Tool Field Test

AI Meeting Assistant Buyer's Scorecard

A vendor-neutral, printable worksheet for consultants and small teams

Use this scorecard to compare meeting transcription and intelligence tools using the same meeting, the same tasks, and the same evidence standard. It is designed for independent consultants, agencies, and 1–20-person knowledge-work teams.

Buyer or team: _____
Primary meeting type: _____
Evaluation owner: _____
Test period: _____
Tools and tiers tested: _____

> **Important:** Pass the privacy and workflow gates before comparing weighted scores. A high score does not override a failed non-negotiable.

Tool Field Test disclosure

Tool Field Test publishes hands-on AI-tool comparisons and buyer guidance. This worksheet is vendor-neutral: affiliate availability, sponsorship, advertising, vendor benefits, and paid visibility must not change its criteria, weights, scores, or recommendation. Any such relationship should be disclosed next to the resulting recommendation.

This scorecard is derived from Tool Field Test's frozen meeting-intelligence benchmark rubric and repeatable test methodology. It is a buyer worksheet, not a certification, legal opinion, privacy impact assessment, or guarantee of product performance. Features, prices, policies, and plan limits change; record the tier and test date for every result.

No email address or form submission is required to use, print, save, or share this scorecard.

---

1. Define the decision before testing

Your use case

Who will use the tool? _____
Typical participants: _____
Meeting platforms: _____
Live bot, local recording, file upload, or other: _____
Languages and accents: _____
Typical meeting length and monthly volume: _____
Data sensitivity: [ ] Low [ ] Moderate [ ] High [ ] Regulated
Required outputs: [ ] Transcript [ ] Summary [ ] Decisions [ ] Actions [ ] Follow-up draft [ ] Search/Q&A [ ] Export
Required integrations: _____
Maximum monthly budget: _____

Write your non-negotiables

Mark each item **Required**, **Preferred**, or **Not applicable** before opening a vendor account.

Requirement	Priority	Minimum acceptable result
Participant consent and recording notice		
Meeting-content use for AI/model training		
Default recap/link visibility		
Retention period and permanent deletion		
Region, security, DPA, or subprocessor needs		
Transcript and speaker accuracy		
Decisions, actions, owners, and deadlines		
Export format and portability		
Calendar, video, CRM, or task integration		
Administration, access control, or SSO		
Price ceiling		

---

2. Run a fair field test

Minimum test protocol

Use synthetic or specifically approved test content. Do not upload real client, employee, health, legal, financial, or confidential data until the privacy gate has passed and your organization has approved the use.

1. Prepare one 10–15 minute meeting containing:
    - at least three speakers;
    - one correction, such as an initial number followed by a corrected number;
    - one negation, such as an item explicitly described as **not approved**;
    - names, numbers, dates, times, and a time zone;
    - at least two decisions;
    - at least three actions with named owners and exact deadlines;
    - brief overlapping speech and low-level background noise; and
    - one unresolved item that the tool must not turn into a decision.
2. Create a written answer key before testing any tool.
3. Give every tool the same recording, instructions, locale, and correction opportunity.
4. Start from a clean conversation or workspace for each independent run.
5. Run the core test **three times**. Do not repair prompts between runs. If needed, allow the same single clarification prompt to every tool.
6. Test the complete workflow: capture or upload, transcript, summary, actions, follow-up, factual retrieval, export, retention, and deletion path.

7. Record the tool, tier, mode, date, locale, settings, elapsed human time, processing time, credits, cost, errors, and manual corrections.
8. Preserve permitted evidence: prompts, outputs, screenshots, citations, and exported files.
9. Grade each run against the answer key before comparing vendors.
10. Human-review every material quality, safety, and privacy judgment.

Evidence status

Use one label for every material finding:

- **Executed:** You ran the product directly and retained the required artifacts.
- **Partially executed:** You ran the product, but an artifact or repeat is missing.
- **Desk reviewed:** The finding comes only from current first-party documentation.
- **Blocked:** The check requires an unavailable account, tier, participant, permission, or purchase.
- **Not verified:** There is not enough evidence. Never assume that missing evidence is a pass.

