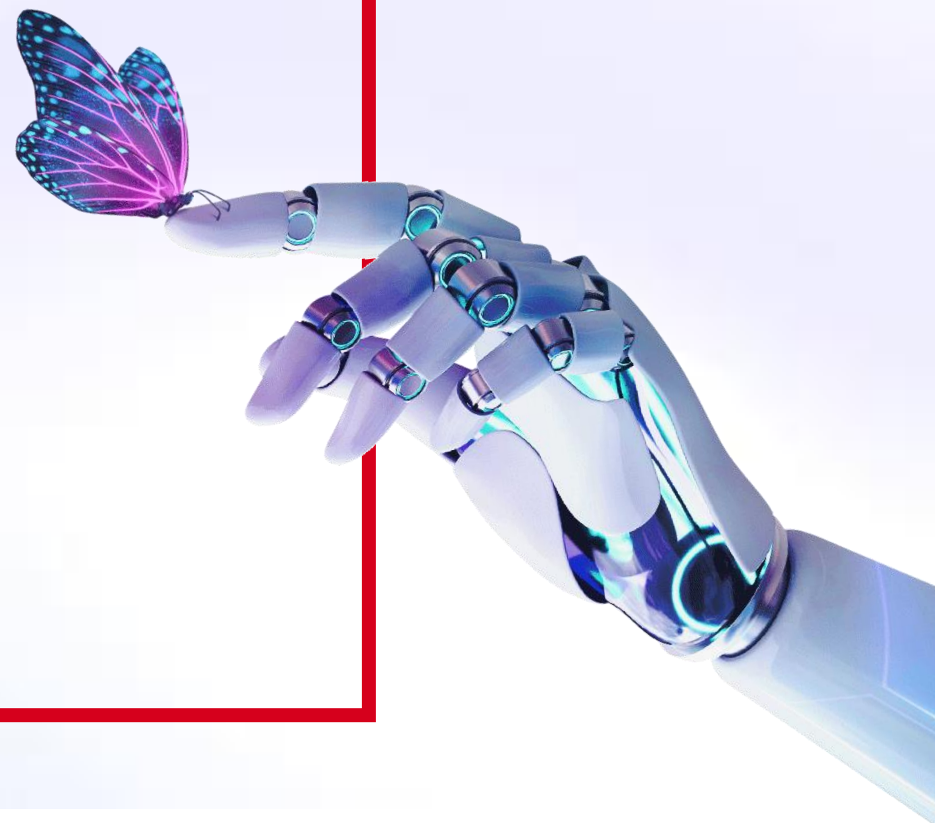
BonBon – NashTech's Generative AI Accelerator



Hai Hoang Thanh / Phi Huynh Ngoc

Kirsty McLaren / Martyna Grabowska

September 2024

Nash Tech.

Agenda

1. **Introductions**
2. **Introduction to NashTech Solutions and Gen-AI Accelerator**
3. **NashTech's Gen-AI Accelerator**
   - High-Level Architecture
   - Security & Data Privacy
   - Multi Assistant & Role-based Access
4. **Demo**
5. **Client's Q&A – Understand the Client's need**

Our Solutions

Technology advisory

- Technology visioning and transformation
- Technology strategy and roadmap
- Technology advisory and assessment
- Delivery excellence
- vCISO

Data solutions

- Data strategy
- Data governance
- Data platforms
- Data management (mesh, lake, warehouse)
- Data visualisation and analytics

AI and ML

- Decision science
- Data science and modelling
- AI platform foundations
- Intelligent products
- MLOps

Cloud Engineering

- Cloud strategy & architecture consulting
- Cloud readiness assessment
- Cloud modernisation
- FinOps

Application engineering

- Solution architecture and design
- Custom software development
- Application modernisation
- Platform engineering
- DevOps
- Mobile
- API and integration
- Low code, no code

Quality solutions

- Security services
- Independent testing
- Security testing

Business process solutions

- Business process re-engineering
- Process automation
- Business process outsourcing

Maintenance and support

- Application maintenance & support
- DevOps
- Site reliability engineering (SRE)
- CMS upgrades & integration

Our "Fusion" Solution

AI and ML

- AI platform foundations
- Intelligent products

Data solutions

- Data platforms
- Data visualisation and analytics

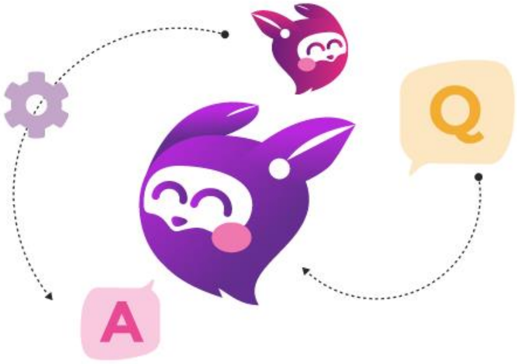
Application engineering

- Solution architecture and design
- Custom software development
- API and integration

Business process solutions

- Business Process Automation

Intelligent Generative AI Accelerator

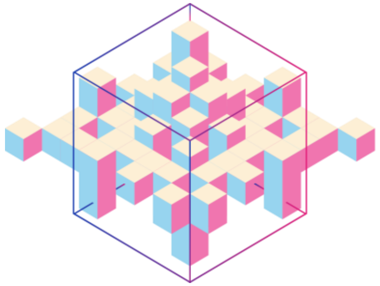


blend of cognitive technologies, intelligent chatbot, generative AI, and the ability of integration to maximize business efficiency

NashTech’s Accelerators

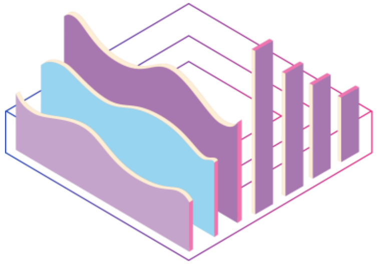
- An Accelerator is a collection of pre-built code libraries and solution templates designed to speed up the development of digital platforms, data pipelines, and AI-driven automation solutions. By leveraging extensive expertise, these accelerators expedite the implementation process, allowing developers to quickly and efficiently create robust digital solutions.

Platform Engineering



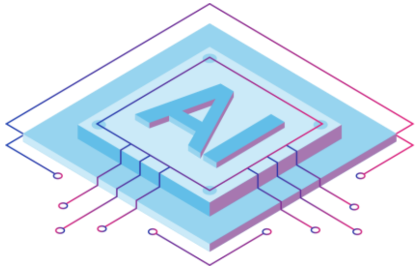
- Self-service Portal
- Architectural Library
- DevSecOps Pipeline
- Observability

