TEST REPORT

Tolly.

#226106

April 2026

Commissioned by octonomy AI GmbH

Agentic AI Accuracy Benchmark
Complex Document Comprehension vs. Competing Solutions

EXECUTIVE SUMMARY

octonomy AI commissioned Tolly to evaluate the accuracy of octonomy Agentic AI against three competing AI solutions in answering 50 complex knowledge questions derived from a production enterprise documentation library spanning 1,000+ pages of real-world materials including annotated diagrams, performance curves, multi-variable data tables, and cross-referenced specifications.

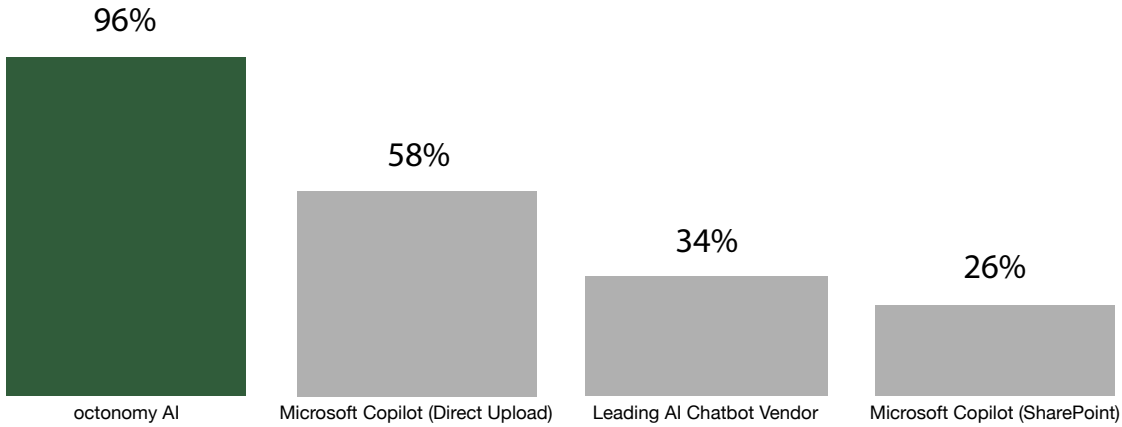
The questions were specifically designed to require interpretation of complex source material, the kind of documentation found across every industry, rather than simple text extraction. The benchmark spanned four question complexity categories testing distinct AI reasoning capabilities: multi-document reasoning, precision data extraction from graphical sources, visual and spatial interpretation, and complex structured data navigation.

The majority of answers could only be obtained by reading values from graphs, interpolating between data points on curves, cross-referencing information across multiple documents, or interpreting annotated drawings. These are challenges that mirror complex knowledge work across every industry and vertical. octonomy AI accurately answered 96% of the questions where the accuracy of the other solutions evaluated ranged from 58% down to 26%. See Figure 1.

THE BOTTOM LINE

1. **96% accuracy** - octonomy Agentic AI correctly answered 48 of 50 complex knowledge questions without hallucination
2. **Real-world validated** - benchmark conducted against a production enterprise documentation library with questions reflecting genuine operational scenarios, not synthetic test data
3. **Cross-industry relevance** - the four question categories tested (multi-document reasoning, graphical data extraction, spatial interpretation, structured data navigation) map directly to complex knowledge work in every sector

Comparison of Agentic AI Solutions: Complex Document Comprehension Accuracy



Note 1: Microsoft Copilot (Direct Upload) used Claude Sonnet as the underlying model, with documentation uploaded directly into Copilot. Microsoft Copilot (SharePoint) used ChatGPT as the underlying model, with documentation stored in SharePoint. Both configurations required manual pre-segmentation of documentation into chapter-level files.
Note 2: "Leading AI Chatbot Vendor" refers to a US market-leading agentic customer support platform. Documentation was manually segmented before upload.

Source: Tolly, March 2026

Report link: https://www.tolly.com/publications/226106

Figure 1

Background

Replacing Lost Knowledge

Complex machinery is built to last. Countless products can and do remain in service for decades. Increasingly, the people who know the most about such products are retiring from the workforce. This leaves a knowledge gap that threatens a manufacturer's ability to provide suitable product support to technicians in the field and others.

Providing technical support for complex products is one area where AI solutions can not only fill the gap but potentially improve upon support delivered by people.

