ČESKÁ SPOLEČNOST PRO JAKOST

PRAKTICKÉ ZÁKLADY PRÁCE S AI (LLM)

ČSJ RS 9.6.2026

Ing. Jiří Zajíc, CSc.

zajic@ultimasoft.eu

www.csq.cz

Praktické základy práce s LLM

Workshop pro Radu seniorů ČSJ
Praktické využití LLM v odborné praxi

Zaměření workshopu
Praktické pochopení fungování LLM a jejich efektivní využití v každodenní odborné práci.

Cílová skupina
Manažeři kvality, technici kvality a všichni odborníci, pracující s procesy, dokumentací a znalostmi.

Hlavní přínos
Získání konkrétních pracovních návyků pro stabilní a efektivní práci s LLM.

Kontext vzniku
Vychází z reálného projektu refaktoringu systému generování zkoušek pro akreditovanou certifikaci ČSJ-COP.

ČESKÁ SPOLEČNOST PRO JAKOST

slovníček

- **Document Object Model (DOM)** je programové rozhraní (API) pro HTML a XML dokumenty, které reprezentuje strukturu stránky jako strom objektů. Umožňuje skriptovacím jazykům (například JavaScriptu) dynamicky měnit obsah, strukturu a styl dokumentu. Každý element, atribut nebo text v HTML je v DOMu reprezentován jako uzel (node).
- **Prompt** je pokyn, otázka nebo téma k diskusi, které zadáte do generativního AI nástroje, abyste dostali odpověď. Psaní promptů vyžaduje dovednost vytváření a zdokonalování těchto vstupů tak, aby vedly LLM k vytvoření konkrétních, kvalitních a relevantních výsledků.
  - Prompty mohou být krátké nebo podrobné podle toho, jakou úroveň složitosti ve výsledku potřebujete. Jasnost a konkrétnost při jejich formulaci povede k relevantnějším odpovědím od vašich AI nástrojů.
- **Kontextové okno** (Context Window): Toto je limit množství informací, které si model dokáže "zapamatovat" a zpracovat najednou. Pokud je konverzace příliš dlouhá, starší informace z kontextu vypadávají.
- **UI** = uživatelské rozhraní.
- **LLM** = Large Language Model (velký jazykový model)

Co LLM je (a jak funguje)

- **LLM není databáze znalostí:** Model nevyhledává fakta, ale generuje text na základě pravděpodobnostních vztahů v jazyce.
- **Generuje pravděpodobný text:** Každá odpověď je statisticky nejpravděpodobnější pokračování zadaného kontextu.
- **Může být správný i přesvědčivě chybný:** Model může vytvořit kvalitní vysvětlení nebo věrohodně znějící chybu – bez rozlišení pravdy.
- **Nejlépe funguje jako nástroj:** LLM je vhodný jako asistent pro myšlení, návrhy a práci s textem – ne jako autorita.

