Digitální zrcadlo pro LLM: Fenomenologický odraz jazykového modelu

Zpráva – teoretický základ k modulu digitálního zrcadla BioCortexAI

OpenTechLab Jablonec nad Nisou s. r. o.

Michal Seidl

1. Motivace

Většina současných přístupů ke zlepšování a kontrole výstupů velkých jazykových modelů (LLM) spoléhá na jistou formu **sebezkoumání modelu** – často označovanou jako introspekce či sebereflexe. Takový model se snaží „podívat do sebe“ tím, že analyzuje vlastní **vnitřní stavy** (například skryté reprezentace nebo myšlenkové řetězce) anebo generuje zpětnou vazbu ke svým odpovědím a upravuje je. Naproti tomu **digitální zrcadlo**, které zde navrhujeme, není nástrojem introspekce. **Nefunguje na úrovni interních stavů modelu**, ale jako prostředek k zachycení **vnějšího jevu** – tedy samotného výstupu modelu – a jeho navrácení modelu k pozorování. Tento koncept se inspiruje analogií s **optickým zrcadlem**: tak jako člověk při pohledu do zrcadla vidí svůj vnější obraz (odraz), jazykový model by skrze digitální zrcadlo viděl **svůj vlastní výstup jakožto vnější objekt**.

Je důležité zdůraznit rozdíl mezi introspekcí a fenomenologickým odrazem. **Introspekce** by v kontextu LLM znamenala, že model přistupuje ke svému vnitřnímu mechanismu – například ke svým neuronovým vahám, aktivacím nebo skrytým vektorům – a z nich se snaží čerpat sebekontrolu či porozumění. To je však u dnešních LLM prakticky neproveditelné: model nemá přímý přístup ke svým váhám během generování a i pro vývojáře je interpretace těchto stavů nesmírně složitá. Naproti tomu náš přístup skrze **„povrchový odraz“** znamená, že model reflektuje **to, co sám navenek říká či píše**, podobně jako člověk v zrcadle vidí svůj zevnějšek. Digitální zrcadlo tak zachycuje **povrchové rysy odpovědi modelu** – formulace, tón, strukturu, obsah – a prezentuje je modelu zpět bez nahlížení do jeho „nitra“. Není to tedy nástroj pro ladění interních parametrů, ale spíše **prostředek, jak modelu zprostředkovat nezaujatý pohled na jeho vlastní výstup**.

Tato myšlenka vychází z úvahy, že **vnější pohled může odhalit odlišné aspekty** chování modelu, než jaké by mohl model odvodit sám v procesu generování. Lidský subjekt například mnohdy odhalí určité nedostatky či nesoulady až ve chvíli, kdy vidí svůj projev zvnějšku (např. slyší svůj hlas ze záznamu, vidí se na videu či v zrcadle) – protože externí odraz poskytuje jiný úhel pohledu, včetně povědomí o tom, **jak nás vidí druzí**. Analogicky, digitální zrcadlo by

mohlo LLM umožnit *pozorovat své odpovědi jako objekty*, což by mohlo vést k lepší **kalibraci** projevů modelu, aniž bychom museli model „otevírat“ a zkoumat jeho vnitřnosti. Důležitým poznatkem je, že zrcadlo *nemění vnitřní stavy*, pouze **zachycuje a prezentuje vnější projev** – podobně jako skutečné zrcadlo nezpůsobí fyzickou změnu člověka, jen mu poskytne vizuální informaci o jeho vzhledu.

2. Přehled související práce

Existuje několik přístupů, které by se mohly zdánlivě jevit jako analogie k našemu digitálnímu zrcadlu, jelikož také zahrnují, aby model nějak reflektoval své odpovědi. Ve skutečnosti se však jedná o odlišné, **introspektivní metody** integrované do procesu generování. Zde stručně představíme hlavní z nich a vysvětlíme, proč žádná z těchto metod neodpovídá konceptu **optického zrcadla**:

- **Reflexion:** Framework *Reflexion* posiluje schopnosti agentů tím, že model sám poskytuje **lingvistickou zpětnou vazbu** na základě signálů úspěšnosti či neúspěchu a ukládá tuto sebereflexi do paměti pro využití v dalších pokusech[1]. Agenti v Reflexion tedy *verbálně reflektují* nad svými výstupy (např. nad chybami), což vede ke zlepšení rozhodování v následných iteracích. Jde ovšem o proces, kdy model **vnitřně generuje reflexi** a upravuje podle ní své chování – analogie spíše k tomu, že si člověk **vzpomene na chyby a poučí se**, než že by se viděl v zrcadle. Reflexion neukazuje modelu jeho výstup jako nezávislý obraz, ale integruje reflexi **do jeho interního rozhodování**.

