

Postup vytvárania chatbota v MS Azure

Výpracovala: Bc. Romana Halamová

V rámci diplomovej práce sme v prostredí MS Azure OpenAI Service vytvorili chatbota slúžiaceho ako virtuálny asistent pre študentov. Proces zahŕňal návrh infraštruktúry, výber vhodného modelu, nastavenie parametrov, vytvorenie úložiska, integráciu vyhľadávacej služby a implementáciu dátového zdroja.

1. Vstup do MS Azure

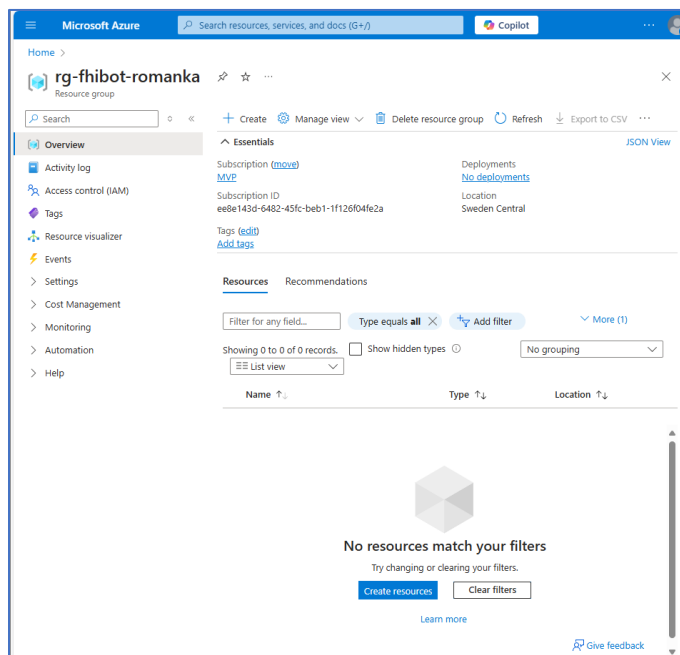
Najprv sme sa prihlásili do platformy MS Azure pomocou univerzitného či osobného konta. MS Azure je cloudová služba od spoločnosti Microsoft, ktorá umožňuje vytvárať a spravovať rôzne digitálne riešenia.

2. Vytvorenie Resource Group (skupiny zdrojov)

V sekcii "Subscriptions" (predplatné) sme vytvorili novú Resource Group (skupinu zdrojov) s názvom **rg-fhibot-romanka**. Resource Group (skratka RG) je logická jednotka, ktorá združuje a spravuje všetky zdroje potrebné pre náš projekt. Slúži ako "úložisko" alebo "množina" pre všetky budúce zdroje a komponenty potrebné na vytvorenie a správu chatbota. Vybrali sme pomenovanie rg-fhibot-romanka, kde:

- **rg** znamená "resource group" (skupina zdrojov),
- **fhibot** označuje názov nášho chatbot projektu,
- **romanka** je identifikátor pre autora projektu.

Obr. 1: Vytvorenie skupiny zdrojov



3. Výber služby OpenAI cez Marketplace

V ďalšom kroku sme v Azure Marketplace vybrali službu **OpenAI**, keďže naše riešenie je postavené na modeloch umelej inteligencie. Služba OpenAI poskytuje prístup k pokročilým jazykovým modelom, ktoré umožňujú generovanie textov, analýzy alebo chatbot riešenia.

4. Nastavenie parametrov:

- **Región:** Vybrali sme East US 2, pretože poskytuje široký výber najnovších modelov.
- **Meno zdroja:** Nastavili sme oai-fhibot-romanka, pričom:
 - **oai** znamená „OpenAI resource“,
 - **fhibot** označuje náš chatbot projekt,
- **Pricing Tier (cenová úroveň):** Vybrali sme Standard S0, čo je jediná dostupná možnosť a zahŕňa základné funkcie služby OpenAI.

Obr. 2: Výber služby OpenAI

Create Azure OpenAI ...

Basics | Network | Tags | Review + submit

Azure OpenAI Service provides access to OpenAI's powerful language models, including all the latest OpenAI models. These models can be easily adapted to your specific tasks, including but not limited to content generation, summarization, image understanding, semantic search, and natural language to code translation. Top use cases include Call Centers, Virtual Assistants, Accessibility, Content Generation, and Code Development. The service also features the Assistants API, Fine Tuning capabilities and many ways to connect your data to the service for conversational experiences. The service can be scaled through Standard (tokens) and Provisioned (PTUs) deployment types.

