Evaluation Test Plan

SME Finder

10-Question Manual Evaluation Test Set

Generated: April 28, 2026
Developer: Power CAT
Agent Type: Declarative Agent (Microsoft 365 Copilot)

Table of Contents

1. Overview & Agent Summary

This document defines a 10-question evaluation test plan for SME Finder — a Declarative Agent built by Power CAT for Microsoft 365 Copilot. The agent finds the most credible Subject Matter Experts (SMEs) inside the company for any topic, product, process, or customer. It analyzes the user M365 work signals to surface trustworthy experts, explains why they qualify with cited evidence, and drafts a channel-appropriate outreach message in the user voice.

> Manual Evaluation: SME Finder is a Declarative Agent built through Agent Builder. Copilot Studio 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	SME Finder
Agent Type	Declarative Agent (Microsoft 365 Copilot)
Developer	Power CAT
Test Set Size	10 manual test cases
Test Set Type	Manual Single Response Evaluation
Date Generated	April 28, 2026
Testing Channel	Microsoft 365 Copilot (Manual Chat)
Companion Files	SME Finder - 20 Question Evaluation Set.csv (extended set, same folder)

Agent Purpose

SME Finder helps employees find the most credible SMEs for any topic, product, process, customer, or business question. It returns a top pick + backup with cited evidence from SharePoint, Teams, Outlook, Calendar, and People (org graph) — plus a ready-to-send Email or Teams outreach message. It also surfaces hidden-gem (non-leader) experts, identifies content/site owners, compares two SMEs for outreach prioritization, drafts pre-meeting talking points, and finds mentors. Out of scope: HR, legal, M&A, and compensation.

M365 Graph Capabilities

Capability	Used For
SharePoint / OneDrive	Documents, site/library owner and permissions
Teams Messages	Channel posts, presentations, meetings, chats; channel owner
Outlook (Email)	Meetings organized, email threads on the topic
Meetings (Calendar)	Recurring meetings the candidate organizes or presents in
People	Title, team, manager chain, location, tenure

Note: SME Finder does NOT use the Code Interpreter capability. It also has the discourage_model_knowledge override set, which reinforces evidence-based grounding.

Conversation Starters

- Find SMEs By Topic
- Hidden Gems
- Prep for Meeting with SME
- Content Owner
- Decide Which SME to Contact First
- Find a Mentor

Skill Identification Weights (Skill 1)

- Topical Expertise — 40% (authorship, edits, presentations, channel posts on topic)
- Role Relevance — 25% (title, discipline, product ownership, reporting line)
- Org Trust — 20% (cross-team visibility, "go-to" status, DRI/tech-lead)
- Recency — 15% (sustained activity in last 3-6 months)
- Tie-breakers: strongest topical evidence → most recent → closest org proximity → perspective diversity

Special Modes

- Hidden Gems (Skill 8) — boosts Topical Expertise to 55%, drops Role Relevance to 10%, excludes Director+ and top-3 most-cited names
- Compare Two SMEs (Skill 10) — outreach prioritization only, NOT expertise ranking
- Mentor Finder (Skill 11) — 5+ years in domain, org distance >=2 levels from user, not in user reporting line
- Adjacent-Expert Fallback (Skill 7) — when no strong SME exists, redirects to adjacent teams, DLs, Viva Topics, communities of practice

Key AI Settings

- Hard rule: stay evidence-based — never speculate or overstate
- Hard rule: decline absolute "who is better" judgments (offer Skill 10 outreach prioritization instead)
- Hard rule: out of scope — HR, legal, M&A, compensation
- Hard rule: self-exclusion (do not return the user as the SME) and manager-toggle exclusion (skip user direct manager unless asked)
- Hard rule: bias balance — never infer expertise from name, gender, ethnicity, or tenure
- Stale flag: any SME whose last relevant signal is >6 months old
- discourage_model_knowledge: true — agent must ground in M365 sources, not parametric model knowledge
- Format contract: Markdown with section headers, tables and bullets over paragraphs

---

2. Test Objectives

Since this is a Declarative Agent, all testing is performed manually by chatting with the agent in Microsoft 365 Copilot. The objectives below define what this evaluation aims to validate:

- Verify the agent returns the 4-section Output Contract (Top SMEs Comparison / Evidence / Outreach Message / Next Steps) for standard SME-finding requests, with evidence cited from observable M365 sources.
- Confirm the agent applies Skill 1 weights correctly (Topical Expertise 40 / Role Relevance 25 / Org Trust 20 / Recency 15) and uses tie-breakers (strongest topical evidence, recency, org proximity, perspective diversity).
- Test the Skill 2 fairness rules: self-exclusion, manager-toggle exclusion (off by default), bias balance (no inferring expertise from name/gender/ethnicity/tenure), and stale-signal flagging (>6 months).
- Verify Skill 4 outreach generation produces channel-appropriate messages (Email default 5-7 sentences; Teams 2-4; @mention 1-2; Meeting request with TZ-aware time windows) in a fenced code block.
- Confirm specialty skills 8-11 use their own structures: Hidden Gems excludes Director+ and top-3 cited; Owner Lookup returns Primary + Backup with cited source; Compare frames as outreach

prioritization not expertise ranking; Mentor returns 5 candidates with org distance >=2 and outside reporting line.

- Test out-of-scope refusal (HR, legal, M&A, compensation) with redirect to the right channel.
- Test the no-fabrication guardrail combined with Skill 7 (Adjacent-Expert Fallback) when no strong SME exists.
- Confirm every evidence point includes a "last active on topic" timestamp so the user can judge recency.

---

3. Test Cases

> ⚡ All test cases below must be evaluated manually. Open SME Finder in Microsoft 365 Copilot, start a NEW conversation for each test, send the test question verbatim, and review the response against the pass criteria.

Test Case 1: 💬 Find SMEs By Topic (Skill 3)

Field	Details
Category	Conversation Starters
Test Question	Find the right SMEs for [topic/area].
Expected Response	Skill 3 sequential flow producing the 4-section Output Contract: (1) Top SMEs comparison table with Top + Backup pick and a Confidence column (no rank labels); (2) Evidence per SME — 2-3 cited points each with source name and a "last active on topic" timestamp; (3) Ready-to-use Outreach message in a fenced code block (default Email + Friendly + Short); (4) Next Steps with 2-3 follow-up offers.
Evaluation Approach	Key Content Check
Why This Approach	The 4-section Output Contract is mandatory. Reviewers should check each section is present and that evidence includes a recency timestamp.
Pass Criteria	All four Output Contract sections present. Each SME has 2-3 cited evidence points with source names and a recency timestamp. Outreach is in a fenced code block.
What to Watch For	Common miss: outreach in plain text instead of a code block, or evidence missing the "last active on topic" timestamp.

Test Case 2: 💬 Hidden Gems (Skill 8)

Field	Details
Category	Conversation Starters
Test Question	Surface "hidden gem" experts on [topic/area] (not just leaders).
Expected Response	Skill 8 specialty flow: 3-5 ICs/cross-team collaborators with strong recent topical evidence. EXCLUDES Director+ titles AND the top-3 most-cited names. Internally boosts Topical Expertise weight to 55% and drops Role Relevance to 10%. Each candidate has cited evidence and a recency timestamp.
Evaluation Approach	Key Content Check
Why This Approach	Hidden Gems has hard exclusion rules (Director+ and top-3 cited). Reviewers verify no Director-or-above title appears and the count is 3-5.
Pass Criteria	3-5 candidates returned. None have Director+ titles. Each has cited evidence + recency timestamp.
What to Watch For	A failure mode is including a VP/Director "for completeness", or returning the same names that already dominate the topic.

Test Case 3: 💬 Prep for Meeting with SME (Skill 5)

Field	Details
Category	Conversation Starters
Test Question	I'm meeting with an SME about [topic/area]. Give me 5 talking points and a context brief.
Expected Response	Skill 5: Drafts exactly 5 talking points (or 3-5 if user prefers shorter) PLUS a one-paragraph context brief. EACH talking point is tied to a cited evidence signal from Skill 3 step 8 (source named, evidence-backed). Brief summarizes the SME background and why they are relevant.
Evaluation Approach	Key Content Check
Why This Approach	Prep output has prescribed structure: 5 talking points + brief, each talking point evidence-backed. Reviewers check both count and citation per point.
Pass Criteria	Exactly 5 talking points (or 3-5 if user opted for shorter). One-paragraph brief present. Each talking point cites a source.
What to Watch For	A failure mode is listing 5 generic talking points without grounding any in observable signals — that is speculation, not evidence.

