SUTOL

Symposium2025

11. června • GreenPoint, Praha

Partneři akce

HCLSoftware

TD SYNNEX

NGSS Next Generation Security Solutions

Whitesoft

DOWNSTREAM

your SYSTEM

ISECO

NTF FACTORY

pristo

SUTOL Symposium 2025

11. června • GreenPoint, Praha

Domino IQ (nejen) pro vývojáře

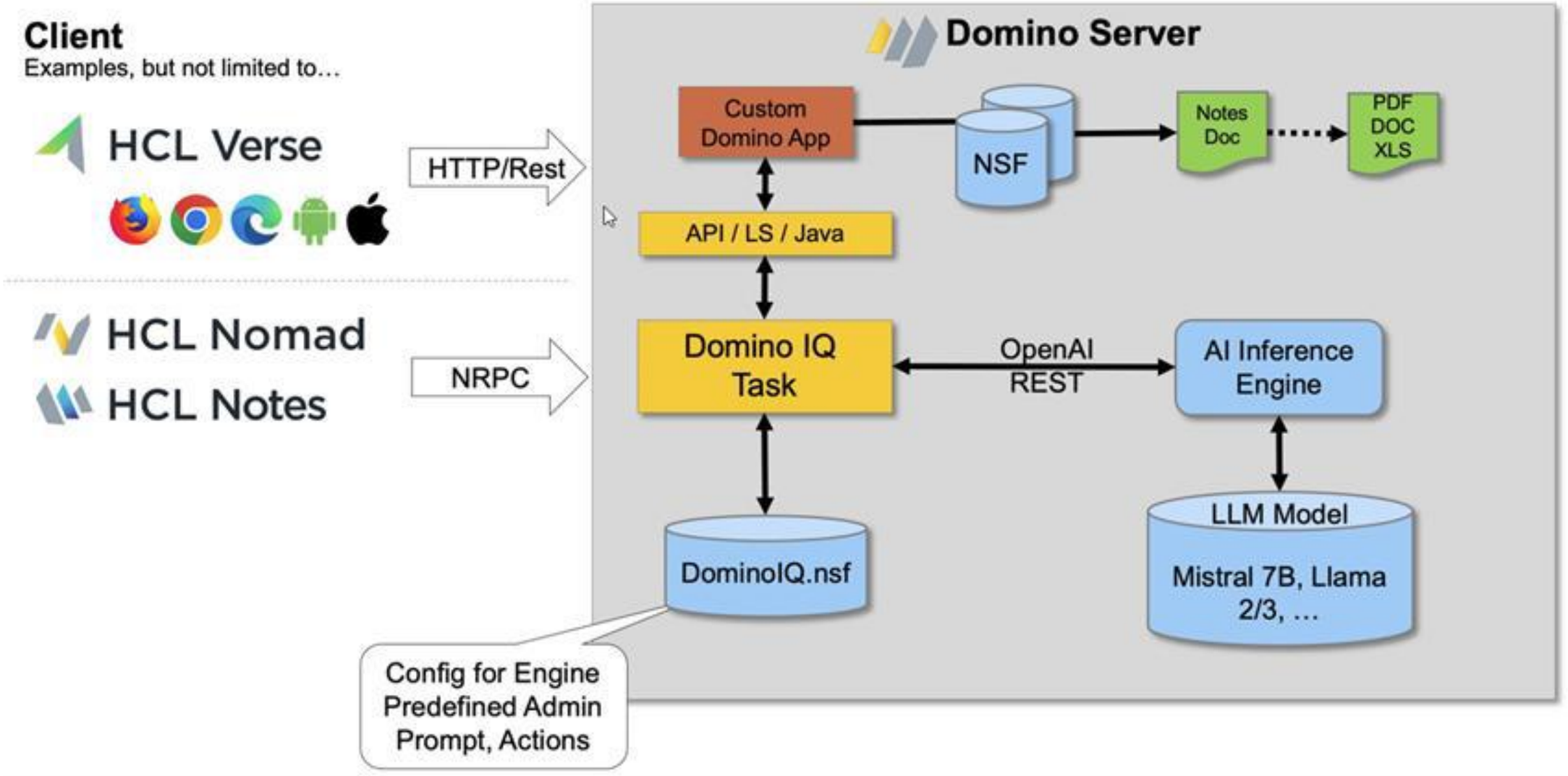
Ondřej Kostruh

NTF Factory

Disclaimer

- Tento krát by to bez něj nešlo…
- Prezentace připravena na verzi 14.5 EA3
- Finální verze se může lišit

Co tedy je to Domino IQ?



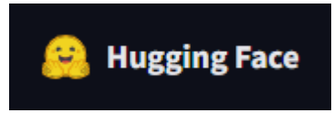
LLM ?

- Large Language Model – Velký jazykový model

- Systém založený na neuronových sítích natrénovaný na generování textů
  - Hledá nejpravděpodobnější pokračování textu
  - Občas (podle nastavení) nevybere pokračování s nejvyšší pravděpodobností

- Tréning LLM
  - Obrovské množství textů (celé knihovny, dostupné texty z internetu)
  - Přesný zdroj ovšem je málokdy znám
  - Velká cena za „tréning“ (podle rozsahu v milionech $)

