quantiphi
Solving What Matters

baioniq

An Enterprise Ready Generative AI Platform, powered by NVIDIA.

NVIDIA | Elite Partner

Service Delivery Partner of the Year

2024

2023

2022

baioniq™ is designed to swiftly onboard Generative AI capabilities for specific organizational tasks and overcome critical challenges encountered with Large Language Models (LLMs). It simplifies and accelerates the process of domain-adapting, and fine-tuning these models to business-specific needs, ensuring alignment with groundedness, ethical and responsible AI practices.

baioniq is powered by NVIDIA AI Enterprise and NVIDIA AI Foundation Models (including leading community - and NVIDIA-built models), leveraging the NeMo Framework for model customization, and TensorRT-LLM for optimization and deployment. The platform allows you to harness the advantages of LLMs while maintaining full control over your data and models on NVIDIA Certified Systems or on your preferred hyperscaler.

Use baioniq™ in your organization to:

Streamline content creation, and enhance images for visually appealing content.

Unlock data insights, and enable data-driven decision-making and forecasting through predictive modeling.

Automate data entry, customer support and contract redlining to enhance operational efficiency, and optimize workflows.

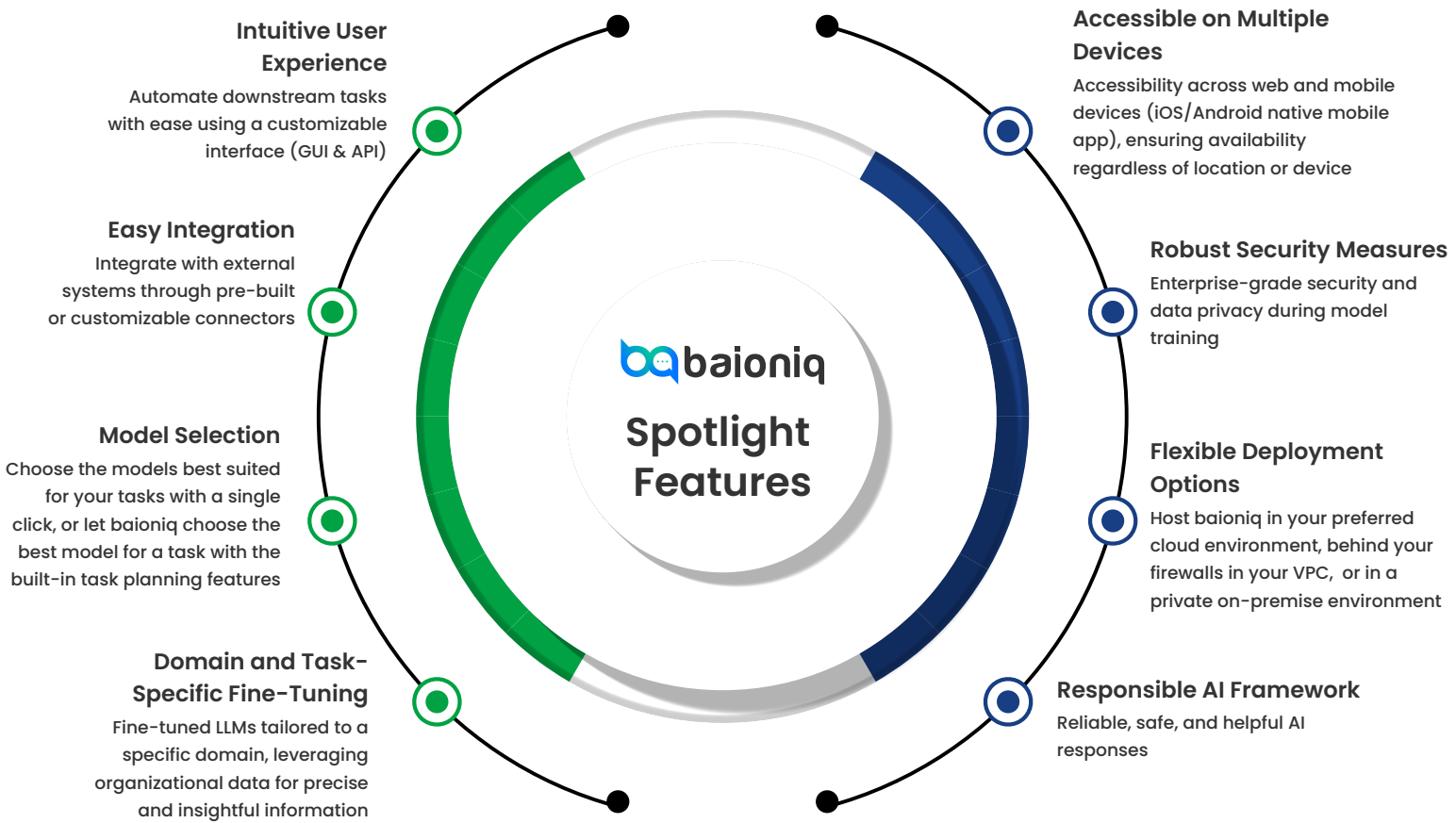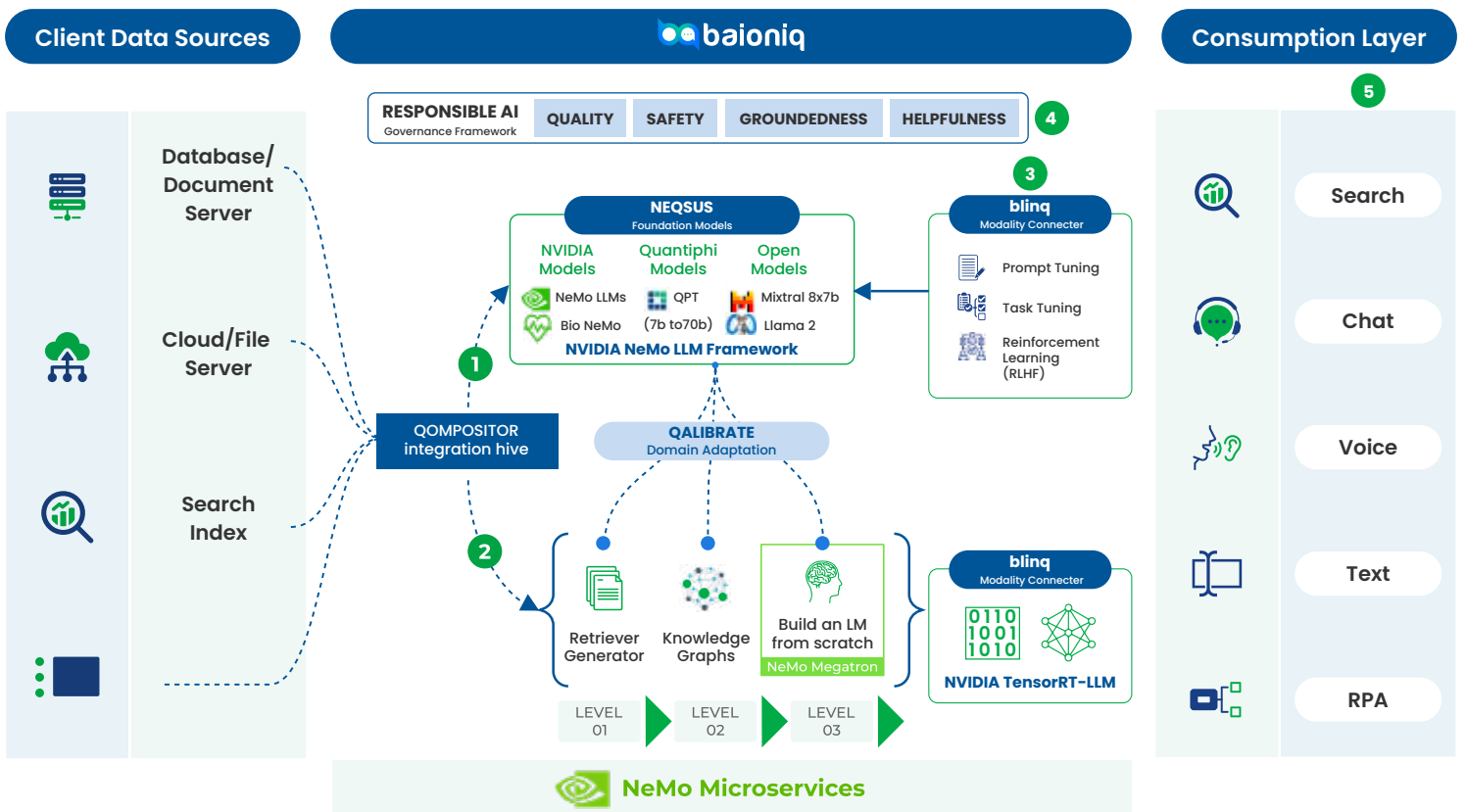Accelerate coding with auto-completion and code refactoring, improving quality and maintainability.

