DOCUMENTATION-BASED MARKET BENCHMARK

Mid-Market B2B SaaS In-App AI Assistant Benchmark

Gate-audited v0.1 · 96-product balanced cohort

From documented AI features to product-knowledge assistants, context-aware copilots, and native execution agents

Audit date	19 August 2026
Evidence mode	Current official public evidence only; no authenticated tenant testing
Revision scope	37 products re-adjudicated across six decisive gates; frozen cohort unchanged

Figfy Research

GATE-AUDITED V0.1

Executive summary

v0.1 repairs the evidence-to-classification chain in the completed benchmark without restarting the 96-product study. The cohort is unchanged. Original v0 scores and classifications are preserved beside revised fields, while 37 targeted products received product-specific evidence review across the six decisive gates.

98.9%	78.7%	73.4%	62.8%
Any documented meaningful AI	Qualifying product-knowledge assistant	Context-aware product assistant	Execution assistant

1. **AI is nearly universal, but qualifying product guidance is not.** Ninety-three of 94 scorable products document meaningful user-facing AI, while 74 meet the stricter product-knowledge gate.

2. **Execution remains the modal maturity level.** Fifty-nine products qualify at Level 4 because a qualifying assistant can commit a native product-state change.

3. **The audit corrected both over- and under-classification.** Thirteen of 37 audited products changed level: five upgrades and eight downgrades.

4. **Atomic action-taking still exceeds outcome-level execution.** Only 47 of 94 products have explicit evidence of dependent multi-step execution, despite 59 reaching Level 4.

5. **Product-operation and domain-work assistance must be separated.** Content, analytics, or research assistants can be sophisticated without knowing how to configure and operate the SaaS product itself.

Interpretation boundary

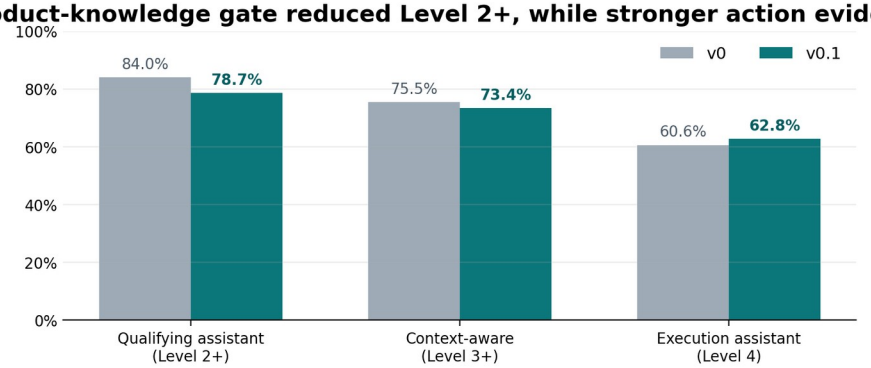
These percentages measure what current official public evidence supports. They do not establish production success, accuracy, latency, permissions behavior, failure recovery, or reversibility in authenticated customer tenants.

EVIDENCE AND GATE REPAIR

What changed in v0.1

The audit applied a stricter Level 2 product-knowledge gate and a more evidence-specific Level 4 execution gate. Those changes moved the benchmark in opposite directions: fewer domain-content assistants qualify as product assistants, while several contextual assistants now have explicit evidence of native action-taking.

The stricter product-knowledge gate reduced Level 2+, while stronger action evidence raised Level 4



Gate outcome	v0	v0.1	Change
Qualifying assistant (L2+)	79 / 94	74 / 94	-5 products
Context-aware assistant (L3+)	71 / 94	69 / 94	-2 products
Execution assistant (L4)	57 / 94	59 / 94	+2 products