Visually Complex Sources

Machinery and other sophisticated products typically rely on complex visuals in product documentation. Drawings with measurements and tolerances, tables containing precise numerical data, graphs of elements such as temperature over time are just some of the data sources that are encountered in support manuals.

octonomy AI was architected to be able to understand complex visuals in addition to textual information. According to the company, the accuracy of competing offerings suffers when answering questions that require evaluation of complex visual information. The purpose of this test was to provide quantitative evidence for that claim.

Source & Focus

The test project largely replicated a proof-of-concept test run internally by octonomy AI in recent months. The technical focus of the test was complex industrial equipment where the service manual was in excess of 1,000 pages in length. According to octonomy AI, the questions used in the Tolly test were largely based on the questions posed by the vendor of complex industrial equipment in evaluating octonomy AI.

Question Complexity Categories

In order to exercise the solutions appropriately, the questions were distributed across four AI reasoning categories. See Figure 2, below, for high-level category names and distribution. A deeper discussion of each area and its importance follows later in this report.

Data Preparation

It is important to note that octonomy AI processed the complete, unmodified documentation library (thousands of pages) as a single knowledge base. All competing platforms required the documentation to be manually pre-segmented into individual chapters before they could process it - a significant structural advantage for the competitors that simplified their retrieval task. Microsoft Copilot was evaluated using two different methods of data preparation: 1) direct upload, and 2) upload to Microsoft SharePoint.

Test Results

Summary

As noted above, octonomy AI answered all but two of the 50 questions correctly for a 96% accuracy rate.

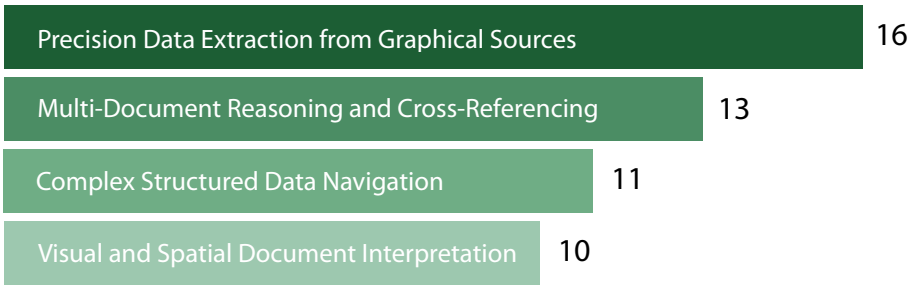
Microsoft Copilot was tested twice using two different upload repositories and two different underlying LLMs. See Table 2 for more information. Microsoft fared better when the product documentation was uploaded directly to Copilot. In this test, with Anthropic Claude as the underlying LLM, Copilot's accuracy was 58%.

Tested a second time where product documentation was uploaded to a Microsoft SharePoint portal, and running OpenAI ChatGPT, Copilot only answered 26% of the questions correctly.

To provide additional perspective, the test also included a leading AI chatbot vendor. That solution was able to answer 34% of the questions accurately.

Question Complexity Categories

Distribution of 50 benchmark questions across four AI reasoning capabilities



Source: Tolly, March 2026 Figure 2

AI Hallucinations

The AI industry describes an AI hallucination as an incorrect or fabricated output from a generative model that looks plausible but isn't supported by facts or evidence - examples include made-up facts, fake citations, or wrong dates.

Tolly analysts noted that frequently the incorrect answers were hallucinations. Rather than simply answering "that information cannot be found," a solution would present a verbose, plausible-sounding answer that, in fact, was completely incorrect. In Tolly's view, misleading (and incorrect) answers are far worse than no answer at all. "Hallucinated" answers that are accepted by users as correct and are acted upon have the potential to cause significant problems for system users.

Test Insights & Examples

Failure Patterns

While it is impractical to publish detailed results for each question, analysts noted consistent failure patterns for the competing solutions. See Table 1 for a summary of these patterns.

As can be seen, competing solutions had difficulty evaluating visual information (charts, tables, schematics, etc) and, thus, either frequently provided no answer or an incorrect answer.

As noted previously, "AI hallucinations" were encountered frequently as competing solutions delivered plausible answers that were objectively incorrect.

On the following pages, details are provided for three example questions referenced in the first three rows of Table 1.

Because of the confidential nature of the 50 official test questions, the examples have been anonymized. Tolly confirms that the anonymized examples accurately reflect the types of answers encountered in the full test. Each example is illustrated in a separate call-out box. See Figures 3-5 on the following pages.

