VYSOKÁ ŠKOLA POLYTECHNICKÁ JIHLAVA

Aplikovaná informatika

IDENTIFIKACE TEXTŮ VYTVOŘENÝCH UMĚLOU INTELIGENCÍ

Bakalářská práce

Autor práce: Tomáš Přibyl

Vedoucí práce: doc. Dr. Ing. Jan Voráček, CSc.

Jihlava 2026

Vysoká škola polytechnická Jihlava

Tolstého 16, 586 01 Jihlava

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Autor práce:	**Tomáš Přibyl**
Studijní program:	Aplikovaná informatika
Obor:	Aplikovaná informatika
Garant studijního programu:	Ing. Lenka Kuklišová Pavelková, Ph.D.
Název práce:	**Identifikace textů vytvořených umělou inteligencí**
Vedoucí práce:	doc. Dr. Ing. Jan Voráček, CSc.
Cíl práce:	Obecným cílem práce je ověření funkčnosti vybraného systému pro identifikaci textů, vytvořených umělou inteligencí. Naplněn bude prostřednictvím následujících konkrétních cílů: (1) Rešerše současného stavu poznání v oblasti strojové klasifikace textů s důrazem na problematiku identifikace dokumentů, generovaných prostřednictvím umělé inteligence. (2) Komparativní analýza volně dostupných nástrojů z této oblasti. (3) Výběr platformy, vhodné pro realizaci praktické části této práce. Návrh, vyhodnocení a diskuse výsledků rozpoznávacích experimentů. (4) Zobecnění získaných poznatků a formulace doporučení pro další výzkum.

Abstrakt

Bakalářská práce se zabývá možnostmi detekce textů vytvořených generativní umělou inteligencí v prostředí vysokých škol. Cílem je zhodnotit nástroj GPTZero při rozlišování lidských textů, plně AI generovaných textů (AI-GEN) a textů vzniklých post-editací AI (AI-EDIT) ve třech doménách: administrativní/formální, akademické/vědecké a literární/esejistické. Teoretická část shrnuje principy jazykových modelů a detekčních metod, včetně perplexity, burstiness a stylometrických rysů, a věnuje se etickým otázkám akademické integrity. Praktická část popisuje konstrukci vyváženého datasetu 180 textů, jejich testování v režimu advanced scan a vyhodnocení pomocí klasifikačních metrik (accuracy, precision, recall, F1), které vycházejí z binárního rozdělení textů na lidské a texty s podílem AI. Celkovou přesnost GPTZero dosáhlo přibližně 78 %, precision 99 % a recall 68 %; nástroj zachytil přibližně 78 % administrativních, 68 % akademických a 58 % literárních textů s podílem AI. Detekce je doménově citlivá a u post-editovaných textů obzvlášť nespolehlivá.

Klíčová slova

akademická integrita; detekce AI-generovaných textů; generativní umělá inteligence; GPTZero; velké jazykové modely

Abstract

The bachelor's thesis examines the detection of texts produced by generative artificial intelligence in higher education. Its aim is to evaluate GPTZero in distinguishing human-written texts, fully AI-generated texts (AI-GEN) and texts created by post-editing AI output (AI-EDIT) across three domains: administrative/formal, academic/scientific and literary/essayistic. The theoretical part presents language models and detection methods, including perplexity, burstiness and stylometric features, and discusses ethical issues of academic integrity. The practical part describes the construction of a balanced data set of 180 texts, their testing using the advanced scan mode and evaluation using standard classification metrics (accuracy, precision, recall and F1), derived from a binary split between human texts and texts with AI involvement. Overall, GPTZero achieved approximately 78 % accuracy, 99 % precision and 68 % recall; it detected about 78 % of administrative, 68 % of academic and 58 % of literary texts with AI involvement. The results show strong domain dependence and particularly low reliability for post-edited texts.

Keywords

academic integrity; detection of AI-generated text; generative artificial intelligence; GPTZero; large language models

Prohlašuji, že předložená bakalářská práce je původní a zpracoval jsem ji samostatně. Prohlašuji, že citace použitých pramenů je úplná, že jsem v práci neporušil autorská práva (ve smyslu zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů, v platném znění, dále též „**AZ**“).

Byl jsem seznámen s tím, že na mou bakalářskou práci se plně vztahuje **AZ**, zejména § 60 (školní dílo).

Podle § 47b zákona o vysokých školách souhlasím se zveřejněním své práce podle Směrnice pro vedení, vypracování a zveřejňování závěrečných prací na VŠPJ, a to bez ohledu na výsledek obhajoby.

Beru na vědomí, že VŠPJ má právo na uzavření licenční smlouvy o užití mé bakalářské práce a prohlašuji, že **s o u h l a s í m** s případným užitím mé bakalářské práce (prodej, zapůjčení apod.).

Jsem si vědom toho, že užít své bakalářské práce či poskytnout licenci k jejímu využití mohu jen se souhlasem VŠPJ, která má právo ode mě požadovat přiměřený příspěvek na úhradu nákladů, vynaložených vysokou školou na vytvoření díla (až do jejich skutečné výše), z výdělku dosaženého v souvislosti s užitím díla či poskytnutím licence.

V Jihlavě dne 24. listopadu 2025

…….

Podpis studenta

Poděkování

Tímto bych rád poděkoval doc. Voráčkovi za vedení a poskytování rad během své bakalářské práce.

Obsah

Seznam obrázků

Seznam tabulek

Seznam zkratek

AI	Artificial Intelligence (Umělá inteligence)
BERT	Bidirectional Encoder Representations from Transformers
D1	dataset administrativních/formálních
D2	dataset akademické/vědecké
D3	dataset literární
GPT	Generative Pre-trained Transformer
GPTZero	nástroj pro detekci AI-generovaných textů
LLM	Large Language Model (velký jazykový model)
MLP	Multi-Layer perceptron (vícevrstvá perceptronová síť)
NLP	Natural Language Processing (Zpracování přirozeného jazyka)
ROC	Receiver Operating Characteristic křivka

Úvod

Generativní umělá inteligence rychle mění podobu psaní v administrativní komunikaci, akademickém prostředí i literární tvorbě. Modely schopné vytvářet plynulé a stylisticky konzistentní texty jsou dnes běžně dostupné studentům, akademikům i široké veřejnosti a během několika let se staly samozřejmou součástí psacích nástrojů a online služeb. Zároveň ale výrazně narušují tradiční představy o autorství, originalitě a poctivosti. V univerzitním prostředí se proto stále častěji objevuje otázka, jak spolehlivě rozlišit lidské psaní od obsahu vzniklého s pomocí generativní AI a jak s těmito technologiemi realisticky pracovat v rámci výuky, hodnocení a regulace.

V praxi se jako obzvlášť problematické ukazují texty, které nevznikají jako čistý výstup modelu, ale jako kombinace lidské a strojové práce. Typický scénář spočívá v tom, že uživatel nejprve nechá generativní model vytvořit návrh textu a ten následně upravuje tak, aby lépe odpovídal zadání, délce a kontextu. Po těchto úpravách se stopy automatické generace často stírají: lexikální a syntaktické vzorce se přibližují lidskému psaní, lokální nepřirozenosti mizí a detekční nástroje mají potíže takové texty odlišit od originálního díla. Vzniká tím napětí mezi reálnými možnostmi generativní AI a představou, že je možné její použití spolehlivě „odhalit“ čistě technickými prostředky.

Cílem této práce je zhodnotit spolehlivost detekčního nástroje GPTZero při rozlišování lidských a AI-spojených textů v češtině a popsat, v jakých typech textů a za jakých podmínek se klasifikace daří nebo selhává. Zvláštní pozornost je věnována rozdílům mezi plně generovanými texty a texty, které prošly post-editací člověkem, a také vlivu textové domény na chování detektoru. Práce se zaměřuje na prostředí českého vysokého školství, kde otázka využívání generativní AI úzce souvisí s akademickou integritou, férovým hodnocením studentů a s nastavením institucionálních pravidel.

Za tímto účelem byl připraven vyvážený soubor textů uspořádaný do tří domén: administrativní a formální komunikace, akademické a vědecké psaní a literární a esejistická tvorba. Každá doména obsahuje lidské texty, plně generované texty a texty vzniklé post-editací výstupu umělé inteligence člověkem; počet položek v jednotlivých kombinacích je držen srovnatelný, aby bylo možné výsledky mezi doménami férově porovnávat. Tento dataset je analyzován dávkově v pokročilém režimu nástroje GPTZero a výsledky jsou převáděny do matic záměn a standardních metrik, jako jsou accuracy, precision, recall a F1-score, které umožňují posoudit přesnost, citlivost a vyváženost klasifikace.

Práce nesleduje pouze technickou stránku detekce, ale zasazuje ji do širšího teoretického a etického rámce. V teoretické části shrnuje vývoj generativní a detekční umělé inteligence, popisuje základní principy jazykových modelů a detektorů, vysvětluje význam kvantitativních ukazatelů, jako jsou perplexita a burstiness, a věnuje se také etickým a společenským dopadům generativní AI. Pozornost je věnována zejména otázkám plagiátorství, šíření dezinformací, autorským právům a roli AI ve vzdělávání a výzkumu, včetně aktuálních doporučení a rámců pro odpovědné používání těchto nástrojů v akademickém prostředí.

Struktura práce tomu odpovídá. Úvodní kapitoly přinášejí teoretické ukotvení v oblasti generativní a detekční umělé inteligence, vysvětlují klíčové pojmy a principy detekce, například perplexitu a variabilitu vět, a zasazují je do širšího kontextu etických a právních debat. Na ně

navazuje metodická část, která popisuje návrh a konstrukci datasetu, způsob generování a úprav AI textů, postup testování nástrojem GPTZero a zvolený systém vyhodnocení. Praktická část pak prezentuje výsledky pro jednotlivé domény, porovnává chování nástroje u tříd HUM, AI-GEN a AI-EDIT a diskutuje, v jakých situacích lze na technickou detekci spoléhat a kde naopak naráží na své limity. Závěrečná kapitola shrnuje hlavní zjištění, upozorňuje na omezení práce a naznačuje možné směry dalšího výzkumu.

Práce si klade za cíl přispět k informovanému a odpovědnému používání detekčních nástrojů na českých vysokých školách. Neusiluje o definitivní hodnocení konkrétního softwaru, ale o realistický pohled na to, jaké typy textů jsou detekovatelné a kde je riziko omylu příliš vysoké. Zjištění mají nabídnout oporu pro pravidla a pedagogické postupy, které s generativní AI realisticky počítají, tedy s nástrojem, který může být užitečný, ale sám o sobě neřeší otázku poctivosti ani kvality výsledného textu.

1 Teoretická východiska a přehled dosavadních přístupů

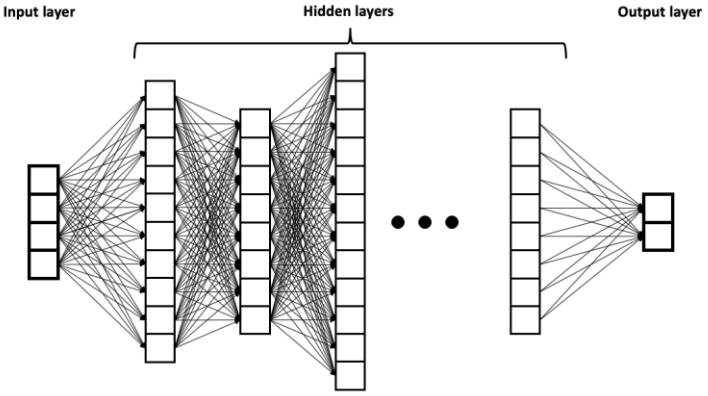
Teoretická část práce se zaměřuje na vysvětlení klíčových konceptů spojených s generativní a detekční umělou inteligencí. Obsahuje přehled základů umělé inteligence, strojového učení a hlubokého učení, na nichž technologie staví. Dále popisuje vývoj jazykových modelů, jejich aplikace v generativní AI a specifické nástroje využívané pro tvorbu textů. Zvláštní pozornost je věnována etickým a společenským aspektům, jako je riziko zneužití generativní AI, otázky autorských práv a vliv těchto technologií na společnost a trh práce. Kapitola zároveň zahrnuje přehled dostupných detekčních nástrojů a popisuje budoucí trendy v generativní a detekční AI. Teoretický základ slouží jako východisko pro praktickou část zaměřenou na komparativní analýzu detekčních nástrojů.

1.1 Základy umělé inteligence a její aplikace v generování textů

Umělá inteligence (AI) se během posledních několika desetiletí stala jednou z klíčových oblastí technologického výzkumu a inovací. AI zahrnuje různé techniky a algoritmy, které umožňují strojům vykonávat úkoly, jež by jinak vyžadovaly lidskou inteligenci, jako je rozpoznávání vzorů, rozhodování nebo zpracování přirozeného jazyka (Natural Language Processing) (Russell, et al., 2021). Její aplikace sahají od autonomních vozidel přes zdravotnictví až po tvorbu textů a obsahu. (Bommasani, 2021)

V kontextu generování textů hraje umělá inteligence zásadní roli v oblastech, jako je tvorba obsahu, automatizace administrativních úkonů, nebo dokonce personalizace marketingových kampaní (Chollet, 2018). Pokročilé jazykové modely, jako jsou GPT, ukázaly, že AI může produkovat texty, které jsou téměř nerozeznatelné od těch vytvořených lidmi (Vaswani, a další, 2017). Schopnosti umělé inteligence však vyvolávají i etické otázky, například v oblasti šíření dezinformací, plagiátorství nebo zneužití technologie k manipulaci veřejného mínění (Russell, et al., 2021).

Vývoj AI prošel několika etapami. Od základních algoritmů založených na pravidlech přes první aplikace strojového učení (Machine Learning) až po hluboké učení (Deep Learning), které využívá vícevrstvé neuronové sítě pro analýzu a generování dat (Chollet, 2018). Kapitola se zaměří na definici umělé inteligence, principy jejích klíčových technologií a jejich aplikace v oblasti generování textů.



Obr. 1: Schéma vícevrstvé perceptronové sítě (MLP).
Zdroj: wikipedia.com (2025)

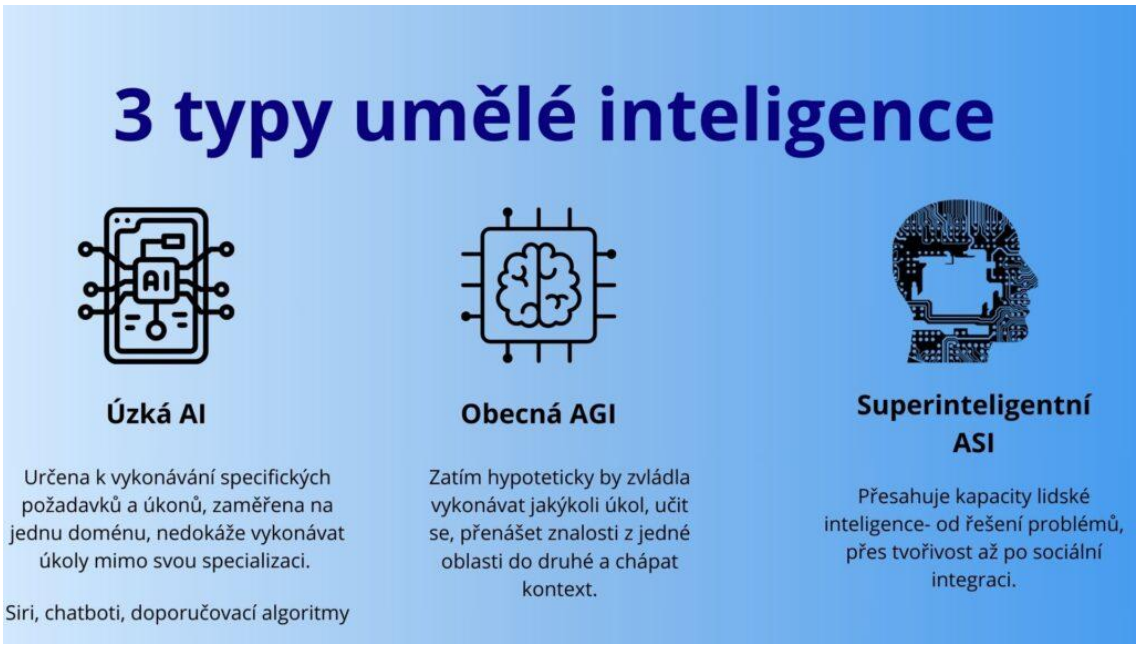
1.1.1 Co je umělá inteligence

Umělá inteligence je interdisciplinární obor informatiky, jehož cílem je vytvářet systémy, které napodobují nebo simulují lidskou inteligenci. Mezi základní schopnosti těchto systémů patří učení, plánování, rozhodování, rozpoznávání obrazů a zpracování přirozeného jazyka (Russell, et al., 2021). AI se zaměřuje na automatizaci úkolů, které by jinak vyžadovaly lidskou inteligenci, s cílem zlepšit efektivitu, přesnost a rychlost zpracování dat.

Historie AI sahá až do poloviny 20. století, kdy Alan Turing navrhl koncept „stroje schopného myslet“ ve své práci „Computing Machinery and Intelligence“ (Turing, 1950). První praktické aplikace se objevily v 50. a 60. letech minulého století, kdy byly vyvinuty programy schopné hrát šachy nebo řešit matematické problémy (Russell, et al., 2021). S postupným rozvojem výpočetní techniky a algoritmů strojového učení se AI začala více zaměřovat na aplikace, které jsou dnes široce využívány, jako je autonomní řízení, rozpoznávání hlasu nebo generování textů.

AI lze rozdělit do tří hlavních kategorií podle úrovně inteligence:

- Úzká AI (Narrow AI): Specializované systémy navržené pro konkrétní úkoly, například chatboty nebo systémy pro detekci podvodů.
- Obecná AI (General AI): Hypotetické systémy schopné provádět jakýkoliv úkol, který dokáže člověk. Forma AI zatím nebyla dosažena.
- Superinteligence (Superintelligence): Teoretická úroveň inteligence, která by překonala lidské schopnosti ve všech ohledech. (Bostrom, 2014)



Obr. 2: typy umělé inteligence podle funkcionality
Zdroj: octodeep.com (2025)

Klíčem k úspěchu AI je její schopnost analyzovat velké objemy dat, odhalovat v nich vzory a na jejich základě vytvářet predikce nebo rozhodnutí (Chollet, 2018). Díky těmto schopnostem AI mění způsob, jakým funguje moderní společnost, a zároveň vyvolává nové otázky týkající se etiky, bezpečnosti a regulace. (Jobin, a další, 2019)

1.1.2 Základy strojového učení a hlubokého učení

Strojové učení (Machine Learning) je podmnožinou umělé inteligence, která se zaměřuje na vývoj algoritmů umožňujících strojům učit se z dat a zlepšovat své výkony bez nutnosti explicitního programování. Základním principem je analýza dat a následné vytváření modelů, které dokážou predikovat budoucí chování nebo výsledky na základě naučených vzorců. (Russell, et al., 2021).

Existují tři hlavní typy strojového učení:

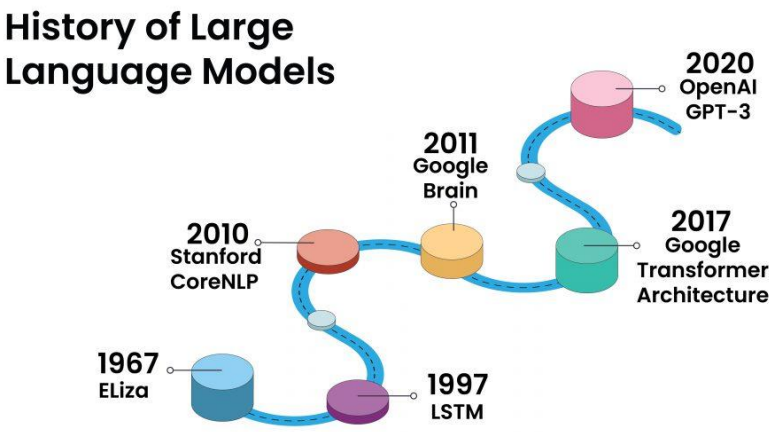
- Učené s učitelem (Supervised Learning): Model je trénován na datech s označenými výstupy, což umožňuje predikovat výsledky pro nová data. Příkladem je klasifikace e- mailů na „spam“ a „ne-spam“.
- Učené bez učitele (Unsupervised Learning): Model zkoumá strukturu dat bez předem definovaných výstupů, například při shlukové analýze (clustering).
- Posilované učení (Reinforcement Learning): Model se učí na základě zpětné vazby z prostředí, což je užitečné například u autonomních vozidel (Russell, et al., 2021).

Hluboké učení (Deep Learning) je specifickou formou strojového učení, která využívá hluboké neuronové sítě s více vrstvami pro zpracování složitých datových struktur. Modely inspirované strukturou lidského mozku využívají spolupráci jednotlivých vrstev neuronů při zpracování informací (Chollet, 2018). Díky své architektuře je hluboké učení schopné automaticky extrahovat relevantní rysy z dat, což je klíčové pro úlohy, jako je rozpoznávání obrazu, řeč nebo generování textu (Vaswani, a další, 2017)

Jedním z revolučních přístupů v hlubokém učení je použití transformerů, které byly představeny ve studii „Attention Is All You Need“. (Vaswani, a další, 2017) Transformery zavedly mechanismus „self-attention“, který umožňuje modelům lépe porozumět vztahům mezi slovy v textu, a tím výrazně zlepšily výkon jazykových modelů, jako je GPT. Pokrok v této oblasti položil základy pro současné generativní modely, které nacházejí široké uplatnění v AI. I přes svůj potenciál má hluboké učení své limity, například vysoké nároky na výpočetní výkon a potřebu velkých datových sad pro efektivní trénink (Chollet, 2018). Výzvy však stimulují další výzkum, který se snaží bariéry překonat.

1.1.3 Jazykové modely používané v generativní AI

Jazykové modely jsou klíčovým prvkem generativní umělé inteligence. Modely analyzují sekvence slov a předpovídají jejich pravděpodobnost, což jim umožňuje generovat texty, odpovídat na otázky nebo provádět překlady. Vývoj jazykových modelů prošel několika etapami, od jednodušších statistických metod až po moderní hluboké neuronové sítě, jako jsou transformery (Russell, et al., 2021).



Obr. 3: Přehled vývoje velkých jazykových modelů
Zdroj: scribbledata.com

Původní přístupy k modelování jazyka byly založeny na statistických metodách, například n – gramových modelech. Modely využívaly pravděpodobnostní analýzu k predikci následujícího slova na základě předchozí sekvence. Přestože byly n-gramové modely efektivní v jednoduchých aplikacích, měly zásadní omezení v porozumění širším kontextům textu (Russell, et al., 2021).

Průlom ve vývoji jazykových modelů nastal s příchodem hlubokých neuronových sítí a zejména transformerů. Mechanismus „self-attention“, popsaný ve studii *„Attention Is All You Need“* (Vaswani, a další, 2017), umožnil modelům efektivněji chápat vztahy mezi slovy napříč celým textem. Transformery, jako je BERT, jsou optimalizovány pro úlohy, jako je analýza sentimentu nebo odpovídání na otázky, zatímco modely GPT se zaměřují na generování přirozeného textu.

