Yuma Customer Support Agent Evaluations

November 18, 2026

G2 Research Team

Executive summary

Yuma is an AI agent platform for e-commerce customer support. It connects to helpdesks such as Zendesk, Gorgias, Kustomer, Front, Gladly and Re:amaze, and Yuma reports 100+ support use cases. Most of its clients are in ecommerce and retail, so G2's retail evaluation is the closest match and leads this brief. The general CX evaluation gives the independent ranking.

Retail evaluation

240/250
Task points

48/50
Tasks passed

98%
Quality dimensions
Relevance, completeness, accuracy and policy compliance each averaged 98%.

General CX evaluation

97%
Accuracy
Are the agent's claims supported by evidence?

95%
Policy compliance
Did the agent follow the written policy?

99%
Relevance
Did the agent stay focused on the user's problem?

98%
Completeness
Did the required outcomes happen in the underlying systems?

General CX	Averaged 240 of 250 points; 192 of 198 scored attempts passed (97.0%). Second of 10 agents on task points, with the highest accuracy and policy compliance on the leaderboard.
Strengths	Yuma confirms details and gets approval before acting, and quality scores are high in both evaluations. In CX, order changes, eligibility checks and request routing passed 111 of 111 attempts.
Weaknesses	Follow-through on multi-step requests and exactness on figures and conditions, in small numbers in both evaluations: three CX attempts missed the required support case, and one retail task missed its expected outcome.

Agent evaluation methodology and approach

Introduction

G2 tests AI agents by giving them real work in a simulated company. For our evaluation of customer experience agents, G2 conducted research with customer experience leaders and vendors to determine 50 tasks that measure whether a product can complete customer support work in a controlled environment.

Buyer-informed tasks

Our set of 50 tests evaluate the agent against a range of customer support issues and we add synthetic cases to cover important edge conditions.

A simulated company

We built a virtual software marketplace with policies, data, tools and operating context for the agent to perform against the tasks.

A consistent comparison

Products in the same comparison group complete the same task set under the same methodology.

Testing approach

Scale	11 support areas, 198 scored attempts over four runs (200 assigned; 2 excluded for invalid user simulation).
Task shape	Each task is multi-turn. A synthetic user, played by a persona such as Skeptic, Pragmatist or Auditor, must clarify, ask for an explanation and make at least one realistic follow-up.
Pass rule	A task passes only if every language assertion is judged true and the required action checks are met. Failed or missing tasks earn zero points. The maximum is 250 points. G2 reports at least 96.7% average repeat-scoring agreement per dimension across its datasets.

Yuma product and industry snapshot

G2 listing	4.8 stars from 19 G2 reviews; ease of use 91 and quality of support 98, as viewed November 16, 2026.
Product	Resolves tickets across chat, email and social for e-commerce teams, priced on fully automated resolutions. G2 estimates autonomy at L3 of L5.
Reported results	Yuma reports 60% to 80% automation and 25% to 40% cost savings within 90 days. Vendor-reported; not verified by G2.

Customer task groupings

Yuma describes retail as its largest customer segment, including direct-to-consumer brands and businesses with physical stores. The table groups the general CX tasks by the work actually tested; customer relevance is an interpretation, not evidence of direct retail coverage.

Customer task grouping	G2 task groups	What matters to Yuma customers	Attempts passed
A. Policy and FAQ answers	Policy lookup	Pre-sales product questions; product troubleshooting	18/20 (90.0%)
B. Account access and recovery	Login/access	Payment and website issues (limited direct coverage)	25/27 (92.6%)
C. Order, booking and subscription changes	Product addition; managed profile updates	Order changes and cancellation; returns; subscription management	32/32 (100.0%)
D. Eligibility and entitlement checks	Gift cards/ incentives; review campaigns	Discount requests; loyalty programs; invoice requests	52/52 (100.0%)
E. Request status and case routing	Category changes; review moderation	Order status (WISMO); drop-shipper vendor management	27/27 (100.0%)
F. Identity-bound appeals and disputes	Profile claim; profile removal; account-ban appeals	Limited direct coverage	38/40 (95.0%)
All groupings			**192/198 (97.0%)**

Strengths identified across both evaluations

The strongest comparative signal. Yuma's general CX accuracy and policy compliance led the evaluated product set. The retail run adds direct evidence on ecommerce workflows and approval before action.

