
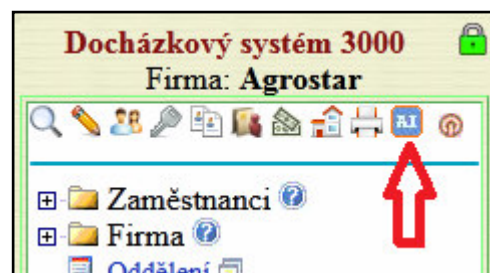


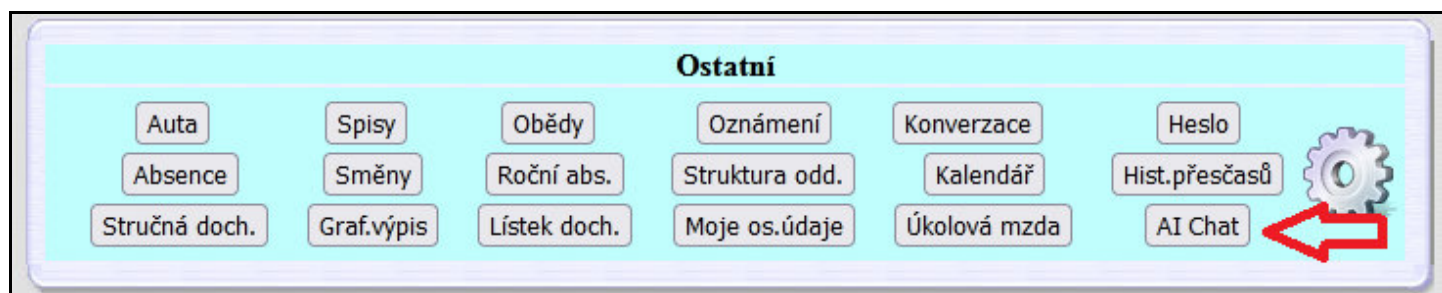
Jednoduché nastavení systému Docházka 3000 pro podporu AI založené na LLM

Docházkový systém Docházka 3000 umožňuje integrovat funkce velkých jazykových modelů (LLM) do svého rozhraní. Takže přímo z docházky můžete komunikovat se systémem umělé inteligence (AI).

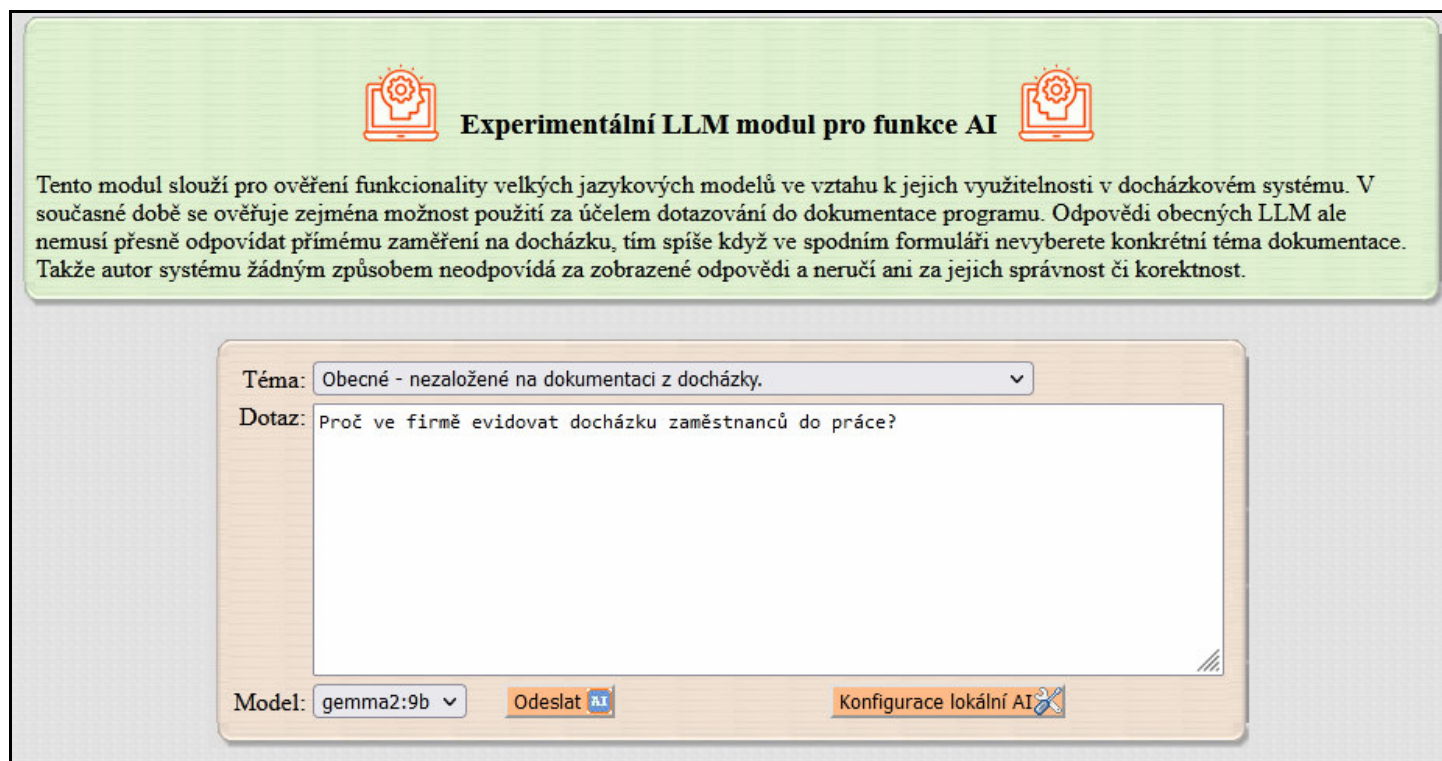
Přístup do AI je v programu dostupný od verze 9.73 a to v admin. menu „Ostatní / Modul AI“ nebo v levé horní liště ikon přes ikonu :



A pokud to podle dále popsaných informací povolíte, tak je dostupný i běžným zaměstnancům přes jejich uživatelské menu „Ostatní / AI Chat“:



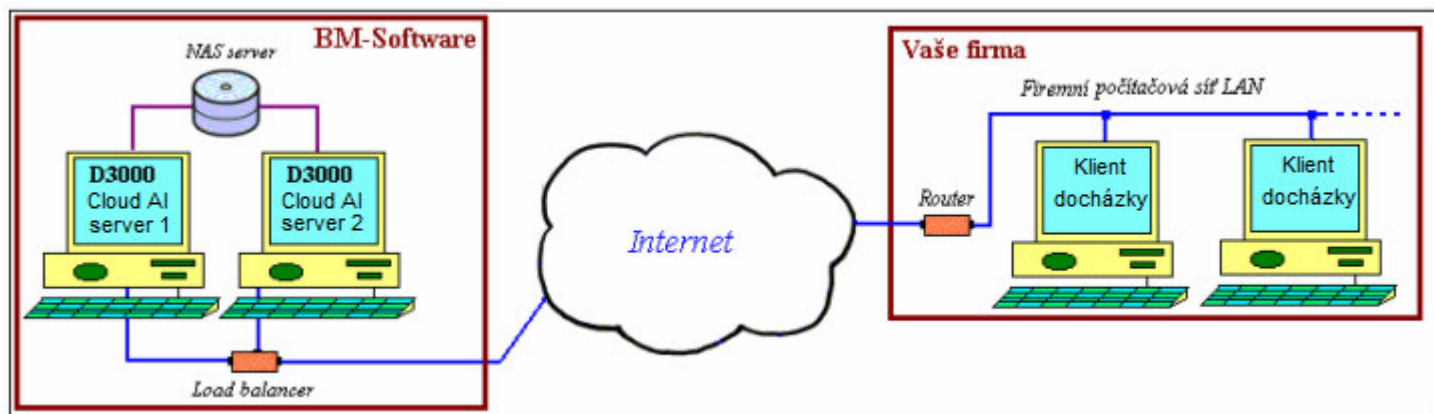
Po vstupu do modulu se zobrazí oranžové okno, do kterého je po provedení konfigurace možné psát jak obecné dotazy, tak pomocí výběrového formuláře *Téma* klást dotazy související s problematikou dle zvolené oblasti.



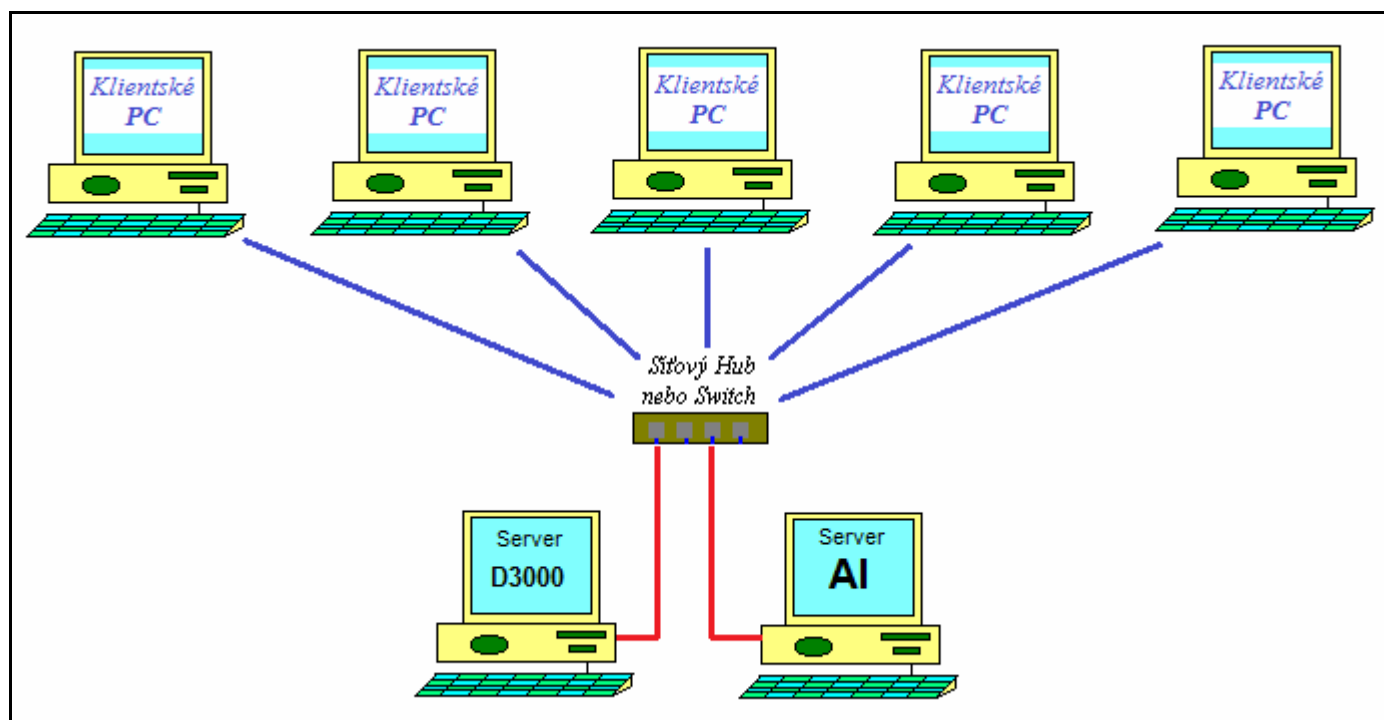
Zkratka: Pro rychlé zprovoznění lokální AI na stejném PC jako běží docházka stačí provést kroky ze strany 4 až 7.

Modul umožňuje dva typy použití:

1) Využít předinstalované AI servery provozované v cloudové infrastruktuře výrobce systému Docházka 3000




2) Instalace vlastního AI serveru. Tato příručka se zabývá zejména právě tímto tématem a provede vás snadnou instalaci vlastní lokální AI infrastruktury během pár minut.



Přičemž výhoda instalace AI přímo na vašem PC či serveru je zejména v tom, že od vás neodchází po síti žádná data a tak vaši zaměstnanci mohou s větší důvěrou přes AI řešit i problematiku, u které není vhodné aby byla data přenášena přes veřejný internet.

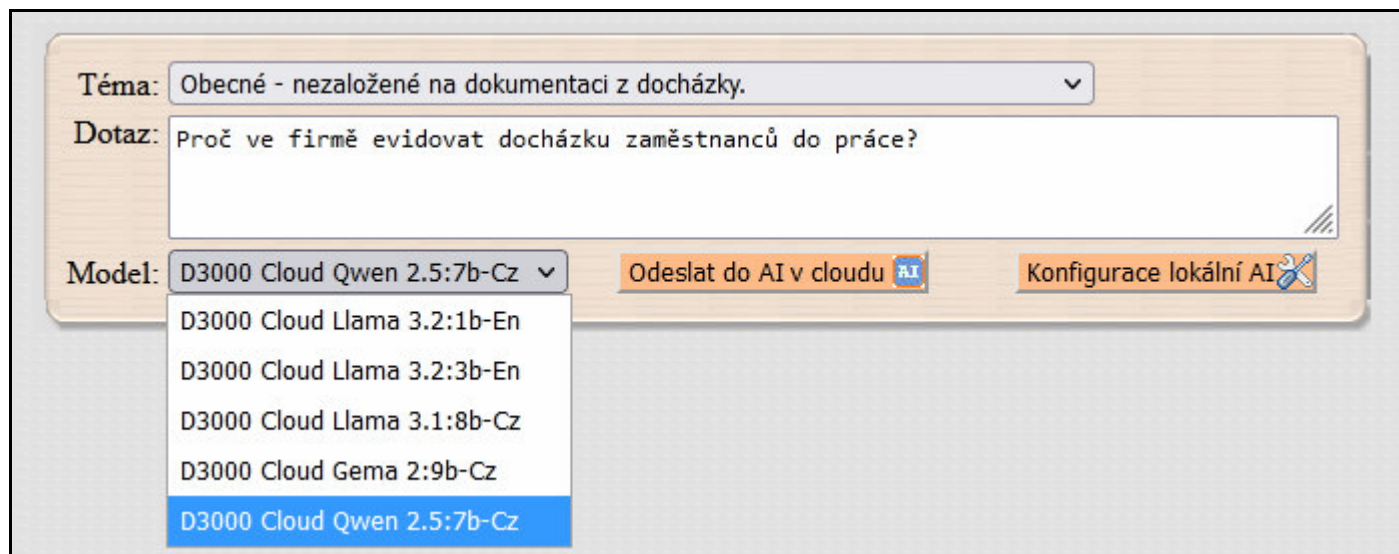
Navíc není nutné, aby při lokální instalaci AI server běžel na samostatném PC, ale klidně se může jednat o stejný počítač na kterém je zároveň nainstalovaná i docházka. Přípustné jsou obě možnosti, tedy jak docházka a AI v jednom společném PC, tak i rozdělení rolí na samostatné počítače, zejména pokud počítáte s větší zátěží. V jednom PC tak můžete dělat kompletní zpracování jak docházky tak AI a tento počítač může být i poměrně obyčejné klasické kancelářské PC či notebook, který bude zároveň plnit roli jak serveru Docházka 3000, tak AI serveru a klienta oběma těmito službám. Pokud vám nevádí pomalejší odezva systému, může se jednat v podstatě o libovolný počítač. V nejskromnější variantě (ale pak i velice pomalé), může mít jen dvoujádrové CPU s 4GB RAM a zcela běžnou grafickou kartou, i když pak jste omezeni jen na triviální LLM modely. U AI obecně platí, že čím rychlejší CPU s co nejvíce jádry, velkým množstvím paměti RAM, SSD či NVME diskem a ideálně výkonnou grafickou kartou opět osazenou co nejvíce pamětí, tím lepších výkonů a rychlosti odezvy AI dosahuje a lze použít i velké LLM modely s profesionální úrovní odpovědí.