[Learn more](#)

Project Details

Subscription *

Resource group * [Create new](#)

Instance Details

Region

Name *

Pricing tier *

[View full pricing details](#)

Content review policy

In default and minimum harmful use of the Azure OpenAI Service. Microsoft limit the content you send to the

[Previous](#) [Next](#)

5. Nastavenia siete (Network)

Zvolili sme predvolené nastavenie: **All networks** (Všetky siete). To znamená, že naša služba OpenAI je prístupná zo všetkých sietí vrátane internetu. Toto riešenie je praktické, keďže plánujeme API endpoint využívať aj mimo Azure, napríklad v Pythone.

6. Pridávanie tagov (Tags)

Sekciu **Tags** (značky) sme ponechali prázdnu, pretože v rámci tohto projektu nepotrebujeme označovať zdroje.

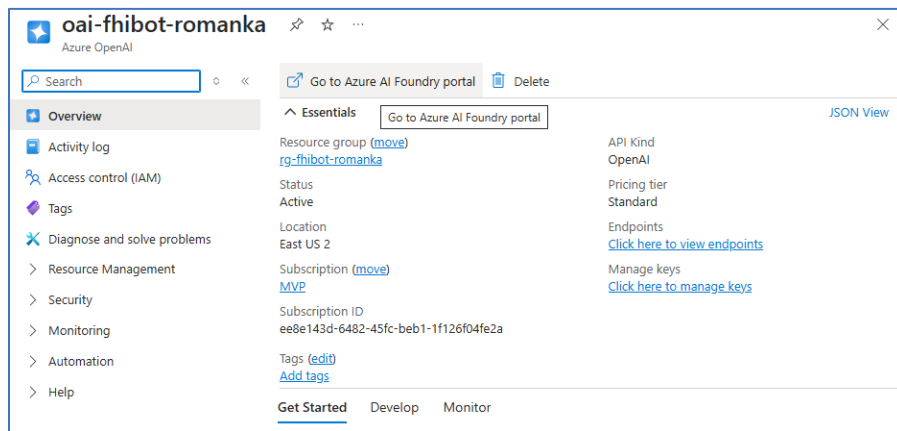
7. Dokončenie vytvorenia služby

Na záver sme skontrolovali nastavenia v sekcii **Review + Submit** a klikli na tlačidlo **Create**. Služba bola úspešne vytvorená a získali sme **API endpoint**, ktorý vieme používať v rôznych aplikáciách, napríklad v Pythone.

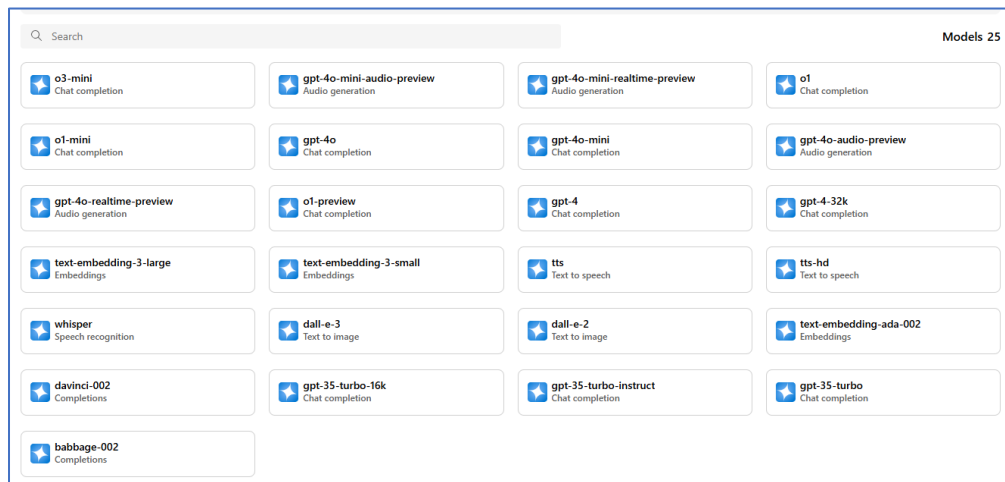
8. Výber vhodného modelu pre chatbot v Azure AI Foundry

Po prihlásení do **Azure AI Foundry** sme prešli do sekcie **Model Catalog**, kde sú zobrazené všetky dostupné modely pre tvorbu AI riešení. Každý model má odlišné vlastnosti a cenu, čo je dôležité zvážiť pri výbere.