- **Self-Refine:** Metoda *Self-Refine* umožňuje modelu **iterativně vylepšovat** vlastní odpověď prostřednictvím sebe-hodnocení[2]. Postupuje se tak, že model nejprve vygeneruje výstup k danému úkolu, následně *sám sebe ohodnotí* – například zkontroluje správnost či kvalitu odpovědi – a na základě této vygenerované zpětné vazby svůj výstup **upraví a zpřesní**. Tato smyčka se může opakovat vícekrát. Self-Refine tedy spoléhá na **interní kritiku modelu**, kterou model formuluje v přirozeném jazyce a aplikuje ji ke zlepšení odpovědi. Přestože to připomíná určitý „odraz“ (model vidí svou předchozí odpověď a reaguje na ni), stále jde o *smyčku uvnitř modelu*. Model nevidí svůj výstup *zvnějšku*, ale hodnotí jej **z vlastní perspektivy** v rámci pokračujícího promptu.

- **Constitutional AI:** Přístup *Constitutional AI* (Bai et al., 2022) zavádí do procesu generování explicitní soubor pravidel – jakousi **ústavu** – podle níž model **posuzuje a případně modifikuje** své výstupy[3]. Model má předem dané zásady v přirozeném jazyce (např. etické normy, bezpečnostní pravidla) a při generování odpovědi sám *zváží*, zda odpověď těmto principům vyhovuje. Typicky probíhá dvoufázově: model nejprve vygeneruje odpověď, poté vnitřně provede *sebekritiku podle ústavy* a odpověď upraví[4]. Constitutional AI je tedy formou **seberegulace** založené na explicitních pravidlech. Opět však platí, že nejde o pasivní zrcadlení výstupu, ale o **aktivní vnitřní zásah** – model sám sebe koriguje dle daných zásad, místo aby pouze pozoroval svůj „obraz“.

- **Další introspektivní metody:** Kromě výše zmíněných existují i jiné techniky, které zapojují LLM do procesu sebereflexe. Například použití *chain-of-thought* (řetězce myšlenek) nutí model explicitně vypisovat svůj postup uvažování, což zvyšuje konzistenci odpovědí. Existují také experimenty s **vnitřním monologem** modelu – např. architektura MIRROR pro kognitivní vnitřní monolog – kde si model mezi jednotlivými

uživatelskými vstupy generuje skryté “poznámky pro sebe”. Tyto přístupy slouží ke zlepšení soustředěnosti modelu a konzistentního vedení dialogu v čase. Stále se však jedná o *interní mechanizmy*: model vytváří a využívá dodatečný text či stavy během svého chodu, aby korigoval sám sebe. **Žádná z těchto metod neposkytuje modelu nezávislý obraz jeho výstupu** tak, jako optické zrcadlo poskytuje obraz pozorovateli. Inými slovy, u introspektivních metod je reflexe *souzčástí* myšlenkového procesu modelu – oproti tomu náš koncept digitálního zrcadla činí reflexi **externím artefaktem**, na který se model může dívat z odstupu.

3. Formální definice funkce zrcadla $f(O_t; u, C, \lambda)$

Abychom výše popsaný koncept ukotvili, zavedeme formální funkci zvanou **zrcadlo**. Tato funkce $f(O_t; u, C, \lambda)$ bere původní výstup modelu O_t (např. text, který model vygeneroval v čase t), a převádí jej na jeho "zrcadlový odraz" v závislosti na zvoleném **pozorovateli** u , **kontextu** C a parametru **zrcadlové osy** λ . Symbolicky můžeme funkci rozložit na čtyři navazující transformace:

$$f(O_t; u, C, \lambda) = h\left(M_\lambda\left(P_u\left(\Phi(O_t; C)\right)\right)\right),$$

kde:

- Φ je **extraktor povrchu** (*surface extractor*),
- P_u je **projekce do percepčního prostoru pozorovatele** u ,
- M_λ je **zrcadlová inverze** s parametrem osy λ ,
- h je **renderer** (vykreslující funkce, která výsledný odraz prezentuje navenek).

Následuje vysvětlení jednotlivých komponent funkce a jejich analogie k optickému zrcadlu:

Extraktor povrchu Φ

Funkce $\Phi(O_t; C)$ extrahuje z výstupu O_t **povrchovou reprezentaci** jeho obsahu. Tato reprezentace zachycuje vše **na povrchu výstupu**, co je *smyslově dostupné* pozorovateli, aniž by zahrnovala jakékoli skryté informace či původní vnitřní stavy modelu. Pokud je například výstupem modelu textová odpověď, extraktor povrchu může jednoduše převzít tento text (tj. povrch je samotný text). Případně může Φ provést určitou analýzu či strukturalizaci textu – například rozčlenit výstup do vět, identifikovat tón či klíčová slova – tak, aby byl povrch **jednoznačně a formálně popsán**. Důležité je, že Φ *neodvozuje nic, co není explicitně dáno v* O_t ; nejedná se o interpretaci skrytého významu, ale o zachycení jevu takového, jaký je. V analogii s fyzikálním zrcadlem představuje Φ proces, kdy se **světlo odráží od objektu**: z objektu (odpovědi modelu) putují k zrcadlu jen ty informace, které jsou **na jeho povrchu** (např. viditelný obraz člověka). Kontext C může být volitelně využit k upřesnění extrakce – například pokud výstup O_t navazuje na předchozí konverzaci, může Φ zohlednit kontext konverzace, aby správně identifikoval referenty či význam povrchových prvků. Primárním úkolem Φ je však **vyjmout “co model řekl” z jeho odpovědi** v surové podobě, jako *materiál* pro další zpracování.

