SOLUTION BRIEF

Turnkey private AI. On your site.

GenBI, powered by Wren AI.
AI server, platform software, and models in one box.
Plain-language answers from your own data.

Powered by WrenAI



Shown: AI Server - 4 GPU (2U). Also available with 8 GPUs.

TURNKEY

One system. Ready to run.

First application live in 14 days.

EXTENDED AI MEMORY

More AI. Fewer GPUs.

2× the sessions on the same GPUs with Pascari aiDAPTIV™, measured.

ENTERPRISE READY

Your data. Your models. Your site.

Works next to your BI and existing databases.

Copyright ©2026 Phison Electronics | All Rights Reserved.

WHY PRIVATE AI NOW

Now is the time to build it right.

CONTROL

Keep AI inside your boundary.

Prompts, data, and models stay on your site. Runs air-gapped when the network cannot connect out.

PERFORMANCE

Meet service targets as use grows.

Built for many users at once, long context, and steady response times.

ECONOMICS

Predictable cost for daily AI.

No per-token fees. Plan capacity and total cost up front.

INSIDE THE SYSTEM

Every layer, one system.

GenBI · Wren AI
Plain-language questions, analysis, charts, reports, and dashboards.

Phison AI Data Platform software
Installs, runs, and monitors the AI models, GPUs, storage, and application.

AI server
GPU compute, system memory, aiDAPTIV Cache Memory, and Pascari Enterprise SSDs.

GenBI is the first solution on the Phison AI Data Platform. More applications are coming soon.

HOW AIDAPTIV WORKS

Same GPUs. Twice the AI work.

GPU and system memory set the limit. aiDAPTIV adds a flash memory tier built for AI, with more room for every session. More users. More AI. Faster first token.*

Concurrent sessions	Reports completed	Follow-up answers	Time to first token
80	**78**	**4.7 min**	**1.9 s**
vs 40 without	vs 38 without	vs 4.8 without	vs 3.8 s without
2× sessions	**2.1× more**	**Same at 2× load**	**2× faster**

Holds on 8 GPUs too: 160 vs 80 concurrent sessions.*

*Phison lab test, Oct–Nov 2026. Same server with and without aiDAPTIV. 4× NVIDIA RTX PRO 6000 Blackwell. Wren AI 2.5.1 GenBI agent, Qwen3.8-27B-NVFP4, 256K context. Session = one person asking for a report, then follow-ups, about 5 minutes apart. Smooth answers = at least 10 tokens/s in 95% of minutes. Time to first token = median, per model call. 8 GPUs: Server Edition; 160 includes 60 simulated sessions.

HOW IT COMPARES

Not a second server. A complete private AI platform.

	Cloud AI	Build it yourself	Phison AI Data Platform
Where it runs	Off site	On site	On site, air-gap ready
Time to first app	Fast	Your project	14 days*
Who stands behind it	Provider	Your team	Phison and partners

*From when the hardware is on site.

Copyright ©2026 Phison Electronics | All Rights Reserved.

WHAT IS GENBI

New question. No ticket. No wait.

Dashboards answer the questions you planned for. A new question usually means a request, an analyst, and a wait.

GenBI lets people ask that question in plain language. The AI writes the query and returns an answer, a chart, or a report. Your analysts define the business terms once, instead of building each report by hand. **GenBI works alongside your BI tools and uses the same databases.**

Connect. Ask. Check.

