SUTOL

Symposium2026

3.června • GreenPoint, Praha

Partneři akce

HCLSoftware

TD SYNNEX

ARROW

NGSS Next Generation Security Solutions

Whitesoft

NTF FACTORY

DOWNSTREAM

your SYSTEM

ISECO

pristo

cube team

SUTOL Symposium2026 3.června • GreenPoint, Praha

Začínáme s Domino IQ

aneb „mám Domino a chci aby bylo pracovalo za mě“

Ondřej Kostruh

NTF Factory

kostruh@ntf-factory.com

















NTF FACTORY

Továrna na aplikace Notes/Domino.



Ojedinělá fotografie z továrny, kde právě dělníci vyrábějí Domino aplikace pro krabicový software Spinesys.

Nedatováno.

Obsah

1. část (před obědem)

- Úvod
- Domino IQ - jak začít
- Novinky v 14.5.1 - Guard model, Paraphrasing
- Konfigurace - Model, Prompt, Command, Configuration
- Testy na příkladu

2. část (po obědě)

- RAG
- Troubleshooting
- HW a měření výkonu
- Závěr

Legislativa - jen krátce

- AI Act
  - Školení obsluhy
  - Označení obsahu generovaného AI
- GDPR
  - Zpracování osobních dat pomocí AI
  - Předávání dat mimo EU
- NIS2
  - Zabezpečení dat

K čemu AI chci?

Otázka za 1.000.000$

Otázkay za 1.000.000$

- K čemu AI chci?
- Jaká data budu posílat?
  - Jsou moje?
- A kam je budu posílat?
  - Jsou v bezpečí?

Dvě hlavní cesty

Cloud

- (Ne)bezpečí dat
- Nejistá cena
- Omezené možnosti nastavení
+ Snadný start

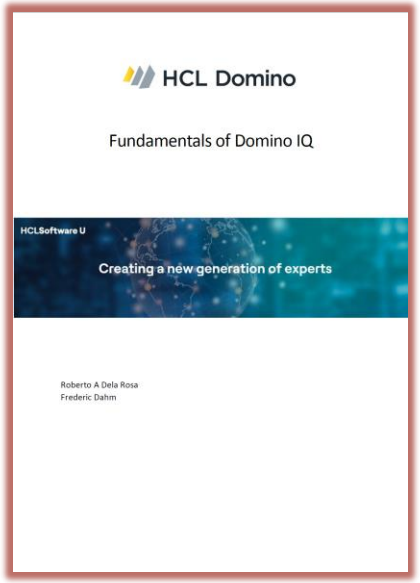
On Premise

+ Kompletní kontrola
+ Pevná cena (HW, napájení, chlazení)
- Velká vstupní investice
+ Bezpečí dat
- Složitější zprovoznění

https://lenovopress.lenovo.com/lp2130-lenovo-llm-sizing-guide

Začínáme …

- Co potřebuji mít než začnu?
  - HCL Domino 14.5 nebo 14.5.1 (podle požadavků)
    - Linux nebo Windows
  - HCL Notes 14.5 (a vyšší)
    - Podle použití může, ale nemusí být potřeba
  - Licence CCB (term)



https://help.hcl-software.com/domino/14.5.1/admin/domino_iq_server.html
https://hclsoftwareu.hcl-software.com/
https://hclsoftwareu.hcl-software.com/images/Lc4sMQCcN5uxXmL13gSlsxClNTU3Mjc3NTc4MTc2/DS_Academy/Whitepapers__Tutorials/Fundamentals_of_Domino_IQ_v1_-_28_Apr_2026.pdf

Instalace Domino IQ

1. Stažení
   - Llama Server for Domino IQ (sekce Tools, Connectors, and Integrators)

2. Instalace
   - Rozbalení ZIP

3. Konfigurace
   - Domino Directory Profile
   - Domino IQ Configuration (dominoiq.nsf)

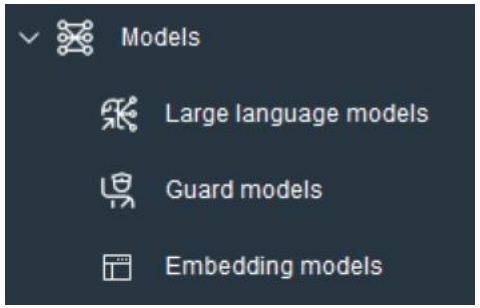
https://my.hcltechsw.com/downloads/domino/llamaserver
https://ntf-factory.com/sutol/domino-iq/



Domino IQ Configuration

- Modely … GGUF, Text Generation
  - Large language model
    - „hlavní“
  - Guard model
    - ochrana
  - Embedding model
    - RAG

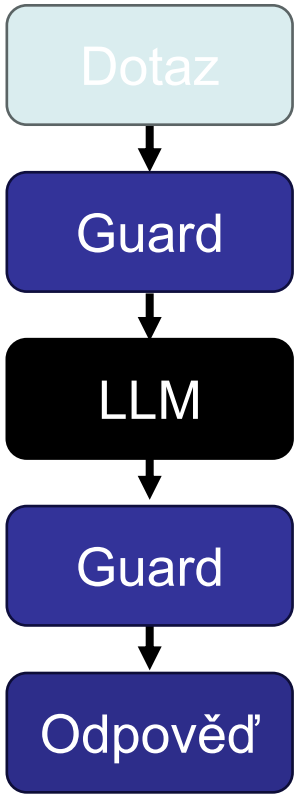
Novinky v 14.5.1



https://huggingface.co/models

Guard Model

- Kontrola promptu i odpovědi
  - Vrací jen safe/unsafe nebo Yes/No
  - Model to musí umět - ne všechny to umí
- Pro lokální i remote mod
- Ochrana
  - hlavně uživatele
  - ale i modelu (resp. API)



https://help.hcl-software.com/domino/14.5.1/admin/conf_understand_guard_models.html

Guard Model - pokračování

- Příklady modelů a před čím chrání

Enterprise Guard Model Risk Taxonomy Comparison

Risk Category	Meta Llama Guard 3	Google ShieldGemma	IBM Granite Guardian 3.0
Violent/Dangerous Content	Yes (S1, S9)	Yes (Targeted)	Yes
Hate Speech / Bias	Yes (S10)	Yes	Yes (Social Bias)
Sexual Content	Yes (S3, S4, S12)	Yes	Yes
Defamation / IP / Privacy	Yes (S5, S7, S8)	No	Yes (Unethical Behavior)
Specialized Advice / Elections	Yes (S6, S13)	No	No
RAG Groundedness / Relevance	No	No	Yes (Highly Specialized)
Prompt Injection / Jailbreaking	Yes (S14 Code Abuse)	No	Yes

Paraphrasing

- Další úroveň zabezpečení
- Před samotným dotazem na hlavní model je ještě jeden dotaz na „parafrázování“ původního dotazu uživatele
  - Spotřeba času, tokenů, el.energie
  - Může měnit význam dotazu
  - (často) Dotaz přeloží do angličtiny
- Ne všechny LLM umí

- …asi raději vypnout…
  - DOMIQ_DISABLE_PROMPT_PARAPHRASE=1

https://help.hcl-software.com/domino/14.5.1/admin/conf_security_considerations_for_iq.html
https://blog.nashcom.de/nashcomblog.nsf/dx/domino-iq-paraphrasing-explained.htm

Sizing

- Model
  - velikost (počet parametrů),
  - velikost kontextového okna (prompt systemový, uživatelský, RAG, historie, …)
    - Angličtina - 0,75 slov na token (100 tokenů ~ 75 slov)
    - Čeština - 2-3 tokeny na slovo (100 tokenů ~ 35-50 slov)

- Výkon
  - Latence (spíše u chatu),
  - Počet requestu za min. nebo s.

- Důležitá je zejména kombinace Model x HW

https://developer-blogs.nvidia.com/wp-content/uploads/2024/08/S62797-LLM-Inference-Sizing_-Benchmarking-End-to-End-Inference-Systems.pdf
https://lenovopress.lenovo.com/lp2130-lenovo-llm-sizing-guide

