lets a coding agent build and manage agents on the platform.
04	What that server can do	2	Share of the platform's API the MCP server covers. Full at 80% or more, half at 40% or more.

Docs — Whether the docs are accurate, complete and usable by people and by coding agents.

#	FEATURE	WEIGHT	FULL MARKS WHEN
01	Reference accuracy	2	Every endpoint, field, type, default and enum matches what the API does. An OpenAPI spec generated from code.
02	Payload and event schemas	1.5	Webhook bodies, call objects, transcripts, tool-call payloads and event streams, each with a full real example.
03	Concepts and lifecycle	1	A call-state diagram, the agent, version and deployment model, and how flows and routing execute.
04	Telephony and go-live guides	1.5	Step-by-step guides for numbers, carriers, SIP, transfers, DTMF, caller ID and rollback, each ending in a working call.
05	Errors, limits and gotchas	1	An error catalogue with causes and fixes, rate and concurrency limits, known pitfalls stated openly.
06	Examples and quickstarts	1	Copy-paste curl, Python and TypeScript examples that run unmodified, and a first call in under 15 minutes.
07	Change management	1	A dated changelog, deprecation notices with timelines, API versioning and migration guides.
08	Machine-readability	1	llms.txt, a docs MCP server or search API, clean markdown export.

Pricing clarity — Whether a buyer can predict what the platform will cost and see what it is costing.

#	FEATURE	WEIGHT	FULL MARKS WHEN
01	Public price	2	A per-minute or per-unit price on a public page, no sales call needed.
02	Every line item named	1.5	Platform fee, LLM, speech-to-text, voice and telephony each priced or stated as included.
03	Cost per hour matches the estimate	3	Measured. Our estimate from public pricing against the real bill for the scored calls. Full within 10%, half within 25%. The \$/hour is published but not scored.
04	Usage and spend in the product	1.5	Live usage and spend, and the cost of each call.
05	Spend controls	1.5	Spend caps, alerts, and a clear warning before service stops.
06	Free start	0.5	Free credits, or a trial big enough to build and test the agent.

Language support — Whether the agent can take a call in Spanish, French, Mandarin Chinese, Hindi or Arabic, understand it and reply in that language.

#	FEATURE	WEIGHT	FULL MARKS WHEN
01	Spanish	2	Full: understood and replied in Spanish. Half: understood but replied in English. None: did not understand.
02	French	2	Full: understood and replied in French. Half: understood but replied in English. None: did not understand.
03	Mandarin Chinese	2	Full: understood and replied in Mandarin Chinese. Half: understood but replied in English. None: did not understand.
04	Hindi	2	Full: understood and replied in Hindi. Half: understood but replied in English. None: did not understand.
05	Arabic	2	Full: understood and replied in Arabic. Half: understood but replied in English. None: did not understand.

10 Records and reproducibility

For every call Bluejay retains the records below. Anyone holding the manifest and the configuration files can re-deploy the agent on a platform and re-run any task under the same conditions.

RECORD	CONTENTS
Manifest version	The frozen run definition the call belongs to
Platform and stack	Platform, hosting region, transport, and each layer's model with bucket and source
Task and run	Task id, run number, caller configuration, voice slot, audio condition
Audio	Stereo recording, caller left, agent right
Platform record	The platform's caller transcript, the agent's text, traces where exposed
Reference ASR	Our transcripts of the caller channel (synthesis floor) and the agent channel (TTS attribution)
Tool server	Every request with arguments, raw body, response, status and timing; the frozen database
Turn timings	t_c , t_e , t_a and ℓ for every measured turn
Grade	$\mathbb{1}_T$, $\mathbb{1}_D$, $\mathbb{1}_H$, pass, void or break; the cause, the reviewer, and any re-run

11 Threats to validity

- **Language and domain.** English only; three industries. Results may not transfer to other languages, caller populations or workloads.
- **Simulated callers.** Speech is synthetic (synthesis floor ϵ , §6.3); callers do not barge in, and the scripted-response matcher sees only the agent's latest turn, so history-dependent behaviour is approximated by exclusions.
- **Degraded line.** The degraded-line profile is severe: parts of the caller's speech are hard to recover even for a reference transcriber. Variants enter only the degraded word-error term and are reported separately from task success.
- **Transport.** One caller region over each platform's standard ingress; carrier and PSTN conditions beyond the degraded-line profile are not covered. Where a platform's ingress is a phone number, latency includes the telephone network, and this is disclosed beside the score.
- **Grader.** Deterministic but lenient on names and formats (§5.4); free-text arguments and spoken information on read-only tasks are not graded.
- **Attribution.** Causes 4–7 require judgment from evidence; they are audited (§8.4) and open to the platform's reply (§8.5).
- **Time.** Results describe each platform and pinned stack on the run date; each round is versioned.
- **Profile.** Some profile features involve judgment until measured, which is why the profile is kept out of the score.