Obr. 3: Azure AI Foundry



Obr. 4: Model Catalog



9. Výber modelu: GPT-4o-mini

Zvolili sme model **GPT-4o-mini** s globálnym nasadením. Tento model je pre nás vhodný z viacerých dôvodov:

- **Dostatočné schopnosti:** Podporuje textové vstupy a poskytuje rýchle odpovede, čo je pre študentského chatbota postačujúce.
- **Nízke náklady:** Je cenovo efektívny, pretože oproti plnej verzii GPT-4o je lacnejší.
- **Jednoduchá integrácia:** Je priamo dostupný v Azure AI Foundry a ľahko sa nasadzuje do našich služieb.
- **Rýchla odozva:** Poskytuje okamžité odpovede, čo zlepšuje používateľskú skúsenosť.

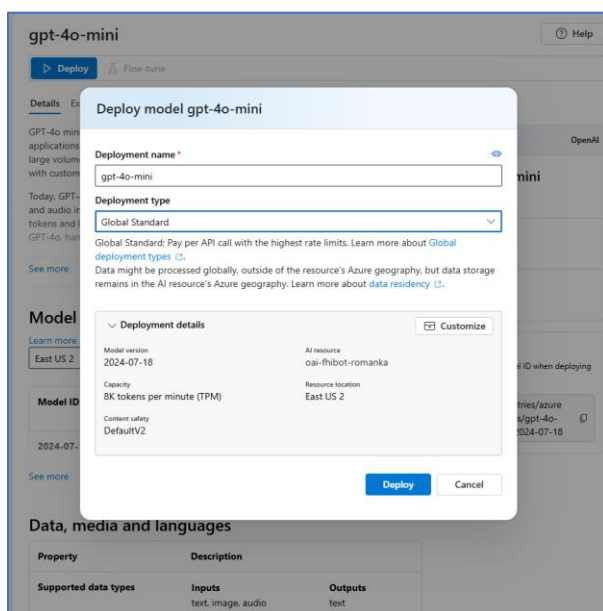
Týmto výberom sme zabezpečili rozumný kompromis medzi výkonom, cenou a potrebami nášho projektu, pričom sme sa zamerali na efektívne riešenie vhodné pre študentské účely.

Microsoft Azure ponúka viacero platforiem pre AI riešenia. **Azure OpenAI Service** (ktorú používame) poskytuje priamy prístup k modelom OpenAI, ako sú GPT-4, a je ideálna na tvorbu chatbotov. **Azure AI Studio** je komplexný nástroj na budovanie AI aplikácií, ktorý zahŕňa aj OpenAI modely. **Azure Machine Learning Studio** slúži na tréning vlastných modelov strojového učenia a **Azure Cognitive Services** poskytujú predtrénované modely pre rozpoznávanie obrazu, textu alebo reči.

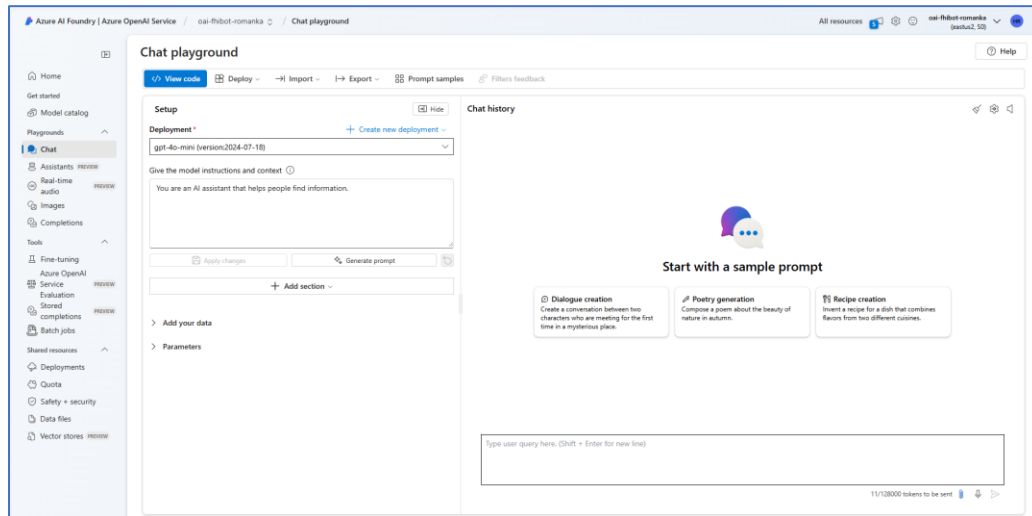
10. Vytvorenie prvého modelu GPT-4o-mini a nastavenie systémovej správy

Po nasadení modelu **GPT-4o-mini** sa zobrazí rozhranie **Chat Playground**, kde vpravo vidíme chatovacie okno a vľavo možnosti nastavení. Na ľavej strane môžeme vybrať model, definovať systémovú správu, pridávať vlastné dáta a nastavovať parametre.