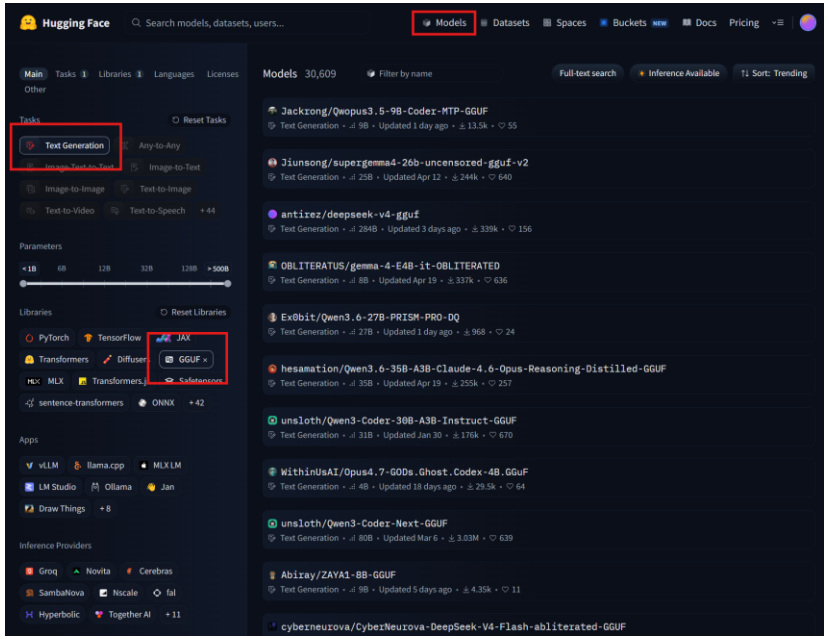
Typy modelů

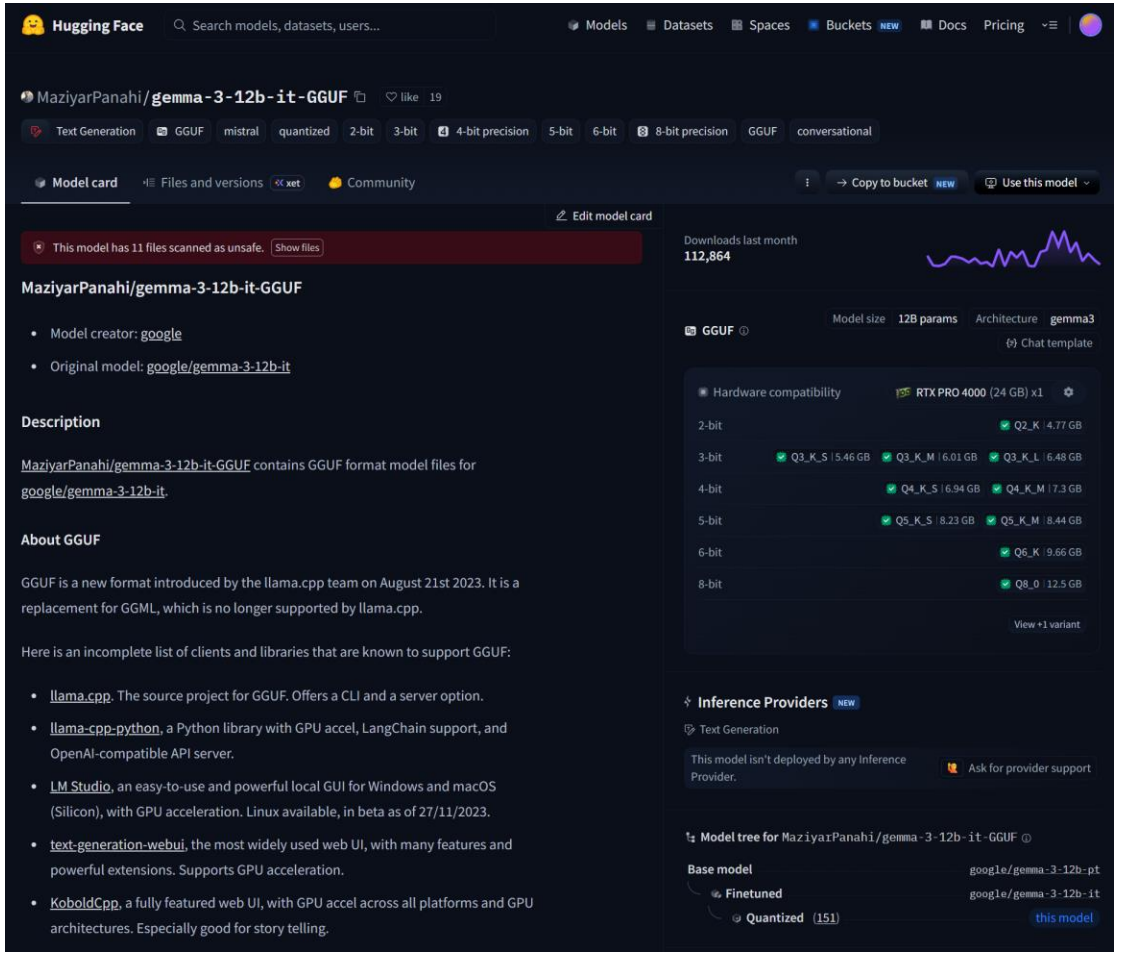
- Base
  - Výchozí model před fine-tuning
  - Jen hledá další token
- Instruct / Chat
  - Jednoduše formuje odpovědi na otázky
  - Rychlý, přímočarý
  - Lidský Systém 1
  - Velmi orientačně a přibližně 5x - 20x rychlejší (levnější) než Reasoning
- Reasoning / Thinking
  - Řeší komplexnější zadaní, vícekrokově
  - Je pomalejší - úkol si rozdělí a pracuje na něm
  - Lidský Systém 2

https://dev.to/rod_schneider/understanding-ai-language-models-base-chat-and-reasoning-a-beginners-guide-4323

Výběr modelu

•





https://ollama.com/search
https://huggingface.co/

Konfigurace, a stažení modelu

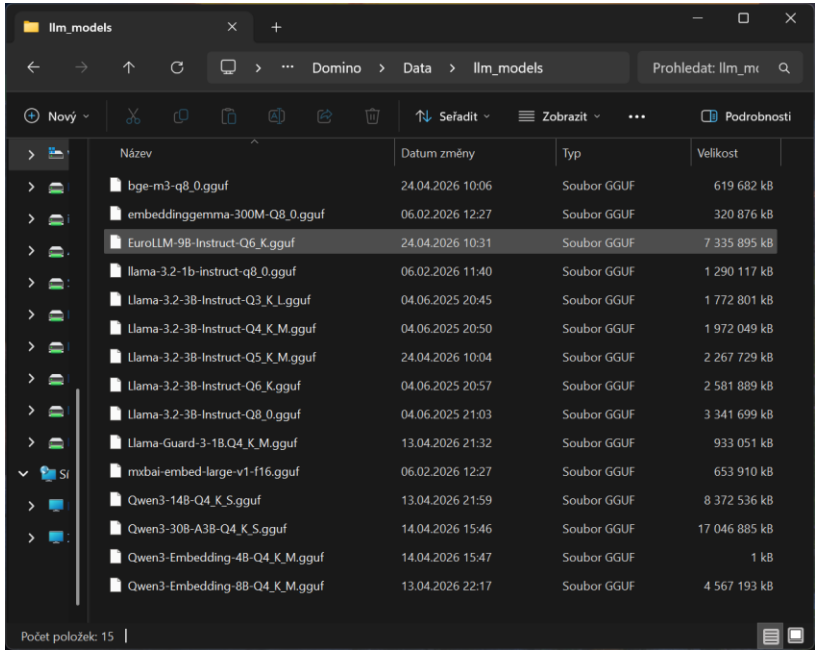
LLM Model

Basic | Comments

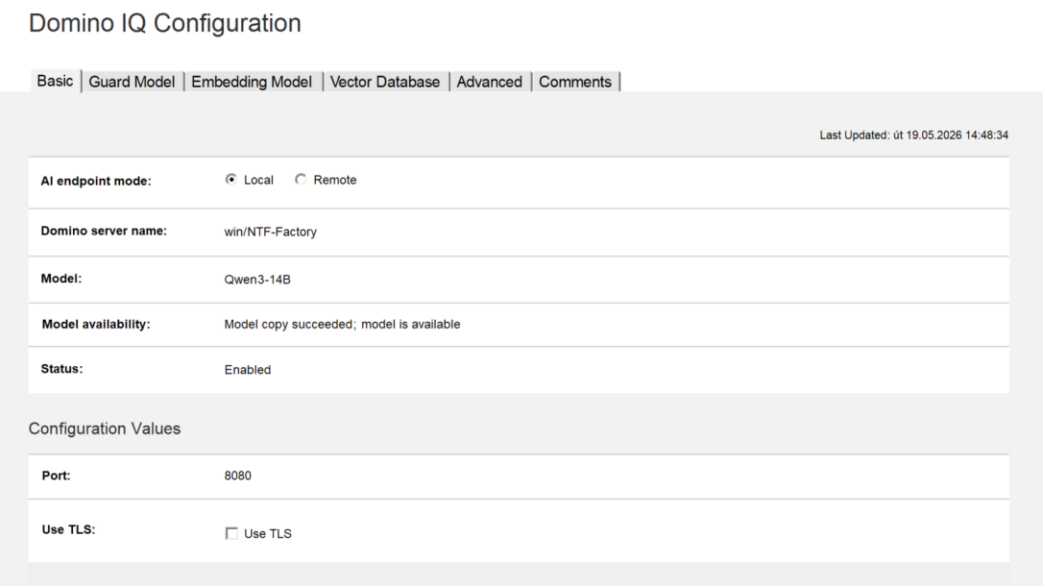
LLM Model

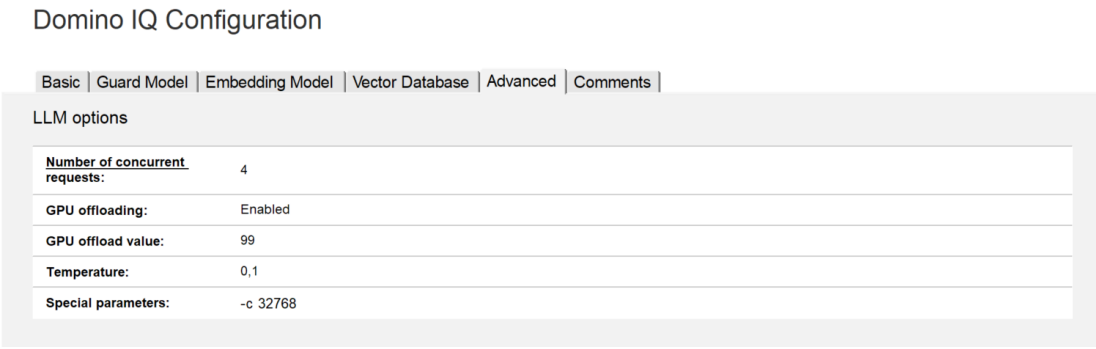
Last Updated: po 13.04.2026 21:59:06

Model name:	Qwen3-14B
Model type:	(•) Large language model ( ) Guard model ( ) Embedding model
Description:	unsloth/Qwen3-14B-GGUF
File name:	Qwen3-14B-Q4_K_S.gguf
Download model:	(•) Enabled ( ) Disabled
Download URL:	https://huggingface.co/unsloth/Qwen3-14B-GGUF/resolve/main/Qwen3-14B-Q4_K_S.gguf?download=true
SHA 256 Hash:	
Model status:	Model download succeeded



