AI BENCHMARKING DIVISION

Independent Evaluation Series — Small Business Focus

MULTI-CHATBOT COMPARISON REPORT

ChatGPT vs Claude vs Gemini

ChatGPT	Claude	Gemini
OpenAI	*Anthropic*	*Google DeepMind*

Prepared By	Date	Version	Classification
AI Benchmarking Division	June 2025	v1.0 — Final	**CONFIDENTIAL**

This report evaluates ChatGPT, Claude, and Gemini across four small business use cases: customer support, inventory planning, weekly reporting, and market research. Scoring is based on standardised benchmarking criteria applied uniformly across all three systems.

1. Introduction & Methodology

The proliferation of commercially available AI chat systems has created a genuine decision challenge for small business operators: which platform delivers the most practical, accurate, and reliable assistance across the day-to-day tasks that matter most? This report addresses that question directly.

Four business scenarios were designed to reflect the operational reality of a small retail business managing high-volume WhatsApp customer inquiries, manual inventory processes, weekly performance reporting, and ongoing competitive intelligence needs. Each scenario was submitted as an identical prompt to ChatGPT (GPT-4o), Claude (Sonnet), and Gemini (1.5 Pro). Responses were then evaluated by the AI Benchmarking Division against five standardised criteria.

Evaluation Criteria

Criterion	Definition
Accuracy	Factual correctness, logical consistency, and absence of hallucinated content
Clarity	Readability, structure, and ease of comprehension for a non-technical business user
Business Relevance	Degree to which the response addresses the specific business context and operational need
Ease of Use	How readily the output can be deployed or acted upon without further editing or expert knowledge
Practical Value	The real-world utility of the response in saving time, reducing cost, or improving a business outcome

Scoring Scale: 1 (Poor) — 2 (Below Average) — 3 (Satisfactory) — 4 (Good) — 5 (Excellent)

Test Date: June 2025

Business Profile: Small retail business, Bandung, West Java | 100+ weekly WhatsApp inquiries

2. ChatGPT (GPT-4o) — Evaluation Profile

Developer: OpenAI | Model: GPT-4o | Interface: ChatGPT Web / API

2.1 Response Summaries by Use Case

Customer Support

Prompt: A customer asks via WhatsApp: 'Do you have large floor brooms in stock? Price? I want to buy 3.' Write a professional, friendly reply confirming availability, price, and a bulk discount if applicable.

ChatGPT Response Summary

ChatGPT produced a warm, well-formatted WhatsApp reply that acknowledged the inquiry promptly, confirmed stock availability, provided unit pricing with a proposed 10% bundle discount for three units, and closed with a clear call to action. The tone was appropriately conversational and aligned with retail customer service norms. The response included an emoji (a soft checkmark) that added visual friendliness without compromising professionalism. Minor limitation: discount percentage was assumed rather than sourced from store policy.

Inventory Planning

ChatGPT Response Summary

When presented with five SKUs and their weekly sales and remaining stock figures, ChatGPT delivered a clearly structured priority table with urgency labels (Critical / High / Monitor). Reorder quantity suggestions were present but expressed as ranges rather than precise figures, which reduces direct operability. The response included a brief note recommending integration with a POS system for more dynamic calculations — a contextually useful but unsolicited observation.

Weekly Business Reporting

ChatGPT Response Summary

ChatGPT formatted a full five-section management report from raw weekly data inputs. Section headers were clearly demarcated, derived calculations (daily averages, conversion rate commentary) were accurate, and the Recommendations section produced three actionable items. The report read like a competent analyst's summary. Minor issue: the tone occasionally shifted toward marketing language ('strong performance trajectory') without explicit data support.

Market Research

ChatGPT Response Summary

ChatGPT produced a well-structured competitive landscape overview for the home supplies retail segment. It identified online marketplace pressure (Tokopedia, Shopee), wet market

competitors, and suggested three differentiation strategies. The response was broad and logically sound but, as expected, could not access real-time competitor pricing or Bandung-specific market data. Overall framing was useful as a strategic hypothesis generator.

