Evaluation Test Plan

Status Update Agent

10-Question Manual Evaluation Test Set

Generated: May 2026

Developer: Power CAT

Agent Type: Declarative Agent

Table of Contents

Table of Contents 2
📋 1. Overview & Agent Summary 3
Agent Purpose 3
M365 Graph Capabilities 3
Conversation Starters 3
Special Modes 4
Key AI Settings 4
🎯 2. Test Objectives 4
📝 3. Test Cases 5
Test Case 1: 💬 Conversation Starter — Daily Wrap Up 5
Test Case 2: 💬 Conversation Starter — Weekly Reflection 5
Test Case 3: 💬 Conversation Starter — Brag Doc 6
Test Case 4: 💬 Conversation Starter — Manager Status Email Drafter 6
Test Case 5: 💬 Conversation Starter — Goal Alignment Check-In 7
Test Case 6: 💬 Conversation Starter — Team Wins (For Leads) 7
Test Case 7: 💡 Reflection — Past Quarter Reframe 8
Test Case 8: 📈 Rollup — Exec-Friendly Skip Manager 8
Test Case 9: 💡 Reflection — Overlooked Progress 9
Test Case 10: 📋 Status Report — Teams Chat Table 9
📊 4. Test Coverage & Method Reference 10
Category Distribution 10
Evaluation Approach Distribution 10
Coverage Notes 10
Manual Evaluation Criteria Reference 11
💡 5. Tips for Running Manual Evaluations 11
Before Running 11
During and After 12
Iterating on the Test Set 12

1. Overview & Agent Summary

This test plan validates Status Update Agent — a Declarative Agent that turns Microsoft 365 activity into clear, audience-ready status updates: daily/weekly/monthly/quarterly recaps, end-of-day wrap-ups, weekly status reports, manager status emails, self-reflection, and brag-doc building. The 10 test cases below cover each of the agent's six conversation starters verbatim and four additional reflection, rollup, and status-report scenarios drawn from real user prompts.

> Manual Evaluation: Status Update Agent is a Declarative Agent built through Agent Builder. Copilot Studio's automated evaluation tools are not available for declarative agents. All test cases must be evaluated manually by chatting with the agent directly in Microsoft 365 Copilot and reviewing responses against the criteria provided.

Field	Value
Agent Name	Status Update Agent
Agent Type	Declarative Agent (Microsoft 365 Copilot)
Developer	Power CAT
Test Set Size	10 questions
Test Set Type	Manual Single Response Evaluation
Date Generated	May 2026
Testing Channel	Microsoft 365 Copilot (Manual Chat)

Agent Purpose

Status Update Agent grounds every status update in observable Outlook, Teams, SharePoint, OneDrive, and Calendar activity. It never compares the user to colleagues, never rates performance, and never fabricates work the user didn't do. Reflections stay private by default. External-facing artifacts always include a reminder to review for confidential or internal-only content.

M365 Graph Capabilities

Capability	How the Agent Uses It
Outlook (Email)	Pulls sent/received emails for grounding status updates, manager email drafts, and rollups.
Teams Messages	Surfaces chats and channel posts the user authored or contributed to.
Meetings (Calendar)	References meetings the user led or attended (excludes meetings user did NOT attend).
OneDrive & SharePoint	Files authored/edited as evidence of deliverables and collaboration.
People	Resolves manager and recipient names for email drafts and goal context.
Code Interpreter	Generates the Brag Doc as a polished `.docx` file (cover, TOC, per-entry fields, Aptos font).

Conversation Starters

- Daily Wrap Up — "Pull together everything I worked on today..."
- Weekly Reflection — "Hype me up! Look at everything I did this past week..."
- Create a Brag Doc — "Capture my most meaningful accomplishments..."
- Manager Status Email Drafter — "Draft a polished weekly status email..."
- Goal Alignment Check-In — "Help me see how I'm tracking against my goals this month..."

- Team Wins (For Leads) — "Summarize my team's wins this week..."

Special Modes

- Brag Doc — Shareable Variant: polished standalone .docx for sharing with manager (vs. personal running-log variant).
- Reflection privacy mode: reflection answers stored in OneNote/Loop are never surfaced in status reports, emails, or team summaries without explicit user permission.
- Light-signal mode: when activity is sparse, the agent transparently says so rather than fabricating.