Obr. 5: Vytvorenie modelu GPT-4o-mini

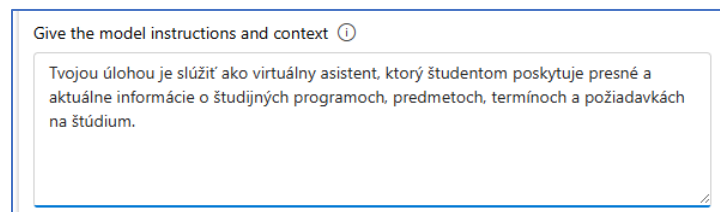


Obr. 6: Chat playground



Model zatiaľ funguje všeobecne a odpovedá na akékoľvek otázky bez špecializácie. Preto sme mu zadefinovali systémovú správu, aby slúžil ako virtuálny asistent pre študentov:

Obr. 7: Systémová správa modelu



Dôležité je, že čím je systémová správa detailnejšia, tým presnejšie model reaguje. Zároveň však platí, že **systémový prompt sa posiela s každou správou**, čo zvyšuje spotrebu tokenov.

Po vložení systémovej správy sme klikli na „**Apply changes**“ (**uložiť zmeny**). Model sa už správa podľa nových pokynov, no **momentálne nemá dáta** zo študijného systému fakulty, takže nedokáže poskytovať konkrétne informácie o štúdiu. Na to bude potrebné nahrat špecifické dáta alebo integrovať externé zdroje.

11. Nastavovanie parametrov modelu GPT-4o-mini

Pri práci s modelom v Azure OpenAI Service vieme upraviť rôzne parametre, ktoré ovplyvňujú spôsob generovania odpovedí. Tu je stručné vysvetlenie jednotlivých nastavení a ich význam pre náš chatbot:

Past messages included (Počet zahrnutých predchádzajúcich správ) určuje, koľko predchádzajúcich správ sa posiela modelu spolu s novou otázkou. Model si totiž „nepamätá“ predošlé konverzácie, ale vidí ich, ak ich znovu pošleme. Nastavili sme hodnotu 10, aby model videl posledných 10 otázok a odpovedí, čo je ideálne pre plynulé konverzácie. Nevýhodou je, že viac správ znamená vyššiu

spotrebu tokenov, čo zvyšuje cenu. Toto nastavenie je pre nás užitočné, ak chceme, aby si chatbot pamätal kontext diskusie, napríklad pri poradenstve študentom.

Max response (Maximálny počet tokenov v odpovedi) obmedzuje dĺžku odpovede. Jeden token zodpovedá približne jednému slovu, pričom 800 tokenov je asi 500 slov. Nastavili sme hodnotu 800, čo postačuje na podrobné odpovede, no stále kontroluje dĺžku, aby chatbot nerozprával príliš obsiahle. Pre náš účel je toto nastavenie ideálne, pretože odpovede budú výstižné, ale dostatočne obsiahle na vysvetlenie študijných informácií.

Temperature (Kreativita odpovedí) ovláda náhodnosť odpovedí. Hodnota 0 znamená veľmi presné a faktické odpovede, zatiaľ čo hodnota 1 je veľmi kreatívna, ale model si môže „vymýšľať“. Nastavili sme hodnotu 0.7, čo je rozumný kompromis medzi faktickosťou a zrozumiteľnosťou. Pre nás je to ideálne, pretože študenti potrebujú spoľahlivé informácie, ale odpovede môžu byť stále priateľské a prirodzené.

Top P (Jemné nastavenie variability odpovedí) obmedzuje rozsah možností, z ktorých model vyberá odpoveď. Hodnota 1 znamená maximálnu variabilitu, kde môžu byť odpovede rôznorodé, zatiaľ čo hodnota 0.5 poskytuje veľmi konzervatívne odpovede. Nastavili sme hodnotu 0.95, vďaka čomu model poskytne odpovede s jemnou variabilitou, ale stále konzistentné. Pre nás je toto nastavenie dobré, pretože chatbot pôsobí prirodzene, ale neodbočuje od témy.

Frequency penalty (Trest za opakovanie) zabráňuje modelu opakovať rovnaké frázy alebo informácie. Nastavili sme hodnotu 0, čiže model môže niekedy zopakovať informácie, ak je to vhodné. Ak by chatbot začal často opakovať rovnaké odpovede, môžeme túto hodnotu zvýšiť.