Generate, Summarize and Enhance Content

Analyze Data

Automate Tasks

Code Completion and Refactoring

baioniq Spotlight Features

Intuitive User Experience
Automate downstream tasks with ease using a customizable interface (GUI & API)

Easy Integration
Integrate with external systems through pre-built or customizable connectors

Model Selection
Choose the models best suited for your tasks with a single click, or let baioniq choose the best model for a task with the built-in task planning features

Domain and Task-Specific Fine-Tuning
Fine-tuned LLMs tailored to a specific domain, leveraging organizational data for precise and insightful information

Accessible on Multiple Devices
Accessibility across web and mobile devices (iOS/Android native mobile app), ensuring availability regardless of location or device

Robust Security Measures
Enterprise-grade security and data privacy during model training

Flexible Deployment Options
Host baioniq in your preferred cloud environment, behind your firewalls in your VPC, or in a private on-premise environment

Responsible AI Framework
Reliable, safe, and helpful AI responses

---

baioniq comprises a suite of modules integrated with the NVIDIA stack that empower businesses to harness the full potential of domain-specific Generative AI.

Client Data Sources

- Database/Document Server
- Cloud/File Server
- Search Index

baioniq

RESPONSIBLE AI Governance Framework — QUALITY | SAFETY | GROUNDEDNESS | HELPFULNESS **4**

NEQSUS Foundation Models
- NVIDIA Models: NeMo LLMs, Bio NeMo
- Quantiphi Models: QPT (7b to70b)
- Open Models: Mixtral 8x7b, Llama 2

NVIDIA NeMo LLM Framework

blinq Modality Connecter **3**
- Prompt Tuning
- Task Tuning
- Reinforcement Learning (RLHF)

QOMPOSITOR integration hive **1**

QALIBRATE Domain Adaptation **2**
- Retriever Generator — LEVEL 01
- Knowledge Graphs — LEVEL 02
- Build an LM from scratch — NeMo Megatron — LEVEL 03

blinq Modality Connecter
NVIDIA TensorRT-LLM

NeMo Microservices

Consumption Layer **5**

- Search
- Chat
- Voice
- Text
- RPA

Get in touch with our experts to learn more.
Email us at: nvidia.practice@quantiphi.com