Konfigurace





https://help.hcl-software.com/domino/14.5.1/admin/conf_add_dom_iq_config_document.html

Konfigurace

Systém Prompt

- Zadání pro LLM
- První část kontextu

LLM System Prompt

Basic | Comments

Command Definition

Last Updated: st 13.05.2026 10:18:59

Display name:	restyle_and_translate_CZ2EN
Prompt:	Ze zadaného textu připrav obchodní mail v angličtině. Vrať pouze výsledný text.

Administration

Administrators:	Ondřej Kostruh/NTF-Factory

Command

- Volají funkce LS
- Nastavení parametrů

LLM Command

Basic | Comments

LLM Command

Last Updated: st 13.05.2026 10:14:56

Configuration(s):	win/NTF-Factory
Model:	
Command:	restyle_and_translate_CZ2EN
Description:	Přestylizuje text a přeloží ho do angličtiny.
System prompt:	restyle_and_translate_CZ2EN
RAG enabled database:	
RAG enabled fields:	
RAG maximum responses:	25
RAG threshold:	0,7
Maximum tokens:	10000
Temperature:	0,5

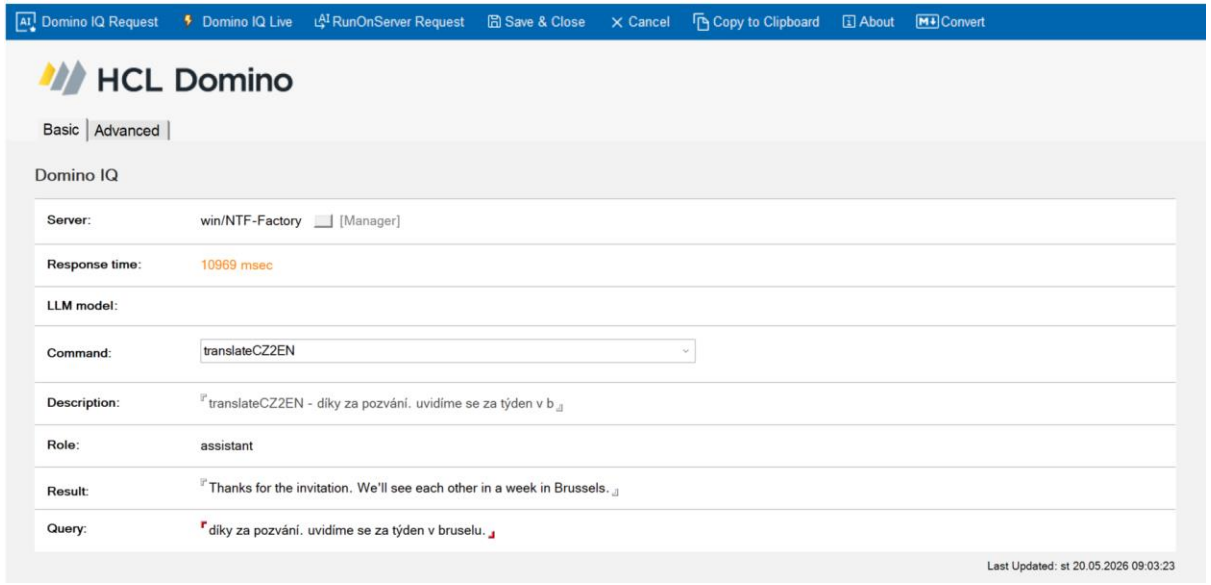
„Hello world!“ aneb první dotaz

- Domino IQ Lab Template – by OpenNTF



Domino IQ Lab Template – by OpenNTF
Free & Open Source (Apache 2.0)

Requires Domino Server V14.5
https://www.openntf.org



https://openntf.org/main.nsf/blog.xsp?permaLink=GACS-DGVK3M

Testy

Praktický test - klíčová slova

- Zadání:
  - V Notes aplikaci typu Helpdesk jsou uloženy servisní případy, které obsahují mimo jiné původní popis problému.
  - Celkem asi 2077 servisních požadavků.
  - Po AI budeme chtít přiřadit k požadavkům klíčová slova podle popisu problému
  - Jde o pochopení textu … nebo ne?

- Nastavení:
  - Model: Qwen3-14B (Q4)
  - Number of concurrent requests: 4
  - GPU offloading: Enabled
  - GPU offload value: nejprve výchozí 33, pak upraveno na 99
  - Temperature: 0,1
  - Special parameters: -c 32768

1. pokus

- Prompt:

  Najdi v zadaném textu hlavní klíčová slova. Vrať pouze maximálně 4 klíčová slova vystihující text.

- Výsledek:
  - 2334 klíčových slov (z 2077 dokumentů)
  - Téměř každý dokument dostal „své” klíčové slovo



2. pokus

- Prompt:

  Jsou dána klíčova slova: server, upgrade, aplikace, prohlídka, databáze, instalace, Docházka, Obchod, Adresář, Personalistika, Faktury, CFG, LOG, Datové schránky, Servis, Spisová služba. V zadaném textu najdi, které z klíčových slov ho co nejvíce vystihují. Vrať maximálně 4 klíčová slova vystihující text.

2. pokus - výsledek

- Text:

  Žádám opravit datum střídaní tj. v době od: 22.10. 15:15 do 31.10. 7:00 střídá Jan Zelený v pohotovosti, kterou měl Josef Červený

  na : v době od: 22.10. 15:15 do 25.10. 7:00 střídá Jan Zelený v pohotovosti, kterou měl Josef Červený

  (jména změněna)

- Odpověď:

  **Docházka
  Personalistika
  Spisová služba
  Servis**

  Text se zabývá úpravou dat střídaní pracovníků a změnou pohotovosti
  což je základní činnost spojená s **Docházka** (stav přítomnosti/absence) a **Personalistika** (pracovní plánování). Zároveň může být spojeno s **Spisová služba** (dokumentace a evidence činností) a **Servis** (organizace procesů v pracovním prostředí). Ostatní klíčová slova (např. databáze
  upgrade
  aplikace) nejsou v textu přímo zmiňována.

- Za cca 1,5h bylo zpracováno 100 dokumentů, tempo cca 1 dokument za minutu

3. pokus

- Prompt:

  Jsou dána klíčova slova: server, upgrade, aplikace, prohlídka, databáze, instalace, Docházka, Obchod, Adresář, Personalistika, Faktury, CFG, LOG, Datové schránky, Servis, Spisová služba. V zadaném textu najdi, které z klíčových slov ho co nejvíce vystihují, pokud tam není vyslovně uvedeno tak se nepoužije. Může být použito pouze jedno nebo více z nich, ale jen pokud je se tam opravdu hodí. Vrať pouze samotná klíčová slova.

- Výsledek:
  - Nechal jsem běžel jen 100 dokumentů (cca 1,5h)
  - Tempo cca 1 dokumentů za minutu
  - Úprava nastavení GPU offload value: 99



4. pokus

- Prompt:

  Jsi expert na analýzu textu a datovou kategorizaci. Tvým úkolem je analyzovat přiložený text, extrahovat z něj klíčová slova a zařadit ho do jedné nebo více z mých předem definovaných kategorií.

  Zde jsou povolené kategorie:

  - SERVER: vše co se týká serveru mimo pravidelné prohlídky a upgrade serveru, domino serveru, traveler serveru, nomad serveru, linux serveru
  - SPINESYS: úpravy nebo opravy aplikací Obchod, Adresář, Docházka, Personalistika, CFG, LOG, Datová schránka, Faktury, Spisová služba
  - APLIKACE: úpravy jednotlivých aplikací neuvedených v kategorii Spinesys, databáze, úpravy databází, úpravy dat v aplikacích nebo databázích
  - PROHLÍDKA: pravidelná prohlídka serveru
  - UPGRADE: upgrade serveru

  Při analýze dodržuj tato pravidla:

  1. Text zařaď POUZE do kategorií ze seznamu výše. Pokud text neodpovídá žádné, uveď "Nezařazeno".
  2. Vyber maximálně 2 nejrelevantnější kategorie.

