

Univerzita Hradec Králové
Fakulta informatiky a managementu
Katedra informačních technologií

Transformery a zpracování textu
Optimalizace dotazování velkých jazykových modelů
Diplomová práce

Autor: Bc. Michael Bartoš
Studijní obor: Aplikovaná informatika

Vedoucí práce: Ing. Karel Mls, Ph.D.

Prohlášení:

Prohlašuji, že jsem diplomovou práci zpracoval samostatně a s použitím uvedené literatury.

V Hradci Králové dne 11.8.2023

vlastnoruční podpis

Jméno a Příjmení

Poděkování:

Děkuji vedoucímu diplomové práce Ing. Karel Mls, Ph.D. za časté a nápomocné konzultace. Dále děkuji svým blízkým za podporu, kterou jsem pociťoval po celou dobu tvorby práce.

Anotace

Diplomová práce je rešerší tvořící holistický pohled na optimalizaci, respektive limitace dotazovacích metod a práce s velkými jazykovými modely, dále jen LLM, s transformer architekturou. Cílem je prezentovat metody použitelné k dosažení co nejlepších výsledků, ale i lokalizovat původce limitací mimo dotazování.

Práce popisuje jednotlivé části LLM od principů zpracování přirozeného jazyka, dále jen NLP, až po detaily designování dotazů a je zakončena souhrnem směru vývoje LLM ve vztahu k jednoduchosti použití. Významnost práce spočívá v pohledu na problematiku skrze komunikaci mezi člověkem a strojem. Oba disponují jiným vnímáním světa a jinými vlastnostmi a schopnostmi fungování v něm.

Výsledky rešerše ukazují, že v budoucnu budou techniky dotazování jiné, než je známe dnes a zároveň budou LLM schopnější díky efektivnějším metodám trénování nebo změnám v transformer architektuře.

Annotation

Title: Transformers and text processing

The diploma thesis is a research work that presents a holistic view of optimizing and addressing the limitations of prompting methods and working with LLMs using transformer architecture. The goal is to present methods applicable to achieving the best results and identify the origins of limitations beyond prompting. The thesis describes various aspects of LLM, from the principles of NLP to details about prompt-engineering, and concludes with a summary of the direction of LLM development concerning ease of use. The significance of the work lies in the perspective on the issue through the communication between humans and machines. Both possess different perceptions of the world and different characteristics and abilities to function within it. The research results indicate that in the future, prompting techniques will be different from what we know today, and LLMs will be more capable due to more efficient training methods or changes in the transformer architecture.

Seznam zkratk

AI – umělá inteligence (z anglického artificial intelligence)

AGI – Obecná umělá inteligence (z anglického artificial general intelligence)

ANN – umělá neuronová síť (z anglického artificial neural networks)

LLM – velký jazykový model (z anglického large language model)

NLP – Zpracování přirozeného jazyka (z anglického natural language processing)

NLU – pochopení přirozeného jazyka (z anglického natural language understanding)

NLG – generování přirozeného jazyka (z anglického natural language generation)

NER – Rozpoznávání entit v textu (volně přeloženo z anglického Named Entity Recognition)

PoS označování – Označování slovními druhy (z anglického Part of Speech tagging)

RNN – rekurentní neuronové sítě (z anglického recurrent neural networks)

LSTM – Dlouhá krátkodobá paměť (volně přeloženo z anglického Long Short-Term Memory)

CoT – Řetězení myšlenek (z anglického chain of thought)

GPT – Generativní předtrénovaný transformer (z anglického Generative Pre-Trained Transformer)

BERT – Oboustranné kodérové reprezentace z transformerů (přeloženo z Bidirectional Encoder Representations from Transformers)

RLHF – Zpětnovazební učení s lidskou zpětnou vazbou (z anglického Reinforcement Learning from Human Feedback)

SotA – Aktuálně nejlepší, v kontextu technologií (volný překlad z anglického State of the Art)

API – Rozhraní pro komunikaci (volný překlad z anglického Application Programming Interface)

Obsah

1	Úvod.....	1
2	Cíl práce.....	2
3	Metodika zpracování	3
4	Vlastní text práce	4
4.1	Umělá inteligence.....	4
4.2	Neuronové sítě.....	4
4.3	Zpracování přirozeného jazyka.....	5
4.3.1	Odstraňování stop slov	6
4.3.2	Stemming a lemmatizace	6
4.3.3	Anotace	6
4.3.4	Tokenizace	6
4.3.5	Vektorizace	7
4.4	Word Embedding	10
4.4.1	Dělení.....	10
4.4.2	Word2vec	11
4.4.3	GloVe.....	11
4.4.4	BERT/GPT embedding.....	11
4.4.5	Limitace.....	11
4.5	Transformer.....	16
4.5.1	Self-attention	17
4.5.2	Architektura	17
4.5.3	Enkodér	17
4.5.4	Dekodér	18
4.5.5	Dotazování	19