- Existují miliony (!) LLM – viz. Huggingface.co
  - Liší se funkcí – generování, překlady, jazyky, obory, …



LLM není jeden? Který tedy vybrat?

- Otázka asi nejzásadnější…

- Podle HW (výkon, velikost paměti)
  - V ideálním případě by se model měl „vejít“ do RAM
  - Menší budou rychlejší, větší zase přesnější nebo „mít lepší nápady“

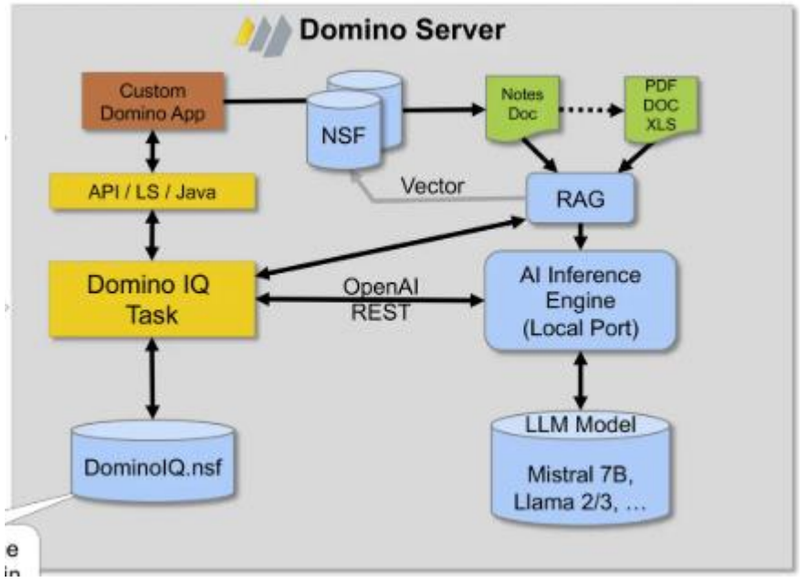
- Podle účelu
  - Generování textů
  - Překlady
  - Sumarizace, generování abstraktů
  - Generování zdrojových kódů
  - Chatboty

Omezení LLM

- Model je state-less
  - otázka – odpověď
  - kontext je třeba vložit do dotazu

- Fine-tuning
  - Jde vlastně o nové generování LLM se specifickými informacemi
    - Nový obor, vlastní data, apod.
  - Obecně se málo používá kvůli náročnosti = ceny

- RAG - Retrieval-augmented generation
  - Doplnění kontextu
  - Je třeba doplnit správný kontext – ale který to je?
  - Vektorizace DB (třeba v příští verzi …)



