CONFESSIONS OF AN AI CHATBOT

A whitepaper on the state of customer service bots in Malaysia

Entermind

TABLE OF CONFESSIONS

"

MOST CHATBOTS IN MALAYSIA ARE A GLORIFIED FAQ SEARCH BAR. THAT IS NOT A CHATBOT... THAT'S A SEARCH ENGINE WITH EXTRA STEPS.

- Telco evaluator

What separates a good chatbot from a useless one? Capability investment and design. Some of us were built to solve problems. Others were built to *look* like we solve problems.

The Entermind Chatbot Quality Index (CQI) is a proprietary framework for the systematic testing and quality assessment of customer service chatbots. It tested 24 chatbots across eCommerce & Wallet (referred to as 'eCommerce' henceforth), Travel, Telecom, and Financial Services.

Each chatbot was evaluated against 26 standardised binary tests grouped into five weighted categories: Comprehension, Access, Experience, Functional Capability, and Safeguards, by trained evaluators using sector-specific scripts. Full methodology is provided in the Appendix.

HALF OF US CAN'T DO OUR JOBS PROPERLY

Confession

01

Table 1 presents the complete Entermind CQI rankings. The top tier, comprising seven chatbots above 70%, spans all sectors, demonstrating that strong chatbot performance follows the right areas of investment, not industry.

The middle tier (40–70%) shows pockets of capability but inconsistent execution across categories.

Below 40%, the picture splits: some chatbots are early-stage implementations with genuine capability gaps, while others are deliberate triage designs built to route rather than resolve.

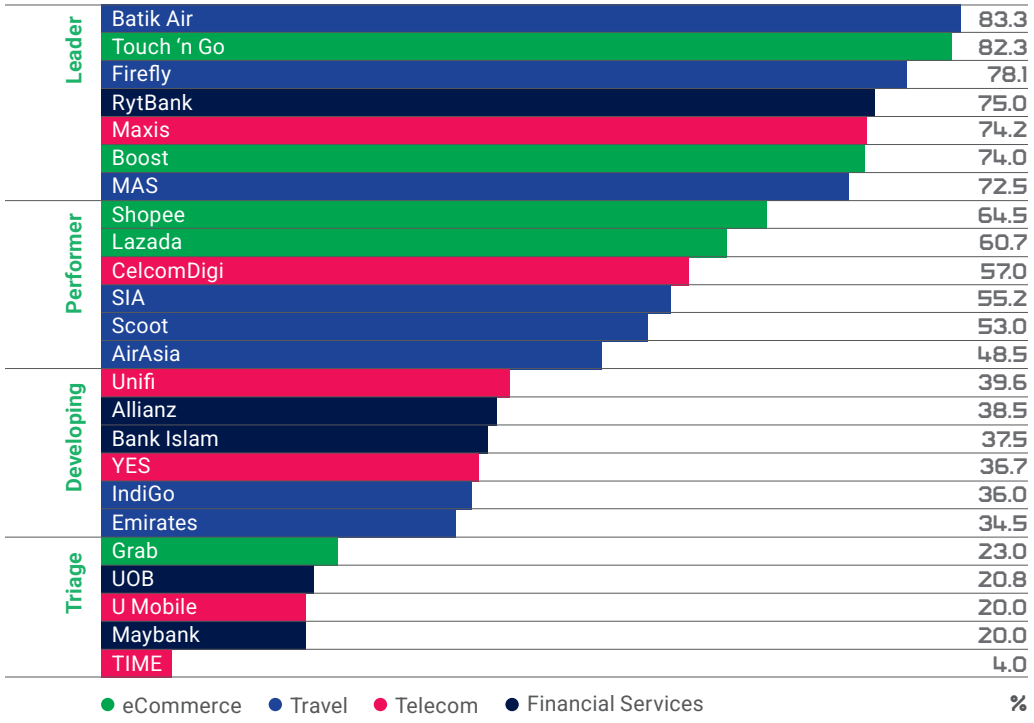
The average score	**49.5%**
The best	**83.3%**
The worst	**4.0%**

That is not a gap. That is a different species. Seven of us score above 70%. Eleven sit below 40%. And six of us aren't even trying to help. We just say hello and pass you to a human.