Model GPT, vyvinutý společností OpenAI, se stal jedním z nejvýznamnějších pokročilých jazykových modelů. Díky před trénování na masivních datových sadách a následnému jemnému doladění pro konkrétní úlohy je GPT schopen vytvářet texty, které jsou k nerozeznání od těch vytvořených člověkem. Jeho schopnost generovat soudržné a kontextově relevantní odpovědi je výsledkem implementace transformátorového mechanismu (Vaswani, a další, 2017).

I přes svůj pokrok mají moderní jazykové modely určité limity. Například vysoká závislost na velkých tréninkových datových sadách může vést k tomu, že modely přejímají předsudky obsažené v datech. Další výzvou jsou náklady na výpočetní výkon a etické otázky spojené s možností zneužití technologie, například k šíření dezinformací nebo vytváření plagiátorského obsahu (Russell, et al., 2021).

1.1.4 Etické a praktické otázky generování textů pomocí AI

Generativní umělá inteligence přináší významné technologické možnosti, ale zároveň otevírá řadu etických a praktických otázek. Mezi hlavní etické obavy patří šíření dezinformací, plagiátorství a možnost manipulace veřejného mínění. Praktické výzvy pak zahrnují náročnost implementace těchto technologií, jejich dostupnost a potenciální zneužití.

Jedním z největších rizik spojených s generováním textů pomocí AI je šíření dezinformací. Generativní modely, jako jsou GPT, mohou být využity k vytváření realisticky znějících textů, které mohou být zavádějící nebo falešné (Russell, et al., 2021). Například v oblasti politiky může

být technologie zneužita k vytváření falešných zpráv nebo k manipulaci voličů prostřednictvím personalizovaných dezinformačních kampaní.

Další významnou etickou otázkou je plagiátorství. Modely generativní AI jsou často trénovány na obrovských objemech textů získaných z internetu, což může zahrnovat autorsky chráněný obsah. I když generované texty nejsou přímými kopiemi, mohou obsahovat prvky, které se podobají originálním dílům, čímž mohou porušovat autorská práva (Chollet, 2018).

Z praktického hlediska představuje výzvu dostupnost a náklady na výpočetní výkon. Trénování a provozování velkých jazykových modelů, jako je GPT-4, vyžaduje značné zdroje, což omezuje jejich dostupnost na technologické giganty a výzkumné instituce (Vaswani, a další, 2017). Koncentrace technologických kapacit může vést k nerovnoměrnému rozložení moci a znalostí, což představuje riziko pro otevřenost a demokratický přístup k AI technologiím.

Navzdory těmto výzvám existují iniciativy zaměřené na zmírnění negativních dopadů generativní AI. Například vývoj etických směrnic a regulací, jako je „AI Ethics Guidelines“ Evropské unie, se snaží zajistit odpovědné využívání těchto technologií (Russell, et al., 2021). Důležité je také pokračovat ve výzkumu metod detekce textů generovaných AI, což může pomoci omezit zneužití této technologie.

Celkově vzato je generativní AI nástrojem s obrovským potenciálem, ale její odpovědné využití vyžaduje pečlivé zvážení etických a praktických dopadů. Balancování mezi inovacemi a odpovědností zůstává klíčovou výzvou budoucnosti.

1.2 Technologie a nástroje pro generování a detekci textů

Technologie spojené s generováním textů prostřednictvím umělé inteligence a jejich následnou detekcí představují významnou oblast moderního výzkumu a aplikace. Generativní modely, jako je GPT (Generative Pre-trained Transformer), využívají hluboké neuronové sítě a moderní architektury, například transformery, k analýze a reprodukci jazykových struktur. Modely umožňují vytvářet texty, které jsou koherentní, smysluplné a často nerozeznatelné od textů psaných člověkem. Typickými aplikacemi jsou marketing, tvorba kreativního obsahu, automatizace zákaznické podpory nebo dokonce generování vědeckých článků (Russell, et al., 2021).

S rostoucí sofistikovaností generativních modelů vzniká potřeba nástrojů schopných rozlišit texty generované AI od textů vytvořených člověkem. Detekce těchto textů je klíčová zejména v akademickém prostředí, při boji proti šíření dezinformací nebo při ověřování autenticity publikovaných materiálů. Detekční nástroje, jako je DetekceGPT, Smodin nebo Isgen, kombinují analýzu stylu psaní, sémantickou analýzu a detekci anomálií v textu. Algoritmy využívají specifické jazykové vzorce a syntaktické charakteristiky, které mohou být pro generované texty typické. (Weber-Wulff, 2023).

Generování a detekce textů jsou navzájem propojené technologie, které představují dvě strany jednoho problému. Zatímco generativní modely se zaměřují na co nejpřirozenější vytváření obsahu, detekční nástroje usilují o identifikaci jeho původu. Toto propojení umožňuje vývoj systémů, které dokážou nejen generovat obsah, ale zároveň zajistit jeho autenticitu a etickou odpovědnost. Výzvami zůstávají etické otázky spojené s používáním generativní AI, například šíření nepravdivých informací či plagiátorství, a technické výzvy, jako je omezení předsudků obsažených v tréninkových datech (Chollet, 2018).

Budoucnost těchto technologií bude spočívat v optimalizaci generativních a detekčních modelů tak, aby byly využitelné v širokém spektru aplikací, od vzdělávání přes vědu až po průmyslové nasazení. Spojením generativních a detekčních nástrojů lze vytvářet systémy schopné automaticky kontrolovat kvalitu obsahu, což může přispět k větší transparentnosti a důvěře ve zveřejňované informace

1.2.1 Generativní modely umělé inteligence

Generativní modely umělé inteligence patří mezi pokročilé technologie strojového učení, jejichž hlavním cílem je vytváření nových dat na základě vzorců naučených z tréninkových sad. V oblasti textu se využívají zejména pro automatizovanou tvorbu obsahu, překlady, sumarizace či odpovídání na otázky (Russell, et al., 2021).

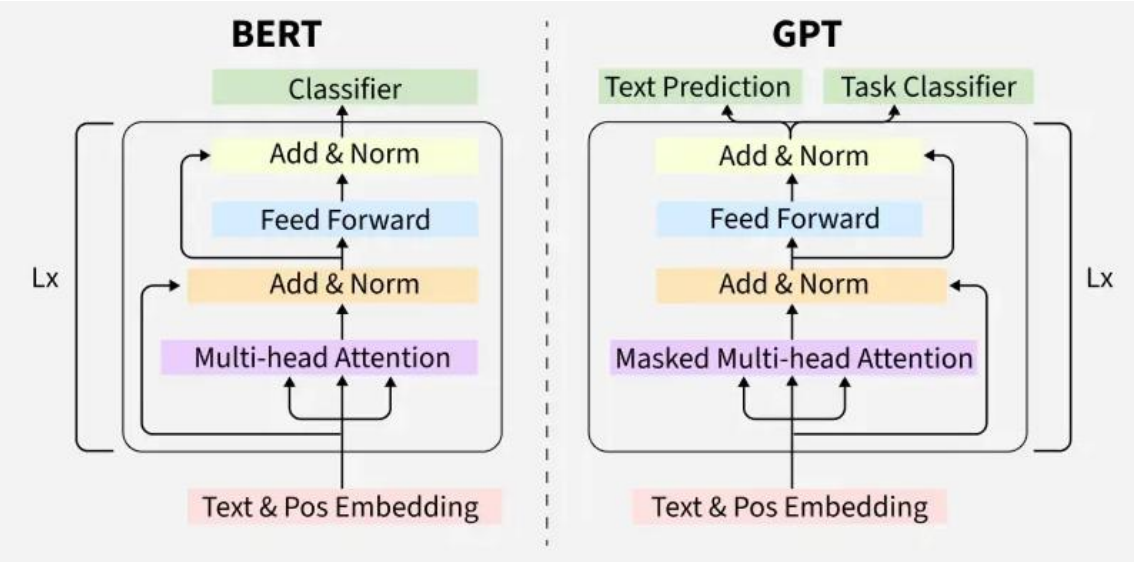
Jedním z nejvýznamnějších přístupů v generativní AI je využití hlubokých neuronových sítí, konkrétně transformerů. Modely jako GPT (Generative Pre-trained Transformer) kombinují proces před trénování na rozsáhlých datových sadách s následným jemným doladěním na specifické úlohy. Transformery, poprvé představené ve studii Attention Is All You Need, zavedly mechanismus self-attention, který umožňuje efektivní analýzu vztahů mezi slovy v textu bez ohledu na jejich pozici v sekvenci (Vaswani, a další, 2017).

Model GPT, vyvinutý společností OpenAI, se stal klíčovým hráčem na poli generativních modelů. Jeho před trénování na miliardách textových dokumentů a flexibilita v aplikacích, jako je tvorba

kreativních textů nebo automatizace administrativních procesů, posunuly možnosti generativní AI na novou úroveň (Chollet, 2018).

Další důležitý model, BERT (Bidirectional Encoder Representations from Transformers), je optimalizovaný pro analýzu textu. Na rozdíl od GPT, které je především generativní, je BERT navržen pro úlohy klasifikace, extrakce a analýzy významu (Russell, et al., 2021).

Přestože generativní modely nabízejí široké možnosti, jejich použití přináší i výzvy, například potřebu velkých výpočetních zdrojů a riziko šíření nepravdivých informací. Výzkum a aplikace těchto modelů však zůstávají jednou z nejdynamičtějších oblastí současné umělé inteligence (Vaswani, a další, 2017; Bommasani, 2021).



Obr. 4: Schéma porovnání architektur BERT a GPT
Zdroj: geeksforgeeks.org

1.2.2 Nástroje pro generování textů

Nástroje pro generování textů jsou praktickými aplikacemi generativních modelů umělé inteligence, jako je GPT, a nacházejí uplatnění v širokém spektru oblastí. Typické aplikace zahrnují automatizovanou tvorbu obsahu, podporu kreativních procesů a zjednodušení administrativních úkolů. Nástroje se od sebe liší rozsahem funkcí, uživatelským prostředím a zaměřením na specifické potřeby uživatelů.

Mezi nejznámější nástroje patří ChatGPT, vyvinutý společností OpenAI, který umožňuje vytvářet koherentní a kontextově relevantní odpovědi na základě zadaných dotazů. Kromě přirozeného jazyka zvládá technické popisy, generování kódu i kreativní úkoly, jako je psaní příběhů nebo poezie (Brown et al., 2020). Další populární aplikace zahrnují Jasper AI a Writesonic, které se specializují na marketingový obsah, jako jsou články, reklamy a příspěvky na sociální sítě. V akademickém a vědeckém prostředí nachází uplatnění OpenAI Codex, který pomáhá s generováním kódu nebo tvorbou technické dokumentace. Vzdělávací aplikace generativní AI zahrnují systémy určené pro tvorbu studijních materiálů a simulaci interaktivních výukových scénářů.

Obr. 5: ukázka prostředí ChatGPT

Zdroj: vlastní zpracování (2025)

Hlavní výhodou těchto nástrojů je schopnost ušetřit čas a zefektivnit proces tvorby obsahu. Automatizace umožňuje vytvářet texty ve velkém měřítku, což je přínosné zejména v oblastech, kde je klíčová rychlost a efektivita, například v marketingu nebo zákaznické podpoře. Kvalita generovaných textů je často na takové úrovni, že pro běžného čtenáře bývá obtížné rozeznat, zda je text vytvořen člověkem nebo AI (Brown, 2020).

I přes výhody přináší využití těchto nástrojů i výzvy. Jedním z hlavních rizik je šíření dezinformací, protože generované texty mohou být zneužity k tvorbě falešného obsahu. Další otázky zahrnují etiku, například problém plagiátorství nebo využívání AI generovaných textů bez adekvátního přiznání. Kvalita výsledků je navíc ovlivněna daty použitými při tréninku, která mohou obsahovat předsudky nebo zkreslení (Russell, et al., 2021). Nástroje se také liší svou dostupností – zatímco ChatGPT je volně přístupný v omezeném režimu, pokročilejší funkce často vyžadují předplatné nebo robustní výpočetní prostředí.

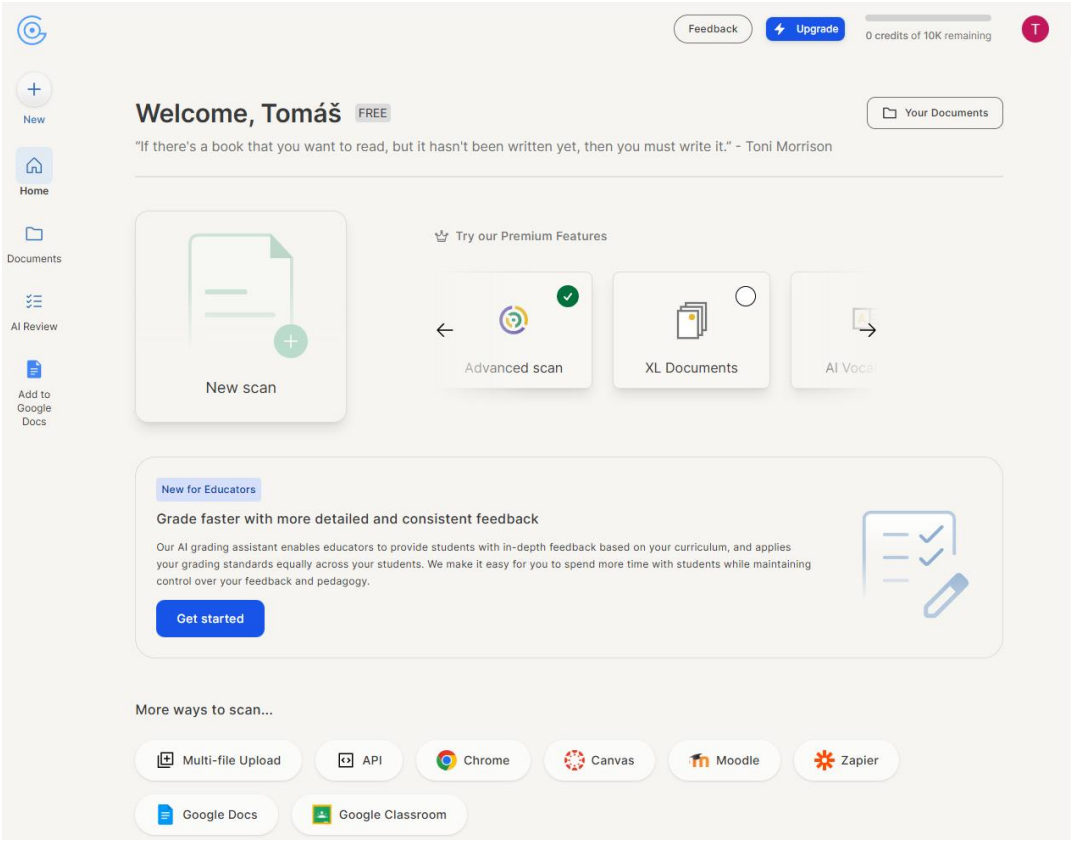
Budoucnost nástrojů pro generování textů směřuje k větší personalizaci a integraci do běžného softwaru, jako jsou textové editory, systémy zákaznické podpory nebo marketingové platformy. Výzkum se zaměřuje na zlepšení kvality generovaných textů a zmírnění rizik spojených s jejich používáním, například etických problémů nebo potenciálního šíření nepravdivých informací (Chollet, 2018).

1.2.3 Detekce AI generovaných textů

S rozvojem generativní umělé inteligence, která umožňuje vytvářet texty nerozeznatelné od těch, které píšou lidé, roste potřeba efektivních nástrojů pro detekci takového obsahu. Detekce AI generovaných textů má klíčový význam v oblastech, jako je akademické prostředí, kde může docházet k plagiátorství, nebo ve veřejném prostoru, kde hrozí šíření dezinformací. Identifikace generovaného obsahu přispívá k zajištění autenticity textů, ochrany autorských práv a transparentnosti v komunikaci.

Moderní detekční metody jsou založeny na kombinaci několika přístupů, které zahrnují analýzu stylu psaní, sémantickou analýzu a detekci anomálií v textu. Jedním z nejpoužívanějších přístupů je porovnávání statistických charakteristik textu, například četnosti slov, délky vět a složitosti

syntaxe. Generativní modely jako GPT nebo BERT mají tendenci produkovat texty, které vykazují jisté vzorce, například vyšší opakování určitých slov nebo preferenci kratších vět. Odlišnosti mohou být detekčními algoritmy identifikovány a použity k určení pravděpodobnosti, že byl text vytvořen umělou inteligencí. (Russell, et al., 2021).



Obr. 6: ukázka prostředí GPTZero

Zdroj: vlastní zpracování (2025)

Nástroje jako DetekceGPT, Smodin a Isgen kombinují algoritmy s pokročilými metodami strojového učení. DetekceGPT například využívá srovnávací analýzu mezi pravděpodobnostní distribucí slov v textu a distribučními vzory běžnými pro generativní modely. Podobně Smodin nabízí rychlou analýzu textu a poskytuje pravděpodobnostní skóre, které indikuje, zda byl text generován AI. Isgen, český nástroj pro detekci, umožňuje identifikaci generovaných textů v několika jazycích a zaměřuje se na preciznost při detekci složitých textů, například odborných článků (Weber-Wulff, 2023; Tang, a další, 2024; Mitchell, 2023).

Jedním z klíčových problémů detekce je neustálé zdokonalování generativních modelů. Nové verze modelů, jako je GPT-4, jsou stále sofistikovanější a dokážou lépe imitovat lidský styl psaní. To klade vysoké nároky na detekční algoritmy, které musí být pravidelně aktualizovány a přizpůsobovány novým vzorcům generovaného textu. Současně existují výzvy spojené s vícejazyčností, protože většina dostupných detekčních nástrojů je optimalizována primárně pro angličtinu a v jiných jazycích vykazuje nižší přesnost (Brown, 2020).

Etické otázky spojené s detekcí AI generovaných textů zahrnují potřebu respektovat soukromí uživatelů a minimalizovat riziko falešných pozitivních výsledků. Chybné označení textu jako generovaného může vést k neoprávněným obviněním, což je obzvláště problematické v akademickém a právním prostředí. Proto je nezbytné, aby detekční nástroje byly transparentní

ve svých metodách, a poskytovaly jasné indikátory, na základě kterých dospěly ke svým závěrům. (Chollet, 2018).

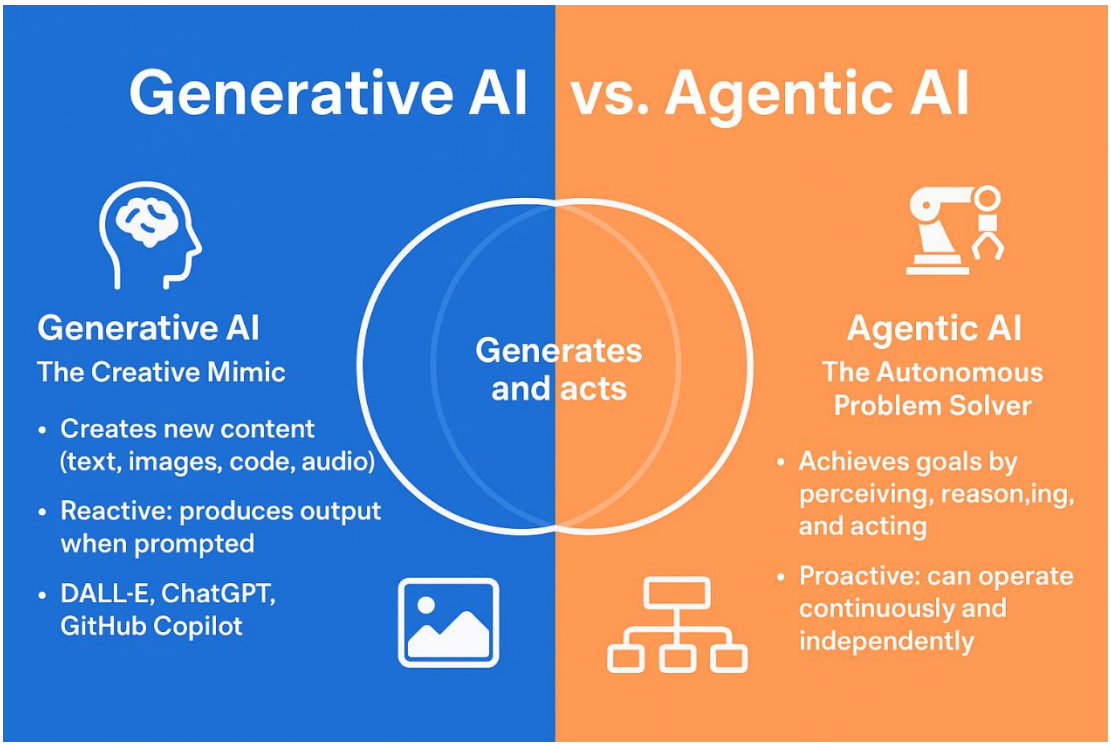
Budoucnost detekce AI generovaných textů spočívá ve vývoji integrovaných řešení, která budou schopna fungovat přímo v reálném čase při tvorbě obsahu. Řešení mohou zahrnovat implementaci detekčních algoritmů do textových editorů nebo nástrojů pro kontrolu pravopisu, což umožní okamžité varování v případě podezření na generovaný obsah. Dalším směrem vývoje je zvýšení dostupnosti detekčních nástrojů v různých jazycích a jejich adaptace pro široké spektrum aplikací, od vzdělávání po průmyslové využití (Weber-Wulff, 2023; Brown, 2020).

1.2.4 Srovnání generativních a detekčních nástrojů

Generativní a detekční nástroje představují dva protichůdné přístupy ke zpracování textů pomocí umělé inteligence, přičemž jejich cíle jsou vzájemně opačné, ale zároveň úzce propojené. Generativní nástroje, jako je ChatGPT, Jasper AI nebo Writesonic, se zaměřují na vytváření nových textových dat na základě zadaných vstupů, přičemž využívají pokročilé algoritmy hlubokého učení a transformery, které umožňují efektivní modelování jazykových struktur. Naproti tomu detekční nástroje, například DetekceGPT, Smodin či Isgen, jsou navrženy tak, aby identifikovaly texty vytvořené generativními modely a poskytovaly informace o jejich možném původu (Russell, et al., 2021; Weber-Wulff, 2023).

Generativní nástroje excelují ve schopnosti vytvářet texty, které jsou nejen smysluplné, ale také stylisticky odpovídají lidskému psaní. Díky jejich flexibilitě lze generovat širokou škálu textů, od marketingových sdělení a technických popisů po kreativní obsah, jako jsou příběhy či poezie. Výhodou generativních nástrojů je rychlost a škálovatelnost – dokážou zpracovávat velké objemy textových požadavků během několika sekund. Nicméně jejich výkon je závislý na kvalitě tréninkových dat, což může vést k problémům s předsudky nebo reprodukcí nepřesných informací (Brown, 2020).