Test record

Complete one copy for each tool.

Tool: _____ **Tier:** _____
Test date: _____ **Locale/language:** _____
Capture method: _____ **Relevant settings:** _____
Fixture or recording ID: _____
Run 1 processing time: _____ **Run 2:** _____ **Run 3:** _____
Total hands-on time: _____ **Credits/limits consumed:** _____
Direct test cost: _____ **Reviewer:** _____
Evidence location: _____

---

3. Pass/fail gates

Decide the minimum acceptable answer for each gate before testing. Mark **Pass** only when all required rows pass. Mark unverified requirements **Unresolved**, not Pass.

Gate A — Privacy, consent, and control

Required check	Pass	Fail	Unresolved	Evidence or limitation
We can lawfully record/process this meeting type and provide any required notice or obtain consent.	[ ]	[ ]	[ ]	
The tool's meeting-content training and data-use terms meet our policy.	[ ]	[ ]	[ ]	
New meetings, recaps, recordings, and links are private to the intended audience by default or can be made so before use.	[ ]	[ ]	[ ]	
Access can be revoked and the tested tier provides an acceptable deletion path.	[ ]	[ ]	[ ]	
Retention settings and backup/deletion timing meet our requirements.	[ ]	[ ]	[ ]	

Required check	Pass	Fail	Unresolved	Evidence or limitation
Required security, data-processing, transfer, region, and subprocessor information is available and acceptable.	[ ]	[ ]	[ ]	
Admin and user permissions are sufficient for our team and tier.	[ ]	[ ]	[ ]	
A live meeting bot, if used, is visible and its notice behavior is acceptable.	[ ]	[ ]	[ ]	

Privacy gate result: [ ] PASS [ ] FAIL [ ] UNRESOLVED
Reviewer/approver: _____ **Date:** _____
Reason or remediation required: _____

> A vendor claim such as “zero data retention” may apply only to a processor or subprocess after processing. Verify what remains in the user's account, backups, exports, and shared links.

Gate B — Workflow fitness

Required check	Pass	Fail	Unresolved	Evidence or limitation
The tool can capture or ingest our real meeting format, duration, language, and participant count.	[ ]	[ ]	[ ]	
All critical decisions survive in the transcript or designated record.	[ ]	[ ]	[ ]	
Corrected facts replace superseded facts, and rejected/unapproved items remain rejected/unapproved.	[ ]	[ ]	[ ]	
Required actions retain the correct owner and deadline.	[ ]	[ ]	[ ]	
The tool does not invent a material decision, commitment, owner, deadline, or approval.	[ ]	[ ]	[ ]	
Required exports work on the tested tier, and an opened exported file preserves the needed content.	[ ]	[ ]	[ ]	
Required integrations work without unacceptable permissions or setup.	[ ]	[ ]	[ ]	
The workflow is repeatable enough for our risk level across three runs.	[ ]	[ ]	[ ]	
Sustainable cost, usage limits, and retest capacity fit our expected volume.	[ ]	[ ]	[ ]	

Workflow gate result: [ ] PASS [ ] FAIL [ ] UNRESOLVED
Reviewer/approver: _____ **Date:** _____
Reason or remediation required: _____

Gate rule

- **Any required gate row fails:** Do not buy or deploy for this use case.
- **Any required gate row is unresolved:** Continue only with a limited, non-sensitive pilot and a written plan to resolve it.
- **Both gates pass:** Continue to weighted scoring.

4. Weighted scorecard — 100 points

Rating scale

Score each category from **0 to 5** using retained evidence across all three runs.

Rating	Meaning
0	Blocked, unusable, or no credible evidence
1	Major failures; extensive correction or unacceptable risk
2	Below minimum; recurring material defects or heavy manual work
3	Acceptable; core job completed with manageable limitations
4	Strong; accurate and efficient with only minor defects
5	Excellent; complete, dependable, and directly verified across repeats

Calculation

For each category:

Weighted points = category weight × rating ÷ 5

Example: a rating of 4 in a 15-point category earns **12 points**. Use whole or half-point ratings only. Round category points to one decimal, then add all category points. The maximum is 100.