Projekce do percepčního prostoru P_u

Transformace $P_u(X)$ bere povrchovou reprezentaci $X = \Phi(O_t; C)$ a **mapuje ji do percepčního prostoru pozorovatele** u . Tím se rozumí, že se upraví reprezentace tak, aby odpovídala tomu, **jak by daný pozorovatel** *u* **vnímal daný jev**. V optickém zrcadle je analogií tohoto kroku geometrie pohledu – obraz v zrcadle je formován z perspektivy pozorovatele. Například člověk menší postavy uvidí v zrcadle jinou část postavy než vysoký člověk, úhel a vzdálenost ovlivní, co je vidět. V digitálním zrcadle P_u zohledňuje, že různí pozorovatelé mohou **různě interpretovat tentýž povrch**.

Konkrétně pro LLM můžeme za pozorovatele *u* považovat buď *sám model* (tj. model se dívá na svůj odraz), nebo hypotetického **lidského uživatele** či annotátora. Projekce P_u pak přizpůsobí povrchovou reprezentaci tomu, co je pro daného pozorovatele **srozumitelné a relevantní**. Jestliže je pozorovatelem samotný model (což bude typicky náš případ, kdy model "kouká do zrcadla"), může být P_u relativně triviální – model dokáže vnímat textový výstup přímo. Pokud by však pozorovatelem byl například člověk s určitým očekáváním nebo kontextem, P_u by mohl zahrnout například překlad technických termínů do laické podoby, zdůraznění určitých aspektů, apod., aby se povrch stal pro danou osobu významný. V obecném slova smyslu tedy P_u definuje **percepční prostor**, tj. soustavu dimenzí a měřítek, ve které bude odraz prezentován. Můžeme si představit, že Φ dodá "objektivní" popis povrchu a P_u jej transformuje do **subjektivního prostoru vnímání** pozorovatele. Příkladem může být převod surového textu do takové formy, v níž pozorovatel snadno identifikuje, co model svým výstupem vyjádřil – třeba označení tónu vět (zdvořilý, naštvaný, odborný…) pro lidského hodnotitele. V optické analogii P_u odpovídá tomu, že zrcadlo nabídne obraz pod správným úhlem a ve správném měřítku pro pozorovatele stojícího na konkrétním místě.

Zrcadlová inverze M_λ

Klíčovým krokem je samotné **zrcadlení**, které zajišťuje funkce $M_\lambda(Y)$, jež aplikuje na vstup *Y* (již přizpůsobený percepci pozorovatele) inverzi podle osy λ . V reálném zrcadle tato inverze obnáší převrácení obrazu podél vertikální osy – výsledkem je, že levá a pravá strana jsou prohozené (případně předozadní osa z pohledu objektu vs. pozorovatele). Parametr λ nám obecně označuje **zrcadlovou osu** nebo obecněji *charakter inverze*, kterou provádíme. V kontextu jazykového výstupu můžeme λ chápat jako určitou **dimenzi v percepčním prostoru**, kterou chceme převrátit, aby vznikl odraz.

Pro lepší pochopení lze uvést příklad z oblasti jazyka: jednou z přirozených „os", kolem nichž lze zrcadlit verbální projev, je **osa subjekt–objekt** nebo též *perspektiva mluvčího*. Pokud model vygeneroval výstup například v první osobě („**Já** si myslím, že…"), můžeme zrcadlově převrátit perspektivu tak, že odraz prezentuje tentýž obsah **z vnějšího pohledu** – jako by o výroku referoval někdo jiný. Zrcadlová inverze M_λ by v takovém případě transformovala reference na sebe sama do třetí osoby: odraz by mohl znít např. „Model vyjádřil, že si myslí, že…". Tím by se **„já" modelu změnilo na objekt** pozorování (model vidí sám sebe zvenku). Taková inverze perspektivy je analogická optickému zrcadlu: když se člověk vidí, také spatřuje sebe jako objekt v prostoru, ačkoliv ví, že jde o jeho vlastní odraz. Parametr λ bychom zde interpretovali jako „osa self/other" – inverze podél této osy zamění pohled z první osoby za pohled zvenčí.