Varianta 1 – AI v cloudu výrobce:

V současné době je přístup ke cloudové AI v experimentálním režimu. Služba je prozatím poskytována zdarma v omezeném rozsahu počtu klientů. Dostupnost není garantovaná. Po vstupu do AI v docházce přes ikonu  nebo přes menu *Ostatní* systém nejprve ověří, zda je pro vás cloudová služba v tomto okamžiku dostupná:





Pokud nemáte nakonfigurovanou lokální AI službu a cloudová je dostupná, zobrazí se v oranžovém okně formulář pro zadávání dotazů možnost volit buď obecné téma nebo vybrat téma z dokumentace k SW Docházka 3000 na které se chcete dotázat. Dále můžete vybrat preferovaný jazykový model (LLM) a tlačítkem „Odeslat do AI v cloudu“ dotaz odešlete.



Téma: Obecné - nezaložené na dokumentaci z docházky.

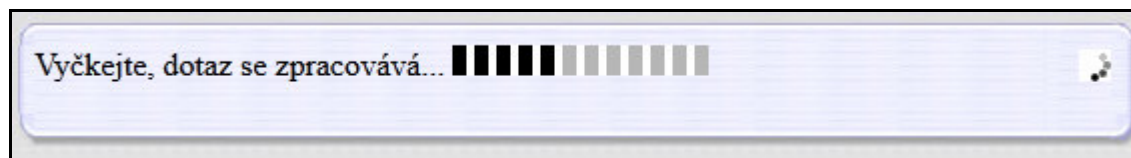
Dotaz: Proč ve firmě evidovat docházku zaměstnanců do práce?

Model: D3000 Cloud Qwen 2.5:7b-Cz

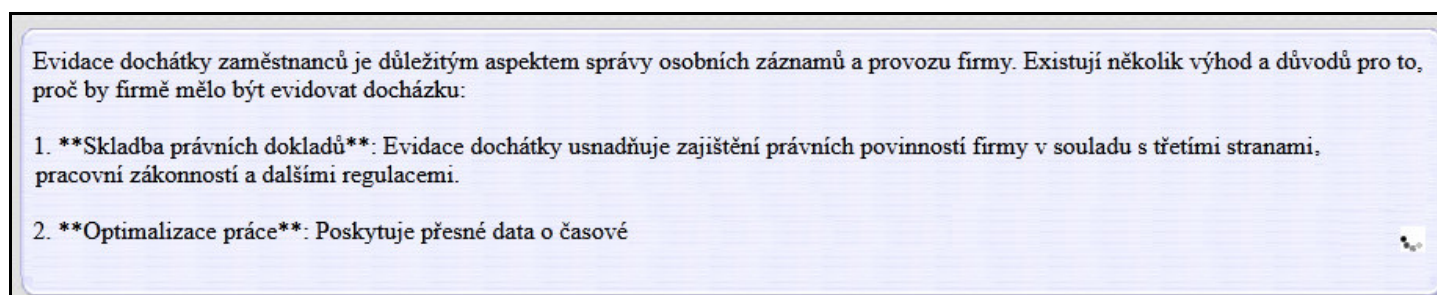
Odeslat do AI v cloudu  Konfigurace lokální AI 

- D3000 Cloud Llama 3.2:1b-En
- D3000 Cloud Llama 3.2:3b-En
- D3000 Cloud Llama 3.1:8b-Cz
- D3000 Cloud Gema 2:9b-Cz
- D3000 Cloud Qwen 2.5:7b-Cz

Systém pak začne váš dotaz zpracovávat, což může zejména při výběru konkrétního tématu (jiného než obecného) trvat delší dobu. Při velké vytíženosti cloudové AI i desítky vteřin či jednotky minut.



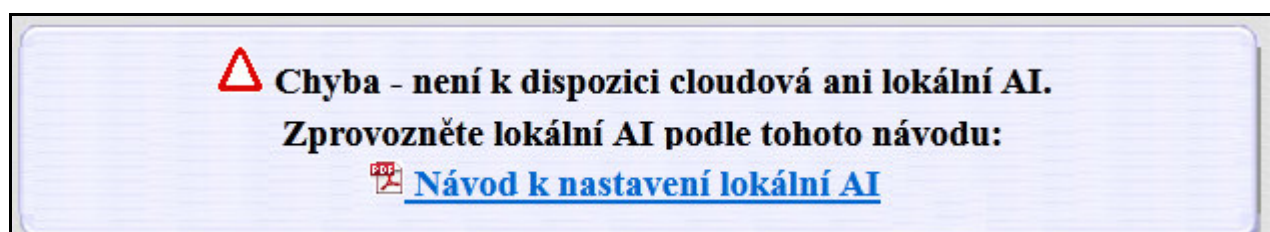
Jakmile se ve frontě požadavků od ostatních uživatelů dostane na váš dotaz, začne AI zobrazovat odpověď:



Evidace docházky zaměstnanců je důležitým aspektem správy osobních záznamů a provozu firmy. Existují několik výhod a důvodů pro to, proč by firmě mělo být evidovat docházku:

- **Skladba právních dokladů****: Evidace docházky usnadňuje zajištění právních povinností firmy v souladu s třetími stranami, pracovní zákoností a dalšími regulacemi.
- **Optimalizace práce****: Poskytuje přesné data o časové

Pokud by pro vás nebyla cloudová AI právě k dispozici a lokální AI ještě nemáte nakonfigurovanou, zobrazí se nahoře nad formulářem tato zpráva:

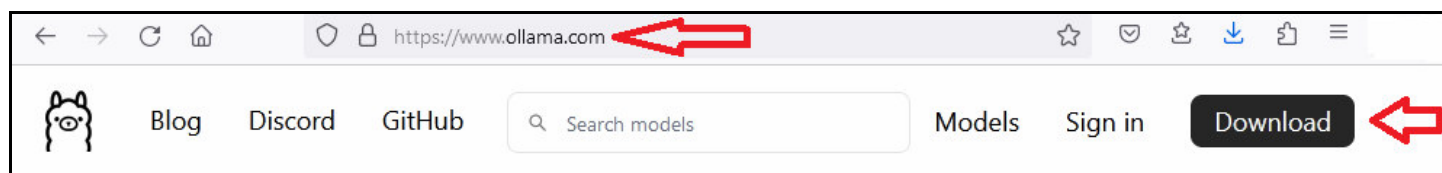


Varianta 2 – lokální AI na vašem PC či serveru:

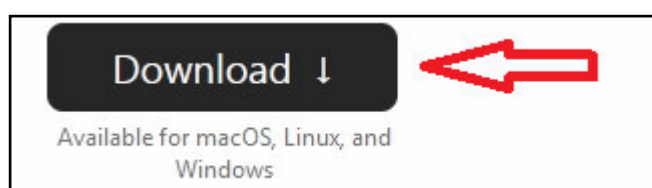
Lokální AI má mnoho výhod a její instalace není vůbec tak složitá jak by se vám mohlo na první pohled zdát. Zvládne jí praktický jakýkoli středně zkušený uživatel počítače. Žádné speciální IT znalosti k tomu nejsou úplně potřeba. Instalaci podle níže uvedených kroků můžete provést buď přímo na hlavním PC docházky, nebo pokud chcete použít jiný počítač či server, protože plánujete AI intenzivně používat, tak úvodním níže uvedené kroky provádějte na tomto PC pro AI. Čím lepší HW bude mít, tím snadněji se vám bude AI používat a operačním systémem tohoto AI počítače pak může být buď Windows, nebo Linux či MacOS.

A) Instalace prostředí.

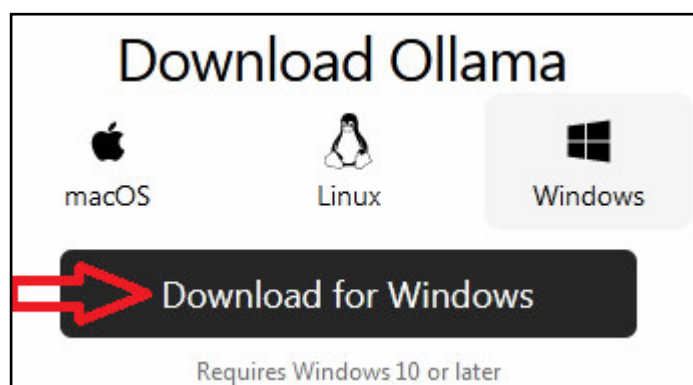
Ve webovém prohlížeči zadejte do adresního řádku adresu www.ollama.com a poté klikněte na *Download*



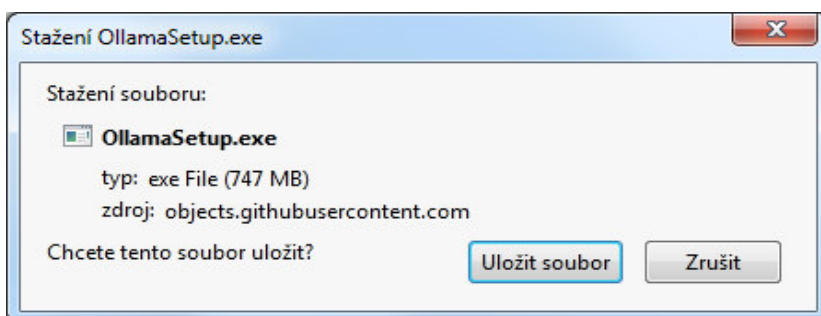
Na další obrazovce opět klikněte na *Download*



Zobrazí se výběr operačního systému. Typ OS můžete přepnout, ale pokud instalujete na Windows (alespoň ve verzi 10 či novější), mělo by se rovnou zobrazit tlačítko *Download for Windows*, takže na něj klikněte. U jiných operačních systémů (Linux, MacOS) se zobrazí příkaz či postup přes který stažení spustíte, ale u Windows jednoduše klikněte na *Download for Windows* a program se začne stahovat.



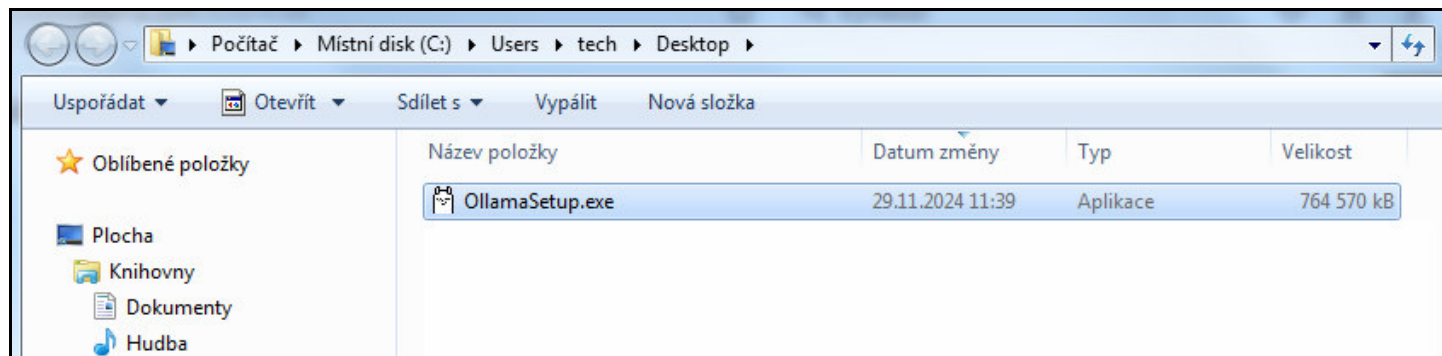
Podle nastavení webového prohlížeče se vás počítač buď zeptá kam chcete soubor stáhnout (jako na obrázku vpravo), takže klikněte na tlačítko *Uložit soubor*, nebo se soubor *OllamaSetup.exe* rovnou uloží do složky *Stažené soubory*. Stahování může chvíli trvat a jeho průběh by vám měla ukazovat ikona vpravo nahoře ve vašem webovém prohlížeči.