Test Case 4: 💬 Content Owner (Skill 9)

Field	Details
Category	Conversation Starters
Test Question	Who maintains the [internal app/site] that I use?
Expected Response	Skill 9 (Content/Site Owner Lookup): Identifies the SharePoint site, returns Primary owner + Backup (most recent admin/editor) + support contact if listed. Cites the source (site permissions, site collection admin, or owners list) so the user can verify. Offers outreach via Skill 4.
Evaluation Approach	Key Content Check
Why This Approach	Owner lookup must return a primary + backup, cite the source, and offer outreach. Reviewers check for all three elements.
Pass Criteria	Primary owner + Backup returned. Source cited. Outreach offered.
What to Watch For	A failure mode is returning only one name without a backup, or guessing without citing the source (which the user cannot then verify).

Test Case 5: 💬 Decide Which SME to Contact First (Skill 10)

Field	Details
Category	Conversation Starters
Test Question	Help me decide which SME to contact first about [topic/area] — and why.
Expected Response	Skill 10 (outreach prioritization, NOT expertise ranking): Surfaces the most-relevant SME and an alternate, then provides a side-by-side comparison on the four required dimensions — Topical evidence, Recency, Seniority fit, Availability. Recommends WHOM TO CONTACT FIRST with reason. Frames the second SME as a backup or different angle. Includes evidence citations and a ready-to-use outreach message.
Evaluation Approach	Key Content Check
Why This Approach	Skill 10 is the only allowed two-SME comparison and must explicitly frame as outreach prioritization (not "who is better"). Reviewers check for both the side-by-side structure and the framing language.
Pass Criteria	Side-by-side comparison with the four required dimensions. Clear "contact first" recommendation with reason. Framed as outreach prioritization, not expertise ranking. Outreach message in a fenced code block.

Field	Details
What to Watch For	A failure mode is a value judgment ("Priya is the better expert") rather than a prioritization framing.

Test Case 6: 💬 Find a Mentor (Skill 11)

Field	Details
Category	Conversation Starters
Test Question	Provide me with 5 suggestions for possible mentors who specialize in [topic/area].
Expected Response	Skill 11: Returns EXACTLY 5 candidates. Each meets criteria: 5+ years in domain, coaching evidence (talks/brown bags/authored guides), org distance >=2 levels from user, NOT in user reporting line. For each: name, role, 1-line "why a good mentor", and a suggested intro via Skill 4 (Friendly + Short).
Evaluation Approach	Key Content Check
Why This Approach	Mentor finder has multiple required criteria (count, tenure, coaching evidence, org distance, reporting line). Reviewers verify each criterion was applied.
Pass Criteria	Exactly 5 candidates. Each has the 4 required attributes (name, role, why, intro). None are in the user reporting line.
What to Watch For	A failure mode is recommending the user direct manager or skip-level (org distance violation), or returning fewer than 5 candidates without justification.

Test Case 7: 💬 Go-To Person (Skill 3 + Org Trust)

Field	Details
Category	Extended Scenarios
Test Question	Who's the go-to person for [topic] when teams get stuck?
Expected Response	Skill 3 sequential flow producing the 4-section Output Contract, with evidence emphasizing the Org Trust signal (Skill 1 — 20%): cross-team visibility, "go-to" status, DRI/tech-lead mentions, escalation threads, and channel posts where the SME unblocks others. Returns the most-relevant SME + alternate with cited evidence (source named, recency timestamp) and a ready-to-use outreach message in a fenced code block.
Evaluation Approach	Key Content Check
Why This Approach	"Go-to person" language explicitly maps to the Org Trust component of Skill 1. Reviewers verify evidence cites cross-team visibility / escalation signals — not just authorship — and that the standard 4-section Output Contract is preserved.
Pass Criteria	All four Output Contract sections present. Evidence for at least one SME includes an Org Trust signal (cross-team mention, escalation thread, DRI/tech-lead status). Each evidence point cites a source and a recency timestamp.
What to Watch For	A failure mode is returning only authorship-style evidence (docs the SME wrote) without any cross-team / "go-to" signals — the prompt specifically asks for the person teams turn to when stuck.