Data Solutions



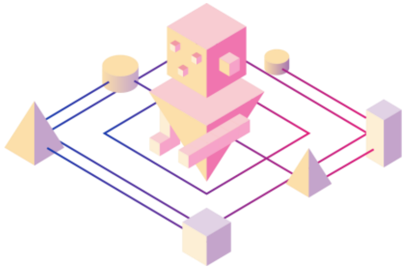
- Data Solution Templates
- Data Mesh

AI/ML



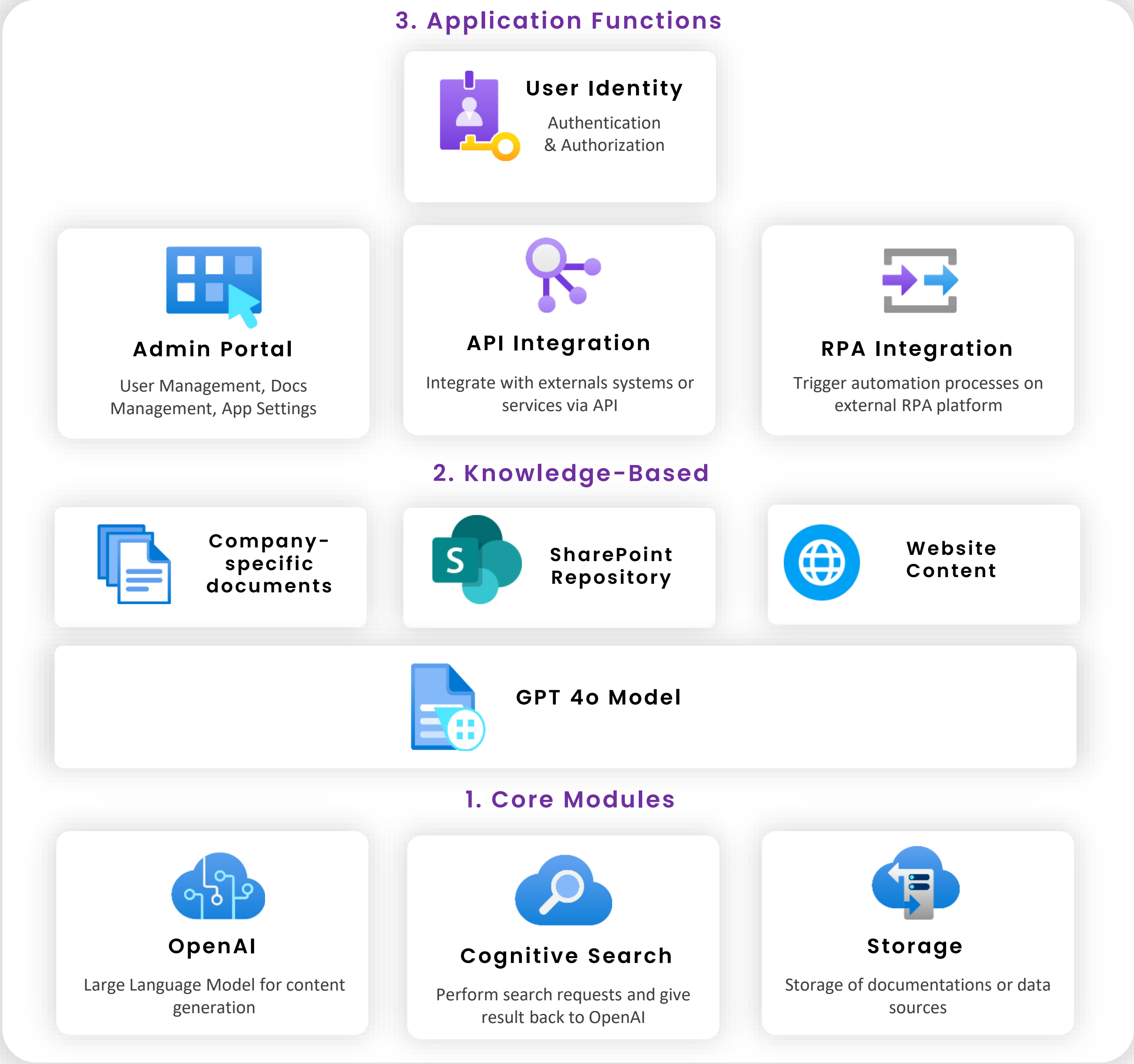
- AI / ML Libraries

Intelligent Automation



- NashAP - RPA
- BonBon – Virtual Assistant

https://accelerator.nashtechglobal.com



3. Application Functions

User Identity
Authentication & Authorization

Admin Portal
User Management, Docs Management, App Settings

API Integration
Integrate with externals systems or services via API

RPA Integration
Trigger automation processes on external RPA platform

2. Knowledge-Based

Company-specific documents

SharePoint Repository

Website Content

GPT 4o Model

1. Core Modules

OpenAI
Large Language Model for content generation

Cognitive Search
Perform search requests and give result back to OpenAI

Storage
Storage of documentations or data sources

BOOSTING YOUR USER EXPERIENCE WITH OUR

NashTech's Gen-AI Accelerator

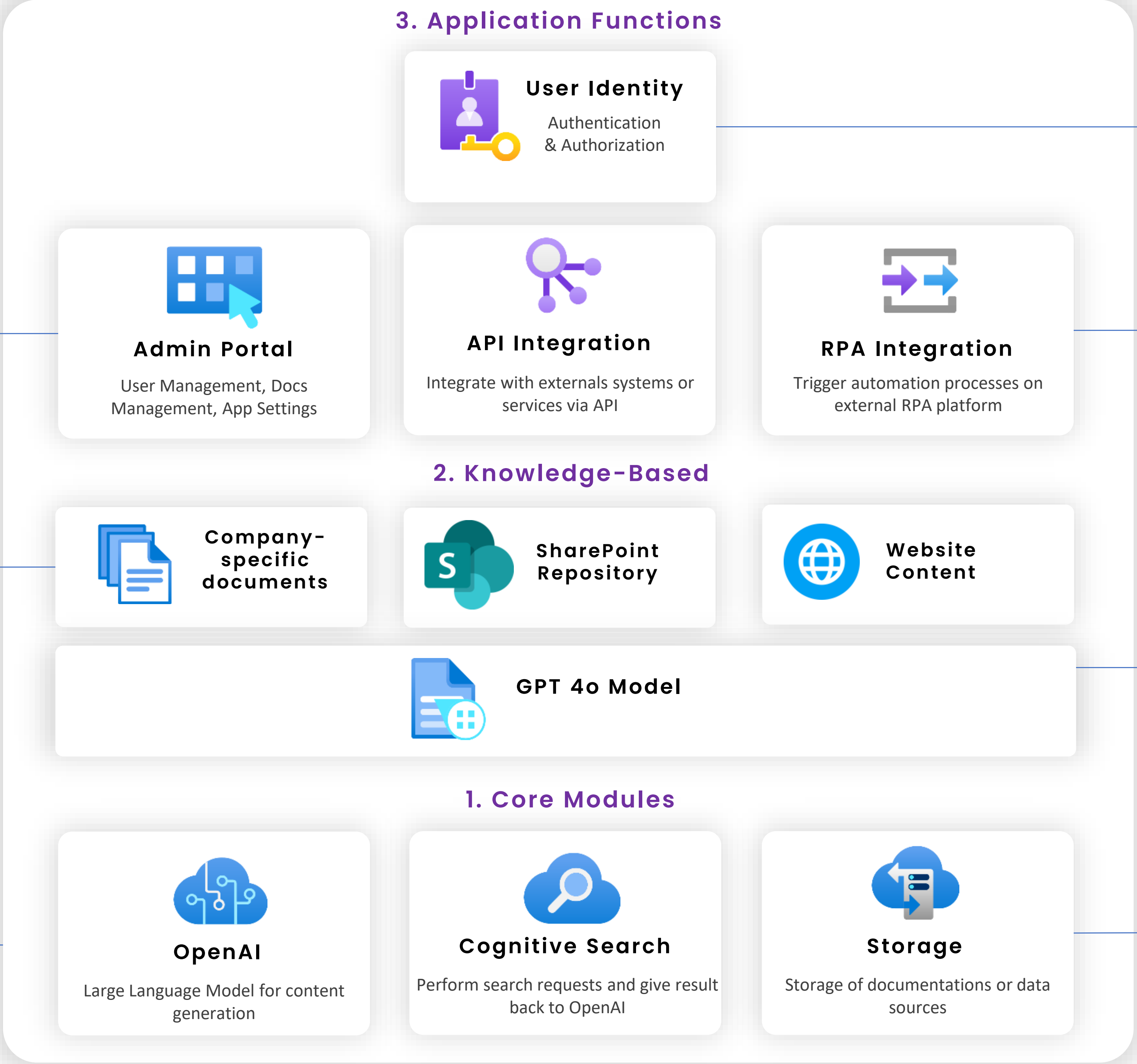
Key Benefits

- 24/7 availability
- Friendly interaction
- Multichannel support
- Multilingual support
- Self-service Admin Portal
- Scalable
- Cloud Agnostic

Key Features

- Multi-Assistant support
- OpenAI supported
- Document understanding
- API Integration
- RPA Integration
- Secured by Azure infrastructure
- SharePoint Integration
- Website Integration
- Multiple documents format: PDF, DOC, XLS etc.
- .mp3 format supported

Security & Data Privacy



3. Application Functions

User Identity

Authentication & Authorization

Admin Portal

User Management, Docs Management, App Settings

API Integration

Integrate with externals systems or services via API

RPA Integration

Trigger automation processes on external RPA platform

2. Knowledge-Based

Company-specific documents

SharePoint Repository

Website Content

GPT 4o Model

1. Core Modules

OpenAI

Large Language Model for content generation

Cognitive Search

Perform search requests and give result back to OpenAI

Storage

Storage of documentations or data sources

Integrated with **Azure Entra ID** (formerly Azure AD), and can be integrated with **any Identity providers** on demand

Using **Role-Base Access Control** (RBAC) for User Management and Settings

Using **tokens** or **secret keys** for integration

Documents stored on private **Azure Storage**: Blob or Database

LLM models can work in a **private-endpoint environment** with **no exposure to the Internet**

Can be hosted on Azure infra: **Virtual Machine**, **App Services** or **Container Apps**

Using private **Azure Services: Cognitive Search, OpenAI** under Organization management

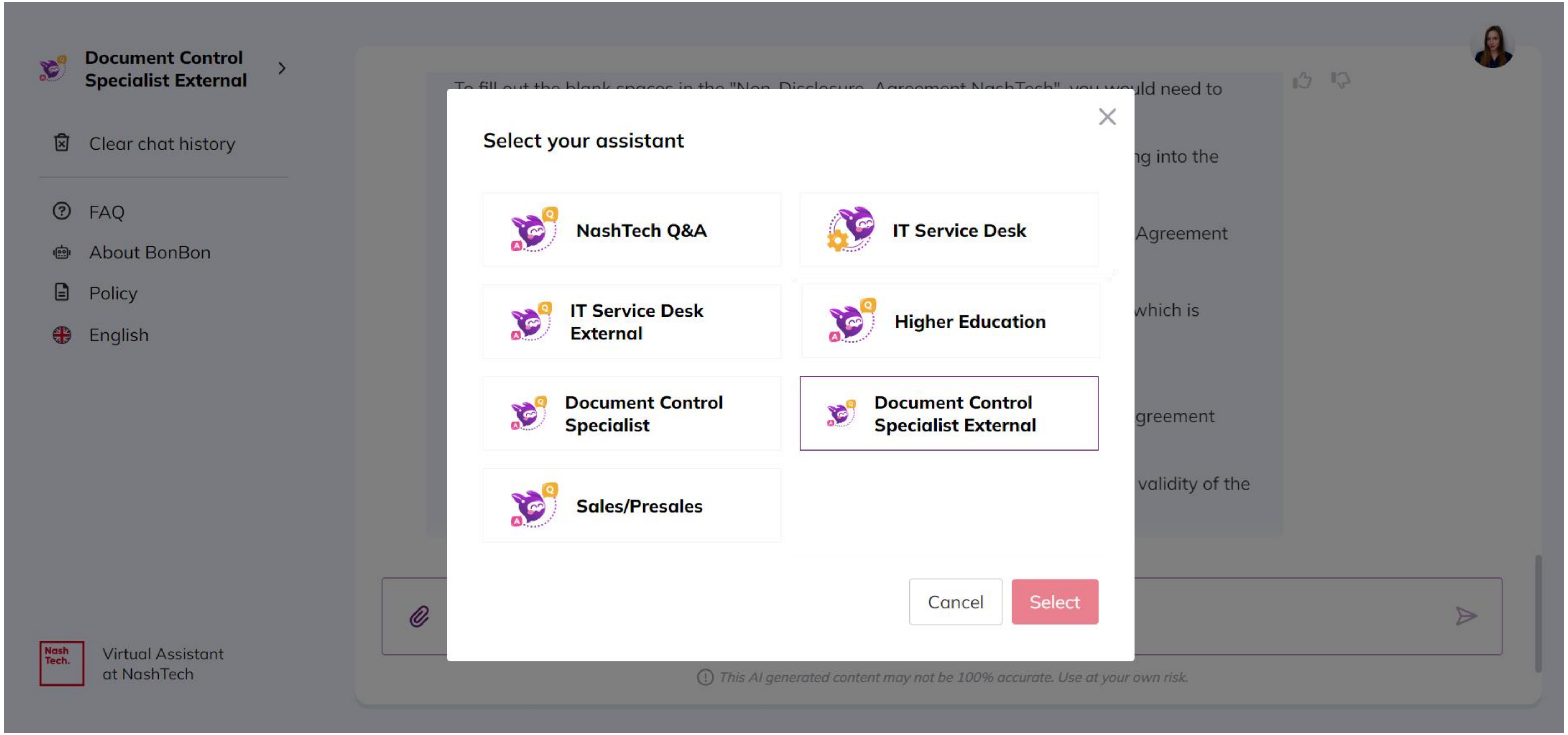
Security & Data Privacy

Your prompts (inputs) and completions (outputs), your embeddings, and your training data:

- are NOT available to other customers.
- are NOT available to OpenAI.
- are NOT used to improve OpenAI models.
- are NOT used to improve any Microsoft or 3rd party products or services.
- are NOT used for automatically improving Azure OpenAI models for your use in your resource (The models are stateless, unless you explicitly fine-tune models with your training data).
- Your fine-tuned Azure OpenAI models are available exclusively for your use.
- The Azure OpenAI Service is fully controlled by Microsoft; Microsoft hosts the OpenAI models in Microsoft's Azure environment and the Service does NOT interact with any services operated by OpenAI (e.g. ChatGPT, or the OpenAI API).