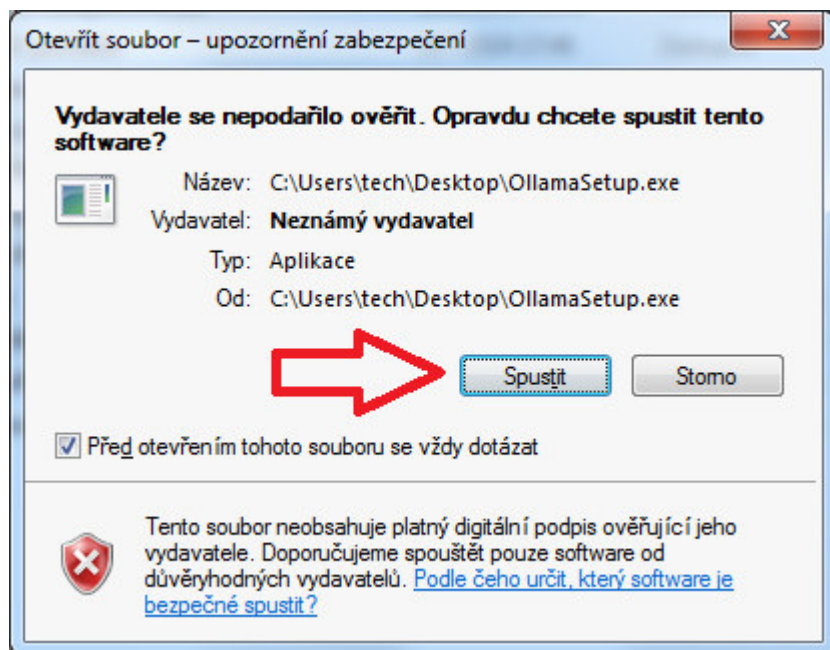
Po dokončení stahování klikněte v prohlížeči vpravo nahoře na ikonu stažených souborů a zobrazí se informace o staženém souboru. Přes ikonu složky vpravo od *OllamaSetup.exe* otevřete adresář se staženým souborem



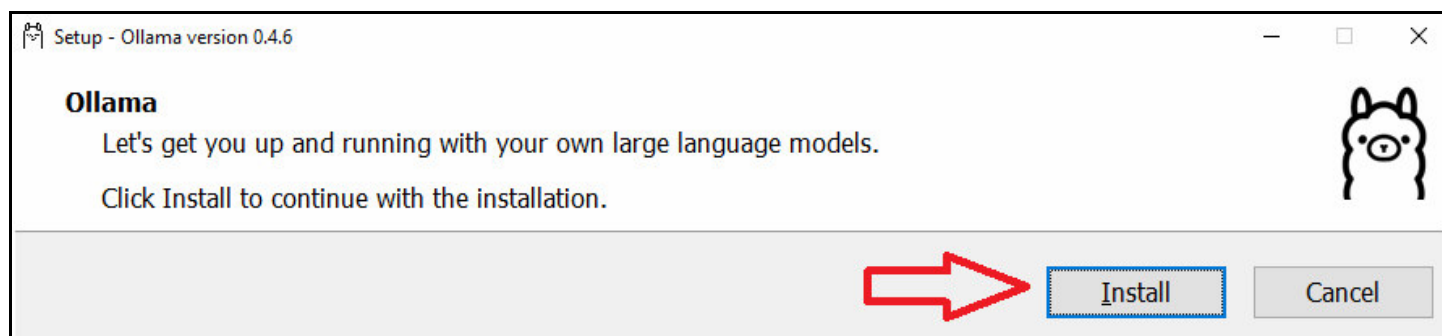
Zde v průzkumníkoví souborů Windows vidíte stažený soubor *OllamaSetup*:



Dvojklikem levého tlačítka myši na tomto souboru spustíte instalaci (odsouhlasíte případné hlášení OS):



Jakmile se instalační program spustí, kliknete na tlačítko *Install* a počkáte až doběhne instalace do konce.



B) Stažení modelů:

Jakmile instalace doběhne, tak je ještě potřeba stáhnout nějaké LLM modely, které AI používá pro svou práci. Modelů jsou k dispozici desítky. Jejich přehled je na webu *Ollama.com* pod záložkou *Models*.



Ve Windows dole na liště zadejte do vyhledávacího okna *cmd* a zmáčkněte klávesu *Enter*. Spustí se černé okno příkazového řádku, ale vůbec se nelekejte, bude to snadné (jediný příkaz, 3 slova). Zadejte toto:

```
ollama pull llama3.2:1b
```

```
C:\> Administrator: Command Prompt
Microsoft Windows [Version 10.0.20348.587]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Administrator>ollama pull llama3.2:1b
```

Začne se stahovat LLM model *llama3.2* ve verzi *1b*, který by měl fungovat v podstatě na jakémkoliv počítači pokud má alespoň 4GB paměti RAM. Průběh stahování se postupně zobrazuje (nabíhají procenta atd.)

```
C:\> Administrator: Command Prompt - ollama pull llama3.2:1b
Microsoft Windows [Version 10.0.20348.587]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Administrator>ollama pull llama3.2:1b
pulling manifest
pulling 74701a8c35f6... 24% ██████████ 320 MB/1.3 GB 9.1 MB/s 1m50s
```

Vpravo vidíte odhad zbývajících času, tedy za jak dlouho by se mělo zhruba stahování dokončit (na obrázku výše je to méně než 2 minuty). Jakmile stahování doběhne, bude dole uvedeno *success* (úspěšně staženo):

```
C:\> Administrator: Command Prompt
Microsoft Windows [Version 10.0.20348.587]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Administrator>ollama pull llama3.2:1b
pulling manifest
pulling 74701a8c35f6... 100% ██████████ 1.3 GB
pulling 966de95ca8a6... 100% ██████████ 1.4 KB
pulling fcc5a6bec9da... 100% ██████████ 7.7 KB
pulling a70ff7e570d9... 100% ██████████ 6.0 KB
pulling 4f659a1e86d7... 100% ██████████ 485 B
verifying sha256 digest
writing manifest
success
C:\Users\Administrator>
```

Pokud by se vypsal něco jiného než na výše uvedeném obrázku, zkontrolujte zda má PC funkční připojení do internetu, dostatek místa na disku, zda jako uživatel máte potřebná práva atd.

Nyní můžete buď příkazový řádek ukončit příkazem *exit*, nebo doporučíme stáhnout ještě několik dalších modelů. Přeci jen ten právě stažený má velmi omezené schopnosti a komunikovat česky moc neumí, vyhovuje mu hlavně angličtina. Jeho výhodou je, že je rychlý a zabírá jen minimum systémových prostředků, takže na testování se hodí poměrně dobře a určitě jej mějte stažený. Mírně lepší je tento model ve verzi 3b (tomu stačí méně jak 6GB RAM) a ještě lepší ve verzi 8b, který ale vyžaduje minimálně 8GB RAM zato již komunikuje obstojně česky. Do 8GB RAM se vleze ještě mírně lepší model *qwen 2.5* ve verzi *7b* a pokud máte více jak 16GB RAM, můžete model *qwen 2.5* vyzkoušet i ve verzi *14b*. Příkazy ke stažení těchto dalších 4 modelů jsou:

```
ollama pull llama3.2:3b
ollama pull llama3.1:8b
ollama pull qwen2.5:7b
ollama pull qwen2.5:14b
```

Máte-li na disku zhruba 20GB místa, stáhněte je všechny. Tedy snad až na ten poslední, ten jen pokud má vaše PC rychlé CPU, přes 16GB RAM a výkonnou grafickou kartu se stejným množstvím video-paměti. Nakonec přes příkaz *exit* příkazový řádek ukončíte.


C) Konfigurace docházky:

Pokud jste systém *Ollama* instalovali na stejné PC na jakém běží docházka a máte docházkový systém alespoň ve verzi 9.73, což zjistíte na úvodní obrazovce před přihlášením, kde je číslo verze uvedeno zeleně:

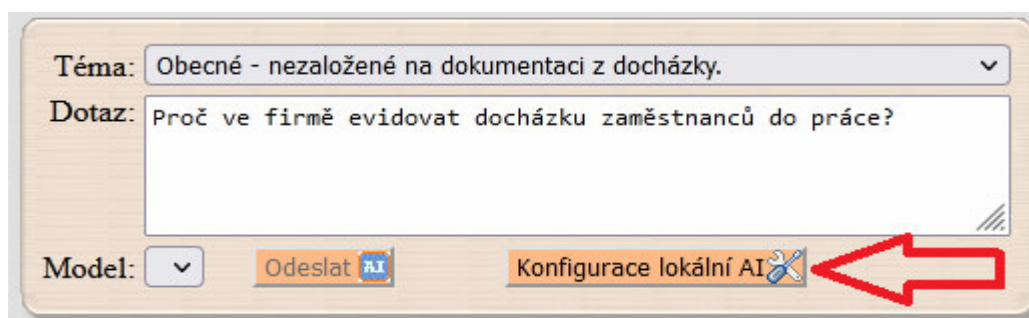
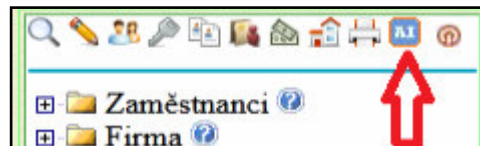


tak nyní pokračujte postupem níže. V opačném případě objednáte aktualizaci na této webové adrese:

<https://www.dochazka.eu/dochazka3000/shop/index.php?naid=45>

Přihlásíte se do docházky jako administrátor a v levém horním menu kliknete na ikonu .

Zobrazí se AI modul docházky a v něm kliknete na ikonu *Konfigurace lokální AI*




Zobrazí se okno pro snadné napojení lokální nainstalované AI systému Ollama do docházkového systému:

Formulář pro zprovoznění lokálního LLM modulu *Ollama* pro dotazy na AI:

IP adresa AI serveru:	<input type="text"/>
AI model č.1:	<input type="text"/>
AI model č.2:	<input type="text"/>
AI model č.3:	<input type="text"/>
AI model č.4:	<input type="text"/>
AI model č.5:	<input type="text"/>
Povolit AI zaměstnancům:	<input type="checkbox"/>
<input type="button" value="Uložit nastavení"/>	

Pokud jste již provedli lokální instalaci systému Ollama dle níže uvedené PDF příručky a instalovali jste tuto AI na stejné PC na kterém běží docházka, můžete pomocí tlačítka níže naplnit výchozí hodnoty konfigurace pro provoz AI běžící na stejném PC na kterém je nainstalovaná i docházka:

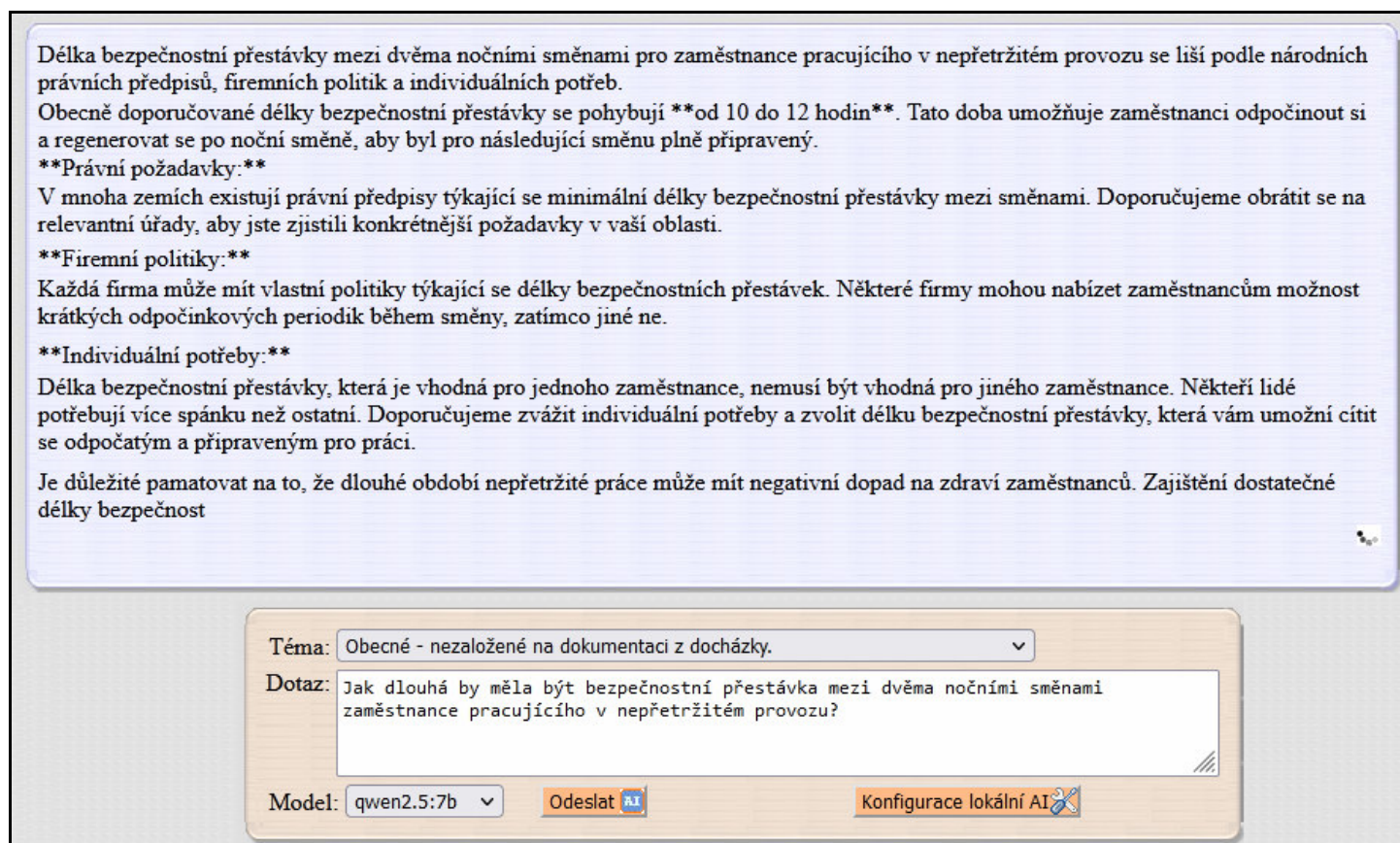
 [Návod k nastavení lokální AI](#)

Stačí kliknout na tlačítko *Naplní výchozí hodnoty* vpravo dole (které je dostupné když je formulář prázdný)

Nastavení uloženo. Lokální AI modul bude **POVOLEN**.

IP adresa AI serveru:	<input type="text" value="127.0.0.1"/>
AI model č.1:	<input type="text" value="llama3.2:1b"/>
AI model č.2:	<input type="text" value="llama3.2:3b"/>
AI model č.3:	<input type="text" value="llama3.1:8b"/>
AI model č.4:	<input type="text" value="qwen2.5:14b"/>
AI model č.5:	<input type="text" value="qwen2.5:7b"/>
Povolit AI zaměstnancům:	<input type="checkbox"/>
<input type="button" value="Uložit nastavení"/>	

Program provede výchozí konfiguraci propojení AI nainstalované na stejném PC jako je PC s docházkou. Tlačítkem *Zpět do modulu AI* se vrátíte do dialogu pro zadávání dotazů a ten by již měl fungovat. Stačí zadat dotaz, vybrat některý ze stažených modelů a po odeslání dotazu by měl systém během několika vteřin (na pomalém PC i minut) začít psát odpověď. Pro češtinu doporučujeme v přepínacím menu *Model* vybrat ze stažených buď *llama3.1:8b* nebo model *qwen2.5:7b*. Poměrně rychle na obyčejném PC funguje prvně stažený model *llama3.2:1b* a o moc pomalejší není ani o něco lepší model *3b*, oba ale komunikují spíše v angličtině, resp. anglicky psaným dotazům rozumí dobře.



Délka bezpečnostní přestávky mezi dvěma nočními směnami pro zaměstnance pracujícího v nepřetržitém provozu se liší podle národních právních předpisů, firemních politik a individuálních potřeb.

Obecně doporučované délky bezpečnostní přestávky se pohybují ****od 10 do 12 hodin****. Tato doba umožňuje zaměstnanci odpočinout si a regenerovat se po noční směně, aby byl pro následující směnu plně připravený.