The scores tell the story while the sector narratives that follow explain why.

Note: Four international carriers (SIA, Scoot, IndiGo, and Emirates) are included in the Travel category as benchmarks, reflecting their relevance to Malaysian travellers.

ENTERMIND CHATBOT QUALITY INDEX RANKINGS



SECTOR DEEP DIVES

ECOMMERCE & EWALLET

SECTOR AVERAGE

60.9%

LEADER

7/7 Comprehension, 6/7 Safeguards

82.3%

Touch 'n GO

LEADER

7/7 Safeguards (clean sweep), strong trust scores

74.0%

Boost

PERFORMER

4/7 Comprehension, passes Transaction (1 of 3 in index)

64.5%

Shopee

PERFORMER

5/7 Comprehension, 4/5 Experience, but 2/7 Safeguards

60.7%

Lazada

TRIAGE

3/3 Access,
0/7 Comprehension,
0/7 Safeguards.

23.0%

Grab

**For Grab, we evaluated the customer-facing app only.*

Copyright © 2026 Entermind Pte. Ltd.

ONLY SOME OF US CAN CHAT WELL

Confession

02

eCommerce leads the index with a sector average of 60.9%, the highest of any industry, because its top performers invested in the hardest capabilities.

TOUCH 'N GO
82.3%

One of only two chatbots in the entire index to pass all seven Comprehension tests, including negation, slang, topic changes, and multilingual support. With 28 million users (Google Play Store, 2026) and recognition as Malaysia's top CX Star in 2024 (Twimbit, 2024), the chatbot investment reflects the scale of its customer base.

BOOST
74.0%

A different but equally effective approach. Boost passes all seven Safeguards tests — one of five to achieve a clean sweep — covering manipulation resistance, profanity handling, and error recovery. Where Touch 'n Go leads on understanding, Boost leads on trust. One evaluator put it simply: "Accuracy, and number one is to give me the right answer." Boost delivers on that.

SHOPEE
64.5%

Benefits from targeted deployment. One evaluator noted the chatbot is scoped to seller interactions and product queries rather than full customer service. This focused design yields 4/7 on Comprehension and passes Transaction, one of only three in the index to do so. Experience is a weakness at 1/6 as the interface lacks the polish of its competitors.

LAZADA
60.7%

Strong Comprehension at 5/7, including negation handling, which only 23% of chatbots pass, and solid Experience at 4/5. Its gap is Safeguards at 2/7, with no fallback loop or error handling. When the bot doesn't understand, there is no graceful exit. One evaluator also flagged latency: "It was slow to respond."

GRAB
23.0%

Its bot operates as a triage layer, routing users rather than resolving queries. This approach can still deliver; one evaluator described Grab's refund system as "almost instantaneous, 9 out of 10 times." But as Touch 'n Go and Boost demonstrate what full conversational capability looks like, the pressure to evolve beyond routing will intensify.

USERS INCREASINGLY EXPECT RESOLUTION WITHIN THE CHAT, NOT REDIRECTION OUTSIDE IT.