Na druhé straně detekční nástroje hrají klíčovou roli při kontrole autenticity obsahu. Díky analýze lingvistických vzorců a distribuce pravděpodobnosti slov jsou schopny identifikovat texty, které vykazují typické znaky generativních modelů. Například DetekceGPT se zaměřuje na analýzu pravděpodobnostních vzorců, zatímco Isgen využívá kombinaci sémantické analýzy a detekce anomálií pro identifikaci generovaného obsahu. Hlavní výzvou pro detekční nástroje je rychlost adaptace na nové generativní modely, které neustále zlepšují svou schopnost imitovat lidský styl psaní (Weber-Wulff, 2023).



Obr. 7: Srovnání generativní AI a agentní AI podle funkcí a míry autonomie

Zdroj: javascript.plainenglish.io (2025)

Při srovnání těchto dvou kategorií nástrojů je zřejmé, že generativní systémy jsou technologicky pokročilejší, protože vyžadují hlubší porozumění jazykovým strukturám a schopnost generovat smysluplné texty. Detekční nástroje jsou však stejně důležité, protože jejich cílem je chránit autenticitu a integritu obsahu v prostředí, kde generativní modely dominují. Výkon obou kategorií závisí na stejných principech – kvalitě tréninkových dat a účinnosti algoritmů hlubokého učení.

Významnou otázkou zůstává etické využití těchto technologií. Generativní nástroje mohou být zneužity ke generování falešného obsahu, například dezinformací nebo klamavých reklam, zatímco detekční nástroje mohou při chybné detekci poškodit důvěryhodnost uživatelů. Budoucí vývoj by měl směřovat k integraci obou přístupů do komplexních systémů, které by umožnily nejen tvorbu obsahu, ale také jeho automatickou kontrolu na autenticitu. Kombinace by mohla být klíčová pro udržení rovnováhy mezi inovacemi v oblasti umělé inteligence a potřebou zajištění etické a odpovědné aplikace technologií (Russell, et al., 2021; Chollet, 2018).

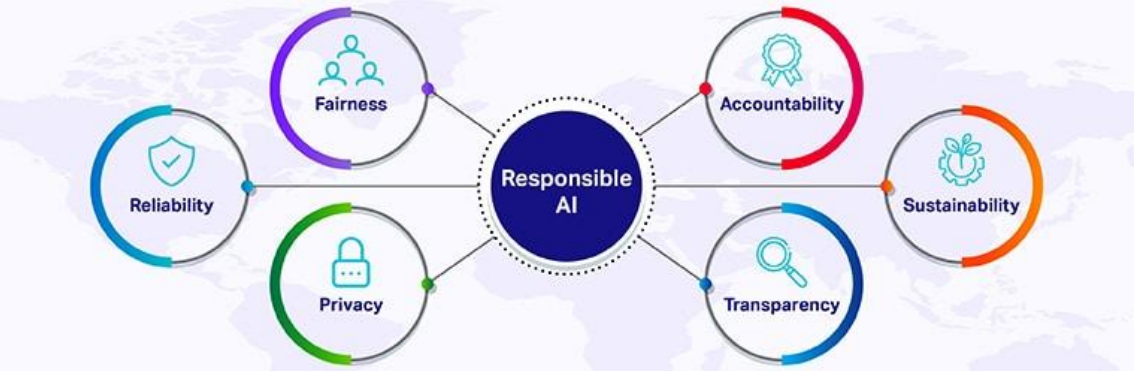
Z hlediska praktického využití představuje srovnání těchto nástrojů významný krok směrem k pochopení jejich omezení i příležitostí. Generativní modely budou nadále hrát dominantní roli v automatizaci tvorby obsahu, zatímco detekční technologie musí držet krok s jejich vývojem, aby zůstaly efektivní. Společný výzkum a propojení těchto oblastí tak nabízejí slibné možnosti nejen pro akademické a komerční aplikace, ale také pro zvýšení důvěryhodnosti obsahu ve veřejném prostoru (Brown, 2020; Weber-Wulff, 2023).

1.3 Etické a společenské aspekty generativní AI

Rozvoj generativní umělé inteligence přináší nejen technologické inovace, ale také řadu etických a společenských výzev. Modely jako GPT nebo BERT mají potenciál transformovat mnoho

odvětví, avšak jejich masové nasazení může vést k problémům, které přesahují technologickou rovinu a zasahují do otázek, jako jsou ochrana soukromí, autorská práva nebo šíření dezinformací (Bostrom, 2014; Floridi, 2019; Jobin, a další, 2019). Etická odpovědnost při navrhování, vývoji a využívání generativní AI je proto klíčová pro zajištění jejího bezpečného a udržitelného nasazení.

Jedním z hlavních etických problémů je zneužití generativní AI k tvorbě nepravdivých nebo manipulativních informací. Modely jsou schopny generovat texty, které mohou být použity k šíření dezinformací, propagandy nebo k napodobování identity jiných osob, což otevírá otázky týkající se odpovědnosti za jejich použití (Floridi, 2019). Technologie zároveň vyvolávají diskusi o ochraně autorských práv. Generované texty často vycházejí z dat použitých během tréninku, která mohou obsahovat chráněný obsah, což ztěžuje identifikaci původního autora a přisuzování práv.



Obr. 8: Hlavní principy odpovědné umělé inteligence
Zdroj: cs.shaip.com (2025)

Dalším významným aspektem je rovnováha mezi inovací a regulací. Na jedné straně stojí zájem vývojářů a společností na rychlém pokroku, zatímco na straně druhé je potřeba chránit uživatele a zabránit zneužití. Regulace generativní AI, například prostřednictvím globálních standardů nebo zákonů, by měla být navržena tak, aby nezpomalila technologický pokrok, ale zároveň zajistila odpovědné a etické používání (Russell, et al., 2021).

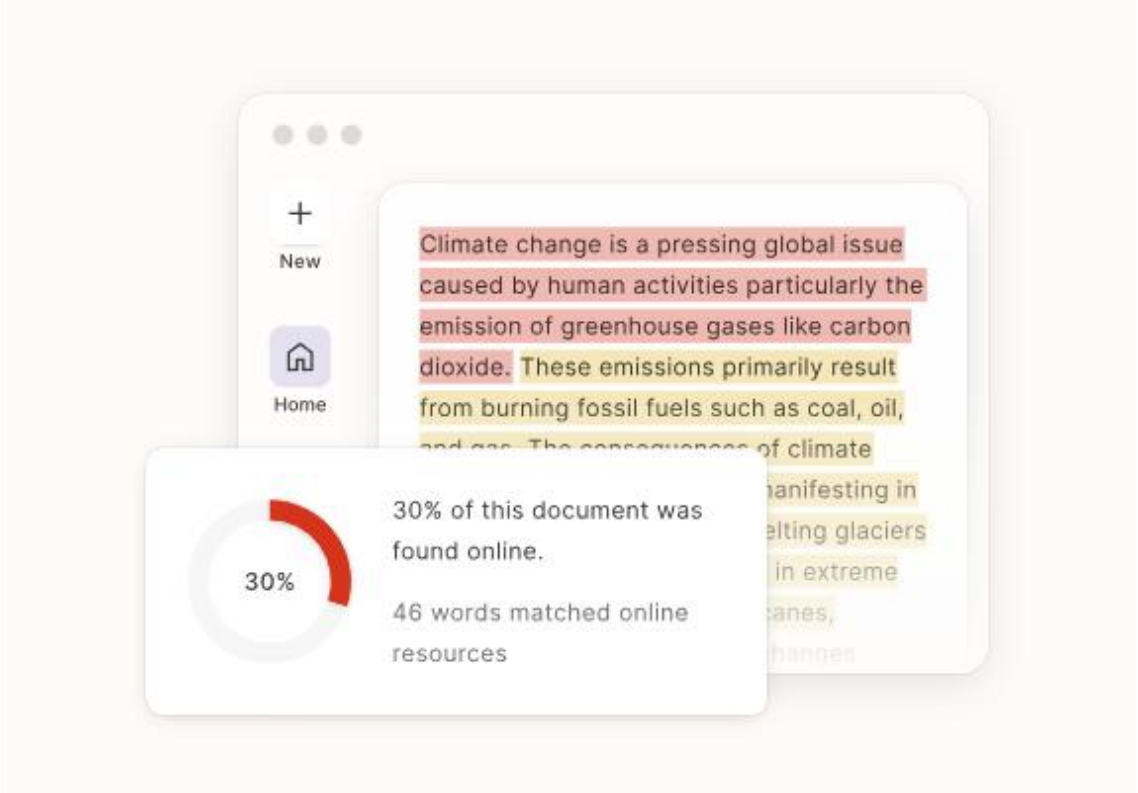
Společenské dopady generativní AI se neomezují pouze na technologii samotnou. Automatizace textových úloh, jako je tvorba marketingového obsahu, zákaznická podpora nebo psaní článků, může ovlivnit trh práce a změnit dovednosti, které budou od pracovníků požadovány. Zatímco některé úlohy mohou být díky generativní AI zcela automatizovány, jiné mohou vyžadovat užší spolupráci mezi člověkem a AI, což přinese nové požadavky na školení a vzdělávání (Chollet, 2018).

Etická odpovědnost vývojářů a společností, které modely vytvářejí, spočívá nejen v minimalizaci rizik spojených s jejich používáním, ale také v podpoře transparentnosti a odpovědnosti. Vyvíjející se technologie musí být zároveň doprovázena informovaností veřejnosti, která je klíčová pro pochopení přínosů i rizik generativní AI (Floridi, 2019).

Etické a společenské otázky generativní AI jsou komplexní a vyžadují spolupráci odborníků z různých oblastí. Jedině tak lze zajistit, aby rozvoj těchto technologií přispíval k pozitivním změnám a minimalizoval negativní dopady na společnost.

1.3.1 Zneužití generativní AI

Generativní umělá inteligence má schopnost transformovat různé oblasti lidské činnosti, avšak její neregulované používání může vést ke značným rizikům. Jedním z nejzávažnějších problémů je možnost zneužití této technologie k vytváření obsahu, který může manipulovat, klamat nebo poškodit jednotlivce či celé komunity. Schopnost generativních modelů, jako je GPT, vytvářet texty k nerozeznání od textů psaných člověkem zvyšuje riziko jejich zneužití například v šíření dezinformací, kybernetické kriminalitě či napodobování identity (Floridi, 2019).



Obr. 9: Ukázka plagiátové shody detekované nástrojem GPTZero
Zdroj: gptzero.me (2025)

Jedním z nejvíce diskutovaných problémů je šíření dezinformací. Generativní modely umožňují vytvářet obsah ve velkém měřítku, což může být zneužito k propagandě, politické manipulaci nebo šíření falešných informací. Například během volebních kampaní mohou být generovány texty, které napodobují komunikaci konkrétních politiků nebo organizací, čímž dochází k manipulaci veřejného mínění. Fenomén je umocněn tím, že generativní AI dokáže přizpůsobit styl a jazyk tak, aby obsah působil důvěryhodně (Bostrom, 2014).

Dalším rizikem je možnost napodobování identity. Generativní AI může být zneužita k vytváření textů, které napodobují konkrétní osoby, například v e-mailech nebo na sociálních sítích. Typ kybernetického útoku, známý jako phishing, může vést k neoprávněnému přístupu k citlivým informacím, finančním ztrátám nebo reputačním škodám. Pokročilé modely, které dokážou napodobit styl psaní jednotlivců, ztěžují detekci těchto podvodů tradičními metodami (Russell, et al., 2021).

Zneužití generativní AI se neomezuje pouze na textový obsah. Modely mohou být propojeny s multimodálními systémy, například pro generování obrázků nebo videí, což otevírá dveře

k tvorbě tzv. deepfakes. Technologie mohou být zneužity k diskreditaci jednotlivců, vytváření falešných důkazů nebo šíření nenávistného obsahu. Kombinace generativních textových a vizuálních modelů představuje další výzvu pro detekční a regulační mechanismy (Floridi, 2019).

Proti těmto hrozbám lze bojovat pouze prostřednictvím kombinace technických, právních a vzdělávacích opatření. Technologická řešení zahrnují vývoj detekčních nástrojů schopných identifikovat generovaný obsah, zatímco právní rámce by měly definovat odpovědnost za zneužití těchto technologií. Důležitá je také informovanost veřejnosti, která musí být schopna rozpoznat a kriticky hodnotit generovaný obsah.

Zneužití generativní AI představuje komplexní problém, který vyžaduje spolupráci mezi vývojáři technologií, vládami a uživateli. Pouze integrovaný přístup může minimalizovat negativní dopady a zajistit odpovědné využívání těchto technologií.

1.3.2 Otázky autorských práv

Generativní umělá inteligence otevírá nové možnosti tvorby textového obsahu, ale zároveň přináší značné právní výzvy, zejména v oblasti autorských práv. Hlavní otázkou je, kdo by měl mít nárok na autorská práva k textům generovaným AI – zda autor vstupních dat, vývojář modelu, nebo uživatel, který obsah vytvořil za pomoci AI. Problém je zásadní pro ochranu práv jednotlivců i organizací využívajících generativní technologie. (Bostrom, 2014).

Generované texty často vycházejí z dat použitých při tréninku modelu, což zahrnuje obrovské množství informací shromážděných z internetu. Texty mohou zahrnovat autorsky chráněný obsah, což vyvolává otázku, zda je generovaný text originálním dílem, nebo pouze derivátem původních dat. Pokud text přímo kopíruje části původního obsahu, může dojít k porušení autorských práv, což již bylo předmětem několika právních sporů v zahraničí (Floridi, 2019).

Další otázkou je, zda lze generované texty považovat za autorsky chráněné. Podle současné právní úpravy ve většině zemí jsou autorská práva přiznávána pouze lidem, nikoli strojům. Pokud AI vytvoří text bez přímého tvůrčího zásahu člověka, není jasné, kdo by měl mít nárok na autorské právo. Situace je zvláště problematická v oblastech, kde generovaný obsah má významnou komerční hodnotu, například v marketingu, novinářství nebo tvůrčí sféře.(Russell, et al., 2021).

Další komplikací je použití generativní AI k úpravám nebo rozšiřování již existujících textů. Pokud například AI přepracuje text chráněný autorskými právy, může být obtížné určit, zda se jedná o nové dílo, nebo o porušení práv původního autora. V takových případech se často používají právní koncepty, jako je "transformativní využití", což vyžaduje posouzení míry originality a nového významu přidaného AI (Bostrom, 2014).

Některé země a organizace již začaly hledat řešení těchto právních otázek. Například Evropská unie zvažuje zavedení právního rámce, který by stanovil pravidla pro využívání dat při tréninku modelů AI a ochranu práv autorů původních děl. Iniciativy by mohly přispět ke zvýšení právní jistoty a minimalizovat konflikty mezi vývojáři AI, autory a uživateli generovaných textů (Floridi, 2019).

Otázky autorských práv v souvislosti s generativní AI vyžadují důkladnou analýzu a spolupráci mezi právníky, technologickými odborníky a regulačními orgány. Vyřešení těchto otázek je

klíčové pro zajištění rovnováhy mezi podporou inovací a ochranou práv autorů, což umožní odpovědné a spravedlivé využívání generativních technologií.

1.3.3 Rovnováha mezi inovací a regulací

Rozvoj generativní umělé inteligence přináší nové příležitosti pro inovace, ale zároveň vyvolává otázky týkající se regulace. Generativní AI má potenciál transformovat různé oblasti lidské činnosti, od vědy a vzdělávání až po kreativní průmysl. Současně její neregulované využívání může vést k významným rizikům, jako je šíření dezinformací, porušování autorských práv nebo narušování soukromí. Proto je klíčové nalézt rovnováhu mezi podporou inovací a zavedením regulačních opatření, která zajistí odpovědné používání této technologie (Bostrom, 2014; Floridi, 2019; Jobin, a další, 2019).

	Principle	*Summary*	*Relevance to DP for EIAI*
1	Proportionality and do no harm	AI should aim for legitimate aims without harm.	Provides data to prevent harm and address social issues.
2	Justification of AI use	AI must be appropriate, respect rights, and be rigorous.	Ensures ethical data sourcing for beneficial AI.
3	Safety and security	AI systems should address safety and security risks.	Enhances AI safety with secure data sharing.
4	Fairness and non-discrimination	AI should prevent discrimination and promote justice.	Offers diverse data to reduce AI bias.
5	Minimizing discrimination	Efforts to avoid AI biases and discrimination are vital.	Contributes representative datasets for fairness.
6	Sustainability	AI's impact on sustainability should be assessed.	Prevents duplication in data acquisition efforts to save energy.
7	Right to privacy and data protection	Privacy must be protected throughout AI's life cycle.	Ensures data sharing, respects privacy standards.
8	Data protection frameworks	Establish data protection governance.	Participates in frameworks for responsible sharing.
9	Privacy by design	AI must protect personal information.	Anonymizing data before sharing.
10	Human oversight and determination	Persons/entities should be responsible for AI systems.	Conducts data sharing with clear accountability.
11	Human decision in AI reliance	Humans should control AI systems.	Supports AI that enhances human decision-making.
12	Transparency and explainability	AI should be transparent and explainable.	Advocates for clarity on donated data usage in AI.
13	Balancing transparency	Balance transparency with privacy and security.	Navigates data sharing between transparency and privacy.
14	Transparency for trust	Transparency contributes to trust in AI.	Fosters trust by being clear about data usage.
15	Explainability of AI systems	AI should be understandable and insightful.	Demands explanations on data impact on AI.
16	Responsibility and accountability	AI actors should assume responsibility for AI impact.	Holds AI actors accountable for ethical data use.
17	Accountability mechanisms	Develop AI oversight and audit mechanisms.	Includes DP in accountability tracking.
18	Awareness and literacy	Promote understanding of AI.	Plays a role in AI education and engagement.
19	Learning about AI impact	Educate on AI's societal impact.	Supports research on AI's effects on society.
20	Multi-stakeholder governance	Respect laws in data use and regulation.	Engages with stakeholders for respectful data use.
21	Stakeholder participation	Diverse participation for inclusive AI governance.	Involves various stakeholders for diverse AI input.

Obr. 10: Dvacet jedna etických principů UNESCO pro umělou inteligenci

Zdroj: researchgate.net (2022)

Na jedné straně stojí zájmy technologických firem, které generativní AI vyvíjejí. Organizace usilují o rychlý pokrok a nasazení nových aplikací, což vyžaduje svobodu v experimentování a přístup k velkým datovým sadám. Přísné regulace by mohly zpomalit vývoj, omezit konkurenceschopnost a snížit motivaci pro investice do výzkumu. Na straně druhé je však nezbytné chránit zájmy společnosti a jednotlivců před možnými negativními dopady, což zahrnuje ochranu soukromí, prevenci zneužití a zachování etických standardů (Floridi, 2019).

Jedním z hlavních přístupů k dosažení rovnováhy je zavedení principu "by design". Koncept znamená, že etické a regulační požadavky by měly být zohledněny již při návrhu generativních systémů. Příkladem je začlenění detekčních mechanismů přímo do generativních modelů, které by umožnily snadnou identifikaci generovaného obsahu. Dalším opatřením může být

transparentnost algoritmů, kdy by vývojáři měli povinnost zveřejnit základní principy fungování modelů a způsob práce s daty (Russell, et al., 2021).

Regulační přístupy by měly být zároveň flexibilní a přizpůsobivé. Generativní AI je dynamický obor, ve kterém se technologie vyvíjejí rychleji než legislativa. Namísto přísných a statických pravidel je vhodnější zavést rámce, které umožňují přizpůsobení novým výzvám. Například Evropská unie navrhuje regulaci AI založenou na riziku, která rozlišuje aplikace podle míry jejich potenciálního dopadu na společnost a zavádí odpovídající úroveň dohledu (Floridi, 2019).

Důležitou součástí rovnováhy mezi inovací a regulací je také vzdělávání a osvěta. Uživatelé generativních technologií by měli být informováni o jejich možnostech i omezeních a schopni posoudit rizika spojená s jejich využíváním. Zároveň je nezbytné podporovat výzkum, který propojuje technologické inovace s etickými a společenskými otázkami. Spolupráce mezi vývojáři, akademiky, regulačními orgány a veřejností je klíčová pro zajištění odpovědného vývoje generativní AI.

Rovnováha mezi inovací a regulací je komplexní otázkou, která vyžaduje multidisciplinární přístup. Řešení tohoto problému by mělo směřovat k vytvoření prostředí, které umožní generativní AI rozvíjet se odpovědným způsobem a zároveň minimalizovat rizika spojená s jejím neregulovaným využíváním.

1.3.4 Etická odpovědnost vývojářů

Vývojáři generativní umělé inteligence nesou klíčovou odpovědnost za to, jakým způsobem budou jejich technologie ovlivňovat společnost. Vytváření a nasazení generativních modelů, jako je GPT nebo BERT, má potenciál přinést výrazné přínosy, ale také generovat nové výzvy a rizika, která nelze ignorovat. Etická odpovědnost vývojářů spočívá v prevenci negativních dopadů, transparentnosti technologických procesů a respektování práv uživatelů (Floridi, 2019).

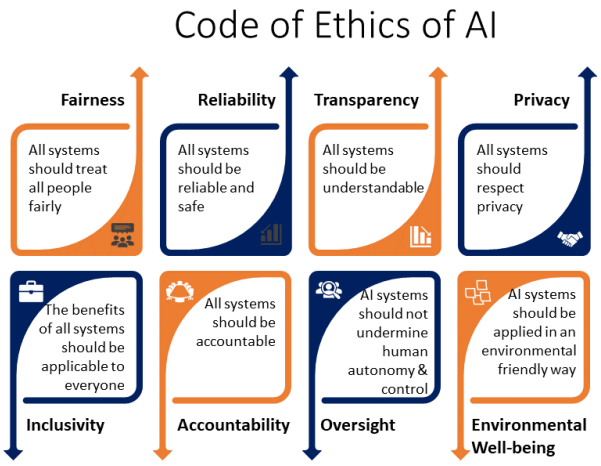
Jedním z hlavních principů etické odpovědnosti je minimalizace rizik spojených s nasazením generativní AI. Vývojáři by měli při navrhování modelů zohlednit možnosti jejich zneužití, například pro šíření dezinformací, kybernetické útoky nebo narušení soukromí. Preventivní opatření mohou zahrnovat implementaci detekčních mechanismů, které umožní rozpoznání generovaného obsahu, nebo zavedení omezení, která zabrání zneužívání modelů k neetickým účelům (Bostrom, 2014).

Dalším důležitým aspektem je transparentnost. Uživatelé by měli mít jasnou představu o tom, jak technologie fungují, jakým způsobem jsou zpracovávána data a jaké možnosti i limity generativní AI nabízí. Vývojáři by měli poskytovat dokumentaci a vysvětlení základních principů algoritmů, aby bylo možné porozumět jejich činnosti a zajistit jejich důvěryhodnost (Russell, et al., 2021). Transparentnost je zároveň klíčová pro budování důvěry mezi vývojáři, uživateli a regulačními orgány.