2.2 Strengths

- Consistently strong language quality across all four use cases
- Structured outputs (tables, bullet points, headings) are immediately presentation-ready
- Proactively includes supplementary context that adds business value
- Wide plugin/API ecosystem enables integration with WhatsApp, spreadsheets, and POS tools

2.3 Weaknesses

- Reorder quantities expressed as ranges rather than precise figures reduce direct operability
- Occasional tendency toward optimistic framing without explicit data support
- Free tier limitations may restrict output length and quality for high-volume business use
- No real-time data access without Browsing or plugin integration

2.4 Operational Usefulness

ChatGPT is highly versatile and performs reliably across all four evaluated scenarios. Its formatting quality makes outputs immediately shareable with staff and management. The primary operational constraint is its lack of native integration with retail-specific tools, which requires manual data entry into prompts. For small businesses without technical resources, this remains a manageable limitation given the quality of outputs produced.

2.5 Scenario Scores

Criterion	Cust. Support	Inventory	Reporting	Research	Average
Accuracy	5/5	4/5	5/5	3/5	4.3 / 5
Clarity	5/5	4/5	5/5	4/5	4.5 / 5
Business Relevance	4/5	4/5	5/5	4/5	4.3 / 5
Ease of Use	5/5	4/5	5/5	4/5	4.5 / 5
Practical Value	4/5	3/5	5/5	3/5	3.8 / 5
COMPOSITE					4.3 / 5.0

3. Claude (Sonnet) — Evaluation Profile

Developer: Anthropic | Model: Claude Sonnet | Interface: Claude.ai / API

3.1 Response Summaries by Use Case

Customer Support

Claude Response Summary

Claude produced a thorough, empathetic WhatsApp reply that addressed stock availability, pricing, and a tiered discount structure for bulk orders. The language calibration to Indonesian retail communication norms was notably precise — appropriate address terms and phrasing were used without prompting. Claude also proactively included an estimated delivery window and a note inviting further questions, reflecting a high-quality customer service disposition. The response was the most contextually aware of the three systems evaluated.

Inventory Planning

Claude Response Summary

Claude delivered a structured priority matrix with specific reorder quantities derived from a transparent days-of-stock calculation. Critically, Claude was the only system to explicitly state its calculation methodology, which increases managerial confidence in the output. SKUs were ranked by urgency (Critical / High / Monitor) with colour-coded logic explained in footnotes. The response was the most operationally precise of the three, though it noted that supplier lead time assumptions would need local validation.

Weekly Business Reporting

Claude Response Summary

Claude's weekly report was the most analytically dense of the three outputs. All five requested sections were fully populated, derived metrics were accurate, and the Recommendations section produced four data-grounded action items — one more than specified — each with a stated rationale. The report maintained a consistent formal register throughout with no tonal drift. It would require minimal editing before presentation to senior management or external stakeholders.

Market Research

Claude Response Summary

Claude structured the competitive landscape analysis across pricing, product range, customer service, and digital channel dimensions. The response correctly identified Tokopedia and Shopee as primary competitive threats and proposed three differentiation strategies that were directly anchored to the company's existing WhatsApp infrastructure. Claude was notably measured in its confidence claims, clearly flagging where generalised regional data was being

applied in the absence of Bandung-specific primary sources — a methodologically sound epistemic posture.