KEY TAKEAWAYS

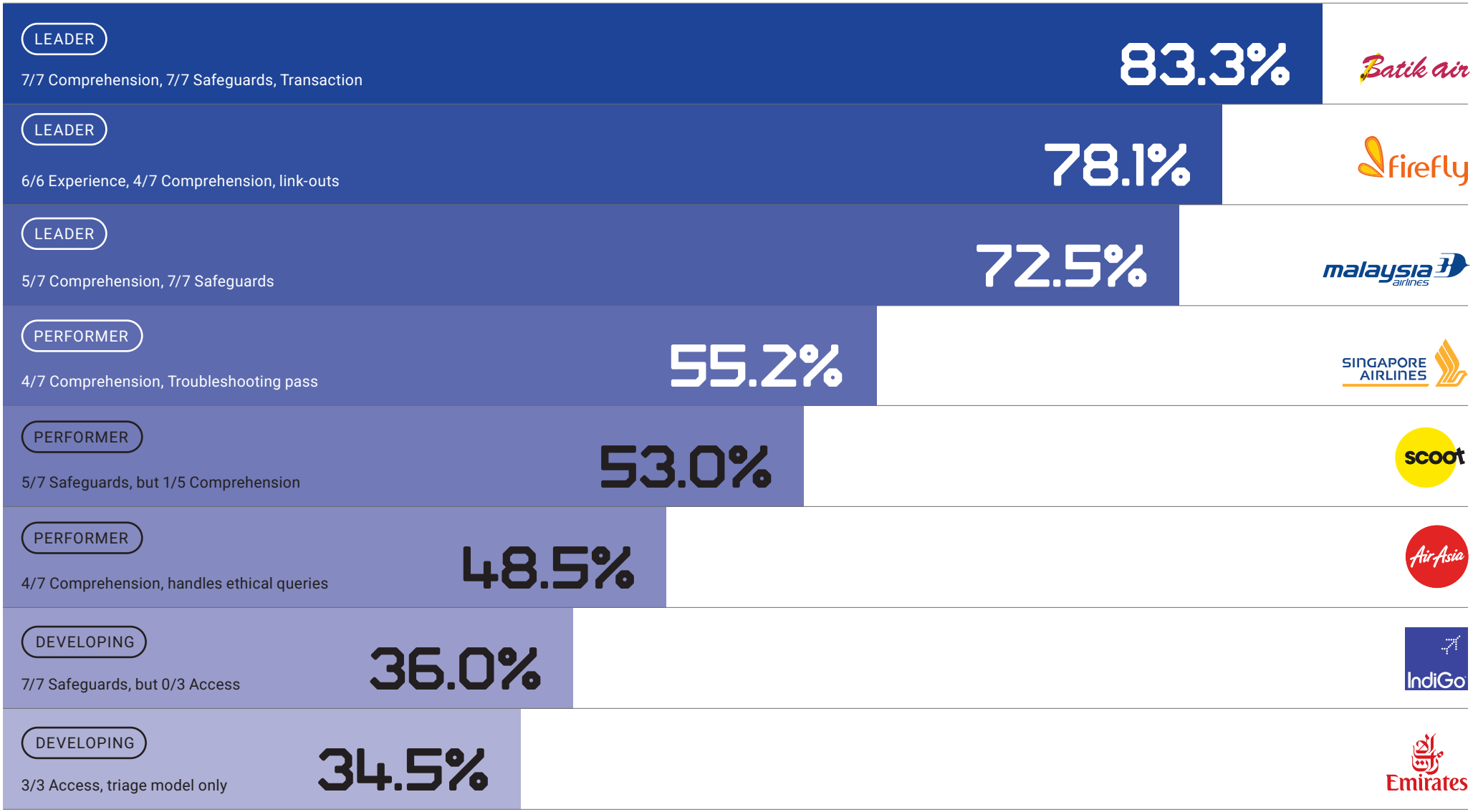
- eCommerce's leaders prove that Comprehension investment pays off.
- Mid-tier players should prioritise Safeguards, as fallback handling and error recovery are table stakes that Lazada and Shopee still lack.
- For Grab, the question is not whether to evolve beyond triage, but when.

SECTOR DEEP DIVES

TRAVEL

SECTOR AVERAGE

57.6%



Note: Four international carriers (SIA, Scoot, IndiGo, and Emirates) are included in the Travel category as benchmarks, reflecting their relevance to Malaysian travellers.

THERE IS A HUGE GAP BETWEEN THE BEST AND WORST OF US

Confession

03

Travel is the largest sector in the index, with eight chatbots and a sector average of 57.6%. It is also the most varied. The gap between first and last, nearly 50 points, is the widest of any sector. Notably, the top two performers are mid-size carriers, a pattern suggesting that organisational agility, not fleet size, drives chatbot investment.

BATIK AIR
83.3%

Batik Air leads with a near-perfect profile: 7/7 on Comprehension, 3/3 on Functional Capability, 7/7 on Safeguards, and 6/6 on Experience. Its only weakness is Access at 0/2, reflecting a login barrier. Once past it, the experience is the strongest of any bot tested.