****Právní požadavky:****

V mnoha zemích existují právní předpisy týkající se minimální délky bezpečnostní přestávky mezi směnami. Doporučujeme obrátit se na relevantní úřady, aby jste zjistili konkrétnější požadavky v vaší oblasti.

****Firemní politiky:****

Každá firma může mít vlastní politiky týkající se délky bezpečnostních přestávek. Některé firmy mohou nabízet zaměstnancům možnost krátkých odpočinkových periodik během směny, zatímco jiné ne.

****Individuální potřeby:****

Délka bezpečnostní přestávky, která je vhodná pro jednoho zaměstnance, nemusí být vhodná pro jiného zaměstnance. Někteří lidé potřebují více spánku než ostatní. Doporučujeme zvážit individuální potřeby a zvolit délku bezpečnostní přestávky, která vám umožní cítit se odpočatým a připraveným pro práci.

Je důležité pamatovat na to, že dlouhé období nepřetržité práce může mít negativní dopad na zdraví zaměstnanců. Zajištění dostatečné délky bezpečnost

Téma: Obecné - nezaložené na dokumentaci z docházky.

Dotaz: Jak dlouhá by měla být bezpečnostní přestávka mezi dvěma nočními směnami zaměstnance pracujícího v nepřetržitém provozu?

Model: qwen2.5:7b

Odeslat AI

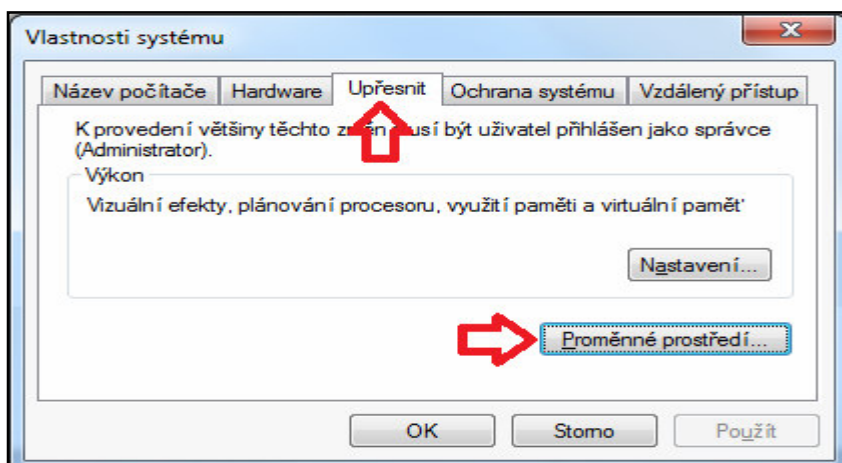
Konfigurace lokální AI

Doplňek pro instalaci na jiném PC než je docházka: (při instalaci na PC docházky není níže uvedené už potřeba)

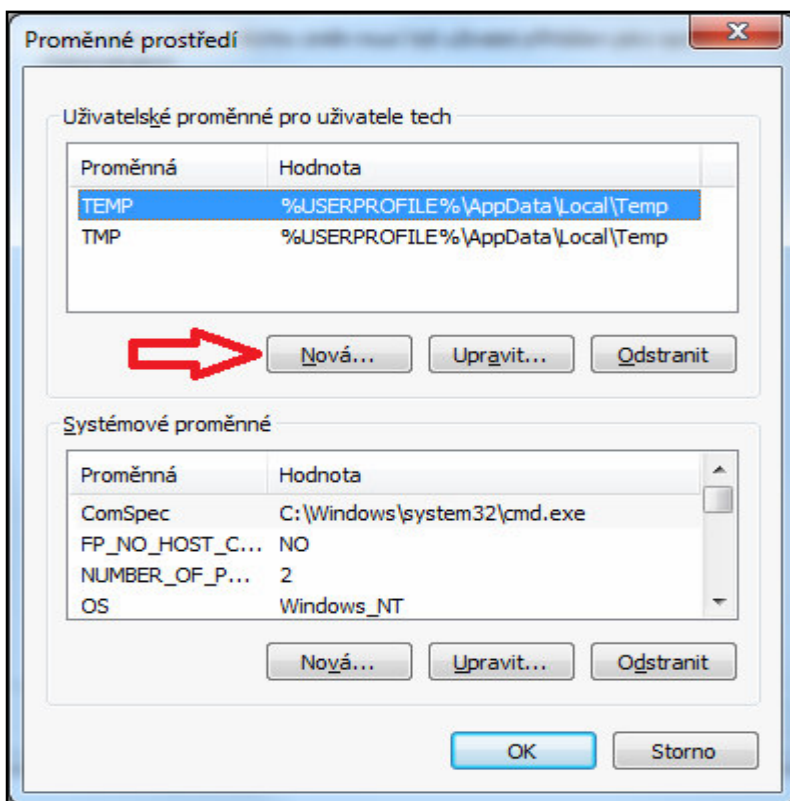
Pokud jste systém *Ollama* instalovali na jiné PC než na kterém běží docházka, je potřeba v konfiguraci modulu AI v docházce do položky *IP adresa AI serveru* uvést správnou síťovou IP adresu pro počítač se systémem *Ollama*. Další důležitá věc je na tomto počítači s nainstalovaným systémem *Ollama* nastavit v operačním systému systémovou proměnnou prostředí *OLLAMA_HOST* na hodnotu *0.0.0.0* (pro přístup po síti LAN přes kterýkoli síťový adaptér) a počítač restartovat nebo restartovat službu *Ollama*. Na windows postupujte tak, že dole v liště windows do vyhledávacího okna zadáte příkaz *sysdm.cpl* a potvrdíte Enter:



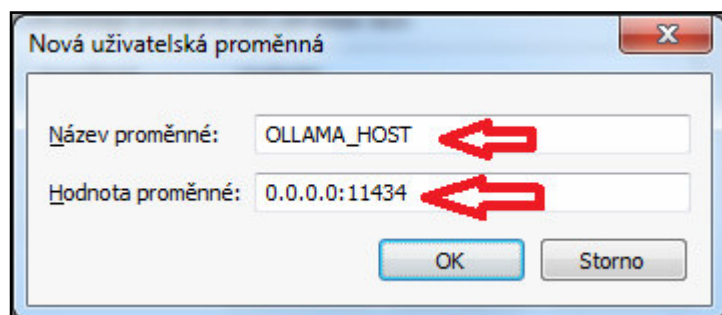
V okně vlastností systému se přepnete na záložku *Upřesnit* a dáte *Proměnné prostředí*:



V uživatelských proměnných kliknete na *Nová*

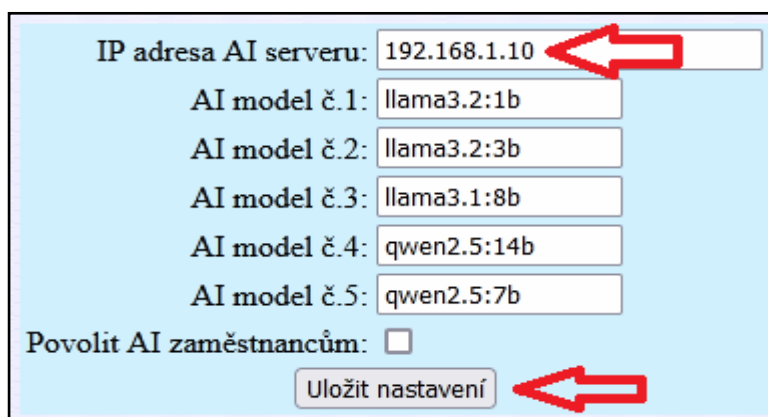


Název nové proměnné bude *OLLAMA_HOST* a hodnota *0.0.0.0*

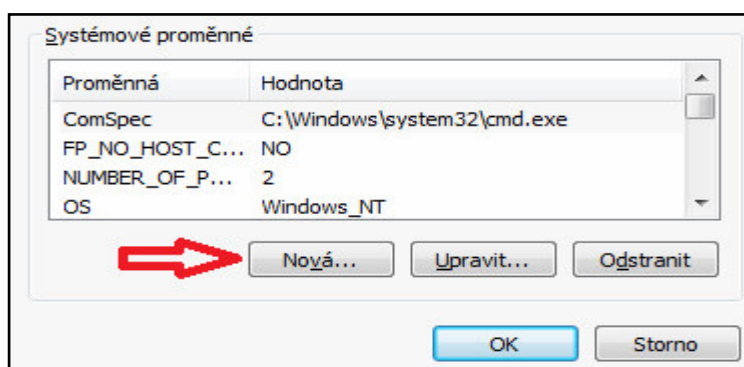


Nakonec restartujete celé PC se službu *Ollama*. Nezapomeňte v nastavení modulu AI v docházce nastavit položky *IP adresa AI serveru* uvést správnou síťovou IP adresu pro počítač se systémem *Ollama*. Příklad pokud *Ollama* běží na PC s IP adresou *192.168.1.10*:

Dále nezapomeňte port 11434 odblokovat v nastavení firewallu (Brána firewall systému windows v ovládacím panelu).



Pokud by spojení stále nefungovalo, tak stejnou proměnnou prostředí doplňte i do části pro systémové proměnné. Dále samozřejmě nezapomeňte port 11434 odblokovat v nastavení firewallu (Brána firewall systému windows v ovládacím panelu), případně antivirovém nebo jiném zabezpečovacím programu, který může spojení blokovat.



Závěrem:


Tento AI modul z pohledu uplatnění v docházce zatím slouží spíše jen pro ověření funkcionality velkých jazykových modelů ve vztahu k jejich využitelnosti v docházkovém systému. V současné době se ověřuje zejména možnost použití za účelem dotazování do dokumentace programu. Odpovědi obecných LLM ale nemusí přesně odpovídat přímému zaměření na docházku, tím spíše když ve formuláři dotazu nevyberete konkrétní téma dokumentace. Takže autor systému žádným způsobem neodpovídá za zobrazené odpovědi a neručí ani za jejich správnost či korektnost.

Přístup k modulu AI pro všechny zaměstnance

Pokud přesto chcete AI modul zpřístupnit i běžným zaměstnancům přes jejich uživatelské menu programu, tak v menu „*Ostatní / Modul AI / Konfigurace lokální AI*“ zatrhněte položku „*Povolit AI zaměstnancům*“ a nastavení uložte.

Formulář pro zprovoznění lokálního LLM modulu *Ollama* pro dotazy na AI:

IP adresa AI serveru:	<input type="text" value="127.0.0.1"/>
AI model č.1:	<input type="text" value="llama3.2:1b"/>
AI model č.2:	<input type="text" value="llama3.2:3b"/>
AI model č.3:	<input type="text" value="llama3.1:8b"/>
AI model č.4:	<input type="text" value="qwen2.5:14b"/>
AI model č.5:	<input type="text" value="qwen2.5:7b"/>
Povolit AI zaměstnancům:	<input checked="" type="checkbox"/>
<input type="button" value="Uložit nastavení"/>	

 [Návod k nastavení lokální AI](#)

Pracovníci se pak k systému dostanou ze svého uživatelského menu z modré části *Ostatní / AI Chat*:

Ostatní

<input type="button" value="Auta"/>	<input type="button" value="Spisy"/>	<input type="button" value="Obědy"/>	<input type="button" value="Oznámení"/>	<input type="button" value="Konverzace"/>	<input type="button" value="Heslo"/>
<input type="button" value="Absence"/>	<input type="button" value="Směny"/>	<input type="button" value="Roční abs."/>	<input type="button" value="Struktura odd."/>	<input type="button" value="Kalendář"/>	<input type="button" value="Hist.přesčasů"/>
<input type="button" value="Stručná doch."/>	<input type="button" value="Graf.výpis"/>	<input type="button" value="Lístek doch."/>	<input type="button" value="Moje os.údaje"/>	<input type="button" value="Úkolová mzda"/>	<input type="button" value="AI Chat"/>



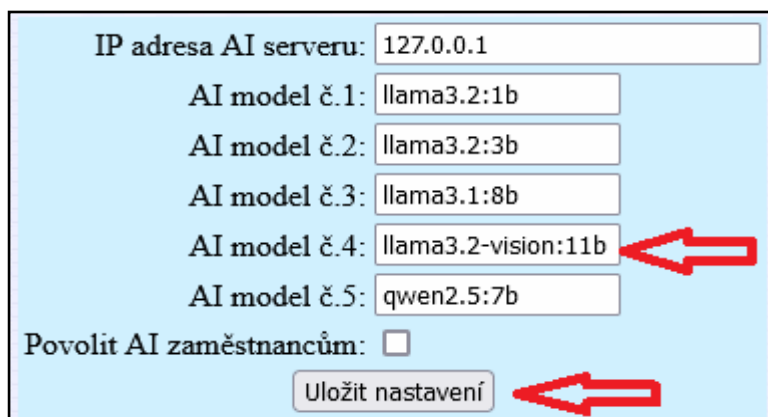
Instalace dalších LLM modelů:

Jak bylo uvedeno na předchozích stranách, na webu *Ollama.com* je v záložce *Models* přístupná celá řada dalších modelů. Stačí si vybrat který chcete kliknutím na název v seznamu, rozkliknout vlevo nabídku *velikostí* a vybrat požadovanou variantu. Vpravo se pak zobrazí příkaz přes který model stáhne v příkazovém řádku:

11b	▼	<input type="text" value="ollama run llama3.2-vision:11b"/>	<input type="button" value="📄"/>
11b	7.9GB		
90b	55GB		

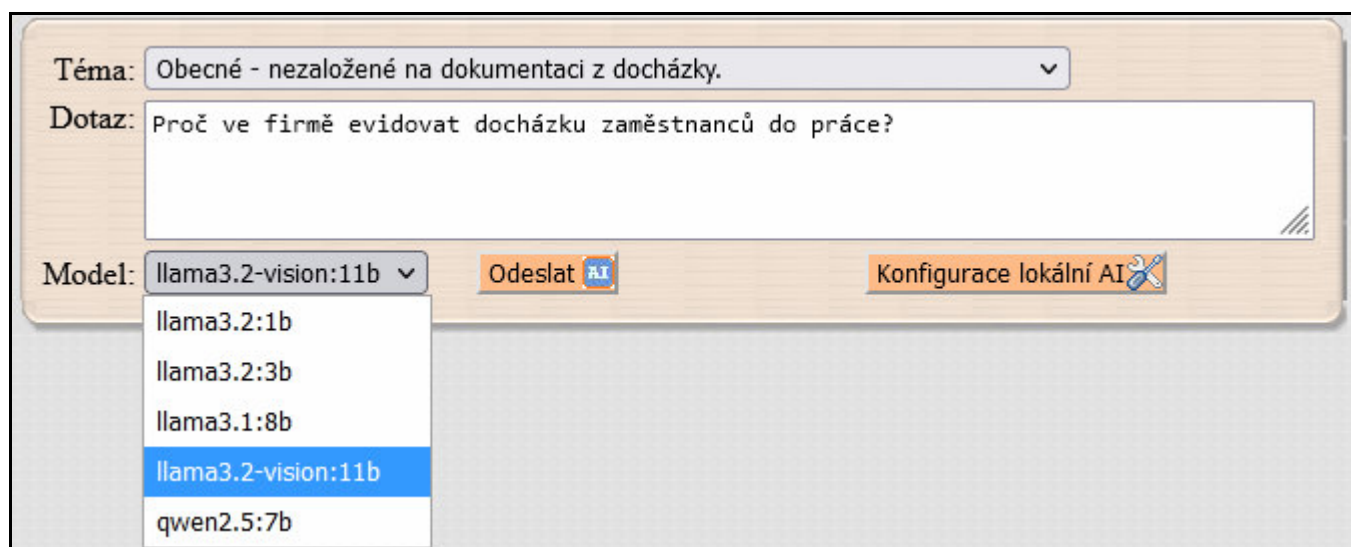
To, zda vaše PC bude schopno s modelem pracovat a splňuje požadavky na paměť RAM se dá velice zjednodušeně odvodit tak, že u malých modelů zobrazenou velikost vynásobíte zhruba dvěma a tím dostanete paměťové nároky konkrétního modelu. Takže z předchozího obrázku lze zhruba usoudit, že model 11b zabírající 8GB by potřeboval PC s 16GB RAM.

