



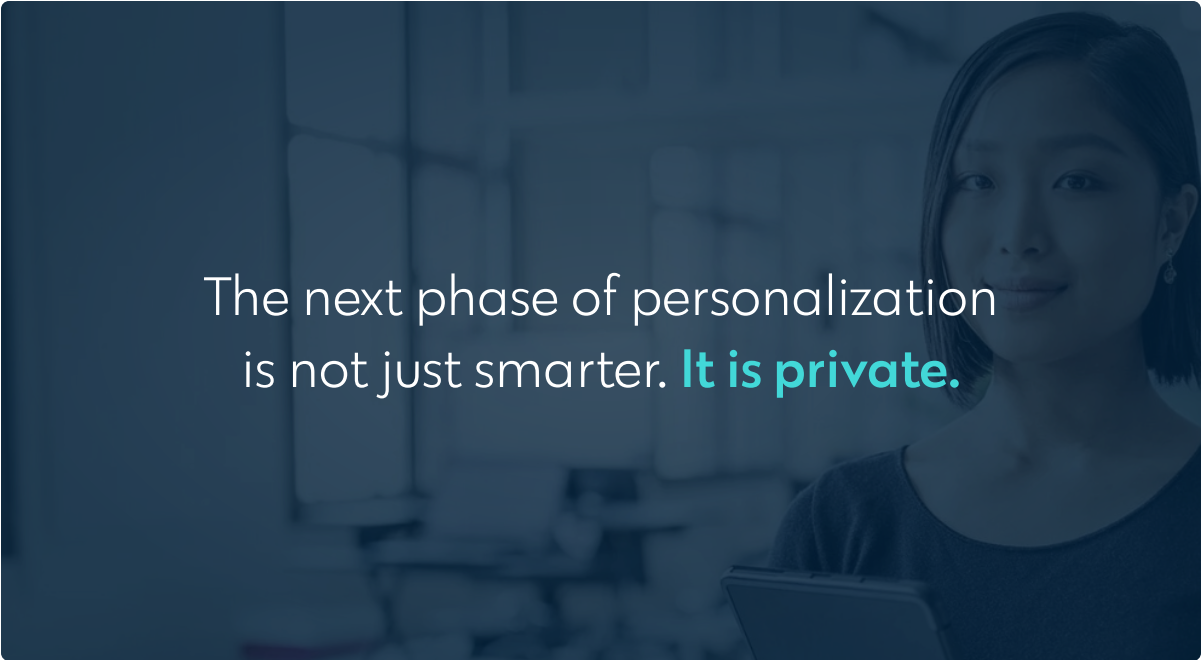
WHITE PAPER

Private AI Personalization

Delivering 1:1 experiences without sending customer data to the cloud

Inside this paper

01	The Next Phase of Personalization Is Private	3	09	What Private AI Personalization Looks Like in Practice	12
02	Why Personalization Has Become a Data-Governance Problem	4	10	Privacy Does Not Have to Mean Less Personalization	13
03	From Personalization Profiles to Real-Time Intent	5	11	The Business Case for Private AI Personalization	14
04	How RAG Changes Personalization	6	12	The Emerging Personalization Stack	15
05	Private LLMs Bring AI Closer to the Data	7	13	What Does This Customer Need Next?	16
06	Personalization Without Per-Token Cost	8	14	Before You Implement Private AI Personalization	17
07	The Private Personalization Architecture	9	15	Private AI Personalization Is a Competitive Architecture	18
08	Recommendations and Intelligent Personalization	11			



The next phase of personalization
is not just smarter. **It is private.**

The Next Phase of Personalization Is Private

For years, enterprises have been told that personalization requires collecting more customer data, building larger profiles, and sending more behavioral signals into cloud-based analytics and AI platforms.

That model is increasingly difficult to sustain.

Customers expect relevant, individualized experiences, but they also expect organizations to protect their personal information. Meanwhile, enterprises are under growing pressure to control where data is processed, how it is retained, which systems can access it, and how AI models use it.