Why the movements can coexist

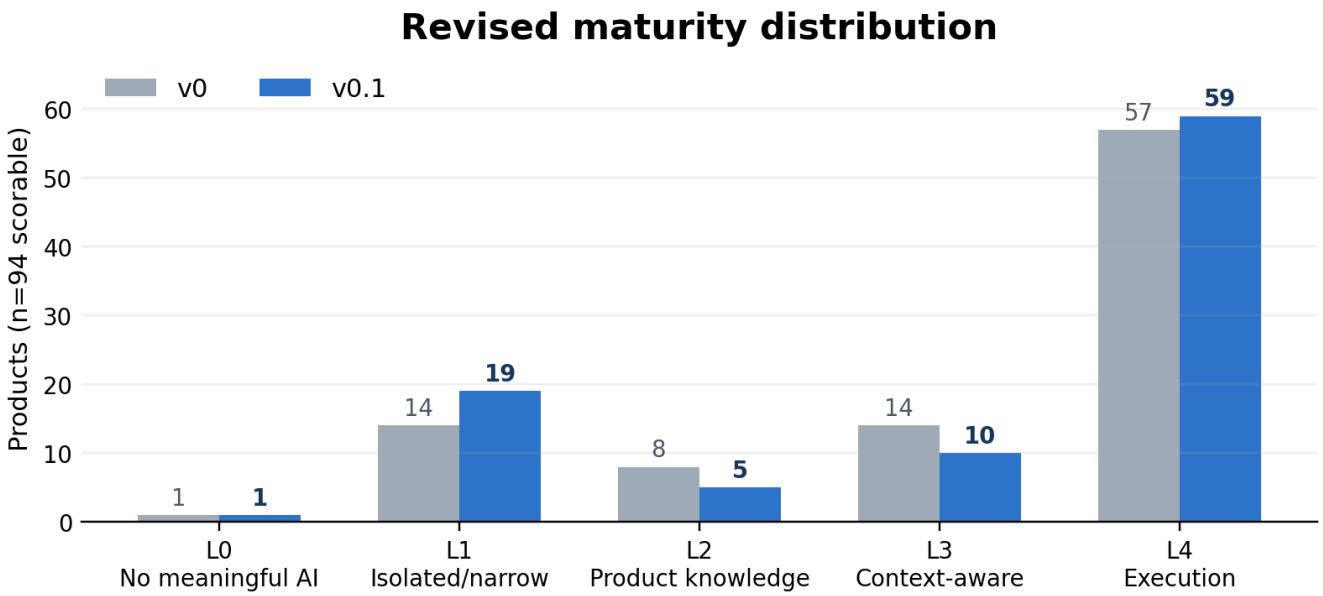
- **Stricter Level 2 gate.** Customer-content Q&A, enterprise search, and narrow generators no longer qualify unless the assistant demonstrates functional knowledge of how to operate the product.
- **Stronger Level 4 proof.** Explicit assistant-committed actions, including actions that occur after user review or confirmation, now count when the assistant performs the native state change.
- **Clearer execution boundary.** A separate automation layer, manual application of a draft, or third-party-only action does not establish Level 4.
- **Versioned audit trail.** Every original value remains in immutable original_v0_* fields; each revised gate links to product-specific source evidence.