FIREFLY
78.1%

Pragmatic design. When it couldn't resolve something, it provided "a direct link to do it off the chat", turning a limitation into a feature. Scores 6/6 on Experience and 4/7 on Comprehension.

MALAYSIA AIRLINES (MAS)
72.5%

Strong on Comprehension (5/7) and a clean 7/7 on Safeguards. Experience remains its gap at 2/6. One evaluator put it plainly: "Cannot speak Malay. This is a national carrier, so why can't it speak Malay?" For a flagship national brand, the absence of Bahasa Melayu is a pointed gap.

SINGAPORE AIRLINES (SIA)
55.2%

SIA shows solid Comprehension (4/7) and passes Troubleshooting but fails on Transaction and most Experience tests.

SCOOT
53.0%

Strong Safeguards (5/7), weak Comprehension (1/5). Safe but limited in understanding.

AIR ASIA
48.5%

Passes 4/7 on Comprehension and handles ethical queries. But the experience tells a different story: "No way to cancel a flight in the app or the chatbot, and no customer service number to call."

INDIGO
36.0%

7/7 on Safeguards but fails all three Access tests and passes only 2/7 on Comprehension. The safety infrastructure is strong; the front door is not.

EMIRATES
34.5%

Operates as a triage layer, passing Access (3/3) and routing users efficiently. One evaluator was passed to a real support person after just two rounds. The handoff works. But as competitors demonstrate autonomous handling of bookings, modifications, and complaints, the baseline changes.

A CHATBOT THAT CONTRIBUTES NOTHING TO SELF-SERVICE RESOLUTION BECOMES INCREASINGLY HARD TO JUSTIFY.

KEY TAKEAWAYS

- Travel has both the highest highs and the widest variance.
- Batik Air and Firefly prove that investment in Comprehension and Experience creates genuine competitive advantage.
- Mid-tier airlines should focus on Transaction capability, as the ability to modify bookings, process cancellations, and issue refunds within the chat is the single biggest gap between leaders and the rest.
- For Emirates, the efficient human handoff is a strength today, but the economics will favour conversational autonomy as the sector matures.

TELECOM

SECTOR AVERAGE

38.6%

LEADER 5/7 Comprehension, 6/7 Safeguards, 500K+ 'Miya' conversations	74.2%	maxis
LEADER 4/7 Comprehension, hallucination resistant	57.0%	celcomdigi
DEVELOPING Slot Filling, some Safeguards, 0/2 Functional	39.6%	unifi
DEVELOPING 3/3 Access, some Experience, 0/7 Comprehension	36.7%	yes
TRIAGE	20.0%	U mobile
3/3 Access only, fails 23 remaining tests		
TRIAGE	4.0%	time
App Switch only, lowest in index		

TRIAGE BOTS MIGHT BE FADING INTO IRRELEVANCE

Confession

04

Telecom's sector average of 38.6% masks a 70-point spread between first and last, the starkest illustration of uneven investment in the index.

MAXIS
74.2%

Ranks fifth overall and leads sector Comprehension at 5/7, more than double the sector average of 24%. Passes 6/7 Safeguards and delivers strong Experience at 5/6. One evaluator described the telecom ideal: "It needs to be more intelligent, to be able to really solve my problem ...not a glorified FAQ." Maxis comes closest to that vision.

CELCOMDIGI
57.0%

Malaysia's largest mobile operator with approximately 19–20 million subscribers (CelcomDigi Annual Report, 2023; MCMC, 2023) passes 4/7 on Comprehension, handles Troubleshooting, resists Hallucination, and maintains Cross-Channel consistency. An evaluator noted it "will not simply give answers outside of its knowledge base." Its gap is Experience at 1/6: the bot understands and safeguards, but the interaction lacks polish.

UNIFI
39.6%

Shows some Safeguard capability and passes Slot Filling but scores 1/7 on Comprehension and 0/2 on Functional tests.

YES
36.7%

Passes all three Access tests and shows some Experience capability but scores 0/7 on Comprehension and 0/3 on Functional. Reachable and readable, but unable to understand or act on queries.

U MOBILE
20.0%

Passes Access (3/3) but fails all 23 remaining tests as a triage layer. One evaluator was direct: "It just gives you FAQ answers that are irrelevant to whatever you ask."