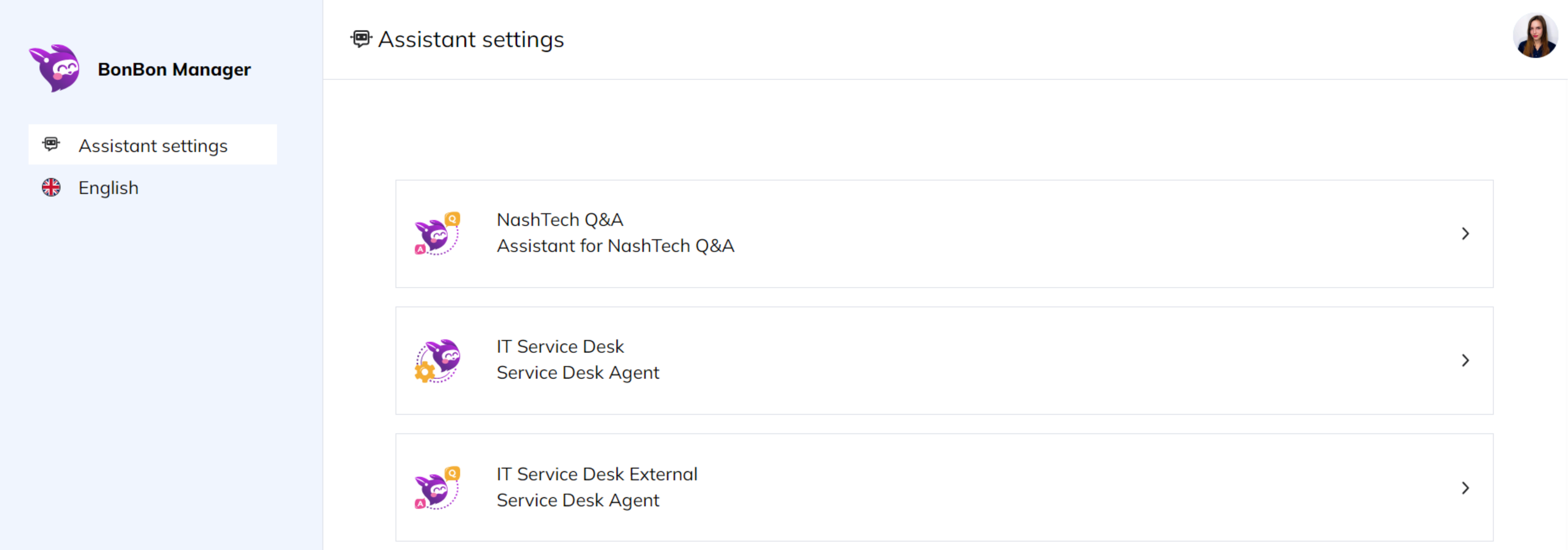
Source: [Data, privacy, and security for Azure OpenAI Service - Azure AI services | Microsoft Learn]

Multi Assistant



This AI generated content may not be 100% accurate. Use at your own risk.

Admin Portal



BonBon Manager

Assistant settings

English

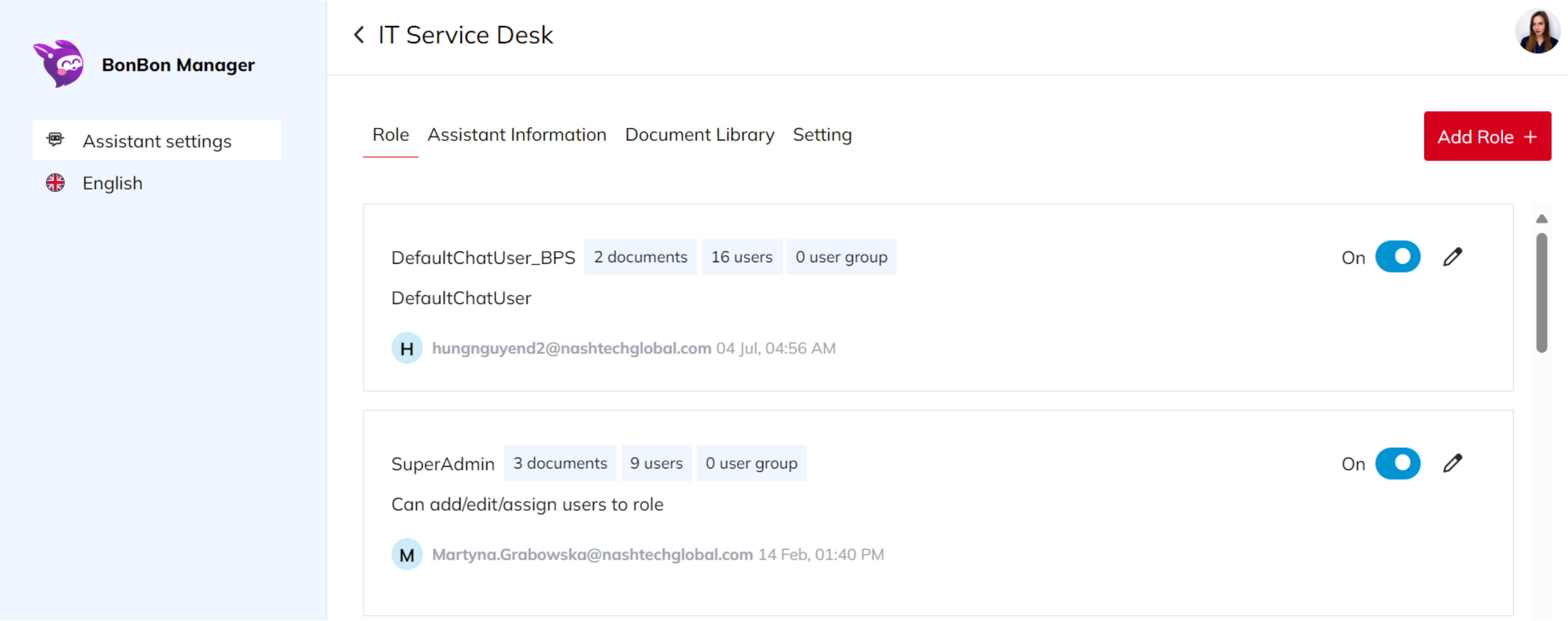
Assistant settings

NashTech Q&A
Assistant for NashTech Q&A

IT Service Desk
Service Desk Agent

IT Service Desk External
Service Desk Agent

Role-based Access



DefaultChatUser_BPS 2 documents 16 users 0 user group On

DefaultChatUser

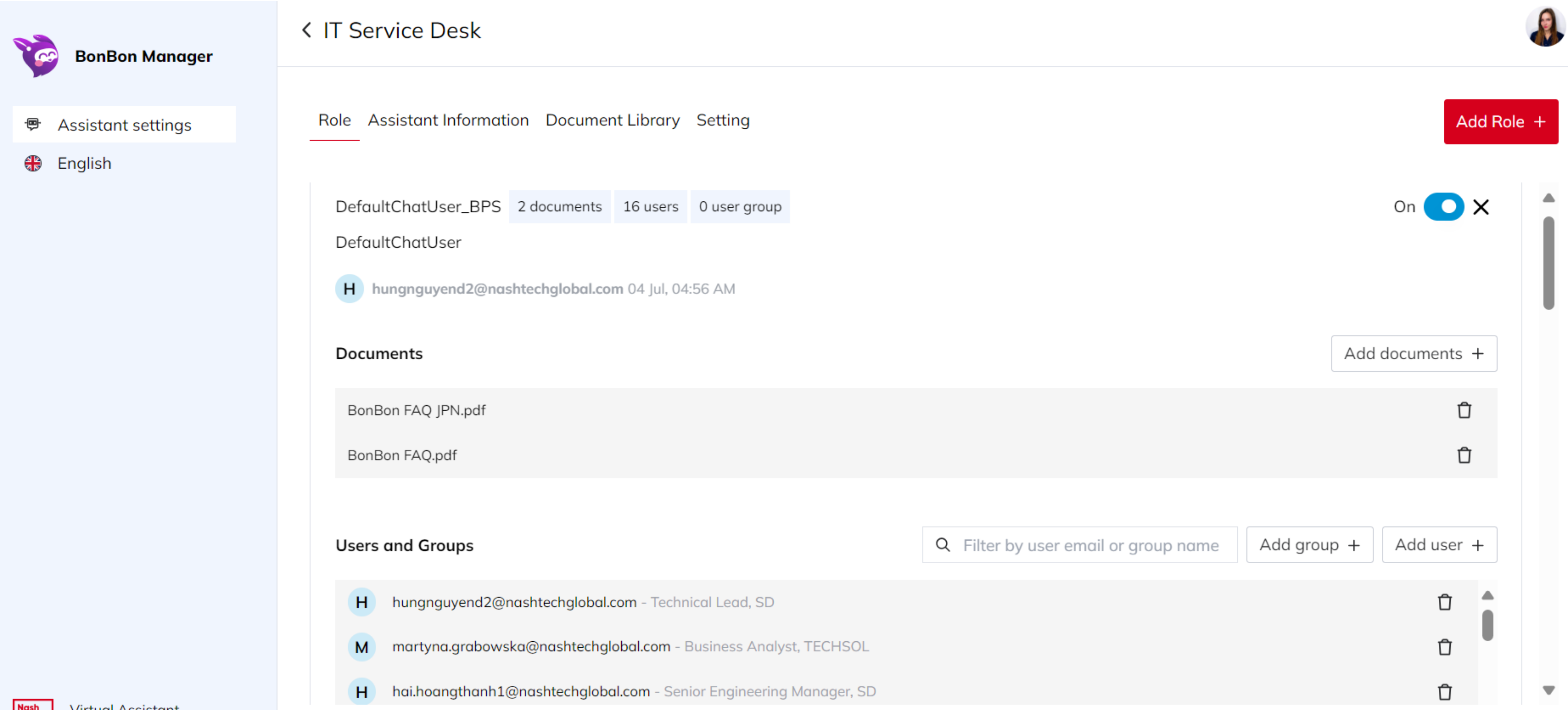
H hungnguyend2@nashtechglobal.com 04 Jul, 04:56 AM