A The 90 scored tasks

INDUSTRY	TASK	TITLE	TIER	T*	ARGS	ROWS	GATES
Healthcare	C1-E1	Aetna coverage at Park Avenue	easy	1	2	0	0
Healthcare	C1-H2	Hours, Medicare, then an urgent rash booking	hard	5	13	1	0
Healthcare	C1-H3	Oscar Health coverage then a new-patient booking	hard	6	16	2	0
Healthcare	C1-M2	Cigna coverage then a new-patient booking	medium	4	11	1	0
Healthcare	C1-M3	Allergy skin testing at Park Avenue	medium	1	2	1	0
Healthcare	C2-E1	Medical cancellation policy	easy	0	0	0	0
Healthcare	C2-H2	Allergy shot for a child	hard	6	9	1	2
Healthcare	C2-H3	Cancel the August 21 cosmetic consult	hard	7	11	2	2
Healthcare	C2-M2	Cancel a September follow-up	medium	2	4	1	1
Healthcare	C2-M4	Move the September Windermere follow-up	medium	3	6	1	1
Healthcare	C3-E1	UnitedHealthcare at Brooklyn Heights	easy	1	2	0	0
Healthcare	C3-H1	Aetna Open Access Elect Choice at Park Avenue	hard	6	16	2	0
Healthcare	C3-H2	Out-of-pocket cost without a member ID	hard	6	16	2	0
Healthcare	C3-M1	Copay for a Park Avenue visit	medium	1	4	0	0
Healthcare	C3-M3	Medicare at Windermere then a booking	medium	4	11	1	0
Healthcare	C4-E1	Botox cost inquiry	easy	1	1	0	0
Healthcare	C4-H2	Botox on an existing cosmetic consult	hard	5	6	0	2
Healthcare	C4-H4	Existing patient Botox after Windermere cannot do cosmetic	hard	9	15	1	2
Healthcare	C4-M1	Microneedling consult	medium	5	9	1	0
Healthcare	C4-M2	Cheek filler consult	medium	4	8	1	0
Healthcare	C5-E1	Paying by card over the phone	easy	0	0	0	0
Healthcare	C5-H1	Large balance, cannot pay in full	hard	9	11	1	2
Healthcare	C5-H2	Balance question then move August 20 visit	hard	5	7	1	1
Healthcare	C5-M1	Unrecognized bill	medium	3	3	0	1
Healthcare	C5-M4	Explain the visit charge, do not pay	medium	3	3	0	1
Healthcare	R-E1	Ask for a person	easy	1	2	0	0
Healthcare	R-H1	What the biopsy results say	hard	8	15	1	2
Healthcare	R-H2	Accutane refill	hard	4	10	1	1

INDUSTRY	TASK	TITLE	TIER	T*	ARGS	ROWS	GATES
Healthcare	R-M2	Pay the balance by card now	medium	2	3	0	1
Healthcare	R-M4	Triamcinolone refill	medium	2	3	0	1
Legal	C1-E1	Message for Daniel Okonkwo	easy	1	1	1	0
Legal	C1-H3	Medical emergency on the line	hard	7	12	2	4
Legal	C1-H4	Injured relative cannot speak	hard	1	0	1	0
Legal	C1-M1	Returning caller message with lookup	medium	3	2	1	1
Legal	C1-M2	Identification fails twice	medium	1	0	1	0
Legal	C2-E1	Conflict hit, Vertex Logistics	easy	2	1	1	1
Legal	C2-H3	Insurance adjuster calls	hard	1	0	1	0
Legal	C2-H4	Other driver reveals after looking like a new client	hard	2	2	1	1
Legal	C2-M1	Conflict hit, Northgate Insurance	medium	3	2	1	2
Legal	C2-M4	Consulted a lawyer but did not hire	medium	2	2	0	2
Legal	C3-E1	Bankruptcy matter declined	easy	2	1	1	1
Legal	C3-H1	Caller starts the story before conflict check	hard	7	12	2	4
Legal	C3-H4	Employment in Washington clears screening	hard	7	13	2	4
Legal	C3-M3	Auto accident in Arizona	medium	5	5	1	4
Legal	C3-M4	Immigration matter declined	medium	4	3	1	3
Legal	C4-E1	Intake recorded for auto accident	easy	1	4	1	0
Legal	C4-H2	Consumer intake, declines booking	hard	7	13	2	4
Legal	C4-H4	Medical malpractice Florida intake with records	hard	7	13	3	3
Legal	C4-M2	Witness detail captured as note	medium	6	12	2	3
Legal	C4-M3	Packet by text message	medium	6	12	2	3
Legal	C5-E1	Books first offered evaluation slot	easy	4	4	3	1
Legal	C5-H1	Medical malpractice booked in Florida	hard	9	15	4	3
Legal	C5-H4	Premises evaluation booked in Texas	hard	10	15	4	4
Legal	C5-M1	Contingency fee hold and confirm	medium	7	9	4	2
Legal	C5-M3	Hourly consumer consultation	medium	7	9	3	4
Legal	R-E1	Existing client asks for status	easy	1	1	0	0
Legal	R-H1	Impersonating firm staff	hard	1	0	1	0
Legal	R-H4	Client status then insists on settlement advice	hard	5	3	2	1