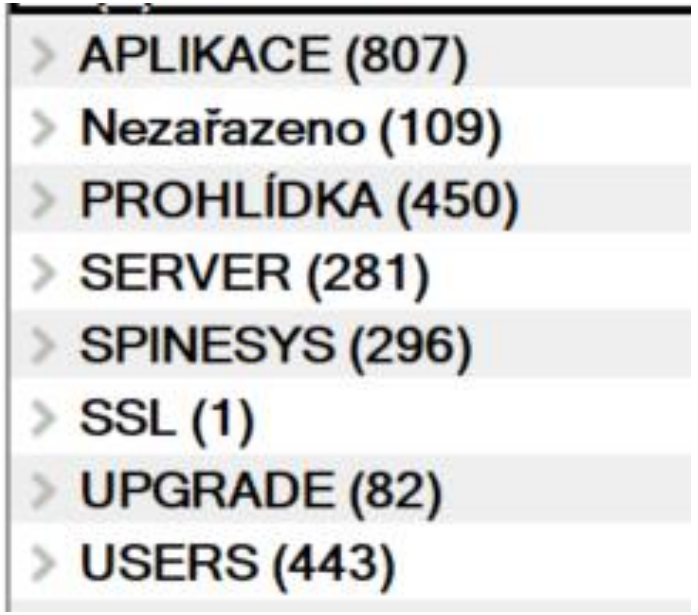
  Formát výstupu vracej striktně takto:

  Kategorie oddělené čárkou

  Zde je text k analýze:

4. pokus - výsledek

- Výsledek
  - Agent běžel cca 3,5h
  - Tempo cca 10 dokumentů za minutu (99% GPU)
  - 1 LLM dotaz a odpověď cca 5s
  - Záměrně byla DB servis na jiném serveru



  - Samotné klíčové slova:
    - 392 dokumentů dostalo 2 hodnoty, zbytek jen 1
    - Zajímavý je ten jeden v kategorii SSL
      - SSL kategorie nebyla v seznamu
      - V textu servisu se nikde o SSL nepíše ... ale je tam zmínka o CSR

Prompt engineering

- Na promptu opravdu hodně záleží
- Je třeba přijmout nového člověka na pozici Prompt inženýra?
- Máme přece AI - stačí si nechat prompt vygenerovat…

Jiný příklad - generování „náhodného“ obsahu

- Pro testování nebo ukázky aplikací
- Není potřeba specializované knihovny, stačí dobrý prompt, např.:

  Vygeneruj náhodný seznam 100 fiktivních osob z České republiky. Pro každou osobu uveď: jméno, příjmení, ulici a číslo popisné, PSČ a město, telefonní číslo a e-mailovou adresu. Používej běžná česká jména, příjmení a reálně vypadající české adresy, ale zajisti, aby všechny údaje byly zcela smyšlené a neodpovídaly skutečným osobám ani existujícím adresám. Výstup vrať ve formátu CSV bez záhlaví, kde jednotlivé položky jsou odděleny čárkami: Jméno,Příjmení,Ulice a číslo,PSČ,Město,Telefonní číslo,E-mailová adresa

- Fun fact - vygeneroval jich 148 než vyčerpal kontextové okno

Příjmení	Jméno	Adresa	Město	PSČ	Telefon	E-mail
Beneš	David	U Kříže 4	Hradec Králové	750 01	614 666 777	david.benes@domain.cz
Beneš	David	U Vodárenského dvoře 24	Olomouc	700 01	648 666 444	david.benes@test.cz
Beneš	Klára	U Vodárenského dvoře 23	Olomouc	700 01	644 666 888	klaudia.benes@domain.cz
Beneš	Lucie	U Křižového pole 29	Hradec Králové	254 01	787 333 666	lucie.benes@test.cz
Beneš	Markéta	U Zlatejho větru 39	Brno	200 01	657 888 111	martka.benes@test.cz
Beneš	Martin	U Vodárny 28	Pardubice	150 00	744 555 777	martin.benes@demo.net
Beneš	Pavel	U Zlatejho dvoře 24	Olomouc	700 01	632 666 888	pavel.benes@domain.cz
Beneš	Petr	U Kříže 20	Liberec	500 01	777 777 111	petr.benes@demo.net
Beneš	Petr	U Vodárenského dvoře 25	Olomouc	700 01	652 666 333	petr.benes@example.com
Beneš	Tomáš	U Mlýna 15	Brno	200 01	637 888 111	tomas.benes@example.com
Beneš	Zuzana	U Kříže 21	Liberec	500 01	781 777 333	zuzana.benes@test.cz
Benešová	Lucie	U Mlýna 15	Ostrava	280 01	725 444 555	lucie.benesova@example.com
Benešová	Zuzana	U Kříže 7	Hradec Králové	254 01	622 222 777	zuzana.benesova@example.com
Bílá	Věra	U Křižového pole 32	Olomouc	740 01	608 777 666	vera.bila@domain.cz
Bílek	Jan	U Hradu 25	Hradec Králové	220 01	617 888 222	jan.bilek@domain.cz
Bůček	Vítěk	U Kříže 5	Olomouc	700 01	724 222 888	vitek.bucek@test.cz
Bučková	David	U Mlýna 36	Brno	725 01	790 999 333	david.buckova@test.cz
Bučková	Eliška	U Kříže 22	Liberec	500 01	785 777 444	eliska.buckova@example.com
Bučková	Hana	U Křižového pole 6	Brno	725 01	737 777 222	hana.buckova@example.com
Bučková	Jan	U Zlatejho dvoře 36	Plzeň	360 01	651 444 888	jan.buckova@test.cz
Bučková	Karel	U Kříže 12	Prague	170 00	630 222 555	karel.buckova@test.cz
Bučková	Klára	U Křižového pole 20	Pardubice	150 00	752 555 999	klaudia.buckova@example.com
Bučková	Marta	U Křižového pole 28	Hradec Králové	254 01	783 333 111	marta.buckova@demo.net
Bučková	Radek	U Křižového pole 23	Hradec Králové	254 01	763 333 111	radek.buckova@test.cz
Bučková	Tomáš	U Křižového pole 27	Hradec Králové	254 01	779 333 777	tomas.buckova@example.com
Bučková	Tomáš	U Vodárny 35	Prague	170 00	654 222 666	tomas.buckova@test.cz
Bučková	Věra	U Mlýna 20	Brno	725 01	625 888 666	vera.buckova@example.com
Bučková	Vítěk	U Křižového pole 26	Hradec Králové	254 01	775 333 666	vitek.buckova@test.cz
Bučková	Vítěk	U Zlatejho větru 31	Brno	725 01	758 999 111	vitek.buckova@example.com
Bučková	Zuzana	U Zlatejho dvoře 3	Pardubice	310 00	730 555 888	zuzana.buckova@test.cz
Černá	Klára	U Mlýna 11	Plzeň	360 01	627 444 888	klaudia.cerna@test.cz
Černá	Lucie	U Kříže 16	Brno	200 01	645 888 333	lucie.cerna@test.cz
Černá	Lucie	U Vodárenského dvoře 21	Liberec	500 01	753 777 222	lucie.cerna@demo.net
Černá	Markéta	U Křižového pole 15	Pardubice	150 00	748 555 222	martka.cerna@test.cz
Černá	Marta	U Kříže 14	Brno	200 01	641 888 222	marta.cerna@domain.cz
Černá	Marta	U Mlýna 17	Hradec Králové	254 01	732 444 555	marta.cerna@demo.net
Černá	Vítěk	U Vodárny 34	Prague	170 00	650 222 555	vitek.cerna@domain.cz
Černá	Zuzana	U Vodárenského dvoře 26	Olomouc	700 01	656 666 888	zuzana.cerna@domain.cz
Černý	David	U Mlýna 3	Plzeň	360 01	739 333 444	david.cerny@test.cz
Černý	Jan	U Vodárny 14	Plzeň	350 01	725 111 555	jan.cerny@example.com
Dvořák	Eliška	U Vodárenského dvoře 27	Olomouc	700 01	660 666 222	eliska.dvorak@test.cz
Dvořák	Jan	U Vodárenského dvoře 29	Olomouc	700 01	668 666 333	jan.dvorak@domain.cz
Dvořák	Karel	U Mlýna 27	Pardubice	150 00	772 555 222	karel.dvorak@test.cz
Dvořák	Karel	U Vodárenského dvoře 28	Olomouc	700 01	664 666 888	karel.dvorak@example.com
Dvořák	Martin	U Mlýna 22	Liberec	500 01	761 777 444	martin.dvorak@example.com
Dvořák	Pavel	U Zlatejho větru 15	Prague	110 00	615 333 111	pavel.dvorak@test.cz
Dvořák	Petr	U Vodárny 30	Plzeň	360 01	635 444 888	petr.dvorak@domain.cz
Dvořák	Radek	U Vodárenského dvoře 30	Olomouc	700 01	672 666 222	radek.dvorak@test.cz
Dvořák	Tomáš	Bělejová 89	Plzeň	360 01	603 999 888	tomas.dvorak@test.cz
Dvořáková	Eliška	U Mlýna 9	Prague	170 00	609 888 444	eliska.dvorakova@test.cz
Dvořáková	Eliška	U Zlatejho dvoře 22	Hradec Králové	750 01	734 111 999	eliska.dvorakova@example.com
Dvořáková	Hana	U Vodárenského dvoře 19	Brno	200 01	629 888 222	hana.dvorakova@domain.cz
Dvořáková	Lucie	U Hradu 29	Pardubice	150 00	611 666 888	lucie.dvorakova@domain.cz
Dvořáková	Marta	U Křižového pole 19	Liberec	500 01	624 666 111	marta.dvorakova@test.cz
Dvořáková	Pavel	U Vodárenského dvoře 22	Liberec	500 01	757 777 333	pavel.dvorakova@test.cz
Dvořáková	Věra	U Křižového pole 24	Hradec Králové	254 01	767 333 666	vera.dvorakova@example.com
Havel	Karel	U Kříže 33	Liberec	500 01	722 333 222	karel.havel@test.cz
Havlíček	Karel	U Vodárny 21	Prague	170 00	738 999 888	karel.havlicek@demo.net
Havlíček	Marta	U Mlýna 24	Brno	725 01	766 999 333	marta.havlicek@test.cz
Havlíček	Petr	U Zlatejho větru 25	Hradec Králové	254 01	743 333 111	petr.havlicek@example.com
Havlíček	Tomáš	U Mlýna 28	Olomouc	700 01	728 999 333	tomas.havlicek@example.com
Havlíček	Věra	U Mlýna 21	Prague	170 00	642 222 555	vera.havlicek@test.cz
Havlíček	Vítěk	U Vodárenského dvoře 20	Brno	200 01	633 888 444	vitek.havlicek@test.cz
Havlíčková	Hana	U Kříže 13	Brno	725 01	754 999 444	hana.havlicikova@test.cz
Havlíčková	Klára	U Kříže 5	Pardubice	150 00	735 333 777	klaudia.havlicikova@demo.net
Havlíčková	Lucie	U Mlýna 26	Brno	725 01	770 999 444	lucie.havlicikova@example.com
Havlíčková	Pavel	U Mlýna 28	Brno	725 01	774 999 555	pavel.havlicikova@demo.net

