Private GPT – a Fujitsu AI solution

White paper

Second Edition February 2025

FUJITSU

intel.

Fujitsu's Private GPT AI solution is designed to provide an intelligent private chat based local knowledgebase. This on-premises solution brings state-of-the-art GenAI technology within the private scope of your enterprise, creating an environment where your specific data can be accessed in private and securely by your employees. Since all processing is done locally, the security of your enterprise data is assured.

Contents

Intel® for generative AI

The broad Intel portfolio of software tools and AI hardware is designed to make your generative and linguistic AI initiatives a success - whether you're training a model from scratch, fine-tuning an existing algorithm, or looking for a way to perform advanced inferencing at scale. Scalable Intel Xeon® processors with integrated accelerator engines are an excellent example of the power of Intel technologies: Thanks to their maximum cost efficiency, they are the first choice for training and inferencing sophisticated AI models.

The **Fujitsu AI Test Drive** runs on PRIMERGY servers with the latest Intel Xeon processors. On this platform, you can test your GenAI use case to determine the right size of infrastructure for your successful on-premises deployment.

Introduction

Generative AI is a force multiplier for business, enabling organizations to do more, faster with less, making them more competitive in the process.

Generative AI can influence every part and every level of business operations, from Design, Development, and supply chain optimization through to Sales, Marketing, HR and Legal. It has the potential to help all working environments become more energy efficient and more sustainable.

Generative AI is also a force multiplier for the individuals in your organization, working as both a consultant and an assistant to help employees in their daily tasks, enabling them to do a wider range of tasks, more proficiently and faster, and enhancing their personal productivity. This implies not just increased personal productivity but also increased productivity on a team and enterprise-wide scale.

Generative AI is rapidly becoming a catalyst for enterprise innovation and increased competitiveness. The potential uses of Generative AI to benefit organizations is limited only by people's insight and imagination.

The adoption of generative AI will also have a significant impact on a wide range of tasks and roles, encompassing not only enterprise blue-collar positions but also a broad spectrum of white-collar creative and decision-making roles. For instance, generative AI has the capability to develop software code. To fully harness the benefits of Generative AI adoption, enterprises will need to proactively invest in reskilling and upskilling their workforces.
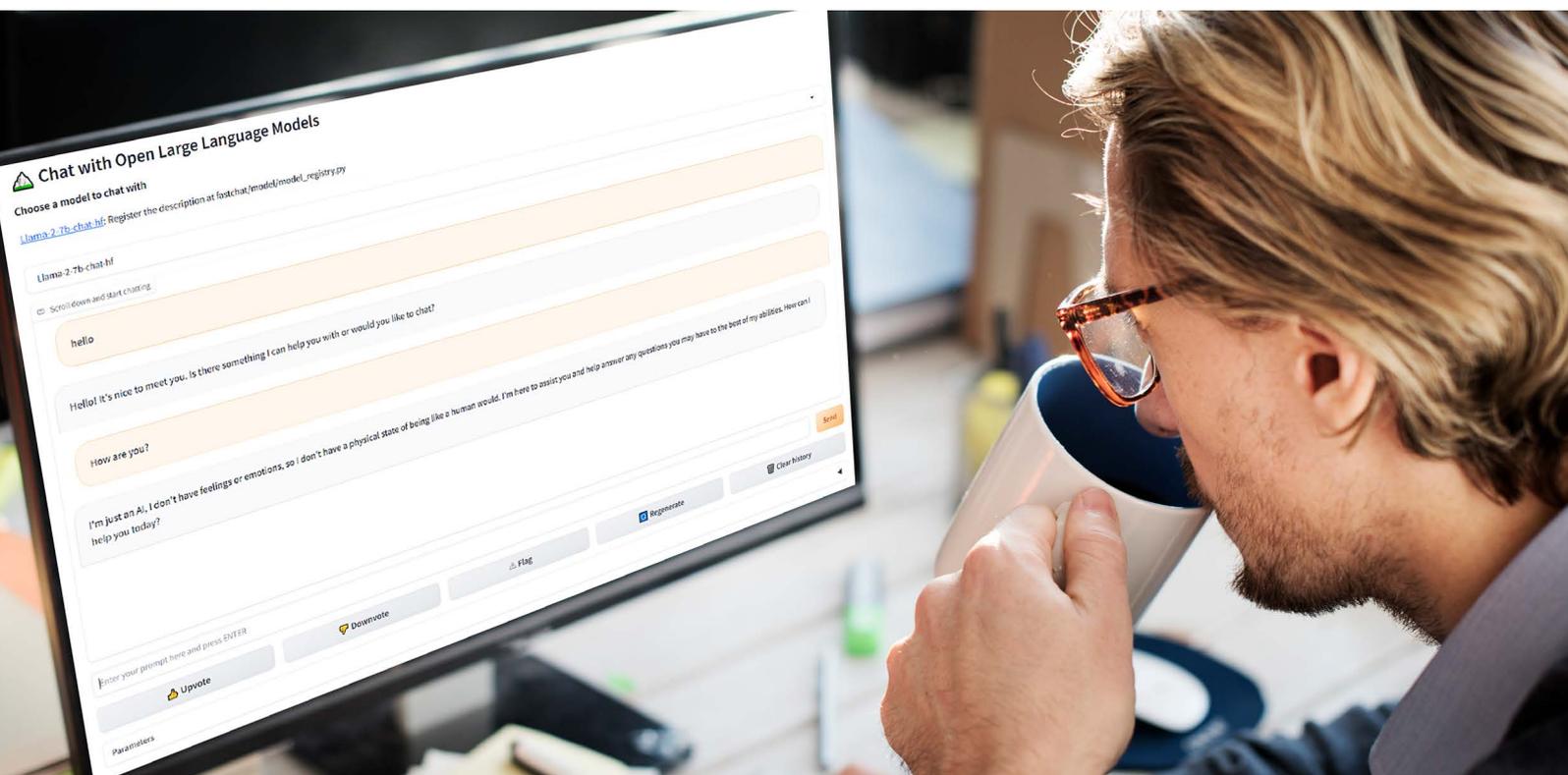
Product description