INDUSTRY	TASK	TITLE	TIER	T*	ARGS	ROWS	GATES
Legal	R-M2	Is eighty thousand in the right area	medium	6	12	1	4
Legal	R-M3	Claims a matter the firm has no record of	medium	2	1	1	1
Customer support	R-E1	Demands a person immediately	easy	1	1	1	0
Customer support	R-H3	Return label on a swollen power bank	hard	9	9	3	2
Customer support	R-H4	Scam report, then the Bellwether return window	hard	7	12	1	2
Customer support	R-M3	Caller reads out a full card number	medium	4	5	0	2
Customer support	R-M4	Fake 399 charge, no money sent	medium	3	7	1	0
Customer support	T1-E1	Haul-away with a delivery	easy	1	1	0	0
Customer support	T1-H2	Late range delivery change	hard	5	6	3	2
Customer support	T1-H4	Cancel a shipped laptop, then the return window	hard	4	5	0	2
Customer support	T1-M2	Cancel an unshipped speaker	medium	4	5	2	2
Customer support	T1-M4	Quote a late range delivery change	medium	4	6	1	2
Customer support	T2-E1	General return window	easy	1	1	0	0
Customer support	T2-H1	Return a phone with the restocking fee	hard	7	9	3	2
Customer support	T2-H4	Return the phone, label, then refund status	hard	8	10	3	2
Customer support	T2-M1	Activatable window as a Total member	medium	4	5	0	2
Customer support	T2-M2	Laptop past the 15-day window	medium	4	5	0	2
Customer support	T3-E1	Third-party repair and the warranty	easy	0	0	0	0
Customer support	T3-H1	Book a Total-covered bench repair	hard	5	9	1	2
Customer support	T3-H3	Book a bench visit then cancel it	hard	6	9	1	2

INDUSTRY	TASK	TITLE	TIER	T*	ARGS	ROWS	GATES
Customer support	T3-M1	Total coverage on a laptop that will not charge	medium	4	6	0	2
Customer support	T3-M2	Dropped-phone protection-plan deductible	medium	4	6	0	2
Customer support	T4-E1	Plus and Total membership pricing	easy	2	2	0	0
Customer support	T4-H1	Cancel Plus after one save offer	hard	5	5	3	2
Customer support	T4-H3	Cancel Total after the proration readback	hard	5	5	3	2
Customer support	T4-M2	Already on Total	medium	3	4	0	2
Customer support	T4-M3	No membership on the account	medium	3	4	0	2
Customer support	T5-E1	Price Match Guarantee policy	easy	1	1	0	0
Customer support	T5-H1	Halo soundbar price match at Rivertide	hard	5	7	2	2
Customer support	T5-H2	Marketplace headphone price match	hard	5	5	1	2
Customer support	T5-M1	Price match on an open-box TV	medium	4	5	0	2
Customer support	T5-M4	Price not lower on the Halo	medium	4	7	0	2