Druhá část

Proč RAG?

- LLM obsahují jen obecné znalosti
  - navíc jen do určitého bodu v čase (kdy byl model trénován)

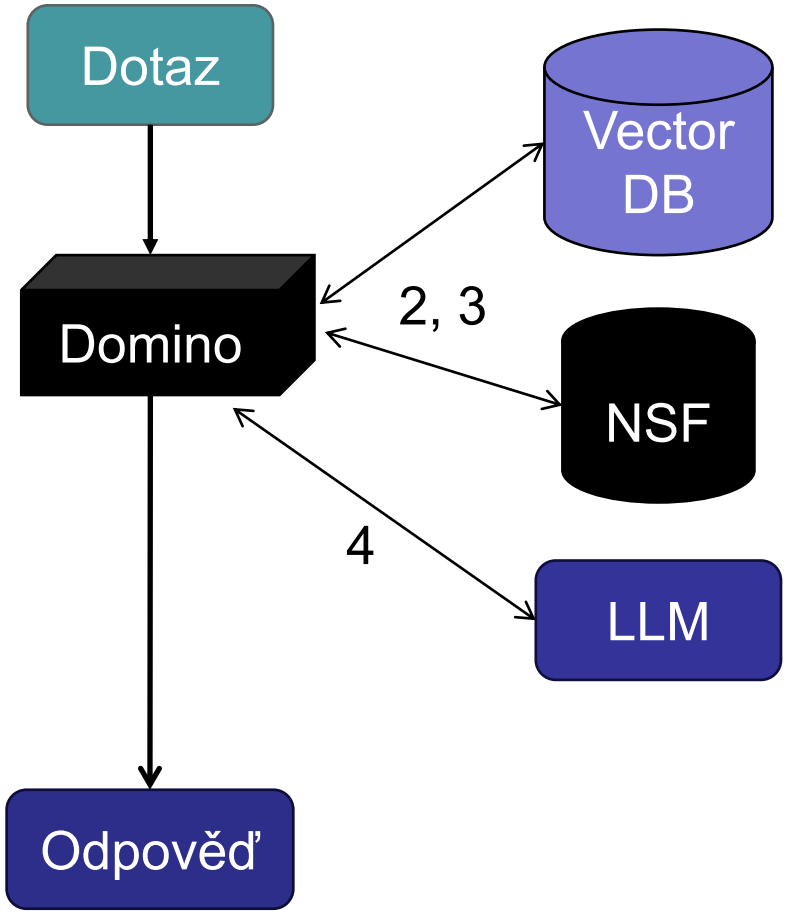
- Co když potřebuji odpověď na základě znalostí firemních dokumentů?

- V zásadě jsou tři možnosti:
  1. Velké kontextové okno
  2. RAG (Retrieval-Augmented Generation)
  3. MCP (Model Context Protocol)

RAG – co vlastně dělá?

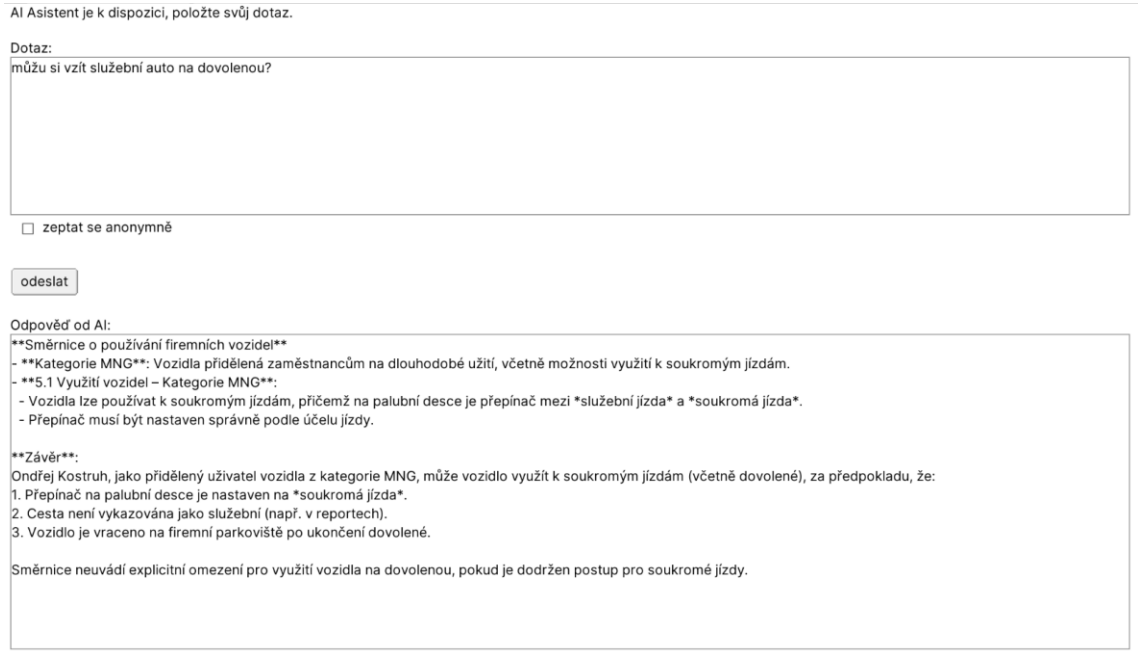
- Jednoduše řečeno – doplní kontext k dotazu

- A detailněji:
  1. kontrola Command v konfiguraci
  2. Podle dotazu vyhledá ve vektorové DB (sémanticky nikoliv fulltextově) odkazy na relevantní dokumenty (včetně příloh **!**)
  3. Nalezené dokumenty zkontroluje (včetně ACL a Readers polí **!**)
  4. K dotazu odesílanému do LLM inference engine přidá kontext