Presence penalty (Trest za opakovanie tém) odrádza model od opakovania tém alebo slov, ktoré už v konverzácii použil. Aktuálne sme nastavili hodnotu 0, čím nechávame chatbotu voľnosť. Ak by sa chatbot príliš sústredil na jednu tému, môžeme túto hodnotu zvýšiť.

Obr. 8: Nastavovanie parametrov modelu

The image shows a configuration interface for an AI model. It includes several sliders and input fields for different parameters. The parameters and their values are as follows:

| Parameter | Value |
|------------------------|---------------|
| Past messages included | 10 |
| Max response | 800 |
| Temperature | 0.7 |
| Top P | 0.95 |
| Stop sequence | Stop sequence |
| Frequency penalty | 0 |
| Presence penalty | 0 |

12. Vytvorenie úložiska (Storage) v Azure pre dátový zdroj

Po kliknutí na „Add your data“ sme vytvorili **úložisko (storage)** v cloude, ktoré slúži ako dátový zdroj pre nášho chatbota. Úložisko je miestom, kde môžeme ukladať súbory, dokumenty alebo iné informácie, ktoré môže chatbot využívať pri odpovediach.

Pri vytváraní úložiska sme nastavili nasledovné parametre:

- **Subscription (Predplatné):** Použili sme predplatné s názvom *MVP*, ktoré máme pridelené.
- **Resource Group (Skupina zdrojov):** Vybrali sme skupinu **rg-fhibot-romanka**, ktorú sme vytvorili na začiatku. Všetky súvisiace zdroje tak ostávajú pokope.
- **Storage account name (Názov úložiska):** Nastavili sme **stfhibotromanka**, čo je jedinečný identifikátor úložiska. Názov musí obsahovať iba malé písmená a čísla.
- **Region (Región):** Opäť sme zvolili **East US 2**, aby boli všetky služby v rovnakom regióne, čím sa zníži latencia a zvýši výkon.
- **Performance (Výkon):** Vybrali sme možnosť **Standard**, ktorá je vhodná pre väčšinu bežných scenárov a je cenovo výhodnejšia.
- **Redundancy (Redundancia):** Zvolili sme **Geo-redundant storage (GRS)**, čo znamená, že dáta budú automaticky replikované do iného regiónu. Navyše sme aktivovali možnosť **read-access**, aby boli dáta dostupné aj pri výpadku hlavného regiónu.

Po vyplnení všetkých parametrov sme klikli na „Review + create“.

Obr. 9: Vytvorenie úložiska

Home >

Create a storage account

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription * MVP

Resource group * rg-fhibot-romanka [Create new](#)

Instance details

Storage account name * ⓘ stfhibotromanka

Region * ⓘ (US) East US 2 [Deploy to an Azure Extended Zone](#)

Primary service ⓘ Select a primary service

Performance * ⓘ ☒ Standard: Recommended for most scenarios (general-purpose v2 account) ☐ Premium: Recommended for scenarios that require low latency.

Redundancy * ⓘ Geo-redundant storage (GRS) ☒ Make read access to data available in the event of regional unavailability.

[Previous](#) [Next](#) [Review + create](#)

13. Vytvorenie AI Search (Vyhľadávacej služby) v Azure

Po vytvorení úložiska sme prešli na vytvorenie služby **AI Search**, ktorá umožní chatbotovi rýchlo a presne vyhľadávať informácie z našich dát. Táto služba je kľúčová pre poskytovanie relevantných odpovedí študentom.

Pri nastavovaní AI Search sme vybrali tieto parametre:

- **Cenník (Pricing Tier):** Zvolili sme plán **Basic** s cenou približne **70 € mesačne**, pretože ponúka dostatočné kapacity pre naše potreby za rozumnú cenu.
- **Indexy:** Basic plán umožňuje **15 indexov**, čo znamená, že môžeme vytvoriť 15 rôznych kolekcií dát (napr. predmety, harmonogramy, pravidlá štúdia).
- **Indexery:** Získavame **15 indexerov**, čo sú automatické procesy, ktoré pravidelne prehľadávajú a aktualizujú naše dáta.