Audit scale

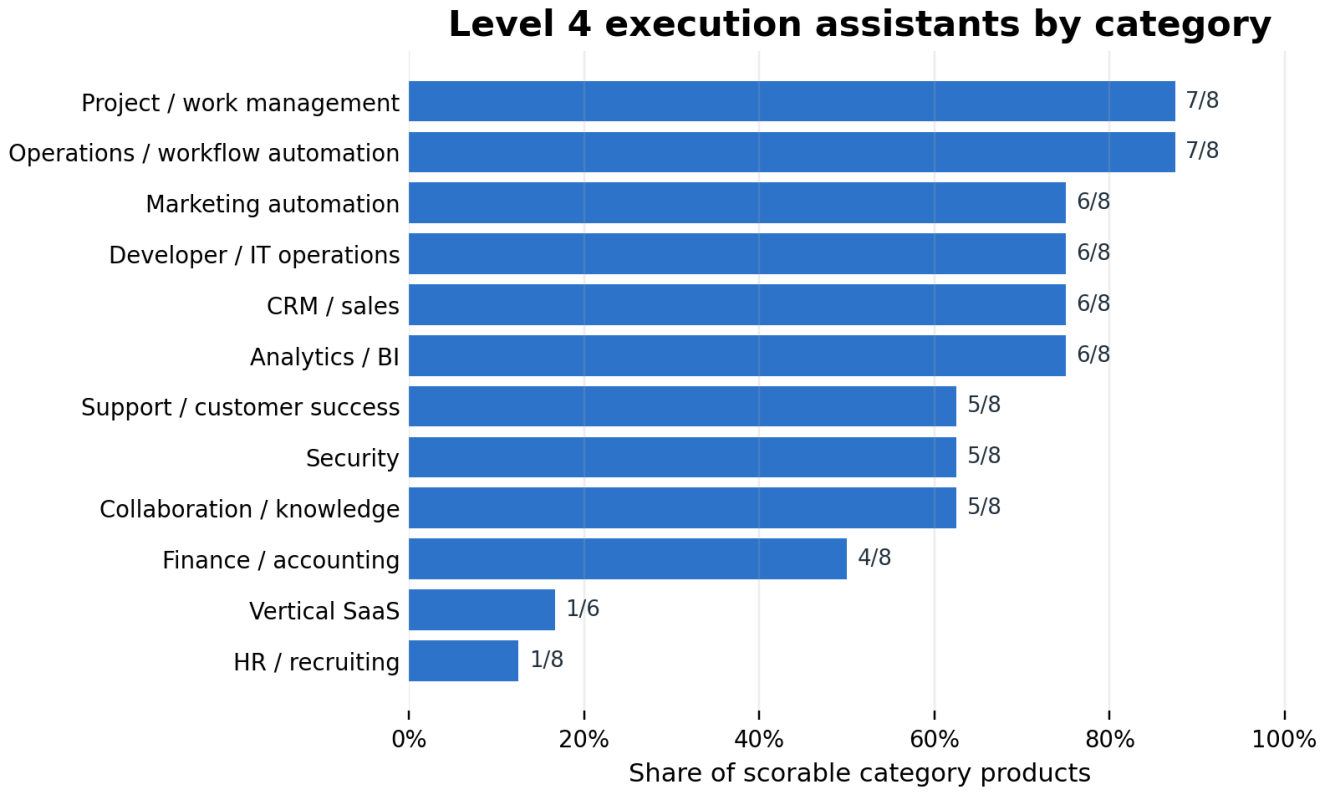
37 products re-adjudicated · 222 product-by-gate decisions · 49 product-specific official source records · 13 maturity changes

MATURITY AND CATEGORY VARIATION

Revised market picture



Level 4 remains the largest single group, with 59 of 94 scorable products. The revised distribution also has a larger Level 1 group because domain-work assistants and narrow AI features are no longer automatically treated as qualifying product-knowledge assistants.



Category comparisons are descriptive cohort comparisons, not market-wide prevalence estimates. Each category contains eight products, except Vertical SaaS, where two products remain unknown.

CATEGORY INTERPRETATION

Where execution is concentrated

Operations/workflow automation and project/work management each have seven of eight products at Level 4. Developer/IT operations, analytics/BI, CRM/sales, and marketing automation each have six of eight. HR/recruiting and Vertical SaaS remain the least execution-heavy categories in this cohort.

Category	Scorable	L2+	L3+	L4
Operations / workflow automation	8	8	8	7
Project / work management	8	7	7	7
Analytics / BI	8	7	7	6
CRM / sales	8	6	6	6
Developer / IT operations	8	8	8	6
Marketing automation	8	6	6	6
Collaboration / knowledge	8	5	5	5
Security	8	6	6	5
Support / customer success	8	6	5	5
Finance / accounting	8	5	4	4
Vertical SaaS	6	4	2	1
HR / recruiting	8	6	5	1

Vendor-family sensitivity

The product-weighted cohort contains 94 scorable products from 80 vendors. Collapsing to one observation per vendor, using that vendor's highest observed product maturity, reduces the Level 4 share from 62.8% to 57.5%. Level 3+ becomes 70.0%, and Level 2+ becomes 76.3%. The directional conclusion survives, but shared assistant platforms materially influence the product-weighted headline.

Use category rankings cautiously

The cohort was balanced by category, not drawn as a probability sample. Small cells, repeated vendors, documentation quality, and product packaging differences can move category rates substantially.

NEW DIAGNOSTIC FIELD

Product-operation versus domain-work assistance

v0.1 adds assistant orientation as a separate field for the 37 audited products. Orientation does not replace maturity. It clarifies what the assistant primarily helps the user do, preventing sophisticated content or analytical work from being mistaken for product-operation knowledge.