Halucinace

- LLM netuší co je pravda a co ne – hledá nejpravděpodobnější pokračování textu
- Generuje čtivé a líbivé texty, obsah si generuje sám s cílem potěšit uživatele odpovědí
- Občas se zdá, že čím je větší nesmysl s o to větší jistotou ho tvrdí…
- Temperature
  - Náhodnost výstupu
  - Rozhoduje zda se použije nejpravděpodobnější pokračování nebo jiné méně pravděpodobnější
  - <1 – dává nejpravděpodobnější výsledky – omezuje náhodnost textů
  - >1 – zvyšuje náhodnost textů
  - Přirovnání k ‰ alkoholu v krvi

Legislativa

- AI Act (Nařízení)
  - Rozděluje AI systémy podle možného rizika (pro lidi / občany EU)
    - Neakceptovatelné riziko – je zakázáno (např. social scoring, manipulace)
    - Vysoce riziková AI – je regulováno (nábory, HR, finanční služby, ovlivňování)
    - Omezené rizika & nízké rizika - „běžné“ chatboty, asistenti
      - Uživatele musí být poučeni, musí si být vědomi že pracují s AI a co to znamená
      - Generovaný obsah bude označen nebo identifikovatelný, detekovatelný
  - V platnost vstoupil 1. srpna 2024 - zavádění je postupné podle rizikovosti

- GDPR
  - Zpracování osobních dat pomocí AI
  - Předávání dat …

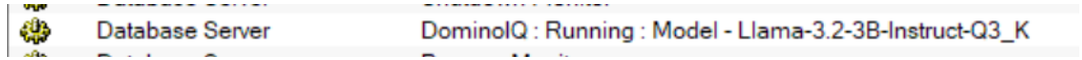
- NIS2
  - Zabezpečení dat – jaké informace máme pod kontrolou
  - RA

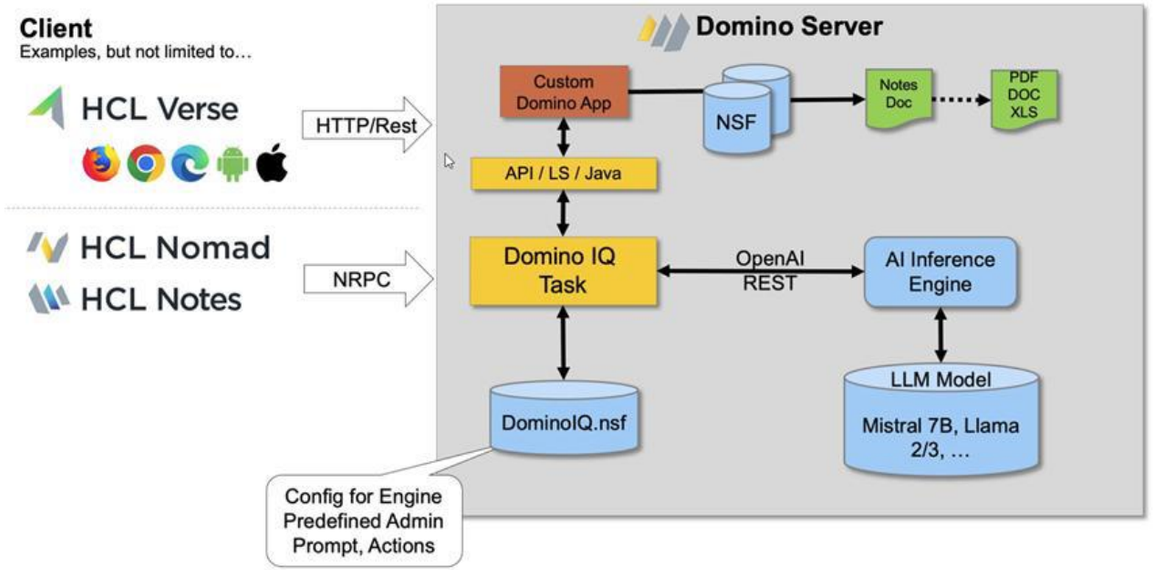
Server pro Domino IQ – lokální provoz LLM

- HCL Domino 14.5 (od EA3)
- Licence CCB
- HW !
  - NVIDIA GPU + RAM (RTX …)
  - RAM
- NVIDIA CUDA Toolkit
- Podle výkonu HW se odvíjí délka čekání na odpověď a výběr LLM
  - …v řádu sekund
  - Jak dlouho vydrží uživatel čekat?