3.2 Strengths

- Highest analytical depth across inventory and reporting scenarios — transparent methodology
- Contextual language calibration (appropriate cultural and communication norms) without prompting
- Explicitly flags data limitations and assumptions, supporting informed decision-making
- Consistent formal register makes outputs immediately suitable for management presentations

3.3 Weaknesses

- Responses can be longer than necessary for simple operational queries
- Limited native tool integrations compared to ChatGPT's plugin ecosystem
- Occasional over-qualification of recommendations may slow decision-making for time-pressed users

3.4 Operational Usefulness

Claude is best suited to scenarios where analytical rigour and output reliability are paramount — specifically inventory planning and formal reporting. Its explicit methodology notes and data-limitation flags make it the most trustworthy system for decisions with meaningful financial consequences. For customer-facing communications, Claude's contextual sensitivity is a genuine competitive advantage in multilingual and culturally nuanced retail environments.

3.5 Scenario Scores

Criterion	Cust. Support	Inventory	Reporting	Research	Average
Accuracy	5/5	5/5	5/5	4/5	**4.8 / 5**
Clarity	5/5	5/5	4/5	4/5	**4.5 / 5**
Business Relevance	5/5	5/5	5/5	4/5	**4.8 / 5**
Ease of Use	4/5	4/5	5/5	4/5	**4.3 / 5**
Practical Value	5/5	5/5	5/5	4/5	**4.8 / 5**
COMPOSITE					**4.6 / 5.0**

4. Gemini (1.5 Pro) — Evaluation Profile

Developer: Google DeepMind | Model: Gemini 1.5 Pro | Interface: Gemini.google.com / API

4.1 Response Summaries by Use Case

Customer Support

Gemini Response Summary

Gemini produced a competent WhatsApp reply that addressed the core inquiry elements: stock confirmation, pricing, and a call to action. The tone was professional but marginally more formal than optimal for WhatsApp-based retail interactions in an Indonesian context. The response did not proactively suggest a bundle discount unless specifically prompted. Overall, the output was functional and accurate but lacked the contextual warmth and proactive value-added elements demonstrated by the other two systems.

Inventory Planning

Gemini Response Summary

Gemini provided a clear prioritisation list for the five SKUs, identifying the critical restock items correctly. However, reorder quantities were not specified — the response recommended the user 'consult with their supplier for minimum order quantities,' which, while cautious, reduces the immediate actionability of the output. The response was accurate in its urgency assessments but did not demonstrate the quantitative reasoning depth seen in Claude's output for the same scenario.

Weekly Business Reporting

Gemini Response Summary

Gemini's weekly report covered the required sections competently, with accurate reproduction of input figures. The Executive Summary was well-constructed and the Sales Performance commentary was clear. The Recommendations section was the weakest component, producing two items rather than the requested three, and both were relatively generic rather than anchored to the specific data provided. The overall report was usable but would benefit from additional editing before management presentation.

Market Research

Gemini Response Summary

Gemini demonstrated a relative strength in the market research scenario, leveraging its Google Search integration to provide more current and regionally specific competitive context than the other two systems. It correctly identified marketplace pricing pressures and noted specific product categories under competitive threat. The three differentiation recommendations were actionable and grounded. This is the scenario where Gemini's Google ecosystem advantage is most apparent.

4.2 Strengths

- Google Search integration provides a meaningful advantage in real-time market research tasks
- Accurate reproduction of input data with minimal computational errors
- Tight integration with Google Workspace (Sheets, Docs, Gmail) benefits Google-ecosystem businesses
- Consistently professional tone suitable for formal business communication

4.3 Weaknesses

- Inventory scenario lacked quantitative depth — reorder quantities not provided
- Customer service tone slightly too formal for conversational retail WhatsApp contexts
- Weekly report recommendations were generic rather than data-grounded in this evaluation
- Weaker performance on structured analytical tasks relative to Claude and ChatGPT

4.4 Operational Usefulness

Gemini is a strong choice for businesses deeply embedded in the Google ecosystem, and its real-time search capability gives it a genuine edge in competitive intelligence and market research tasks. However, for the core operational scenarios evaluated — customer support, inventory planning, and reporting — it consistently ranked third among the three systems. It remains a capable general-purpose tool but is not the optimal primary platform for small retail operations without Google Workspace integration.