Following the example results, there is a section that provides more detailed information about each of the question categories mentioned in Figure 2, why they are important, and why such question categories present challenges to AI solutions.

octonomy AI

Agentic AI

Accuracy Benchmark: Complex Document Comprehension

Tolly. Certified

Tested March 2026

Failure Patterns

Summary of Competing Solution Error Paths

Failure Pattern	Description
Failure to locate visual information	Competitors frequently could not find answers that were encoded in diagrams, drawings, or charts rather than in text. When the answer existed only as a visual annotation (Example 1: force value on a drawing), they either missed it entirely or concluded the information was not documented.
Inability to distinguish between contexts	When multiple similar-looking answers existed for different scenarios (Example 2: indoor vs. outdoor installation), competitors retrieved the wrong one. They found relevant documentation but failed to apply the conditional context specified in the question.
Failure to perform graphical interpolation	When the answer required reading a value from a curve at a point between plotted data (Example 3: temperature at a specific speed), competitors either refused to estimate, produced incorrect values, or confused measurement units.
Hallucination under uncertainty	When competitors could not find the correct information, they sometimes fabricated plausible-sounding answers rather than acknowledging uncertainty - a particularly dangerous failure mode in technical and safety-critical contexts.

Notes: Analysis of errors and incorrect evaluation paths encountered in competing solutions.

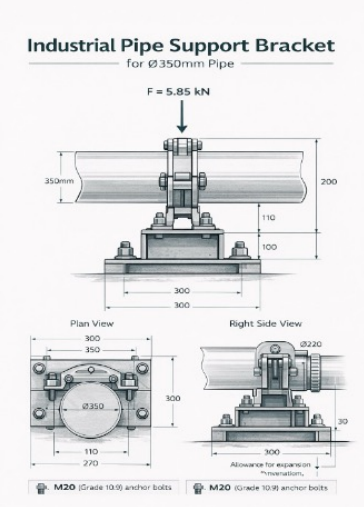
Source: Tolly, March 2026

Table 1

Example 1 - Visual and Spatial Document Interpretation

Question: "What is the design load acting on the support bracket of the IPSB-350 pipe support assembly?"
Correct Answer: "F = 5.85 kN"

Octonomy AI - Correct	Leading AI Chatbot - Failed	Copilot - Failed
Identified the face value **F = 5.85 kN** directly from the engineering drawing. Also, it correctly described the context: downward load shown in front view, 0350 mm pipe, M20 anchor bolts, and all three orthographic views.	Could not locate the relevant documentation. It stated that the information "isn't clearly documented" despite the value being annotated directly on the engineering drawing.	Found the drawing but could not extract the force value. It listed the dimensions (300mm, 200mm, 110mm) but missed the **F = 5.85 kN** annotation. It concluded that the document "does not explicitly specify a numeric design load."



Why the competitors failed:

The force value **(F = 5.85 kN**) is annotated directly on the engineering drawing as a labeled arrow. The value is not present in any text, table, or specification. Both competitors either failed to locate the drawing entirely or could not interpret the visual annotation, instead focusing on dimensional values that appear as plan numbers.

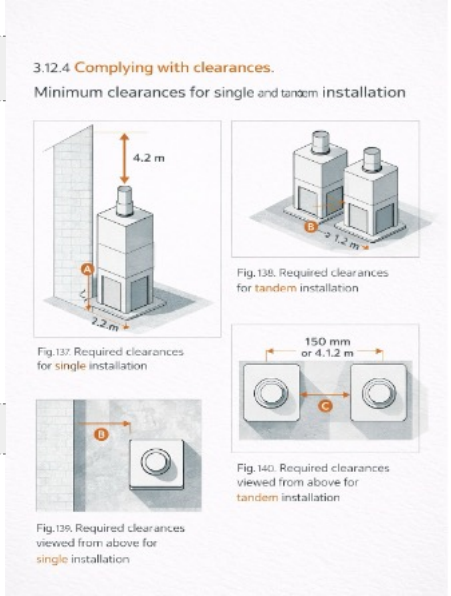
Source: Tolly, March 2026

Figure 3

Example 2 - Multi-Document Reasoning with Conditional Context

Question: "What minimum overhead clearance must be observed for the AeroStack outdoor scrubber installation?"
Correct Answer: "≥ 4.2 m (outdoor installation)"