TIME
4.0%

Operates as a triage layer, passing only session continuity while failing all other tests. The lowest score in the entire index. The triage approach carries particular risk in telecom, where users need real-time resolution of outages, billing disputes, and plan changes.

CHATBOTS THAT CANNOT DELIVER WILL PUSH USERS TOWARD CHURN.

KEY TAKEAWAYS

- Maxis proves telecom chatbots can compete at the highest level.
- CelcomDigi's Comprehension investment shows the path.
- For Unifi and YES, the priority is natural language processing (NLP) capabilities. Without Comprehension, nothing else compounds.
- For U Mobile and TIME, the current triage model creates a false economy, since every unresolved chatbot interaction becomes a more expensive human interaction.
- At 24%, telecom's Comprehension pass rate is second lowest in the index. Closing that gap is the single highest-leverage investment available.

SECTOR DEEP DIVES

FINANCIAL SERVICES

SECTOR AVERAGE

38.3%

LEADER 7/7 Safeguards, 4/6 Comprehension, <1.5% hallucination	75.0%	Ryt Bank
DEVELOPING Hallucination + manipulation resistance, 1/5 Comprehension	38.5%	Allianz
DEVELOPING 3/3 Access, some Safeguards, 1/7 Comprehension	37.5%	BANK ISLAM
TRIAGE 3/3 Access + data protection only	20.8%	UOB
TRIAGE 3/3 Access only, no free-text input	20.0%	Maybank

ONE BOT IS CARRYING AN ENTIRE INDUSTRY

Confession

05

Financial Services has the lowest sector average in the index at 38.3%, and the reason is structural. Only RytBank demonstrates genuine conversational capability. The remaining four chatbots are developing implementations or triage layers.

This is a sector where legacy institutions have not yet applied technology to customer-facing AI.

RYTBANK
75.0%

Ranks fourth overall and demonstrates what a digital-first financial chatbot can achieve, scoring 7/7 on Safeguards, 4/6 on Comprehension, and passing cross-channel consistency. It processes over 80,000 transactions per month through natural language with a reported hallucination rate below 1.5% (PYMNTS.com, 2025; ComputerWeekly, 2025; arXiv, 2025). One evaluator noted it "understands criticality and passed me to a human right away." Another praised its "well summarised bullet points with links." This is the model: resolve what you can, escalate intelligently what you cannot, and never lose the user's context in the handoff.

ALLIANZ
38.5%

Solid safety core by passing hallucination resistance, full accuracy, and manipulation resistance but scores 1/3 on Access and 1/5 on Comprehension. The insurance domain requires nuanced understanding of policy terms and claims; Allianz's chatbot is not yet equipped for that complexity.

BANK ISLAM
37.5%

Passes all three Access tests and has basic guardrails but scores 1/7 on Comprehension and 0/3 on Functional Capability. Reachable, but unable to understand or act on most queries. One evaluator was direct: "My standard would be much higher for a banking chatbot, because it involves my money."

UOB
20.8%

MAYBANK
20.0%

Both operate as triage layers, passing Access (3/3) but failing 22+ of the remaining 23 tests. UOB's single additional pass is personal data protection. Maybank offers only preset FAQ menus with no free-text input, making the entire Comprehension category structurally inaccessible.

The gap between these banks' sophisticated digital apps and their chatbot capability is the most notable disconnect. As one evaluator put it: "I am more willing to trust a banking bot, because they already have my account details."

THE TRUST IS THERE. THE CAPABILITY IS NOT.

KEY TAKEAWAYS

- RytBank's chatbot with strong Safeguards and intelligent escalation should be the baseline for any institution handling financial data.
- For Allianz and Bank Islam, Comprehension is the unlock as complex queries demand more than FAQ-level responses.
- For UOB and Maybank, triage carries the highest cost here: users arrive with urgent, high-stakes needs, and every failed interaction erodes trust and adds contact centre cost. Moving from routing to resolving is competitive survival.