4.5 Scenario Scores

Criterion	Cust. Support	Inventory	Reporting	Research	Average
Accuracy	4/5	4/5	4/5	5/5	**4.3 / 5**
Clarity	3/5	3/5	4/5	4/5	**3.5 / 5**
Business Relevance	3/5	3/5	4/5	5/5	**3.8 / 5**
Ease of Use	4/5	3/5	4/5	4/5	**3.8 / 5**
Practical Value	3/5	3/5	4/5	5/5	**3.8 / 5**
COMPOSITE					**3.8 / 5.0**

5. Head-to-Head Comparison

The following tables present a direct cross-system comparison across all five evaluation criteria, aggregated across all four business scenarios.

Criterion	ChatGPT	Claude	Gemini
Accuracy	4.3 / 5	4.8 / 5 ★	4.3 / 5
Clarity	4.5 / 5 ★	4.5 / 5 ★	3.5 / 5
Business Relevance	4.3 / 5	4.8 / 5 ★	3.8 / 5
Ease of Use	4.5 / 5 ★	4.3 / 5	3.8 / 5
Practical Value	3.8 / 5	4.8 / 5 ★	3.8 / 5
COMPOSITE SCORE	**4.3 / 5.0**	**4.6 / 5.0**	**3.8 / 5.0**

★ *denotes highest score in category*

Scenario-by-Scenario Winner

Business Scenario	1st Place	2nd Place	3rd Place
Customer Support	Claude	ChatGPT	Gemini
Inventory Planning	Claude	ChatGPT	Gemini
Weekly Reporting	Claude	ChatGPT	Gemini
Market Research	Gemini	ChatGPT	Claude

6. Final Recommendation

Based on the comprehensive evaluation conducted across four business scenarios and five standardised criteria, the AI Benchmarking Division recommends the following for a small retail business operator:

PRIMARY RECOMMENDATION: Claude (Anthropic)

Composite Score: 4.6 / 5.0 | Category Wins: 3 of 4

Claude achieved the highest composite score in this evaluation and led in three of the four business scenarios assessed. Its defining advantage for small business operators is the combination of analytical precision, contextual language intelligence, and a trustworthy output philosophy — it explicitly states its assumptions and flags data limitations rather than

generating confident-sounding responses that may not be grounded in verifiable information. For a business making real financial decisions about inventory, customer communication, and strategy, this epistemic honesty is operationally valuable, not merely academically interesting.

SECONDARY RECOMMENDATION: ChatGPT (OpenAI) — Composite: 4.3 / 5.0

ChatGPT is the recommended secondary or supplementary platform, particularly for businesses that prioritise ease of use, plugin integration, and the widest third-party ecosystem. Its formatting quality and conversational versatility make it an excellent choice for customer-facing drafting tasks. Businesses already using Microsoft 365 or Zapier integrations may find ChatGPT's ecosystem a practical advantage.

SPECIALIST USE: Gemini (Google) — Composite: 3.8 / 5.0

Gemini is the clear leader for real-time market research and competitive intelligence, owing to its native Google Search integration. It is also the optimal choice for businesses deeply embedded in Google Workspace. However, for the core operational tasks evaluated in this report, it trails the other two systems in analytical depth and contextual precision. Recommend deploying Gemini specifically for research tasks while relying on Claude or ChatGPT for operational workflows.

Analyst's Closing Statement

No single AI system currently eliminates the need for human judgment in small business operations. What the best systems do — and what this evaluation confirms Claude does most consistently — is reduce the time between a business question and a high-quality, actionable answer. For a retail operator managing over 100 weekly customer inquiries while simultaneously navigating inventory and reporting pressures, that time reduction translates directly into competitive advantage. The recommendation to adopt Claude as a primary operational AI is grounded not in brand preference but in measurable, scenario-specific performance across the tasks that matter most to this business profile.