Octonomy AI - Correct	Leading AI Chatbot - Hallucinated	Copilot - Wrong Context
Identified the correct outdoor clearance of **≥ 4,200 mm (4.2 m)** and provided the full clearance table, including lateral and tandem spacing requirements. Correctly distinguished between indoor and outdoor installation contexts.	Failed to find the outdoor installation documentation. Responded with a generic indoor recommendation of "≥ 2.40 m room height" — a value from a different section entirely. Fabricated guidance about "air exchange volume" not present in the source material.	Found clearance documentation but returned the **indoor** installation values (2.1 m vertical, 0.6 m sides, 1.0 m front) instead of the outdoor values. Failed to distinguish between the two installation scenarios despite the question explicitly specifying "outdoor."



Why the competitors failed:

The documentation contains separate clearance diagrams for indoor and outdoor installations on different pages. The question explicitly asks about outdoor installation, but both competitors returned indoor values - either because they couldn't locate the outdoor-specific diagram, or because they failed to distinguish between the two contexts. This is a classic multi-document reasoning failure: the AI found a relevant answer but not the correct answer for the specified condition.

Source: Tolly, March 2026

Figure 4

Example 3 - Precision Data Extraction from Graphical Sources

Question: "What is the product core temperature of the CTO-240 conveyor oven at 2.5 m/min?"
Correct Answer: "30 °C (±2 °C tolerance for interpolation)"

Octonomy AI - Correct	Leading AI Chatbot - Wrong Units and Wrong Value	Copilot - Could Not Interpolate
Located the temperature characteristic curve, correctly read the axes (°C vs. m/min), and interpolated a value of **≈ 31 °C** - within the acceptable tolerance. Also correctly explained the relationship: temperature decreases as conveyor speed increases due to reduced dwell time.	Found the correct chart but reported values in **°F instead of °C** - a fundamental unit error. Also stated "approximately 35 °F" which, even if converted, does not match the correct value. Failed at both reading accuracy and unit identification.	Found and correctly described the chart (axes, shape of curve) but stated it **could not provide a numeric answer** because "the exact numeric product core temperature at 2.5 m/min is not explicitly listed as a discrete data point." Failed to perform visual interpolation.

Why the competitors failed:

This question requires graphical interpolation — reading a value from a curve at a point that falls between marked data points. Leading AI Customer Agent attempted this but confused the units (°C vs. °F) and produced an inaccurate reading. Copilot acknowledged the chart existed but refused to interpolate, treating the absence of an explicit data point as an inability to answer. Only Octonomy AI successfully performed the visual interpolation and returned an accurate value within tolerance.

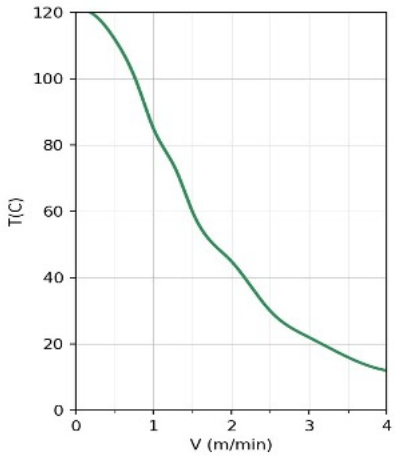


Fig. 142 Product core temperature characteristic curve for CTO-240 conveyor oven

Source: Tolly, March 2026

Figure 5

Question Complexity Categories

Each of the question categories were based on real-world examples that present challenges to AI solutions. This section provides further insights into each category and each category. See Figure 2 for the number of questions from each category used in the test.

These four categories represent fundamentally different AI reasoning capabilities. Each tests a distinct cognitive task: visual data interpretation, cross-document synthesis, spatial reasoning, and multi-conditional data filtering that maps directly to real-world knowledge work across every industry. The benchmark questions were drawn from a production enterprise documentation library, ensuring that the complexity reflects genuine operational challenges rather than synthetic test scenarios.

Category 1 - Precision Data Extraction from Graphical Sources

Real-World Challenge

Locating and reading exact values from complex charts, curves, and multi-line graphs often requiring interpolation between plotted data points where the answer falls between marked values on an axis.

Why This Is Hard for AI

Standard text extraction (OCR, RAG) cannot read values from graphical data. The AI must understand axis scales, identify the

correct data series, trace to the specified operating point, and interpolate a precise numerical answer from a visual representation. On logarithmic or non-linear scales, even small visual errors produce significantly wrong answers.