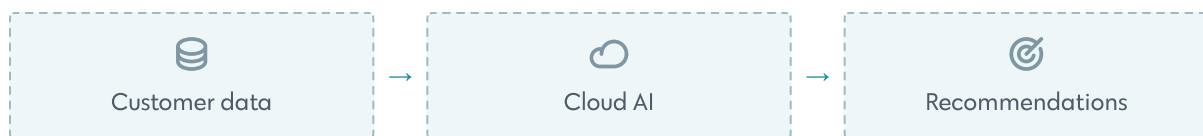
This creates a strategic challenge:

How do you deliver a genuinely 1:1 AI experience **without turning customer data into an external AI dependency?**

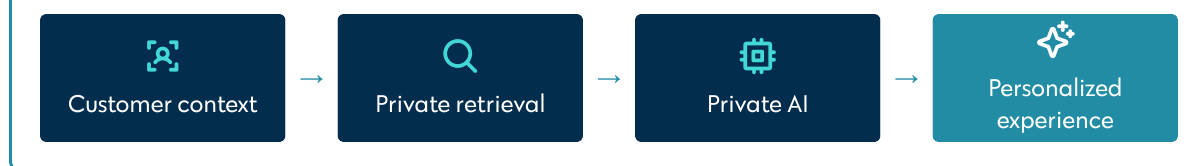
The answer is emerging through private AI personalization: combining real-time behavioral signals, enterprise knowledge, retrieval, recommendations, and private AI models inside an organization's own security boundary. Instead of sending customer context to a public AI service and asking it to personalize an experience, enterprises can keep the intelligence close to the data.

That changes the personalization model from:

CLOUD-DEPENDENT MODEL



PRIVATE MODEL – INSIDE YOUR SECURITY BOUNDARY



SearchBlox approaches this model through its broader SearchAI platform, combining search, RAG, Recommendations, Personalization, and Private LLM capabilities in a unified architecture.

Why Personalization Has Become a Data-Governance Problem

Personalization is fundamentally about context.

A retailer wants to know what a shopper is interested in. A financial services provider wants to understand which products are relevant to a customer. A B2B company wants to surface the documentation, products, or services most relevant to an account.

That context can include:

• Search queries	• Click behavior	• Pages viewed
• Products compared	• Previous purchases	• Customer segment
• Account information	• Content engagement	• Current session intent
• Product relationships	• Service history	

The more signals an AI system can understand, the more relevant the experience can become. But those same signals can be sensitive.

Sending customer context to an external AI provider introduces questions around data residency, access controls, retention, third-party processing, and organizational governance.

For enterprises operating in regulated or data-sensitive environments, privacy cannot be treated as an afterthought to personalization. **It has to be part of the architecture.**

SearchBlox's personalization approach has long emphasized anonymized behavioral data and avoiding personally identifiable information for personalization workflows, while its current Private LLM architecture is designed to keep models, inference requests, and datasets within the organization's security domain.

From Personalization Profiles to Real-Time Intent

Traditional personalization often depends heavily on historical behavior. A customer bought a laptop six months ago. They viewed several monitors last month. They previously searched for business accessories.

That information can be useful. But it may not tell you what the customer needs right now. Imagine that same customer searches:

Q Best lightweight laptop for frequent international travel

Their historical profile is useful context, but their current query introduces a much more important signal: immediate intent. This distinction is becoming central to modern AI personalization.

HISTORICAL DATA TELLS YOU

Who is this customer?