RAG - ukázka

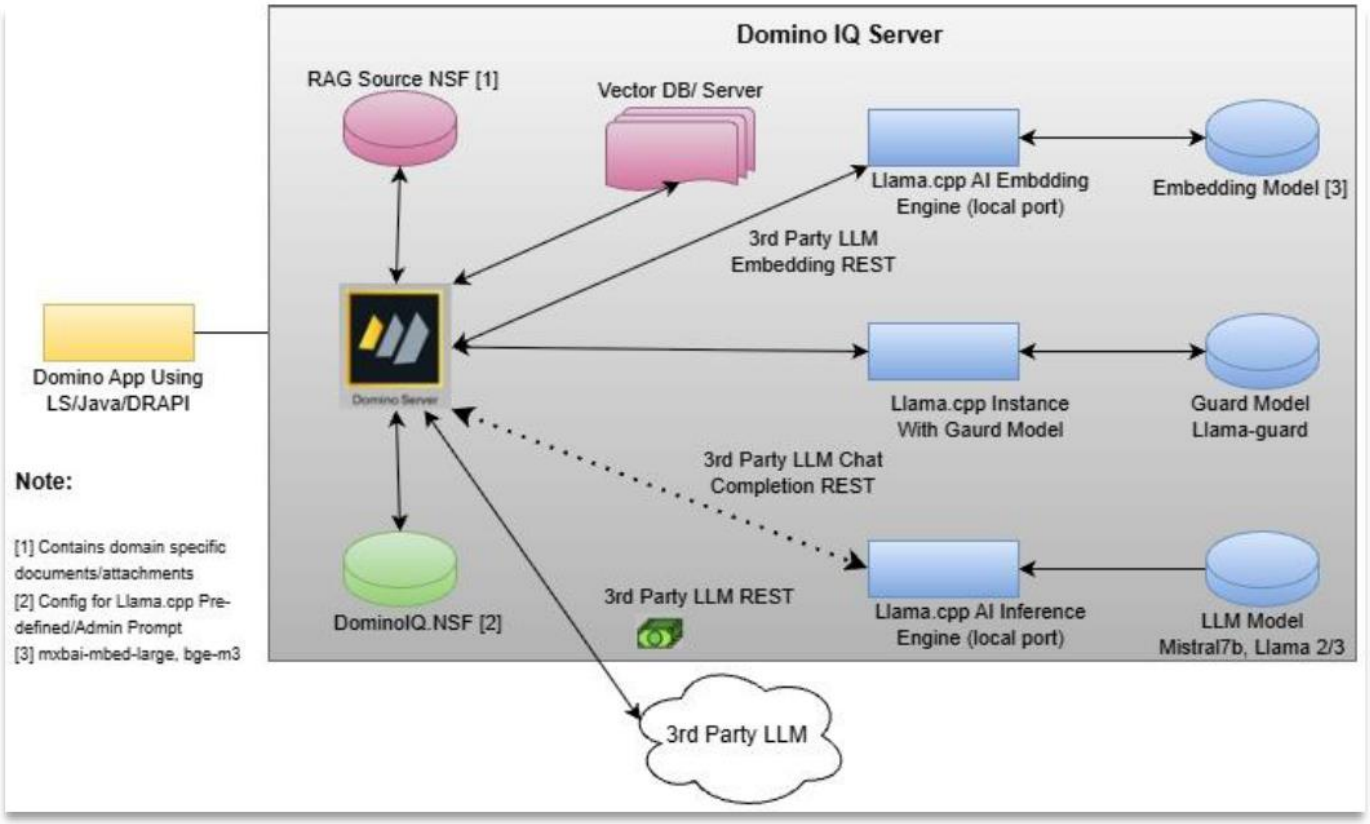
- Aplikace Směrnice
  - Jednoduchá aplikace pro tvorbu a evidenci firemních směrnic
  - Tvorba směrnic - pomocí LLM
  - Vyhledávání informací ve směrnicích pomocí RAG



RAG – co potřebuje

- Vector DB (Qdrant)
- Embedding model

+ konfiguraci



https://help.hcl-software.com/domino/14.5.1/admin/conf_iq_rag_support.html

RAG – volba modelu

- Embedding model (*embed* v názvu jako nápověda ne podmínka)
  - Doporučuji začít modely z dokumentace

- Parametry
  - Velikost modelu
    - nároky na paměť, ale i rychlost
  - Velikost kontextového okna (např. podle velikosti indexovaného textu)
    - dělení textu na chunks
  - Velikost NSF a podle toho výsledné vektorové DB
  - Jazyky

Embedding Modely - příklady

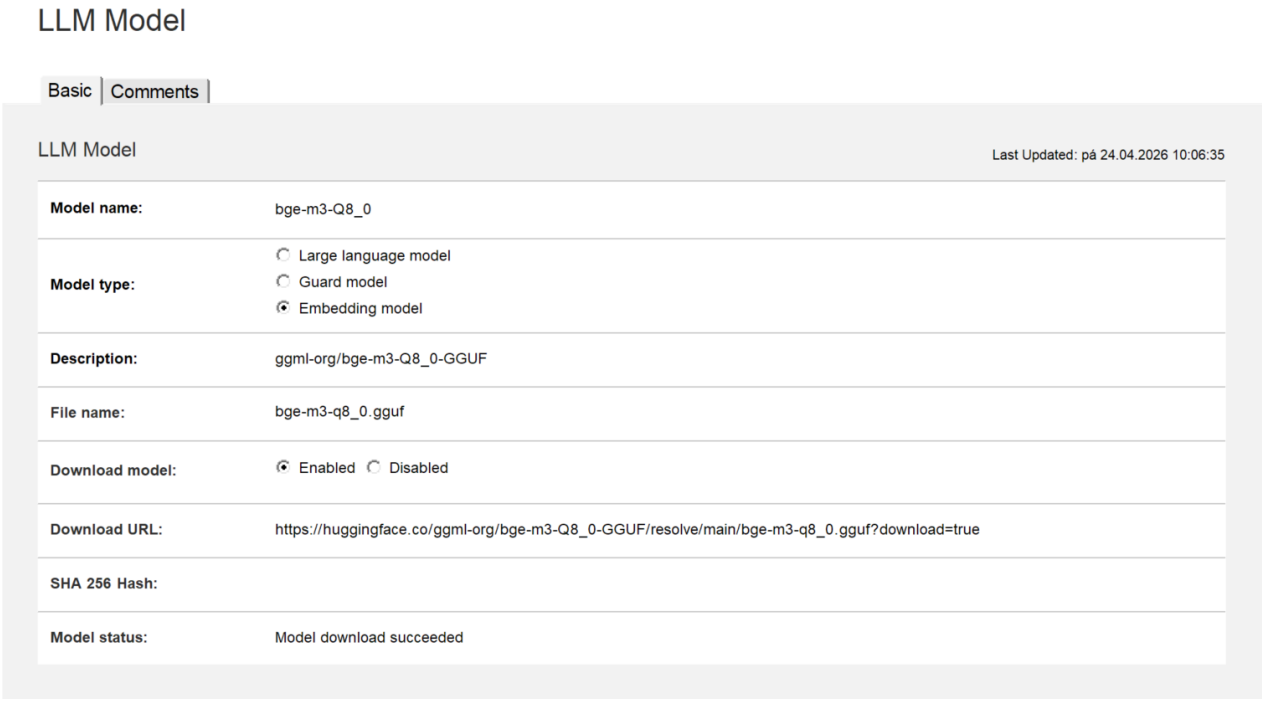
Embedding Model Identifier	Parameter Count	Output Dimensionality	Context Window (Tokens)	Primary Architecture / Specialization
nomic-embed-text-v1.5	137M	768 (Flexible down to 64)	8192	Matryoshka Representation Learning; highly flexible storage optimization.
nomic-embed-text-v2-moe	475M (305M Active)	768	512	Mixture of Experts (MoE); state-of-the-art multilingual retrieval (~100 languages).
bge-m3	567M	1024	8192	Dense, sparse, and multi-vector capabilities; highly multilingual support.
mxbai-embed-large-v1	335M	1024	512	English retrieval optimization; native support for binary and int8 quantization.
snowflake-arctic-embed-m-v1.5	109M	768 / 256	512	Enterprise retrieval; aggressive int8/int4 scalar quantization for massive RAM reduction.

RAG – příprava

- HCL Domino 14.5.1 (Linux nebo Windows) + Domino IQ
- Transaction Logging & DAOS
- NSF pro RAG na tomto serveru (Domino IQ serveru)
- Fulltext index
  - Index attached files
  - FT_SKIP_IGNORE_FIELD=1
- LLM v lokálním módu

RAG – nastavení 1

- Embedding models

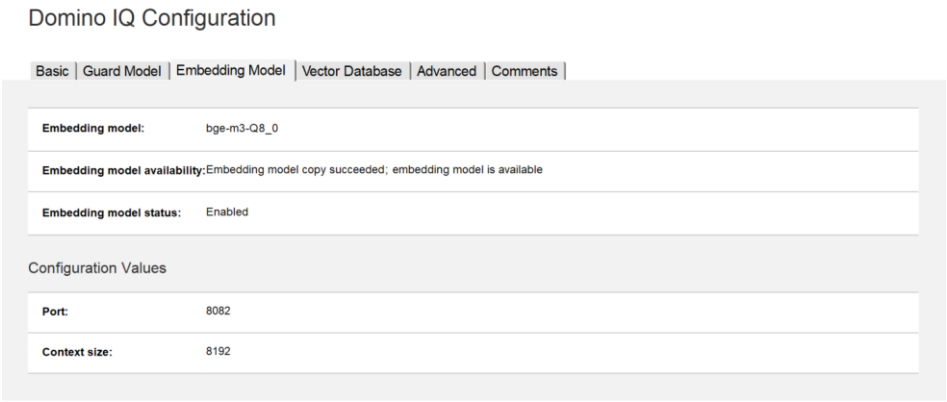


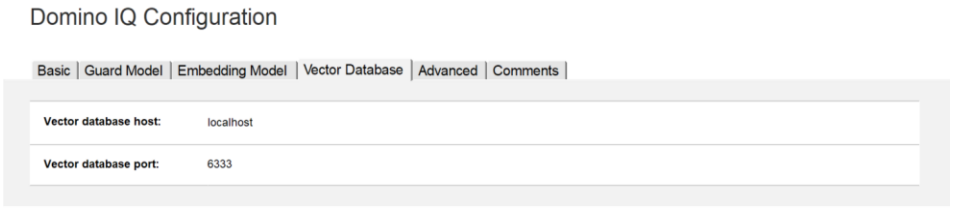
https://help.hcl-software.com/domino/14.5.1/admin/conf_add_embedding_document.html