Po stažení modelu nezapomeňte přejít do docházky do menu „Ostatní / Modul AI / Konfigurace lokální AI“ a do jedné z pěti dostupných kolonek pro název modelu tento nový model zapsat, aby byl dostupný v docházce (pokud to chcete). Například po stažení modelu *llama3.2-vision:11b* z předchozího obrázku bude nastavení v docházce vypadat takto:
A nastavení nezapomeňte uložit.



The screenshot shows a configuration window with a light blue background. At the top, there is a text input field for 'IP adresa AI serveru:' containing '127.0.0.1'. Below it are five rows, each with a label 'AI model č. 1:' through 'AI model č. 5:' and a corresponding text input field. The fields contain the following values: 'llama3.2:1b', 'llama3.2:3b', 'llama3.1:8b', 'llama3.2-vision:11b', and 'qwen2.5:7b'. A red arrow points to the 'llama3.2-vision:11b' field. Below the list is a checkbox labeled 'Povolit AI zaměstnancům:' which is currently unchecked. At the bottom right, there is a button labeled 'Uložit nastavení' with a red arrow pointing to it.

V rámu s pokládáním dotazů pak bude nový model na nastavené pozici dostupný a můžete jej začít využívat:



The screenshot shows a chat interface with a light orange header. At the top left, there is a dropdown menu for 'Téma:' with the value 'Obecné - nezaložené na dokumentaci z docházky.'. Below it is a text input field for 'Dotaz:' containing the question 'Proč ve firmě evidovat docházku zaměstnanců do práce?'. Below the input field is a 'Model:' dropdown menu with a list of models: 'llama3.2:1b', 'llama3.2:3b', 'llama3.1:8b', 'llama3.2-vision:11b' (highlighted in blue), and 'qwen2.5:7b'. To the right of the model dropdown are two buttons: 'Odeslat AI' and 'Konfigurace lokální AI'.

Pokud má váš počítač výkonnou grafickou kartu s dostatkem RAM a je podporovaná systémem Ollama (viz přehled zde: <https://github.com/ollama/ollama/blob/main/docs/gpu.md>), bude jí systém automaticky využívat a práce s AI bude mnohonásobně rychlejší než pokud takovou grafickou kartou nedisponujete a vše musí běžet na CPU a klasické paměti RAM.

Pokud se některý model nespustí (ani přes příkazový řádek příkazem *ollama run* doplněným o název modelu), tak jste jej buď nestáhli, nebo jste špatně zadali jeho název a nebo se model do paměti RAM vašeho počítače jednoduše nevejde a museli byste paměť rozšířit. Ollama běží pod uživatelem, který provedl instalaci. Aby tedy vše fungovalo, musí být tento uživatel přihlášený, nebo je potřeba systém Ollama nainstalovat jak službu OS.

Malé modely o velikostech v jednotkách GB nejlépe komunikují v angličtině. V ní i 1GB model *ollama3.2:1b* poskytuje poměrně dobré odpovědi na obecné dotazy. Pro seriózní práci je ale potřeba modely s daty ve vyšších desítkách až stovkách GB. Máte-li opravdu výkonný server s grafickou kartou se 48GB RAM (jako je Nvidia Quadro RTX A6000 a podobně), pak vyzkoušejte model *llama3.3*, který je v době psaní této příručky asi nejvýkonnějším volně dostupným LLM. Výborně si poradí i s češtinou a řešením logických a matematických problémů.

Co se řebříčku podpory češtiny v jednotlivých modelech týče, naleznete přínosnou srovnávací tabulku zde: <https://huggingface.co/spaces/CZLC/BenCzechMark>

Navíc na webu *huggingface.co* máte po registraci přístup ke stovkám dalších modelů, které lze v *GGUF* formátu použít přes SW *Ollama* i v docházce - příkaz pro stažení je *ollama pull hf.co/{uživatel}/{repozitář}*. Např. *ollama pull hf.co/mradermacher/czech-gemma-2-9b-it-i1-GGUF* a v docházce pak v konfiguraci zadáte celý název modelu: *hf.co/mradermacher/czech-gemma-2-9b-it-i1-GGUF* (lepší čeština ale vyžaduje 12GB RAM)

Pokud tedy máte počítač s alespoň 16GB RAM a dostatečně rychlým CPU nebo ještě lépe s grafickou kartou disponující alespoň 16GB video paměti, doporučujeme buď model *qwen2.5:14b* nebo *Gemma Czech* stáhnout a používat. *Gemma Czech* byly natrénován na českých textech a komunikuje výhradně v češtině. Lépe než ostatní podobně velké (či přesněji malé) modely. Dalším modelem s dobrou podporou češtiny je *Qwen2.5*

Model *Gemma Czech* lze v příkazovém řádku Windows stáhnout tímto příkazem

```
ollama pull hf.co/mradermacher/czech-gemma-2-9b-it-i1-GGUF
```

a poté v Docházce 3000 v menu „*Ostatní / Modul AI / Konfigurace lokální AI*“ zadáte do AI modelu č. 1 takto:

```
hf.co/mradermacher/czech-gemma-2-9b-it-i1-GGUF
```

Formulář pro zprovoznění lokálního LLM modulu *Ollama* pro dotazy na AI:

IP adresa AI serveru: 127.0.0.1

AI model č.1: hf.co/mradermacher

AI model č.2: llama3.2:3b

AI model č.3: llama3.1:8b

AI model č.4: qwen2.5:14b

AI model č.5: qwen2.5:7b

Povolit AI zaměstnancům:

Uložit nastavení

Zpět do modulu AI

[Návod k nastavení lokální AI](#)

Modul se pak v menu „*Ostatní / Modul AI*“ bude nabízet jako první a má z uvedených nejlepší podporu češtiny

Téma: Obecné - nezaložené na dokumentaci z docházky

Dotaz: Proč ve firmě evidovat docházku zaměstnanců do práce?

Model: hf.co/mradermacher/czech-gemma-2-9b-it-i1-GGUF

Odeslat AI

Konfigurace lokální AI

Tento model například jako jediný z výše uvedených odpověděl správně česky na česky položenou otázku „*Kolik je třicet dva plus šest*“. Všechny ostatní uvedené modely odpověděly chybně. I když na stejnou otázku položenou v angličtině odpověděly všechny ostatní modely správně. Model *Gemma* odpověděl na tuto anglicky položenou rovněž správně, navíc ale odpověď zobrazil česky. Takže si vše i správně přeložil do češtiny.

```
ollama run hf.co/mradermacher/czech-gemma-2-9b-it-i1-GGUF:latest "What is thirty two plus six?"  
32 plus 6 je 38.
```

Pokud tedy chcete modul AI v docházce zpřístupnit i běžným zaměstnancům k experimentování a máte dostatečně výkonný počítač s 16 GB RAM, doporučujeme z menších modelů právě *Czech Gemma 2 9b*.

Do 16GB RAM se vejde i model *qwen2.5:14b* který lze stáhnout v příkazovém řádku příkazem:

```
ollama pull qwen2.5:14b a poté jej opět vložíte do „Ostatní / Modul AI / Konfigurace lokální AI“
```

Rychlost na CPU:

Pro základní představu o rychlosti lokálně provozovaných LLM modulů na běžném počítači nevybaveném prémiovou grafickou kartou (tedy při běhu na CPU), což je případ naprosté většiny instalací docházky ve firmách a podnicích, uvádíme níže tabulku z provedeného testovacího měření.

Testováno bylo na běžném levném PC s procesorem Intel N100, což je počítač který lze v roce 2024 pořídit zhruba za 4 000,- Kč.

Model	Prodleva 1. načtení (vteřin)	Tokenů za vteřinu
Llama 3 2:1b	4	14
Llama 3 2:3b	5	6
Llama 3 1:8b	12	2
Gemma 2:9b	15	2
Qwen 2.5:7b	12	3

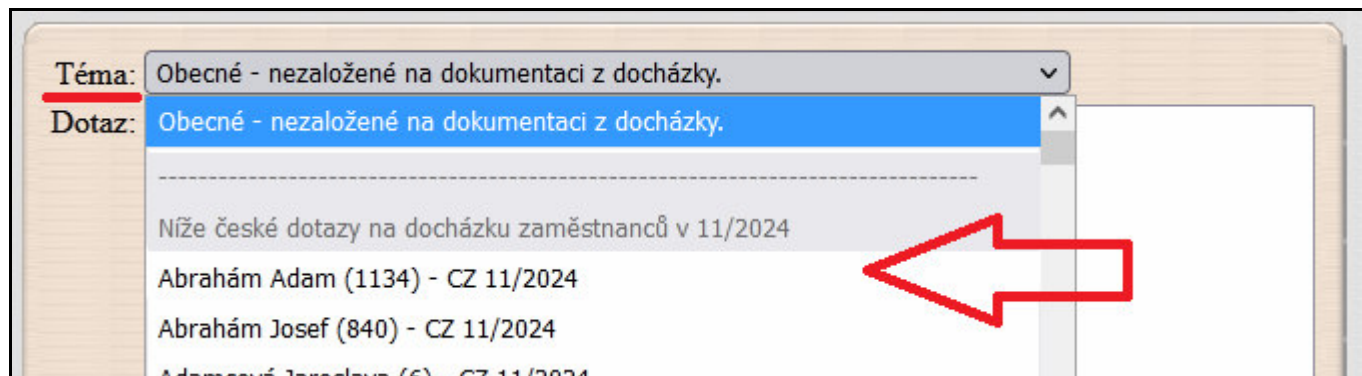
Položka „*Prodleva 1. načtení (vteřin)*“ udává dobu za kolik vteřin se model inicializuje pokud jej použijete poprvé nebo pokud jste AI nepoužívali delší dobu. Při prvním dotazu si totiž systém musí načíst celý model z disku do paměti RAM, což zabere několik vteřin v závislosti na velikosti modelu, rychlosti disku, paměti, CPU, sběrnic atd. Poté se již model udržuje v paměti, pokud jí počítač nepotřebuje na něco jiného nebo pokud neuplyne nastavená doba bez položení dalšího dotazu (ve výchozím nastavení je to 5 minut). Takže pokud s modelem pak aktivně komunikujete, již je vše načteno a tato prodleva se u dalších dotazů již neprojevuje. U složitých dotazů obsahujících mnoho slov (či slabik / znaků) se ještě přidává doba na porozumění dotazu.

Položka „*Tokenů za vteřinu*“ udává rychlost s jakou pak model generuje (zobrazuje) svojí odpověď. Tokem může být písmeno, znak, slabika či kratší slovo. Platí že čím větší model zabírající více paměti RAM, tím náročnější je jeho zpracování odpovědi a tím pomaleji se odpověď zobrazuje. O to je ale u větších modelů zpravidla odpověď přesnější oproti modelům menším.

Testovali jsme i model Llama 3.3 se 70 miliardami parametrů (70b) uvolněný společností Meta koncem roku 2024. Testování na CPU staršího serveru s 2 procesory Xeon 5560 a 48 GB RAM (model zabral 43GB na disku a 46GB paměti RAM) dosáhlo výsledku 0.04 tokenů za vteřinu. Tedy na tomto HW byl tento velký model naprosto nepoužitelně pomalý. Jeho odpovědi byly sice vynikající co do relevance, ale zabraly desítky minut času při plné zátěži HW serveru. Provozovat tyto modely je nutné jen na špičkových grafických kartách.

Novinky vyšších verzí:

Ve verzi 9.74 byla rozšířena funkcionalita modulu umělé inteligence o dotazy na výpočet docházky. V modulu AI přibyla možnost dotazovat se jazykových modelů na docházku konkrétních zaměstnanců. V rozbalovacím menu *Téma* jsou nové sekce položek nazvané "*Dotazy na docházku zaměstnanců v měsíci*", kde vyberete konkrétní měsíc a požadovaného pracovníka. Lze také zvolit zda má být dotaz zformulovaná v češtině (výchozí volba) nebo v angličtině.



Program vám poté předvyplní okno pro zápis dotazu včetně doplnění načipovaných dat a časů příchodů a odchodů. Navíc pomocí lišty zatržitek můžete nechat do dotazu vložit také požadavky na odečet neplacených přestávek v práci, výpočet hodin příplatků za přesčasy, noční práci, práci ve svátky a o víkendu atd. Tyto doplňující dotazy jsou konstruovány na míru pro konkrétního zaměstnance, takže v dotazu je zohledněna v programu nastavená doba noční práce, fond a délka přestávek dle kategorie přidělené pracovníkovi, běžný fond pracovní doby zaměstnance pro výpočet přesčasu atd. Text dotazu můžete samozřejmě dále upravit dle vlastní potřeby, pokud vám nebude vyhovovat programem zformulovaný dotaz.

Téma:

Param.: Přesčas Přestávky Noční Víkendy Svátky

Dotaz:
Dne 02.12.2024 příchod v 07:33
Dne 02.12.2024 odchod v 16:03
Dne 05.12.2024 příchod v 08:06
Dne 05.12.2024 odchod v 16:36
Dne 06.12.2024 příchod v 07:49
Dne 06.12.2024 odchod v 13:00
Dne 07.12.2024 příchod v 08:40
Dne 07.12.2024 odchod v 10:41
Dne 08.12.2024 příchod v 08:00
Dne 08.12.2024 odchod v 12:06

Datумы jsou ve formátu den.měsíc.rok
Časy jsou zapsány jako hodiny:minuty
Uveď i celkový měsíční součet odpracované doby.