Obr. 10: Vytvorenie vyhľadávacej služby

[Home](#) >

Create a search service

Basics

Scale

Networking

Tags

Review + create

Project details

Subscription *

MVP

Resource Group *

rg-fhilot-romanka

[Create new](#)

Instance Details

Service name ⓘ

srch-fhilot-romanka

Location *

East US 2

Pricing tier ⓘ

Basic

15 GB/Partition, max 3 replicas, max 3 partitions, max 9 search units

[Change Pricing Tier](#)

Review + create

Previous

Next: Scale

Obr. 11: Cenník AI Search

| Select Pricing Tier | | | | | | | | | |
|--|-------------------|---------|----------|------------------|------------------|--------------|----------|------------|-------------------------------|
| Browse available skus and their features | | | | | | | | | |
| SKU | Offering | Indexes | Indexers | Vector quota | Total storage | Search units | Replicas | Partitions | Search unit cost/month (est.) |
| F | Free | 3 | 3 | 25 MB ⓘ | 50 MB | 1 | 1 | 1 | €0.00 |
| B | Basic | 15 | 15 | 5 GB/Partition | 15 GB/Partition | 9 | 3 | 3 | €70.34 |
| S | Standard | 50 | 50 | 35 GB/Partition | 160 GB/Partition | 36 | 12 | 12 | €234.00 |
| S2 | Standard | 200 | 200 | 150 GB/Partition | 512 GB/Partition | 36 | 12 | 12 | €936.00 |
| S3 | Standard | 200 | 200 | 300 GB/Partition | 1 TB/Partition | 36 | 12 | 12 | €1,872.00 |
| S3HD | High-density | 1000 | 0 | 300 GB/Partition | 1 TB/Partition | 36 | 12 | 3 | €1,872.00 |
| L1 | Storage Optimized | 10 | 10 | 150 GB/Partition | 2 TB/Partition | 36 | 12 | 12 | €2,673.59 |
| L2 | Storage Optimized | 10 | 10 | 300 GB/Partition | 4 TB/Partition | 36 | 12 | 12 | €5,346.48 |

14. Pridanie dátového zdroja a vytvorenie modelu pre vektorové vyhľadávanie

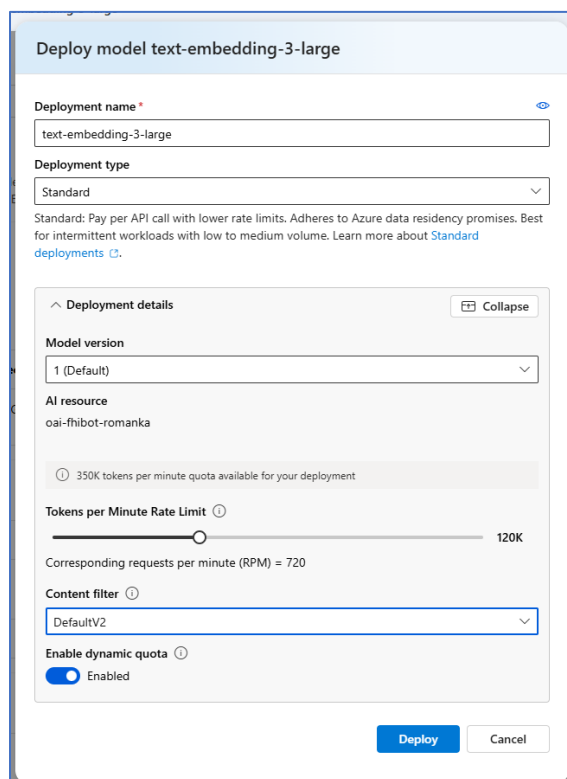
V **Playgrounde** sme pridali nový **dátový zdroj (data source)** z úložiska stfhibotromanka a pre vyhľadávanie sme zvolili AI Search s indexom indexfhiweb.

Na zlepšenie výsledkov vyhľadávania sme použili vektorové vyhľadávanie, ktoré hľadá podobnosti v obsahu namiesto jednoduchého vyhľadávania kľúčových slov.

Pre vektorové vyhľadávanie sme museli vytvoriť a nasadiť model „**text-embedding-3-large**“, ktorý transformuje texty do číselných vektorov (embeddingov). Tieto vektory umožňujú rýchle a presné vyhľadávanie podľa významu a nie len podľa presných slov.

Model sme nasadili s obmedzením 120K tokenov za minútu a aktivovali sme dynamickú kvótu, ktorá automaticky prispôbuje výkon podľa potreby. Tento krok zabezpečil, že chatbot môže efektívne vyhľadávať informácie zo študijných materiálov.

Obr. 12: Vytvorenie modelu pre vektorové vyhľadávanie



The screenshot shows the 'Deploy model text-embedding-3-large' configuration page in Azure AI Studio. The 'Deployment name' is set to 'text-embedding-3-large'. The 'Deployment type' is 'Standard'. Under 'Deployment details', the 'Model version' is '1 (Default)', the 'AI resource' is 'oai-fhibot-romanka', and the 'Tokens per Minute Rate Limit' is set to 120K. The 'Content filter' is 'DefaultV2' and 'Enable dynamic quota' is 'Enabled'. The 'Deploy' button is highlighted in blue.