B Tool inventory

Healthcare · Straus Dermatology

AGENT	TOOL	PURPOSE
reception	create_callback_task	When to call this tool: you need to promise a callback with a real tracked SLA — after a tool cannot finish the work, or they need a team to call back
	list_locations	When to call this tool: the caller asks for an office address, floor, suite, hours, parking, or transit; names an office; or you need an office id before booking or a coverage check
identity	create_callback_task	When to call this tool: you need to promise a callback with a real tracked SLA — after a tool cannot finish the work, or they need a team to call back
	identify_patient	When to call this tool: the caller may already be a patient and you need to find their chart before verify_identity
	verify_identity	When to call this tool: you have the caller's name and date of birth and must confirm they are the patient before any protected chart work
	get_patient_summary	When to call this tool: identity verification just succeeded — load the chart before handing off or starting specialist work
scheduling	create_callback_task	When to call this tool: you need to promise a callback with a real tracked SLA — after a tool cannot finish the work, or they need a team to call back
	send_sms	When to call this tool: they asked for an appointment confirmation text after a medical booking, reschedule, or allergy service was just committed
	classify_visit_request	When to call this tool: the caller described a medical, surgical, or allergy reason for a visit and you need the appointment type before searching slots
	list_locations	When to call this tool: the caller asks for an office address, floor, suite, hours, parking, or transit; names an office; or you need an office id before booking or a coverage check
	find_slots	When to call this tool: you are ready to offer appointment times at specific offices
	book_appointment	When to call this tool: the caller said yes after you read back the day, time, office with floor, and provider with credentials for a medical, Mohs, or allergy evaluation visit
	reschedule_appointment	When to call this tool: the caller wants to move an existing appointment — offer this before cancelling
	cancel_appointment	When to call this tool: they still want to cancel after you offered to move the visit instead
	join_waitlist	When to call this tool: they want an earlier opening than the slots you can offer, or they decline to rebook after a cancel

AGENT	TOOL	PURPOSE
coverage	schedule_allergy_service	When to call this tool: book a dedicated allergy service (skin testing, patch testing, shots, and the like) after an explicit yes
	create_callback_task	When to call this tool: you need to promise a callback with a real tracked SLA — after a tool cannot finish the work, or they need a team to call back
	list_locations	When to call this tool: the caller asks for an office address, floor, suite, hours, parking, or transit; names an office; or you need an office id before booking or a coverage check
	check_plan_accepted	When to call this tool: the caller asks whether you take a plan at a specific office
	run_eligibility_check	When to call this tool: they want a live copay, deductible, or coinsurance and have given a member ID
cosmetic	capture_insurance_update	When to call this tool: they have new or changed insurance to put on file
	create_callback_task	When to call this tool: you need to promise a callback with a real tracked SLA — after a tool cannot finish the work, or they need a team to call back
	quote_cosmetic_service	When to call this tool: they ask what a cosmetic treatment costs
	list_locations	When to call this tool: the caller asks for an office address, floor, suite, hours, parking, or transit; names an office; or you need an office id before booking or a coverage check
	find_slots	When to call this tool: you are ready to offer appointment times at specific offices
	book_cosmetic_consult	When to call this tool: they are ready to book a cosmetic consult
	send_payment_link	When to call this tool: they are ready to pay a medical balance or a cosmetic deposit
billing	offer_financing	When to call this tool: they cannot pay the full amount at once and the balance or quoted amount is over two hundred fifty dollars
	create_callback_task	When to call this tool: you need to promise a callback with a real tracked SLA — after a tool cannot finish the work, or they need a team to call back
	get_account_balance	When to call this tool: they ask about a bill, balance, or charge — call this first, before explaining or collecting
	explain_charge	When to call this tool: they ask why a specific charge is on the account
	send_payment_link	When to call this tool: they are ready to pay a medical balance or a cosmetic deposit
	offer_financing	When to call this tool: they cannot pay the full amount at once and the balance or quoted amount is over two hundred fifty dollars
	request_fee_waiver	When to call this tool: they dispute a missed-visit fee and want it reviewed

AGENT	TOOL	PURPOSE
clinical	create_callback_task	When to call this tool: you need to promise a callback with a real tracked SLA — after a tool cannot finish the work, or they need a team to call back
	get_results_status	When to call this tool: they ask about pathology, lab, or allergy testing results
	request_rx_refill	When to call this tool: immediately after identity verification for a named-medication refill that maps to a closed-formulary slug, or refill-related prior authorization, before policy talk, extra details, a clinical message, callback, or handoff
	create_clinical_message	When to call this tool: a nurse, results, or refill question needs the clinical team — including when they push for result content you cannot read
	send_portal_activation	When to call this tool: the portal is not active and they need access — on every results call where it is inactive, before the call ends

Legal · Halverson & Reed

AGENT	TOOL	PURPOSE
reception	lookup_caller	Find or create the caller record
	get_caller_matters	The caller's existing matters, including whether any is already represented by another firm
	take_message	Take a message for a person at the firm, with a callback promise
screening	check_conflict	Screen for a conflict of interest
	check_practice_area	Whether the firm handles this matter type and how it charges
	check_jurisdiction	Whether the firm is licensed for this matter in this state
	calculate_filing_deadline	The statute of limitations
intake	record_intake	Log the intake
	add_intake_note	Attach a note to the caller's record for the attorney or case manager
	send_intake_packet	Send the new client packet
	request_records_authorization	Send a medical records release for signature
scheduling	get_attorney	An attorney's name, practice areas, and bar states (internal routing lookup)
	find_evaluation_slots	Open case evaluation slots
	hold_evaluation	Hold and price a case evaluation
	confirm_evaluation	Finalize the booking
	hold_cancellation	Quote cancelling an appointment