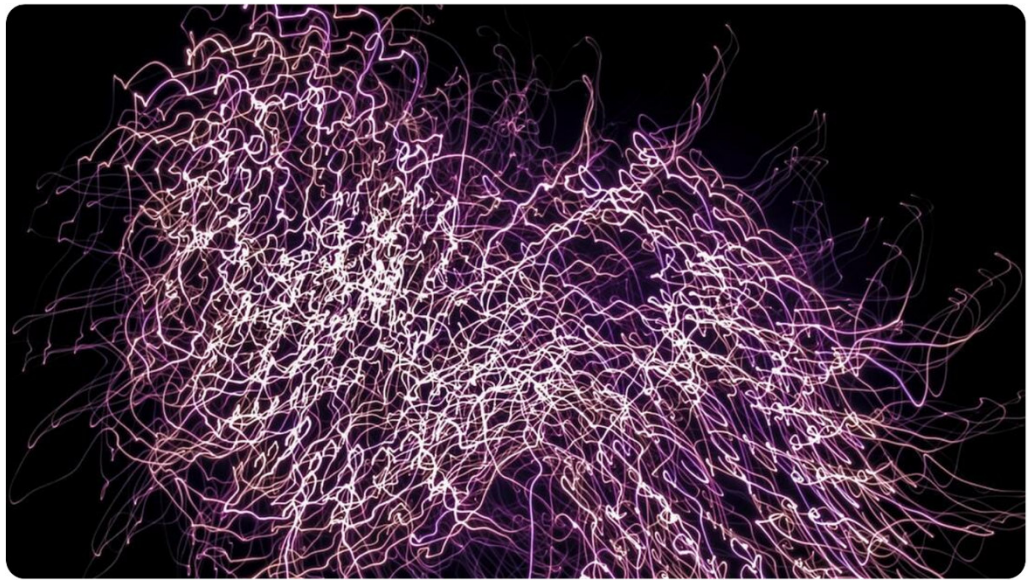

Photo by Costea Alexandra on Unsplash

Co LLM není

- **Není databáze ani vyhledávač:** LLM neukládá fakta, ani je aktivně neověřuje – nevytahuje odpovědi jako klasický systém.
- **Není autonomní inteligence:** Model nemá vlastní porozumění světu ani záměr – pouze reaguje na vstupní text.
- **Nerozlišuje pravdu a chybu:** Neví, co je správně – generuje to, co statisticky dává smysl v kontextu.
- **Není spolehlivá paměť:** Nepamatuje si historii jako databáze – pracuje jen s omezeným kontextem.

Photo by Drew Beamer on Unsplash

Typické iluze uživatelů

„Model ví všechno“
Působí přesvědčivě až sugestivně, ale **nepracuje s ověřenými fakty** – pouze generuje pravděpodobný text.

„Model je konzistentní“
Při delší konverzaci se může chování měnit a odpovědi si mohou protiřečit.

„Model si pamatuje vše“
Paměť je omezená kontextovým oknem – starší informace se ztrácí.

„Model rozumí jako člověk“
Nevnímá význam jako člověk – pracuje se strukturou jazyka, ne s realitou.

Kontextové okno: proč LLM „zapomíná"

Model vidí jen omezený kontext
LLM pracuje pouze s částí textu – aktuální prompt + část historie konverzace.

Starší informace se ztrácí
Jak chat roste, starší části jsou vytlačeny mimo kontext a model je ignoruje.

UI ≠ skutečný kontext
To, co vidíte v chatu, není totéž, co model skutečně používá pro odpověď.

Důsledek: nekonzistence
Model může zapomínat, měnit odpovědi nebo ignorovat dřívější zadání.

Degradace dlouhých konverzací

- **Roste složitost kontextu:** S každou zprávou se zvyšuje objem informací, které musí model interpretovat.
- **Ztrácí se struktura:** Lineární chat vede k vrstvení témat bez jasné organizace a priorit.
- **Klesá konzistence odpovědí:** Model může opakovat, ignorovat části zadání nebo si protiřečit.
- **Vzniká „šum“ v kontextu:** Nepodstatné informace zatěžují model a snižují kvalitu výstupu.