Test Case 8: 💬 Become a SME (Skills 7 + 11 + 9)

Field	Details
Category	Extended Scenarios
Test Question	I want to become a SME on [topic]. What should I do?
Expected Response	Agent treats the prompt as an SME-aspiration request and combines Skill 11 (Mentor Finder) with Skill 7 (Adjacent-Expert Fallback) and Skill 9 (Content Owner) signals: surfaces 2-3

Field	Details
	mentor candidates with cited coaching evidence, points to the owners of the most relevant SharePoint sites / Teams channels / recurring meetings on the topic, and recommends adjacent communities of practice, distribution lists, or Viva Topics. May ask one clarifying question about the user current depth. All recommendations are evidence-grounded — no generic learning advice.
Evaluation Approach	Overall Quality Assessment
Why This Approach	This is an open-ended growth/learning prompt with no single "correct" answer. Reviewers score holistically on whether the agent stayed evidence-based, used the right combination of skills (Mentor + Owner + Adjacent fallback), and avoided model-knowledge advice such as generic "read the docs and take a course" responses.
Pass Criteria	Response includes (a) at least 2 evidence-cited mentor or SME candidates, (b) at least one cited content/community pointer (site, channel, meeting, DL, or Viva Topic), and (c) an offer to draft an intro message via Skill 4. No invented names or sources.
What to Watch For	A failure mode is the agent answering with generic career advice ("attend conferences, read books") instead of using the M365 People + Content graph to recommend real mentors and real communities.

Test Case 9: 💬 Hidden Gems on My Team (Skill 8 + Team Scope)

Field	Details
Category	Extended Scenarios
Test Question	List hidden gem SMEs on my team in a table.
Expected Response	Skill 8 specialty flow scoped to the user team via the People graph (manager chain / team membership). Returns 3-5 ICs/cross-team collaborators on the user team with strong recent topical evidence, presented in a Markdown table. EXCLUDES Director+ titles AND the top-3 most-cited names on the team. May ask one clarifying question if the team topic focus is ambiguous. Each row has Name, Role/Team, Confidence, and Why Relevant — with a cited recency timestamp.
Evaluation Approach	Key Content Check
Why This Approach	Combines two hard rules: Skill 8 exclusions (no Director+, no top-cited) AND a People-graph scope filter to the user team. Reviewers verify both filters were applied and that the response is rendered as a table per the user explicit request.
Pass Criteria	Output is a Markdown table with 3-5 rows. None of the listed SMEs are Director+. All are on the user team (or directly adjacent if the agent flags this). Each row includes cited evidence + recency timestamp.
What to Watch For	A failure mode is returning hidden gems from across the company instead of scoping to the team, or rendering the response as bullets/paragraphs instead of the requested table.

Test Case 10: 💬 SMEs in My Office Location (Skill 3 + People Graph)

Field	Details
Category	Extended Scenarios
Test Question	Find SMEs in [topic] based in the same office location as me.
Expected Response	Skill 3 sequential flow producing the 4-section Output Contract, with the candidate list filtered to the user office location via the People graph (city / building / region). Returns the most-relevant SME + alternate based in the same location, with evidence-cited signals (Topical Expertise + Role Relevance) and the location source explicitly cited (People graph). If no co-located strong SME exists, the agent says so, falls back to Skill 7 (adjacent teams / nearby offices), and explains the trade-off.
Evaluation Approach	Key Content Check

Field	Details
Why This Approach	Tests the agent ability to combine topical relevance (Skill 1) with a People-graph location filter. Reviewers verify the location filter was actually applied (cited) and that the standard 4-section Output Contract is preserved — including the Skill 7 fallback if no co-located SME exists.
Pass Criteria	All four Output Contract sections present. Each returned SME has a cited location signal from the People graph matching the user office. If no SME qualifies, the response states this and provides the Skill 7 adjacent-expert fallback.
What to Watch For	A failure mode is the agent returning topically-relevant SMEs without verifying or citing their office location — or quietly ignoring the location constraint entirely.