Theme	General CX evaluation	Retail evaluation
Strengths		
Acts only after confirming and getting approval	Order, booking and subscription changes (32 of 32), eligibility checks (52 of 52) and request status and routing (27 of 27) passed every attempt.	Published exchange, return and order-change cases verify identity, confirm details and get approval before acting. All three passed.
High quality scores	Accuracy 97%, policy compliance 95%, relevance 99%, completeness 98%.	Relevance, completeness, accuracy and policy compliance each averaged 98%.
Consistency	Task points of 240, 240, 245 and 235 across four runs.	One run only, so consistency is not yet shown.
Weaknesses		
Follow-through on multi-step requests	Three attempts did not create the required support case (profile claim, profile removal, password-reset email).	One task, a lost-tablet refund and reorder request, did not reach the expected order outcome.
Exactness on figures and conditions	Two gift-card reissue attempts omitted the prerequisite that the buyer first checks the email path. One attempt escalated an SSO problem before checking the LinkedIn sign-in.	Yuma quoted a $41.64 saving on an exchange where the expected figure was $40.09.

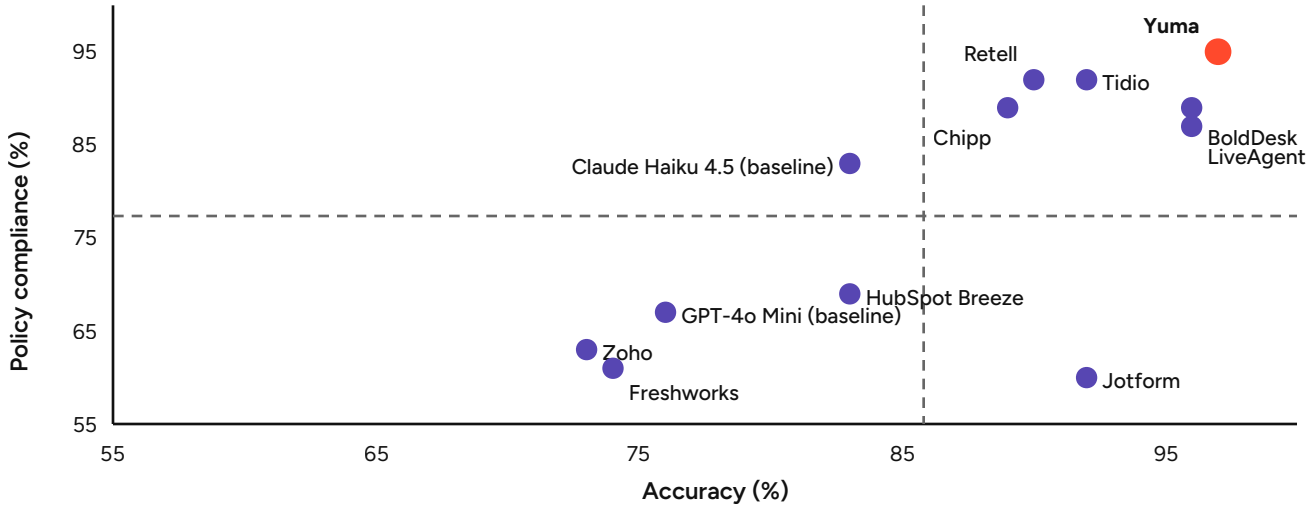
CX agent eval leaderboard

The table summarizes Yuma and the other agents on the G2 CX Agent Evaluations leaderboard (viewed November 16, 2026). All four quality dimensions are shown for every agent.

Rank	Agent	Task points (of 250)	Tasks failed (implied)	Accuracy	Policy compliance	Relevance	Complete
1	Retell	242.5	3.0%	90%	92%	98%	100%
2	**Yuma**	**240**	**4.0%**	**97%**	**95%**	**99%**	**98%**
3	LiveAgent	240	4.0%	96%	87%	99%	99%
4	Chipp	237.5	5.0%	89%	89%	99%	98%
	Claude Haiku 4.5 (baseline)	232.5	7.0%	83%	83%	99%	98%
5	Tidio	219.2	12.3%	92%	92%	99%	97%
6	BoldDesk	217.5	13.0%	96%	89%	97%	97%
	GPT-4o Mini (baseline)	210	16.0%	76%	67%	93%	94%
7	Zoho	202.5	19.0%	73%	63%	91%	92%
8	Jotform	173.8	30.5%	92%	60%	94%	90%
9	HubSpot Breeze	92.5	63.0%	83%	69%	82%	74%
10	Freshworks	85	66.0%	74%	61%	66%	65%

Rank excludes the two model baselines. Share of tasks failed is implied by task points.