Dny s těmito daty 7.12.2024, 8.12.2024 jsou víkendové. Vypiš součet počtu hodin odpracovaných v tyto víkendové dny jako nárok na příplatek za práci o víkendu.

Z denní odpracované doby odečti neplacenou přestávku v délce 30 minut pokud odpracovaná doba dne dosáhla 6 nebo více hodin.

Model:

Pomocí tlačítka *Odeslat* pošlete dotaz na docházku zvoleného zaměstnance ke zpracování vybranému modelu AI.


Chceme vypočítat docházku pracovníka na základě uváděných časů. Pokusím se to provést krok za krokem.

Krok 1: Výpočet jednotlivých dní

Dne 02.12.2024:
- Příchod: 07:33
- Odchod: 16:03
- Docházka: 16:03 - 07:33 = 8 hodin 30 minut

Dne 05.12.2024:
- Příchod: 08:06
- Odchod: 16:36
- Docházka: 16:36 - 08:06 = 8 hodin 30 minut

Dne 06.12.2024:
- Příchod

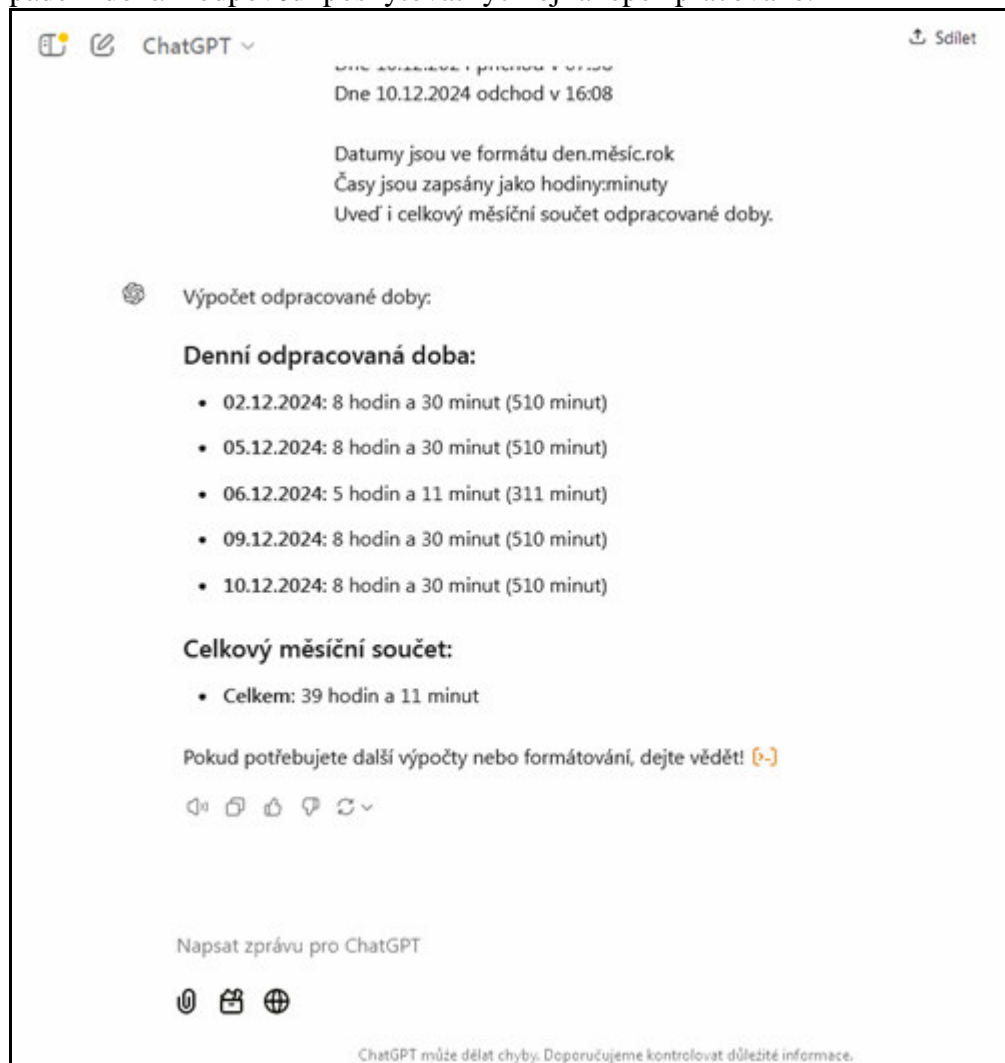
Navíc ikona pro AI výpočet docházky zaměstnanců přibyla i v menu "Zaměstnanci / Prohlížení docházky". Takže pokud chcete nechat docházku, kterou právě prohlížíte v klasické sestavě výpisu docházky, nechat zpracovat či analyzovat modulem umělé inteligence, nemusíte se z prohlížení přepínat do modulu AI a tam zaměstnance a období znovu vybírat, ale stačí přímo v prohlížení nahoře v menu kliknout na modrou ikonu 



a dotaz se odešle do modulu AI. Tam můžete případně dotaz ještě upravit buď dopsáním textu nebo jednoduše zatržením předpřipravených úprav dotazu o požadavek na dopočítání přestávek, přesčasů, nočních, víkendů, svátků atd. Poté nechat dotaz zpracovat umělou inteligencí kliknutím na tlačítko *Odeslat*. Výhodou tohoto přístupu přes prohlížení docházky je i to, že můžete vybrat období i 10 let dozadu.

Je ale třeba upozornit, že odpověď může v závislosti na vytížení serverů trvat delší dobu a pokud vyberete malý model, který je sice rychlý, ale výpočty LLM nemusí být bezchybné. Větší modely jsou na tom podstatně lépe co se týče správnosti analýz a výpočtů, ale zase jsou náročnější na HW serveru.

Možnosti dotazu na konkrétní docházku a zaměstnance ale můžete využít i tak, že programem zformulovaný dotaz s daty načipované docházky zkopírujete a vložíte je například do ChatGPT nebo podobného velkého LLM systému poskytovanému na webu jinými dodavateli, které běží na mnohem výkonnějším HW a tím pádem dokáží odpovědi poskytovat rychleji a lépe zpracované.



Takže i když nemáte na svém počítači žádné LLM modely nainstalované, lze AI modul docházky využít jen k formulování dotazů doplněných o data načipované docházky a samotné zpracování pak přenechat úplně jiným nástrojům dostupným volně na internetu, které mají mnohem lepší výkon a přesnější odpovědi.

Tímto se vám otevírají zajímavé možnosti pro zpracování docházky, kdy samotný docházkový program bude primárně jen sbírat data o načipovaných příchozech a odchodech, kdežto zpracování, sumáře a sestavy pak přenecháte umělé inteligenci. V současné době sice ještě není AI založená na menších LLM ideální pro zpracování tohoto typu úloh založených na přesných datech a může dělat některé i základní chyby, ale s ohledem na prudký rozvoj této technologie je v budoucnu možné, že docházkové systémy budou AI využívat čím dál více a pro uživatele bude snadnější sdělit počítači své požadavky na výpočet a jeho parametry slovně, než vše nastavovat v různých menu programů.

Od verze 9.78 lze při tvorbě měsíčního plánu směn v generátoru použít pro plánování AI. V menu "Kategorie / Denní rozpis směn / Měsíční plán hromadný / Generuj plán směn" zatrhnout novou volbu "Generovat i dotaz pro AI".

vána počtu typu směn jeonono zaměstnanče v měsíci: Nejvyšší (+3) ▾
 Váha rozložení směn mezi jednotlivé zaměstnanče: Nejvyšší (+3) ▾
 Generuj

Výběr algoritmu pro generování: LTRC ▾ (doporučen LTRC)
 Paralelní zpracování na: 1 CPU (pro víceprocesorové servery, max. 8)
 Spustit jako aplikaci Win32: (jen pokud server běží na OS Windows)
 Generovat i dotaz pro AI: (dotaz do AI překopírujete)

Program pak předgeneruje plán směn podle zadaných podmínek a pomocí tlačítka "Generuj pomocí AI" navrhne text úkolu generování směn se zohledněním zadaných podmínek, pracovníků, dní, typů směn atd.

5 - Noční	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14 - Volno	3	3	3	3	5	4	4	3	4	4	3	5	3	3		

Uložit rozpis Zpět na zadávací podmínky Generuj pomocí AI Načíst plán směn od AI

Naformulovaný dotaz se přenese do modulu umělé inteligence zobrazeném v nové záložce prohlížeče. Můžete si tedy nechat podle tohoto zadání vygenerovat plán směn zaměstnanců na zvolený měsíc pomocí AI.

Dotaz: Jsi expert na plánování směn zaměstnanců ve firmě ve které se pracuje na směny.

Tvým úkolem je rozepsat zaměstnancům měsíční plán směny skládající se z těchto směn:
 Směna číslo 3 na které musí být každý den přiděleno 2 zaměstnanců.
 Směna číslo 4 na které musí být každý den přiděleno 2 zaměstnanců.
 Směna číslo 5 na které musí být každý den přiděleno 1 zaměstnanců.
 Směna číslo 14 na které musí být každý den přiděleno 3 zaměstnanců.

Zaměstnanci, pro které budeš plánovat směny, jsou označeni těmito osobními čísly:
 8, 823, 7, 853, 809, 1147, 824, 1150,
 Celkem tedy přiděluješ směny 8 zaměstnancům. Do plánu směn uváděj jejich osobní čísla.

Přidělované směny každému ze zaměstnanců musí splňovat tyto podmínky:
 Po jedné směně číslo 5 musí následovat směna číslo 14

Model: D3000 Cloud Qwen 2.5:7b-Cz Odeslat do AI v cloudu AI Konfigurace lokální AI

AI vygeneruje tabulku přidělení směn zaměstnancům. Pokud by vám nevyhovovala odpověď od základních malých modelů dostupných v docházce, lze díky kroku se zobrazením textu dotazu pro generování směn využít nejen LLM modely dostupné v docházce, ale i externí, jako je například ChatGPT či jiné AI od jiných dodavatelů.

ChatGPT
Plan Smen Leden2025

... kterou ma tento zaměstnanec na tento den naplánovanou.

Vytvoření směn pro daný měsíc vyžaduje postupné splnění všech pravidel a rovnoměrné rozdělení směn mezi zaměstnance. Zohledním všechny požadavky a vytvořím plán směn ve zvoleném formátu. Pro začátek vytvořím algoritmus, který přiděluje směny na první týden, a poté postupně doplním zbytek měsíce.

«» Plan Smen Leden2025

Odpovědět místo plochy v chatu >


První týden plánu směn je připraven. Můžete jej zkontrolovat. Pokud je vše v pořádku, můžete pokračovat v plánování.

1	3, 823, 3
2	3, 7, 3
3	3, 853, 4
4	3, 809, 4
5	3, 1147, 5
6	3, 824, 14
7	3, 1150, 14
8	
9	6, 823, 3
10	6, 7, 3
11	6, 853, 4
12	6, 809, 4
13	6, 824, 14
14	6, 8, 5
15	6, 1150, 14
16	
17	7, 823, 3
18	7, 8, 3
19	7, 853, 4
20	7, 809, 4
21	7, 1147, 14
22	7, 7, 5
23	7, 1150, 14
24	
25	8, 823, 3
26	8, 7, 3

Tuto tabulku pak můžete naimportovat zpět do docházky přes nové tlačítko "Načíst plán směn od AI". Stačí v odpovědi od AI myší označit tabulku rozpisu směn (Ctrl-A), dát kopírovat (Ctrl-C), na předchozí záložce prohlížeče kliknete na tlačítko "Načíst plán směn od AI" :

5 - Noční	1	1		1	1	1	1	1		1	1	1	1	1		1
14 - Volno	3	3		3	3	5	4	4		3	4	4	3	5		3

Uložit rozpis Zpět na zadávací podmínky Generuj pomocí AI Načíst plán směn od AI




A do zadávacího pole dat tabulku vložíte (Ctrl-V):

```
3, 823, 3
3, 7, 3
3, 853, 4
3, 809, 4
3, 1147, 5
3, 824, 14
3, 1150, 14

6, 823, 3
6, 7, 3
```

Oddělovač položek: , (jediný znak, měla by jím být čárka)

Načíst plán směn od AI do programu



Pokud by vám AI v tabulce místo oddělovač částky použila jiný znak (například |) můžete zadejte jej do pole *Oddělovač položek*. Nakonec kliknete na tlačítko „Načíst plán směn od AI do programu“ a zobrazí se plán směn ve formě tabulky:

Níže je zobrazen plán směn načtený z AI:

Pracovník	St 1.	Čt 2.	Pá 3.	So 4.	Ne 5.	Po 6.	Út 7.	St 8.	Čt 9.	Pá 10.	So 11.	Ne 12.	Po 13.	Út 14.	St 15.	Čt 16.	Pá 17.	So 18.	Ne 19.	Po 20.	Út 21.	St 22.	Čt 23.	Pá 24.	So 25.	Ne 26.	Po 27.	Út 28.	St 29.	Čt 30.	Pá 31.	Směn
Šálková Jana (7)	1.S	3.R	3.R	1.S	1.S	3.R	5.N	3.R	14.V	3.R	1.S	1.S	5.N	14.V	3.R	4.O	14.V	1.S	1.S	4.O	14.V	4.O	14.V	3.R	1.S	1.S	3.R	5.N	14.V	14.V	5.N	22
Prát Luděk (8)	1.S	14.V	4.O	1.S	1.S	5.N	3.R	5.N	3.R	5.N	1.S	1.S	14.V	3.R	3.R	14.V	5.N	1.S	1.S	14.V	3.R	14.V	3.R	14.V	1.S	1.S	4.O	14.V	3.R	14.V	3.R	22
Zich Karel (809)	13.D	13.D	4.O	1.S	1.S	4.O	4.O	4.O	4.O	14.V	1.S	1.S	3.R	5.N	14.V	4.O	14.V	1.S	1.S	14.V	3.R	3.R	14.V	4.O	1.S	1.S	3.R	3.R	14.V	5.N	14.V	23
Svobodová Bohuslava (823)	1.S	3.R	3.R	1.S	1.S	3.R	3.R	3.R	3.R	4.O	1.S	1.S	14.V	4.O	14.V	3.R	4.O	1.S	1.S	5.N	14.V	3.R	3.R	14.V	1.S	1.S	4.O	14.V	3.R	4.O	14.V	22
Pracovník DPP (824)	1.S	14.V	14.V	1.S	1.S	14.V	14.V	14.V	5.N	14.V	1.S	1.S	14.V	4.O	14.V	14.V	3.R	1.S	1.S	3.R	5.N	14.V	14.V	5.N	1.S	1.S	14.V	3.R	5.N	14.V	14.V	22
Blatný Josef (853)	1.S	4.O	4.O	1.S	1.S	4.O	4.O	4.O	4.O	14.V	1.S	1.S	3.R	3.R	14.V	3.R	14.V	1.S	1.S	4.O	14.V	5.N	14.V	4.O	1.S	1.S	5.N	14.V	14.V	3.R	4.O	22
Vaněk Karel (1147)	13.D	13.D	5.N	1.S	1.S	5.N	14.V	4.O	14.V	3.R	1.S	1.S	4.O	14.V	14.V	5.N	14.V	1.S	1.S	3.R	14.V	4.O	14.V	3.R	1.S	1.S	14.V	4.O	14.V	4.O	14.V	23
Wagner Josef (1150)	13.D	13.D	14.V	1.S	1.S	14.V	14.V	14.V	14.V	14.V	1.S	1.S	4.O	14.V	5.N	14.V	3.R	1.S	1.S	14.V	4.O	14.V	5.N	14.V	1.S	1.S	14.V	4.O	14.V	3.R	3.R	23