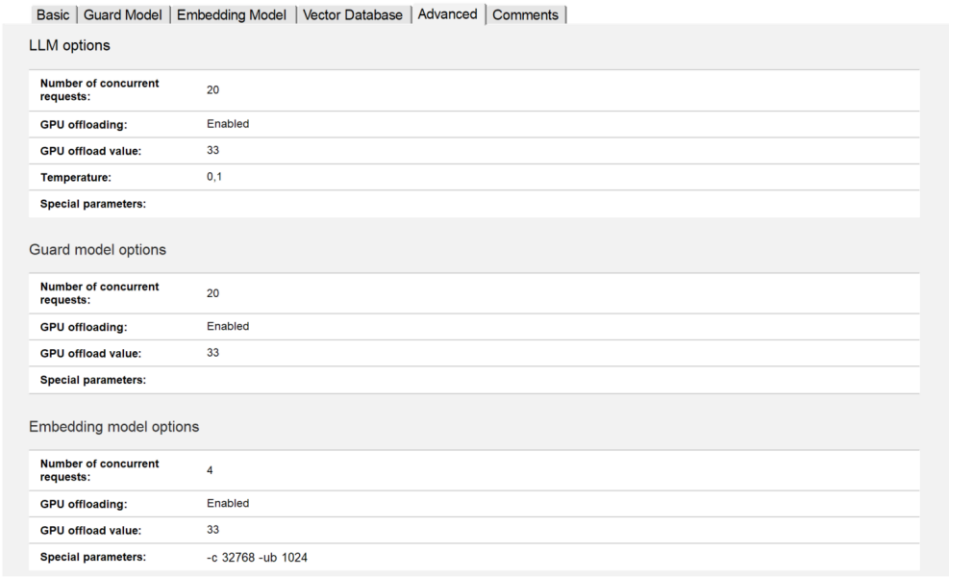
RAG – nastavení 2

- Configurations
  - Embedding Model
  - Vector
  - Advanced
    - GPU offloading
    - Special parameters

https://help.hcl-software.com/domino/14.5.1/admin/conf_create_embed_and_vector_dbs.html

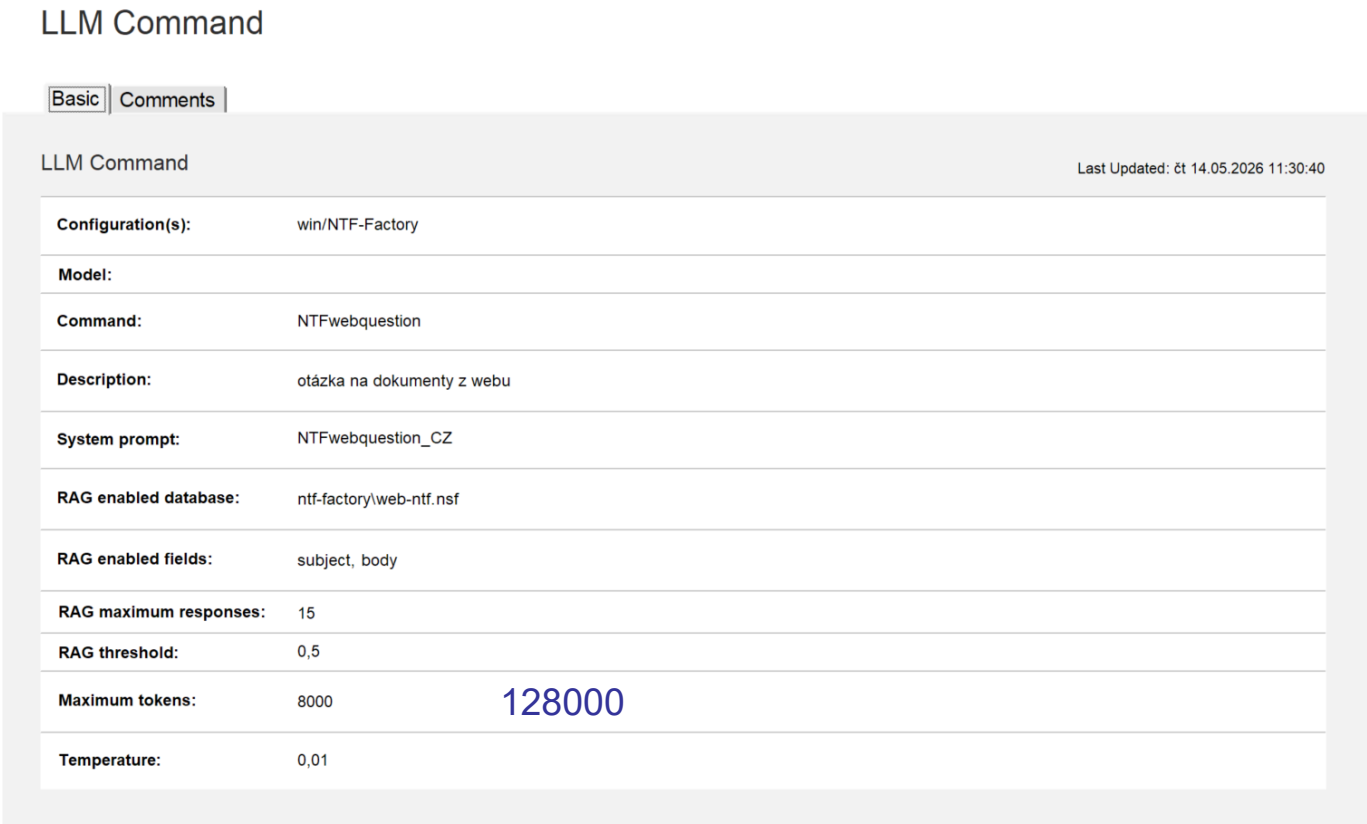






RAG – nastavení 3

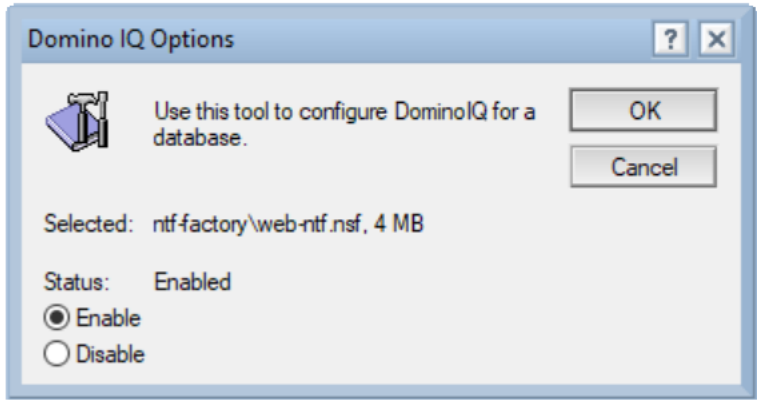
- Command
  - DB (nsf)
  - Pole (fileds)
  - Nastavení
- Prompt



https://help.hcl-software.com/domino/14.5.1/admin/conf_command_doc_for_rag.html

RAG – povolit DB pro

- Je třeba povolit DB pro AI
  - Admin klient 14.5.1



https://help.hcl-software.com/domino/14.5.1/admin/conf_enable_disable_vector_db.html

RAG – indexace - updall

- První indexace
  - load updall -w <database.nsf>
  - Vytvoří adresář <database>.vec
- Incremental update (změna v nsf)
  - load updall -k <database.nsf>
- Smazání indexu (změna modelu nebo jeho konfigurace)
  - load updall -w -d <database.nsf>
- Neprobíhají automatické obnovy indexu
  - volat ručně
  - naplánovat pomocí Programs v Domino Directory
  - vhodné pro použití IND (indirect files)

https://help.hcl-software.com/domino/14.5.1/admin/conf_enable_disable_vector_db.html
https://help.hcl-software.com/domino/14.5.1/admin/admn_settingascheduleforupdallinaprogramdocument_t.html
https://help.hcl-software.com/domino/14.5.1/admin/admn_using_indirect_files_to_run_database_maintenance_tasks_t.html

https://marketplace.hcl-software.com/domino/catalog/domino-ind-file-editor

Update - 14.5.1 Interim Fix 1

- Fixed an issue where the HCL DominoIQ RAG feature is affected by a Broken Access Control vulnerability. Under certain circumstances, document level access restrictions will be ignored when determining what data to return from an AI query. This could enable an authenticated attacker to view sensitive data.