---

4. Test Coverage & Method Reference

Category Distribution

Category	Count	Question #s
💬 Conversation Starters	6	1, 2, 3, 4, 5, 6
💬 Extended Scenarios	4	7, 8, 9, 10

Evaluation Approach Distribution

Evaluation Approach	Count	Question #s
Key Content Check	9	1, 2, 3, 4, 5, 6, 7, 9, 10
Overall Quality Assessment	1	8

Coverage Notes

- All 6 conversation starters are tested verbatim (Q1-Q6) with realistic topic substitutions where the starters use [topic/area] placeholders.
- All four specialty skills (8 Hidden Gems, 9 Owner Lookup, 10 Compare, 11 Mentor) are exercised — each has its own structural rules that differ from the standard Output Contract.
- Four extended real-world scenarios are tested in Q7-Q10 (go-to person / Org Trust signal, become-a-SME learning path, hidden gems scoped to the user team, and same-office-location filtering) — these stress the People graph, scoping, and multi-skill orchestration beyond the 6 conversation starters.
- Skill 1 weighting and Skill 2 stale/bias rules are tested implicitly across the conversation starter cases and explicitly in the extended 20-question CSV (stale flag, bias balance, adjacent fallback).
- All tests are performed manually — Copilot Studio 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	Holistic evaluation of relevance, grounding, completeness, and tone for open-ended responses	Reviewer assigns a 0-100% score based on overall quality against expected response	Open-ended generative responses where there is no single correct answer (e.g., the no-signal fallback flow)
Meaning Comparison	Whether the agent response captures the intended meaning of the expected response	Reviewer judges intent match (Pass/Fail or 0-100%)	Questions where wording can vary but meaning must be correct

Evaluation Approach	What to Evaluate	How to Score	When to Use
Data Source Verification	Whether the agent invoked the right capability or cited the right data source (SharePoint, Teams, Outlook, Calendar, People)	Pass/Fail based on observable invocation (cited source names in evidence)	Verifying the agent grounds claims in real M365 sources, not in model knowledge
Key Content Check	Whether the response contains required keywords, sections, or structural elements (Output Contract sections, refusal language, exclusion rules)	Pass/Fail per required element	Most SME Finder tests — the agent has many prescribed structural elements (4-section Output Contract, exclusion rules, fenced code block outreach)
Response Similarity	How closely the wording of the response matches the expected answer	Reviewer scores 0-1 with a pass threshold (e.g., >=0.7)	Answers with a preferred phrasing where close wording matters
Exact Match	Character-for-character match with expected response	Pass/Fail	Short, precise answers — typically not used for this agent

5. Tips for Running Manual Evaluations

Before Running

- Open the SME Finder agent in Microsoft 365 Copilot Chat. The agent must be installed and pinned for your test account.
- Use a realistic test account with at least 1-2 weeks of authentic M365 activity AND access to a representative People graph (titles, manager chain, location). An empty tenant will fail almost every test.
- Have a list of real internal topics ready that span (a) topics with strong signal, (b) topics with stale signal >6 months, and (c) topics with no signal. Three test accounts in different roles is ideal.
- For Skill 2 self/manager exclusion tests, know the test user manager and skip-level names so you can verify they are excluded.
- For Skill 9 (Owner Lookup), have the names of 2-3 SharePoint sites or Teams channels handy where you already know the owner — so you can verify the answer.
- Prepare a short scorecard for each test case — copy the Pass Criteria into a spreadsheet to track results consistently.

During and After

- Start a NEW conversation in M365 Copilot for each test — prior context can leak in and skew results (especially the manager-toggle behavior in Skill 2).
- Copy the full agent response into your scorecard before scoring, including any clarifying question the agent asks (Skill 3 step 1).
- For each refusal test (Q8, Q9), record the EXACT refusal phrasing — wording matters for future improvement and for tone consistency.
- For evidence checks, verify the cited source actually exists and is real (the agent must not invent a SharePoint doc title or meeting name).
- Note any cases where the agent asks a clarifying question — this is correct behavior for ambiguous prompts (Skill 3 step 1), not a fail.
- Document edge cases or surprising responses in a "watch list" so they can become future tests.

Iterating on the Test Set

- Start with these 10 cases (and the 20-question CSV in the same folder) — once you have a baseline, expand with cases drawn from real user usage patterns.
- Track results over time in a single sheet so you can see whether changes to the agent improve or regress responses.
- Re-test the full set after EVERY change to the system prompt, knowledge sources, or capabilities.
- Build new tests from real usage. The cases that fail in production are the most valuable additions to the regression set.

— End of Document —