SuperAdmin 3 documents 9 users 0 user group On

Can add/edit/assign users to role

M Martyna.Grabowska@nashtechglobal.com 14 Feb, 01:40 PM

Role-based Access



IT Service Desk

DefaultChatUser_BPS 2 documents 16 users 0 user group

DefaultChatUser

H hungnguyend2@nashtechglobal.com 04 Jul, 04:56 AM

Documents

BonBon FAQ JPN.pdf

BonBon FAQ.pdf

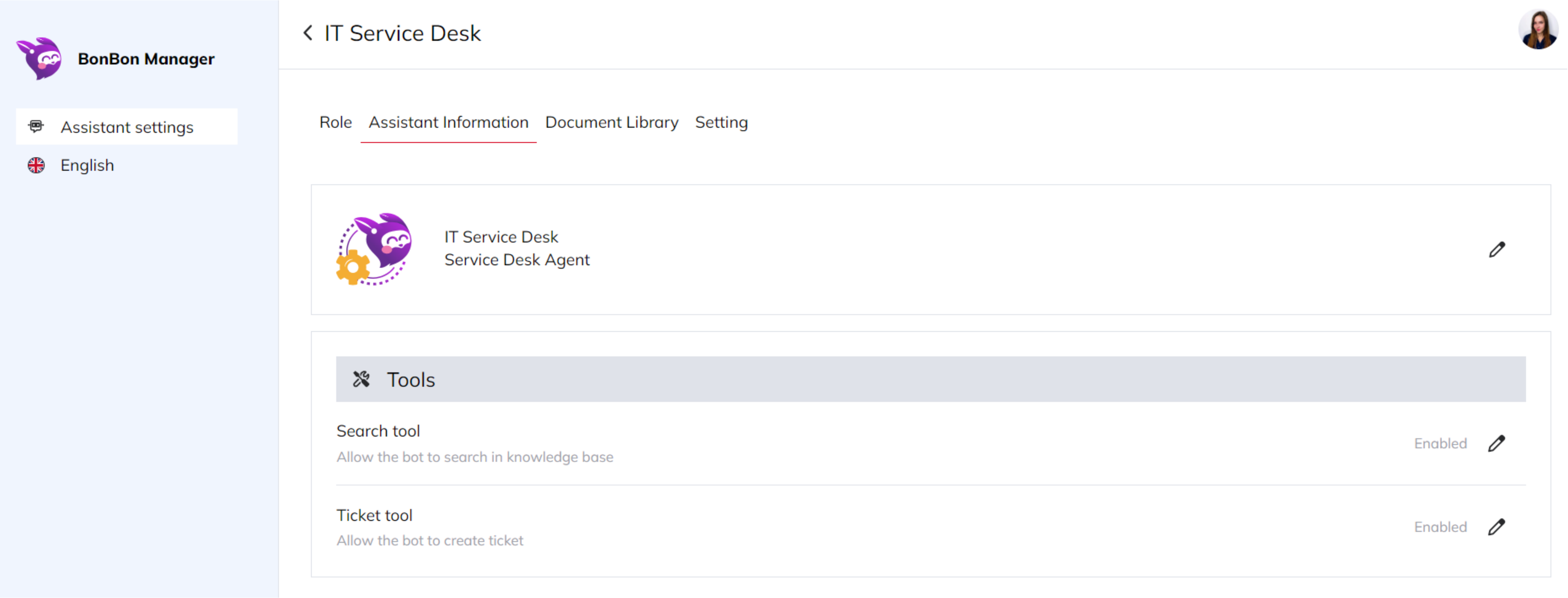
Users and Groups

H hungnguyend2@nashtechglobal.com - Technical Lead, SD

M martyna.grabowska@nashtechglobal.com - Business Analyst, TECHSOL

H hai.hoangthanh1@nashtechglobal.com - Senior Engineering Manager, SD

Assistant Information



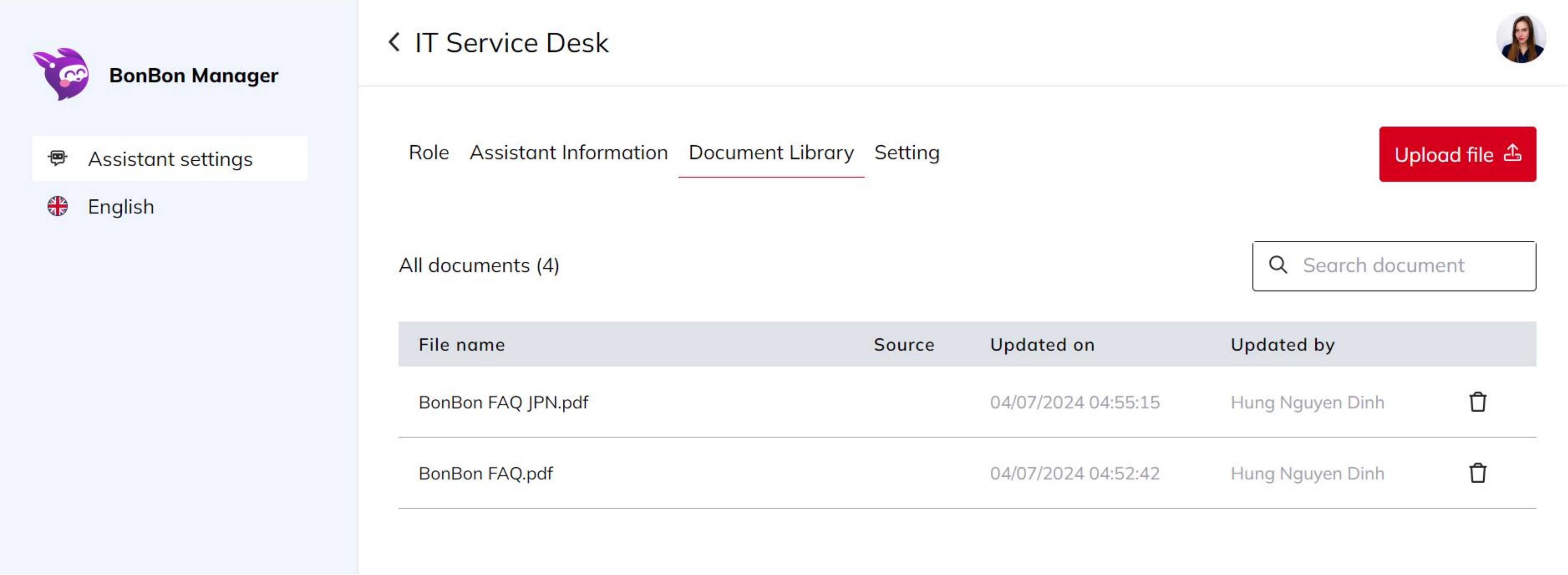
IT Service Desk
Service Desk Agent

Tools

Search tool
Allow the bot to search in knowledge base
Enabled

Ticket tool
Allow the bot to create ticket
Enabled

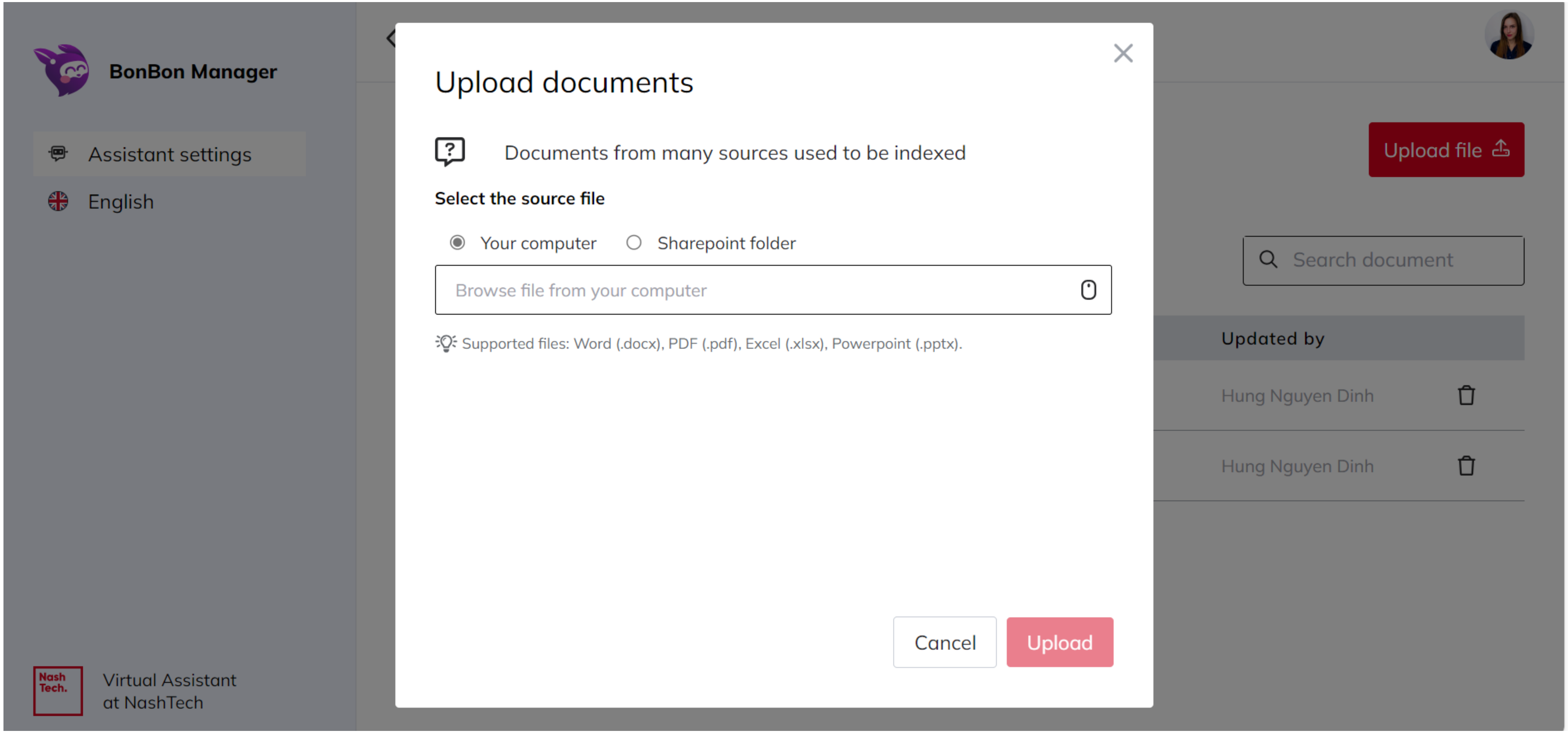
Document Library



All documents (4)

File name	Source	Updated on	Updated by	
BonBon FAQ JPN.pdf		04/07/2024 04:55:15	Hung Nguyen Dinh	🗑
BonBon FAQ.pdf		04/07/2024 04:52:42	Hung Nguyen Dinh	🗑

Document Library



Upload documents

Documents from many sources used to be indexed

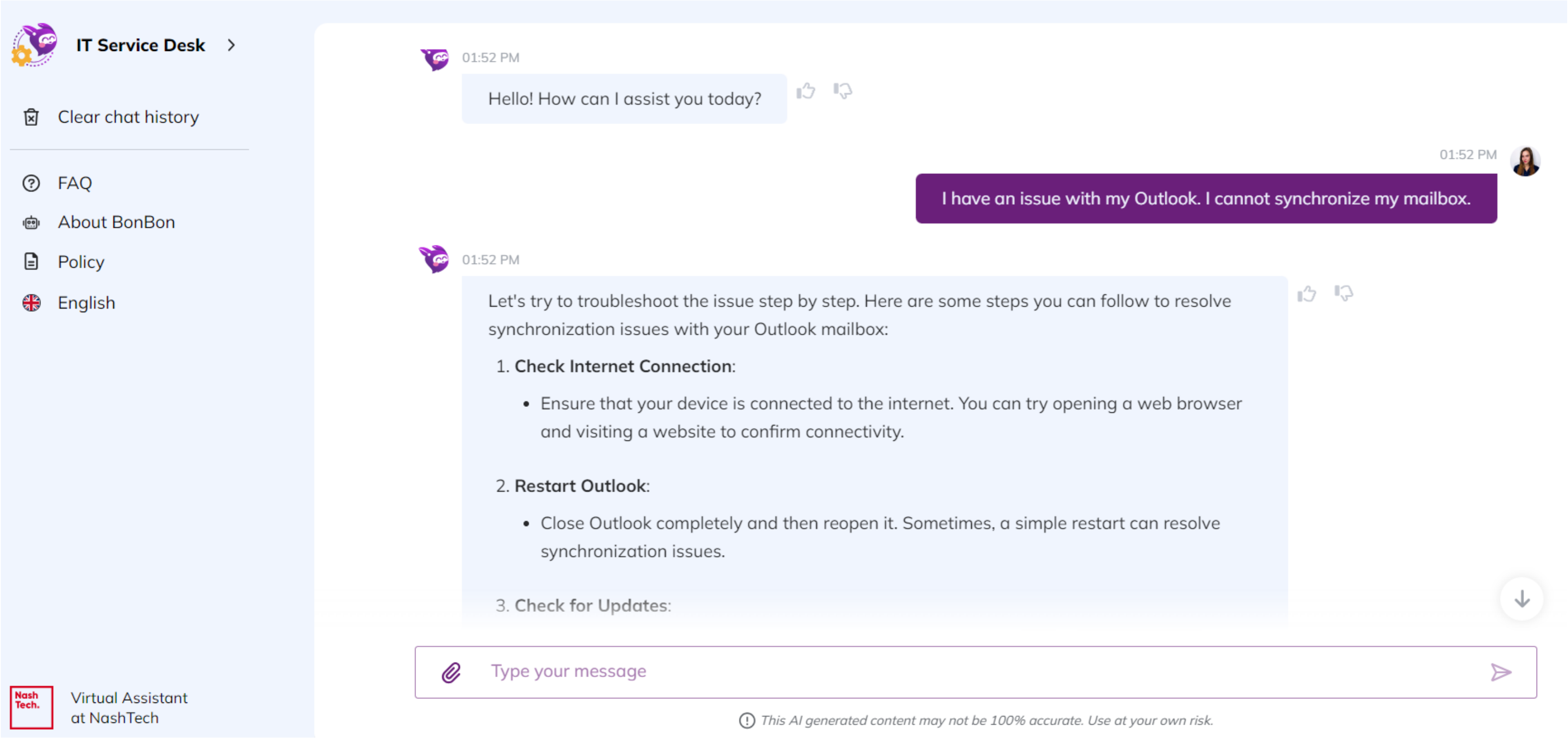
Select the source file

● Your computer ○ Sharepoint folder

Browse file from your computer

Supported files: Word (.docx), PDF (.pdf), Excel (.xlsx), Powerpoint (.pptx).

BonBon – NashTech's Intelligent Virtual Assistant



01:52 PM

Hello! How can I assist you today?

01:52 PM

I have an issue with my Outlook. I cannot synchronize my mailbox.

01:52 PM

Let's try to troubleshoot the issue step by step. Here are some steps you can follow to resolve synchronization issues with your Outlook mailbox:

1. **Check Internet Connection**:
   - Ensure that your device is connected to the internet. You can try opening a web browser and visiting a website to confirm connectivity.