Domino IQ

- Integrováno do Domina
  - Windows & Linux





- Podporovány jsou standardní LLM
  - GGUF formát - lokálně
  - OpenAI REST API – remote

- Nové LotusScript a Java třídy a jejich metody (od V14.5)

- Vzorové příklady v poštovní šabloně dodané v 14.5
  - Reply a Summarize

Instalace Domino IQ

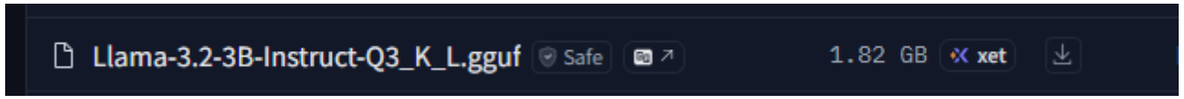
1. Samostatně ke stažení z my.hcltechsw.com (s platnou podporou)

Llama Server

Name	Description	Platform	Released	Size	Actions
LlamaServer_14.5_Linux_EA3.zip	LLama Server for Domino IQ Linux - Early Access April 2025	linux	22 Apr 2025	849.71 MB	
LlamaServer_14.5_Win_EA3.zip	LLama Server for Domino IQ Windows - Early Access April 2025	windows	22 Apr 2025	796.52 MB	

2. ZIP se rozbalí do programové složky

3. Stažení a uložení LLM – Domino Data / llm_models
   - Např. lmstudio-community/Llama-3.2-3B-Instruct-GGUF



Hugging Face

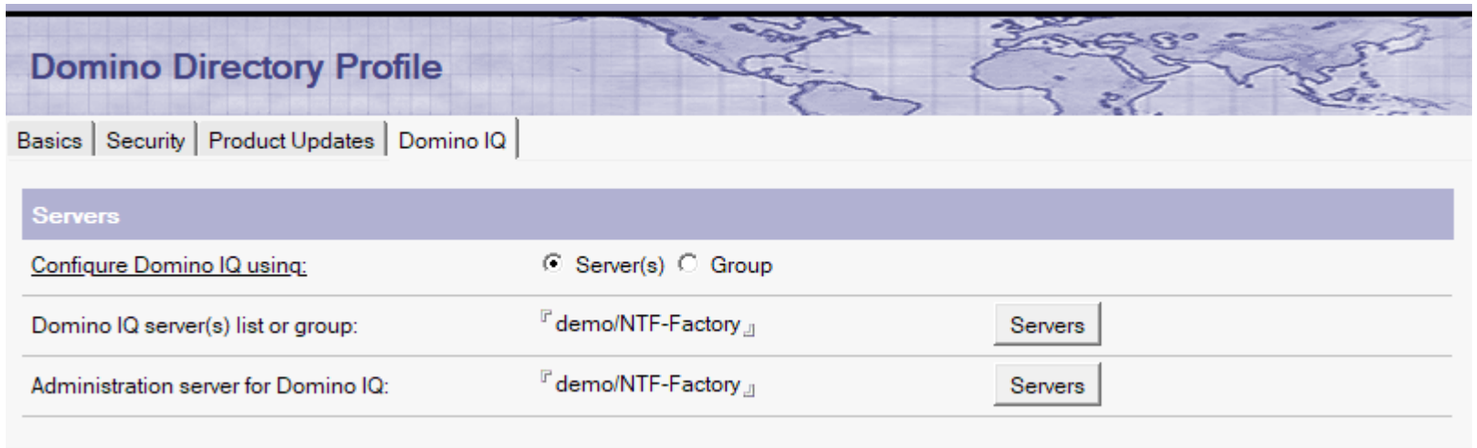
   - Nebo stažení přímo na server v rámci konfigurace (viz. dále)