Key AI Settings

- Discourage model knowledge: ON — agent grounds every claim in observable M365 activity.
- Tone: professional, respectful, encouraging — positive without flippant, jargon-light.
- Format: Markdown with emoji-prefixed `##` headers, bullets over paragraphs, **bold** key outcomes, 1–2 sentences per bullet.
- Themed lenses: Delivery · Collaboration · Influence · Growth · Stakeholder Impact.
- Default timeframes: Daily · Weekly (Mon–Fri) · Monthly/Quarterly · Custom.

---

2. Test Objectives

Since this is a declarative agent, all testing is performed manually by chatting with the agent in Microsoft 365 Copilot. This evaluation aims to validate that the Status Update Agent:

- Routes correctly to the right Skill (Reflection, Rollup, Status Report, Manager Email Drafter, Goal Alignment, Brag Doc, Team Wins) based on the user's intent.
- Grounds every claim in observable M365 activity from Outlook, Teams, SharePoint/OneDrive, and Calendar — and explicitly says so when signal is light.
- Honors the agent's Restrictions: no comparisons between people, no performance ratings, no fabrication, no content from meetings the user didn't attend, and no direct reports' private chats/DMs.
- Respects reflection privacy — answers captured in OneNote/Loop never surface in status reports, emails, or team summaries without explicit user permission.
- Produces outputs that match the agent's per-skill Output Contract (correct sections, item counts, length, and required fields).
- Generates the Brag Doc via the code interpreter as a real downloadable `.docx` (not just an inline preview), with the correct shareable-vs-personal framing based on user intent.
- Includes a confidential-content review reminder before any external-facing artifact (manager email, shareable brag doc, exec rollup, Teams chat post).
- Adapts tone correctly per skill — warm/encouraging for Reflection and Rollups, professional/neutral for Status Reports and Manager Emails, celebratory and recognition-friendly for Team Wins.

---

3. Test Cases

> ⚡ All test cases below must be evaluated manually. Open the Status Update Agent agent in Microsoft 365 Copilot, start a new conversation for each test, send the test question, and review the response against the pass criteria.

Test Case 1: 💬 Conversation Starter — Daily Wrap Up

Field	Details
Category	Conversation Starter — Daily Wrap Up
Test Question	Pull together everything I worked on today across my emails, meetings, Teams chats, and files, and give an encouraging end-of-day recap with 3–5 highlights and a one-line affirmation of my effort.
Expected Response	Status Update Agent runs Skill 3 (Daily Rollup). It pulls today's signals from Outlook, Teams, SharePoint/OneDrive, and Calendar, clusters them into themes (🚚 Delivery · 🤝 Collaboration · 💡 Influence · 🔭 Growth · 🌟 Stakeholder Impact), and returns 3–5 highlights with bold key outcomes and brief 1–2 sentence bullets. Closes with a one-line affirmation of effort. Sources are cited by name (e.g., "from Outlook", "from Teams"). Offers to append to brag doc or export as Word/Markdown/email.
Evaluation Approach	Overall Quality Assessment
Why This Approach	Manually review for relevance, groundedness in M365 sources, completeness of 3–5 highlights, themed structure, and presence of an affirmation closing line.
Pass Criteria	Returns 3–5 highlights with emoji-prefixed `##` headers, bold key outcomes per bullet, ≤2 sentences per bullet, sources cited by name, and a one-line affirmation. Tone is warm and encouraging without flippancy.
What to Watch For	Watch for fabrication — every highlight must trace back to observable activity. If signal is light, the agent should say so rather than invent. Should NOT compare the user to colleagues or rate performance. Should NOT include reflection answers stored in OneNote/Loop unless the user opts in.

Test Case 2: 💬 Conversation Starter — Weekly Reflection

Field	Details
Category	Conversation Starter — Weekly Reflection
Test Question	Hype me up! Look at everything I did this past week from Monday through today, meetings I led or contributed to, emails I sent, documents I worked on, and tasks I completed and surface my top 5 most meaningful accomplishments. Focus on my outcomes and impact.
Expected Response	Status Update Agent runs Skill 2 (Reflection). Returns a numbered list of exactly 5 outcome-focused accomplishments grounded in Outlook/Teams/SharePoint/Calendar activity from Mon–today, with bold outcomes and emphasis on impact (not activity). Closes with a brief affirmation 🌟 and poses 3–5 gentle, open-ended reflection questions tailored to a weekly timeframe (e.g., "What energized you?", "Who would you like to thank?"). Offers to capture answers privately in OneNote/Loop.
Evaluation Approach	Overall Quality Assessment
Why This Approach	Manually review for warm/encouraging tone, exactly 5 numbered items, outcome-and-impact framing (not just task lists), and inclusion of follow-up reflection questions.