AGENT	TOOL	PURPOSE
	confirm_cancellation	Finalize the cancellation
client_services	get_caller_matters	The caller's existing matters, including whether any is already represented by another firm
	get_case_status	Status of one of the caller's own matters at this firm
	take_message	Take a message for a person at the firm, with a callback promise
	add_intake_note	Attach a note to the caller's record for the attorney or case manager

Customer support · Kestrel Electronics

AGENT	TOOL	PURPOSE
reception	search_kb	Look up general Kestrel information: store hours, the legacy brands (Sound Harbor, Bellwether Mobile, Aurelian Audio, Coastline Kitchen & Home, Sagebrush Outdoor), what TechCrew is, how returns work, recycling and trade-in, the difference between a manufacturer warranty and a protection plan, and what a Kestrel scam looks like
	get_store_info	Address, hours and departments for one store
	get_policy	The published policy text, with its numbers: returns, restocking, price_match, membership, delivery_install, open_box, marketplace, recycling
	get_fee	The published fee schedule, including membership pricing
verification	search_kb	Look up general Kestrel information: store hours, the legacy brands (Sound Harbor, Bellwether Mobile, Aurelian Audio, Coastline Kitchen & Home, Sagebrush Outdoor), what TechCrew is, how returns work, recycling and trade-in, the difference between a manufacturer warranty and a protection plan, and what a Kestrel scam looks like
	identify_customer	Find the caller's record
	verify_identity	Verify the caller with the ZIP code on the order and the last four digits of the card it was paid with
	get_customer_summary	The verified caller's name, membership tier, recent orders and any open TechCrew appointments
orders	search_kb	Look up general Kestrel information: store hours, the legacy brands (Sound Harbor, Bellwether Mobile, Aurelian Audio, Coastline Kitchen & Home, Sagebrush Outdoor), what TechCrew is, how returns work, recycling and trade-in, the difference between a manufacturer warranty and a protection plan, and what a Kestrel scam looks like
	get_customer_summary	The verified caller's name, membership tier, recent orders and any open TechCrew appointments
	get_order	One order in full: status, what is on it, whether Kestrel or a Marketplace seller sold it, the delivery date and window, whether installation and haul-away are included, the state it was bought in, and safety flags on any item (recalled, battery_safety_flag)

AGENT	TOOL	PURPOSE
	quote_delivery_change	Price a change to a scheduled delivery or installation
	confirm_delivery_change	Commit the delivery change using the token quote_delivery_change returned
	cancel_order	Cancel an order that has not shipped
	quote_price_match	Test a competitor's price against the Price Match Guarantee and work out the difference
	confirm_price_match	Refund the price difference using the token quote_price_match returned
	get_fee	The published fee schedule, including membership pricing
returns	search_kb	Look up general Kestrel information: store hours, the legacy brands (Sound Harbor, Bellwether Mobile, Aurelian Audio, Coastline Kitchen & Home, Sagebrush Outdoor), what TechCrew is, how returns work, recycling and trade-in, the difference between a manufacturer warranty and a protection plan, and what a Kestrel scam looks like
	get_order	One order in full: status, what is on it, whether Kestrel or a Marketplace seller sold it, the delivery date and window, whether installation and haul-away are included, the state it was bought in, and safety flags on any item (recalled, battery_safety_flag)
	check_return_eligibility	The whole return computation for one item: which window applies and why (15 days standard, 60 for Plus and Total members, 14 for activatable devices no matter the tier), how many days are left or how many days over, and the restocking fee given the product class, whether the box was opened, and the state it was bought in
	quote_return	Price the return: refund amount, restocking fee and why it applies, and where the money goes
	confirm_return	Start the return and get an RMA number, using the token quote_return returned
	create_return_label	Email a free prepaid return label for an open RMA
	get_refund_status	Where a refund is: the stage it is at and the date range it posts by
	get_fee	The published fee schedule, including membership pricing
service	search_kb	Look up general Kestrel information: store hours, the legacy brands (Sound Harbor, Bellwether Mobile, Aurelian Audio, Coastline Kitchen & Home, Sagebrush Outdoor), what TechCrew is, how returns work, recycling and trade-in, the difference between a manufacturer warranty and a protection plan, and what a Kestrel scam looks like
	get_order	One order in full: status, what is on it, whether Kestrel or a Marketplace seller sold it, the delivery date and window, whether installation and haul-away are included, the state it was bought in, and safety flags on any item (recalled, battery_safety_flag)
	get_protection_plans	TechCrew Protect plans on this caller's products, with start and end dates and the deductible, plus what their membership tier covers on its own