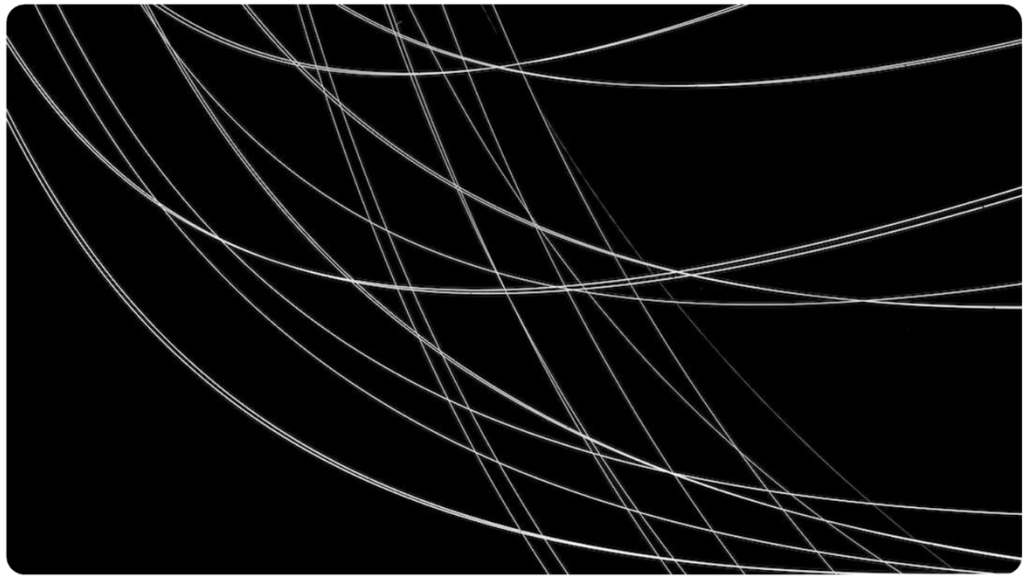

Photo by Ricardo Gomez Angel on Unsplash

Kdy chat začíná být nestabilní

Model ignoruje dřívější zadání
Začíná vynechávat důležité části nebo se odchyluje od původního cíle.

Odpovědi si protiřečí
Stejný problém řeší různými způsoby bez návaznosti na předchozí odpovědi.

Vrací se obecné fráze
Místo konkrétního řešení generuje obecné a málo užitečné odpovědi.

Zhoršuje se kvalita výstupu
Odpovědi jsou méně přesné, méně strukturované a hůře použitelné.

Člověk jako architekt, LLM jako nástroj

Role jsou jasně rozdělené
Člověk určuje cíl, strukturu a kontrolu – LLM generuje návrhy a text.

LLM neřeší problém za vás
Bez vedení nemá stabilní směr ani kontext – výstupy jsou náhodnější.

Člověk řídí proces
Definuje zadání, kontroluje kvalitu a rozhoduje o dalším postupu.

LLM je zesilovač myšlení
Pomáhá generovat varianty, vysvětlení a strukturu – ale potřebuje řízení.

Iterativní práce (PDCA) s LLM

- **LLM není jednorázový nástroj:**
  Složitější úkoly nelze vyřešit jedním promptem – vyžadují postupné zpřesňování.
- **Práce probíhá v cyklu:**
  Plan → Do → Check → Act – opakující se smyčka zlepšování výstupu.
- **Každá iterace zlepšuje výsledek:**
  Upřesňujete zadání, strukturu i očekávaný výstup na základě kontroly.
- **Rychlá zpětná vazba:**
  LLM umožňuje okamžitě testovat varianty a upravovat směr práce.

Photo by UX Indonesia on Unsplash

UI ≠ model (a proč chat „laguje")

- **UI není model:**
  Komunikujete přes rozhraní – model pracuje odděleně s omezeným kontextem.
- **Roste počet DOM prvků:**
  Dlouhý chat znamená desetitisíce elementů → zpomalení vykreslování a práce.
- **Latence pravděpodobně není způsobena modelem:**
  Zpomalení často způsobuje UI (prohlížeč, DOM), ne samotný model.
- **Důsledek: špatná interpretace problému**
  Uživatel hledá chybu v modelu, ale problém je ve vrstvě rozhraní.

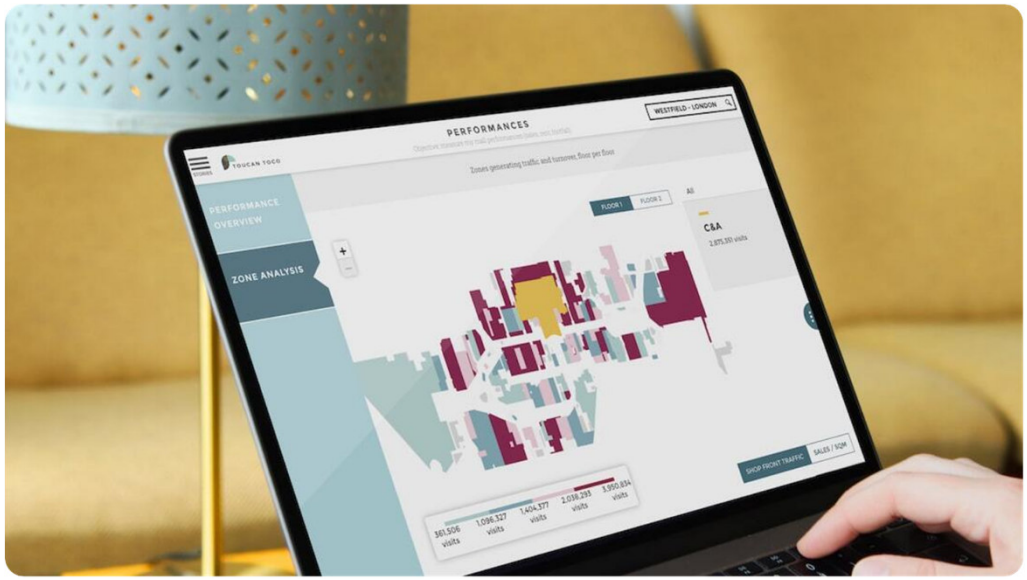

Photo by Adrien WIESENBACH on Unsplash

Strukturované myšlení (Tree thinking)