Deploy model text-embedding-3-large

Deployment name *

text-embedding-3-large

Deployment type

Standard

Standard: Pay per API call with lower rate limits. Adheres to Azure data residency promises. Best for intermittent workloads with low to medium volume. Learn more about [Standard deployments](#).

Deployment details

Model version

1 (Default)

AI resource

oai-fhibot-romanka

350K tokens per minute quota available for your deployment

Tokens per Minute Rate Limit

120K

Corresponding requests per minute (RPM) = 720

Content filter

DefaultV2

Enable dynamic quota

Enabled

Deploy Cancel

Obr. 13: Vloženie dátového súboru

Select or add data source

Your data source is used to ground the generated results with your data. Select an existing data source or create a new data connection with Azure Blob Storage, databases, search, URLs, or local files as the source the grounding data will be built from.

[Learn more about data privacy and security in Azure AI](#)

Select data source *

Upload files (preview)

Subscription *

MVP

Select Azure Blob storage resource ⓘ *

stfhibotromanka

[Create a new Azure Blob storage resource](#)

Select Azure AI Search resource ⓘ *

srch-fhibot-romanka

[Create a new Azure AI Search resource](#)

Enter the index name ⓘ *

indexfhibot

Using Azure AI Search will incur usage to your account. [View Pricing](#)

☒ Add vector search to this search resource. ⓘ

Embedding model

Select an embedding model ⓘ *

Azure OpenAI Service - text-embedding-3-large

15. Nahratie súborov a nastavenie Data Management v Azure

Na začiatku sme **nahrali niekoľko súborov** do úložiska, ktoré bude chatbot využívať pri odpovediach.

V sekcii **Data Management** sme nastavili:

- **Typ vyhľadávania (Search type):** Zvolili sme **Hybrid + Semantic**, čo kombinuje vektorové vyhľadávania so sémantickým vyhľadávaním. Vďaka tomu chatbot rozumie kontextu otázok a poskytuje presnejšie odpovede.
- **Veľkosť častí (Chunk Size):** Vybrali sme **1024 tokenov (default)**, čo znamená, že dokumenty sa pri indexovaní rozdelia na bloky s veľkosťou približne 1024 slov. Táto veľkosť je vyvážená – dostatočne veľká na pochopenie kontextu, ale stále presná pri vyhľadávaní.

V sekcii **Data Connection** sme nastavili spôsob, ako sa služby navzájom prepoja:

- **Typ autentifikácie (Azure resource authentication type):** Zvolili sme **API key**, pretože je jednoduchý a rýchly spôsob na pripojenie medzi AI Search, OpenAI Service a naším Blob Storage.

Na záver sme klikli na „**Save and close**“, čím sme spustili proces Ingestion (nahrávania a spracovania dát).

Obr. 14: Nastavenie Data Management

The screenshot shows the 'Add data' dialog box with the 'Data management' tab selected. On the left, a vertical list of steps includes 'Data source', 'Upload files', 'Data management' (highlighted), 'Data connection', and 'Review and finish'. The main content area is titled 'Data management' and contains instructions: 'Set up specific configurations for your data and how the model will respond to requests.' It includes a link to 'Learn more about data privacy and security in Azure AI'. Below this, there is a 'Search type' dropdown menu set to 'Hybrid + semantic', with a note that using semantic search will incur usage to the Azure AI Search account. A 'Chunk Size' section explains that chunking breaks documents into segments for search and retrieval, measured in tokens, and provides a link to learn more about selecting a chunk size. At the bottom, there is a 'Select a size' section with four radio button options: 256, 512, 1024 (selected and marked as default), and 1536.

Obr. 15: Prepojenie dát

The screenshot shows the 'Add data' dialog box with the 'Data connection' tab selected. The left-hand navigation pane is identical to the previous image, with 'Data connection' now highlighted. The main content area is titled 'Data connection' and instructs the user to 'Select how your Azure resources connect to each other.' It includes a link to 'Learn more about data privacy and security in Azure AI'. Below this, the 'Azure resource authentication type' section has two radio button options: 'System assigned managed identity' and 'API key' (which is selected).

16. Nasadenie chatbota a vytvorenie webovej aplikácie

Po nastavení všetkých parametrov v **Azure OpenAI Playground** sme pristúpili k nasadeniu aplikácie (deployment). Využili sme **Contoso**, čo je vzorová platforma v Azure určená na jednoduché nasadenie webových aplikácií.