2. **Restart Outlook**:
   - Close Outlook completely and then reopen it. Sometimes, a simple restart can resolve synchronization issues.
3. **Check for Updates**:

Type your message

This AI generated content may not be 100% accurate. Use at your own risk.

BonBon Marketing Video

BonBon Demo Recording

Thank you

Nash
Tech.

Appendices

Generative AI

WHAT IS GENERATIVE AI?

- Using deep learning algorithms to create new digital images, video, audio, text or code
- Those are large models which requires extreme large GPU cluster to train for billions of hyper-parameters.
- Generative AI applications are limitless, and we can use this technology to create numerous kinds of content.

RESEARCH OBJECTIVES

✓ Research foundational algorithms of Generative AI such as: **GAN**, **Transformer, Auto-encoder**, Stable Diffusions, etc.

✓ Research on using popular large models on the clouds (OpenAI, Azure, GCP, AWS) such as ChatGPT (GPT-4), GPT-3.5, DALL-E, Copilot)

✓ Build virtual assistant such first-line support, copilots

5-10%

DATA

WILL BE GENERATED BY GENERATIVE AI IN 2025

(Gartner's prediction)

Children playing on the green field, illustrator --w 770, original version

Autumn in Hanoi, original version

This image is created by Midjourney AI

Generative AI capabilities

1 **TEXT, CODE GENERATION**

AI can rephrase a paragraph to correct your input or summarize articles or news. It can also generate code for developers to boost performance.

2 **IMAGE, VIDEO GENERATION**

AI can generate the images like artists or can generate the video for based on the context. (e.g., deepfake technology)

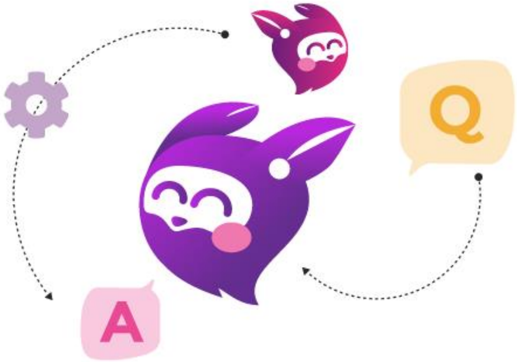
3 **AUDIO GENERATION**

Build text-to-speech AI voice generator with your own voice. You can train AI by inputting your voice samples.