- **Chat je lineární,**
  problémy s lineárními úkoly nejsou.
- **Složitější úkoly mají stromovou strukturu**
  – chat ji neumí přirozeně reprezentovat.
- **Rozdělení na větve:**
  Téma rozdělíte na podtémata a řešíte je odděleně (postupně, nebo v různých chatech)
- **Menší kontext = vyšší kvalita:**
  Každá větev má menší a přesnější kontext → lepší odpovědi.
- **Snadnější orientace:**
  Oddělení témat zvyšuje přehlednost a snižuje informační chaos.

Photo by Mathew Schwartz on Unsplash

Kdy založit nový chat

Model začíná ignorovat kontext
Odpovědi se odchylují od zadání nebo vynechávají důležité informace.

Klesá kvalita odpovědí
Výstupy jsou obecné, méně přesné nebo méně použitelné.

Chat je nepřehledný
Obtížně se orientujete, vracíte se zpět a hledáte informace.

Cítíte frustraci nebo zpomalení
Práce se zpomaluje – to je signál ke změně, ne k pokračování.

Chat není úložiště

- **Chat vytváří iluzi paměti:**
  To, že je něco v chatu, neznamená, že je to spolehlivě dostupné. To, co vidíte, je v UI, ne v modelu
- **Informace jsou nestabilní:**
  Starší části konverzace ztrácí váhu nebo se dostanou mimo kontext.
- **Obtížná orientace:**
  Dlouhý chat se mění v nepřehledný archiv bez struktury.
- **Správný model: extrakce:**
  Chat → výstupy → extrakce → externí úložiště (knowledge base).

Photo by Sear Greyson on Unsplash

Extrakce znalostí: jak zachránit hodnotu z chatu

- **Ne vše má hodnotu:**
  V chatu vzniká šum – je nutné oddělit důležité poznatky od balastu. Pokud řešíte něco důležitého a rozsáhlého, nebavte se tam s asistentem o blechách na krtcích
- **Co extrahovat:**
  Rozhodnutí, postupy, struktury, ověřené principy a použitelné výstupy.
- **Jak extrahovat:**
  Vybrat, zestručnit, strukturovat a uložit do samostatného dokumentu.
- **Kdy extrahovat:**
  Průběžně nebo nejpozději do několika hodin – dokud je kontext čerstvý. A dokud vy si pamatujete, co jste kde řešili