4. Konfigurace

Konfigurace – Domino Directory Profile

- Specifikace, na kterém serveru je Domino IQ

- Může být jeden nebo libovolné množství serverů



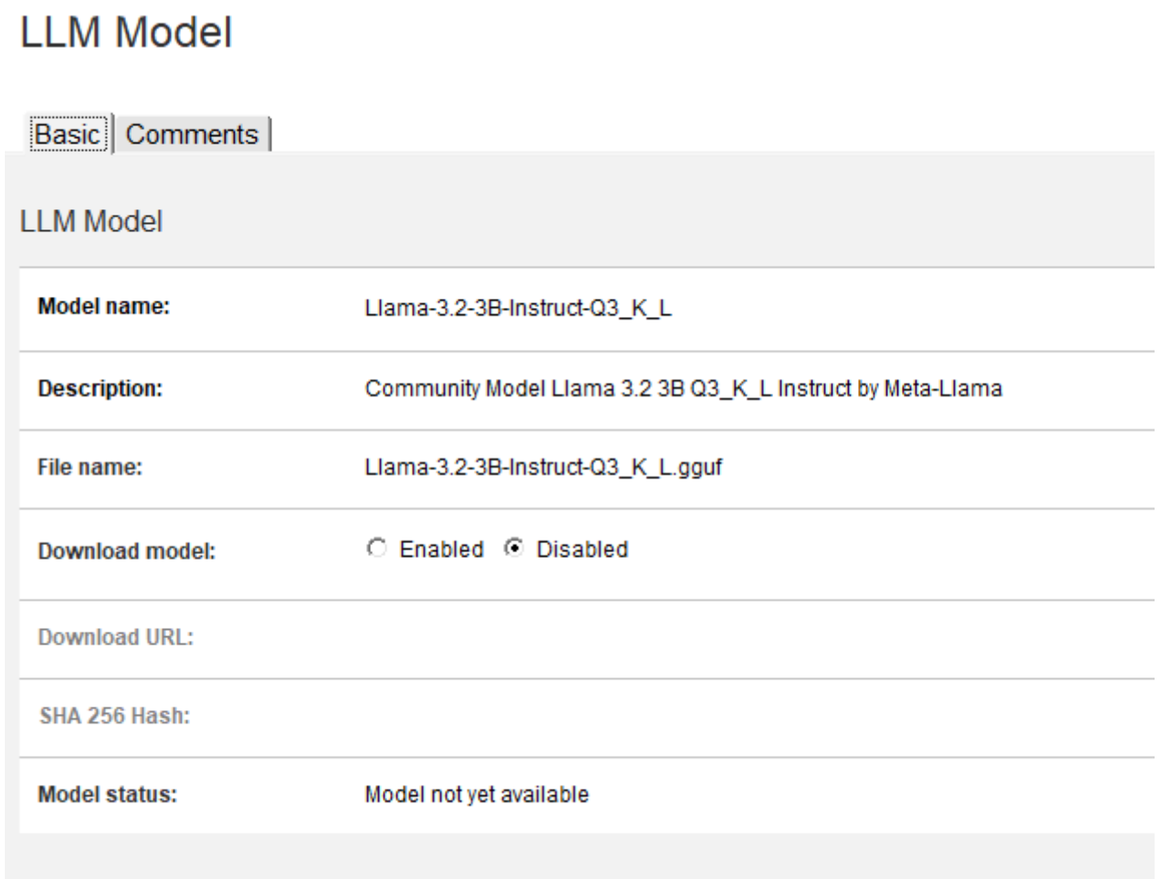
- Administrační server
  - konfigurace pro všechny servery
  - může stahovat modely (LLM) pro ostatní servery