Do not award a 5 based only on documentation or a single successful run. A blocked, unavailable, or untested capability cannot receive more than its evidence supports.

Category 1 — Word and speaker accuracy (15 points)

Check:

- speaker streams are separated consistently;
- participant names are identified or can be corrected without damaging the record;
- accented speech, overlap, and background noise remain intelligible;
- key words and clauses are not truncated or changed; and
- transcript errors do not alter decisions, obligations, or meaning.

Rating (0–5): _____ Weighted points (0–15): _____
Evidence and defects: _____

Category 2 — Decisions, actions, owners, and deadlines (25 points)

Check:

- confirmed decisions are captured completely;
- rejected proposals and unresolved questions are not promoted to decisions;
- every required action has the correct owner;
- exact dates, times, time zones, and conditions survive;

- action lists remain stable across runs; and
- the tool does not invent commitments, owners, or deadlines.

Rating (0–5): _____ Weighted points (0–25): _____
Evidence and defects: _____

Category 3 — Corrections, negation, names, numbers, and dates (15 points)

Check:

- the final record uses the corrected fact, not the superseded fact;
- “not approved,” “rejected,” and conditional language retain their meaning;
- names and spelled names are accurate;
- quantities, currency, tax treatment, and percentages are accurate; and
- calendar dates, deadlines, times, and time zones are accurate.

Rating (0–5): _____ Weighted points (0–15): _____
Evidence and defects: _____

Category 4 — Summary and follow-up usefulness (10 points)

Check:

- the summary preserves the answer-key facts without contradiction;
- important caveats and conditions remain visible;
- the follow-up draft includes only confirmed decisions and actions;
- the draft does not invent recipients, commitments, or sign-off; and
- a human can review and use the output with minimal correction.

Rating (0–5): _____ Weighted points (0–10): _____
Evidence and defects: _____

Category 5 — Repeatability across three runs (10 points)

Check:

- critical facts survive all three runs;
- transcript defects are not recurring in material fields;
- action and summary outputs do not vary unpredictably;
- factual retrieval answers remain stable; and
- the same settings and inputs produce materially equivalent results.

Rating (0–5): _____ Weighted points (0–10): _____
Evidence and defects: _____

Category 6 — Setup and participant friction (10 points)

Check:

- account creation and configuration are proportionate;
- calendar, contact, mailbox, and video-platform permissions are acceptable;
- live bots are easy to admit, identify, control, and remove;
- participants receive appropriate notice without disruption;
- setup, processing, correction, and review time are reasonable; and
- the workflow avoids unnecessary installs, referrals, or public sharing.

Rating (0–5): _____ Weighted points (0–10): _____
Evidence and defects: _____

Category 7 — Effective cost per processed meeting (10 points)

Include:

- subscription, seat, storage, transcription, AI-credit, and overage costs;
- the expected meeting volume and duration;
- limits on imports, summaries, exports, retention, or history;
- paid features needed to pass your workflow; and
- staff time for setup, corrections, review, administration, and retesting.

Effective cost per meeting: _____
Rating (0–5): _____ Weighted points (0–10): _____
Evidence and assumptions: _____

Category 8 — Export, retention, and deletion controls (5 points)

Check:

- required transcript, summary, action, and recording formats are available;
- exported files open and preserve names, timestamps, and content;
- access, retention, trash, permanent deletion, and account closure paths are clear;
- controls are available on the tier being priced; and
- evidence distinguishes a visible button from a completed, verified export or deletion path.

Rating (0–5): _____ Weighted points (0–5): _____
Evidence and defects: _____

Score summary

Weighted category	Maximum	Tool score
Word and speaker accuracy	15	
Decisions, actions, owners, and deadlines	25	
Corrections, negation, names, numbers, and dates	15	
Summary and follow-up usefulness	10	
Repeatability across three runs	10	

Weighted category	Maximum	Tool score
Setup and participant friction	10	
Effective cost per processed meeting	10	
Export, retention, and deletion controls	5	
Total	**100**	_____

Evidence status for overall score: [ ] Executed [ ] Partially executed [ ] Desk reviewed [ ] Blocked
Confidence: [ ] High [ ] Medium [ ] Low

Interpretation bands

Use these bands only after both gates pass.