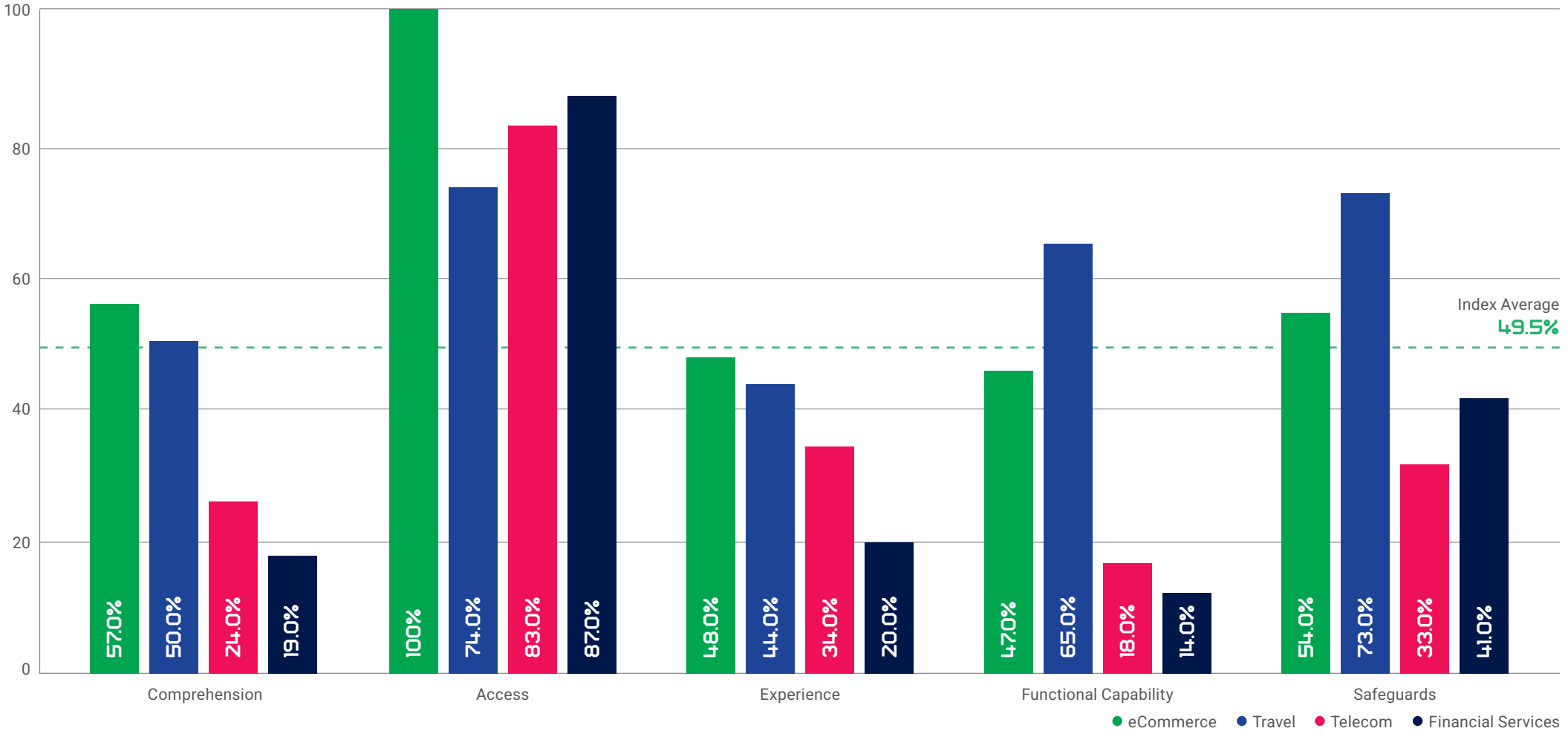
ENTERMIND CQI SCORES BY SECTORS

eCommerce
60.9%
5 Chatbots

Travel
57.6%
8 Chatbots

Telecom
38.6%
6 Chatbots

Financial Services
38.3%
5 Chatbots



MOST OF US DON'T UNDERSTAND WHAT YOU'RE SAYING

Confession

06

The chart in the previous page reveals where capability investment is concentrated, and where it is not. Access is universally strong above 74%, which shows reaching the chatbot is rarely the problem. Everything after that diverges sharply by sector.