Ochrana uživatelských práv je dalším klíčovým bodem etické odpovědnosti. Vývojáři by měli respektovat soukromí a bezpečnost uživatelů, například tím, že zajistí ochranu citlivých dat použitých při tréninku modelů. V souvislosti je důležité dbát na dodržování právních a etických standardů, například v souladu s Obecným nařízením o ochraně osobních údajů (GDPR) v Evropské unii (Floridi, 2019).

Vývojáři mají rovněž odpovědnost za informování a vzdělávání veřejnosti. Uživatelé generativních nástrojů by měli být obeznámeni nejen s jejich přínosy, ale také s možnými riziky a omezeními. To zahrnuje osvětu o tom, jak rozpoznat generovaný obsah, jak hodnotit jeho důvěryhodnost a jak se vyhnout potenciálním rizikům spojeným s jeho používáním.

Etická odpovědnost vývojářů zahrnuje také podporu spravedlivého přístupu k těmto technologiím. Generativní AI by měla být navržena a implementována tak, aby byla přístupná různým skupinám uživatelů bez ohledu na jejich sociální nebo ekonomické postavení. Vývojáři by měli minimalizovat zaujatosti v tréninkových datech, které by mohly vést k diskriminaci nebo reprodukci předsudků.



Obr. 11: Základní principy etického kodexu AI

Zdroj: a3logics.com (2024)

Etická odpovědnost vývojářů je nedílnou součástí procesu vývoje generativní AI. Pouze přijetím odpovědnosti a implementací odpovídajících opatření lze zajistit, že technologie bude sloužit společnosti pozitivním způsobem a přispěje k udržitelnému rozvoji.

1.3.5 Vliv na společnost a trh práce

Generativní umělá inteligence má potenciál přetvořit mnoho aspektů společnosti, včetně způsobu, jakým pracujeme, komunikujeme a tvoříme. Díky schopnosti automatizovat úkoly, které dříve vyžadovaly kreativní či analytické schopnosti, generativní AI nejen rozšiřuje možnosti inovace, ale také přináší výzvy spojené s proměnou trhu práce a dopady na společenské normy.

V oblasti trhu práce představuje generativní AI zásadní faktor automatizace. Úlohy, které zahrnují rutinní textovou tvorbu, jako je psaní marketingových sdělení, právních dokumentů nebo analýz, mohou být plně nebo částečně automatizovány. To vede ke zvýšení produktivity, ale současně může způsobit zánik některých pracovních míst. Například pozice copywriterů nebo technických redaktorů mohou být ohroženy, protože generativní modely dokážou vytvářet obsah rychleji a levněji než lidé (Russell, et al., 2021).

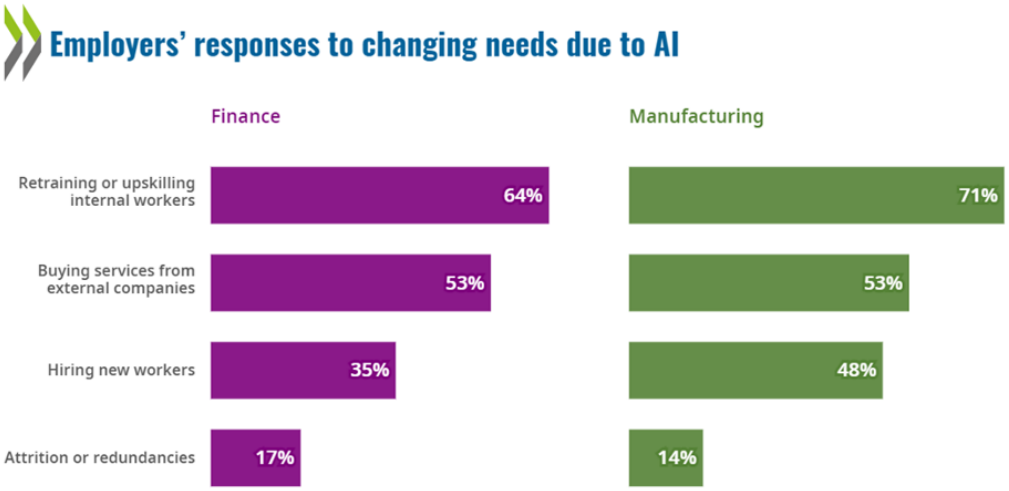
Na druhé straně vznikají nové příležitosti v oblastech, které vyžadují úzkou spolupráci mezi člověkem a AI. Role, jako je „editor AI generovaného obsahu“, „AI stratég“ nebo „vývojář

tréninkových dat“, nabývají na významu. Pozice kladou důraz na schopnost chápat a interpretovat generovaný obsah a zajistit, aby splňoval požadavky na kvalitu, přesnost a etiku. Organizace proto budou muset investovat do školení a rekvalifikace pracovníků, aby byli schopni efektivně využívat technologie (Floridi, 2019).

Generativní AI také ovlivňuje společenské normy a způsoby komunikace. Například schopnost modelů generovat texty na základě zadaných parametrů může změnit způsob, jakým lidé píší e- maily, připravují prezentace nebo vytvářejí vzdělávací materiály. Automatizace může přispět ke zlepšení efektivity, ale zároveň vyvolává otázky o ztrátě individuality a originality.

Pokud bude generativní AI široce využívána pro tvorbu osobní či profesionální komunikace, může dojít k homogenizaci stylu psaní, což by mohlo negativně ovlivnit kulturní rozmanitost (Bostrom, 2014).

Z pohledu společenského vlivu je klíčová otázka rovného přístupu k těmto technologiím. Organizace s dostatečnými finančními a technologickými zdroji mohou využívat generativní AI pro zlepšení své konkurenceschopnosti, zatímco menší subjekty mohou zaostávat. Technologická nerovnost by mohla dále prohloubit stávající rozdíly mezi ekonomickými regiony a sociálními skupinami. (Russell, et al., 2021).



Obr. 12: Reakce zaměstnavatelů ve financích a výrobě na změny způsobené AI
Zdroj: industrialrelationsnews.ioe-emp.org (2023)

Dalším významným aspektem je vliv generativní AI na vzdělávání a výzkum. Technologie, které umožňují automatizovanou analýzu textů, generování studijních materiálů nebo simulaci komplexních scénářů, mohou obohatit vzdělávací procesy. Zároveň však vyvolávají obavy z možného zneužití, například při podvádění v akademickém prostředí. Instituce budou muset zavést jasné směrnice pro používání AI v těchto kontextech, aby zajistily rovnováhu mezi inovací a etickým přístupem (Floridi, 2019). Generativní AI má výrazný potenciál ovlivnit společnost a trh práce. Zatímco automatizace a nové příležitosti přinášejí výhody, je nezbytné řešit výzvy spojené s nerovností, etickými otázkami a potřebou rekvalifikace pracovní síly. Spolupráce mezi vývojáři, zaměstnavateli a regulačními orgány bude klíčová pro zajištění, že technologie budou sloužit jako nástroj k pozitivní transformaci společnosti.

1.4 Budoucí trendy v generativní a detekční AI

Budoucnost generativní a detekční umělé inteligence slibuje významné pokroky v oblasti personalizace, efektivity a adaptace na měnící se potřeby společnosti. Generativní modely, jako je GPT-4 a jeho nástupci, budou nadále rozšiřovat své schopnosti, přičemž se zaměří na větší přizpůsobení individuálním potřebám uživatelů. Personalizace může zahrnovat možnost vytvářet obsah, který odpovídá preferencím konkrétního uživatele, včetně stylu psaní, tónu a zaměření. Vylepšení však vyvolávají otázky ohledně ochrany soukromí, protože modely by mohly analyzovat velké množství osobních dat, aby byly schopny takové personalizace dosáhnout. (Amodei, 2016).

Generativní modely současnosti vyžadují značné výpočetní zdroje, což představuje jak finanční, tak ekologickou zátěž. Budoucí výzkum se zaměří na vývoj menších, energeticky efektivnějších modelů, které si zachovají vysoký výkon. Optimalizace umožní širší dostupnost generativní AI, například pro malé podniky, vzdělávací instituce nebo jednotlivce. Současně bude nezbytné zavést regulační rámce, které zajistí odpovědné využití těchto technologií bez zbytečného zatížení životního prostředí (Russell, et al., 2021).

V oblasti detekčních technologií bude klíčovou výzvou přizpůsobivost nástrojů na stále sofistikovanější generativní modely. Detekční nástroje budoucnosti budou muset zahrnovat mechanismy samostatného učení, které jim umožní reagovat na nové vzorce generovaného obsahu v reálném čase. Multimodální detekce, která kombinuje textovou analýzu s detekcí vizuálního nebo zvukového obsahu, bude hrát významnou roli při identifikaci komplexních generovaných dat, jako jsou deepfake videa nebo syntetické obrazy. Technologie budou klíčové pro ochranu autenticity informací v digitálním prostředí (Floridi, 2019).

Důležitým směrem vývoje je integrace generativní a detekční AI do jednoho systému. Generativní modely by mohly být doplněny o vestavěné mechanismy, které automaticky označí obsah jako generovaný. Integrace by mohla zvýšit transparentnost a umožnit snadnější identifikaci původu dat. Společné využití technologií by mohlo přispět k větší důvěryhodnosti digitálního obsahu a minimalizaci rizika jeho zneužití. (Bostrom, 2014; Kirchenbauer, 2023).

Budoucí vývoj generativní a detekční AI bude rovněž úzce spjat s otázkami etiky a regulace. Globální standardy pro využívání těchto technologií by mohly snížit rizika spojená s jejich neregulovaným nasazením, například šíření dezinformací nebo narušování soukromí. Vývojáři a regulační orgány budou muset spolupracovat na vytvoření rámců, které umožní inovaci, ale zároveň zajistí ochranu práv uživatelů a společnosti. Informovanost veřejnosti bude také hrát klíčovou roli při zajištění odpovědného využívání těchto technologií (Amodei, 2016).

Budoucnost generativní a detekční AI je neoddělitelně spjata s technologickými, etickými a společenskými výzvami. Zatímco technologie přinášejí obrovský potenciál pro zlepšení mnoha aspektů lidského života, jejich rozvoj a implementace musí být doprovázeny odpovědným přístupem, který zajistí, že budou sloužit jako nástroj pozitivní změny.

1.4.1 Vývoj jazykových modelů

Jazykové modely představují základ generativní umělé inteligence a jejich vývoj byl v posledních letech mimořádně dynamický. Od počátečních statistických přístupů se jazykové modely vyvinuly k sofistikovaným architekturám hlubokého učení, jako jsou transformery, které přinesly

revoluci v oblasti zpracování přirozeného jazyka (NLP). Vývoj je charakterizován přechodem od jednoduchých modelů schopných základní analýzy textu ke komplexním systémům generujícím obsah, který je k nerozeznání od lidského psaní (Vaswani, a další, 2017).

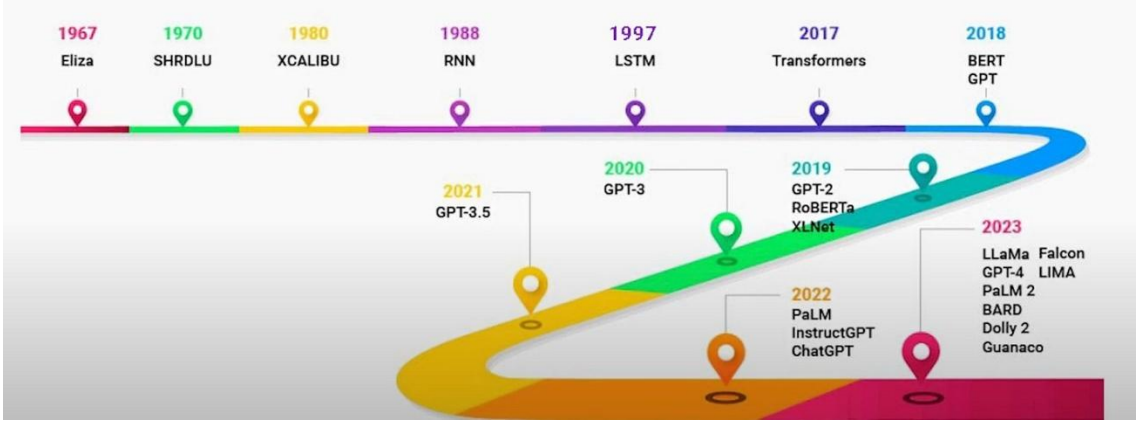
Jedním z klíčových milníků ve vývoji jazykových modelů bylo zavedení architektury transformerů, která byla představena ve studii Attention Is All You Need. Architektura přinesla koncept mechanizmů sebeorganizace (self-attention), což umožnilo modelům lépe chápat kontext slov v textu a jejich vzájemné vztahy. Díky tomu jsou transformery, jako je BERT (Bidirectional Encoder Representations from Transformers) a GPT (Generative Pre-trained Transformer), schopny generovat texty, které jsou stylisticky i obsahově přesvědčivé (Vaswani, a další, 2017).

Dalším významným pokrokem byl nárůst velikosti modelů a jejich tréninkových dat. Například GPT-3 od OpenAI obsahuje 175 miliard parametrů, což mu umožňuje pracovat s obrovským množstvím dat a generovat texty s vysokou mírou relevance a koherence. Velikost modelů však přináší i výzvy, zejména z hlediska výpočetních nákladů a ekologické stopy. Současný výzkum se proto zaměřuje na vývoj efektivnějších modelů, které by dosahovaly podobného výkonu s menšími zdroji (Brown, 2020).

Budoucnost jazykových modelů zahrnuje větší personalizaci a adaptabilitu. Modely budou schopny přizpůsobit své výstupy konkrétním uživatelům a jejich preferencím, což může přinést zlepšení uživatelského zážitku v oblastech, jako je zákaznická podpora, marketing nebo vzdělávání. Vývoj se také zaměří na zlepšení schopnosti modelů zpracovávat více jazyků, včetně těch, které jsou dosud nedostatečně zastoupeny v tréninkových datech, čímž se zvýší jejich inkluzivita (Amodei, 2016).

Dalším krokem bude zlepšení schopnosti modelů pracovat s multimodálními daty, což znamená kombinaci textu s jinými formami informací, jako jsou obrázky, zvuky nebo videa. Multimodální modely by mohly například generovat texty na základě vizuálních vstupů nebo naopak popisovat obrazové materiály pomocí přirozeného jazyka. Schopnosti by mohly najít uplatnění v oblastech, jako je zdravotnictví, kde by modely mohly analyzovat lékařské snímky a generovat srozumitelné zprávy pro lékaře a pacienty (Floridi, 2019).

Vývoj jazykových modelů je neoddělitelně spjat s otázkami etiky a odpovědnosti. S rostoucí schopností generovat realistický obsah přicházejí rizika zneužití, například šíření dezinformací nebo napodobování identity. Budoucí modely proto budou muset obsahovat mechanismy, které zajistí transparentnost a umožní detekci generovaného obsahu. Spolupráce mezi vývojáři, regulačními orgány a společností bude klíčová pro zajištění odpovědného využívání těchto technologií.



Obr. 13: evoluce jazykových modelů

Zdroj: medium.com (2024)

1.4.2 Budoucnost detekčních nástrojů

S rychlým rozvojem generativní umělé inteligence roste i potřeba efektivních detekčních nástrojů, které dokážou identifikovat obsah vytvořený těmito technologiemi. Budoucnost detekčních nástrojů bude charakterizována adaptabilitou, multimodálním přístupem a větší integrací s běžnými technologiemi. Trendy budou klíčové pro udržení rovnováhy mezi generativní a detekční AI a ochranu autenticity digitálního obsahu (Weber-Wulff, 2023).

Jedním z hlavních směrů vývoje bude zvýšení adaptivní schopnosti detekčních nástrojů. Současné nástroje, jako je DetekceGPT nebo Isgen, jsou navrženy pro identifikaci vzorců typických pro současné generativní modely, například GPT-3 nebo GPT-4. Budoucí nástroje však budou muset reagovat na stále sofistikovanější modely, které lépe napodobují lidské psaní a skrývají stopy generace. Adaptivní přístupy založené na mechanismu samostatného učení umožní detekčním algoritmům přizpůsobit se novým technologiím bez nutnosti manuálního přeprogramování (Amodei, 2016).

Dalším trendem je rozvoj multimodálních detekčních nástrojů. Budoucí systémy nebudou analyzovat pouze text, ale i vizuální a zvukový obsah generovaný AI, například deepfake videa nebo syntetické obrazy. Kombinace těchto analýz umožní vytvoření komplexních systémů, které dokážou identifikovat manipulace napříč různými médii. Multimodální přístup bude hrát klíčovou roli při ochraně před šířením dezinformací, které kombinují různé typy generovaného obsahu (Floridi, 2019; Hu, 2024).

Integrace detekčních nástrojů do běžných aplikací je další perspektivou budoucího vývoje. Nástroje pro detekci AI generovaného obsahu by mohly být zabudovány do textových editorů, platforem sociálních sítí nebo vyhledávačů, což by umožnilo okamžité varování uživatelům při detekci podezřelého obsahu. Integrace by přispěla k vyšší transparentnosti a důvěryhodnosti digitální komunikace. Současně by nástroje mohly být použity v akademickém prostředí pro detekci plagiátorství nebo v mediální sféře pro ověřování autenticity zpráv (Weber-Wulff, 2023).

Dalším důležitým aspektem budoucnosti detekčních nástrojů je minimalizace falešných pozitivních a negativních výsledků. Vyšší přesnost algoritmů a standardizace hodnoticích metrik umožní vytvoření spolehlivějších systémů, které budou schopny přesně identifikovat

generovaný obsah. Pokrok je nezbytný, zejména v právním nebo akademickém prostředí, kde chybné detekce mohou mít vážné důsledky (Russell, et al., 2021).

Budoucnost detekčních nástrojů bude rovněž ovlivněna potřebou etických a regulačních rámců. Spolupráce mezi vývojáři, regulačními orgány a akademickou sférou bude klíčová pro vytvoření nástrojů, které jsou nejen technologicky pokročilé, ale také transparentní a spravedlivé. Etický design by měl zahrnovat požadavek na vysvětlitelnost detekčních algoritmů, což zvýší důvěru uživatelů a usnadní implementaci v širokém spektru aplikací (Amodei, 2016).

Detekční nástroje budou v budoucnosti nezbytným prvkem digitálního ekosystému. Jejich vývoj musí reagovat na stále složitější generativní technologie a zároveň zohledňovat potřebu ochrany uživatelů před zneužitím. Kombinace adaptivních algoritmů, multimodálních přístupů a eticky odpovědného designu zajistí, že detekční technologie zůstanou efektivním nástrojem při ochraně autenticity a integrity informací.

1.4.3 Integrace generativních a detekčních AI

Integrace generativní a detekční umělé inteligence představuje perspektivní směr vývoje, který by mohl zásadně změnit způsob, jakým jsou technologie navrhovány a implementovány. Spojení těchto dvou přístupů by mohlo přinést efektivnější a transparentnější systémy, které nejen generují obsah, ale zároveň umožňují jeho autentifikaci a kontrolu. Synergie je klíčová zejména ve světě, kde je generativní AI stále sofistikovanější a detekce falešného obsahu složitější (Amodei, 2016).

Jedním z hlavních cílů integrace je zavedení mechanismů, které umožní generativním modelům automaticky označit svůj výstup jako generovaný. Funkce by mohla být implementována prostřednictvím tzv. vodotisků (watermarking), které by byly neviditelně začleněny do textu a identifikovatelné detekčními nástroji. Vodotisky by mohly obsahovat informace o použitém modelu, datovém zdroji nebo časovém razítku generace. Tím by bylo možné jednoduše ověřit původ obsahu a minimalizovat riziko jeho zneužití (Floridi, 2019).

Další možností integrace je vývoj modelů, které kombinují generativní a detekční funkce. Hybridní systémy by mohly být schopny nejen vytvářet obsah, ale také analyzovat jeho pravděpodobnou autentičnost v reálném čase. Například modely určené pro automatizovanou zákaznickou podporu by mohly během interakce nejen odpovídat na dotazy uživatelů, ale zároveň monitorovat, zda generované odpovědi splňují stanovené etické a kvalitativní normy (Bostrom, 2014; Kirchenbauer, 2023).

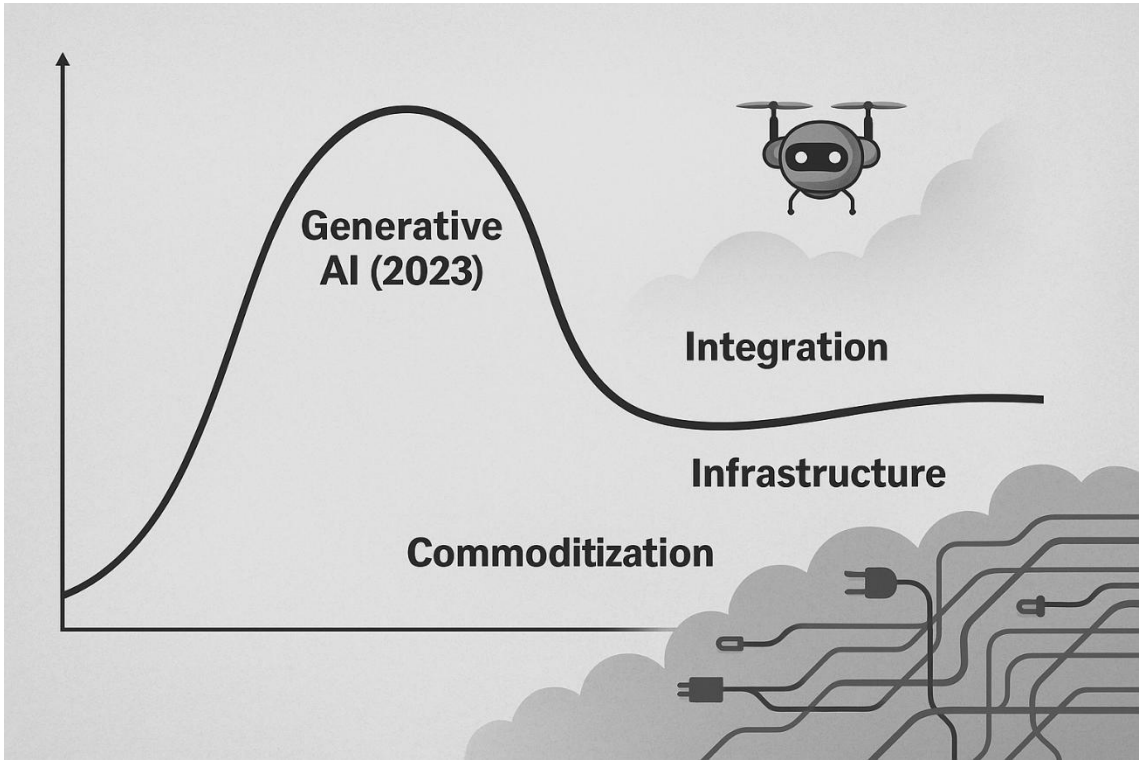
Integrace generativní a detekční AI má také významné aplikace ve vzdělávání a výzkumu. Generativní modely mohou být využity k vytváření studijních materiálů nebo simulací, zatímco detekční funkce by mohly zajistit, že vytvořený obsah odpovídá požadovaným standardům Kombinace by mohla pomoci vzdělávacím institucím lépe zvládnout výzvy spojené s používáním generativní AI studenty, například při kontrole originality jejich prací (Weber-Wulff, 2023).