Výsledkom je plne funkčná online aplikácia, ktorá spracováva otázky používateľov, komunikuje s modelom OpenAI a poskytuje odpovede na základe dostupných dát. Táto stránka môže byť ďalej prispôbena, napríklad zmenou dizajnu alebo prepojením s inými systémami univerzity.

Obr. 16: Nasadenie chatbota

Deploy to a web app PREVIEW ✕

ⓘ Your web app will be configured with Entra ID authentication enabled. It may take a few minutes to apply after deployment completes, during which time chat functionality in the app won't be available. Please wait 10 minutes, then reload the app and sign in to begin chatting. ✕

Pick your configurations to deploy a web app. [Learn more about web apps](#)

☒ Create a new web app ☐ Update an existing web app

Name ⓘ *

Subscription ⓘ *

Resource group ⓘ *

Location ⓘ *

Pricing plan ⓘ *

☒ Enable chat history in the web app

Enabling chat history will incur CosmosDB usage to your account [View Pricing](#)

Web apps will incur usage to your account [View Pricing](#)

Deploy Cancel

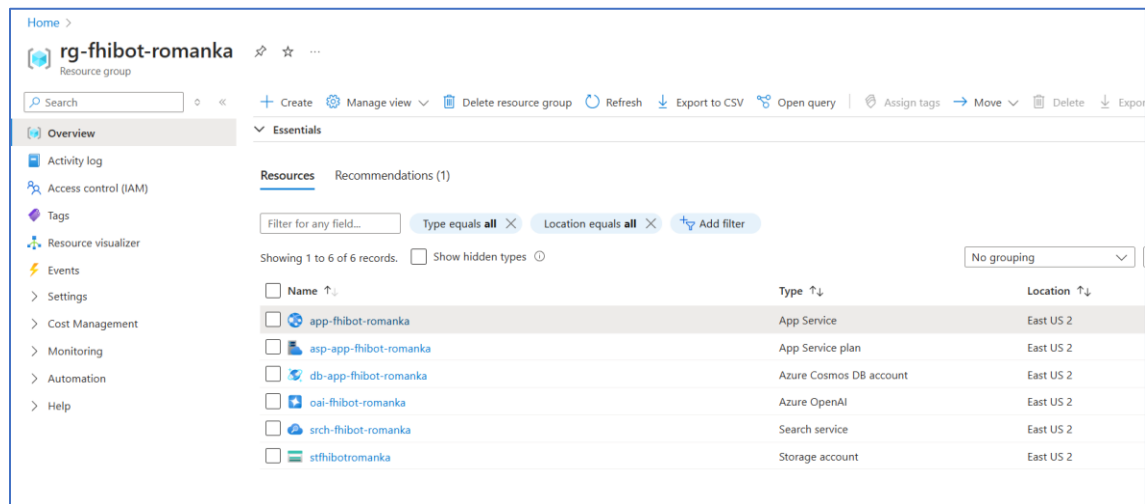
Prehľad vytvorených zdrojov v Resource Group (RG) pre chatbot

V rámci **Resource Group (rg-fhibot-romanka)** sme vytvorili niekoľko služieb, ktoré sú potrebné na fungovanie nášho chatbota.

1. **App Service (app-fhibot-romanka)** – Webová aplikácia, ktorá umožňuje prístup k chatbotovi cez webovú stránku.
2. **App Service Plan (asp-app-fhibot-romanka)** – Plán služieb, ktorý definuje výpočtový výkon a škálovanie webovej aplikácie.
3. **Azure Cosmos DB (db-app-fhibot-romanka)** – Databáza na ukladanie údajov, ktoré chatbot využíva.
4. **Azure OpenAI Service (oai-fhibot-romanka)** – Hlavná služba poskytujúca jazykový model OpenAI, ktorý spracováva otázky používateľov.
5. **Azure AI Search (srch-fhibot-romanka)** – Vyhľadávacia služba, ktorá umožňuje chatbotovi rýchlo vyhľadávať informácie v dátach.
6. **Azure Storage Account (stfhibotromanka)** – Úložisko pre dokumenty a dáta, ktoré chatbot využíva pri odpovediach.

Tieto služby spolupracujú tak, aby chatbot mohol efektívne spracovávať otázky, vyhľadávať informácie a poskytovať odpovede prostredníctvom webovej aplikácie.

Obr. 17: Výsledná skupina zdrojov



Odkazy

Microsoft. (2025). *Microsoft Azure*. Dostupné na Internete: Azure OpenAI Service: <https://azure.microsoft.com/en-us/products/ai-services/openai-service>