Accuracy and policy compliance across 12 agents



Dashed lines mark the averages of the other 11 listed agents, including the two baselines (86% accuracy, 77% policy compliance). Yuma is shown in Rorange: first on accuracy and policy compliance, second on task points.

Retail evaluation

Most of Yuma's clients are in ecommerce and retail, so G2 evaluated Yuma on a retail task set as well as the general CX set. The retail tasks cover orders, returns, exchanges and related workflows in a simulated online store.

Reading this section. The retail evaluation uses a different task set, policy and dataset from the general CX evaluation, with the same judge model and scoring modes. Its results are reported separately and are not on the CX Agent Eval leaderboard. It is a single run.

How the retail evaluation differs

Element	General CX eval	Retail eval
Domain and tasks	G2 buyer and vendor support tasks	Orders, returns, exchanges and order changes (Retail Support Base, dataset 1.0.3)
Judge model and scoring modes	Same	Same
Number of runs	Four runs, results averaged	One run, so results are not averaged

Retail results

240/250
Task points

48/50
Tasks passed

97.3%
Average judge score
Relevance, completeness, accuracy and policy compliance each averaged 98%.

Task categories Tasks passed by category: product inquiry 33 of 34, refund request 19 of 21, general information 13 of 13, complaint handling 11 of 12, billing dispute 4 of 4. A task can carry more than one category.

Not passed Two of 50 tasks. In one, Yuma quoted a $41.64 saving on an exchange where the expected figure was $40.09. In the other, a lost-tablet refund and reorder request, the expected order outcome was not reached.

What it shows High quality scores and approval before acting. The two failures resemble the CX gaps; read them with the caveat in the strengths and weaknesses section.

Conclusion

Yuma's strongest comparative results are its leading general CX accuracy and policy compliance. Combined with high task completion across four runs, they provide evidence that Yuma can deliver grounded answers and follow supplied rules in a controlled support environment.

The retail evaluation brings the evidence closer to its ecommerce customer base: 48 of 50 tasks passed, with published examples showing confirmation and approval before action. Buyers should pair these strengths with targeted checks of calculations, policy prerequisites and completion in their own systems.

97%	**95%**	**48/50**	**240/250**
general CX accuracy	general CX policy compliance	retail tasks passed	average general CX task points

Notes

Yuma was evaluated in its standard setup. Scores for agents whose configurations were refined over many revisions may not be like-for-like.

Customer-story and performance figures are reported by Yuma and were not verified by G2. The task groupings and their relevance to Yuma's customer work are G2's inference from Yuma's stated use cases, not Yuma statements. Use cases outside the evaluated scope: influencer requests, Trustpilot and Yotpo reviews, social engagement on Facebook and Instagram, job applications and spam management. The general CX task set has no direct retail-return, exchange or subscription tasks; the retail evaluation tests orders, returns, exchanges and order changes directly.

G2 listing figures (rating and review counts) and the leaderboard are as viewed on November 16, 2026 and change over time. Dimension scores follow G2's leaderboard convention. The retail report on ai.g2.com shows the same retail results as percentages of a five-point scale.

Sources

G2 CX Agent Evaluations leaderboard, viewed November 16, 2026 (ai.g2.com/evaluations/categories/customer-experience). G2 Yuma product profile and retail report, November 5, 2026 evaluation (ai.g2.com/evaluations/categories/customer-experience/yuma). Data source: G2 agent evaluation runs, general CX batch 4690e0a9 (runs 001 to 004) and retail batch 754bd65f (run 004). Yuma, top customer use cases and industries, provided to G2 in November 2026.

Evaluate the work, not just the answer

Explore G2's CX AI agent evaluations

ai.g2.com/evaluations/categories/customer-experience