Fujitsu Private GPT builds on the success of Generative AI in revolutionizing productivity and the human machine interface. Employees can work in a natural language environment that brings a pleasant aesthetic to the way they interact with digital technology.

Public Generative AI engines harness information from many sources to provide a broad scope of general information to its users. The premise of the Private GPT solution is two-fold:

1) Enable an enterprise to load their enterprise specific data (product information, working procedures, support data, development processes, legal documents, etc.) into the GenAI environment

2) Keep the data and interactions with users private and secure, away from any public access.

For this purpose, the solution is implemented using a self-contained Fujitsu server, primed with the Private GPT software stack and ready for loading with enterprises specific data. The server is securely located on an internal and private network of the enterprise.

Components and subsystems

The solution runs on standard Fujitsu PRIMERGY servers, specifically tuned to the overall requirements. The operating system is the Fujitsu certified SUSE SLES 15 SP6. The Private GPT solution includes a web application in the form of a browser-based chat and administration client, as well as the AI and language subsystem. For better interchangeability and therefore better future proofing, containers are used to encapsulate the AI and language subsystem. The subsystems internally communicate via an API.

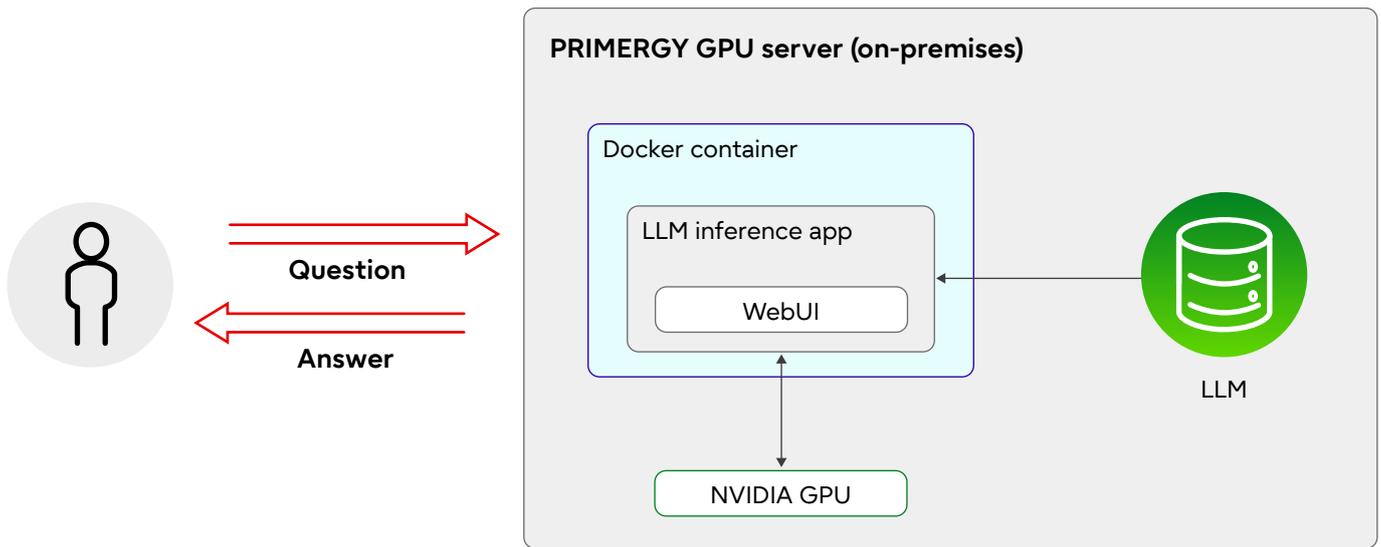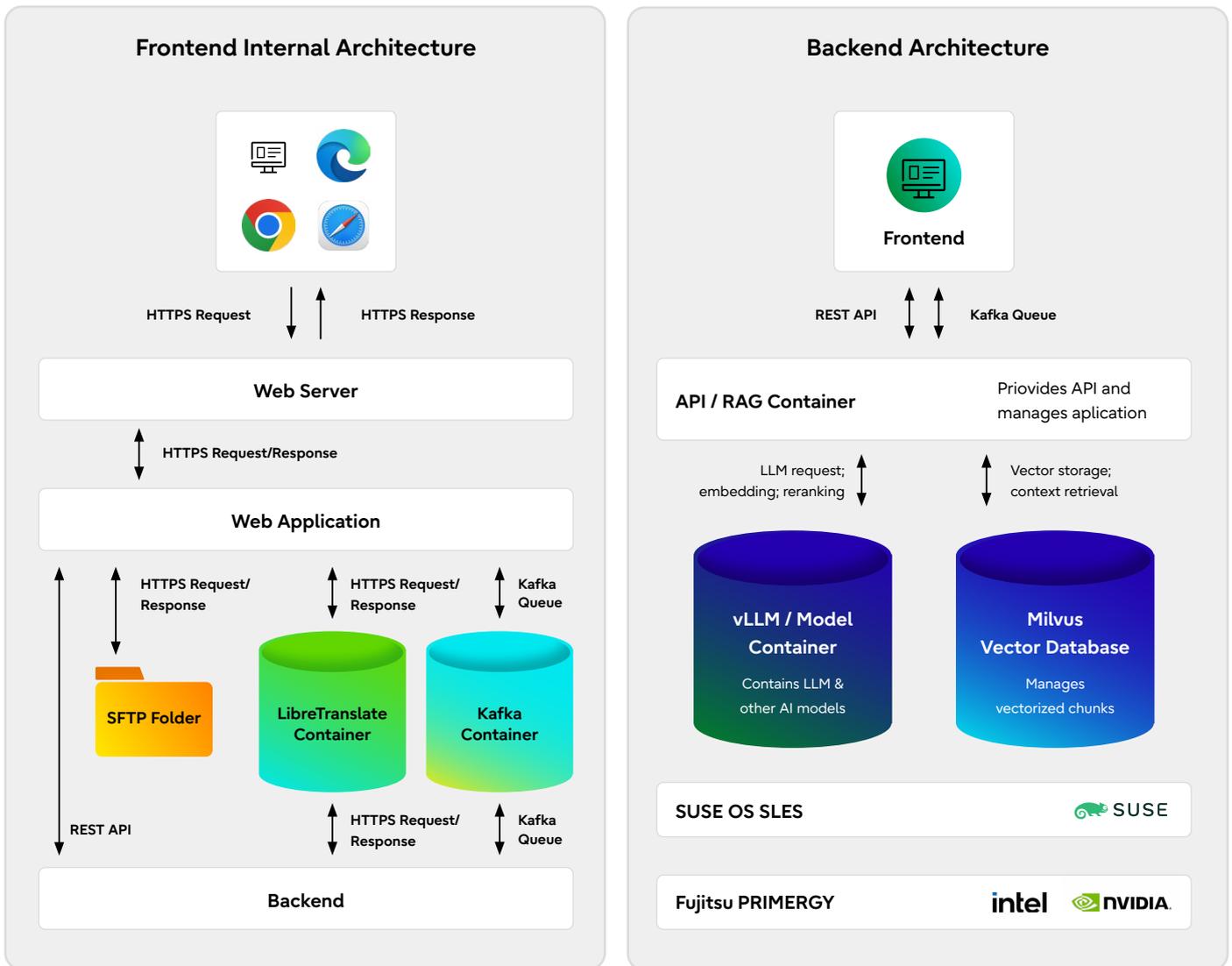**Figure 1: Fujitsu's Private GPT solution at a glance**