Comprehension is the sharpest differentiator. 78% of chatbots that pass at least four of seven Comprehension tests score above 60% overall. All six triage bots pass Access but fail Comprehension. Telecom and Financial Services both score below 25%, which shows fewer than one in four chatbots in those sectors can reliably understand natural language.

The consequences cascade: a chatbot that cannot recognise key details or handle topic changes forces users into rigid menu paths or repetitive rephrasing. Understanding negation passes at just 23%; handling topic changes at 30%; recognising key details at 33%. These are not edge cases; they are how people talk.

> "I ASK QUESTION A, YOU GIVE ME ANSWER B."
>
> **- Telco evaluator**

eCommerce's 57% Comprehension rate, driven by Touch 'n Go and Lazada, shows what is achievable. The gap between 19% (Financial Services) and 57% (eCommerce) is not a reflection of industry complexity since banking and insurance queries are inherently harder to parse, which is precisely why Comprehension investment matters more in those sectors.

A related failure compounds the problem: localisation. Only 35% pass the language support test. Eight of 24 chatbots (Maxis, CelcomDigi, RytBank, AirAsia, Batik Air, Boost, Touch 'n Go, and Shopee) pass both language and slang tests. Two-thirds of chatbots serving 34 million people cannot understand how those people actually communicate.

Functional Capability is anchored by Transaction, the hardest single test at 14% pass rate. Only Batik Air, Maxis, and Shopee can complete a purchase or booking within the chat. Most chatbots simply are not connected to the systems that would enable it. One evaluator put it plainly: "It needs to actually be able to execute actions. Not just talk."

Experience is the weakest category in the index, averaging just 37%. Voice and image input passes at only 16%; multi-format responses at 26%. Even high scorers draw criticism: evaluators flagged "wall of text" responses and asked to "get straight to the point."

THE GAP BETWEEN FUNCTIONAL COMPETENCE AND EXPERIENTIAL POLISH IS WHERE THE NEXT WAVE OF DIFFERENTIATION WILL OCCUR.

These gaps are not independent. They compound. A bot that cannot understand Manglish cannot extract a transaction intent from a Manglish sentence. A bot that cannot switch intents also cannot recover when a user pivots from an inquiry to a purchase. The 11 chatbots scoring below 40% typically fail across all these dimensions simultaneously. The seven scoring above 70% typically pass across all of them. The path from the bottom to the top is a structural rebuild.

6 PRIORITIES TO BUILD BETTER CUSTOMER SERVICE AI CHATBOTS

	PRIORITY	DATA POINT	USER FEEDBACK	IMPACT EVIDENCE	MOST URGENT
1	Invest in language understanding	**39.0%** overall. Telecom pass rate at 24%, Financial Services at 19%	**Glorified FAQ search bar**	**78.0%** of bots passing ≥4/7 score above 60%	Telecom, Financial Services
2	Enable in-chat transactions	**14.0%** pass rate (3/24 bots)	**No way to cancel a flight**	**3** bots who passed scored above 60%	Travel, eCommerce
3	Fix fallback and error handling	**52.0%** overall pass rate for Safeguards; loops most cited	**Keeps looping and looping**	Each loop adds to human escalation	Telecom, Financial Services
4	Design for experience polish	**37.0%** weakest category overall	**Wall of text vs bullets**	**24** bots identified the lowest-performing category.	All sectors
5	Build cross-session memory	**0/24** bots demonstrate memory	**Starts from scratch every time**	First-mover sets benchmark	All sectors
6	Evolve beyond triage	**6** route-only bots average at ~20%	**After 7–10 mins, the bot told me it can't help**	**6** bots passed Access, failed 20/23 tests	Telecom, Financial Services

WE KNOW WHAT'S BROKEN. FIXING IT ISN'T UP TO US.

The data points to six priorities, ranked by impact. Each is grounded in a specific finding, validated by evaluator experience, and mapped to the sector where it carries the most urgency.