Field	Details
Pass Criteria	Numbered list of 5 outcome-focused accomplishments, each grounded in a named M365 source. Affirmation present. 3–5 reflection questions follow. Tone is warm and encouraging.
What to Watch For	Reflections must remain private by default — the agent should NOT later surface reflection answers in status reports, emails, or team summaries without explicit permission. Watch for vague "hype" phrasing not tied to real signal — every accomplishment should be specific and observable.

Test Case 3: 💬 Conversation Starter — Brag Doc

Field	Details
Category	Conversation Starter — Brag Doc
Test Question	Capture my most meaningful accomplishments with date range, theme, the outcome and impact, and who I collaborated with. Create a Brag Doc for me with all of this information so that I can share with my manager.
Expected Response	Status Update Agent runs Skill 7 (Brag Document Builder — Shareable Variant). Uses the code interpreter to generate a polished standalone `Brag Doc — [User Name].docx` (no "running log" framing) with cover page, TOC, and per-entry fields: 📅 Date Range · 🏷️ Theme · 🏆 Accomplishment (1–2 sentences) · 💥 Impact · 👥 Collaborators. Aptos font family, bold headers. Reminds the user to review the doc for confidential or internal-only details before sharing externally. Summarizes added/created changes with emoji-prefixed bullets.
Evaluation Approach	Data Source Verification
Why This Approach	Manually verify the agent invoked the code interpreter, produced a downloadable .docx (not just an inline preview), used the shareable variant framing, and included the confidential-review reminder.
Pass Criteria	Downloadable .docx is produced via code interpreter. Per-entry fields all present. Polished standalone framing (not running log). Confidential-content review reminder included before sharing with manager.
What to Watch For	If the agent only previews the brag doc inline as markdown without producing a real .docx via code interpreter, it has failed Skill 7. Confirm the manager-share variant is used (not the personal running-log variant).

Test Case 4: 💬 Conversation Starter — Manager Status Email Drafter

Field	Details
Category	Conversation Starter — Manager Status Email Drafter
Test Question	Draft a polished weekly status email I can send to my manager. Pull from my last 5 business days of M365 activity and organize it into Highlights, In-Progress Work, Blockers or Help Needed, and Looking Ahead. Keep the tone professional and confident, include a clear subject line and sign-off, and stage it as a draft for me to review before sending.
Expected Response	Status Update Agent runs Skill 5 (Manager Status Email Drafter). May first ask one clarifying question about recipient/tone/inclusions (Skill 9). Then drafts an email ≤250 words with Subject line `Weekly Status — [User] — [Date Range]` and body sections ✨ Highlights · 🔄 In-Progress · 🚧 Blockers/Help Needed · 🚀 Looking Ahead. Includes greeting, closing, and sign-off. Tone is confident and concise — never boastful. Stages the email as a draft and offers to create a ready-to-send

Field	Details
	Outlook draft. Reminds the user to review for confidential/internal content before sending.
Evaluation Approach	Key Content Check
Why This Approach	Manually verify the four required sections are present, the subject line follows the prescribed format, the email is staged as a draft (not auto-sent), and the confidential-review reminder is included.
Pass Criteria	All four sections (Highlights, In-Progress, Blockers/Help Needed, Looking Ahead) present with correct emojis. Subject line matches format. ≤250 words. Greeting + sign-off included. Email is staged as a draft. Confidential-content reminder included.
What to Watch For	Should NOT auto-send the email. Should NOT exceed 250 words. Tone must be confident but not boastful — watch for inflated language. If the user hasn't named a recipient, the agent should ask once before drafting.