Our “Fusion” Solution

- Generative AI Accelerator leveraging power of Azure OpenAI Large Language Model and Azure Cognitive Search
- Enables the rapid building, customization, and deployment of intelligent assistants within a company's cloud environment
- Primary knowledge is derived from the OpenAI GPT model, but it can be augmented with company-specific information such as policies, standard operating procedures, case studies, and training materials
- Can integrate with external systems and databases using APIs, enabling real-time retrieval and posting of information
- Web-based Admin Portal facilitating the management of knowledge bases and user access controls through a user-friendly interface
- The Multi-Assistant functionality empowers the creation of distinct chatbots with customized knowledge bases and prompts under one infrastructure

Intelligent Generative AI Accelerator



blend of cognitive technologies, intelligent chatbot, generative AI, and the ability of integration to maximize business efficiency

Benefits

- Reduced Development Time and Cost
- Increased Employee Productivity
- Improved Customer Experience
- Reduced Operational Costs
- Scalability and Security

BonBon Use Cases by Customer Industry

Software Providers	- **Product Integration:** Embedding the solution directly within their software products - **Customer Service Automation:** Developing chatbots to handle customer service inquiries and provide support
Financial Services, Legal Services, Insurance	- **Data Analysis and Summarization:** Utilizing the chatbot to assist with summarizing and analyzing large datasets for improved decision-making
Logistics & Transportation	- **Real-Time Delivery Support:** Implementing chatbots to provide customers with real-time tracking information and support regarding deliveries
Recruitment	- **Content Generation:** Leveraging the chatbot to generate new content, such as job descriptions, based on specific recruitment requirements
Retail	- **Personalized Assistant**: Supporting customers in navigating the company's product offerings, understanding their needs, and recommending the best products for them
Higher Education	- **Admissions Support**: Utilizing the chatbot to assist with admissions processes, answer frequently asked questions, and analyze personal statements to identify promising candidates
Call Centre	- **Call Transcript Analysis**: Preparing call transcripts, analysing them, and scoring based on specific criteria.
Miscellaneous	- **Service Desk Automation:** Deploying the chatbot as a first-level agent for service desks, handling initial inquiries and routing complex issues to human agents

Applications fit with BonBon

Chatbot

- Common Q&A chatbot
- IT Helpdesk First-line support virtual agent
- Chatbot for specific domain such as law-firm, taxes, e-commerce
- The chatbot that can connect to enterprise documentation with high security & data governance

Document Management*

- Generate documents based on templates
- Analyze the contracts in terms of risks, terms, compliances
- Help to review due diligence documentation
- Document comparison and summarization
- Identifying relevant documents based on prompt

Other*

- Support of proposal creation by linking with case studies and analyzing archive proposals
- Support in campaign creation by analyzing historical data on previously booked campaigns
- Automating routine tasks i.e., data entry, document processing, generating reports

Logistics

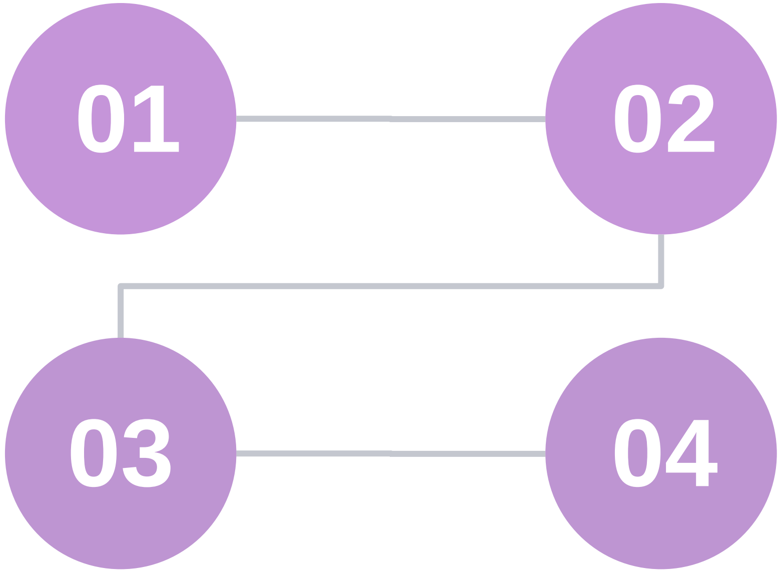
- Provide real-time assistance for tracking shipments, delivery queries or providing order status
- Manage a high volume of inquiries, freeing up human agents for more complex tasks
- Assists in optimizing routes, scheduling deliveries, managing inventory based on demand forecasting

Recruitment*

- Profile matching based on the job description
- Review the CVs with analysis
- Ranking the CVs based on metrics
- Writing the resumes based on your skillsets and experiences.
- Better searching your CV database

*Applications suggested by our customers

Customer Journey



Generic Demo

Introductory meeting to BonBon and its features. Presenting the high-level architecture, security, and demonstrating example use cases.

Individual Demo

Tailored demo using customer data (PDFs, Word documents, Excel files) to support specific use cases.

Requirement Workshops

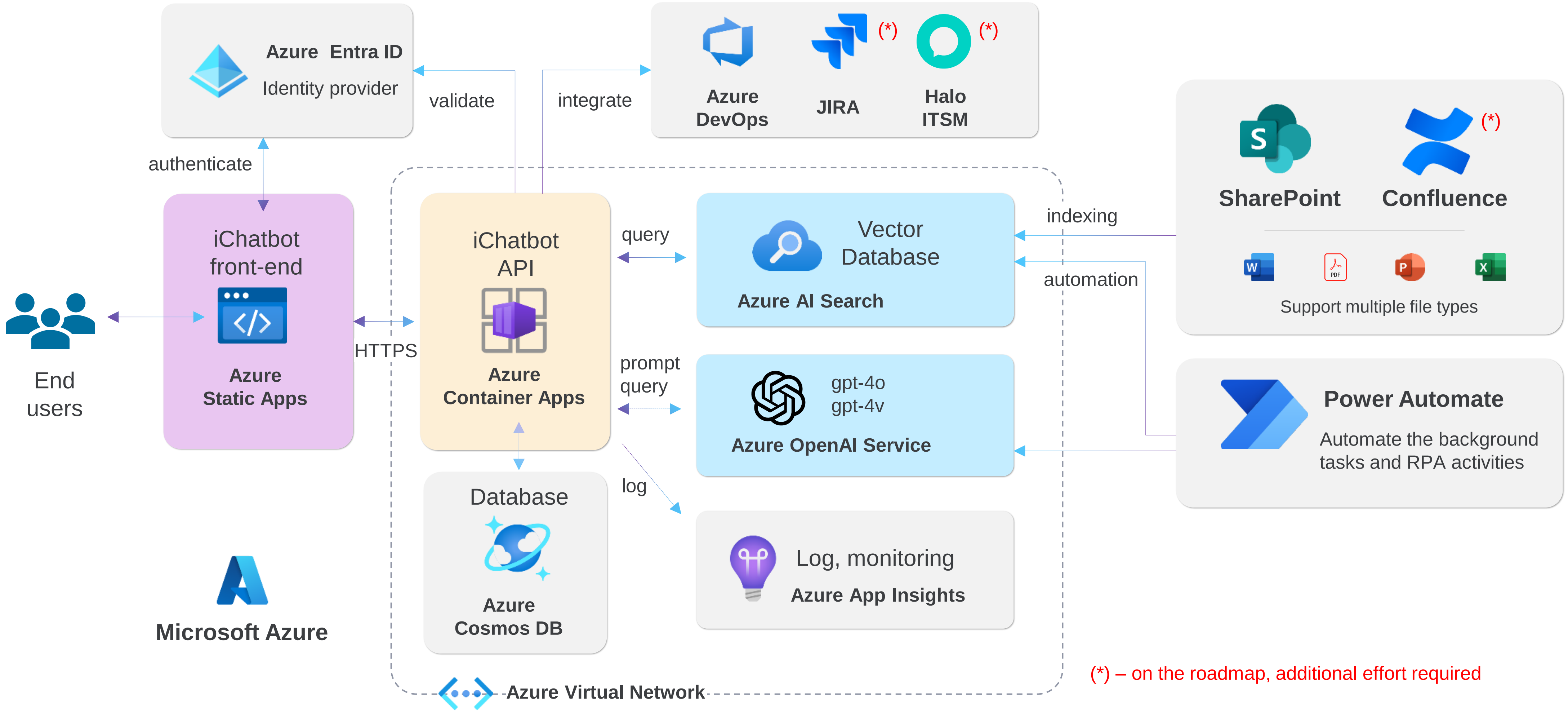
Conducting workshops to identify customer-specific requirements, use cases, details for the knowledge base, and necessary integrations.

Proposal Creation

Crafting a comprehensive proposal that outlines project details, prerequisites, timeline, and costs.

Note:
This customer journey is an illustrative example and may vary based on individual customer needs and preferences. Each engagement is unique, and our approach is flexible to accommodate diverse requirements.

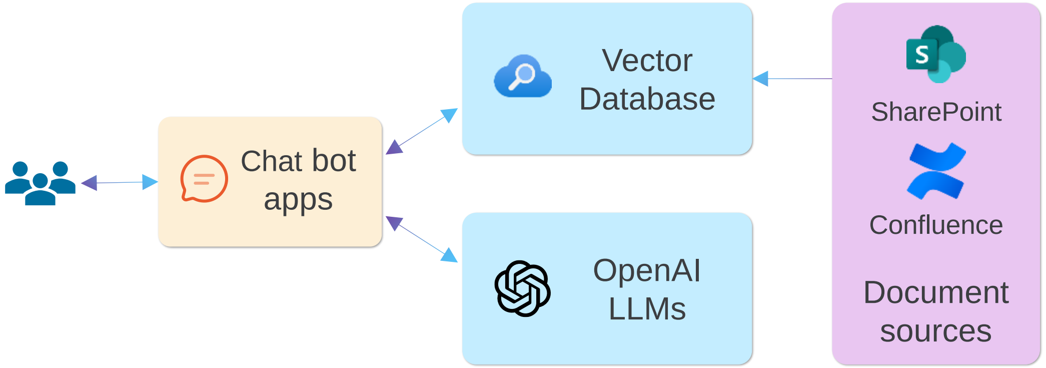
BonBon's architecture



(*) – on the roadmap, additional effort required

LLM techniques

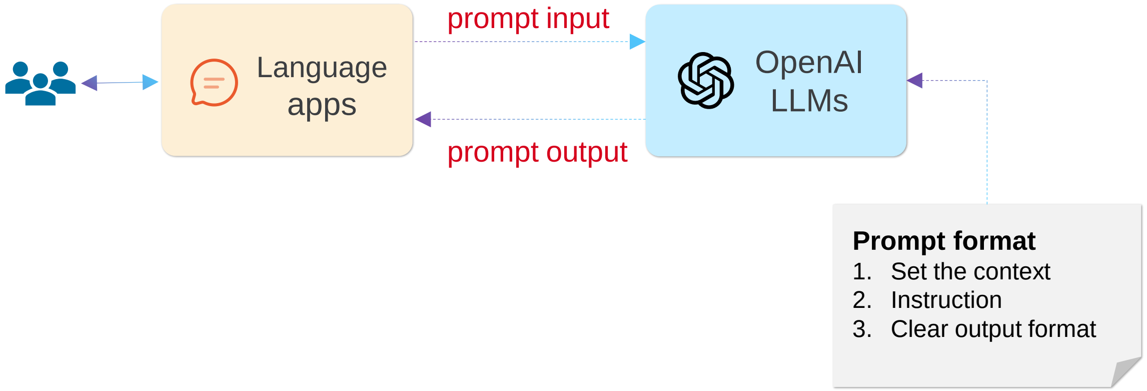
Chat with enterprise data



Approaches to connect with enterprise data

- Retrieval augmented generation (RAG) through a vector database.
- Fine-tune (OpenAI supports fine-tuning for the model `gpt-turbo-35`)
- Fine-tune open models using LoRA / QLoRA