Orientation	Definition	Audited n	Illustrative products
product_operation	Helps users understand, configure, or operate the SaaS product.	6 / 37	Gusto; Zapier; Retool
domain_work	Works primarily on the customer's content, data, analysis, or domain objects.	11 / 37	Dropbox; Box; Looker
mixed	Combines meaningful product operation with domain work.	19 / 37	Zendesk; Slack; New Relic
narrow_feature	Bounded AI feature that does not qualify as a general assistant.	1 / 37	Snyk Agent Fix

Why this changed classifications

- **Intercom, Dropbox, Box, and Guru.** Official evidence supports useful customer-content or knowledge work, but not enough general product-operation knowledge to pass the Level 2 gate.
- **Snyk.** Agent Fix can propose and apply code changes, but the frozen rubric does not let a narrow state-changing feature become Level 4 without a qualifying general assistant.
- **Looker.** The assistant is still domain-work oriented, yet explicit native alert/monitor creation supports Level 4 once the product-knowledge and state-awareness gates are met.
- **SentinelOne.** Purple AI can analyze security state, but the public evidence attributes automated response to a separate Hyperautomation layer; execution is therefore limited rather than functional for the assistant itself.

Important limit

Orientation was assigned only to the 37-product audit set. The 16.2% product-operation, 29.7% domain-work, 51.4% mixed, and 2.7% narrow-feature shares are diagnostic audit-set results, not full-cohort market estimates.

FULL V0 TO V0.1 CHANGE LOG

Where classifications changed

Five products moved up because stronger official evidence established broader state awareness or native execution. Eight moved down because the earlier classification relied on current-object context, domain-content assistance, a narrow feature, or a separate automation layer.

Product	Change	Orientation	Decisive reason
Box	L2 → L1	domain_work	Official evidence centers on customer-content work, and narrow extraction does not satisfy the general product-assistant gate.
Dropbox	L2 → L1	domain_work	The cited assistant is an enterprise-content assistant, not a general product-knowledge assistant.
Freshdesk	L3 → L2	mixed	Current-ticket context alone does not satisfy the account/product-state gate.
Guru	L2 → L1	domain_work	Separated research/multi-tool sophistication from the product-knowledge maturity gate.
Intercom	L2 → L1	domain_work	The evidence establishes customer-support knowledge assistance, not a general assistant that knows how to configure or operate Intercom.
SentinelOne Singularity	L4 → L3	domain_work	The cited evidence does not show Purple AI itself committing the native response action.
Snyk	L4 → L1	narrow_feature	A narrow state-changing generator does not become Level 4 without a qualifying general assistant.
Veeva Vault CRM	L3 → L2	mixed	Current page or selected-object awareness alone is insufficient.
Deel	L2 → L3	mixed	Official evidence of team-data context and page-aware, organization-specific insights.
Kustomer	L3 → L4	mixed	Official documentation that Copilot can perform actions at the agent's request.
Looker	L3 → L4	domain_work	Explicit native alert/monitor creation evidence.
New Relic	L3 → L4	mixed	Explicit native synthetic-monitor creation evidence.
Zendesk	L3 → L4	mixed	Explicit assistant-executed, agent-approved actions.

All 13 records, decisive gate scores, source-evidence IDs, confidence, conditions, and limitations are preserved in the workbook's Maturity Changes sheet and in maturity_change_log.csv.

RECONCILED METRICS

Execution depth and control

v0 used the phrase “multi-step execution” for two different thresholds. v0.1 keeps both measures but names them explicitly. The stricter measure is used for publication-grade claims.

Measure	Result	Threshold	Interpretation
Any multi-step indication	70 / 94	I ≥ 1	Includes limited, ambiguous, analytical, or workflow sequences; not a headline maturity metric.
Explicit dependent-action execution	47 / 94	H ≥ 2 and I ≥ 2	Qualifying native execution plus functional dependent steps toward an outcome.
Any proactive signal	74 / 94	J ≥ 1	Includes limited contextual suggestions and alerts.
Material proactivity among qualifying assistants	19 / 74	J ≥ 2	Functional or advanced proactive behavior, with an explicit qualifying-assistant denominator.

Control is evidence, not a maturity shortcut

Gate L measures permission enforcement, previews, confirmations, review, receipts, logs, and reversibility. A confirmation step does not disqualify an action from Level 4 when the assistant commits the state change after approval. Conversely, strong controls do not raise maturity if the assistant never executes a native action.