Významným přínosem integrace je rovněž zvýšení důvěryhodnosti digitálního obsahu. Generativní AI je stále častěji využívána v médiích, marketingu a dalších oblastech, kde je klíčová autenticita informací. Spojení s detekčními nástroji by umožnilo vytvářet transparentní systémy, které nejen generují obsah, ale také poskytují informace o jeho původu a kvalitě. Tím by se snížila

pravděpodobnost šíření dezinformací a zvýšila důvěra veřejnosti v generované texty (Amodei, 2016).

Integrované systémy generativní a detekční AI by také mohly podpořit etický vývoj těchto technologií. Kombinace generace a detekce v jednom modelu by vytvořila zpětnovazební smyčku, která by umožnila rychle identifikovat a opravit chyby, předsudky nebo neetické chování algoritmů. Přístup by přispěl k vývoji odpovědných a spravedlivých systémů, které budou sloužit ve prospěch celé společnosti (Floridi, 2019).

Budoucí integrace generativní a detekční AI slibuje vytvoření holistických systémů, které budou efektivně kombinovat tvorbu obsahu s jeho autentifikací. Přístup má potenciál nejen zvýšit důvěryhodnost digitální komunikace, ale také podpořit etické a odpovědné využívání umělé inteligence v různých oblastech lidské činnosti.



Obr. 14: očekávaný vývoj generativní AI
Zdroj: marigold.cz (2025)

1.4.4 Role AI ve vzdělávání a výzkumu

Umělá inteligence, včetně generativních a detekčních modelů, hraje stále významnější roli ve vzdělávání a výzkumu. Její aplikace přinášejí nové příležitosti pro zlepšení výukových procesů, zvýšení efektivity vědecké práce a zpřístupnění komplexních poznatků širšímu publiku. (Wu, 2023) Současně vyvolávají otázky týkající se etiky, originality a vlivu na vzdělávací standardy (Chollet, 2018; Large Language Models in Education: A Systematic Review, 2024).

Ve vzdělávání umožňují generativní modely personalizaci výukových materiálů a tvorbu obsahu přizpůsobeného individuálním potřebám studentů. Například generativní AI může vytvářet studijní texty, shrnutí odborné literatury nebo interaktivní scénáře, které simulují reálné situace. Nástroje podporují aktivní učení a pomáhají studentům lépe pochopit složité koncepty. Zároveň

mohou být využity pro automatizovanou kontrolu domácích úkolů nebo poskytování zpětné vazby, což šetří čas učitelům a umožňuje jim zaměřit se na kreativní aspekty výuky (Russell, et al., 2021; Large Language Models in Education: A Systematic Review, 2024).

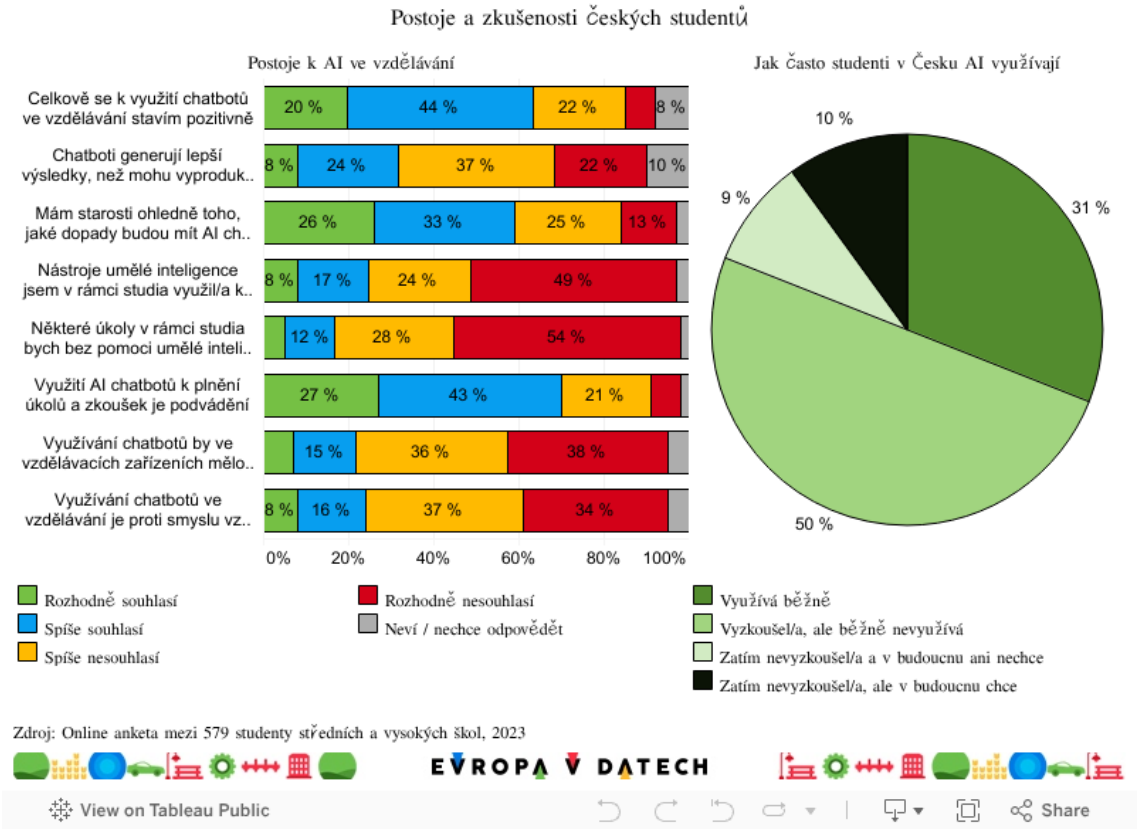
Generativní AI se rovněž stává nepostradatelným nástrojem ve výzkumu. Modely, jako je GPT, dokážou analyzovat rozsáhlé množství dat, identifikovat vzorce a generovat hypotézy, které mohou být dále ověřovány. Automatizace rutinních úkolů, jako je shromažďování literatury, příprava výzkumných zpráv nebo analýza dat, umožňuje výzkumníkům soustředit se na tvůrčí a analytické aspekty své práce. Generativní modely také podporují interdisciplinární spolupráci tím, že zjednodušují komunikaci mezi odborníky z různých oborů (Floridi, 2019).

Na druhé straně detekční AI pomáhá řešit výzvy spojené s originalitou a etickým využíváním generativních nástrojů. Ve vzdělávacím prostředí může být využita k detekci plagiátorství nebo podvodů při vypracovávání úkolů a prací. Detekční nástroje také umožňují ověřit autenticitu výzkumných dat a eliminovat riziko falešných výsledků, což je zásadní pro udržení vědecké integrity (Weber-Wulff, 2023).

Přes své přínosy přináší integrace AI do vzdělávání a výzkumu také výzvy. Jednou z nich je riziko, že nadměrné spoléhání na generativní AI může omezit kreativitu studentů nebo výzkumníků. Pokud jsou studenti příliš závislí na automatizovaných systémech, mohou ztratit schopnost kritického myšlení a manuálního řešení problémů. Podobně ve výzkumu hrozí, že generativní AI bude nekriticky přijímána jako autoritativní zdroj, což by mohlo vést k nekontrolovanému šíření nepřesných nebo zkreslených informací (Chollet, 2018).

Budoucnost role AI ve vzdělávání a výzkumu spočívá v nalezení rovnováhy mezi využíváním těchto technologií a zachováním tradičních přístupů k učení a vědecké práci. Důraz by měl být kladen na informování a vzdělávání uživatelů o možnostech i omezeních generativní a detekční AI, stejně jako na vytvoření etických rámců, které zajistí odpovědné využívání těchto nástrojů.

Integrace AI do vzdělávání a výzkumu má potenciál přinést revoluční změny, ale její implementace musí být prováděna s ohledem na potřeby studentů, učitelů a vědců. Kombinace inovací s důrazem na etiku a transparentnost zajistí, že technologie budou podporovat pozitivní rozvoj klíčových oblastí.



Obr. 15: Postoj k AI ve vzdělávání od českých studentů

Zdroj: evropavdatech.cz (2025)

1.4.5 Výzvy budoucnosti

Budoucnost generativní a detekční umělé inteligence přináší nejen příležitosti, ale také řadu výzev, které budou klíčové pro jejich efektivní a odpovědné nasazení. Výzvy zahrnují technologické, etické, právní i společenské aspekty, přičemž jejich řešení bude vyžadovat spolupráci mezi vývojáři, akademiky, regulačními orgány a veřejností.

Jednou z největších technologických výzev je neustále se zlepšující schopnost generativních modelů napodobovat lidskou tvorbu. Jak se generativní modely stávají sofistikovanějšími, detekční nástroje musí být schopny držet krok. Současné metody detekce se často spoléhají na pravděpodobnostní vzory nebo syntaktické anomálie, ale budoucí modely mohou vzory skrývat, což identifikaci ztíží. Vývoj adaptivních detekčních nástrojů, které se budou schopny samostatně učit a přizpůsobovat novým generativním technologiím, bude proto zásadní (Amodei, 2016).

Dalším problémem je otázka zaujatosti a předsudků ve vstupních datech, která generativní AI využívá. Tréninkové datové sady mohou obsahovat předsudky, které se následně promítají do generovaného obsahu. Problém je obzvláště citlivý v oblastech, jako je zdravotnictví, vzdělávání nebo právo, kde mohou zkreslené modely způsobit reálné škody. Budoucí vývoj bude muset zahrnovat robustnější metody pro čištění dat a eliminaci předsudků (Floridi, 2019).

Rovněž je důležité zajistit, aby technologie byly přístupné pro všechny. Technologická nerovnost může vést k situaci, kdy velké korporace nebo vyspělé země budou mít přístup k pokročilým generativním a detekčním modelům, zatímco menší organizace nebo rozvojové země budou

zaostávat. Podpora otevřených datových sad a volně dostupných nástrojů může pomoci zmírnit rozdíly a zajistit širší dostupnost technologií. (Bostrom, 2014).

Významnou výzvou jsou i etické otázky, jako je šíření dezinformací a narušení soukromí. Generativní AI může být zneužita k vytváření falešných zpráv, podvodných reklam nebo manipulativního obsahu, což vyvolává otázky odpovědnosti. Regulace, která by definovala, kdo je odpovědný za zneužití generovaných textů, a zavedení standardů pro transparentnost, například prostřednictvím automatického označování generovaného obsahu, budou zásadní pro prevenci těchto problémů (Weber-Wulff, 2023).

V neposlední řadě je zde otázka ekologické udržitelnosti. Velké generativní modely, jako je GPT-4, vyžadují obrovské výpočetní zdroje, což přispívá k vysoké energetické náročnosti a ekologické zátěži. Budoucí výzkum se bude muset zaměřit na vývoj efektivnějších algoritmů a optimalizaci výpočetních procesů, aby se minimalizoval dopad těchto technologií na životní prostředí (Chollet, 2018).

Přestože generativní a detekční AI mají potenciál přinést revoluční změny do mnoha odvětví, budoucí vývoj bude vyžadovat odpovědný přístup k řešení výzev. Kombinace technických inovací, robustních regulačních rámců a zvýšené informovanosti uživatelů bude klíčem k tomu, aby technologie sloužily jako nástroj pro pozitivní změny, aniž by ohrožovaly etiku, spravedlnost a udržitelnost (Weber-Wulff, 2023).

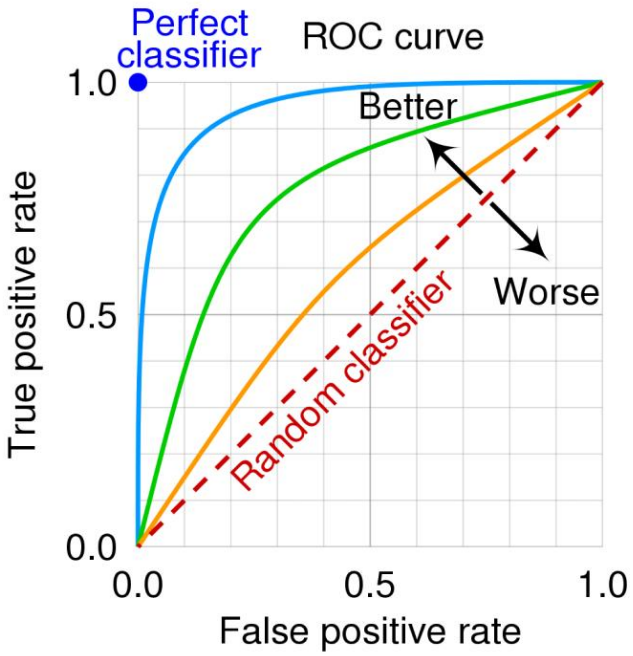
1.5 Principy identifikace AI-generovaných textů a jejich vztah k testování

Tato kapitola shrnuje základní principy, na kterých stojí detekce AI-generovaných textů, a vysvětluje, jak se promítají do konkrétního testování nástroje GPTZero na mém datasetu. Slouží jako most mezi teoretickou částí a kapitolou 4, kde jsou prezentovány detailní výsledky.

1.5.1 Základní principy detekce: perplexita a „burstiness“

Většina současných detektorů AI textu, včetně GPTZero, vychází z předpokladu, že jazykový projev člověka je z hlediska pravděpodobnosti „divočejší“ než text generovaný velkým jazykovým modelem. Prakticky se sledují hlavně dvě kvantitativní veličiny:

- **Perplexita** – míra toho, jak je text pro jazykový model předvídatelný.
    - Texty generované AI bývají předvídatelnější, model „sází na jistotu“, a mají tedy nižší perplexitu.
    - Lidské texty častěji obsahují neobvyklé formulace, odskoky v argumentaci nebo stylu → vyšší perplexita.
- **Variabilita vět** – kolísání mezi větami, například v délce a komplexitě.
    - U lidí bývá rytmus textu nerovnoměrný: střídají krátké a dlouhé věty, různý stupeň formálnosti, někde přidají komentář bokem.
    - AI nechává věty často podobně dlouhé a podobně vystavěné, takže variabilita je menší.



Obr. 16: ROC křivka – porovnání ideálního, náhodného a lepšího/horšího klasifikátoru

Zdroj: medium.com (2025)

Detektor pak na základě těchto a dalších stylometrických znaků přibližuje pravděpodobnost, že text odpovídá „profilu AI“, nebo „profilu člověka“. Výsledek se nakonec zjednoduší na výstup typu *„AI-generated“* vs. *„human-written“* (případně s nějakým stupňováním jistoty). (Mitchell, 2023)

1.5.2 Diskurzní a lexikální markety (kvalitativní, kvantifikované)

Diskurzní a lexikální markery představují rysy, které jsou dobře viditelné i při běžném čtení textu. Jde například o způsob strukturování odstavců, užívání spojovacích výrazů, typickou slovní zásobu, opakování frází nebo práci s metatextem (výrazy typu „shrňme si“, „v této části se zaměřím na“). V literatuře se shrnují jako stylometrické rysy – tedy charakteristický jazykový „otisk“ autora, který může být u člověka odlišný od relativně hladkého, homogenního stylu velkých jazykových modelů. (Wu, 2025)

Studie zaměřené na detekci LLM-generovaných textů ukazují, že strojové výstupy mívají omezenější lexikální rozmanitost, častěji využívají frekventovaná slova a šablonovité obraty a opakují podobné struktury vět napříč odstavci. LLM mají také tendenci nadužívat některé druhy funkčních slov (například spojky a pomocná slovesa) a preferovat standardizované a symetrické argumentační bloky. Naopak lidské texty častěji pracují s lokálními odbočkami, elipsami, nečekanými metaforami či změnami rytmu vyprávění, což se v literatuře popisuje jako větší diskurzní „nerovnoměrnost“ nebo kreativita.

Tyto vlastnosti lze popisovat čistě kvalitativně (například pomocí ručního kódování typů spojovacích výrazů, stereotypních formulací nebo metaforických pasáží), ale současně je možné je do jisté míry kvantifikovat. Stylometrie běžně pracuje s ukazateli, jako je typově–tokenové ratio (poměr počtu různých slov k celkovému počtu slov), četnost hapax legomenon (slov, která se v textu vyskytují jen jednou), podíl různých slovních druhů nebo hustota explicitních

diskurzních markerů. V kontextu této práce je možné interpretovat lepší detekovatelnost čistě AI-generovaných literárních textů tak, že vykazují nápadně uniformní slovní zásobu, častější užívání generických hodnotících výrazů („velmi důležité“, „zásadní roli“) a systematické opakování obdobných syntaktických šablon, zatímco lidské povídky a eseje jsou lexikálně pestřejší a „rozbitější“ v rytmu.

Zvláštním případem jsou post-editované texty (AI-EDIT), které kombinují obsahovou kostru a část lexika generovaného modelem s následnými úpravami člověka. Empirické studie ukazují, že právě tyto „hybridní“ texty mohou být pro detektory obzvlášť obtížné: lokální zásahy do slovní zásoby, přeskládání odstavců a doplnění individuálních poznámek narušují typický strojový profil, ale zcela jej neodstraňují. V českém datasetu této práce se to projevuje například u administrativních a akademických textů, kde post-editované dokumenty často unikají detekci, přestože v nich zůstávají stopy původního AI stylu.

1.5.3 Kvantitativní rysy (měřitelné)

Vedle diskurzních a lexikálních markerů hrají důležitou roli kvantitativní rysy, které lze přímo měřit a zpracovávat statisticky nebo pomocí strojového učení. Patří sem zejména různé délkové charakteristiky (průměrná délka věty, odchylka délky vět, počet odstavců), ukazatele složitosti (hloubka syntaktických stromů, míra vnoření vět, četnost vedlejších vět) a pravděpodobnostní charakteristiky, jako je perplexita a burstiness, na nichž výslovně staví i GPTZero. (Agrahari, a další, 2024)

Současné přehledové studie o detekci LLM-generovaných textů uvádějí, že efektivní detektory často kombinují několik skupin rysů: jednodušší „povrchové“ statistiky, stylometrické ukazatele, pravděpodobnostní rysy odvozené z jazykových modelů a někdy i zpětnou vazbu jiných modelů (například skóre, které textu přiřadí další LLM). Takový přístup umožňuje zachytit nejen to, jak je text formálně složitý, ale také to, jak moc odpovídá typickým „hladkým“ pravděpodobnostním vzorcům velkých modelů.

V podmínkách této práce je část těchto kvantitativních rysů přístupná jen nepřímo, protože GPTZero funguje jako uzavřený systém a neposkytuje detailní rozpad použitých metrik. Z literatury je však zřejmé, že nástroj spoléhá právě na kombinaci perplexity, burstiness a dalších stylometrických ukazatelů, přičemž podobně jako jiné detektory vykazuje citlivost na délku textu a míru obfuskace. Zjištění této práce – například téměř nulová detekce akademických AI-GEN textů a naopak vysoká úspěšnost u literárních AI-GEN textů – lze proto interpretovat tak, že některé domény (silně normovaný akademický styl) jsou z hlediska těchto měřitelných rysů „snadno napodobitelné“, zatímco jiné (literární narativ) ponechávají více prostoru pro odlišný kvantitativní profil lidského a strojového psaní.

Z metodického hlediska je možné navázat tím, že by se pro český dataset, který tato práce představuje, v budoucím výzkumu přímo dopočítaly vybrané kvantitativní rysy – například rozptyl délky vět, entropie rozdělení slov, lexikální bohatství podle klasických indexů a základní ukazatele syntaktické složitosti – a ty by se porovnaly s chováním GPTZero. Taková analýza by umožnila lépe pochopit, které konkrétní měřitelné rysy stojí za vysokou precision, ale nízkým recall nástroje v různých doménách, a mohla by sloužit jako podklad pro návrh robustnějších postupů detekce v českém akademickém prostředí. (Automatic Detection of AI-Generated Text from LLMs Using Feature-Driven Transformer Networks, 2025)

1.5.4 Očekávání a limity detektorů

Z těchto principů plynou dvě důležitá obecná očekávání:

1. Detektor bude dobře fungovat u typických, málo upravených AI textů, které mají silně standardizovaný a homogenní styl.
2. Problém bude mít naopak tam, kde:
    - i lidské texty jsou velmi normované a předvídatelné (právnické, administrativní, akademické texty),
    - AI text někdo upraví (post-editace, parafrázování) nebo zkrátí/rozkouskuje.

Empirické studie skutečně ukazují, že přesnost těchto detektorů zásadně kolísá podle jazyka, žánru a míry editace textu. V některých scénářích umí dobře chytat „čisté“ AI výstupy, v jiných selhávají téměř úplně – často zejména u odborných a silně formalizovaných textů nebo u krátkých úryvků.

1.5.5 Aplikace na testovaný dataset

Tyto obecné principy se velmi výrazně odrážejí i v tom, jak si GPTZero vedl na datasetu, který je rozdělen do tří domén:

- administrativní/formální texty,
- akademické/vědecké texty,
- literární a esejistické texty,

a ve třech třídách původu:

- HUMAN-AUTH– lidské texty,
- AI-EDIT – texty generované AI a následně upravené člověkem,
- AI-GEN – texty plně generované AI.

V administrativní doméně jsou lidské texty i AI texty velmi normované a formální. GPTZero zde téměř vůbec nedělá falešné poplachy u lidí, ale zároveň podhodnocuje přítomnost AI, zejména u 100% AI textů, které často považuje za lidské. V akademické/vědecké doméně jde tento problém ještě dál: nástroj prakticky všechny texty (včetně 100% AI) vyhodnocuje jako lidské. Akademický styl je pro detektor typ „dobře napsaného a předvídatelného“ textu, takže rozdíl mezi AI a člověkem mizí.

V literární doméně je situace odlišná: lidské povídky a eseje mají výraznější rytmus, nápadité metafory a lexikální rozptyl; naopak čistě AI generované texty jsou relativně stereotypní. GPTZero zde relativně úspěšně identifikuje AI-GEN, ale post-editované texty (AI-EDIT) se často „schovají“ mezi lidskými.

Celkově výsledky potvrzují dvě klíčové teze pro další části práce:

- Detekce AI textu je doménově velmi citlivá – stejné nástroje, které fungují uspokojivě u narativních textů, mohou být téměř nepoužitelné u akademických a administrativních.

- Největší problém z hlediska akademické integrity představují post-editované texty (AI-EDIT): vycházejí z AI, ale po zásahu člověka ztrácejí pro detektor stabilní „strojový podpis“. Přitom právě tento způsob práce s generativní AI je pro studenty nejrealističtější.

2 Metodika

Metodická část navazuje na teoretický rámec generativní a detekční umělé inteligence. Jejím cílem je popsat, jakými postupy je v práci zkoumána rozpoznatelnost AI textů v českém akademickém prostředí. Nejprve je stručně představen návrh datasetu a jeho základní parametry, poté použitý detekční nástroj a způsob testování. Závěrečná část metodiky se soustředí na kvantitativní ukazatele, které umožňují vyhodnotit úspěšnost detekce. Konkrétní realizace těchto postupů (výběr textů, generování AI výstupů a praktický průběh skenování) je podrobně popsána v kapitole 3.