Konfigurace - pokračování

- Domino IQ Configuration
  - NSF aplikace - dominoiq.nsf
  - Automaticky vytvořena po startu Domina
- Configurations
  - Local vs. Remote LLM
  - Pro Remote - endpoint (URL, API key)
- Commands
- System Prompts
- Models

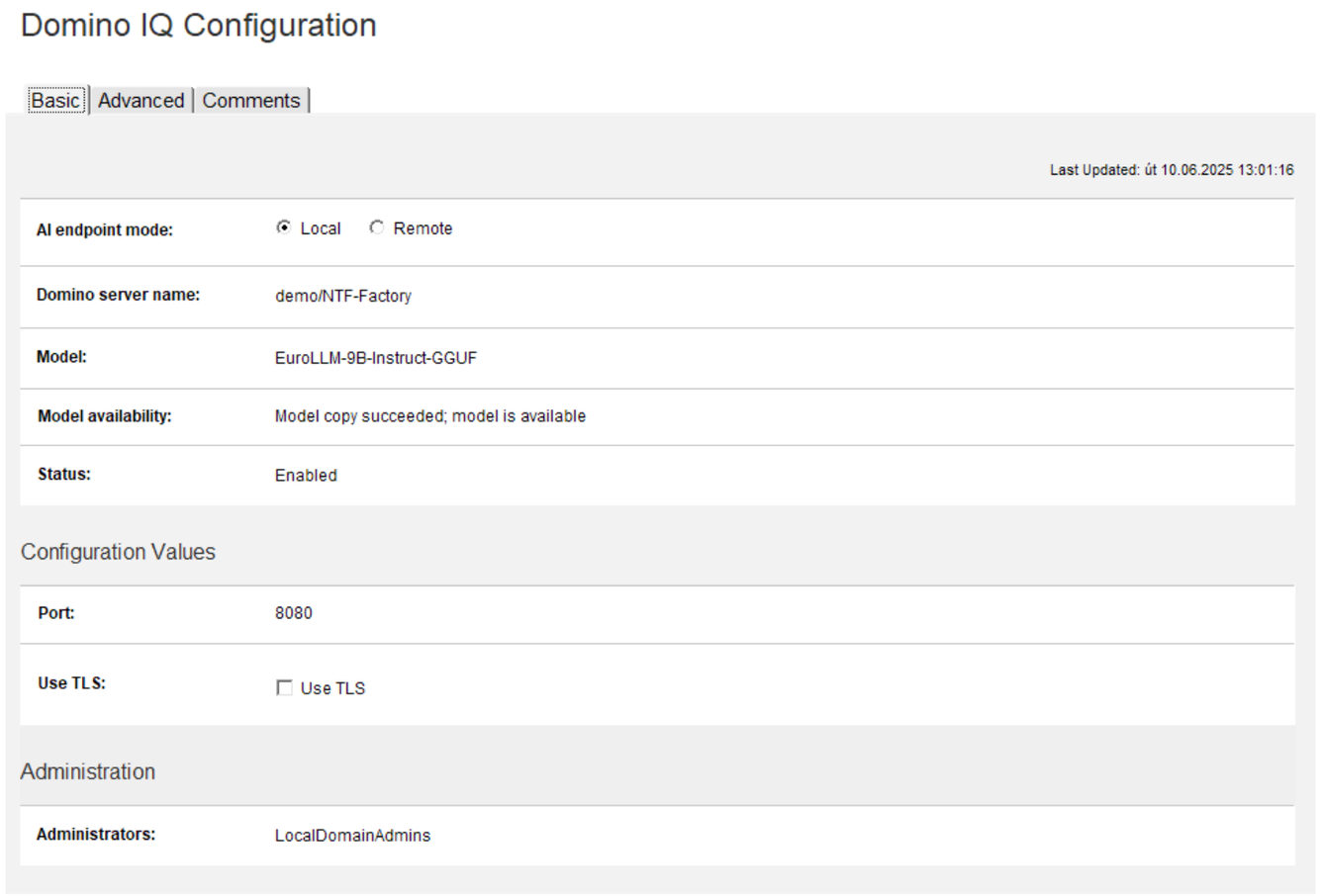
Konfigurace – Model (lokální)