1 Connect

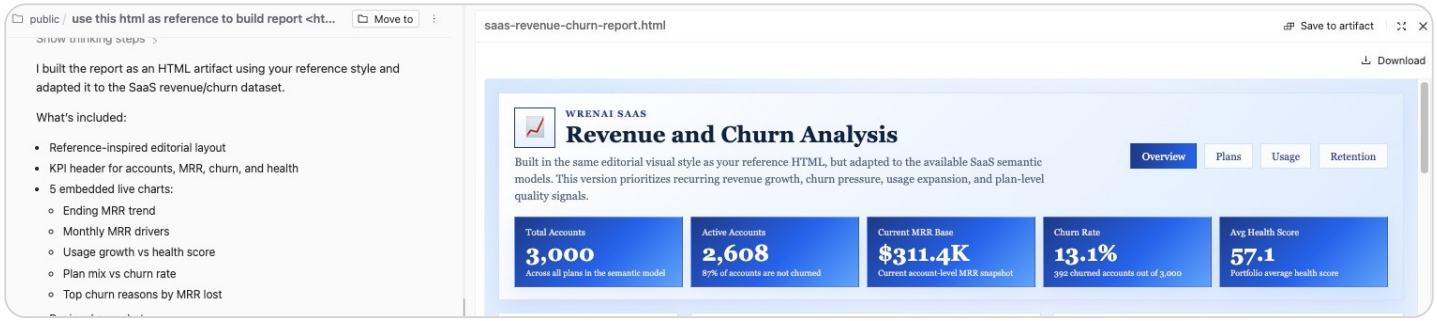
Wren AI queries your databases in place: PostgreSQL, SQL Server, Oracle, MySQL, Snowflake, and more.

2 Ask

Business users ask questions and follow-ups in plain language. No SQL.

3 Check

Get a chart, dashboard, or report. Each answer shows the query behind it.



Ask for a report in chat. Wren AI builds it from your data. Sample data.

What would your team ask?

HEALTHCARE

"Emergency visits by department and triage level?"
A chart by department.

"Which districts have declining patient visits?"
A ranked table and a trend.

MANUFACTURING

"Yield and defect rate at each temperature?"
A comparison table.

"Pressure trends in chamber B this month?"
A trend chart you can filter.

Illustrative questions. Actual use depends on your data.

Answers you can trust.

1 · Semantic model sets the scope
AI sees only the tables, columns, and relationships you define.

2 · Query validation
Invalid SQL never reaches the database.

3 · Row- and column-level security
Each person sees only the data they are allowed to see.

4 · Database permissions
Your database's own access controls still apply.

Every answer shows the tables, columns, and SQL used.

Copyright ©2026 Phison Electronics | All Rights Reserved.

SIZING

Sized for your workload. Four or eight GPUs.

Configuration	AI Server - 4 GPU	AI Server - 8 GPU
GenBI concurrent sessions	80 with aiDAPTIV (40 without)*	160 with aiDAPTIV (80 without)*
NVIDIA RTX PRO 6000 Blackwell GPUs	4, Max-Q	8, Server Edition

We size each deployment with you, by users, data, and response targets. Full specs are in the datasheet. *Test conditions on page 2.

Your data stays in your building.

Prompts, answers, and AI inference stay on your site. No cloud LLM or internet required, so it can run air-gapped. No per-token fees.

Backup and restore, built in.

Back up and restore whole workspaces, on a schedule or on demand. Send backups to storage you already have. **Add TrueNAS** to roll back a mistake or a ransomware attack to an hourly snapshot in about a minute.

TrueNAS

DEPLOYMENT

From delivery to first answers: a 14-day plan.

Our target, counted from when the hardware is on site.

Days 1–3

Rack, power, and network. Platform software and your Wren AI subscription running. Data sources connected.

Days 4–7

Wren AI reads your schema. Your team reviews tables, metrics, and access rules. Tune answers until they match trusted numbers.

Days 8–14

Pilot with a user group. Go live and hand over.

Phison leads. Partners help.

Phison

Leads delivery, setup, and support.

Wren AI

Provides the GenBI application.

Partners

Supply servers and optional storage, such as TrueNAS.

Your team

Provides data access and business terms.

WHY PHISON

We were our own first customer.

72% of Phison employees use AI every day

60+ private AI software deployments before the platform

25+ years of storage, controller, and firmware engineering

START WITH A WORKLOAD ASSESSMENT

See it run on your own workload.

Available now. Tell us what your team needs. We size the system with you and plan the rollout. Start with one team. Add more when it works.

AIDP@phison.com to get started.

www.phison.com getwren.ai/phison

Scan to learn more

Wren AI GenBI is licensed as a separate subscription. NVIDIA and RTX are trademarks of NVIDIA Corporation. Phison, Pascari, and aiDAPTIV are trademarks of Phison Electronics Corp. Wren AI is a trademark of Canner, Inc. TrueNAS is a trademark of iXsystems, Inc. All other trademarks are the property of their respective owners.

 Copyright ©2026 Phison Electronics | All Rights Reserved.