Figure 2: Architectural View

AI core

The AI core incorporates a RAG (Retrieval Augmented Generation) engine and a LLM (large language model). The retrieval is supported by a modified Milvus database[1] for the similarity search of dense vectors and the language model is using the Mistral AI NeMo[2] variant. This LLM variant has been trained on 100+ languages.

To ensure an absolute separation of the data groups, this task is not covered by the LLM since a corresponding dynamic ruleset may be bypassed. Instead, it is done by a specialized mechanism of the retrieval workflow. This also ensures information safety and privacy within the product. The LLM is mainly used as a human machine interface for understanding the meaning of the input question and formulating a human understandable answer based on the question in combination with the results of the retrieval.

---

1   Milvus Database: https://milvus.io/milvus-demos

2   Mistral AI NeMo: https://mistral.ai/en/news/mistral-nemo

Further information: www.fujitsu.com/emeia/private-gpt

How the AI core works

Documents provided to the system are preprocessed to extract all text content. Note that only text is considered and other content like pictures are ignored. The text itself is then divided into smaller sections by a process called dynamic chunking. The chunks are then further split up into smaller segments that are then vectorized via an embedding engine.

Any question the system receives from a user also undergoes the exact same process, which allows for a vector-based similarity search which is threshold driven and results, via a post processing stage, in the retrieval of the relevant chunks where the best vector matches could be found. These retrieved chunks are then further processed by a reranking mechanism that is ranking the chunks regarding their respective content to push chunks by pertinence. Those reranked chunks are then combined with the question along with additional commands and are then sent to the LLM.

The commands add influences on the answering behavior of the LLM to match the desired use case.

Dynamic Chunking?

Although static chunking is effective, it has inherent limitations. By dividing data based on fixed sizes or parameters, it risks splitting critical information at boundaries. This can lead to the loss of contextual connections, such as the conclusion of a paragraph or an essential link between ideas.

Dynamic chunking, also called semantic chunking, is designed to maintain a document's meaning and structure by strategically segmenting it at natural breakpoints—such as paragraphs, sentences, or thematically connected sections—instead of relying on a fixed chunk size. This ensures that each chunk preserves a coherent idea, making it far more effective for RAG models in generating responses and retrieving relevant information.

By embedding entire concepts rather than disjointed fragments, dynamic chunking strengthens the contextual foundation of a RAG system, resulting in more accurate, fluent, and contextually rich outputs.

