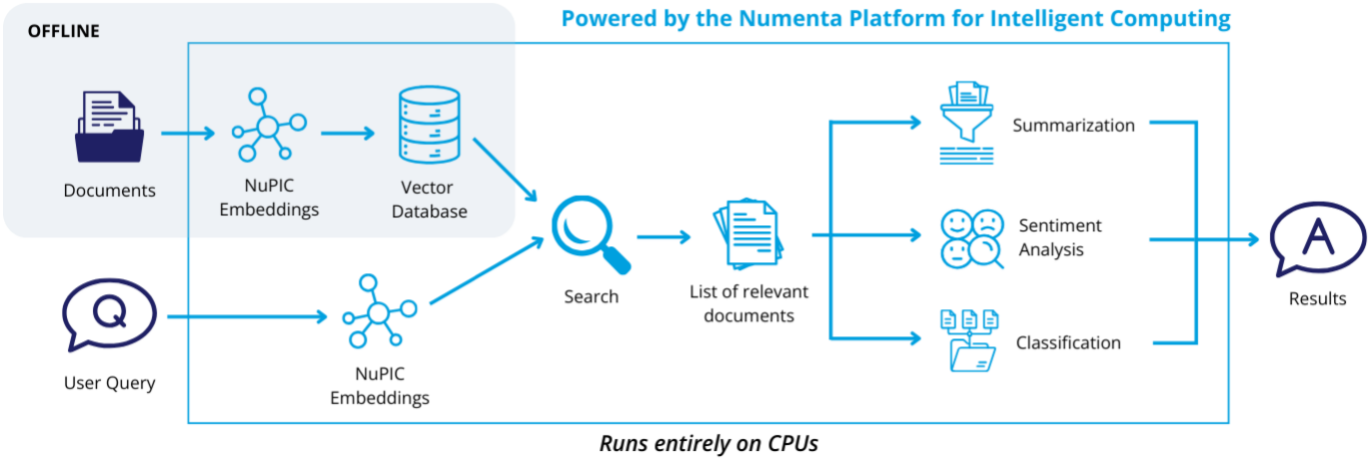
Document Retrieval – Powered by NuPIC™

As businesses scale, effectively managing an ever-growing repository of documents presents a significant challenge. As the volume of documents continues to increase, from emails and reports to legal documents and technical manuals, finding specific information quickly and accurately becomes increasingly difficult. Businesses need a robust solution that can navigate through unstructured data and extract relevant information in real-time.

The Numenta Platform for Intelligent Computing (NuPIC™) is a unique AI software platform that uses neuroscience principles to process large amounts of language data quickly and accurately. It is optimized to run large language models (LLMs) on CPUs with high throughput, low latencies, and higher accuracies than traditional models.



EXAMPLE IMPLEMENTATION

A company is building a smart knowledge base. They can leverage NuPIC models to continually index their growing collection of documents using vector embeddings. The system can then understand and automatically respond to questions by finding relevant information within the knowledge base. Additionally, they can use NuPIC GPT models to automatically summarize the retrieved information. NuPIC enables the company to harness the power of retrieval-augmented generation (RAG) and run a combination of different LLMs catered to different use cases on a single CPU server, which significantly reduces operational and scaling costs.

NuPIC Capabilities	Description
Document Similarity	Compare semantic similarities between different documents and how they relate to each other.
Document Classification	Sort and categorize documents into predefined groups based on different subject matters.
Summarization	Transform lengthy documents or articles into concise, easy-to-digest summaries.
Sentiment Analysis	Analyze opinions or emotions within text and classify them as positive, neutral, or negative.
Question-Answer	Provide accurate responses to user queries based on a knowledge base or previous interactions.

NUPIC BENEFITS

PERFORMANCE IMPROVEMENTS OVER GPUS

- **High Throughput:** NuPIC models are 4-15x faster for inference on Intel Xeon CPUs compared to GPUs.
- **Enhanced Accuracy:** NuPIC minimizes accuracy drop-off while greatly improving speed, ensuring efficient use of your compute budget.
- **Run Multiple Models on the Same Server:** You can run dozens of different models concurrently and asynchronously on a single CPU server.

PLUG-AND-PLAY SCALABILITY

- **Bring Your Own Model**: Incorporate your proprietary models into NuPIC to address specific business needs and domain expertise.
- **Seamless Integration:** NuPIC easily integrates into standard MLOps solutions such as Kubernetes.

SAVINGS AND PRIVACY

- **Delivered via Docker:** You can install NuPIC on any CPU-based server in your infrastructure, on premise, private cloud, or even on your laptop.
- **No GPUs Required:** NuPIC accelerates AI inference on CPUs, leading to advantages in inferences per dollar and inferences per watt.
- **Run in Your Private Network:** Your data and models reside completely on your private storage systems, giving you complete control over updates and data governance policies.

HOW IT WORKS

With a library of production-ready pretrained models that can be customized to a variety of natural language processing use cases, you can directly run the models in the highly-optimized NuPIC Inference Server. You can also fine-tune the models on your data with the NuPIC Training Module, then deploy your custom fine-tuned model to the NuPIC Model Library and run it in the inference server.

TECHNICAL CONFIGURATIONS

Models:	NuPIC BERT, NuPIC GPT (similar to LLAMA2)	**Operating System:**	Recent version of Linux (Ubuntu 22.04) is required to ensure kernel support for AMX
Minimum Processor:	Any Intel-compatible server with AVX512/AVX2/AMX instruction set support	**Software:**	Docker, Python 3.8 or later
Recommended Processor:	AMX enabled server such as AWS m7i.4xlarge	**Training:**	GPU with at least 12GB RAM
Memory & Storage:	16GB RAM, 200GB of storage space		

GET STARTED TODAY

Discover the power of AI-driven customer support with NuPIC. Contact us for a demo to discuss how we can tailor our solutions to meet your business needs: numenta.com/demo

Numenta

889 Winslow Street, 4th Floor
Redwood City, CA 94063

info@numenta.com

As a world leader in deploying large AI models on CPUs, Numenta has mapped its neuroscience-based advances to modern CPU architectures to redefine what's possible in AI. Numenta's AI platform NuPIC, the Numenta Platform for Intelligent Computing, helps businesses leverage the flexibility of CPUs to build robust AI applications that are efficient, scalable, and secure.