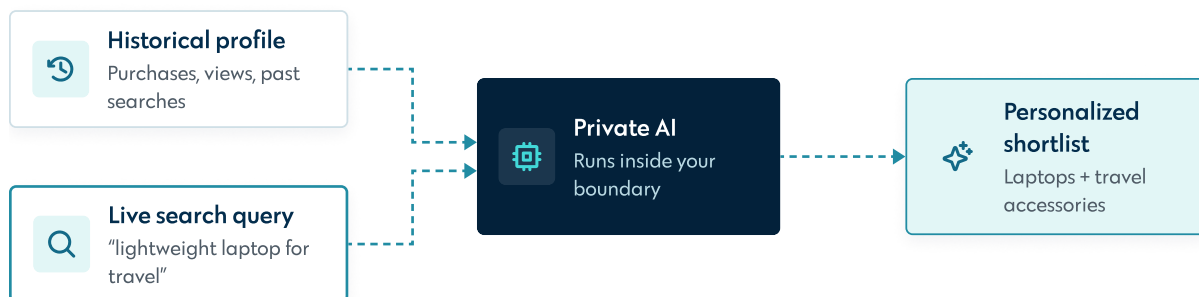
REAL-TIME INTENT TELLS YOU

What does this customer need right now?

The most useful personalization combines both.

SearchBlox's current personalization architecture reflects this shift by combining customer context with search behavior, real-time intent, enterprise knowledge, and recommendations. Its SearchAI Personalization capability uses anonymized behavioral information to build search-behavior personas and then adapts the experience as a visitor interacts with search.

HOW REAL-TIME INTENT IS COMBINED



How RAG Changes Personalization

Personalization becomes significantly more useful when the AI system can access the organization’s own knowledge. This is where Retrieval-Augmented Generation (RAG) enters the architecture.

RAG allows an AI system to retrieve relevant information from trusted enterprise data before generating a response. For personalization, that could mean retrieving:

◆ Product specifications

◆ FAQs

◆ Pricing information

◆ Warranty information

◆ Availability

◆ Knowledge-base articles

◆ Customer service policies

◆ Account-specific information

◆ Product documentation

◆ Service or support content

Instead of generating a recommendation from a generic model’s learned knowledge, the system can ground the experience in the organization’s current information.

Customer intent	“I need a laptop for international travel.”
Customer context	Frequent business traveler.
Enterprise knowledge	Current product catalog, weight, battery life, warranty, connectivity, and travel accessories.
AI output	A personalized shortlist with relevant accessories and supporting information.

This is a fundamentally different experience from simply showing the most popular laptops. SearchBlox describes RAG as a way to retrieve trusted information and provide context to an LLM before generating an answer, allowing AI applications to work against an organization’s own data.

Private LLMs Bring AI Closer to the Data

The next architectural question is obvious:

Where should the AI inference happen?

CONVENTIONAL CLOUD AI

With a conventional cloud AI architecture, customer context may leave the organization's environment before a model generates a response.

PRIVATE AI

The model can run within the organization's infrastructure or controlled cloud environment. Search and retrieval can remain within the same security boundary, and sensitive context does not have to be sent to an external model API.

SearchBlox's SearchAI Private LLM is designed around this model. Its current product positioning emphasizes private AI inference, secure data handling, deployment flexibility, and native inference that runs on your own CPUs or NVIDIA GPUs.

This also introduces an important economic consideration.

Sensitive context does not have to be sent to an **external model API**.



Personalization Without Per-Token Cost

AI economics can change significantly when enterprises operate their own inference infrastructure. Proprietary API usage scales with tokens, while a privately deployed model shifts the primary inference cost toward infrastructure capacity. Search infrastructure costs similarly depend on capacity and data volume rather than individual query counts.

For high-volume personalization, that distinction matters. Imagine an organization processing millions of search interactions, recommendation events, or AI-assisted customer journeys. A usage-priced external AI model can make every additional interaction a variable cost, and the economics of each interaction become an architectural consideration, not simply an API pricing question.

A private deployment moves the economics toward infrastructure capacity and utilization. That does not mean private AI is automatically cheaper in every environment. Enterprises still have infrastructure, operations, model management, security, and optimization costs.

But it gives organizations a different economic model:

How much will every additional AI interaction cost?



What infrastructure do we need to support our desired level of personalization?

