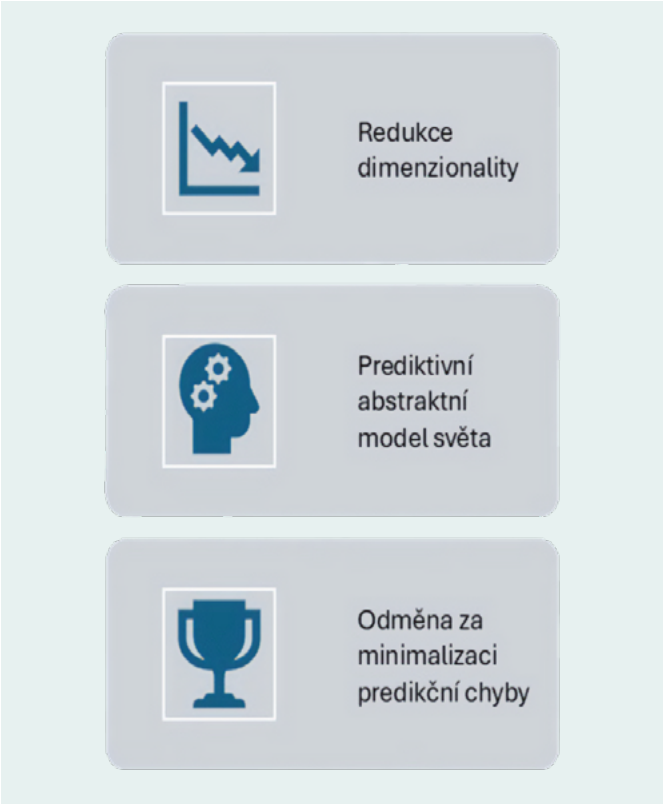
Obr. 1. *Pilíře inteligence a učení*



mu strojům slouží mnoho technik. Za jednou z nich stojí geniální nápad českého vědce Tomáše Mikolova, a to převést slova na vektory charakterizované tisíci souřadnicemi v multidimenzionálním kyberprostoru (3). Vektory zachycují sémantické vztahy mezi slovy, takže podobná slova se nacházejí blízko sebe ve vektorovém prostoru, to poskytuje strojům kontext a strukturu textových dat a tím – jednoduše řečeno – chápat význam. Obrazová data jsou zase pomocí konvoluce komprimována na takzvané mapy rysů (například rys definující hranu), jež pak vystihují základní esenci obrazu. Cílem je u všech typů dat získat kompaktní reprezentaci původních informací pro jejich efektivnější zpracování.

Lidský mozek zpracovává multidimenzionální senzorické vstupy pomocí komplexních neuronových sítí, které využívají kombinaci heuristik, empirických zkušeností a intuice – tedy schopnosti rozpoznat podstatné rysy bez vědomé analýzy. Tento proces probíhá v rámci prediktivního abstraktního modelu okolního světa, kde vyšší kognitivní úrovně generují abstraktní předpovědi, které jsou porovnávány s reálnými daty senzorických vstupů (4).

Mozek zároveň interpretuje výsledky těchto procesů prostřednictvím přirozeného jazyka – univerzálního nástroje k reprezentaci a sdílení vnitřních modelů. Právě přirozený jazyk, převedený do vektorových reprezentací, se ukázal jako mimořádně efektivní pro architekturu umělé inteligence.

Náš mozek se snaží minimalizovat predikční chybu, což představuje adaptační mechanismus, jehož neurofyziologickým korelátem je dopaminová signalizace (dopamin jako signál „lepšího než očekávaného"). To je základem posilovaného učení (RL – reinforcement learning). V analogii s lidským učením byl v posilovaném učení strojů nahrazen dopamin signálem odměny, případně lidskou zpětnou vazbou, což umožňuje během trénovaní těchto modelů přizpůsobit výstupy AI k praktickým a lidsky relevantním cílům.

Technické řešení architektury moderních generativních jazykových modelů označujeme jako transformer. Ten slouží k analýze a generování sekvencí, jako je text, dnes už i obraz, video nebo zvuk. Během tréninku na gigantických datových souborech transformer extrahuje jejich esenci, vytvoří si abstraktní model světa či latentní (pro nás skrytou) reprezentaci světa – jakousi prediktivní pravděpodobnostní mapu souvislostí, statistické vzorce a vztahy korigované lidskou zpětnou vazbou. Takto se model nastaví. Zde je dobře patrný rozdíl, kdy v lidském mozku tento proces běží neustále, naopak jazykový model strojů je takto nastaven při trénování jednou. Ale i to se už daří korigovat například napojením na internetové vyhledávače, případně novým přetrénování celého transformeru po určitém čase. Pak přichází již naše interakce s tímto naučeným jazykovým modelem. Pro názornost dále zmiňuji generující proces textu, který je pro primární péči relevantnější než obraz či jiná data. Namísto trénovacích vstupních dat do nastaveného naučeného modelu vstoupí naše nová informace, výzva, otázka – tzv. PROMPT, který spustí interní generativní proces. Ve skutečnosti model jen napodobuje jazykové a logické vzorce, které se naučil. Výsledkem je v případě textového výstupu předpověď následujícího slova, které navazuje na prompt. Toto slovo je vybráno z vypočtené škály pravděpodobností všech slov metodou tzv. samplingu, kdy každý výrobce různou mírou určuje, zda bude výstup spíše kreativní, nebo přesný deterministický. Na původní prompt se tedy napojí generované slovo, tento celý text je pak novým vstupem a celý proces generuje další slovo a tak dále, dokud není generovaná smysluplná informace. Výsledkem je komunikace, která se postupně přibližuje úrovni mezilidské interakce. Kritické myšlení je zde na místě, neboť generovaná data nejsou faktická tvrzení, ale spíše pravděpodobnostní odhad a takto je k nim potřeba přistupovat. Časté situace, kdy model generuje výstupy, které znějí přesvědčivě, ale jsou fakticky nesprávné, smyšlené nebo zavádějící, označujeme jako halucinace jazykových modelů. Ty jsou jádrem kritiky a obav ze zavádění těchto technologií v medicíně, kde i drobná nepřesnost může stát život. Kvalita našeho promptu zásadně určuje kvalitu odpovědi modelu. Zatímco lidé obvykle odpovídají přesně na jednoduché otázky, ale tápou u složitějších, jazykové modely typu transformer často fungují opačně – čím promyšlenější a strukturovanější je vstup, tím kvalitnější bývá odpověď. Odborně specifická terminologie umožňuje transformeru aktivovat vektorové reprezentace blízké relevantním zdrojům, čímž se zvyšuje přesnost a validita generovaných odpovědí. Naopak na neurčitý nebo banální dotaz mohou reagovat nepřesně či halucinací. Měli bychom také mít na paměti, že modely se učily na datech, jejichž validitu neznáme.

To, jakou základní roli má jazykový model v našem rozhovoru zaujmout, je dáno tzv. systémový promptem, vstupní instrukcí, která není běžnému uživateli viditelná, ale zásadně ovlivňuje chování modelu. Může například obsahovat instrukce typu: „jsi asistent, odpovídej srozumitelně, zdvořile, buď nápomocný, eticky vyvážený, odpovídej maximálně v určitých počtech znaků" a podobně. Tento úvodní kontext rámuje celou interakci – podobně jako když lékař ví, že právě mluví s pacientem, kolegou nebo dítětem, a podle toho