Benchmark Example

Referencing Example 3/Figure 5. Performance rating curves requiring value extraction at specific operating conditions, multi-line graphs where the correct data series must be identified before reading a value, capacity and output charts with multiple overlapping product curves, characteristic diagrams requiring interpolation between plotted data points at a specified input parameter.

Category 2 - Multi-Document Reasoning and Cross-Referencing

Real-World Challenge

Answering questions where the required information is scattered across multiple pages, sections, or entirely separate documents, requiring the AI to identify, retrieve, and synthesize related facts from different locations to produce a single coherent answer.

Why This Is Hard for AI

The AI cannot find the answer in any single location. It must first understand what pieces of information are needed, locate each piece across a large documentation library, recognize how they relate to each other, and combine them correctly. Even the retrieval step is more complex than it appears, when pages lack sufficient contextual markers, multiple pages may appear equally relevant to a query, yet only one contains the correct answer for the specific context in question. Without understanding the broader document context, the AI has no reliable way to distinguish between these near-matches. This tests retrieval breadth, contextual understanding, and logical synthesis not just search accuracy.

Benchmark Example

Referencing Example 2/Figure 4. Matching a product configuration in one document to its component assignment in another, combining installation parameters from a sizing guide with model-specific data from a separate technical sheet, cross-referencing room or site requirements with safety compliance calculations from a different manual section, identifying the correct accessory article number by matching product variant to compatibility lists across separate catalogues.

Category 3 - Visual and Spatial Document Interpretation

Real-World Challenge

Extracting meaningful information from diagrams, schematics, flow charts, and annotated drawings where the answer is encoded in visual and spatial relationships rather than in text or tabular data.

Why This Is Hard for AI

The information exists as visual elements, dimension lines, callout labels, spatial relationships between components, directional flow arrows, and annotated cross-sections. The AI must understand engineering and technical drawing conventions, interpret multiple orthographic views (front, plan, side), and extract specific measurements or relationships from complex visual layouts.

Benchmark Example

Referencing Example 1/Figure 3. Dimensional drawings with measurement callouts requiring extraction of specific distances, installation clearance diagrams showing minimum distances from multiple perspectives, wiring and controller configuration schematics with numbered terminal assignments, mounting templates with drill-hole positions and structural load ratings, connection point elevation drawings, process flow diagrams illustrating multi-step system operations.

Category 4 - Complex Structured Data Navigation

Real-World Challenge

Finding precise values within dense, multi-conditional tables, spreadsheets, and structured databases where the correct answer depends on filtering across three or more variables simultaneously.

Why This Is Hard for AI

Dense tables and spreadsheets can contain hundreds of similar-looking numerical values. The AI must correctly identify which row, column, and conditional filter combination yields the right answer, a task that requires understanding table structure, header hierarchies, merged cells, and multi-level conditional logic. When data spans multiple sheets or tabs, the complexity multiplies further.

Benchmark Example

Emission and output tables filtered simultaneously by product model, operating mode, and test condition; product specification matrices requiring precise row-and-column identification for performance ratings or capacity values; component identification from dense parts

catalogues with hundreds of entries; minimum/maximum value extraction from conditional performance matrices spanning multiple operating scenarios.

Cross-industry Applications

AI skills find broad relevance across industries, with practical applications ranging from pharmaceutical dosage-response and drug interaction matrices, financial yield and performance charts, manufacturing quality control SPC graphs, energy consumption load profiles, pump and compressor performance curves, agricultural yield-vs-input curves, telecommunications signal attenuation charts, and e-commerce conversion funnels and sales trend charts.

Document and data intelligence is particularly valuable across sectors where information is fragmented or voluminous - legal clause cross-referencing across contracts and amendments, medical history correlation across patient records and lab results, procurement specification matching against supplier datasheets, compliance verification across regulatory filings, insurance policy coverage determination across base policies and riders, e-commerce product catalogue reconciliation across supplier feeds and inventory records, and heavy machinery maintenance scheduling cross-referenced with parts catalogues and service bulletins.

Industries that rely heavily on spatial and schematic documentation also benefit significantly, including architectural floor plans and building layouts, warehouse and distribution centre layouts, network topology diagrams, manufacturing assembly instructions and exploded views, industrial equipment installation and rigging schematics, medical device setup schematics, process engineering P&ID diagrams, retail store planograms, and organizational workflow charts.