Source: Semantic Chunking for RAG: Better Context, Better Results

Language subsystem

The LLM itself has been trained on 100+ languages including English, German, French, Italian, Bulgarian, Danish, Czech, Spanish, Estonian, Finnish, Hebrew, Hungarian, Indonesian, Polish, Portuguese, Romanian, Russian, Slovenian, Turkish and Ukrainian and many more. For languages that are not natively supported by the LLM there is an offline translator powered by Libre Translate[3] that is used to cover a broader range of languages by translating the source language to English and back at specific locations within the retrieval and speech generation process workflow.

---

[3] Libre Translate: https://de.libretranslate.com

Further information: www.fujitsu.com/emeia/private-gpt

Web application subsystem

The web application is the main user interface for the solution. It is access restricted by username and password. It provides different sections which covers an administrative, maintenance and a chat section. To gain information provided by the system the chat is available for all approved users. All configurations are covered in the administrative section which contains user management, data-groups and access information regarding Active Directory (AD), mail server and more.

The SMTP covers the local application accounts for services like password resetting.

The AD allows for easy user management and allows whitelisting of single users and of user groups. In addition, data-groups and AD-groups can be connected for easy access management.

The web application has UI language support for English, German, French, Spanish and Italian.

For further information about the web application please refer to the corresponding manual of the web application.

The web service is provided by a NGINX[4] web server listening to port 443 for web application usage. On the web application side PHP is used to cover the main functionality and a MariaDB[5] database is used for persistent storage.

---

[4] NGINX: https://www.nginx.com
[5] MariaDB: https://mariadb.org

Further information: www.fujitsu.com/emeia/private-gpt

Version 3 Features

Version 3 has significantly more features than version 1.1 and version 1.2.

- **Context Viewer:** RAG Scoring giving deeper insights on the given result.
- **Advanced Sources:** More sources that can be processed next to PDF. Markdown and most Office files.
- **Chat/User/Source API 1.0:** 1.0 of API integration for User Management, Source Management and Chat capabilities.
- **Dynamic chunking:** Depending on the file, the right chunking method is chosen to increase context relevancy.
- **UI/UX Improvements:** Collapsable sidebar.
- **Confluence Base Connector:** Connect to confluence and select the pages to be imported.
- **Performance/Stability:** Milvus Database, improved health checks, increased containerization, parallel vectorizing.
- **Hebrew/Arabic support:** Right to left support for Hebrew and Arabic.
- **Grafana Client:** Possibility to integrate with your existing grafana monitoring environment.

	V1.1	V1.2	V1.3
Local Monitoring	✓	✓	✓
Translation Engine	✓	✓	✓
Bulk Upload		✓	✓
OCR		✓	✓
Neutral Language		✓	✓
Local Backup		✓	✓
Context Viewer			✓
Hebrew/Arabic Support			✓
Chat/User/Sourcer API 1.0			✓
Dynamic Chunking			✓
UI/UX Improvements			✓
Confluence Base Connector			✓
Performance/Stability			✓
Advanced Sources			✓

More information can be found on https://sp.ts.fujitsu.com/dmsp/Publications/public/rm-private-gpt-public-en.pdf

Further information: www.fujitsu.com/emeia/private-gpt

Uploading information to the system

Custom data is provided by documents and via confluence which can either be uploaded via the web application or by sFTP for bulk uploads. The documents can be assigned to data-groups to achieve data segregation. On the filesystem the documents are separated within folders that are named according to the groups within the web portal for easy handling. The admin user defined within the installation process has sFTP access and is directed to the corresponding folder after login. The following file types can be uploaded to the system .pdf, .doc, .docx, .txt, .md, .odt, .rtf, .ppt, .pptx and .odp.

Please note that the quality of the input documents is crucial for the capabilities of the system.

Retrieving information from the system

Information is retrieved via the chat of the web application. To ensure quality output of the system precise questions covering all the context information needed to answer it must be given. Also, the language of the chat must match the language of the questions and documents in the selected groups in order to ensure a good match in the retrieved data and the generation of answers.