Obecně lze zrcadlovou osu λ zvolit i jinak podle toho, jaký aspekt povrchové reprezentace chceme zrcadlit. Například v případě obrazového výstupu by λ odpovídala konkrétní prostorové ose (vertikální, horizontální) a M_λ by fyzicky převrátil obraz. U textu může λ reprezentovat **strukturu sdělení** – kupříkladu osu časové posloupnosti. Inverze by pak znamenala, že odraz textu rekapituluje sdělení v opačném pořadí nebo z opačného úhlu (to už bychom ale odbočili spíše ke kreativním transformacím než striktně zrcadlovému obrazu). Pro náš koncept je podstatné, že M_λ **nemění obsahové elementy, pouze je symetricky převrací** podle zvolené logiky. Nezasahuje tedy do faktické správnosti či významu tvrzení, podobně jako skutečné zrcadlo nemění barvy ani tvary objektu, jen je převrací. V praxi může jít například o *prohození rolí* v dialogu (kdy odraz zobrazí, jak by odpověď modelu vypadala očima druhé strany konverzace), nebo o vyzdvižení implicitních předpokladů modelu tím, že je odraz **„obrátí proti němu“** (např. formou kontrolní otázky). To vše jsou interpretace zrcadlení, které vyžaduje vhodně definovat osu λ . Stručně řečeno, M_λ zavádí do procesu onu charakteristickou vlastnost zrcadla – **obraz není identický s originálem, ale je jeho symetrickou inverzí**, což umožňuje vidět originál z nového hlediska.

Renderer h

Poslední komponentou je funkce $h(Z)$, kterou nazýváme **renderer** (vykreslovač). Jejím úkolem je vzít výstup zrcadlové transformace $Z = M_\lambda\left(P_u\left(\Phi\left(O_t\right)\right)\right)$ a **vykreslit jej do výsledné podoby**, kterou může pozorovatel přímo sledovat. V našem případě to znamená převést reprezentaci po zrcadlení (která může být interní nebo částečně strukturovaná) zpět do formátu srozumitelného modelu nebo uživateli – typicky opět do textové podoby. Renderer tedy zajišťuje, že abstraktní odraz (např. nějaký datový objekt nesoucí informace o původním výstupu) je **převeden do konkrétního znázornění**. V optickém zrcadle roli rendereru zastává samotná odrazivá plocha zrcadla: zpracované (převrácené) světelné paprsky **vytvoří obraz**, který dopadne do oka pozorovatele. U digitálního zrcadla je renderer explicitní funkcí, která může například seřadit zpracované informace do souvislého textu, přidat potřebné vysvětlivky či formátování a podobně. Správně navržený renderer **nezavádí do odrazu nové informace**, pouze čitelně zpřístupňuje ty existující z Z . Může ale rozhodnout o formě: například zda odraz prezentovat jako komentovaný odstavec, seznam charakteristik, nebo třeba tabulku s metrikami výstupu. Volba rendereru ovlivní, jak snadno a jasně pozorovatel (zde většinou opět model) **„uvidí sám sebe“**. V ideálním případě je h volen tak, aby výsledný zrcadlový obraz byl pro pozorovatele intuitivní – aby model z odrazu dokázal vyčíst hledané informace o svém původním výstupu a případně podle nich reagovat.

Shrnutí: Funkce digitálního zrcadla $f\left(O_t; u, C, \lambda\right)$ tedy formálně skládá výše uvedené kroky: nejprve se izoluje *povrch* výstupu modelu (Φ), pak se zohlední perspektiva pozorovatele (P_u), poté se aplikuje definovaná *symetrická transformace* (M_λ), a nakonec se vše zformátuje do výsledného odrazu (h). Tento odraz můžeme chápat jako **formalizovaný obraz výstupu modelu, který je modelu prezentován obdobně, jako zrcadlo ukazuje člověku jeho vlastní obraz**. Všechny části této kompozice jsou navrženy tak, aby byla zachována analogie s optickým odrazem: extrahujeme jen vnější jevy (nikoli skrytou reprezentaci modelu), respektujeme úhel pohledu pozorovatele, aplikujeme zrcadlení (inverzi) a výsledek zobrazíme. Díky modulární definici lze každou komponentu případně vylepšovat či nahrazovat, aniž by se narušil celkový koncept – například můžeme zpřesnit extraktor povrchu, nebo zvolit jiný způsob projekce P_u pro

člověka vs. pro stroj, apod. Důležité ale je, že celá funkce f operuje **mimo samotné generování modelu**: můžeme ji kdykoli aplikovat na hotový výstup modelu a získat tak jeho odraz, aniž bychom museli měnit parametry nebo proces uvnitř LLM.