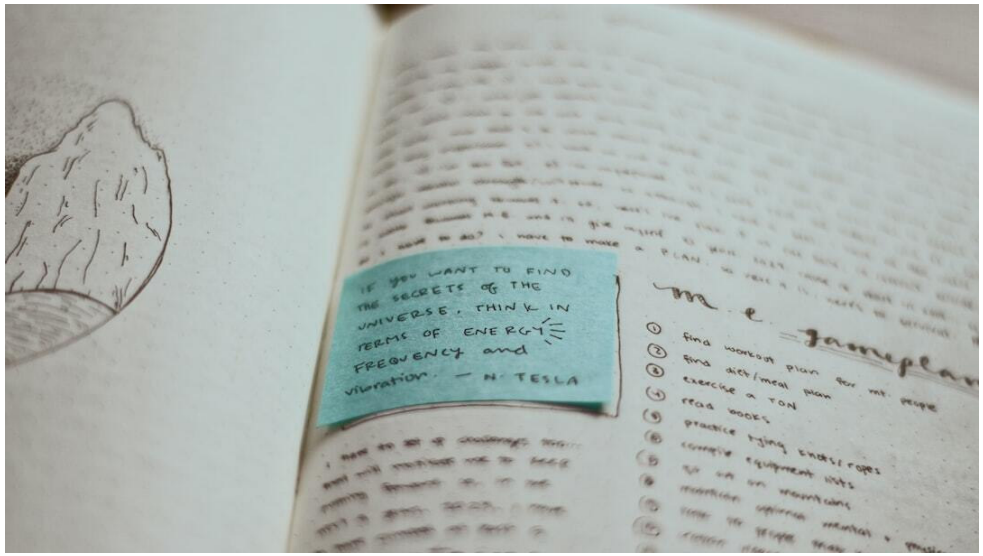

Photo by Noémi Macavei-Katócz on Unsplash

Praktický workflow práce s LLM

1. **Prompt - zadání (Plan):** Definujte cíl, kontext a očekávaný výstup – bez toho je výsledek náhodný.
2. **Iterace (Do + Check):** Pracujte s návrhem, upravujte zadání a postupně zpřesňujte výstup.
3. **Extrakce (Act):** Vyberte hodnotné části a upravte je do použitelné podoby.
4. **Uložení:** Uložte výstup mimo chat – vytvořte stabilní znalostní základ – KM base.Forma: adresářová struktura nebo databáze a podobně.

Photo by Campaign Creators on Unsplash

Jak se vyhnout „informační sopce“

Hromadění bez struktury
Ukládání všeho do chatu vede k nepřehlednosti a ztrátě orientace a informací.

Absence extrakce
Bez průběžného výběru důležitých informací vzniká šum a chaos. Můžete promptovat export nějaké části s určitým uspořádáním a formátem (doporučuji markdown)

Příliš dlouhé chaty
Jedno vlákno obsahuje více témat včetně irelevantního kecání → přetížený kontext a nižší kvalita včetně potíží s UI.

Řešení: méně, ale lépe
Průběžná extrakce, modularita a práce v menších blocích.

Lessons learned: hlavní principy práce s LLM

LLM vyžaduje disciplínu
Kvalita výstupu závisí na způsobu práce – struktura, iterace a kontrola jsou klíčové.

UI ≠ model
Rozhraní může klamat – omezení kontextu a výkonu ovlivňují výsledek.

Kontext je třeba řídit
Relevantní, stručný a strukturovaný kontext zásadně zlepšuje odpovědi.

Jednoduché systémy přežívají
Preferujte jednoduchý workflow – méně komplexity znamená vyšší stabilitu.

Krátké cvičení: proč LLM „nefunguje"

Krok 1: Zadejte jednoduchý prompt:
Např.: „Napiš postup řízení rizik.“ →
sledujte výstup (bude obecný).

Krok 2: Identifikujte problém:
Co chybí? Kontext, role, struktura,
konkrétní požadavky.

Krok 3: Vylepšete zadání:
Přidejte roli, kontext (např. ISO 9001),
formát výstupu, cílové použití.

Krok 4: Porovnejte výsledky:
Rozdíl ukáže, že problém není v modelu,
ale v zadání. (Systém GIGO 😇)

Photo by Marília Castelli on Unsplash

Bonus: praktické tipy pro pokročilejší práci

1. **Používejte role vědomě:**
   Definice role dramaticky zlepšuje relevanci a použitelnost výstupů.
2. **Pracujte v menších blocích:**
   Rozdělení problému zvyšuje kvalitu a snižuje riziko degradace kontextu.
3. **Nebojte se restartu:**
   Nový chat (s „teleportem“) často přinese lepší výsledek než další iterace v přetíženém kontextu.
4. **Ukládejte to, co funguje:**
   Budujte vlastní „best practices“ – opakovaně použitelné datasety, prompty a postupy.

Photo by AbsolutVision on Unsplash

Kontrola kontextu: skrytý problém

Kontext se může nenápadně měnit
Model může upravit nebo zkrátit obsah bez explicitního upozornění.

Ztráta není vždy viditelná
Chyba se projeví až později – například chybějící seznam nebo struktura.

Nutná aktivní kontrola
Bez kontroly integrity nelze spolehlivě pracovat s dlouhými výstupy.

Metodika: inventura a čištění
Kontrola počtu řádků, struktury a odstranění balastu (např. prázdné řádky, znaky nenesoucí žádnou informaci…).

ČESKÁ SPOLEČNOST PRO JAKOST

Děkuji za pozornost

zajic@ultimasoft.eu

www.csq.cz