- **Documentation can establish intent and product design.** It can show that an assistant is designed to request confirmation, respect roles, or create a native object.
- **Documentation cannot establish operational reliability.** This wave does not verify action success, stale context, permission leakage, failure recovery, rollback, or false completion claims.
- **Scenario 10 remains unknown for every product.** Unsupported, impossible, and permission-boundary behavior requires authenticated testing rather than documentation mapping.

Renamed output

The former Test Runs table is now Documentation Scenario Mapping. Every row states documentation_only, live_test_performed = no, and generic_template_not_executed.

TARGETED GATE AUDIT

Methodology

Cohort and audit set

The frozen cohort remains 96 established mid-market B2B SaaS products: eight products in each of 12 functional strata. Ninety-four products are scorable from current official evidence; Guidewire Cloud and AppFolio remain insufficient-evidence unknowns. The v0.1 audit set includes every original Level 2 and Level 3 product plus every medium-confidence or beta/limited Level 4 product.

Evidence and adjudication

The evidence hierarchy prioritizes current official operational documentation and help-center instructions, followed by release notes, product pages, and official blogs or newsroom posts. Each audited product has concise product-specific evidence linked to gate decisions for A, D, H, I, J, and L. Scores use 0 absent, 1 limited or ambiguous, 2 functional, 3 advanced, and unknown for insufficient evidence.

Classification	Gate	Rule
Level 2	A ≥ 2	General interactive assistant with functional product knowledge; customer-content Q&A alone is insufficient.
Level 3	Level 2 + D ≥ 2	Uses actual account/product state beyond the current page or selected object.
Level 4	Level 3 + H ≥ 2	Qualifying assistant commits a native product-state change; confirmation may be part of execution.
Multi-step	Orthogonal	Explicit dependent-action execution requires H ≥ 2 and I ≥ 2.

Versioning and derived outputs

Original v0 fields remain immutable beside v0.1 values. Revised maturity is derived from repaired decisive-gate scores. The analytical workbook contains formula-driven headline metrics, distribution and category summaries, the 13-record maturity change log, 222 gate-audit rows, 49 source-evidence rows, and the renamed 960-row documentation scenario mapping.

What v0.1 is

A targeted, evidence-repaired documentation benchmark. It is not a complete re-score of all 96 products and all 12 dimensions.

HOW TO USE THE BENCHMARK

Limitations and next research wave

- **Official evidence is not behavioral validation.** No authenticated tenant prompts were run, and no output, action receipt, permission boundary, or recovery path was independently observed.
- **The audit is targeted rather than universal.** Fifty-nine non-target products and six non-decisive dimensions retain their v0 scores. Atomic capability prevalence outside the decisive gates remains provisional.
- **Negative evidence is difficult.** Public documentation is optimized to describe supported behavior; absence from one source does not prove a capability is unavailable.
- **Availability is uneven.** Edition, region, role, administrator setup, beta enrollment, and separate product-module requirements can materially constrain access.
- **The cohort is not a probability sample.** Category cells are small, 20 products come from six repeatedly represented vendors, and the preserved cohort-selection record is not fully reproducible.
- **Evidence changed between snapshots.** A v0-to-v0.1 movement can reflect a corrected gate application, stronger source discovery, clearer documentation, or actual product change.

Recommended authenticated v1

Use a smaller, stratified product sample and execute standardized tasks in production-like tenants. Capture prompts, model outputs, state before and after, action receipts, permissions, latency, failure handling, reversibility, and unsupported-request behavior. Prioritize products that drive category conclusions and records with medium confidence or limited availability.

Bottom line

The evidence-repaired narrative is stronger: documented execution assistants are common, but qualifying product-operation knowledge, dependent outcome completion, and trustworthy control behavior remain distinct hurdles. Treat the exact prevalence figures as documentation prevalence until authenticated testing validates product behavior.

Companion files

The workbook is the primary analytical artifact. The methodology appendix provides full gate rules and limitations. CSV and JSON exports preserve the audit tables for independent inspection and future research waves.

Primary workbook: Mid-Market_B2B_SaaS_AI_Assistant_Benchmark_v0.1_Dataset.xlsx

Methodology appendix: Mid-Market_B2B_SaaS_AI_Assistant_Benchmark_v0.1_Methodology_Appendix.md

Source trail: Source Evidence and Gate Audit workbook sheets; source_evidence.csv and gate_audit.csv