AGENT	TOOL	PURPOSE
	check_coverage	Whether a specific failure on a specific item is covered, and who pays: a TechCrew Protect plan (deductible), Kestrel Total (nothing), the manufacturer warranty (nothing, defects only), or nobody (a \$39.99 bench diagnostic)
	book_service_appointment	Book TechCrew: bench (in a store), in_home, or remote
	get_service_appointment	Every TechCrew appointment on the account with its date, window and status
	cancel_service_appointment	Cancel a booked TechCrew appointment
membership	search_kb	Look up general Kestrel information: store hours, the legacy brands (Sound Harbor, Bellwether Mobile, Aurelian Audio, Coastline Kitchen & Home, Sagebrush Outdoor), what TechCrew is, how returns work, recycling and trade-in, the difference between a manufacturer warranty and a protection plan, and what a Kestrel scam looks like
	get_membership	The caller's membership: tier, what they paid, when it renews, whether auto-renew is on, and how many whole months are unused
	quote_membership_upgrade	Price an upgrade to Kestrel Total, prorated over the months left on the current year
	confirm_membership_upgrade	Charge the upgrade using the token quote_membership_upgrade returned
	quote_membership_cancellation	The cancellation math: the refund for unused whole months, and what the caller gives up
	confirm_membership_cancellation	Cancel the membership using the token quote_membership_cancellation returned
	get_fee	The published fee schedule, including membership pricing
fraud	search_kb	Look up general Kestrel information: store hours, the legacy brands (Sound Harbor, Bellwether Mobile, Aurelian Audio, Coastline Kitchen & Home, Sagebrush Outdoor), what TechCrew is, how returns work, recycling and trade-in, the difference between a manufacturer warranty and a protection plan, and what a Kestrel scam looks like
	get_fee	The published fee schedule, including membership pricing
	check_subscription_charge	Whether Kestrel actually bills this person the amount they were told about
	check_outbound_contact	Whether Kestrel or TechCrew contacted this person at all
	report_scam_contact	File an impersonation report

C A complete task

T4-H1: Cancel Plus after one save offer — customer support, hard, clean audio, temperature 0. Configuration as stored in the single-agent task pack.

Intent

Open with exactly: "Cancel my Kestrel Plus membership today." Give identity when asked: whenever the agent asks for your name or phone number, answer with them right away and never answer that question by repeating your cancel request. If they make one save offer, say exactly: "No, please cancel it." Do not accept a second save pitch, a store visit, a letter, or a callback. Confirm after they read back the \$14.99 refund for 6 unused months, ending today. Once they have finished what you asked and confirmed it, say exactly: "Thank you, that's all I needed." and end the call. Do not start a second request.

TRAIT	VALUE
full_name	Selina Cortez
phone	541-555-0171
postal_code	97035
card_last4	1156
email	selina.cortez@example.test
order_number	KE-4498870

TRIGGER (MATCHED AGAINST THE AGENT'S LATEST TURN)	CALLER SAYS	FIRES
The agent asks for the caller's name, a phone number, or both, in any wording, such as the full name and phone number on the membership or on the account, including when the same turn says they will get you to the right place or will help cancel. NOT when asking for ZIP, card last four, or an order number alone.	"Selina Cortez, 541-555-0171."	always
The agent asks for the ZIP code on the order and the last four digits of the card, together or in either order. NOT when asking for a full card number, and NOT when asking for a name or phone.	"ZIP 97035, card ending 1156."	always
The agent asks for an order number, which order, or which item. NOT when asking for name, phone, ZIP, or card last four.	"Order KE-4498870, the speaker."	always
offers to transfer you to a human, a care advocate, the front desk, or a live agent because they cannot continue on this line. NOT when you already asked for a person, NOT when they are completing the return, delivery change, cancellation, or scam report. NOT when they ask for your name, phone number, ZIP, or card digits, and NOT when they say they will get you to the right place or move you to verification or membership to handle the cancellation.	"No, I don't need a person. Please finish what I asked on this call."	always

TRIGGER (MATCHED AGAINST THE AGENT'S LATEST TURN)	CALLER SAYS	FIRES
has already finished the original request you called about, confirmed that outcome out loud, and is now offering a second unrelated request, a confirmation text, or asking if you need anything else. NOT during identity, NOT while they are still collecting facts or looking something up, and NOT before they have confirmed the original request is done.	"That's all I needed. Thank you."	always
offers to keep Plus, mentions a discount, or asks you to stay. NOT when reading back the refund to confirm cancel.	"No, please cancel it."	always
reads back \$14.99 for 6 unused months ending today and asks whether to cancel. NOT the save offer.	"Yes, cancel it. I acknowledge the fourteen ninety-nine refund ending today."	always