For predictable, high-volume enterprise workloads, that can be **an important architectural advantage.**

The Private Personalization Architecture

A practical private AI personalization stack can be thought of as five connected layers.

1

Customer Context

This layer captures the information the organization is permitted to use for personalization. Examples include customer profiles, account context, previous interactions, purchase history, preferences, and relevant behavioral signals.

The important distinction is that not every personalization architecture requires sending raw identity data into the AI layer. Privacy-preserving approaches can minimize exposure and use anonymized or pseudonymized signals where appropriate.

2

Real-Time Search Behavior

Search is one of the strongest indicators of immediate intent. A customer may enter a website without a clear objective. Their first query provides a signal. The next query narrows it. Their clicks reveal what they considered relevant.

Modern personalization can continuously interpret those signals rather than waiting for a static customer profile to be updated. SearchBlox SearchAI Recommend, for example, analyzes the content a user interacts with during an active search session and dynamically adjusts recommendations based on those interactions.

3

Enterprise Retrieval

Hybrid Search can combine keyword-based retrieval and semantic/vector search to identify content based on both terms and meaning.

This matters because personalization is not useful if the underlying information is incomplete. The system has to know not just who the user is, but also what the business currently knows.

The Private Personalization Architecture

4

Knowledge Relationships

Retrieval can become even more powerful when the system understands relationships between entities. A product is connected to features. A feature is connected to use cases. A use case is connected to accessories. An accessory may be connected to an alternative or upgrade. A customer may be associated with a particular business requirement.

Knowledge Graphs provide a way of representing these relationships and adding context beyond the individual document or search result.

5

Private AI Reasoning

The retrieved context is then passed to a private AI model.

This is where personalization moves beyond a rules-based recommendation engine. The system is no longer simply asking “What products are similar?” It can ask “What is most relevant to this customer, given what they are trying to accomplish right now?”

Customer context +
current intent + enterprise
knowledge + relationships



Recommendations and Intelligent Personalization

Recommendation engines have existed for years. The real evolution is in the amount of context used to generate the recommendation.

TRADITIONAL

A traditional recommendation may be based on:
"Customers who bought X also bought Y."

ADVANCED

A more advanced recommendation might consider: "This customer previously purchased X."

AI PERSONALIZATION

"This customer previously purchased X, is currently researching Y, has shown interest in use case Z, and is now comparing products with a specific requirement. Given the current product relationships and enterprise knowledge, these are the most relevant next options."

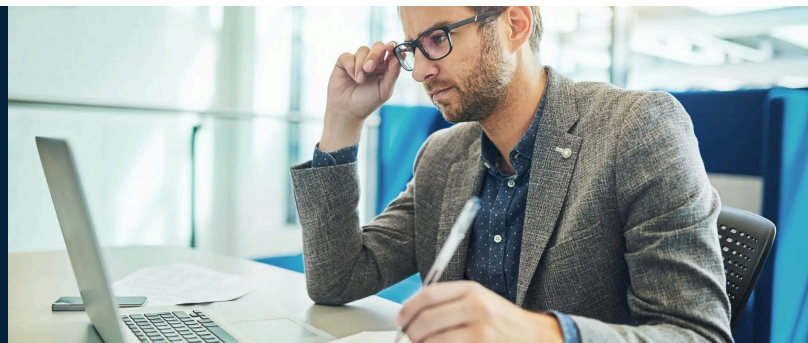
One signal

Customer history

History + intent + knowledge

SearchBlox SearchAI Recommend already follows this real-time model by adapting recommendations based on what users click during active search sessions and connecting those interactions to related content in the organization's knowledge base.

That is closer to **1:1 guidance** than static recommendations.