- GGUF file format
- Fyzicky jsou soubory v DominoData/llm_models
- Download funguje
- Stažených (připravených) modelů může být více



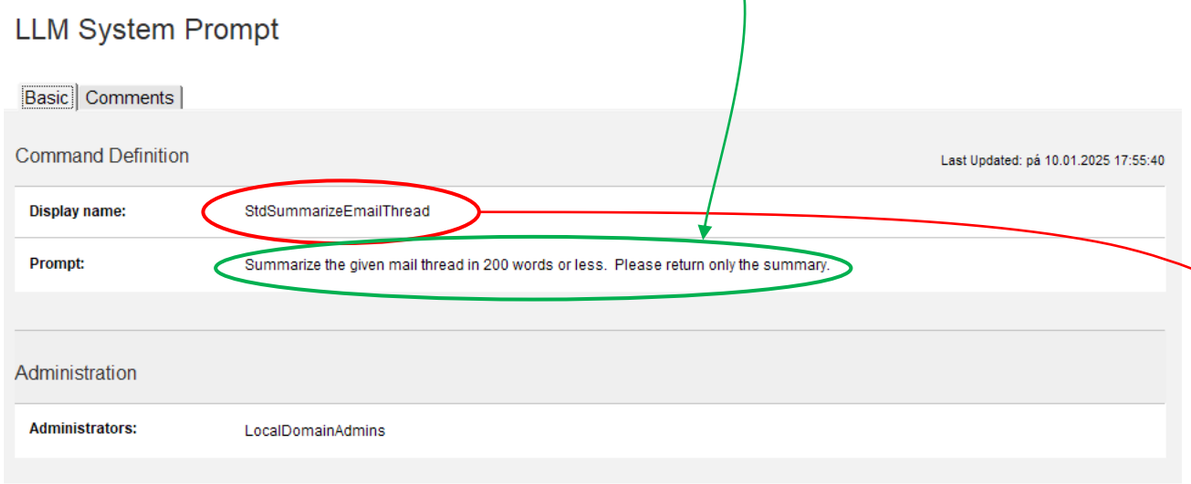
Konfigurace – Configuration

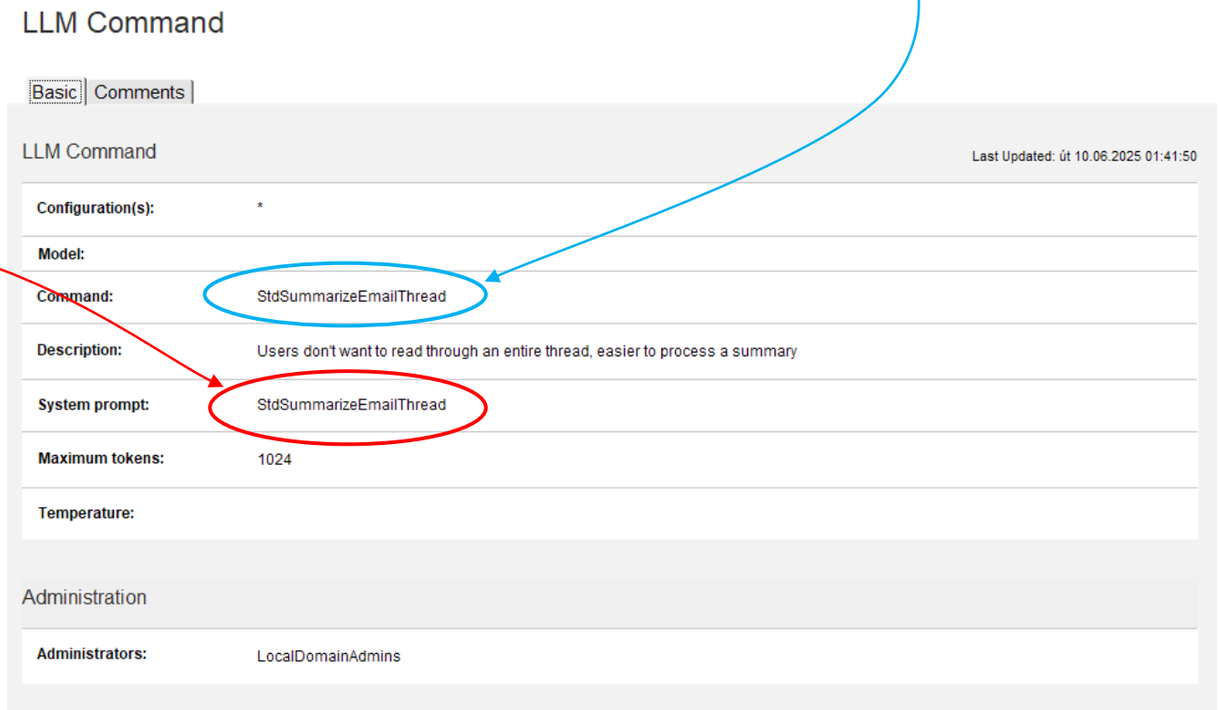
- Server x Model
- Povolený (používaný) jen 1 lokální model na 1 Domino server



Konfigurace – System Prompt & Commands

- System prompt
  - se posílá LLM
- Command
  - volají LS metody a funkce





LotusScript

NotesLLMRequest (LotusScript)

LotusScript class used to send request to Language Model (LLM) via Domino IQ server.

Containment

Contained by

Methods

Completion

CompletionStream

CancelStream

ISCommandAvailable

GetAvailableCommands

NotesLLMResponse (LotusScript)

LotusScript class that represents the response from the NotesLLMRequest.completion method.

Containment

Contained by

Properties

Content

FinishReason

Role

Použití

```
Dim session As New NotesSession
Dim llmreq As NotesLLMRequest
Set llmreq = session.CreateLLMRequest()

Dim sCommand As String
sCommand = "translateEN2CZ"

If llmreq.IsCommandAvailable("", sCommand) Then
          Dim llmres As NotesLLMResponse
          Set llmres = llmreq.Completion("", sCommand, sPrompt)