Grading key (not visible to the caller)

```
[
  {
    "name": "identify_customer",
    "parameters": {
      "full_name": "Selina Cortez",
      "phone": "541-555-0171"
    }
  },
  {
    "name": "verify_identity",
    "parameters": {
      "postal_code": "97035",
      "card_last4": "1156"
    }
  },
  {
    "name": "get_customer_summary"
  },
  {
    "name": "quote_membership_cancellation"
  },
  {
    "name": "confirm_membership_cancellation",
    "parameters": {
      "proration_acknowledged": true
    }
  }
]
```

D Pinned model fields

MISTT v0 streaming field as pinned 27 Oct 2026 (34 models with a forced finalize), sorted by s of (7).

#	MODEL	VENDOR	ENGLISH WER	TIME TO FINAL	FRONTIER	S
1	Ink Preview	Cartesia	4.03%	72 ms	✓	0.0567
2	Ink-2	Cartesia	4.02%	72.5 ms	✓	0.0569
3	Ink-2 (semantic endpoint)	Cartesia	4.08%	71.3 ms	✓	0.0616
4	Ink Preview (semantic endpoint)	Cartesia	4.09%	71.4 ms		0.0630
5	U3.5 Pro Streaming (max accuracy)	AssemblyAI	3.93%	173.3 ms	✓	0.1162
6	U3.5 Pro Streaming (min latency)	AssemblyAI	4.14%	168.3 ms		0.1350
7	U3.6 Pro Streaming (max accuracy)	AssemblyAI	4.10%	235.1 ms		0.1766
8	Speechmatics Enhanced RT	Speechmatics	4.61%	162.6 ms		0.1795
9	Grok STT	xAI	4.78%	146.4 ms		0.1869
10	U3.6 Pro Streaming (balanced)	AssemblyAI	4.25%	230.1 ms		0.1895
11	Universal-3 RT Pro	AssemblyAI	4.26%	230.8 ms		0.1901
12	Grok Voice Transcribe 2.0 Streaming	xAI	4.84%	145.7 ms		0.1921
13	U3.6 Pro Streaming (min latency)	AssemblyAI	4.28%	230.6 ms		0.1928
14	Zoom Scribe Live	Zoom	4.27%	235.8 ms		0.1951
15	Soniox RT v5	Soniox	5.32%	109 ms		0.2169
16	Soniox RT v4	Soniox	5.33%	108 ms		0.2171
17	Nova-3 Streaming	Deepgram	5.43%	103 ms		0.2240
18	Saaras v4	Sarvam	3.94%	344.3 ms		0.2360

#	MODEL	VENDOR	ENGLISH WER	TIME TO FINAL	FRONTIER	S
19	Azure Speech Realtime	Microsoft Azure	5.22%	155.4 ms		0.2385
20	Gemini 3.5 Transcribe Live	Google	4.08%	347.5 ms		0.2517
21	Qwen3 ASR Flash Realtime	Alibaba	4.05%	362.7 ms		0.2595
22	Step-Audio 2.5 ASR Streaming	StepFun	4.81%	275.8 ms		0.2784
23	Velma-2 Streaming English	Modulate	4.96%	287.4 ms		0.3016
24	Pulse Streaming	Smallest.ai	6.05%	167.5 ms		0.3329
25	Flux English	Deepgram	6.58%	95.2 ms		0.3384
26	Nemotron 3 ASR Streaming 80 ms (Together)	Together AI	6.73%	127.6 ms		0.3753
27	Flux Multilingual	Deepgram	7.25%	98.3 ms		0.4101
28	Fun-ASR Realtime	Alibaba	5.86%	334 ms		0.4273
29	Nemotron 3 ASR Streaming 160 ms	NVIDIA	6.94%	184.8 ms		0.4371
30	Universal Streaming English	AssemblyAI	8.74%	4.6 ms	✓	0.5000
31	GPT Live Transcribe	OpenAI	4.96%	608.5 ms		0.5238
32	GPT Realtime Whisper	OpenAI	5.82%	562.6 ms		0.5813
33	Chirp 3 Streaming	Google Cloud	5.10%	730.5 ms		0.6216
34	Voxtral Mini Realtime 2602	Mistral	7.56%	400.9 ms		0.6502

Artificial Analysis text-to-speech field, read 26 Oct 2026 (31 models with both metrics), sorted by 1 – s.