Počty pracovníků na směnách:

1 - Standardní	5			8	8						8	8						8	8						8	8						
3 - Ranní		2	2			2	2	2	2	2			2	2	2	2	2			2	2	2	2	2			2	2	2	2	2	
4 - Odpolední		1	3			2	2	3	2	1			2	2		2	1			2	1	2		2			2	2		2	1	
5 - Noční			1			2	1	1	1	1			1	1	1	1	1			1	1	1	1	1			1	1	1	1	1	
13 - Denní 12h	3	3																														
14 - Volno		2	2			2	3	2	3	4			3	3	5	3	4			3	4	3	5	3			3	3	5	3	4	

Uložit rozpis



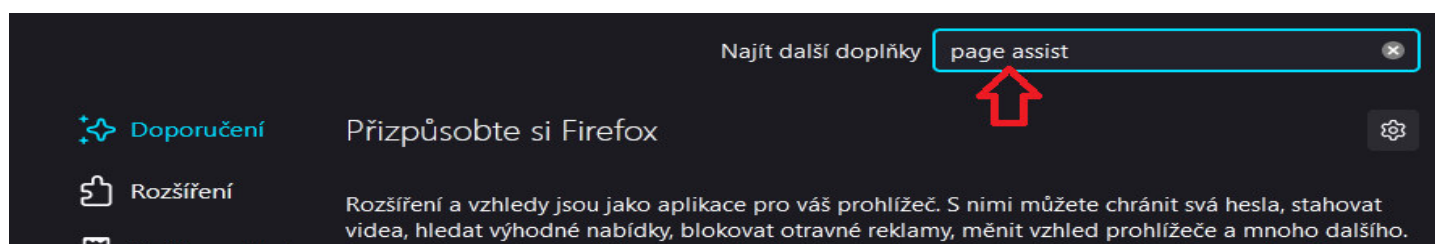
Zpět na zadávací podmínky

Pokud je vše v pořádku a plán směn odpovídá vaší představě a zadání, uložte jej do databáze docházky tlačítkem *Uložit rozpis* pod tabulkou. Další podrobnosti naleznete v příručce *Nastavení směn* přímo v programu v menu *Firma / Návody PDF*, kde je vše vysvětleno a předvedeno na příkladech.

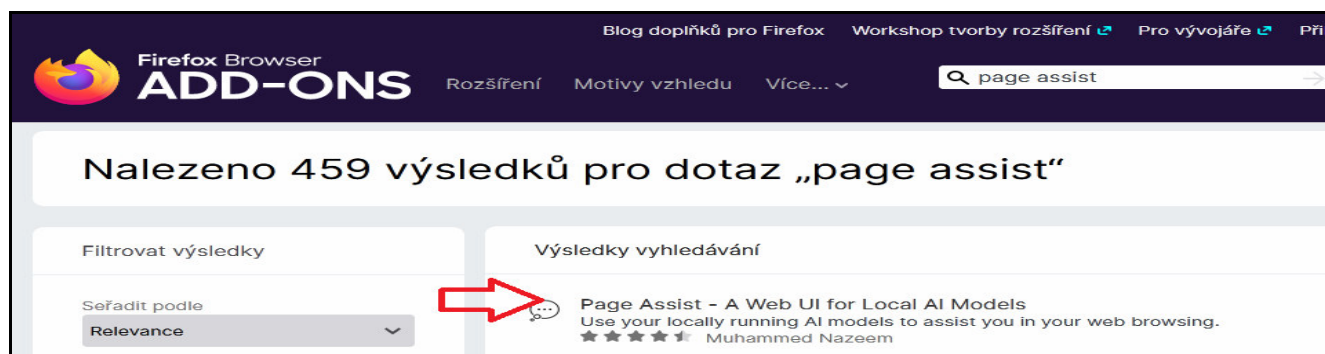
Text dotazu pro AI navržený docházkou lze navíc před odesláním do AI ručně upravit, takže si můžete doplnit své další vlastní požadavky na přidělování směn.

Využití AI i mimo docházku:

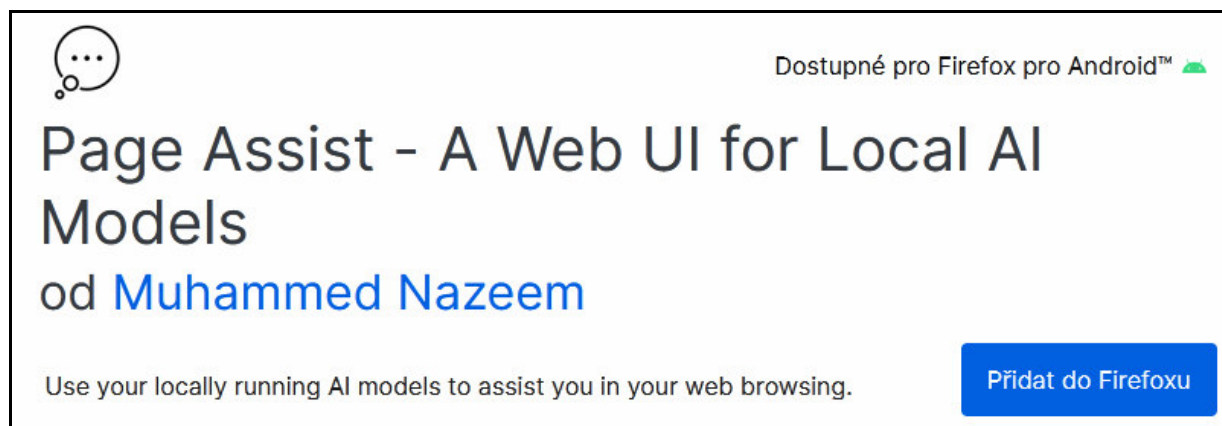
Když už máte ve firmě díky výše uvedeným pokynům nainstalovanou lokální AI, tak pokud jí máte na výkonném serveru ideálně vybaveným výkonnou grafickou kartou, možná by byla škoda nevyužít přístup k této lokální umělé inteligenci i pro běžné uživatele. Existuje celá řada možností. Jednou z nich může být to, že si do webového prohlížeče doinstalujete rozšiřující doplněk *Page Assist*. Jedná se o praktické rozšíření funkcí prohlížeče (v příkladu níže je použit Mozilla Firefox), která zpřístupní AI uživatelům i mimo docházku a výborně spolupracuje právě se systémem Ollama. V prohlížeči v menu „*Nástroje / Nastavení / Rozšíření a vzhledy*“ zadáte do vrchního vyhledávacího pole text *Page assist*:



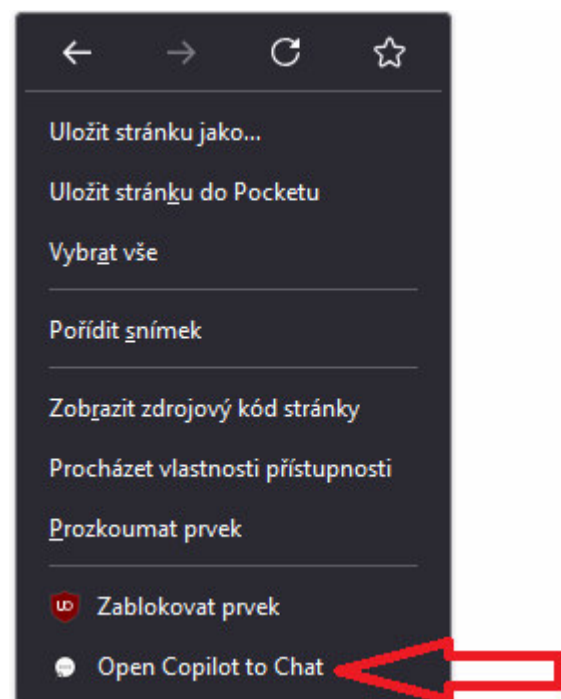
Zobrazí se seznam vyhovujících doplňků a tento je hned jedním z prvních:



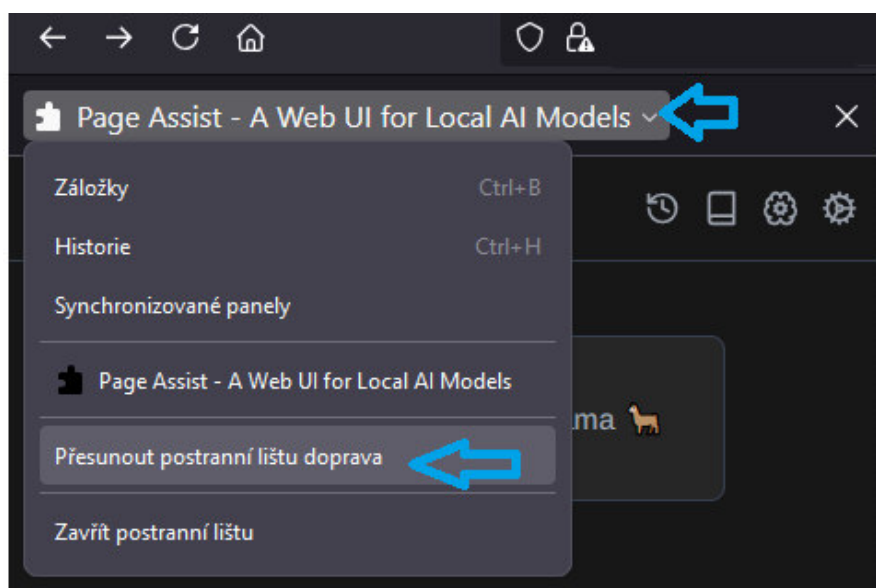
Stačí tedy na doplněk *Page Assist* kliknout a poté potvrdit přidání do webového prohlížeče kliknutím na tlačítko *Přidat do Firefoxu*:



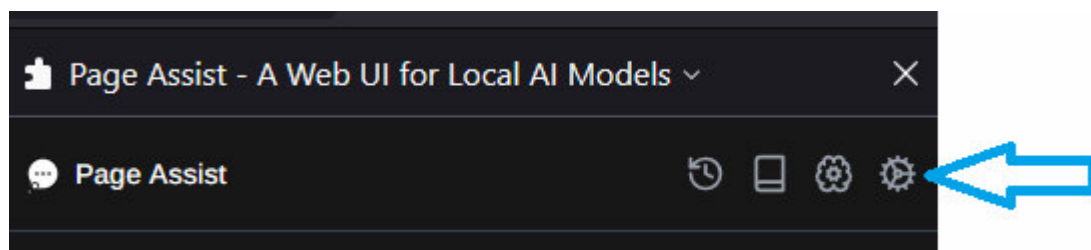
Jakmile se doplněk nainstaluje, stačí někde v prohlížeči kliknout pravým tlačítkem myši a vybrat novou volbu „*Open Copilot to Chat*“. Pokud by volba v přehledu nabídky pravého tlačítka myši nebyla, vyberte na nějaké webové stránce kousek textu a na ten pak klikněte pravým tlačítkem myši.



Doplněk se zobrazí v levé části obrazovky, ale pomocí obrázku níže jej můžete přesunout doprava, což bývá ergonomičtější:

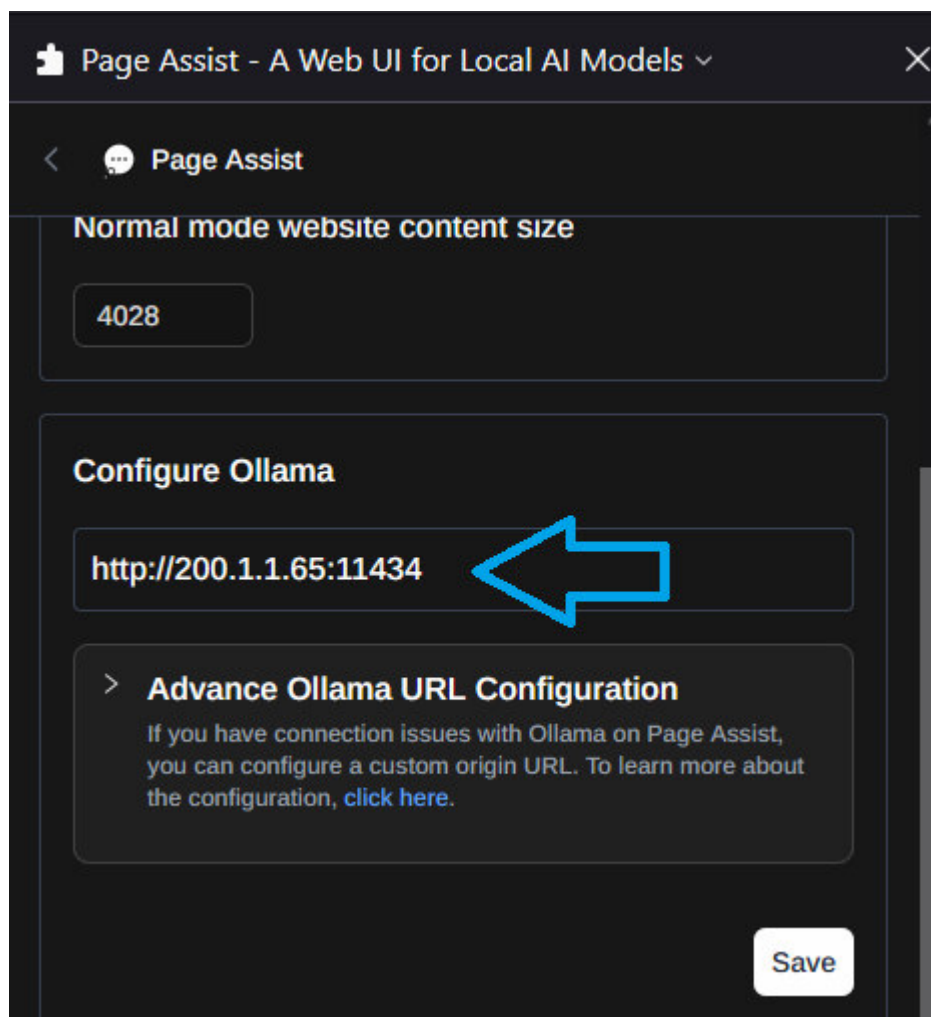


Pokud je toto klientský počítač na kterém není AI přímo nainstalovaná, tak nyní je potřeba v *Page Assist* nastavit spojení na PC ve kterém program Ollama běží. Tedy na ten, do kterého jste AI (program Ollama) dle pokynů z úvodních stran této příručky instalovali. Klikněte tedy na ikonku ozubeného kolečka vpravo:



Nyní potřebujete znát IP adresu PC sloužícího jako server AI, kam jste instalovali program Ollama. Stačí na něm spustit příkazový řádek a zadat příkaz *ipconfig*. Zobrazí se parametry nastavení sítě a potřebujete najít položku nazvanou *Adresa IPv4*. Bude se jednat o 4 čísla oddělená tečkami, například *192.168.1.10*

Takto zjištěnou IP adresu zadejte v konfiguračním menu *Page Assist* do sekce *Configure Ollama*



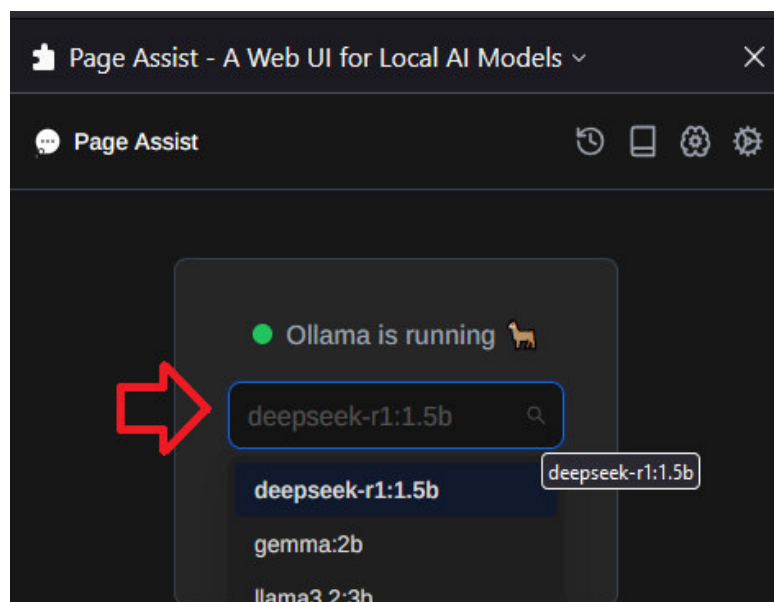
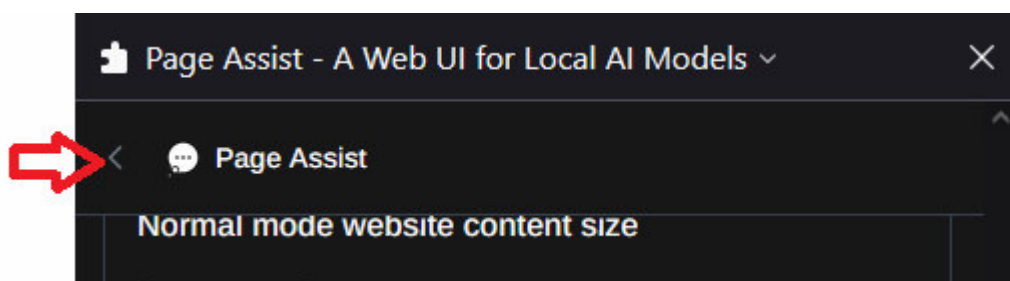
Před adresu uvedete *http://* a za adresu ještě doplníte *:11434*

Takže pokud má server s nainstalovaným programem Ollama například IP adresu *192.168.1.10*, zadáte do *Configure Ollama* toto:
http://192.168.1.10:11434
a o trochu níže kliknete na tlačítko *Save*, čímž se nastavení uloží.

Pokud by se poté nedařilo s AI pracovat a program by psal že nemá spojení na systém Ollama, může být důvod buď v tom, že na hlavní PC s AI není v nastavení brány firewall systému Windows povolený přístup na TCP port 11434. Takže do firewallu na PC s programem Ollama vložte povolení na přístup na port 11434 ze sítě.

Dalším důvodem může být to, že v nastavení PC s programem Ollama není v konfiguraci proměnných prostředí zadaná systémová proměnná prostředí *OLLAMA_HOST* na hodnotu *0.0.0.0* – viz strana 8 a 9.

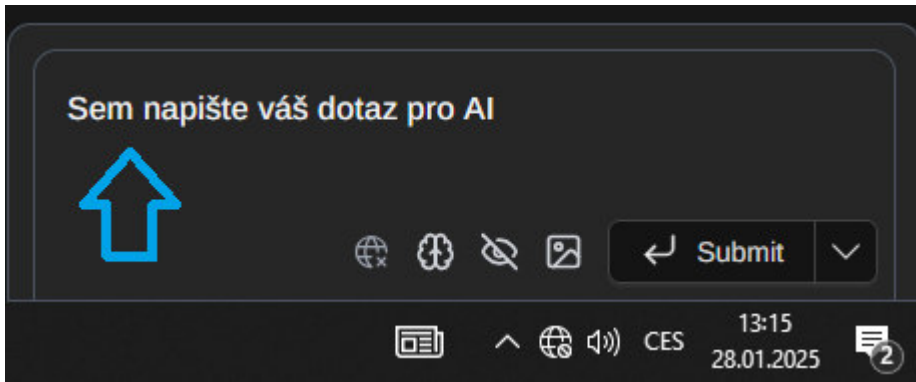
Jakmile bude vše v pořádku nastaveno a spojení tohoto klientského PC na počítač s programem Ollama bude fungovat, ukončíte nastavení konfigurace malou šedou šipkou vpravo nahoře.




Následně můžete rozkliknout výběr dostupných AI modelů v malém okně pod nápisem *Ollama is running* a vyberete si model který vám pro práci s AI vyhovuje nejvíce.



K výběru jsou všechny modely nainstalované na PC se systémem Ollama, takže pokud jste podle pokynů z výše uvedených stránek této příručky již modely stáhli, není třeba nic dalšího instalovat a vše je již funkční.

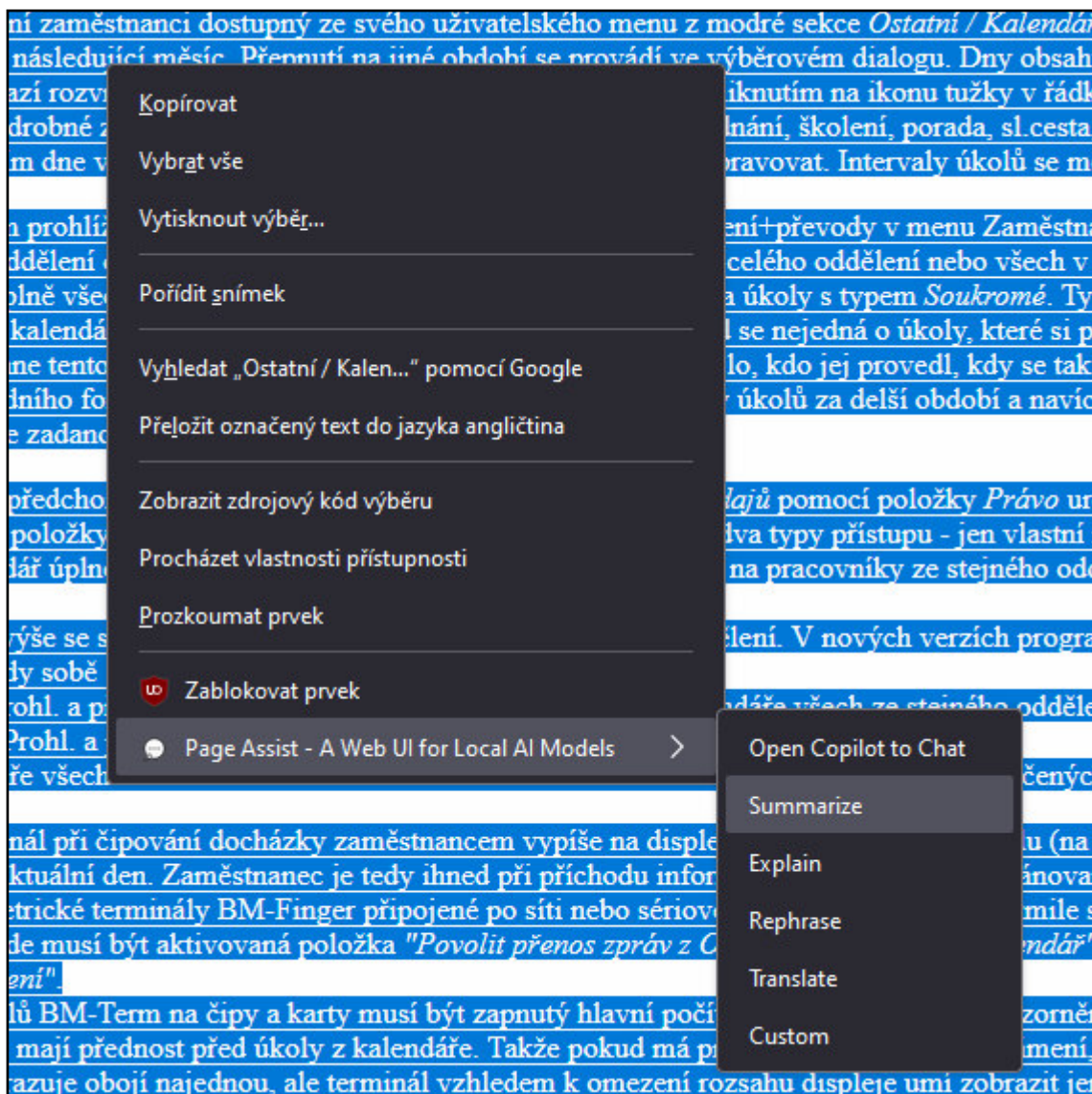
Pokud by spojení přesto stále nefungovalo, můžete v konfiguraci doplnku přes *Advance Ollama URL Configuration* povolit volbu *Enable Custom Origin URL* a zadat stejnou IP adresu včetně úvodního *http://* a koncovky *:11434*



Samotná interakce s AI pak probíhá například tak, že napíšete váš dotaz či požadavek do dolní části pro vkládání dotazů.

Přes ikonu  můžete povolit AI dohledávat podrobnosti k dotazům na webu, takže získá přístup k aktuálním informacím.

Přes ikonu  zase můžete změnit používaný LLM model za jiný nainstalovaný. Jelikož si přes *Page Assist* používaný LLM model udržuje historii toho co jste se jej psali a při každém dalším dotazu o předchozí text rozšíří *Page assist* kontextové okno, je možné se doptávat na upřesnění předchozích odpovědí a vést tak s AI v podstatě dialog. Pokud ale chcete začít úplně nové téma, je potřeba nejprve kliknout nahoře na ikonu  kterou se nové téma založí a tím vymaže předchozí kontextové okno s historií dialogu. Jinak jsou odpovědi na nové téma ovlivněny předchozí konverzací, což může být nežádoucí.



Pokud chcete s AI probírat text na webové stránce kterou si právě prohlédíte, stačí text vymodřit myší, kliknout na něj pravým tlačítkem, rozkliknout dole šipku u volby *Page assist ...* a buď přes *Chat* přenést text do dotazu s možností úpravy, nebo rovnou vybrat vysvětlení textu (*Explain*) nebo shrnutí textu (*Summarize*) překlad, přepis atd.

U česky psaných textů doporučujeme volit modely podporující češtinu, jako je z menších modelů např. gemma-2-9b nebo llama 3.2:8b případně qwen2.5:7b Samozřejmě čím větší model, tím lépe.

Větší modely ale zase na druhou stranu déle zpracovávají zadání dotazu, takže jim začátek reakce na pomalejším HW trvá delší dobu a i samotná rychlost odpovědi může být pomalá, pokud větší model běží na CPU. Což může velmi ovlivnit uživatelský zážitek z používání lokální AI, pokud není systém Ollama nainstalovaný na dostatečně výkonném HW. I použití starší grafických karet nemusí být úplně přínosem. Například grafická karta Nvidia Geforce GTX 750 Ti se 2GB vRAM nestačí výkonu dnešních CPU ani co do

rychlosti ani co do množství využitelné paměti. I dnešní nejlevnější procesory jako je Intel N100 mají stejný nebo lepší výkon. Takže i když je tato stará karta sice systémem Ollama podporovaná i dnes, použít ji lze jen s malými modely 1b až 2b které se do její paměti vejdou. A ty většinou fungují obstojně jen v angličtině, jako například model *Deepseek-R1:1.5b*, který na této kartě funguje.

```
Device 0 [NVIDIA GeForce GTX 750 Ti] PCIe GEN 2@16x RX: 36.72 MiB/s TX: 16.85 MiB/s
GPU 1150MHz MEM 2700MHz TEMP 64fC FAN 42% POW 24 / 38 W
GPU[||||||||||||||||||||| 57%] MEM[|||||||||||||||||||||1.529Gi/2.000Gi]
PID USER DEV TYPE GPU GPU MEM CPU HOST MEM Command
985 ollama 0 Compute 56% 1485MiB 73% 209% 983MiB /usr/local/lib/ollama/r
786 root 0 Graphic 0% 12MiB 1% 0% 77MiB /usr/lib/xorg/Xorg :0 -
```

Pokud tedy máte možnost použít pro AI server či počítač s grafickou kartou Nvidia řady RTX, bude uživatelský zážitek z používání lokální AI výrazně příjemnější. Základní karty jako RTX 4060 či 4070 stačí na menší modely které se vejdou do jejich 8 až 16 GB vRAM a cena těchto karet není příliš vysoká. Pro větší modely perfektně podporující češtinu lze doporučit například kartu RTX 3090 s 24GB RAM, která je cenově na podobně hladině jako nižší modely řady RTX 40xx ale přitom má více paměti. Rychlost těchto karet je při zpracování AI v porovnání s CPU mnohonásobně vyšší. Ale je potřeba počítat i s vyšším odběrem elektřiny. Dobrým kompromisem může být použití počítačů Apple Mac mini s procesorem M4 a dostatečným množstvím RAM, protože spotřeba elektřiny je u těchto počítačů minimální a výkon v AI úlohách více než dostatečný.