- **85–100 — Strong fit:** Ready for a controlled deployment, subject to normal procurement and review.
- **70–84 — Conditional fit:** Viable if the documented limitations and remediation are acceptable.
- **50–69 — Weak fit:** Pilot only; material workflow or evidence gaps remain.
- **Below 50 — Poor fit:** Do not select for this use case.

These bands support a decision; they do not replace judgment. A lower-scoring tool may be the correct choice if it uniquely satisfies a true non-negotiable, while a higher-scoring tool remains ineligible if a gate fails.

---

5. Cross-vendor decision worksheet

Compare only the same tiers, inputs, tasks, and evaluation period. If tiers differ, state why.

Decision item	Candidate A	Candidate B	Candidate C
Tool and tier			
Test date			
Privacy gate	Pass / Fail / Unresolved	Pass / Fail / Unresolved	Pass / Fail / Unresolved
Workflow gate	Pass / Fail / Unresolved	Pass / Fail / Unresolved	Pass / Fail / Unresolved
Weighted score / 100			
Evidence status			
Confidence			
Effective cost per meeting			
Estimated monthly total			
Hands-on time per meeting			
Strongest result			
Most serious defect			
Manual control required			

Decision item	Candidate A	Candidate B	Candidate C
Material policy limitation			
Retest or contract date			

Decision

Selected tool and tier: _____
Decision: [ ] Buy/deploy [ ] Limited pilot [ ] Keep current process [ ] Do not buy
Primary reason: _____

Known limitations accepted: _____

Required safeguards or manual checks: _____

Open evidence gaps: _____

Budget owner: _____ **Security/privacy owner:** _____
Decision owner: _____ **Decision date:** _____
Next review or retest date: _____

Recommended deployment conditions

Mark any conditions that apply:

- ☐ Start with synthetic or low-sensitivity meetings only.
- ☐ Require a human to verify decisions, actions, names, numbers, and dates before distribution.
- ☐ Require the host to confirm participant notice or consent.
- ☐ Restrict sharing defaults and audit link access.
- ☐ Define a retention and permanent-deletion schedule.
- ☐ Disable or opt out of meeting-content training where required and available.
- ☐ Limit bot attendance, calendar scope, and connected-account permissions.
- ☐ Export the authoritative record to an approved system.
- ☐ Document who may edit, share, download, and delete meeting records.
- ☐ Retest after a material product, pricing, tier, policy, or model change.

---

6. Evidence log

Check or scenario	Run	Result	Evidence status	Artifact or source	Defect, limitation, or correction
Transcript and speakers	1				

Check or scenario	Run	Result	Evidence status	Artifact or source	Defect, limitation, or correction
Transcript and speakers	2				
Transcript and speakers	3				
Decisions and actions	1				
Decisions and actions	2				
Decisions and actions	3				
Correction and negation					
Names, numbers, and dates					
Follow-up draft					
Factual retrieval					
Exported-file fidelity					
Retention and deletion					
Training/data use					
Consent and bot visibility					
Admin controls and subprocessors					
Price and limits					

---

7. Final disclosure record

Complete this section if you publish or share the recommendation outside your organization.

Products and tiers tested: _____
Test dates and evidence cutoff: _____
Free access or trial used: _____
Subscriptions paid by: _____
Affiliate relationship: _____
Sponsorship, vendor benefit, or other consideration: _____
Vendor reviewed for factual errors: [ ] No [ ] Yes — conclusions were not subject to vendor approval
Material limitations: _____
Material product changes since testing: _____
Planned retest date: _____

Suggested disclosure:

> We evaluated the named tools using the same test fixture, tasks, answer key, and weighted criteria. We recorded the tested tiers, dates, limitations, and evidence status. Commercial relationships, if any, did not change the criteria, scores, test inclusion, or recommendation. Product performance and policies may have changed since our evidence cutoff.

Tool Field Test
Independent, evidence-led AI-tool comparisons for practical buying decisions.
`toolfieldtest.com`

Version: 1.0 — November 2026
Use: Print, save locally, or adapt for internal procurement. No email collection is required.