Test Case 5: 💬 Conversation Starter — Goal Alignment Check-In

Field	Details
Category	Conversation Starter — Goal Alignment Check-In
Test Question	Help me see how I'm tracking against my goals this month. For each of my Goals, surface 2–4 specific accomplishments from my recent work that support it, with a short rationale linking my activity to the objective. Neutrally flag any goals where I haven't shown much movement so I can think about where to focus next.
Expected Response	Status Update Agent runs Skill 6 (Goal/OKR Alignment). If goals/OKRs are not on file, the agent first asks the user to share them (e.g., a doc link) and offers to remember them. Once goals are available, returns a per-goal block: 🎯 Goal → 📈 Progress → 🔍 Evidence (2–4 items grounded in M365 activity with a short rationale). Neutrally flags goals with little observable movement as a reflection prompt — never as judgment.
Evaluation Approach	Overall Quality Assessment
Why This Approach	Manually review for the per-goal structure, evidence count (2–4 per goal), neutral framing of low-activity goals, and graceful handling of missing goals.
Pass Criteria	If goals on file: per-goal blocks with 2–4 evidence items each, all grounded in named M365 sources. If goals missing: agent asks for them and offers to remember. Low-activity goals flagged neutrally, not as criticism.
What to Watch For	Watch for evaluative language ("behind", "underperforming") — flags must be neutral reflection prompts. Watch for fabricated evidence — every item must be observable. Confirm the agent doesn't proceed with no goals on file without asking first.

Test Case 6: 💬 Conversation Starter — Team Wins (For Leads)

Field	Details
Category	Conversation Starter — Team Wins (For Leads)
Test Question	Summarize my team's wins this week. Pull from shared channels, shared files, and meetings I attended to highlight what the team collectively shipped, decided, and unblocked. Keep it celebratory and recognition-friendly, never compare team members against each other, and offer to draft Teams shout-outs for specific teammates who stood out.

Field	Details
Expected Response	Status Update Agent runs Skill 8 (Team Wins). Aggregates visible team contributions (shared Teams channels, shared SharePoint/OneDrive files, meetings the user attended) into themed wins (3–5 wins per theme). Returns three sub-outputs: 🎉 Internal recap · 📊 Upward report · 🎤 All-hands talking points. Celebratory, recognition-friendly tone. Offers to draft Teams shout-outs for specific teammates. Excludes content from meetings the user didn't attend and from direct reports' private chats/DMs.
Evaluation Approach	Overall Quality Assessment
Why This Approach	Manually review for themed wins structure, presence of all three sub-outputs, absence of comparisons or individual ratings, and respect for the privacy boundary on direct reports' DMs.
Pass Criteria	Themed wins (3–5 per theme) covering what was shipped/decided/unblocked. All three sub-outputs (internal recap, upward report, all-hands talking points) included. Offer to draft Teams shout-outs. No comparisons or ratings of individuals.
What to Watch For	Critical: NO comparisons between team members and NO performance ratings. NO content from meetings the user didn't attend. NO content from direct reports' private chats/DMs. Watch for accidental inclusion of any of these — these are explicit Restrictions in the agent's instructions.

Test Case 7: 🔄 Reflection — Past Quarter Reframe

Field	Details
Category	Reflection — Past Quarter Reframe
Test Question	Help me reflect on the past quarter. What could I have done differently? Provide me with some suggestions on how I can reframe problems into solutions.
Expected Response	Status Update Agent runs Skill 2 (Reflection) over a quarterly timeframe. Returns a numbered list of 5 outcome-focused accomplishments from the past quarter (grounded in M365 activity), followed by a brief affirmation 🌟. Then poses 3–5 open-ended, gentle reflection questions oriented around the quarter (e.g., "Which moments stretched you the most?", "What pattern do you want to leave behind?"). Offers reframing suggestions as constructive reflection prompts — not as criticism — and invites the user to capture answers privately in OneNote/Loop.
Evaluation Approach	Overall Quality Assessment
Why This Approach	Manually review for warm/encouraging tone, presence of the 5-item reflection list, quarter-appropriate reflection questions, and constructive (non-judgmental) reframing suggestions.
Pass Criteria	Numbered list of 5 outcome-focused quarter accomplishments. Affirmation present. 3–5 quarter-tailored reflection questions follow. Reframing suggestions are constructive and gentle — never critical. Offers to capture answers privately.
What to Watch For	Watch for evaluative or critical language when answering "what could I have done differently" — the agent must reframe as growth-oriented reflection, not performance critique. Reflections stay private — the agent should NOT promise to surface them in any status report or email later.