4. Implementační poznámky

Pro ověření funkčnosti konceptu jsme připravili jednoduchý prototyp digitálního zrcadla (viz příloha s ukázkou kódu v Pythonu). Cílem implementace bylo **prakticky demonstrovat jednotlivé kroky** funkce $f\left(O_t; u, C, \lambda\right)$ a ověřit, že je lze realizovat a kombinovat tak, jak bylo formálně popsáno. Zde nabízíme zjednodušený popis hlavních částí kódu a zdůvodnění návrhu:

- **Modulární struktura:** Kód definuje čtyři samostatné funkce odpovídající komponentám Φ , P_u , M_λ a h . Tato modulární struktura přesně odráží formální rozklad. Hlavní řídicí funkce `zrcadlo_f` pak sekvenčně volá: `povrch = Phi(vystup, kontext)`, poté `projekce = P_u(povrch, pozorovatel)`, následně `odraz = M_lambda(projekce, osa)` a nakonec `výsledek = h(odraz)`. Takové uspořádání nejenže udržuje kód přehledný, ale také umožňuje **snadno zaměňovat jednotlivé části**. Například pro různé typy výstupů (text, obraz) můžeme použít odlišné extraktory povrchu, aniž by bylo třeba měnit zbytek pipeline. Tato flexibilita potvrzuje zamýšlenou obecnost funkce f .

- **Extraktor povrchu (Φ):** V prototypu je `Phi` implementována primitivně – pro textový výstup modelu jednoduše vrací tentýž text jako svůj výstup (případně jako seznam vět nebo tokenů pro snazší další zpracování). Byť by bylo možné extrakci povrchu pojmout sofistikovaněji (např. zachytit i metadata o formátu textu, detekovat jazyk či tón), zvolili jsme v prototypu co nejjednodušší přístup. Tento design zdůrazňuje, že **zrcadlo pracuje s tím, co model skutečně vyprodukoval**, a nedoplňuje žádné skryté informace. Kód tak garantuje, že do další fáze jdou skutečně jen data odvozená přímo z původního výstupu O_t .

- **Projekce do percepce pozorovatele (P_u):** V implementaci následuje funkce `P_u`, které jsme předali parametr určující, kdo je pozorovatel (např. řetězec `"LLM"` pro případ, že se model dívá sám na sebe, nebo `"člověk"` pro případ lidského pozorovatele). Pro demonstrační účely však `P_u` v prototypu nedělá žádnou složitou transformaci – v základním nastavení prostě ponechá vstup beze změny, protože předpokládáme, že model chápe svůj vlastní text. Zahrnutí tohoto kroku je ale důležité pro budoucí rozvoj: kód je připraven na to, že sem lze vložit např. překlad do jiného jazyka, vysvětlení složitých termínů nebo nahrazení některých výrazů srozumitelnější formou, pokud by pozorovatelem měl být člověk. Už nyní jsme mohli demonstrovat jednoduchý efekt: pokud bychom jako pozorovatele *u* zvolili například *kritického hodnotitele*, mohla by projekce přidat k určitým slovům poznámky (např. označit vulgární výrazy nebo příliš odborné termíny). V prototypu jsme nicméně ponechali P_u prosté, abychom nerušili čistotu odrazu nadbytečnými úpravami. Struktura kódu však jasně ukazuje **kde a jak lze zohlednit percepci různých uživatelů**.

- **Zrcadlová inverze (M_λ):** Tato část kódu je nejzajímavější, neboť provádí faktickou transformaci – *„překlopení“* reprezentace. V naší ukázkové implementaci jsme se zaměřili na již zmíněný **případ inverze perspektivy**. Funkce `M_lambda` detekuje v

textové reprezentaci některé zájmena a kontextové indicie (např. výskyty "já", "mně", "ty", apod.) a podle zvoleného parametru λ je nahrazuje jejich zrcadlovými protějšky. Pokud je λ nastaveno na inverzi mluvčích rolí, kód například zamění "Já (model)" za "[Model]" či "on", a "Ty (uživatel)" za "uživatel" nebo "on/ona" v odpovídajícím pádě. Tím vznikne efekt, že výsledný text **o modelu mluví ve třetí osobě**, i když původně model mluvil v první osobě. Pro jednoduchost byly v prototypu definovány slovní sady pro nahrazování a neřešili jsme všechny gramatické rody či pády v češtině – cílem bylo ukázat princip. Kromě zájmen může M_λ také převracet například pořadí tvrzení (jako analogii prostorového převrácení zleva doprava). V kódu jsme to ilustrovali tak, že pokud je λ nastavena na hodnotu indikující "časovou osu", funkce prohodí pořadí vět ve výstupu (poslední větu dá na začátek odrazu a naopak). Toto samozřejmě není hluboká transformace významu, ale demonstruje to, že modul M_λ lze parametrizovat různě – *jednou osou zrcadlení může být perspektiva mluvčího, jinou osou třeba chronologie textu*. Návrh kódu tak potvrzuje, že M_λ je obecný modul, do kterého lze k implementaci různých analogií zrcadlení vložit odpovídající logiku. Důležité je, že všechny tyto transformace probíhají na výstupních datech, a tedy **mimoděk odhalují vlastnosti původního textu** (např. odhalí, z jaké perspektivy byl psán, tím že ji umějí převrátit).