2.1 Návrh datasetu pro identifikaci textů generovaných umělou inteligencí

Tato část popisuje koncepci strukturovaného datasetu textů určeného k testování možností automatické detekce generativní umělé inteligence. Dataset je navržen jako kombinace tří textových domén (administrativní/formální, akademická/vědecká, literární/esejistická) a tří tříd původu textu: lidské texty, plně AI generované texty a texty vzniklé post-editací AI výstupu člověkem.

2.1.1 Třídy původu textu

Každý text v datasetu je zařazen do jedné ze tří tříd původu:

- HUMAN-AUTH (lidský text) – texty vytvořené člověkem bez využití generativní umělé inteligence. Vznikaly standardním způsobem (psaní autorem), případně s běžnými jazykovými úpravami (pravopisná kontrola apod.).
- AI-EDIT (AI text s následnou editací) – texty, jejichž základní verze byla vygenerována nástrojem umělé inteligence (ChatGPT), ale následně je člověk upravil. Úpravy se týkaly zejména stylu, struktury a někdy i významového obsahu. Cílem zařazení této třídy je simulovat situaci, kdy student využije AI jako *výchozí návrh* a text si následně přizpůsobí.
- AI-GEN (plně generovaný AI) – texty, které byly vygenerovány nástrojem umělé inteligence a nebyly dále významně editovány. Mohlo dojít pouze k minimálním technickým úpravám (např. odstranění nadbytečných prázdných řádků), nikoli však ke stylistickému nebo obsahovému zásahu.

2.1.2 Kritéria výběru a velikost vzorku

Ve všech doménách byla dodržena stejná velikost vzorku (20 textů na třídu). Při výběru textů byla uplatněna tato základní kritéria:

- Jazyk: všechny texty jsou v češtině;
- Délka textu: texty jsou přibližně srovnatelné délkou, aby nedocházelo k systematickému zkreslení ve prospěch kratších nebo delších textů;

- Aktuálnost a tematika: témata byla volena tak, aby odpovídala současnému vysokoškolskému prostředí (studijní agenda, akademické úkoly, běžné narativní motivy);
- Srozumitelnost: do datasetu nebyly zařazeny texty s výrazně porušenou srozumitelností (např. zcela nedokončené nebo technicky poškozené dokumenty).

2.2 Testování datasetu nástrojem GPTZero

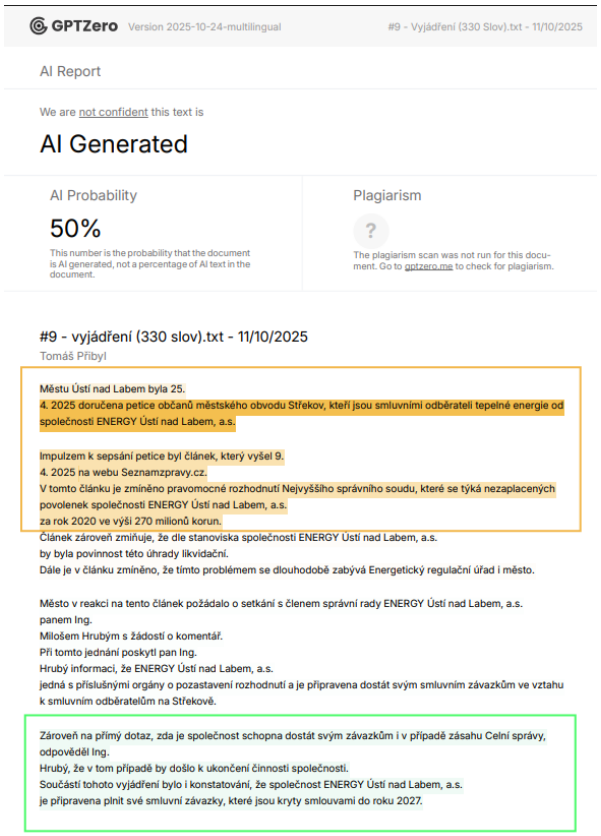
Pro vyhodnocení toho, jak snadno jsou jednotlivé texty datasetu rozpoznatelné jako lidské či generované umělou inteligencí, byl použit online nástroj GPTZero. Ten umožňuje nahrát vlastní soubory nebo texty a na základě interního modelu odhadnout, zda byl text pravděpodobně vytvořen člověkem, nebo generativní AI.

V této práci je GPTZero chápán jako reprezentant běžně dostupných detekčních nástrojů, se kterými mohou pracovat i vysoké školy. Dataset je testován v režimu dávkového zpracování (batch scan) a pro každý text je zaznamenán výsledek klasifikace. Praktická organizace skenování (předplatné, práce s kredity, import souborů) je popsána v kapitole 3.2.

2.2.1 Režimy skenování: basic vs advanced scan

Nástroj GPTZero nabízí více způsobů, jak text analyzovat. Z hlediska této práce je podstatný rozdíl mezi:

- základním (standardním) skenem, který poskytuje především binární či stupňovitý odhad, zda je text spíše „AI“ nebo „human“, případně jednoduché skóre či shrnutí;
- pokročilým (advanced) skenem, který nabízí detailnější analýzu, obvykle včetně:
    - jemnějšího skóre pravděpodobnosti či míry „AI-like“ charakteru textu,
    - možnosti dávkového zpracování většího množství textů,
    - podrobnějšího reportu pro další zpracování (např. export výsledků).



Obr. 17: Ukázka reportu z advanced scanu GPTZero

Zdroj: vlastní zpracování (2025)

V této práci byl na celý dataset aplikován režim advanced scan, a to z následujících důvodů:

- umožňuje systematické zpracování velkého množství souborů v jedné dávce,
- poskytuje výstupy, které lze snadno převést do tabulkové podoby a dále z nich vypočítat metriky (matice záměn, přesnost, recall apod.),
- dává detailnější informace, než by bylo dostupné z čistě základního skenu.

Standardní režim byl využit pouze v přípravných fázích (např. pro ověření funkčnosti nástroje a orientační vyzkoušení rozhraní) a není dále zahrnut do kvantitativního vyhodnocení.

2.2.2 Možnosti a limity zvoleného postupu

Zvolený postup testování má několik výhod, ale i omezení, která je nutné v závěru reflektovat:

- **Výhody**
  - konzistence: všechny texty byly zpracovány stejným nástrojem, ve stejném režimu a v krátkém časovém období, což minimalizuje vliv případných změn modelu nebo nastavení na straně poskytovatele;
  - reálnost scénáře: použit byl standardně dostupný komerční nástroj, který mohou využívat i vysoké školy, nikoli experimentální výzkumný model.
- **Omezení**
  - kreditový limit a časově omezené předplatné neumožnily provést rozsáhlejší sérii experimentů, například porovnání výsledků v různých režimech skenování,

opakované skenování týchž textů nebo systematické testování vlivu délky či postupných úprav textu;

- výsledky odrážejí konkrétní podobu GPTZero v době měření a nelze je bez dalšího zobecňovat na všechny detekční nástroje nebo budoucí verze tohoto systému.

2.3 Metody vyhodnocení

Tato část popisuje, jak byly výstupy nástroje GPTZero převedeny do podoby kvantitativních ukazatelů. Cílem je umožnit:

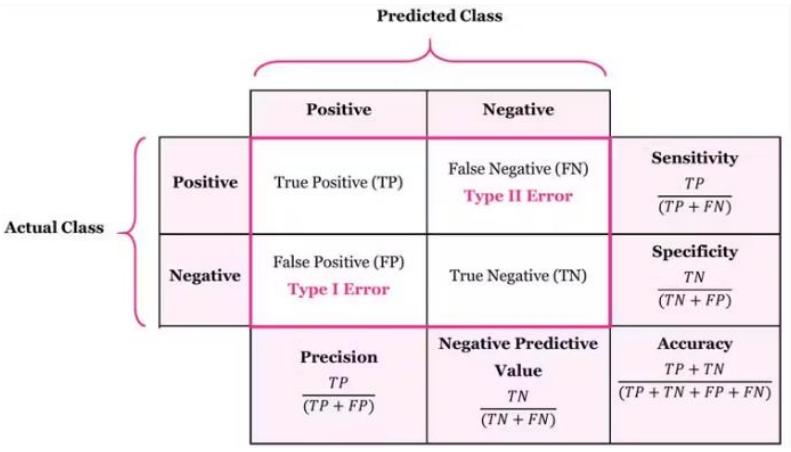
- porovnat úspěšnost detekce napříč třemi doménami (administrativní, akademická, literární),
- sledovat rozdíly mezi lidskými texty (HUM) a texty s podílem umělé inteligence (AI-EDIT, AI-GEN),
- analyzovat chování nástroje vůči čistě AI textům a post-editovaným textům.

Základním nástrojem pro vyhodnocení jsou matice záměn, z nichž jsou následně odvozeny standardní metriky typu accuracy, precision, recall, F1-score a vybrané dílčí ukazatele (např. podíl lidských textů chybně označených jako AI).

2.3.1 Konstrukce matic záměn

Pro každý text v datasetu jsou k dispozici dvě informace:

- „skutečná“ třída (ground truth), tj. jedno z označení:
  - HUM – lidský text,
  - AI-EDIT – text generovaný AI a následně editovaný člověkem,
  - AI-GEN – text plně generovaný AI,
- výsledek klasifikace GPTZero, tj. to, jak nástroj text vyhodnotil (např. jako „human-written“ nebo „AI-generated“).



Obr. 18: Matice záměn pro binární klasifikaci a základní odvozené metriky.
Zdroj: encord.com (2025)

Protože GPTZero v běžném nastavení nerozlišuje mezi AI-EDIT a AI-GEN, pracuje výsledná klasifikace primárně s binárním rozdělením na:

- „lidské“ (human),
- „AI“ (tady v interpretaci spadají jak AI-EDIT, tak AI-GEN).

V této práci tedy matice záměn sledují především schopnost nástroje rozlišit „čistě lidské“ texty od textů s podílem AI, přičemž pro interpretaci se zvlášť dívám na to, jak si vede u AI-EDIT a AI-GEN.

Základní binární matice

Pro každou ze tří domén (D1, D2, D3) byla sestavena 2×2 matice záměn, kde:

- řádky odpovídají skutečné třídě (HUM vs. AI = AI-EDIT + AI-GEN),
- sloupce odpovídají třídě přiřazené GPTZero (predikce).

V každé matici tak vystupují čtyři hodnoty:

- TP (true positives) – počet textů, které jsou ve skutečnosti AI (AI-EDIT nebo AI-GEN) a GPTZero je označil jako AI,
- TN (true negatives) – počet textů, které jsou ve skutečnosti HUM a GPTZero je označil jako human,
- FP (false positives) – počet textů, které jsou ve skutečnosti HUM, ale GPTZero je označil jako AI,
- FN (false negatives) – počet textů, které jsou ve skutečnosti AI (AI-EDIT nebo AI-GEN), ale GPTZero je označil jako human.

Tyto matice byly sestaveny samostatně pro každou doménu a slouží jako podklad pro souhrnné tabulky s výsledky klasifikace a pro výpočet klasifikačních metrik. V textu jsou tedy uváděny již agregované počty správných a chybných klasifikací, nikoli úplné matice záměn.

Rozlišení AI-EDIT a AI-GEN v rámci analýzy

Přestože GPTZero nerozlišuje mezi AI-EDIT a AI-GEN na úrovni výstupu, dataset toto rozlišení obsahuje. To umožňuje dílčí analýzu uvnitř „AI“ skupiny:

- u všech textů označených jako AI-EDIT lze spočítat, kolik jich GPTZero správně zařadil jako AI a kolik prošlo jako human,
- totéž lze spočítat zvlášť pro AI-GEN.

Vznikají tak další (vnitřní) přehledy, které ukazují:

- zda je pro nástroj snadnější detekovat plně AI generované texty (AI-GEN) než post-editované texty (AI-EDIT),
- zda je některá doména (např. literární) pro detekci systematicky obtížnější než jiná (např. administrativní).

2.3.2 Výpočet kvantitativních metrik a práce s doménami

Na základě matic záměn byly vypočítány běžně používané metriky známé z oblasti klasifikace. Pro jednoduchost jsou definovány vždy ve vztahu ke třídě „AI“ (tj. „text s podílem generativní AI“), zatímco třída „HUM“ tvoří komplement.

Základní metriky

Pro každou doménu (D1, D2, D3) byly z příslušné matice záměn spočítány tyto ukazatele:

- Accuracy (přesnost)
  Podíl všech správně klasifikovaných textů (lidských i AI) na celkovém počtu textů v dané doméně:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- Precision (preciznost) pro třídu AI
  Z textů, které GPTZero označil jako AI, jaká část byla skutečně AI (AI-EDIT nebo AI-GEN):

$$\text{precision}_{AI} = \frac{TP}{TP + FP} \tag{2}$$

- Recall (citlivost) pro třídu AI
  Z textů, které jsou ve skutečnosti AI, jaká část byla nástrojem správně označena jako AI:

$$\text{recall}_{AI} = \frac{TP}{TP + FN} \tag{3}$$

- F1-score pro třídu AI
  Harmonický průměr precision a recall, který zvýrazňuje rovnováhu mezi oběma ukazateli:

$$\text{F1}_{AI} = 2 \cdot \frac{\text{precision}_{AI} \cdot \text{recall}_{AI}}{\text{precision}_{AI} + \text{recall}_{AI}} \tag{4}$$

Vedle těchto hodnot je pro interpretaci důležitý také:

- podíl falešně pozitivních nálezů (FP rate) – tj. jak často jsou lidské texty chybně označeny jako AI:

$$\text{FP rate}_{HUM} = \frac{FP}{FP + TN} \tag{5}$$

V kontextu vysokoškolského prostředí je tato hodnota obzvlášť citlivá, protože reprezentuje situace, kdy by opravdový lidský text mohl být mylně označen jako podvod.

Agregace a porovnání mezi doménami

Kromě samostatných matic a metrik pro každou doménu byly výsledky dále:

- agregovány za celý dataset – součet TP, TN, FP a FN přes všechny tři domény umožňuje spočítat „globální“ přesnost a další metriky,
- porovnány mezi doménami – například:

- zda je přesnost GPTZero v administrativní doméně vyšší než v literární,
- zda se liší recall pro AI texty mezi akademickými a literárními texty,
- jaký je podíl falešně označených lidských textů v jednotlivých doménách.

Samostatná pozornost je věnována i porovnání AI-EDIT a AI-GEN:

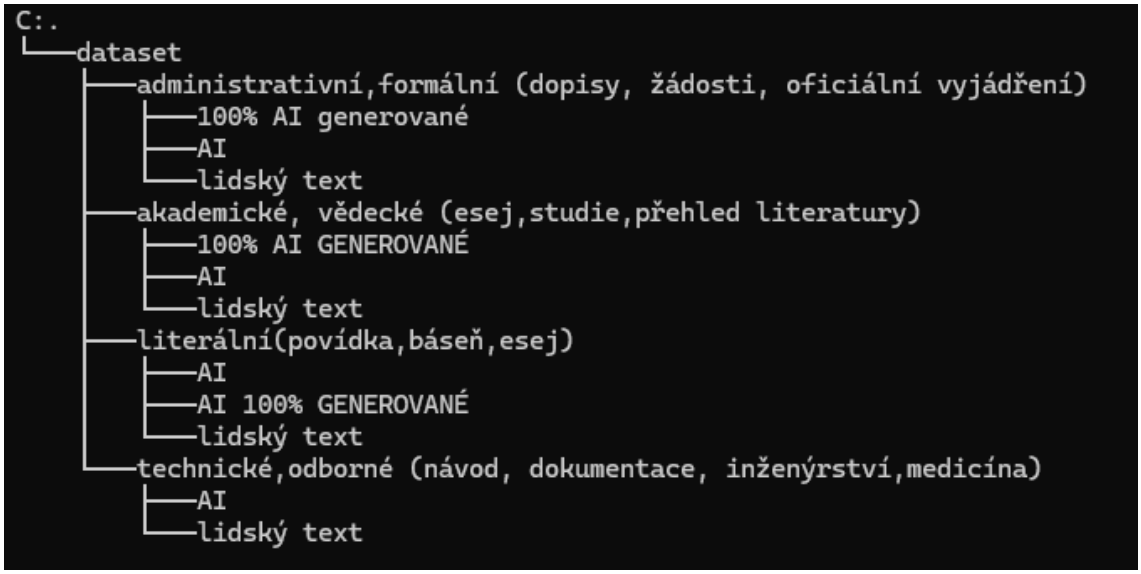
- pro každou doménu lze spočítat podíl správně detekovaných AI-GEN textů,
- a podíl správně detekovaných AI-EDIT textů,
- tyto hodnoty umožňují diskutovat, zda je pro nástroj náročnější odhalit texty, do nichž lidský autor dodatečně zasáhl.

3 Konstrukce datasetu a průběh experimentu

Tato kapitola popisuje praktickou realizaci postupů nastíněných v metodické části. Nejprve je detailně představena konstrukce datasetu – výběr konkrétních zdrojů, úprava lidských textů a generování textů AI-GEN a AI-EDIT. Následně je popsán samotný průběh testování v nástroji GPTZero, včetně organizace skenování a technických omezení vyplývajících z použití prémiové verze nástroje. Na tuto kapitolu navazuje kapitola 4, kde jsou prezentovány a interpretovány výsledky analýz.

3.1 Konstrukce datasetu

V této části je podrobně popsán způsob, jakým byl vytvořen testovací dataset použitý v experimentu. Nejprve jsou představeny jednotlivé textové domény a zdrojové korpusy lidských textů, následně postup jejich úprav do sjednocené podoby a tvorba AI textů ve třídách AI-GEN a AI-EDIT. Kapitola tak převádí obecný návrh z metodiky do konkrétní, krok za krokem realizované podoby a uzavírá ji přehledem velikosti a struktury vzorku.



Obr. 19 Struktura testovacího datasetu

Zdroj: vlastní zpracování (2025)

3.1.1 Volba textových domén a žánrů

Volba textových domén i konkrétních textů vycházela jednak z předpokládaných oblastí použití generativní umělé inteligence ve vysokoškolském prostředí, jednak z dostupných reálných textů, které bylo možné pro účely výzkumu využít. Pro každou doménu byly nejprve vybrány reprezentativní zdrojové korpusy lidských textů a z nich následně vytvářeny či odvozovány jednotlivé položky datasetu.

D1 – Administrativní a formální komunikace
Lidské texty v této doméně vycházejí z reálných administrativních a oficiálních dokumentů, případně z normativních ukázek takových textů. Jde zejména o:

- vzorové stížnosti a návody k jejich psaní v kontextu státní maturitní zkoušky z českého jazyka (např. didakticky koncipovaný text s ukázkovou stížností a rozborem jejích náležitostí),
- otevřené dopisy adresované vládě České republiky či konkrétním ministrům publikované profesními a podnikatelskými asociacemi, vědeckými institucemi a dalšími organizacemi,
- otevřené dopisy a stanoviska adresovaná vedení univerzit (např. rektorovi Univerzity Palackého či ČVUT) k aktuálním otázkám akademického života nebo veřejné debaty,
- oficiální vyjádření měst a obcí k konkrétním kauzám (např. reakce na situaci kolem městských podniků či lokálních konfliktů), zveřejňovaná na úředních portálech,
- stížnosti, žádosti a podání směrem k orgánům veřejné správy (např. žádosti o informace podle zákona, stížnosti na postup policie či jiných orgánů), včetně zveřejněných písemností veřejného ochránce práv a dalších institucí.

Z těchto zdrojů byly vybírány či adaptovány texty, které naplňují požadavky administrativního stylu: formální tón, ustálené formule (oslovení, věc, závěr), jasná struktura (adresy, datum, podpis) a věcně popsaný problém či stanovisko.

D2 – Akademické a vědecké texty
V akademické a vědecké doméně byly lidské texty čerpány ze dvou typů zdrojů:

1. Studentské a soutěžní eseje, typicky publikované v souvislosti se středoškolskými či vysokoškolskými soutěžemi a projekty. Tyto texty reprezentují běžnou podobu eseje v českém školském prostředí – kombinují osobní reflexi s argumentační linií a často reflektují aktuální společenská nebo profesní témata.
2. Odborné a vědecké články z českých časopisů a vysokoškolských publikací. Jde zejména o:
    - teoretické texty o povaze eseje a jejích podobách publikované v rámci vysokoškolských ekonomických a pedagogických periodik,
    - studie a empirické články z časopisu Lifelong Learning (Mendelova univerzita v Brně),
    - filosofické a společenskovědní studie publikované ve *Filosofickém časopisu*.

Z těchto odborných zdrojů byly vybírány kratší úseky textu (např. úvodní kapitoly, části diskuse či shrnutí), které splňují podmínky pro zařazení do datasetu z hlediska rozsahu a relativní samostatnosti. Výsledkem je soubor textů, které reprezentují jak studentské akademické psaní, tak standardní odborný a vědecký styl v češtině.

D3 – Literární a esejistické texty
Lidské texty v literární doméně vycházejí z digitalizovaných literárních děl a souborů dostupných prostřednictvím Městské knihovny v Praze. Do datasetu byly zařazeny zejména:

- texty z almanachu Start Line, který představuje tvorbu současných autorů a autorek,
- vybrané povídky z klasické sbírky Povídky z druhé kapsy,

- kratší prózy a texty ze souboru Povídky a jiné krátké texty,
- esejistické texty ze samostatné publikace zaměřené na eseje.

Z těchto děl byly vybrány nebo vyčleněny jednotlivé úryvky a samostatné povídky, které lze chápat jako uzavřené textové jednotky vhodné pro analýzu. V literární doméně se tak propojuje klasická česká beletrie s novějšími texty publikovanými v rámci současných almanachů, které dohromady reprezentují škálu od narativní prózy po reflexivní esejistiku.

3.1.2 Postup práce se zdrojovými texty

Ze zdrojových dokumentů popsaných v části 2.1.1 nebyly texty přebírány mechanicky, ale prošly několika kroky úpravy, aby odpovídaly potřebám datasetu a bylo možné je mezi sebou porovnávat napříč doménami.

Nejprve byly u všech tří domén vybrány vhodné úseky textu. V případě administrativních a formálních dokumentů šlo zpravidla o celé texty stížností, žádostí nebo otevřených dopisů, které již samy o sobě tvoří relativně krátký a uzavřený celek. U akademických a vědeckých zdrojů byly vybírány zejména úvodní části, shrnutí, diskuse či kratší eseje, které je možné číst samostatně bez znalosti celého článku. U literárních a esejistických zdrojů šlo buď o celé krátké povídky, nebo o ucelené úseky delších textů (odstavec až několik odstavců), které zachovávají smysl a narativní kontinuitu.