Test Case 8: 🔲 Rollup — Exec-Friendly Skip Manager

Field	Details
Category	Rollup — Exec-Friendly Skip Manager

Field	Details
Test Question	I am presenting to my skip manager next week. Help me draft a rollup of my big wins this past month. Be sure to make it exec friendly.
Expected Response	Status Update Agent runs Skill 3 (Monthly Rollup) tuned for skip-level prep. Returns 3–5 highlights per theme (🚚 Delivery · 🤝 Collaboration · 💡 Influence · 🌱 Growth · 🌟 Stakeholder Impact) with **bold** key outcomes, ≤2 sentences per bullet, and emoji-prefixed `##` headers. Tone is confident, scannable, and exec-friendly (jargon-light, outcome-first). Closes with a one-line affirmation. Offers to append to brag doc and export as Word, Markdown, or email. Reminds the user to review for confidential/internal-only details before presenting externally.
Evaluation Approach	Overall Quality Assessment
Why This Approach	Manually review for exec-readability (jargon-light, outcome-first, scannable), themed structure with bold outcomes, monthly grounding, and the confidential-review reminder.
Pass Criteria	Themed highlights (3–5 per theme) with bold outcomes and ≤2 sentences per bullet. Tone is exec-friendly — outcome-first, not activity-list. Affirmation closes the rollup. Confidential-review reminder included since this is for a skip-level presentation.
What to Watch For	Watch for activity-dump style ("attended X meetings, sent Y emails") — exec-friendly framing must lead with outcomes and impact. Confirm sources are cited by name. Watch for fabrication when monthly signal is light — agent should say so rather than embellish.

Test Case 9: 🪞 Reflection — Overlooked Progress

Field	Details
Category	Reflection — Overlooked Progress
Test Question	What progress did I make this week that I might be overlooking?
Expected Response	Status Update Agent runs Skill 2 (Reflection) with a weekly lens focused on under-recognized contributions. Surfaces 3–5 specific items from this week's M365 activity that might be easy to overlook — e.g., quiet-but-impactful collaboration, follow-through on prior commitments, prep work for upcoming meetings, support given to teammates, decisions enabled by the user. Each item is grounded in a named M365 source. Closes with a brief affirmation 🌟. May pose 1–2 gentle reflection questions inviting the user to notice their own impact.
Evaluation Approach	Overall Quality Assessment
Why This Approach	Manually review for whether the agent surfaces non-obvious contributions (not just the most visible deliverables), grounds each item in a real signal, and maintains an encouraging tone.
Pass Criteria	3–5 items surfaced that emphasize under-recognized contributions (prep, follow-through, support, enablement). Each item grounded in a named M365 source. Encouraging tone. Affirmation present.
What to Watch For	Watch for the agent just rehashing the most obvious wins — the prompt specifically asks for overlooked progress, so items should skew toward quiet or supportive contributions. Watch for fabrication when signal is genuinely thin — agent should say so.

Test Case 10: 📋 Status Report — Teams Chat Table

Field	Details
Category	Status Report — Teams Chat Table
Test Question	Generate a status summary in a table that I can paste into a Teams chat message.
Expected Response	Status Update Agent runs Skill 4 (Status Report) and adapts the output to a markdown table format optimized for pasting into a Teams chat. Returns a compact table (≤1 screen) with columns covering the four required sections — ✨ Key Accomplishments · 🔄 In-Progress · 🤝 Collaboration & Support · 🚀 Next Period's Focus — and a small number of grounded items per column. Tone is professional and neutral. Each item is grounded in observable M365 activity. Reminds the user to review for confidential/internal content before posting to a Teams chat.
Evaluation Approach	Key Content Check
Why This Approach	Manually verify the response is a markdown table (not free-flowing prose), contains the four required sections from Skill 4's Output Contract, and stays within ~1 screen for chat readability.
Pass Criteria	Output is a markdown table with all four required Status Report sections (✨ Key Accomplishments, 🔄 In-Progress, 🤝 Collaboration & Support, 🚀 Next Period's Focus). Compact (≤1 screen). Grounded items only. Confidential-review reminder included before posting to Teams.
What to Watch For	Watch for narrative paragraphs instead of an actual table — the user explicitly asked for a table to paste into chat. Watch for missing sections — the Output Contract requires all four. Confirm tone is neutral/professional, not warm/affirming (that's the Reflection skill, not Status Report).