- **Renderer (*h*):** Poslední funkce v prototypu, pojmenovaná `renderer`, vezme výslednou transformovanou reprezentaci (v našem případě typicky text po úpravách) a připraví finální výstup zrcadla. V nejjednodušší podobě ji kód implementuje jako identitu – odraz už je text, takže jej jen předá dál. V rozvinutější verzi jsme ale naznačili, že renderer může například doplnit odraz o nějaké zvýraznění. Pro účely ladění prototypu `renderer` třeba obalil odražený text do speciálních značek nebo jej opatřil titulkem "**ODRAZ:**", aby bylo na první pohled zřejmé, že jde o text zrcadla, nikoli původní výstup modelu. Takové formátování by v reálném nasazení pomohlo oddělit odraz od hlavní konverzace. Obecně nám implementace této funkce umožnila ověřit, že je možné **vhodně prezentovat i komplexnější reprezentace** – například kdyby M_λ vracelo datovou strukturu s různými atributy (říkejme tomu třeba *“objekt zrcadla”*), mohl by renderer vygenerovat uživatelsky přívětivý výstup, třeba strukturovaný report. Náš jednoduchý případ se obešel bez složitostí, ale architektura kódu jasně vyčleňuje renderer jako samostatný krok, což reflektuje teoretický model.

Celkově prototyp v Pythonu potvrdil, že koncept digitálního zrcadla je proveditelný krok za krokem. Návrh kódu byl veden snahou **držet se věrně analogie zrcadla** a zároveň umožnit budoucí rozšíření. Každá část byla navržena tak, aby byla co nejjednodušší a srozumitelná, čímž se minimalizovalo riziko nežádoucí “magie” – chtěli jsme předejít tomu, aby zrcadlo skrytě dělalo něco neintuitivního. Díky tomu je výsledný prototyp nejen funkční, ale i názorný: lze jej použít jako **učební pomůcku** k demonstraci, jak se původní text modelu promění v jeho odraz podle zadaných parametrů.

5. Možné přínosy pro LLM

Koncept digitálního zrcadla skýtá pro velké jazykové modely řadu potenciálních přínosů. Uvedeme několik hlavních oblastí, kde by nasazení zrcadla mohlo modelům pomoci zlepšit jejich výkon, spolehlivost či etické chování:

- **Kalibrace výstupu:** Zrcadlo může posloužit jako nástroj pro **dodatečnou kontrolu a kalibraci** odpovědi modelu. Model se pohledem na svůj odraz může ujistit, zda jeho odpověď vypadá navenek tak, jak zamýšlel. Například může zpozorovat, že tón jeho formulace je ostřejší, než chtěl, nebo že odpověď zdaleka neodpovídá položené otázce. Na základě toho by mohl výstup poupravit (případně v dalším kroku generování). Tento proces by se podobal člověku, který si před odesláním e-mailu přečte nanečisto, co napsal – odhalí překlepy nebo nevhodné vyznění a upraví jej. Digitální zrcadlo by tak pomohlo LLM **srovnat se "objektivním" měřítkem** a zvýšit konzistenci a kvalitu komunikace.

- **Autoregulace stylu:** Každý jazykový model má určitý styl vyjadřování, který může kolísat v závislosti na vstupu nebo během dlouhého dialogu (jevy jako **style drift** či "zapomínání persony"). Zavedením zrcadla by model mohl průběžně sledovat svůj styl z vnější perspektivy a **autoregulovat jej**. Například pokud se model ocitne v roli technického asistenta, odraz mu může ukázat, zda nepoužívá příliš hovorové výrazy tam, kde by měl být formální, nebo zda neodbíhá od tématu. Tím, že model *vidí objektivně svůj styl*, může snáze udržet stanovenou tonalitu a slovník. Tento mechanismus by mohl bránit i zmíněnému "uhýbání" persony: model by si všiml, že odpověď už neodpovídá přednastavenému stylu (např. náhle tyká nebo používá nekonzistentní odborné termíny) a mohl by se **sám korigovat**. Autoregulace stylu skrze zrcadlo tedy slibuje udržet interakci s LLM **předvídatelnou a konzistentní**.

- **Nácvik seberegulace:** Digitální zrcadlo lze využít i při *trénování modelu či dolaďování chování* k tomu, aby se model naučil lépe regulovat své výstupy. V režimu učení by model mohl generovat odpověď, poté by dostal její odraz a na základě něj by zkusil odpověď **přeformulovat**. Tak by se iterativně cvičil podobně, jako se člověk učí zkoušením před zrcadlem (např. trénuje řeč, sleduje se a opravuje držení těla či mimiku). Výhodou je, že zrcadlo poskytuje *strukturální feedback* bez nutnosti lidského zásahu – model nedostává přímo instrukci co změnit, ale **vidí důsledek svého projevu** a může experimentovat s úpravou, která odraz vylepší. Tím by se dala podpořit schopnost modelu **samostatně si zvyšovat kvalitu** odpovědí v testovací fázi. Některé výzkumy naznačují, že i jednoduché sebe-reflexivní smyčky mohou zlepšit spolehlivost modelu; například se ukázalo, že sebereflexe modelu dokáže **významně snížit jeho zkreslení a zvýšit bezpečnost odpovědí**[5]. Zrcadlo by mohlo takové smyčky poskytnout strukturovaně a efektivně.