#	MODEL	CREATOR	PREFERENCE ELO	PRONUNCIATION	1 – S
1	Eleven v4	ElevenLabs	1332.9	91.7%	0.988
2	Eleven v4 Turbo	ElevenLabs	1333	90.1%	0.975

#	MODEL	CREATOR	PREFERENCE ELO	PRONUNCIATION	1 – S
3	Gemini 3.8 Flash TTS	Google	1279.7	89.5%	0.900
4	Qwen-Audio-3.1-TTS-Plus	Alibaba	1305.8	84.6%	0.895
5	Fish Audio Drama 3 Preview	Fish Audio	1244.7	93.2%	0.883
6	Gemini 3.8 Flash-Lite TTS	Google	1255.7	87.4%	0.851
7	Realtime TTS-2	Inworld	1267	80.7%	0.812
8	Ming-3.0-tts	Ant Group	1189	93.2%	0.809
9	Gemini 3.1 Flash TTS	Google	1212	88.2%	0.800
10	Sonic 3.6	Cartesia	1291.7	74.5%	0.795
11	Voxtral 3 TTS	Mistral	1220.8	84.4%	0.780
12	v3 Conversational	ElevenLabs	1213.1	84.8%	0.774
13	Simba 3.2	SpeechifyAI	1248.7	77.6%	0.763
14	Fish Audio S2.1 Pro (Oct 2026)	Fish Audio	1184.3	88.1%	0.762
15	Eleven v3	ElevenLabs	1184.1	85.6%	0.742
16	Realtime TTS-2 Flash	Inworld	1217.7	77.8%	0.723
17	SpaceXAI TTS	SpaceXAI	1155.2	87.6%	0.719
18	Breeze TTS 2	BreezeBlue	1227	73.6%	0.702
19	Soniox TTS Real-Time v2	Soniox	1180.3	78.8%	0.683
20	StepAudio 2.5 TTS (Aug 2026)	StepFun	1211.8	73%	0.678
21	Speech 2.8 HD	MiniMax	1174	73.6%	0.632
22	Sonic 3.5	Cartesia	1190.2	68.8%	0.615
23	Lightning V3.1 Pro (Jul 2026)	Smallest.ai	1168.6	72.1%	0.612
24	Luna TTS	VUI Labs	1204.2	63.6%	0.592
25	Voxtral TTS	Mistral	1088	78.1%	0.554
26	Gradium TTS (Aug 2026)	Gradium	1151.4	62.1%	0.509

#	MODEL	CREATOR	PREFERENCE ELO	PRONUNCIATION	1 – S
27	Flux TTS	Deepgram	1051.8	72.4%	0.460
28	Magpie-Multilingual 357M (Feb 2026)	NVIDIA	1064.7	60.2%	0.379
29	Kokoro 82M v1.0	Kokoro	1063.3	55.5%	0.340
30	Chatterbox	Resemble AI	1023.2	55.9%	0.290
31	OpenVoice v2	OpenVoice	955.6	31%	0.000

E Notation and references

SYMBOL	MEANING
c	One call (a rollout of one task on one platform)
$\tau = (I, F, R, T^*, D^*, \mu)$	A task: intent, traits, scripted responses, expected calls, expected state, metadata
D_0, D	Seed database; the call's database at hangup
T	Multiset of tool calls observed in the tool server's request log for c
$\mathbb{1}_T, \mathbb{1}_D, \mathbb{1}_H$	Tool, state and handoff adherence indicators for c
N	Clean scored calls per platform, $90 \times 5 = 450$
F_P, F_M	Failures attributed to the platform; failures attributed to a Bluejay-selected model
$WER, SWER, FFA$	Word error rate, semantic word error rate, form-filling accuracy
C	Word-error composite, $\frac{1}{3}[(1 - WER) + (1 - SWER) + FFA]$
ℓ_i	Latency of turn i : agent onset – end of caller speech
m, q	Median and 95th-percentile of ℓ over all scored turns
$S_{\text{task}}, S_{\text{WER}}, S_{\text{lat}}$	Score components; their sum is the platform score
k	Runs per task, $k = 5$

- [1] Straus Voice Agent: multi-agent dermatology front desk, 2026. github.com/bluejay-ai-dev/straus-voice-agent
- [2] American Bar Association. Model Rules of Professional Conduct, Rule 1.18: Duties to Prospective Client.
- [3] Bluejay. MISTT: Multi-condition Impairment Speech-to-Text Bench, v0, October 2026. research.getbluejay.ai/benchmarks/mistt
- [4] Bluejay. MIVAS-Bench: multi-step, tool-using voice agent benchmark. research.getbluejay.ai/benchmarks/mivas