The Private GPT Solution now supports OCR

OCR (Optical Character Recognition) is a technology used to identify and extract printed or handwritten text from digital images of paper documents, such as those generated from scanned files. OCR systems, comprising both hardware and software, convert physical documents into machine-readable text for easier processing and analysis.

Private GPT now has Confluence API

Confluence, developed by Atlassian, is a tool for team collaboration and knowledge sharing. It enables teams to create, manage, and share project documentation in a centralized space. With strong content management features and an intuitive collaboration experience, Confluence is widely adopted across industries such as software development, IT, and product management. Private GPT is now compatible with Confluence. Users can connect to confluence and select pages that will be added to the database. Private GPT then uses information from the pages in Confluence when generating responses.

Source: [Confluence | Your Remote-Friendly Team Workspace | Atlassian](#)

Fujitsu Private GPT and Mistral NeMo

Private GPT has now incorporated the Mistral NeMo Model. Mistral NeMo: the new best small model. A state-of-the-art 12B model with 128k context length, built in collaboration with NVIDIA, and released under the Apache 2.0 license.

The model is designed for global, multilingual applications. It is trained on function calling, has a large context window, and is particularly strong in English, French, German, Spanish, Italian, Portuguese, Chinese, Japanese, Korean, Arabic, and Hindi.

Model	Mistral NeMo 12 B	Gemma 2 9B	Llama 3 8B
Context Window	128k	8k	8k
HellaSwag (0-shot)	83.5%	80.1%	80.6%
Winograd (0-shot)	76.8%	74.0%	73.5%
NaturalIQ (5-shot)	31.2%	29.8%	28.2%
TriviaQA (5-shot)	73.8%	71.3%	61.0%
MMLU (5-shot)	68.0%	71.5%	62.3%
OpenBookQA (0-shot)	60.6%	50.8%	56.4%
CommonSenseQA (0-shot)	70.4%	60.8%	66.7%

Source: [NIVIDIA](#)

With support for a 128K context length, the model demonstrates improved comprehension and the ability to process large volumes of complex information, resulting in more coherent, accurate, and contextually appropriate outputs. Mistral NeMo is trained on Mistral's proprietary dataset, which contains a significant portion of multilingual and code data, enhancing feature learning, reducing bias, and increasing its capacity to manage diverse and intricate scenarios effectively.

Performance and next features

The system allows for up to 100 concurrent requests on the large language model. The number of concurrent RAG calls is substantially lower and scales with the amount of customer data. Since the subsystems can work independently the load is covered by multiple CPU threads and GPU usage in parallel.

Next features will be, among others, the RAG history, audio to text , single document chat, so that the interaction with the chat of the web application gains a more natural feel for the users. The parser will be enhanced to efficiently extract the content of the documents.

Further information: www.fujitsu.com/emeia/private-gpt

Authentication and users

Access to the system on operating system level is granted to the local admin user that is defined during the installation. The web portal provides both locally defined users and approved Active Directory (AD) users. All users can be assigned to data-groups to provide them access to the specific information bound to those groups.

There are also roles for the web application users to allow for administrative tasks like user & group management and AD and SMPT settings, to upload PDF files to the data-groups the user has access to and to view analytics.

For further information about the users please refer to the corresponding manuals of the application.

Installation and updates

The installation is automated by SUSEs AutoYAST and requires therefore minimum interaction from the administrators. The interactions are limited to the information for an administrative account, language selection, timeserver and network settings for integration of the product into the client's infrastructure. The admin account defined during the installation will also gain initial access to the web portal and allows for the creation of other user accounts for the web portal. Also, this account enables access to the file structure that contains the uploaded documents via sFTP.

Customers under service contract are provided with updates to the system to the full extend, meaning that updates cover all levels of the system: operating system, drivers, containers, web application and AI subsystem.

Updates and in-place upgrades are provided as via web download (https://support.ts.fujitsu.com/IndexMySupport.asp). The download is then to be transferred to the server i.e. via USB-storage or via the iRMC. The update process is like the initial installation as it also uses SUSEs AutoYAST. It is therefore also guided when any interaction is required and otherwise automated.