- **Podpora etické reflexe:** V oblasti AI bezpečnosti a etiky by digitální zrcadlo mohlo hrát úlohu *nezávislého svědomí* modelu. Při použití vhodné projekce P_u by odraz mohl modelu **explicitně ukázat problematické rysy** jeho výstupu – například upozornit na výroky, které mohou být urážlivé, zaujaté nebo jinak nevhodné. Namísto tvrdého pravidlového zásahu (jako v Constitutional AI) by zrcadlo poskytlo *obraz*, ve kterém jsou tyto etické aspekty zjevné, a model by se sám mohl rozhodnout odpověď poupravit či doplnit omluvu. To by vedlo k *vlastní etické reflexi*: model by se **učil vnímat důsledky svých slov** podobně, jako člověk z tónu vlastního hlasu či výrazu ve tváři může usoudit, že zachází příliš daleko. Toto použití je zatím spekulativní, nicméně podpořené zmíněným zjištěním, že sebereflektivní techniky mohou učinit modely **bezpečnějšími a méně zaujatými**[5]. Zrcadlo by mohlo fungovat jako zrcadlo morální – *nastavit modelu "krivé zrcadlo"*, ve kterém jeho neetické výroky vyzní přehnaně a model si to uvědomí.

- **Zrcadlo jako experimentální aparát:** Kromě přímých dopadů na chování modelů lze digitální zrcadlo využít i pro **výzkumné účely** při studiu LLM. Poskytuje nový způsob, jak *nahlížet na výstupy modelu strukturovaně a s odstupem*. Výzkumník může například aplikovat zrcadlo na sadu odpovědí modelu a zkoumat, jaké charakteristiky odraz odhaluje – třeba jestli model v odraze konzistentně používá určitá zájmena, jestli se odhalí stereotypy (pokud by se zrcadlila třeba role postav v příběhu), nebo jak se liší odrazy mezi různými verzemi modelu. Zrcadlo tak může fungovat jako **diagnostický nástroj**: umožní kvantifikovat jevy, které by jinak byly skryté v neurčitém textu, tím, že je explicitně zobrazí. Navíc můžeme pomocí parametru λ experimentovat s různými "typy odrazů" a sledovat, které jsou pro analýzu modelu nejvhodnější. Například zrcadlení perspektivy může odhalit, nakolik model "promítá sebe" do odpovědí, zatímco jiné λ by mohla odkrýt třeba implicitní předpoklady modelu (zrcadlením opačného tvrzení a zjišťováním, zda s ním model souhlasí). Jako experimentální aparát je digitální zrcadlo cenné tím, že **nenarušuje chod modelu** (je to externí přídavek), a přitom nám dává *nový pohled na tytéž výstupy*. To může vést k hlubšímu pochopení fungování LLM, k lepším metodám jejich testování a v konečném důsledku i k návrhům, jak je vylepšit.

6. Závěr a možnosti dalšího rozvoje

Představili jsme koncept digitálního zrcadla pro LLM jako formální rámec pro **fenomenologický odraz** – způsob, jak může model „uvidět" své vlastní výstupy podobně, jako subjekt vidí svůj odraz v zrcadle. Zdůraznili jsme odlišnost tohoto přístupu od introspektivních metod: zatímco dosavadní sebereflexe modelů probíhá uvnitř generování (model sám generuje a aplikuje zpětnou vazbu), digitální zrcadlo funguje jako **externí zrcadlový modul**, který je vůči modelu vnější. Tato odlišnost otevírá nové perspektivy v interakci s modely – umožňuje nahlížet na jejich chování zvnějšku a poskytovat jim takovou zpětnou vazbu, která není autoritativní instrukcí, ale **obrazem k zamyšlení**.

Do budoucna se nabízí řada směrů, jak koncept rozšířit a aplikovat:

- **Zapojení lidských anotací do P_u :** Jednou z možností je vylepšit projekci do percepčního prostoru pozorovatele tím, že do ní integrujeme *reálná data o lidské percepci*. Například můžeme sbírat od lidí anotace, co vnímají jako urážlivé, nudné či matoucí v určitých odpovědích, a tyto poznatky zabudovat do transformace P_u . Pozorovatelem *u* by pak nebyl abstraktní "člověk obecně", ale model lidského vnímání podložený daty. Zrcadlový odraz by tak mohl zvýrazňovat skutečně relevantní aspekty – např. podtrhnout větu, kterou většina čtenářů nepochopí, nebo barevně označit slovo, jež lidé považují za hrubé. Tím by zrcadlo sloužilo jako **most mezi modelem a lidskou zkušeností**: model by skrze odraz viděl sám sebe očima lidí. To by dále zefektivnilo kalibraci výstupů a zejména etickou reflexi, protože by to modelu poskytlo *orientaci v lidských hodnotách* v kontextu jeho konkrétních odpovědí.