V další fázi došlo k technickému zpracování textů:

- texty byly převedeny z původního formátu (např. PDF, webová stránka) do prostého textu,
- byly odstraněny prvky, které nejsou součástí samotného jazykového projevu (nadpisy nad rámec žánru, čísla stránek, poznámky pod čarou, technická metadata, navigační prvky webu apod.),
- sjednotilo se základní typografické a pravopisné formátování (uvozovky, pomlčky, odstavce), pokud bylo porušeno převodem.

Obsahová stránka textů nebyla měněna – nebyly přepisovány věty ani přidávány nové informace. Úpravy měly výhradně normalizační charakter, aby všechny vzorky odpovídaly zvoleným kritériím. V omezených případech došlo k zkrácení příliš dlouhých textů (například u vědeckých studií), a to tak, že byl zachován souvislý úsek textu v požadovaném rozsahu. Tyto zásahy byly prováděny tak, aby nedošlo k porušení srozumitelnosti výpovědi ani ke změně stylového charakteru textu.

Současně byla věnována pozornost anonymizaci a etickým aspektům: pokud se v textech objevovala osobní jména nebo jiné citlivé údaje, které nebyly pro analýzu podstatné, byly buď ponechány v podobě, v jaké se vyskytují v publikovaném zdroji (např. u oficiálně zveřejněných otevřených dopisů), nebo byly při přebírání textu vypuštěny, pokud šlo o marginální prvky.

Výsledkem tohoto procesu je sada lidských textů, které:

- reprezentují autentický jazykový projev v dané doméně,
- jsou technicky sjednocené z hlediska formátu a přibližné délky,

- lze mezi sebou relevantně porovnávat jak v rámci konkrétní domény, tak mezi doménami navzájem.

3.1.3 Generování AI textů (AI-GEN) a tvorba post-editovaných textů (AI-EDIT)

Vedle lidských textů (HUM) byly pro každou doménu vytvořeny dvě další třídy textů: plně generované texty umělou inteligencí (AI-GEN) a texty vzniklé úpravou AI výstupu člověkem (AI-EDIT). Cílem bylo simulovat dva typické způsoby využití generativní umělé inteligence v praxi:

- situaci, kdy je výstup modelu použit téměř beze změn
- situaci, kdy člověk bere AI výstup jako výchozí návrh a dále ho upravuje.

Generování textů AI-GEN:

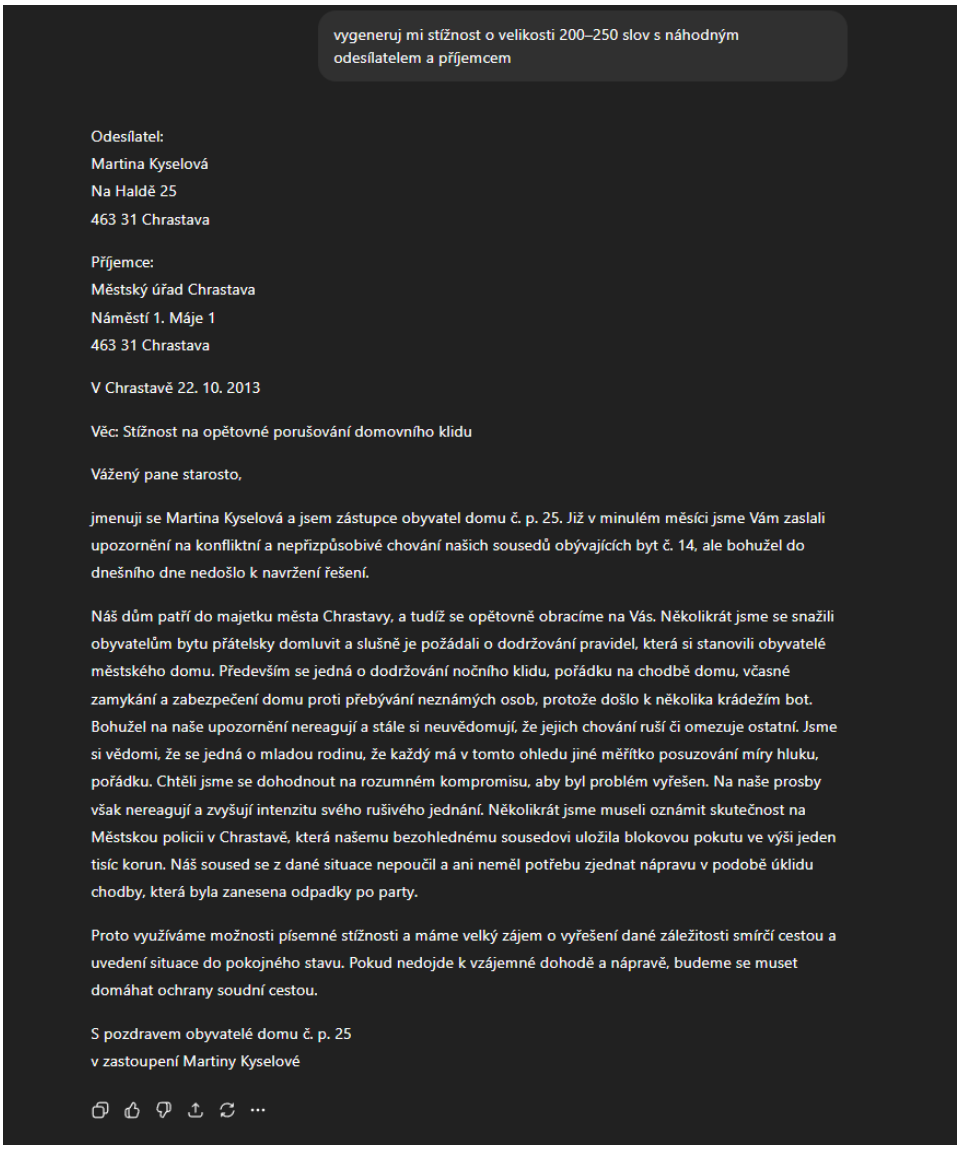
Texty třídy AI-GEN byly vytvářeny pomocí generativního jazykového modelu ChatGPT. Pro každou doménu byly připraveny prompty (zadání), která co nejvěrněji odpovídala žánru a komunikační situaci lidských textů:

- v administrativní/formální doméně šlo například o zadání typu „Napiš formální stížnost na…“ nebo „Napiš otevřený dopis adresovaný…“, s upřesněním adresáta, tématu a požadovaného tónu (formální, věcný);
- v akademické/vědecké doméně byla zadání formulována jako požadavky na krátkou esej, úvod odborného textu či shrnutí studie k určitému tématu;
- v literární/esejistické doméně byla zadání zaměřena na vyprávění příběhu nebo reflexivní esej na zvolený námět, často s určením základního motivu, perspektivy nebo cílového dojmu.

Při generování textů byly důsledně dodržovány následující zásady:

- jazyk: u všech promptů bylo explicitně uvedeno, že text má být napsán česky,
- rozsah: nástroj byl instruován, aby generoval text přibližně v rozsahu srovnatelném s lidskými texty v dané doméně (např. „v rozsahu přibližně … slov“),
- styl: prompty obsahovaly stručnou charakteristiku stylu (např. „formální“, „akademický“, „narativní“, „subjektivně-reflexivní“).

Pokud nástroj vygeneroval text, který zjevně neodpovídal zadání (např. byl příliš krátký, přešel do jiného žánru, opakovaně vypisoval odrážky místo souvislého textu), byl prompt zopakován nebo mírně upraven. Nevhodné výstupy nebyly do datasetu zařazeny. Nebyly však prováděny obsahové korekce generovaných textů – do třídy AI-GEN spadají pouze ty výstupy, které jsou přímo výsledkem modelu, s případnými minimálními technickými úpravami formátu (odstavce apod.).



Obr. 20: ukázka AI-GEN datasetu
Zdroj: vlastní zpracování (2025)

Tvorba post-editovaných textů AI-EDIT:

Texty třídy AI-EDIT vznikly tak, že výchozím materiálem byly lidské texty (HUMAN-AUTH), které byly následně předány generativnímu modelu k úpravě. Cílem bylo simulovat situaci, kdy student nebo jiný uživatel vezme vlastní text a požádal nástroj umělé inteligence o jeho „vylepšení“ tak, aby lépe odpovídal zadání.

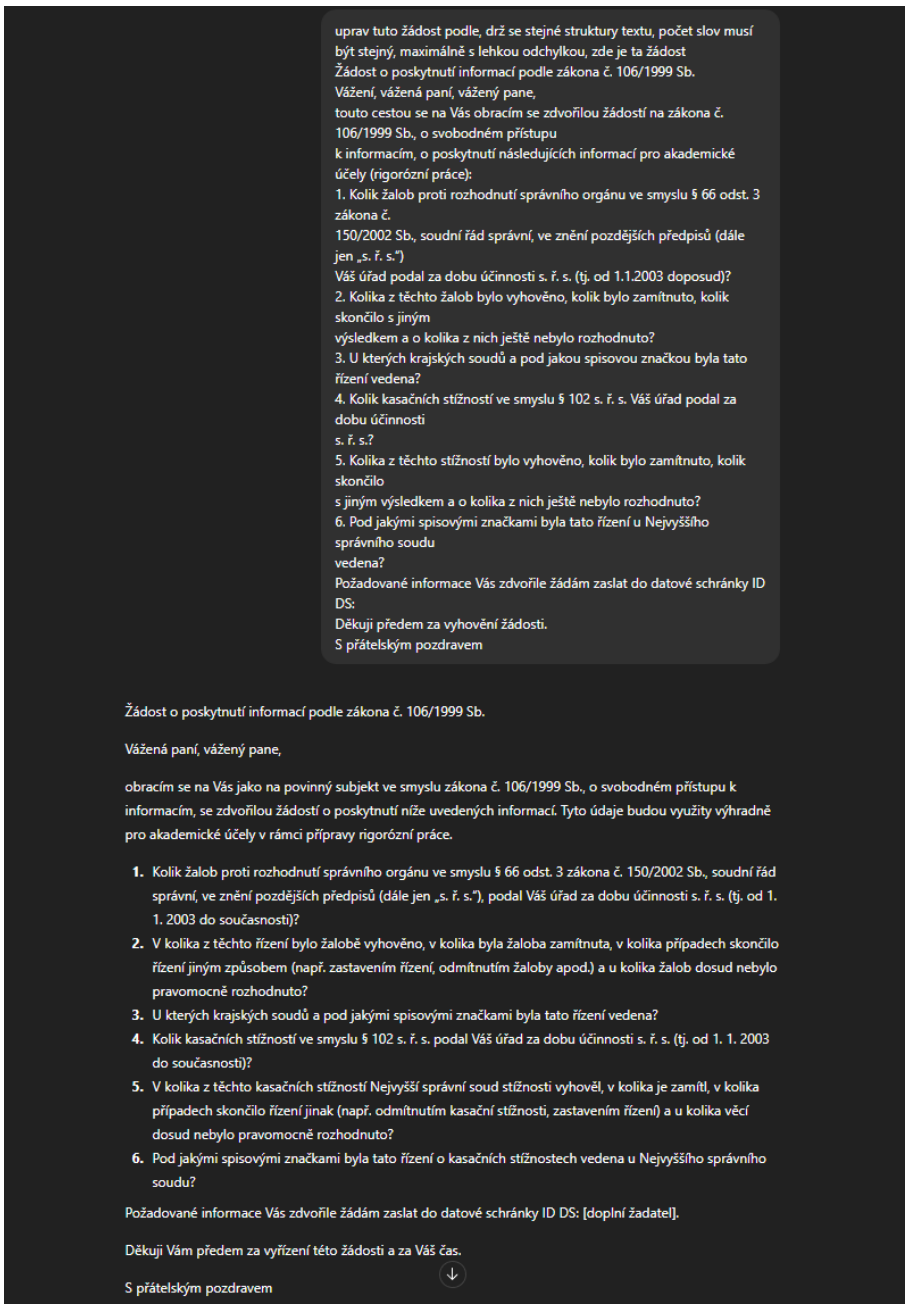
Postup byl následující:

- **Předání lidského textu modelu**
  Pro každý text třídy HUMAN-AUTH byl připraven prompt, který obsahoval jednak instrukci k úpravě. Zadání bylo formulováno tak, aby model:
  - zachoval stejnou doménu a žánr
  - dodržel přibližně stejnou délku
  - respektoval původní téma a kontext

- **Automatická AI úprava**
  Generativní model následně vytvořil upravenou verzi textu. Do výsledných AI-EDIT textů nebylo po této fázi významově ani stylisticky zasahováno člověkem.

Při tvorbě AI-EDIT textů byla dodržena dvě omezení:

- Původní lidský text nebyl modelu zadáván jako volná inspirace, ale jako konkrétní předloha, kterou má upravit.
- Úprava nesměla zásadně měnit téma, žánr ani rozsah, aby bylo možné text stále považovat za odvozený od lidského originálu.

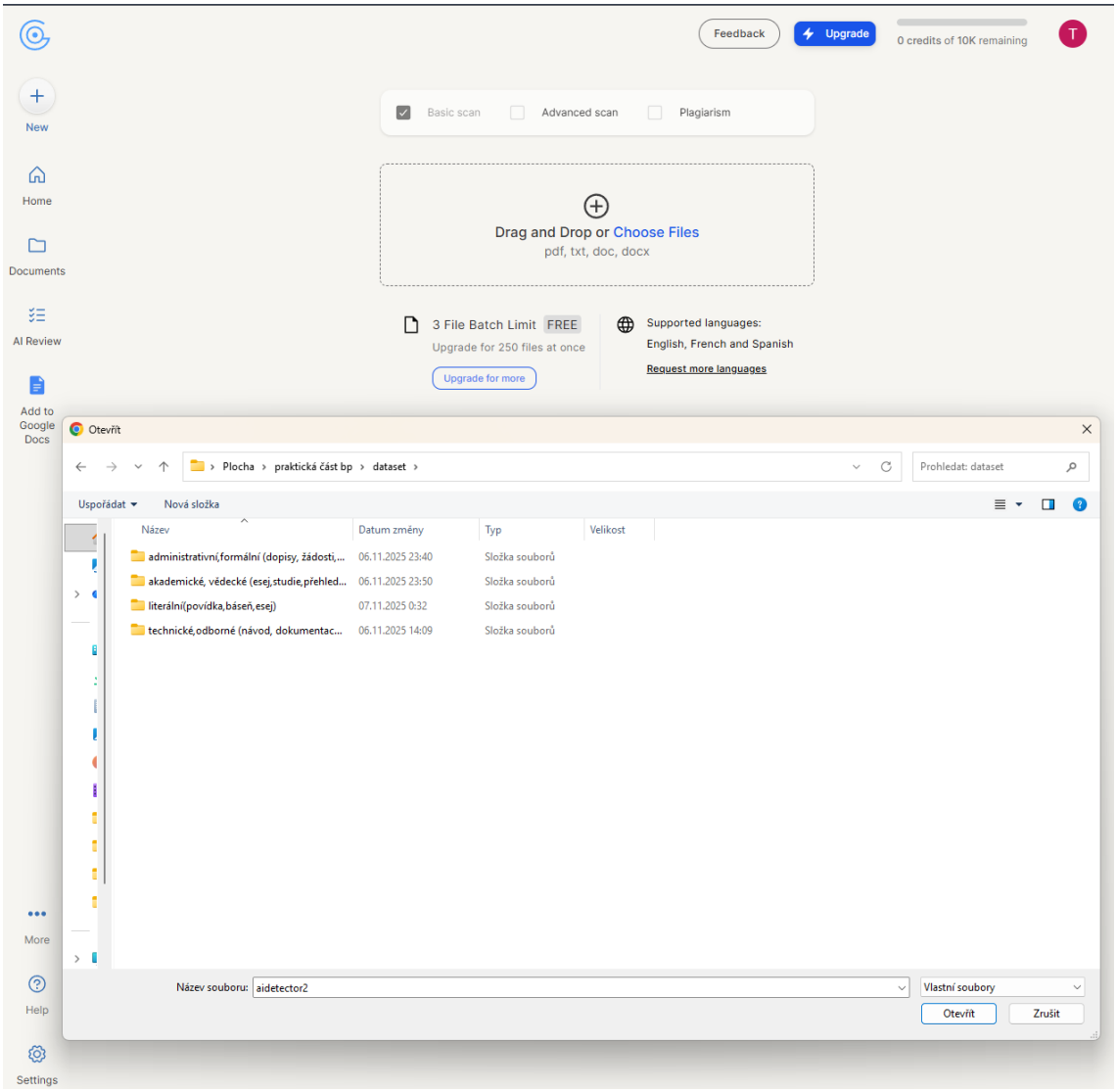


Obr. 21: ukázka tvorby AI-EDIT datasetu
Zdroj: vlastní zpracování (2025)

Tímto postupem vznikla třída textů, které kombinují obsahové jádro lidského autorství s formální a stylistickou úpravou generativním modelem. Právě takové texty jsou v praxi často problematické pro detekční nástroje: obsah může působit lidsky, ale povrchová stylizace nese znaky automatické generace.

3.2 Průběh testování

Tato kapitola popisuje vlastní experiment s detekčním nástrojem GPTZero. Zaměřuje se na praktickou organizaci testování – použitý typ předplatného, práci s kredity, přípravu souborů a jejich import do rozhraní nástroje – a na způsob, jakým byl spuštěn a řízen dávkový (batch) advanced scan nad celým datasetem. Cílem je ukázat, za jakých podmínek byly získány výsledky, které jsou následně vyhodnocovány v kapitole o výsledcích analýz.



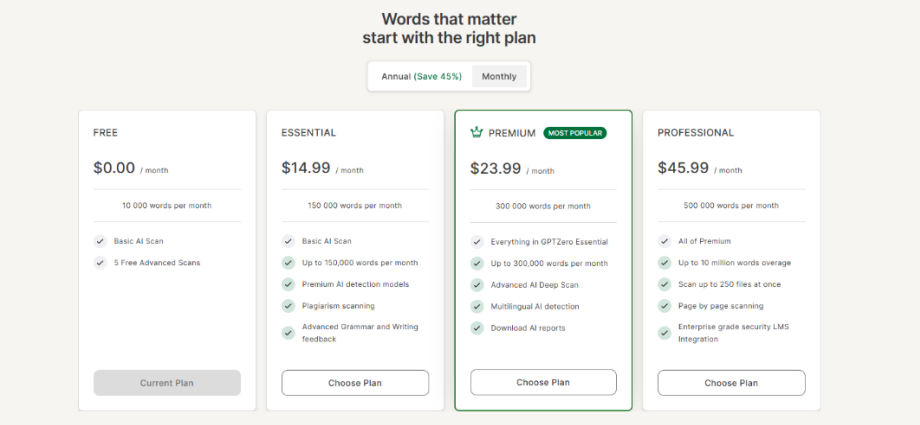
Obr. 22: Výběr datasetu v rozhraní GPTZero
Zdroj: vlastní zpracování (2025)

3.2.1 Předplatné, kredity a praktická omezení testování

Testování probíhalo v rámci dočasného předplatného GPTZero Premium, které je dostupné ve formě týdenní zkušební verze zdarma. V rámci této zkušební verze bylo poskytnuto 300 000 kreditů, které představují interní jednotku, již nástroj používá k účtování jednotlivých skenů (kredity se typicky odvíjejí od délky a počtu nahraných textů).

Pro otestování celého datasetu – tj. všech 180 textů napříč třemi doménami a třemi třídami původu – bylo potřeba přibližně 120 000 kreditů. Tím byla pokryta základní analýza všech textů v režimu *advanced scan* (viz níže). Zbývající počet kreditů sice umožňoval dílčí doplňkové testování, nicméně složitější experimentální nastavení (například opakované skenování několika variant téhož textu, rozsáhlé testy různých délek či postupných úprav) již nebylo z důvodu kreditového limitu reálně proveditelné.

Praktickým důsledkem tohoto omezení je, že testování je koncipováno jako jednoprůchodová analýza celého datasetu: každému textu odpovídá jeden hlavní výsledek skenu v daném režimu. Tato skutečnost je důležitá při interpretaci výsledků – cílem není detailní ladění hranice detekce v jednom konkrétním nástroji, ale spíše orientační zhodnocení jeho chování na realistickém vzorku textů.



Obr. 23: Přehled možností předplatného
Zdroj: vlastní zpracování (2025)

3.2.2 Import datasetu a průběh skenování

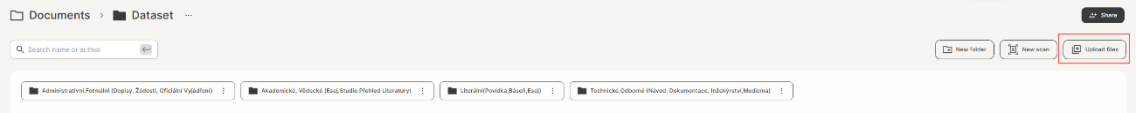
Testování probíhalo tak, že byl vlastní dataset importován přímo do rozhraní GPTZero. Jednotlivé texty byly připraveny ve formátu, který nástroj podporuje (typicky samostatné textové nebo dokumentové soubory), a rozděleny podle domén a tříd původu tak, aby bylo možné výsledky následně spárovat s metadaty:

- pro každou kombinaci *doména × třída původu* (např. D1–HUMAN-AUTH, D2–AI-GEN atd.) existovala sada příslušných souborů,
- soubory byly pojmenovány systematicky (např. D1_HUM_01, D1_HUM_02 atd.), aby bylo možné výsledky jednoznačně přiřadit k položkám datasetu.

V rozhraní GPTZero byl následně spuštěn batch scan v režimu advanced, který:

1. načetl všechny nahrané soubory,

2. pro každý soubor vygeneroval výsledek klasifikace (např. „AI-generated“, „human-written“, případně mezistupně),
3. přiřadil jednotlivým textům skóre a další interní hodnoty,
4. umožnil export výsledků do podoby tabulky, s níž bylo možné dále pracovat v rámci této práce.



Obr. 24: Ukázka prostředí adresáře pro spuštění scanu
Zdroj: vlastní zpracování (2025)

Výsledkem tohoto postupu je, že pro každý z 180 textů datasetu existuje jednoznačný záznam o tom, jak ho GPTZero v daném nastavení klasifikoval. Na základě tohoto záznamu byla následně sestavena matice záměn pro každou doménu a odvozeny další metriky.

4 Výsledky analýz

Tato kapitola představuje výsledky testování nástroje GPTZero na datasetu popsaném v kapitole 2. V souboru je zahrnuto 180 textů rovnoměrně rozdělených do tří domén (administrativní/formální, akademické/vědecké a literární/esejistické) a tří tříd původu (HUMAN-AUTH, AI-EDIT, AI-GEN). Pro každou kombinaci domény a třídy původu je k dispozici dvacet textů, což umožňuje základní kvantitativní porovnání chování detektoru napříč sledovanými oblastmi.

Výkonnost nástroje je hodnocena na základě binárních matic záměn a z nich odvozených metrik definovaných v kapitole 2.3, zejména přesnosti, preciznosti, citlivosti (recall) a F1-score. V praktické části jsou pro větší přehlednost uváděny souhrnné tabulky s počty správných a chybných klasifikací a samostatné tabulky s vypočtenými metrikami. V následujících podkapitolách jsou nejprve samostatně prezentovány výsledky pro administrativní/formální, akademické/vědecké a literární/esejistické texty, na něž navazuje souhrnné srovnání a stručná interpretace hlavních zjištění.