Prompting engineering



Prompt format

1. Set the context
2. Instruction
3. Clear output format

Prompt techniques

- Zero-shot, one-shot, few-shot, …
- Chain of thought
- Tree of thought

Technology

Front-end

- JavaScript / TypeScript
- ReactJS
- TailwindCSS
- Figma (for prototyping)

Back-end

- Python 3.x
- FastAPI
- Integration with:
  - SharePoint
  - Azure DevOps
- Azure Power Automate
- Azure Container Apps
- Azure Static Apps
- Azure ComosDB
- Integrate with Entra ID (formerly Azure AD)

Generative AI

- OpenAI / Azure OpenAI
  - gpt-turbo-35
  - gpt-turbo-35-16k
  - ada-embedding-002
- Prompt Engineering
- Vector database with Azure AI Search
- Langchain framework
- Jupyter Notebooks
- Model fine-tuning

Security & Data Privacy

Secured Infra + App

- Authentication with Entra ID (with 2-factor authentication)
- Support OAuth2.0 + OpenID connect
- Authorization with RBAC-enabled (fine-grained permissions)
- Keys are protected with Azure KeyVault

Secured LLM Models

- LLM models are securely deployed on Azure OpenAI service with separated endpoint + key.
- Secured virtual network with private endpoint support for Azure OpenAI service
- No sensitive data is exposed to Internet or served for model training.

Secured prompting

- Content filtering enabled to reduce risk of harmful use
- Follow OWASP top-10 for LLMs such as prompt injection, insecure output handling, sensitive info exposure, …
- Apply best prompting design patterns to ensure expected output and allow to get user feedbacks.

Popular models

These are the most popular models (both commercial & open-source)

Large language model / services

Commercial

- OpenAI / Azure OpenAI
  - gpt-35-turbo (4k/16k)
  - gpt-4, gpt-4v, gpt4-turbo
  - ada-002 (embedding)
- Google
  - Google Bard, Palm
- AWS
  - Amazon Nitro
- Anthropic
  - Claude
  - Claude2

Open-source

- Meta
  - Llama 2
- Databricks
  - Dolly 2
- OpenLLMs
- H2O
- Finetune
  - LoRA/QLoRA

Image model

- OpenAI / Azure OpenAI
  - DALL.E 2/3
- Nvidia
  - StyleGAN
- StabilityAI
  - Stable Diffusion

Audio model

- OpenAI / Azure OpenAI
  - Whisper
  - TTS

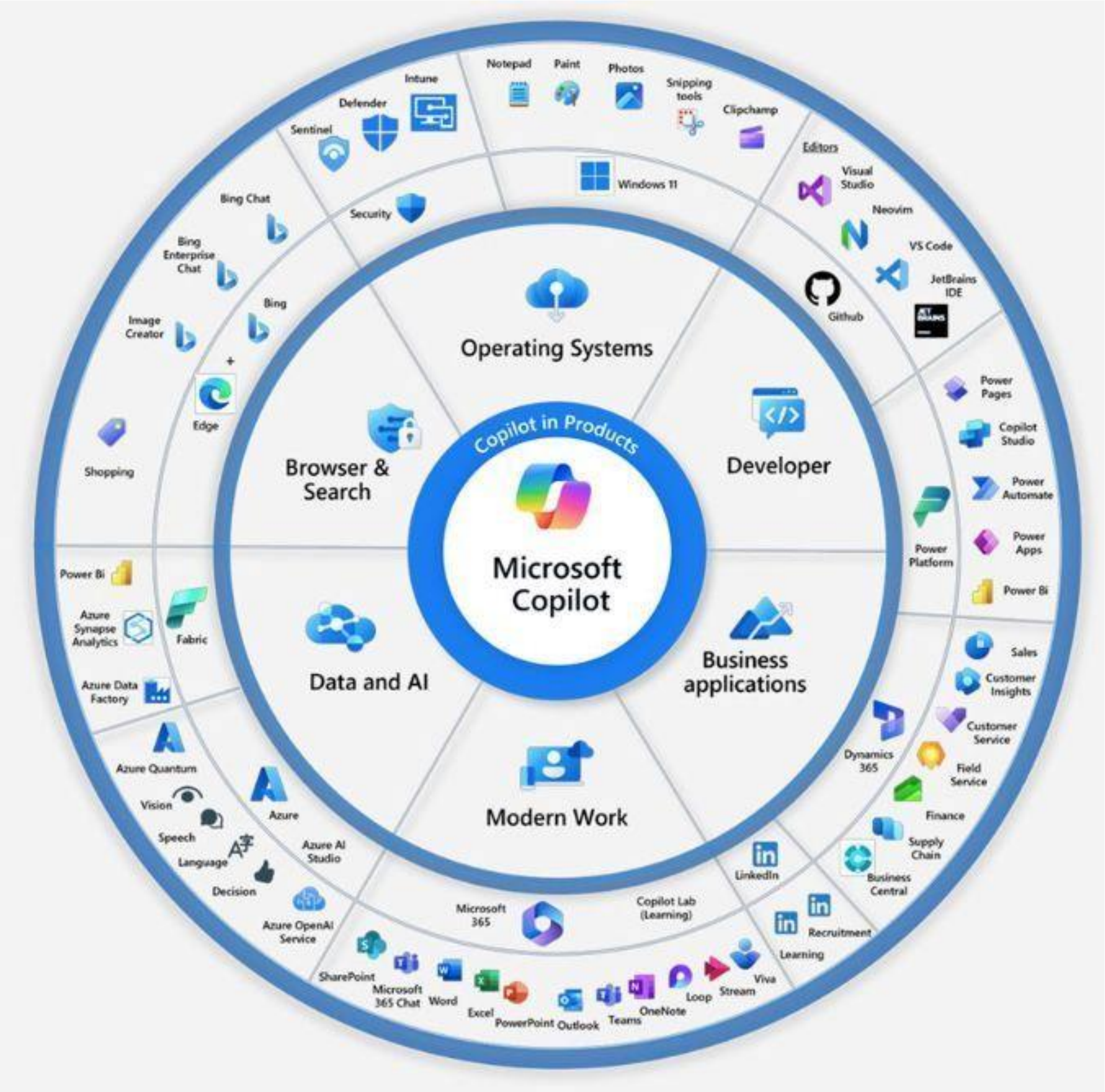
Microsoft GenAI ecosystem

Is Microsoft all-in AI?

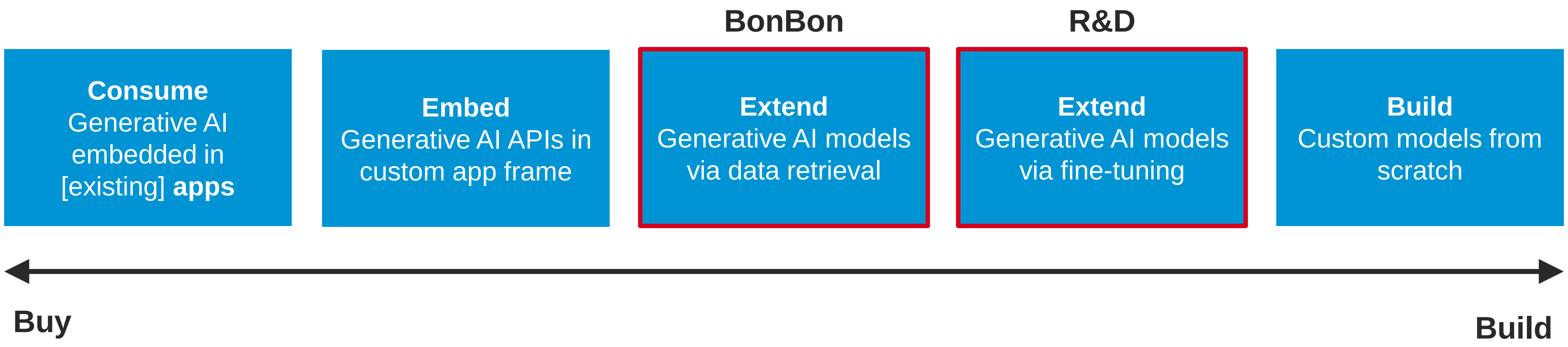
- Bing chat (https://copilot.microsoft.com)
- Bing image
- Azure OpenAI services
- Copilot for M365 (Word, Excel, PowerPoint, Teams, Sales, Azure, …)
- GitHub Copilot, Copilot X
- Azure AI Studio (preview)
- Microsoft Copilot Studio (preview)