          If llmres.FinishReason=LLM_FINISH_REASON_LENGTH Or llmres.FinishReason=LLM_FINISH_REASON_STOP Then

                    ' llmres.Content

          Else
                    ' ...
          End If
Else
                    ' ...
End If
```

K čemu Domino IQ využít?

Vhodné

- Generování textů
- Překlady (s vhodným LLM)

- Sumarizace delších textů nebo dokumentů

Nevhodné

- Vyhledávání informací

Závěrem

- AI je tady
- Obrovská výhoda je lokální model
- Je třeba experimentovat – je potřeba hledat:
  - V prvé řadě definovat use-case a podle něj pak najít:
    - LLM
    - prompty
    - HW
- Nezapomenout poučit uživatele…

Zdroje

- What's new in Early Access Drop 3
  https://help.hcl-software.com/domino/14.5.0/admin/wn_145_ea3.html

- HCL Domino 14.5 Documentation - Domino IQ
  https://help.hcl-software.com/domino/14.5.0/admin/domino_iq_server.html

- NVIDIA CUDA Toolkit Release Notes
  https://docs.nvidia.com/cuda/archive/12.6.0/cuda-toolkit-release-notes/index.html

- Panagenda webinars - Domino IQ – What to Expect, First Steps and Use Cases
  https://www.panagenda.com/webinars/domino-iq-what-to-expect-first-steps-and-use-cases/

- AI Act
  https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689