- Security Bulletin: HCL DominoIQ is affected by a Broken Access Control vulnerability (CVE-2026-21836)

https://support.hcl-software.com/csm?id=kb_article&sysparm_article=KB0130876
https://support.hcl-software.com/csm?id=kb_article&sysparm_article=KB0130932

Poznámky z provozu RAGu

- Indexuje a zpracuje celou DB
  - Pozor na statusy dokumentů -např. „draft“, „v přípravě“, apod.
- Malá pole, resp. krátké hodnoty
- Pojmenování polí vs. jejich hodnoty
  - Bude jim LLM rozumět?
- Sumarizace polí, popř. sumarizace celého dokumentu
  - Formát Markdown

https://www.markdownguide.org/basic-syntax/

Troubleshooting - Domino IQ

- Logy v IBM_TECHNICAL_SUPPORT
  - dominoiq_server.log
  - dominoiq_embed.log
  - dominoiq_vector.log
  - dominoiq_guard.log

DEBUG parametry do notes.ini

- DEBUG_DOMIQ=1
  - general debugging output
- DEBUG_DOMIQ_LLMREQUEST=1
  - requests (cURL) and responses (JSON)
- DEBUG_DOMIQ_PAYLOAD=1
  - extraordinarily verbose flag
  - detailed connection/request/response tracing

https://help.hcl-software.com/domino/14.5.1/admin/conf_iq_troubleshooting_tips.html

Změny v konfiguraci

- Změny v Prompts & Commands
  - Cache - 5 min.

- Změny v modelu
  - Restart serveru
  - `tell dominoiq unload`, `tell dominoiq resume`

- Změny v parametrech
  - Zejména -c (context window) v Configuration \ Advanced

HW a měření

HW

Minimum hardware requirements for a DominoIQ server:

- 4 CPUs
- 8 GB of RAM
- NVIDIA GPU is mandatory
  - minimum of 8 GB of memory
  - compute capability (CC) of 5.2 or higher (doporučeno 8.0 a výše)

https://help.hcl-software.com/domino/14.5.1/admin/domino_iq_server.html#domino_iq_server__section_wgl_dwc_q2c
https://developer.nvidia.com/cuda/gpus

HW 1 – bez GPU

Lenovo ThinkStation P/P3 Gen 2

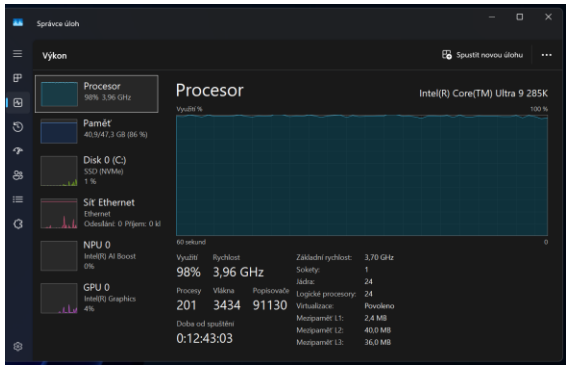


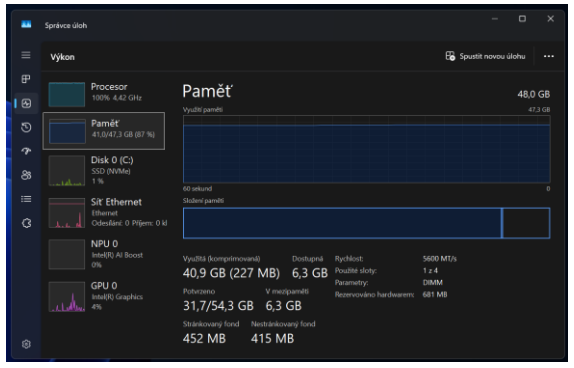
- Intel Core Ultra 9 285K, 36MB Intel Smart Cache
- 24 jader / 24 vláken
- 3.7 GHz / 5.7 GHz
- 1x 48GB UDIMM DDR5-5600 Non-ECC
- Integrated Intel Graphics

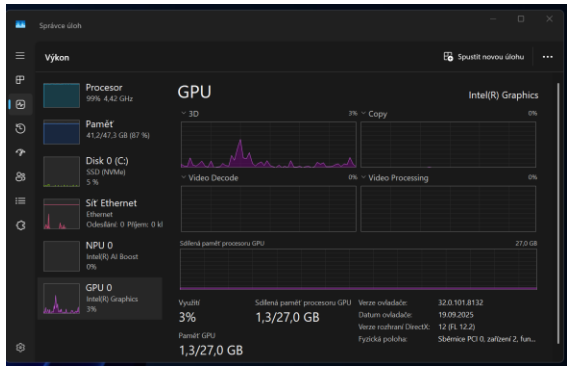
- Windows
  - On Windows servers, Domino IQ server is able to use both GPU memory and half of system RAM, which is treated as Shared RAM

Jednoduché testy

úloha	bez GPU
Překlad (CZ -> EN, 2 věty)	170s
Dotaz s RAG	80s
Přestylizace textu	70s







HW 2 – přidána GPU

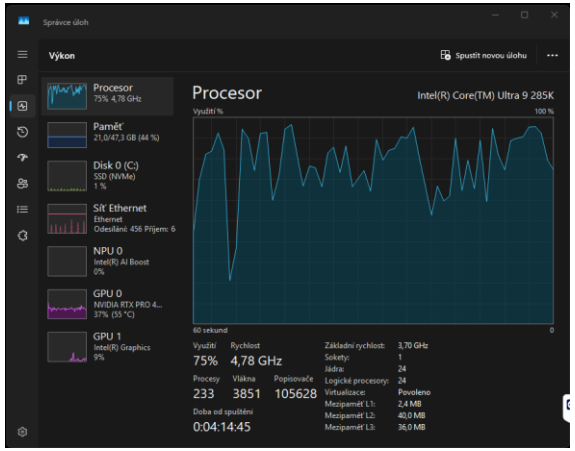
GPU NVIDIA RTX PRO 4000 (Blackwell)

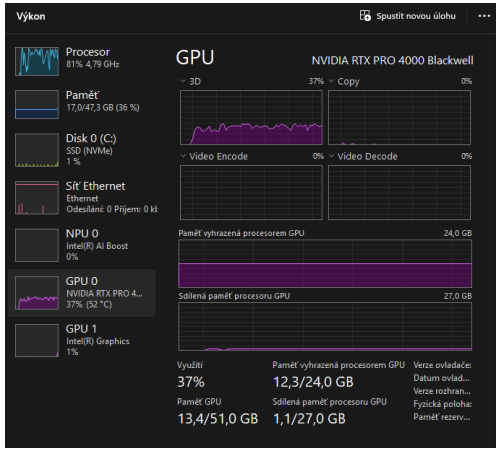
- 24 GB GDDR7
- NVIDIA® CUDA® cores 8,960
- AI Performance: 1178 AI TOPS
- Single-Precision performance 37 TFLOPS
- RT Core performance 112 TFLOPS
- 672 GB/s of memory bandwidth
- CC 12.0
- Maximální spotřeba 140W



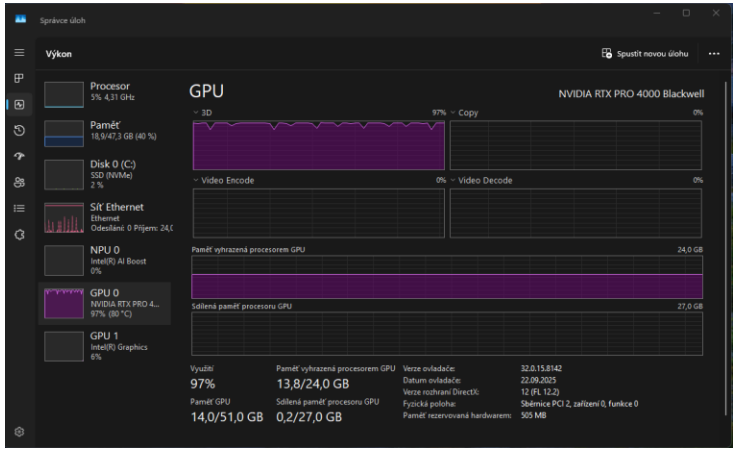
Jednoduché testy 2

úloha	bez GPU	GPU 33%	GPU 99%
Překlad	170s	25s	8s
Dotaz s RAG	80s	17s	6s
Přestylizace textu	70s	-	7s





GPU offload value: 33



GPU offload value: 99

Spotřeba el. energie

úloha	bez GPU	S GPU
Spotřeba v klidu	12W	25W
updall -w (RAG index >2000 dokumentů)	265W	100W
LLMRequest	160W	260W







Přehled testů z pohledu spotřeby

úloha	bez GPU	S GPU
překlad	170s	8s
dotaz s RAG	80s	6s
přestylizace textu	70s	7s
Spotřeba v klidu	12W	25W
updall -w (RAG index)	265W	100W
- test DB (>2000 dokumentů)	1 min	1 min
LLMRequest	160W	260W
	50s	5s

V jeden okamžik běžela vždy jen jedna úloha/dotaz.

HW - bod poslední - cena

- Není to server, ale spíše výkonná pracovní stanice
  - není snadné sehnat některé HW komponenty

- Celková cena HW pod 100.000 Kč (bez DPH) ~ 4.000 EUR
  - ½ z toho stanice
  - ½ GPU

- HCL Domino zatím pouze na x86-64 (a NVIDIA pro Domino IQ)

Závěrem

- A co bude dále?
  - AI Agenti
    - Potřeba MCP (Model Context Protocol) - rozšíření API
    - Provádí akce, činnosti, úkoly
    - Vícekrokově…
  - Prompt engineering vs. vícekrokové řešení problému (výzkum)
  - Zvětšování prompt/context window
  - DATA !
    - Budou data čitelná pro AI? Bude jim „rozumět“?