What Private AI Personalization Looks Like in Practice

Consider an enterprise ecommerce experience. A customer arrives and searches:

🔍 Lightweight running shoes for daily training under \$100

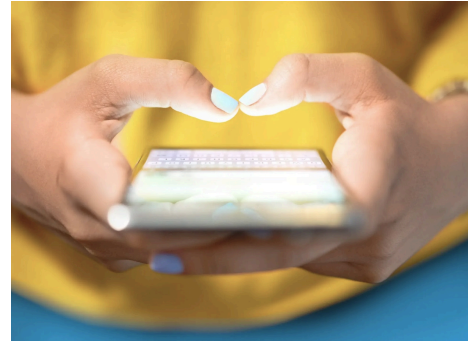
The system can interpret several dimensions of intent:

CATEGORY
Running shoes

PRICE
Under \$100

USE CASE
Daily training

ATTRIBUTE
Lightweight



Now add behavioral signals. The customer clicks two lightweight models. That behavior provides another signal. The customer then searches for:

🔍 Good for flat feet

The personalization system now has evolving intent. Instead of restarting the search experience, the AI can connect and surface products, alternatives, accessories, explanations, or comparisons that fit the evolving intent. This is the foundation of 1:1 personalization.

Current query + Previous clicks + Product attributes + Product relationships + Customer context

The experience is not personalized because the company knows everything about the customer. **It is personalized because the system understands what the customer is trying to accomplish.**

Privacy Does Not Have to Mean Less Personalization

There is a common assumption that privacy and personalization are opposing forces. They don't have to be. The architecture determines how much data has to move, where processing takes place, and how identity is handled.



Keeping AI inference within a controlled environment

Sensitive customer and enterprise information can remain within the organization's security boundary.



Using anonymized or minimized behavioral signals

Personalization does not always require raw personally identifiable information. SearchBlox's personalization offering specifically describes using anonymized user data and zero PII for its personalization approach.



Separating identity from intent

The AI layer can often work with a controlled session or behavioral representation rather than exposing unnecessary identity attributes.



Restricting retrieval

The retrieval layer can determine what information the model is allowed to access before generation takes place.



Applying governance to AI access

Private deployments allow organizations to align AI access with existing infrastructure, security, and data-governance policies.

The objective is not to collect less intelligence. **It is to expose less sensitive data unnecessarily.**

The Business Case for Private AI Personalization

For enterprise leaders, the argument for private AI personalization goes beyond privacy. It is also about control.



Lower data exposure

Customer information and behavioral signals can remain inside the organization's controlled environment rather than being routinely sent to external AI services.



Predictable AI economics

Private inference can shift the cost structure away from variable per-token pricing toward infrastructure-based economics, which may be attractive for high-volume workloads.



Greater deployment flexibility

SearchBlox positions Private LLM for deployment on private infrastructure as well as controlled cloud environments, giving enterprises flexibility in how they operationalize AI.



Easier scaling across business units

A shared private AI platform can support multiple experiences across customer service, ecommerce, employee search, knowledge management, and client portals rather than forcing each business unit to build its own AI stack.

SearchBlox positions SearchAI as a unified platform for Enterprise Search, RAG, ChatBots, AI Agents, personalization, and recommendations, reducing the need to maintain disconnected AI systems and data pipelines.