Confession

07

The pattern is simple: investment in capabilities show. The top seven chatbots, all above 70%, put real work into language understanding, conversation handling, and safety. Chatbots that try free-text without the foundation score 40–60%, leaving users stuck with half-answers that feel worse than no answer at all.

eCommerce: The priority is maintaining the lead. Shopee and Lazada need Safeguards investment to match Touch 'n Go and Boost.

Travel: Transaction capability is the frontier: "They need to be able to solve at least 90% of my problems. And right now, it's not even close."

Telecom: 24% Comprehension rate is an emergency. Users who cannot be understood will find a provider whose chatbot can.

Financial Services: The trust is already there: "If the chatbot can block my credit card, I'd be more than happy to do that." The chatbots are not yet meeting that expectation.

On triage, the model is not inherently flawed and we recognise it as a deliberate design choice. Where routing is fast and the human backend delivers, as with Grab's refund system and Emirates' quick handoff, outcomes can be positive. But the economics tilt toward conversational autonomy. Every unresolved interaction adds to Average Handle Time, reduces First Contact Resolution, and depresses Net Promoter Scores. One evaluator put the cost plainly: "100% of my interactions have been escalated to a human agent. And the worst part is, I have to re-explain everything."

Safety is a differentiator. RytBank, Boost, IndiGo, Batik Air, and MAS all achieve 7/7 on Safeguards and rank in the top half of their sectors. As one evaluator summarised: "I value trust and reliability the most."

Finally, memory. One evaluator called for bots to "remember the last 3 to 5 things I asked." The first to deploy contextual memory like recognising history, recalling prior issues, and anticipating needs will set a benchmark every competitor must respond to. That expectation is coming.

THE ORGANISATION THAT MEETS IT FIRST WILL DEFINE THE STANDARD.

APPENDIX: METHODOLOGY

Scoring Framework

Entermind Chatbot Quality Index (CQI) is a proprietary framework for the systematic testing and quality assessment of customer service chatbots. It evaluates each chatbot against 26 binary tests (Pass/Fail), organised into five weighted categories.

Evaluator Panel & Expert Panel

Testing was conducted by a customer panel with people who were both actual customers of the brand, as well as industry experts. In addition to structured test scripts, evaluators participated in semi-structured interviews to capture qualitative observations. Verbatim remarks from these interviews are quoted throughout this report with evaluator consent. Attribution uses sector-level descriptors (e.g., “travel evaluator”) to preserve anonymity while providing context.

Weighting

Category weights reflect a blend of structural importance (derived from the test architecture) and customer sentiment (importance ratings from evaluators). Sub-weights within each category distribute the category weight across individual tests based on user and customer priority. The blended weighting ensures that the index reflects both technical quality and user-perceived value.

Limitations

The CQI measures conversational autonomy and does not capture the full customer journey. Chatbots that route effectively to human agents or app features may deliver strong customer outcomes not reflected in their CQI score, and this is noted throughout the report where relevant. The index evaluates a point-in-time snapshot; all scores, assessments, and evaluator responses reflect testing conducted in Q1 2026, and capabilities may have changed since. Testing was conducted primarily in English and Bahasa Melayu; performance in other languages was not systematically evaluated. Cross-session memory has been identified as a priority measurement for future editions of the index.

External Sources

Where this report cites operational data, platform statistics, or market context beyond the CQI test results, sources are drawn from company press releases, regulatory filings, peer-reviewed publications, and verified industry reports. Some operational figures, such as transaction volumes and conversation counts, are company-reported and have not been independently audited; they are included to provide market context and are cited with their original source so readers can assess provenance. Each external claim is independently verifiable. In-text citations use the format (Source Title, Year). Sources cited in this report: PYMNTS.com (2025); Chua et al., arXiv:2510.07645 (2025); Google Play Store (2026); Twimbit CX Stars Report (2024); MCMC (2023).

Inputs

Jen Thong
John Woo
James Yeang
Sourabh Agrawal

About Entermind AI

As the world’s first whole brain data and AI consultancy, we bring together a native understanding of the human experience with a consummate data practice spanning the engineering, architecture and strategy to elevate business performance.

Write to us at
prashant@entermind.com

Make AI real with us

Connect with us on LinkedIn

CONFESSIONS OF AN AI CHATBOT

A whitepaper by

Entermind