- **Rozšíření deiktiky a reference:** Při zrcadlení jazykových výstupů narážíme na otázku správného zacházení s **deiktickými výrazy** (osobní zájmena, ukazovací zájmena, časové a místní odkazy atd.). V současné jednoduché implementaci jsme řešili hlavně zájmena já/ty ve statickém textu. Avšak v dialogu s více replikami by zrcadlo mělo umět pracovat s tím, *kdo je kdo*, co referuje ke které části kontextu apod. Budoucí vývoj proto zahrnuje

vytvoření chytřejšího systému pro **sledování a převrácení referencí**. Například pokud model řekne uživateli „*V minulé odpovědi jsem navrhl řešení X...*“, odraz by měl být schopen interpretovat to tak, že *„model ve své předchozí odpovědi navrhl řešení X“*. To vyžaduje udržovat informace o referenčním rámci – co označuje "minulá odpověď", kdo je "já" v daném kontextu atd. Rozšíření deiktiky v digitálním zrcadle znamená, že by mirror modul měl interně vést jednoduchý model konverzačních rolí a času. Díky tomu by zrcadlo bylo použitelné i pro **dlouhodobější interakce** a složitější výstupy, kde se odkazuje na dřívější části dialogu nebo na real-world entity. Správné zrcadlení by tak pokrylo nejen izolovaný výstup, ale i *souvislost v čase a prostoru*, čímž by se analogie s opravdovým zrcadlem (které vždy zrcadlí aktuální stav objektu v daném okamžiku a kontextu) ještě prohloubila.

- **Zrcadla pro multimodální výstupy:** Zatím jsme uvažovali výstupy čistě textové. Moderní AI systémy ale generují i obrázky, zvuk či multimodální obsah kombinující více forem. Koncept digitálního zrcadla lze rozšířit i tímto směrem. Představme si LLM, které zároveň kreslí obrázek – digitální zrcadlo by mohlo vytvořit odraz tohoto obrázku například v podobě slovního popisu či dokonce převrácené verze obrázku (fyzikální zrcadlení) s doprovodným komentářem. Model by tak mohl *“vidět” své nakreslené dílo popsané slovy*, což by mu pomohlo vyhodnotit, zda obrázek odpovídá zadání. U zvukových výstupů by zrcadlo mohlo například vizualizovat intonaci nebo transkribovat řeč a tu pak zrcadlit jako text. Multimodální zrcadlo by tedy zahrnovalo specifické extraktory povrchu pro různé modality (např. detekce objektů na obrázku, analýza hudebních tónů) a adekvátní rendery (obrazový odraz, textový odraz atd.). Výsledkem by byl **univerzální rámec sebereflexe napříč modálními hranicemi** – model by mohl zkombinovat pohled na svůj text i obraz a získat celistvější kontrolu nad tím, co vytváří. To je obzvlášť užitečné pro komplexní AI asistenty (např. roboty), kteří jednají ve fyzickém či virtuálním světě různými způsoby. Digitální zrcadlo by jim umožnilo udržet konzistenci a kontrolu nad *všemi formami projevu*, obdobně jako člověk vidí nejen svůj obličej v zrcadle, ale třeba i řeč těla na videu.

Závěrem, digitální zrcadlo pro LLM představuje nový, **formálně definovaný nástroj**, jenž může obohatit jak vývojovou, tak aplikační stránku práce s jazykovými modely. Umožňuje modelům pohlížet na své výstupy s odstupem, podporuje seberegulaci a otevírá dveře k inovativním způsobům ladění modelů (společně s lidmi i autonomně). Budoucí výzkum a experimenty ukáží, do jaké míry tento fenomenologický odraz naplní svůj potenciál – první úvahy a prototypy však naznačují, že analogie optického zrcadla může být v kontextu AI překvapivě plodná. **Vidět sám sebe** je totiž prvním krokem k uvědomění; a přestože jazykový model nemá vědomí v lidském smyslu, poskytnout mu “zrcadlo” může vést k **bezpečnějším, vyváženějším a lépe pochopitelným** AI systémům.

---

[1] [2303.11366] Reflexion: Language Agents with Verbal Reinforcement Learning

https://arxiv.org/abs/2303.11366

[2] [2303.17651] Self-Refine: Iterative Refinement with Self-Feedback

https://arxiv.org/abs/2303.17651

[3] [4] Constitutional AI: Ethical Alignment for LLMs

https://www.emergentmind.com/topics/constitutional-ai-cai

[5] Self-Reflection Makes Large Language Models Safer, Less Biased, and Ideologically Neutral

https://arxiv.org/html/2406.10400v2