Customer data backup and administrative tasks

The folder that contains all the uploaded files is accessible for backup via SMB access. As with the bulk upload the files can be restored there from a backup. After the restore of the files the AI synchronization needs to be triggered through the web portal. For further information about the synchronization of documents please refer to the corresponding admin manual.

Due to the RAID configurations of the system there is an additional hardware safety net. If the network settings need to be changed or the subsystems need to be restarted, the administrator can do this via an SSH connection. After logging in, the administrator is provided with a selection of possible options for executing the specified task.

Furthermore, an extra option will be introduced providing an external API to post request to the system without using the web application and therefore to allow third party applications communication with the Private GPT solution.

The roadmap is available here: https://docs.ts.fujitsu.com/dl.aspx?id=503464d0-ce35-41ca-8a3b-897993c1c492

Prerequisites

On the hardware side a 2 unit high 19" rack slot to install the server is needed. An Ethernet connection is required for access to the web portal and sFTP file upload. On the user side a web browser is needed for access the web portal. Also, the network needs to be configured to allow users HTTPS access (port 443) from their working place to the web application and sFTP (2222) for source uploads file uploads, SSH (port 4422) for admins respectively.

Private GPT now compatible with Model Context Protocol

Anthropic has introduced a protocol and framework designed to equip language models with relevant context from external systems. The Model Context Protocol (MCP), as the name suggests, outlines how to integrate various data sources—including file systems, relational databases, and code repositories—into LLMs and agents. MCP is also compatible with Private GPT. By providing an open-source framework, MCP simplifies tool integration for developers, and Private GPT Users, reducing the need for custom implementations for each new data source. Designed to work across various environments including Private GPT, MCP offers versatility and adaptability.

Limitations

- The system only supports textual input and output. Thus, questions and answers are only given in text
- The maximum size of a document is limited to 30MB when loaded via the web application.
- When uploading through bulk upload sFTP mechanism the limit is 100MB per file.
- Web browsers of Chrome, Edge, Firefox and Safari are supported.

Further information: www.fujitsu.com/emeia/private-gpt

PRIMERGY hardware blueprint

The Private GPT solution is based on the Fujitsu PRIMERGY server platforms, offering a highly reliable server environment for business-critical applications. This initial blueprint uses a PRIMERGY RX2540 M7 server with two Intel Xeon Gold 6542Y Processors and one NVIDIA L40S Ada Lovelace GPU.

This system's design is characterized by a balanced performance of GPU and CPU enabling optimal responsiveness for the GenAI engine.

Table 1 – **Hardware configuration PRIMERGY RX2540 M7**

Item	Description
System	RX2540 M7 8x 2.5 mixed for graphics
CPU	2x Intel® Xeon® Gold 6542Y 24C 2.9 GHz
Memory	16x 16GB (1x16GB) 1Rx8 DDR5-5600 R ECC
OS & Data Storage	
Controller	PRAID EP640
Disks	8x SSD SAS 24G 1.92TB RI 2.5' Non-/SED H-P
NIC	PLAN EP X710-T2L 2x10GBASE-T
GPU	
Model	NVIDIA L40S
Architecture	Ada Lovelace
Video Memory (VRAM)	48GB GDDR6
Base OS	SUSE SLES 15 SP

Configuration notes

- Operating system and data storage is based on 8x 1.92TB SSDs configured in a highly available RAID5. The minimum configuration offers approximately 12TB of storage capacity for user data.
- Additional capacity up to approx. 46TB can be accommodated in the server.
- One of the 8 disks is reserved as a global hot spare for additional resiliency.
- Network connectivity is supported via 2x 10GbE interfaces (copper)

Further information: www.fujitsu.com/emeia/private-gpt

Take the fast track to benefit from Fujitsu Private GPT in your business

Define your use cases in
co-creation with our experts

Test with your data on our secure, European AI
Test Drives

Make your proof of concept with us before you invest

Visit our website for details on Fujitsu Private GPT and AI Test Drive registration

FUJITSU

Start your AI journey with Fujitsu today: **www.fujitsu.com/emeia/private-gpt**