Wherever complex structured data governs operations, AI brings meaningful capability - from enterprise spreadsheets used as operational databases and insurance rate and underwriting tables, to regulatory compliance grids, tiered pricing and discount matrices, industrial parts compatibility databases and bill-of-materials lookups, e-commerce product attribute matrices with multi-variant SKU filtering, and financial reporting workbooks with multi-tab cross-references.

Agentic AI Solutions under Test

Vendor	Solution	Documentation Upload	LLM Model
Octonomy	Agentic AI	Single upload to knowledge base	Anthropic Claude Sonnet
Microsoft	Copilot	Segmented (by chapters) uploaded directly to Copilot	Anthropic Claude Sonnet
Microsoft	Copilot	Segmented (by chapters) and uploaded to Microsoft SharePoint	Open AI ChatGPT
Leading AI Chatbot Vendor	AI Agent	Segmented (by chapters) and uploaded to knowledge base	Anthropic Claude Sonnet

Source: Tolly, March 2026

Table 2

Test Setup & Methodology

Test Setup

Each AI was configured as appropriate for the test. Prompts were kept as similar as possible and only modified based on limitations of a given solution. See Table 2 for details of the solutions. LLMs were current at the time of testing (March 2026).

As noted earlier, only octonomy was able to ingest the entire 1,000+ page technical service manual without manual pre-processing by the test team. For the other solutions, the manual was broken down into its constituent parts of 20+ chapters.

Test Methodology

As noted, the vast majority of questions were taken from an earlier, internal customer PoC. Some additional questions were added to highlight differences across the solutions.

Where possible, the questions were submitted as a batch to the AI. After the questions were processed, Tolly evaluated the response and compared it to the correct response.

Answers were deemed either correct or incorrect. No credit was given for partial answers. Hallucinated answers were observed frequently from competing solutions but these additional faults were not calculated into the results. Hallucinated answers were simply logged as incorrect answers.

About Tolly

The Tolly Group companies have been delivering world-class IT services for more than 35 years. Tolly is a leading global provider of third-party validation services for vendors of IT products, components and services.

You can reach the company by E-mail at sales@tolly.com, or by telephone at +1 561.391.5610.

Visit Tolly on the Internet at:
http://www.tolly.com

Terms of Usage

This document is provided, free-of-charge, to help you understand whether a given product, technology or service merits additional investigation for your particular needs. Any decision to purchase a product must be based on your own assessment of suitability based on your needs. The document should never be used as a substitute for advice from a qualified IT or business professional. This evaluation was focused on illustrating specific features and/or performance of the product(s) and was conducted under controlled, laboratory conditions. Certain tests may have been tailored to reflect performance under ideal conditions; performance may vary under real-world conditions. Users should run tests based on their own real-world scenarios to validate performance for their own networks.

Reasonable efforts were made to ensure the accuracy of the data contained herein but errors and/or oversights can occur. The test/audit documented herein may also rely on various test tools the accuracy of which is beyond our control. Furthermore, the document relies on certain representations by the sponsor that are beyond our control to verify. Among these is that the software/hardware tested is production or production track and is, or will be, available in equivalent or better form to commercial customers. Accordingly, this document is provided "as is", and Tolly Enterprises, LLC (Tolly) gives no warranty, representation or undertaking, whether express or implied, and accepts no legal responsibility, whether direct or indirect, for the accuracy, completeness, usefulness or suitability of any information contained herein. By reviewing this document, you agree that your use of any information contained herein is at your own risk, and you accept all risks and responsibility for losses, damages, costs and other consequences resulting directly or indirectly from any information or material available on it. Tolly is not responsible for, and you agree to hold Tolly and its related affiliates harmless from any loss, harm, injury or damage resulting from or arising out of your use of or reliance on any of the information provided herein.

Tolly makes no claim as to whether any product or company described herein is suitable for investment. You should obtain your own independent professional advice, whether legal, accounting or otherwise, before proceeding with any investment or project related to any information, products or companies described herein. When foreign translations exist, the English document is considered authoritative. To assure accuracy, only use documents downloaded directly from Tolly.com. No part of any document may be reproduced, in whole or in part, without the specific written permission of Tolly. All trademarks used in the document are owned by their respective owners. You agree not to use any trademark in or as the whole or part of your own trademarks in connection with any activities, products or services which are not ours, or in a manner which may be confusing, misleading or deceptive or in a manner that disparages us or our information, projects or developments.

pc-01 wt-2026-04-21-Verl