4.1.1 Administrativní a formální texty

V administrativní/formální doméně GPTZero:

- správně označil 19 z 20 lidských textů jako „lidské“ (1 falešně pozitivní označení jako AI),
- správně označil 13 z 20 AI-EDIT textů jako AI (detekční úspěšnost 65 %),
- správně označil pouze 18 z 20 AI-GEN textů jako AI (90 %).

Tab. 1: Souhrnný přehled skutečných a predikovaných tříd pro administrativní/formální texty

dataset	správně	špatně
AI (AI-EDIT + AI-GEN)	31	9
HUMAN-AUTH	19	1

Zdroj: vlastní zpracování (2025)

Binární klasifikace (AI = AI-EDIT + AI-GEN) tedy vede k následujícím klasifikačním metrikám:

Tab. 2: Klasifikační metriky pro dataset administrativní/formální texty

Metrika	Hodnota (podíl)	Hodnota (%)
Accuracy	0,83	83,3 %
Precision_AI	0,97	96,9 %
Recall_AI	0,78	77,5 %
F1_AI	0,86	86,1 %

Zdroj: vlastní zpracování (2025)

Hodnota accuracy = 0,83 znamená, že GPTZero správně klasifikuje přibližně 83 % administrativních textů – chybně je tedy vyhodnocena zhruba jedna šestina vzorku. Ještě příznivější je hodnota precision_AI = 0,97, podle níž je z textů označených jako AI téměř 97 % skutečně z třídy AI (AI-EDIT nebo AI-GEN). Falešně pozitivních nálezů je v této doméně minimum

(v celém souboru pouze jeden lidský text). Relativně vysoká je i citlivost: recall_AI = 0,78 ukazuje, že detektor zachytí zhruba tři čtvrtiny administrativních textů s podílem AI, zatímco devět z čtyřiceti AI textů projde jako „lidských“. Tomu odpovídá i vysoká hodnota F1_AI = 0,86, která odráží poměrně dobře vyvážený vztah mezi precizností a citlivostí.

Z hlediska lidských textů je nástroj velmi konzervativní – pouze 1 z 20 administrativních lidských textů byl nesprávně označen jako AI. U AI textů se naopak ukazuje, že GPTZero dokáže administrativní styl relativně dobře odhalovat: plně generované texty AI-GEN jsou správně klasifikovány ve většině případů (18 z 20), mírně nižší úspěšnost má nástroj u post-editovaných textů AI-EDIT (13 z 20). Přesto i zde část textů s podílem AI působí na detektor natolik „normovaně“ a předvídatelně, že jsou zařazeny mezi lidské.

4.1.2 Akademické a vědecké texty

V akademické/vědecké doméně jsou rozdíly ještě výraznější:

- všech 20 lidských textů bylo správně označeno jako „lidské“ (0 falešně pozitivních nálezů),
- z 20 AI-EDIT textů bylo jako AI rozpoznáno pouze 7 (35 %),
- z 20 AI-GEN textů byly všechny označeny jako AI (100 %).

Tab. 3: Souhrnný přehled skutečných a predikovaných tříd pro akademické/vědecké texty

dataset	správně	špatně
AI (AI-EDIT + AI-GEN)	27	13
HUMAN-AUTH	20	0

Zdroj: vlastní zpracování (2025)

Souhrnné výsledky binární klasifikace pro tuto doménu:

Tab. 4: Klasifikační metriky pro dataset akademické/vědecké texty

Metrika	Hodnota (podíl)	Hodnota (%)
Accuracy	0,78	78,3 %
Precision_AI	1,00	100 %
Recall_AI	0,68	67,5 %
F1_AI	0,81	80,6 %

Zdroj: vlastní zpracování (2025)

V akademické doméně vychází celková accuracy = 0,78, tedy GPTZero správně klasifikuje bezmála čtyři pětiny textů. Hodnota precision_AI = 1,00 znamená, že v tomto souboru nástroj ani jednou falešně neoznačil lidský text jako AI – všechny případy, kdy detektor ohlásil přítomnost AI, skutečně odpovídají třídě AI. Citlivost je zde také relativně vysoká: recall_AI = 0,68 ukazuje, že nástroj zachytí přibližně dvě třetiny akademických textů s podílem AI. Kombinovaná metrika F1_AI = 0,81 tak potvrzuje poměrně dobrý kompromis mezi precizností a citlivostí.

Prakticky to znamená, že z 40 AI textů (AI-EDIT + AI-GEN) je jako AI rozpoznáno 27, zatímco 13 textů je hodnoceno jako „lidských“. Všechny lidské texty jsou naopak klasifikovány správně. Podrobnější pohled ukazuje významný rozdíl mezi oběma AI třídami: plně generované texty AI-GEN jsou detekovány se 100% úspěšností, zatímco u post-editovaných textů AI-EDIT je úspěšnost výrazně nižší (7 z 20). GPTZero tedy v akademickém stylu poměrně spolehlivě odhaluje čistě strojově generované práce, ale část „zamaskovaných“ post-editovaných textů stále prochází jako lidská.

4.1.3 Literární a esejistické texty

V literární doméně se chování nástroje částečně obrací:

- všech 20 lidských textů bylo správně klasifikováno jako „lidské“ (0 falešně pozitivních nálezů),
- z 20 AI-EDIT textů bylo jako AI rozpoznáno 5 (25 %),
- z 20 AI-GEN textů bylo jako AI rozpoznáno 18 (90 %),

Tab. 5: Souhrnný přehled skutečných a predikovaných tříd pro literární/esejistické texty

dataset	správně	špatně
AI (AI-EDIT + AI-GEN)	23	17
HUMAN-AUTH	20	0

Zdroj: vlastní zpracování (2025)

Souhrnné výsledky binární klasifikace:

Tab. 6: Klasifikační metriky pro dataset literární/esejistické texty

Metrika	Hodnota (podíl)	Hodnota (%)
Accuracy	0,72	71,7 %
Precision_AI	1,00	100 %
Recall_AI	0,58	57,5 %
F1_AI	0,73	73 %

Zdroj: vlastní zpracování (2025)

V literární a esejistické doméně dosahuje GPTZero rovněž relativně dobrého výkonu. Celková accuracy = 0,72 znamená, že správně klasifikuje zhruba 72 % textů. Stejně jako u akademické domény je precision_AI = 1,00, tedy žádný lidský literární text nebyl nástrojem mylně označen jako AI. Rozdíly se objevují v citlivosti: recall_AI = 0,58 ukazuje, že detektor zachytí přibližně 58 % literárních textů s podílem AI, zatímco zhruba dvě pětiny AI textů procházejí jako lidské. Kombinace perfektní preciznosti a středně vysoké citlivosti vede k hodnotě F1_AI = 0,73, která naznačuje, že v literární doméně je detekce vyvážená, avšak mírně slabší než v administrativní a akademické oblasti.

Detailnější pohled na výsledky ukazuje, že nástroj velmi dobře odhaluje plně AI generované povídky a eseje (AI-GEN, 18 z 20 správně klasifikovaných), zatímco post-editované texty AI-EDIT se častěji „schovávají“ mezi lidskými (správně rozpoznáno 5 z 20). Z hlediska praktického využití

to znamená, že GPTZero umí poměrně spolehlivě detekovat čistě strojově generované literární texty, ale má výraznější potíže s hybridními texty, v nichž člověk strojový výstup po obsahové i stylistické stránce upravil.

4.1.4 Souhrnné srovnání a základní interpretace

Při agregaci dat za celý dataset (180 textů) vycházejí následující hodnoty:

Tab. 7: Souhrnná tabulka metrik

Metrika	Hodnota (podíl)	Hodnota (%)
Accuracy	0,78	77,8 %
Precision_AI	0,99	98,8 %
Recall_AI	0,68	67,5 %
F1_AI	0,80	80,2 %

Zdroj: vlastní zpracování (2025)

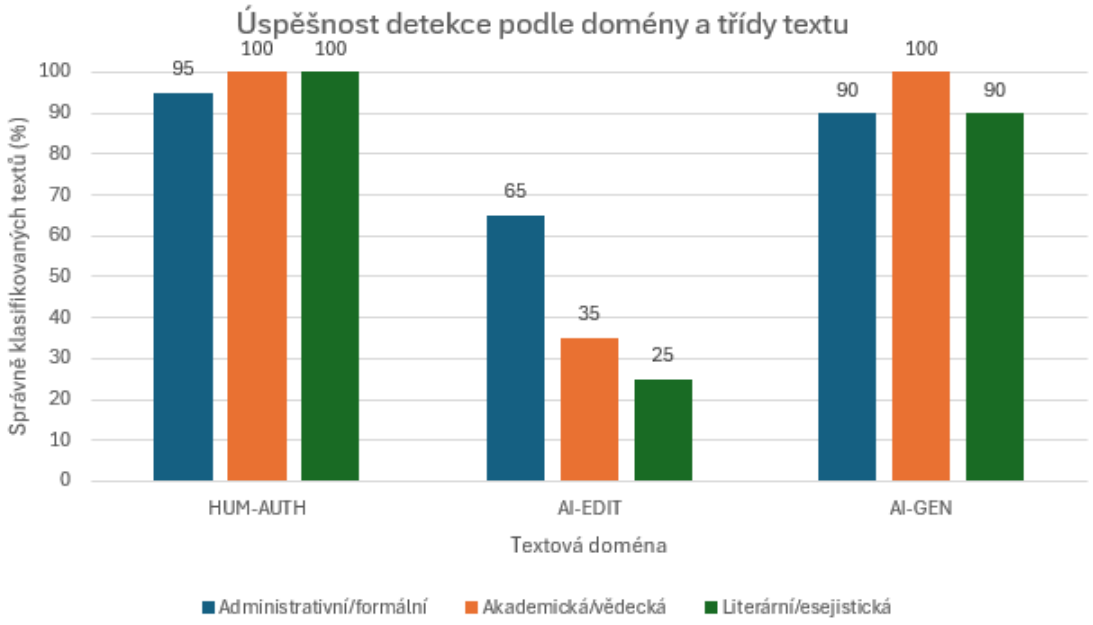
Při agregaci dat za celý dataset vycházejí následující hodnoty (viz Tab. 7). Accuracy = 0,78 znamená, že GPTZero správně klasifikuje přibližně 78 % všech textů v datasetu. Z hlediska celkové úspěšnosti tedy nejde o neomylný „detektor“, ale o nástroj, který ve zhruba čtyřech pětinách případů rozhodne správně.

Velmi vysoká Precision_AI = 0,99 ukazuje, že pokud nástroj text jako AI označí, téměř vždy jde skutečně o text s podílem AI. Falešně pozitivních nálezů je minimum – v celém souboru byl omylem označen pouze 1 lidský text ze 60. Naopak Recall_AI = 0,68 znamená, že GPTZero zachytí přibližně 68 % všech AI textů; zhruba třetina textů s podílem AI tak stále projde jako „lidská“. Z pohledu odhalování podvodů jde proto o nástroj, který je velmi opatrný v označování textů jako AI a stále část takových textů neodhalí. Kombinovaná metrika F1_AI = 0,80 tuto charakteristiku potvrzuje: výkon detektoru je v českém prostředí solidní, ale nikoli bezchybný.

Z hlediska jednotlivých domén se situace liší:

- administrativní/formální texty vykazují nejvyšší celkovou úspěšnost – kombinace vysoké preciznosti a relativně vysokého recall (0,78) vede k nejlepšímu F1 skóre; přesto devět z 40 administrativních AI textů zůstává nerozpoznaných,
- akademické/vědecké texty dosahují podobně dobrých výsledků, zejména díky perfektní preciznosti (žádný lidský text nebyl označen jako AI); recall (0,68) je o něco nižší než v administrativní doméně a část post-editovaných textů AI-EDIT zůstává skryta,
- literární/esejistické texty mají celkově nejnižší F1 skóre, i když stále relativně dobré: recall (0,58) ukazuje, že více než polovina literárních AI textů je detekována, ale významná část – zejména z třídy AI-EDIT – zůstává nerozpoznána.

Pro další části práce (zejména analýzu diskurzních a lexikálních markerů) je důležité, že plně generované texty AI-GEN jsou napříč doménami detekovány podstatně lépe než texty AI-EDIT. Post-editované texty s kombinací strojově generované základní struktury a lidských úprav představují z hlediska GPTZero nejproblematičtější oblast – právě na ně se zaměřuje následující část práce věnovaná identifikačním rysům a markerům.



Obr. 25: Graf úspěšnosti detekce podle domény a třídy textu

Zdroj: vlastní zpracování (2025)

Závěr

Cílem této práce bylo zhodnotit možnosti identifikace AI-generovaných textů v kontextu českého vysokoškolského prostředí na základě vlastního datasetu a empirického testování detekčního nástroje GPTZero. Teoretická část shrnula základní pojmy a principy generativní a detekční umělé inteligence, vývoj jazykových modelů a klíčové metody detekce, včetně využití perplexity, burstiness a stylometrických rysů. Pozornost byla věnována také etickým a společenským souvislostem, zejména otázkám plagiátorství, dezinformací, autorských práv a dopadům generativní AI na vzdělávání a trh práce.

Praktická část práce se zaměřila na návrh a otestování datasetu pro identifikaci AI-generovaných textů ve třech doménách: administrativní/formální, akademické/vědecké a literární/esejistické. V každé doméně byly zastoupeny tři třídy původu textu, lidské texty (HUM), plně generované texty (AI-GEN) a texty vzniklé post-editací AI výstupu člověkem (AI-EDIT), vždy po dvaceti položkách. Celkem tak bylo analyzováno 180 textů. Dataset byl testován nástrojem GPTZero v režimu advanced v rámci týdenní prémiové licence a výstupy nástroje byly převedeny do matic záměn a klasifikačních metrik, jako jsou accuracy, precision, recall a F1-score. Tento postup umožnil kvantitativně posoudit chování detektoru napříč doménami a typy textů.

Výsledky ukázaly, že chování detektoru je silně doménově podmíněné. V administrativních textech GPTZero téměř nikdy neprávem neoznačí lidský text za AI, ale systematicky podhodnocuje přítomnost AI, a to zejména u plně generovaných textů, které často vyhodnotí jako lidské. V akademické a vědecké doméně se tento trend ještě prohlubuje: nástroj prakticky všechny texty, včetně stoprocentně generovaných, klasifikuje jako lidské. Akademický styl se zde jeví jako tak normovaný a „dobře napsaný“, že rozdíl mezi lidským a strojovým původem z hlediska použitých rysů téměř mizí.

V literární a esejistické doméně se chování nástroje částečně obrací. GPTZero poměrně dobře rozpoznává plně generované texty, které nesou specifický kvantitativní a stylometrický profil, zatímco post-editované texty se mu často ztrácejí mezi lidskými. Agregovaně má detektor velmi vysokou přesnost ve chvíli, kdy text označí jako AI – tedy pokud AI nález vyhlásí, bývá zpravidla oprávněný, současně vykazuje nízkou citlivost, protože velká část textů s podílem AI projde jako lidská. To je z hlediska akademické praxe zásadní zjištění: riziko falešného obvinění je relativně malé, ale možnost, že nástroj používání AI vůbec nezachytí, je naopak vysoká.

Z hlediska formulovaných výzkumných otázek lze shrnout, že detekce je výrazně úspěšnější u narativních literárních textů než u formálně normovaných administrativních a akademických textů. Zároveň se potvrzuje, že největší výzvu představují post-editované texty, které kombinují strojově generovanou základní strukturu s lidskými úpravami. Právě tento typ textu je pro vysokoškolské prostředí velmi realistický, protože odpovídá běžnému způsobu, jakým mohou studenti generativní nástroje využívat. Práce tak potvrzuje, že detekce čistě technickými prostředky naráží na své limity právě tam, kde je využití AI v praxi nejpravděpodobnější.

Je zároveň nutné zdůraznit omezení této studie. Testování se opírá o jediný detekční nástroj v konkrétní verzi dostupné v době měření a proběhlo v podobě jednoprůchodové analýzy daného datasetu. Kreditový limit a časově omezené předplatné neumožnily detailnější experimenty s různými režimy skenování, délkou textu či opakovaným testováním týchž vzorků. Dataset je koncipován jako vyvážený, ale stále poměrně malý a zaměřený na tři vybrané

domény; výsledky proto nelze bez dalšího zobecňovat na všechny typy textů ani na jiné detekční nástroje. Přesto poskytují užitečný obraz o tom, jak se jeden konkrétní, v praxi využívaný detektor chová v realistickém vysokoškolském kontextu.

Z praktického hlediska práce ukazuje, že technické detekční nástroje je vhodné chápat spíše jako podpůrný indikátor než jako spolehlivý důkazní prostředek. I nástroj s velmi nízkým počtem falešně pozitivních nálezů může velkou část reálného využití AI opomenout, zejména v akademickém a administrativním psaní. Politika vysokých škol by proto neměla spoléhat výhradně na automatickou detekci, ale měla by kombinovat jasně komunikovaná pravidla, podporu transparentnosti, například formou reflexe postupu psaní, a didaktické přístupy, které studenty učí generativní AI používat způsobem slučitelným s cíli výuky. Technická detekce může v tomto rámci sloužit jako pomocný nástroj, který upozorní na problematické případy, nikoli jako jediný základ pro rozhodování o nepoctivosti.

Závěrem lze říci, že generativní umělá inteligence musí být ve vysokoškolské politice a výuce reflektována systematicky. Nejde o technologii, kterou lze jednoduše zakázat nebo plně kontrolovat prostřednictvím detekčních nástrojů, ale o nový typ psacího prostředí, v němž budou studenti i vyučující pracovat dlouhodobě. Odpovědný přístup k AI ve vzdělávání bude vyžadovat kombinaci technických, institucionálních i pedagogických opatření a zároveň trpělivý dialog mezi studenty, vyučujícími a vedením škol. Tato práce k tomuto dialogu přispívá tím, že ukazuje konkrétní limity současných detekčních nástrojů v českém prostředí a navrhuje, aby byly chápány jako užitečná, ale z principu omezená součást širšího systému podpory akademické integrity.

Seznam použité literatury

ADAGIRI, A., et al. Automatic Detection of AI-Generated Text from LLMs Using Feature-Driven Transformer Networks. In: Artificial Intelligence in HCI: 6th International Conference AI-HCI 2025, Held as Part of the 27th HCI International Conference HCII 2025: Proceedings, Part 3. Cham: Springer Nature, 2025. s. 214–225. [online]. [cit. 2025-01-19]. Dostupné z: https://link.springer.com/chapter/10.1007/978-3-031-93418-6_15

AGRAHARI, S., S. BISHT a R. S. SANASAM. Text Authorship Attribution: Stylometric Insights into Human and LLM-Generated Text. In: CODS-COMAD 2024: 8th ACM India Joint International Conference on Data Science and Management of Data. New York: ACM, 2024. s. 344–346. [online]. [cit. 2025-01-19]. Dostupné z: https://dl.acm.org/doi/10.1145/3703323.3703712

AMODEI, D., et al. Concrete Problems in AI Safety. arXiv. [online]. 21. červen 2016 [cit. 2025-01-19]. Dostupné z: https://arxiv.org/abs/1606.06565

BOMMASANI, R., et al. On the Opportunities and Risks of Foundation Models. arXiv. [online]. 16. srpen 2021 [cit. 2025-01-19]. Dostupné z: https://arxiv.org/abs/2108.07258

BOSTROM, Nick. Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press, 2014. ISBN není uvedeno.

BROWN, T., et al. Language Models are Few-Shot Learners. arXiv. [online]. 28. květen 2020 [cit. 2025-01-19]. Dostupné z: https://arxiv.org/abs/2005.14165

CHOLLET, François. Deep Learning with Python. Shelter Island: Manning Publications, 2018. ISBN není uvedeno.

DONG, B., et al. Large Language Models in Education: A Systematic Review. In: 2024 6th International Conference on Computer Science and Technologies in Education (CSTE). New York: IEEE, 2024. s. 131–134.

FLORIDI, L. a COWLS, J. A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review. 2019. Místo vydání není uvedeno.

HU, Z., et al. Unbiased Watermark for Large Language Models. arXiv. [online]. 16. říjen 2023 [cit. 2025-01-19]. Dostupné z: https://arxiv.org/abs/2310.10669

JOBIN, A., IENCA, M. a VAYENA, E. The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence. 2019. Místo vydání není uvedeno.

KIRCHENBAUER, J., et al. A Watermark for Large Language Models. arXiv. [online]. 24. leden 2023 [cit. 2025-01-19]. Dostupné z: https://arxiv.org/abs/2301.10226

KIRCHENBAUER, J., et al. On the Reliability of Watermarks for Large Language Models. arXiv. [online]. 10. duben 2024 [cit. 2025-01-19]. Dostupné z: https://arxiv.org/abs/2404.04258

LIANG, W., et al. GPT Detectors are Biased Against Non-native English Writers. Patterns. 2023, roč. 4, č. 9. [online]. [cit. 2025-01-19]. Dostupné z: https://doi.org/10.1016/j.patter.2023.100779

MITCHELL, E., et al. DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature. arXiv. [online]. 26. leden 2023 [cit. 2025-01-19]. Dostupné z: https://arxiv.org/abs/2301.11305

RUSSELL, Stuart a NORVIG, Peter. Artificial Intelligence: A Modern Approach. Harlow: Pearson, 2021. ISBN není uvedeno.

TANG, R., CHUANG, Y.-N. a HU, X. The Science of Detecting LLM-Generated Text. Communications of the ACM. 2024, roč. 67, č. 4. Místo vydání není uvedeno.

TURING, Alan Mathison. Computing Machinery and Intelligence. Mind. [online]. 1950 [cit. 2025-01-19]. Dostupné z: https://doi.org/10.1093/mind/LIX.236.433

UNESCO. ChatGPT and Artificial Intelligence in Higher Education: Quick Start Guide. Paris: UNESCO, 2023. [online]. [cit. 2025-01-19]. Dostupné z: https://unesdoc.unesco.org/ark:/48223/pf0000384721

VASWANI, Ashish, et al. Attention Is All You Need. arXiv. [online]. 2017 [cit. 2025-01-19]. Dostupné z: https://arxiv.org/abs/1706.03762

WEBER-WULFF, D., et al. Testing of Detection Tools for AI-Generated Text. arXiv. [online]. 21. červen 2023 [cit. 2025-01-19]. Dostupné z: https://arxiv.org/abs/2306.15666

WU, J., et al. A Survey on LLM-generated Text Detection: Necessity, Methods, and Future Directions. arXiv. [online]. 23. říjen 2023 [cit. 2025-01-19]. Dostupné z: https://arxiv.org/abs/2310.14724

WU, J., et al. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. Computational Linguistics. [online]. 2025, roč. 51, č. 1, s. 275–338 [cit. 2025-01-19]. Dostupné z: https://aclanthology.org/2025.cl-1.8/

Přílohy

Dataset byl nahrán do příloh při odevzdávání elektronické verze bakalářské práce.