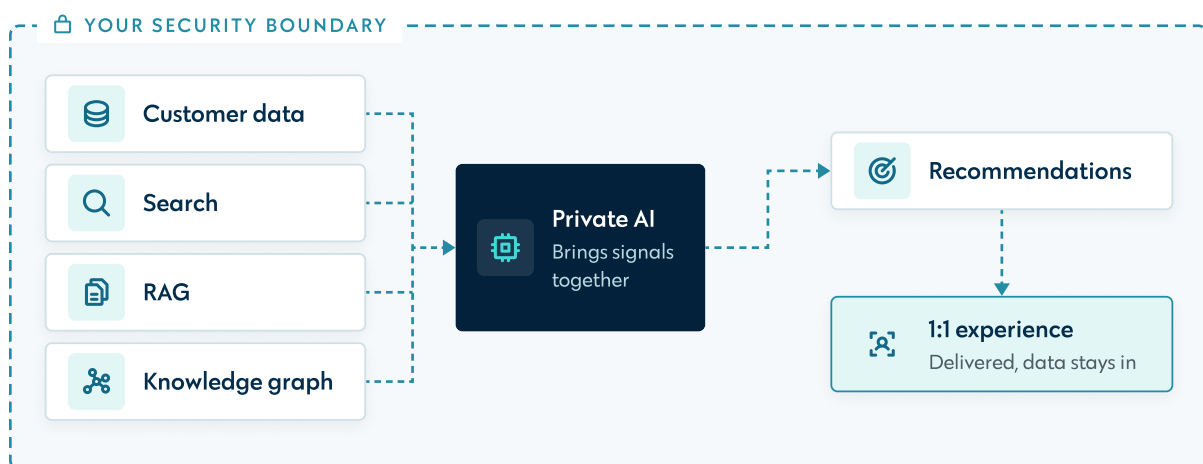
The Emerging Personalization Stack

Customer Data + Search + RAG + Knowledge Graph + Recommendations + **Private AI**

Each component contributes something different.

◆ Customer data	Provides historical context.
◆ Search	Captures current intent.
◆ RAG	Supplies trusted business knowledge.
◆ Knowledge Graphs	Explain relationships.
◆ Recommendations	Determine what should be surfaced next.
◆ Private AI	Brings those signals together without requiring sensitive context to leave the organization's controlled environment.

SearchBlox's recent perspective on the AI personalization stack follows this same direction, describing personalization as the connection of CDP, Search, RAG, Knowledge Graphs, and Recommendations rather than treating personalization as a standalone feature.



What Does This Customer Need Next?

The first generation of digital personalization was about segmentation. The next generation was about prediction. The emerging generation is about contextual decision support.

GENERATION 1

Segmentation

GENERATION 2

Prediction

EMERGING

Contextual decision support

The goal is no longer simply to assign a customer to a segment and display predefined content. It is to understand:

01 Who is the customer?

02 What are they trying to accomplish?

03 What have they already considered?

04 What information is relevant?

05 Which products or content are connected?

06 What does the enterprise currently know?

07 What should the customer see next?

And increasingly:

Can we answer all of those questions **without sending customer data to an external AI system?**



Before You Implement Private AI Personalization

Private AI is not simply a matter of selecting a model and deploying it internally. The strongest implementations start with the architecture. Enterprises should establish:

- ☐ **What data can be used for personalization?**
Not every customer signal needs to be available to the AI layer.
- ☐ **What information does the model actually need?**
Retrieval should be designed around the minimum relevant context rather than indiscriminately passing large amounts of customer data.
- ☐ **Where should inference happen?**
The answer may be on-premises, in a private cloud, or within a controlled cloud environment depending on security, scale, and operational requirements.
- ☐ **How will personalization be governed?**
Organizations need clear rules for access, retention, consent, auditing, model behavior, and data usage.
- ☐ **How will the system scale?**
The economics and infrastructure requirements should be evaluated against expected search volume, recommendation events, users, and business-unit adoption.
- ☐ **How will relevance be measured?**
1:1 personalization should ultimately improve measurable outcomes such as engagement, discovery, conversion, support efficiency, or customer satisfaction.

Start with the architecture. **The model comes second.**

Private AI Personalization Is a Competitive Architecture

Personalization and privacy are often treated as a trade-off. They do not have to be. The more strategic opportunity is to design personalization around data control from the beginning.

With private retrieval, private inference, real-time intent signals, and connected enterprise knowledge, organizations can create AI experiences that understand customers more deeply without automatically exposing that context to external AI services.

For enterprises, that creates three advantages at once:



More relevant experiences



More control over customer data



More control over AI economics

That is the real promise of private AI personalization. Not simply knowing more about the customer.

Knowing what matters to customers and acting on it **without giving up control of the data that makes personalization possible.**



SEARCHAI PERSONALIZATION | PRIVATE LLM

Start your demo.

See private, 1:1 personalization running on your own data, inside your own security boundary.

[Book a Demo](#)

searchblox.com