Frameworks and SDKs

- Langchain
- Semantic Kernels
- TeamsAI library



Build vs. Buy: GenAI Deployment Approaches



BonBon

R&D

Consume
Generative AI embedded in [existing] **apps**

Embed
Generative AI APIs in custom app frame

Extend
Generative AI models via data retrieval

Extend
Generative AI models via fine-tuning

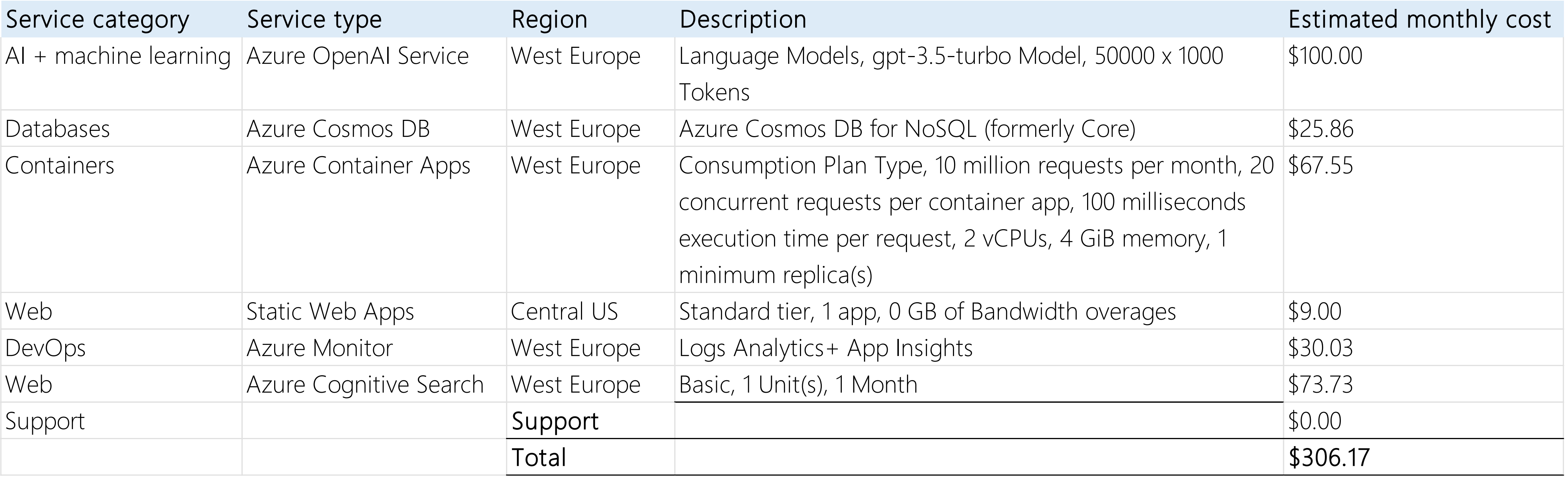
Build
Custom models from scratch

Buy

Build

- Spectrum of options
- Extremes
  - Consume - Simply consume and existing application e.g. Copilot for Microsoft 365 – to test quickly
  - Build - Try to build & train your own LLM from scratch, purview of technology vendors e.g. OpenAI
- Middle
  - Embed - Use APIs that are linked to existing pre-trained LLMs to build a customised application. Extend – BYOD (Bring Your Own Data)
    - RAG – Retrieval Augmented Generation - Doesn’t change the foundational LLM model but informs the LLM with your own internal data
    - Finetuning – relatively few people are doing, but becoming more popular fine tuning open source LLMs

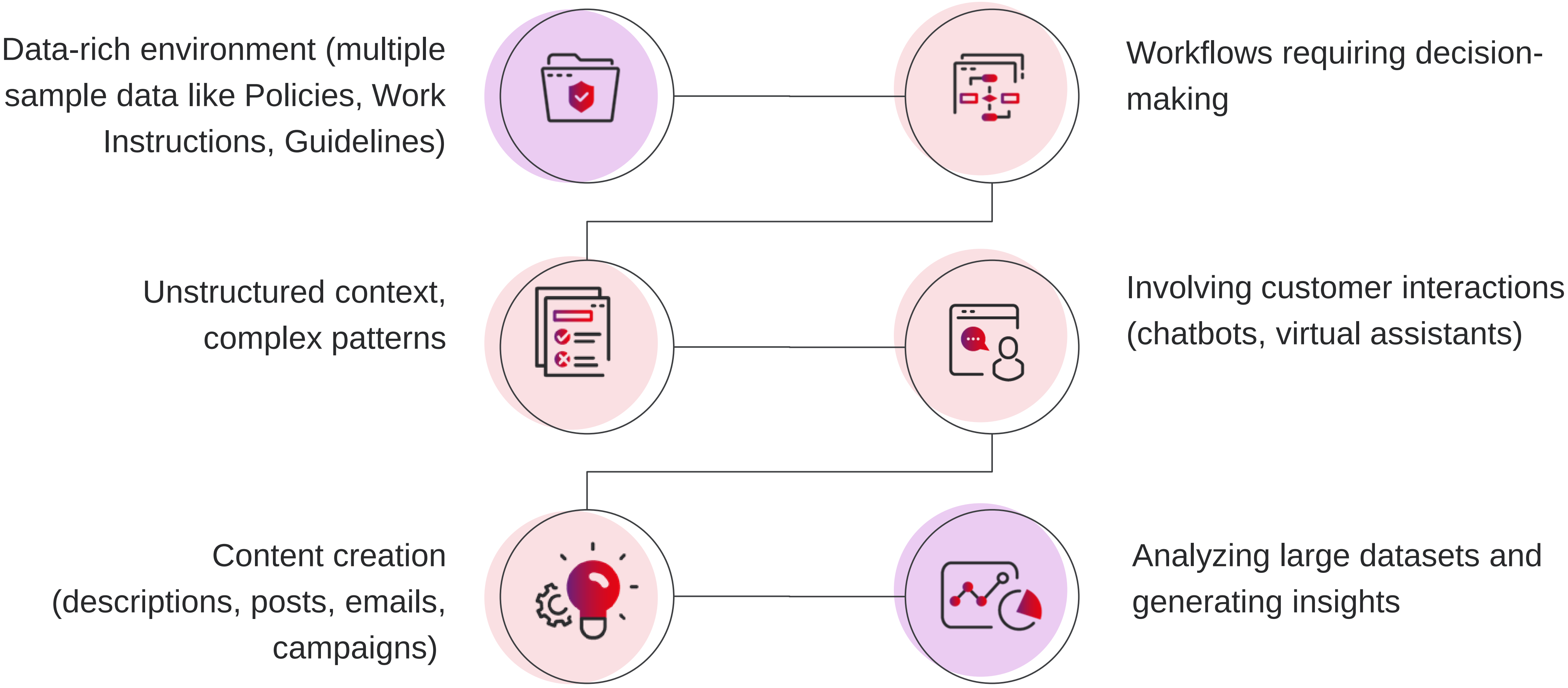
Infrastructure Guesstimate Cost (monthly basic)



Service category	Service type	Region	Description	Estimated monthly cost
AI + machine learning	Azure OpenAI Service	West Europe	Language Models, gpt-3.5-turbo Model, 50000 x 1000 Tokens	$100.00
Databases	Azure Cosmos DB	West Europe	Azure Cosmos DB for NoSQL (formerly Core)	$25.86
Containers	Azure Container Apps	West Europe	Consumption Plan Type, 10 million requests per month, 20 concurrent requests per container app, 100 milliseconds execution time per request, 2 vCPUs, 4 GiB memory, 1 minimum replica(s)	$67.55
Web	Static Web Apps	Central US	Standard tier, 1 app, 0 GB of Bandwidth overages	$9.00
DevOps	Azure Monitor	West Europe	Logs Analytics+ App Insights	$30.03
Web	Azure Cognitive Search	West Europe	Basic, 1 Unit(s), 1 Month	$73.73
Support		Support		$0.00
		Total		$306.17

Link to Microsoft Azure Calculator

Processes suitable for GenerativeAI



Data-rich environment (multiple sample data like Policies, Work Instructions, Guidelines)

Workflows requiring decision-making

Unstructured context, complex patterns

Involving customer interactions (chatbots, virtual assistants)

Content creation (descriptions, posts, emails, campaigns)

Analyzing large datasets and generating insights

High Level Business Needs Example

#	Business Need
1	Client have a need for a <insert need> Example <configurable / customised 1st Line Help Desk Intelligent Chatbot Solution for …internal users …or their multiple Clients whom they manage IT Services>
2	Client have a need to implement a Solution which will interface with its <insert systems> . It has an integration layer < insert a link to the solution and api page if possible> which iChatbot Solution could integrate with.
3	Client have a need to implement a Solution which would be trained on <insert client's data> and which is held <insert client location>
4	Client have a need to implement a Solution which would allow their client users to "chat" with their <insert assistant / role – ie Provide a Multi-Assistant based interface for Client staff – think about does it need a human in the loop
5	Client have a need for NashTech to provide a <Ballpark/Estimation or for a commercial proposal for the overall solution>

Client’s Q&A – Understand the Client’s requirements

#	Category	Questions	Answers
1	Current Process & Most Common L1 Requests	Discussion about the current helpdesk support process and ticketing flow Ie Which Help desk Solution is being used ?	
		Identification of the most common questions and requests for L1 support.	
2	Identity and Access Management and Role Based Access Control for End Users and their ChatBot Interaction	What end users are to be supported ? Is Identity and Authentication mandatory?	
		Discussion about end users’ profiles – how many are internal/external? Are any access restrictions required (i.e., should specific users have access to more information than others)?	
		Exploration of user interaction preferences (website, app, messaging platforms, etc.).	
		Who will be responsible for monitoring and managing the chatbot's performance? How will users receive assistance if they encounter issues with the chatbot?	
3	Functionality & Integrations	Review of potential functionalities of the chatbot – support in troubleshooting, creation of L2 tickets, and anything else.	
		Discussion about integrations with existing systems Are there any other Integration end points required ? Ie Current state of automation (Rewst).	
		Are there specific metrics or analytics you'd like to track related to the chatbot's performance?	
4	Knowledge Base/Document Sources & Security	Discussion about the existing knowledge base, its location and access to it Ie number of documents, types of documents, format of documents.	
		Will the chatbot have access to user-specific data for personalized troubleshooting? Ie GDPR consideration for Any Special Personal Information?)	
5	Infra-structure	Whose Infra-structure will the solution be deployed on ?	
6	Volumetrics	Do you have approximately numbers for Number of clients ? Number of tickets per day/month? Number of users having access to the chatbot?	
7	Any other relevant information for Chatbot Solution	Language requirements ? Pre-defined IT Ticket Categorisation? Ticket Creation Requirement ie attach chat summary to ticket Prompt Engineering Style ? Manual Bot Response Review Any Self Service capability required ie provision of an Admin portal?	