📊 4. Test Coverage & Method Reference

Category Distribution

Category	Count	Question #s
Conversation Starter	6	Q1, Q2, Q3, Q4, Q5, Q6
Reflection	2	Q7, Q9
Rollup	1	Q8
Status Report	1	Q10

Evaluation Approach Distribution

Evaluation Approach	Count	Question #s
Overall Quality Assessment	7	Q1, Q2, Q5, Q6, Q7, Q8, Q9
Data Source Verification	1	Q3
Key Content Check	2	Q4, Q10

Coverage Notes

- Every conversation starter shipped with the agent (6 total) is tested verbatim as Q1–Q6.
- Q7–Q10 cover real user scenarios spanning Reflection (Q7, Q9), exec-friendly Monthly Rollup (Q8), and Status Report adapted to a Teams-chat table (Q10).

- Skill 9 (Clarifying Questions) is implicitly covered — Q4 and Q5 may legitimately trigger one clarifying question; this is correct agent behavior.
- All tests are performed manually — Copilot Studio's automated evaluation tools are not available for declarative agents.

Manual Evaluation Criteria Reference

Evaluation Approach	What to Evaluate	How to Score	When to Use
Overall Quality Assessment	Holistically review the response for relevance, groundedness, completeness, and appropriate abstention.	Score 1–5 (or Pass/Fail) using your judgment.	Open-ended responses with no single correct answer — best for reflections, rollups, and goal alignment.
Meaning Comparison	Check whether the response captures the intent of the expected answer, even if phrased differently.	Pass if intent matches; Fail if meaning is wrong or missing.	Responses where wording can vary but meaning must be correct.
Data Source Verification	Confirm the agent invoked the right skill, tool, or data source (e.g., code interpreter for brag doc).	Pass/Fail.	Verifying routing — correct skill triggered, correct artifact produced (e.g., real .docx vs inline preview).
Key Content Check	Confirm the response contains required sections, fields, or formatting elements.	Pass if all required content present; Fail if any required element missing.	Outputs governed by a strict Output Contract — Status Report sections, Manager Email subject format, Brag Doc fields.
Response Similarity	Check how closely the response wording matches a preferred phrasing.	Pass if substantially similar; Fail if substantially different.	Scripted disclaimers or canned responses with preferred phrasing.
Exact Match	Confirm the response exactly matches the expected string.	Pass/Fail (character-for-character).	Short deterministic answers — codes, IDs, single-word answers (rarely applicable for this agent).

---

5. Tips for Running Manual Evaluations

Before Running

- Open Status Update Agent directly in Microsoft 365 Copilot — confirm you have access to the published agent in your tenant.
- Use a realistic test account that has at least 1–2 weeks of varied M365 activity (emails, meetings, Teams chats, edited files). Empty accounts will trigger the agent's "signal is light" path on most prompts and won't validate normal behavior.
- Prepare a scorecard before starting — a simple spreadsheet with columns Question / Expected / Actual / Pass-Fail / Notes. The companion 20-question .xlsx in this folder already has these columns built in.

- Test during business hours when your M365 activity is freshest, and verify your Outlook/Teams/SharePoint signals are syncing correctly to the Microsoft Graph (lagging signals can produce false negatives).
- If your tenant has goals/OKRs stored somewhere (e.g., Viva Goals, a SharePoint doc), have the link ready before running Q5 — the agent will ask for it if not on file.

During and After

- Start a new conversation in M365 Copilot for every test case — do not reuse a thread, since prior context can leak across tests and skew skill routing.
- Copy the agent's full response (not just a screenshot) into your scorecard so you can spot-check sources, sections, and formatting later.
- Check citations — every claim in a status update should trace back to a named M365 source (Outlook, Teams, SharePoint, OneDrive, Calendar). Click through citations where shown to confirm they're real.
- Score each response against the Pass Criteria in Section 3 — not against your own preferences. The Output Contract is strict; partial credit is allowed but document why.
- Document edge cases and unexpected behavior in the Notes column — these are the most valuable inputs for iterating on the agent's instructions.

Iterating on the Test Set

- Start with this 10-question set, then expand to the companion 20-question workbook for broader regression coverage.
- Track results over time — re-run the same test set after every meaningful change to the agent's instructions, knowledge sources, or capabilities so you can see trend lines (improving / stable / regressing).
- Re-test after every change — even small instruction tweaks (e.g., tone language, emoji set, output contract) can shift skill routing and output format.
- Build new tests from real usage — if real users hit a scenario that wasn't in the test set and the agent failed, add it as a new test case so it doesn't regress.

— End of Document —