4.5.6	Limitace architektury	20
4.5.7	Úpravy transformerů	20
4.6	Velké jazykové modely	21
4.6.1	Dělení modelů	22
4.6.2	Základní funkce	24
4.6.3	Komunikace s LLM	24
4.6.4	Typové aplikace	24
4.6.5	Kontextový rámec	28
4.6.6	Emergentní chování	28
4.6.7	Limitace	29
4.6.8	Nastavování	31
4.6.9	Finetuning	32
4.6.10	Trénování	36
4.7	Benchmarky	37
4.7.1	BigBench	37
4.7.2	GSM8K	37
4.7.3	SuperGLUE	37
4.7.4	SQUAD2.0	38
4.7.5	LAMBADA	38
4.7.6	HellaSwag	39
4.7.7	CoQA	39
4.7.8	LogiQA	39
4.7.9	MMLU	40
4.8	GPT modely	40
4.8.1	GPT-1 a GPT-2	40
4.8.2	GPT-3	41

4.8.3	InstructGPT.....	41
4.8.4	GPT-3.5 (ChatGPT).....	41
4.8.5	GPT-4.....	41
4.8.6	BARD jako alternativa k GPT.....	42
4.9	Aplikace s GPT modely	42
4.9.1	ChatGPT.....	42
4.9.2	Bing.....	44
4.10	Design dotazu	45
4.10.1	Důležitost designování dotazů	45
4.10.2	Dělení a typy dotazů	46
4.10.3	Metody dotazování.....	47
4.10.4	Vnímání textu.....	51
4.10.5	Pravidla efektivních dotazů	52
4.10.6	Na co si dát pozor.....	54
4.10.7	Automatizované objevování dotazů	56
4.11	Nebezpečí	60
4.11.1	Špatné rady	60
4.11.2	Nerozeznatelnost původu textů.....	61
4.11.3	Únik citlivých údajů	61
4.11.4	Phishing.....	61
4.11.5	Vytváření malwaru.....	62
4.11.6	Prompt injection.....	62
4.11.7	Ochrana.....	64
4.11.8	Obcházení filtrů	64
4.12	Budoucnost.....	66
4.12.1	Důsledky očekávání společnosti (GPT-3 komentář).....	67

4.12.2	Zákon škálování.....	67
4.12.3	LLM	68
4.12.4	Trénování.....	69
4.12.5	Cenzura.....	70
4.12.6	Interakce ve světě.....	70
5	Závěry a doporučení	72
6	Seznam použité literatury	74

Seznam obrázků

Obrázek 1: Ilustrace plně propojené neuronové sítě.....	4
Obrázek 2: Ukázka umělého neuronu	5
Obrázek 3: Ukázka Bag of Words metody bez odstranění stopslov	8
Obrázek 4: Tabulka nenormalizovaného Tf-idf, kde f = frekvence slov v dokumentu, l = délka dokumentu. N = celkový počet dokumentů, DF =počet dokumentů kde se slovo vyskytuje.	9
Obrázek 5: Zobrazení Word2vec embedding prostoru pro slovo „mouse“ v grafickém nástroji. Modře podbarvená jsou nejbližší sousední slova k „mouse“, zeleně podbarvené slovo „mickey“ je první nejbližší soused.....	12
Obrázek 6: Volba slov ChatGPT při doplňování do věty	15
Obrázek 7: Schéma architektury prvního transformer modelu. Enkodér se nachází na levé straně a dekodér na straně pravé.....	19
Obrázek 8: Paralelní zpracování dotazu modelem BERT, který doplňuje chybějící text.....	23
Obrázek 9: Vizualizace sekvenční generace textu v auto-regresním modelu. Ilustrace ukazuje sekvenci pěti kroků, respektive slov. Jednotlivá navržená slova jsou na sebe vázaná, což znamená jiný výběr při změně předchozího slova. Pravděpodobnost slova je vyznačena barevně, jako vhodné slovo je v každém kroku vybíráno slovo s největší pravděpodobností.....	23
Obrázek 10: Generování e-mailu z heslovitého zadání, e-mail vypadá profesionálně, ale obsahuje zvláště zvolené předložky jako „na školu“ a „ve vašem předmětu“.	25
Obrázek 11: Ukázka rozdílu výstupu z modelu bez a s RLHF. V prvním případě je prompt, který je jednoznačným zadáním nepochopen a zpracován jako seznam úloh. V druhém případě je výsledkem e-mail přesně podle lidského očekávání.	33
Obrázek 12: Diagram paradigmatu prompt-based finetuning, v tomto případě je výstup z menšího modelu spojen s původním dotazem a použit jako vstup do LLM.	36
Obrázek 13: Ukázka few-shot learning k získání názvu země původu značky automobilů Citroen. Do prázdné kolony na odpověď model dosadí podle vzoru z předchozích příkladů název státu.....	48

Obrázek 14: Ilustrace vícekrokového řešení ze zero-shot CoT na slovní úloze z 5. třídy základní školy. Odpověď je chybná správný výsledek je rok 2012.....	48
Obrázek 15: Ilustrace sebekritiky na vyžádání a generování nového textu na základě kritiky. Odstavec s kritikou byl zkrácen na výpis kritizovaných vlastností textu....	50
Obrázek 16: Ukázka nejednoznačné reference v anglickém jazyce a chybné zpracování modelem BARD. Zájmeno „he“ může ve větě dotazu referovat na slovíčko „lawyer“ i „witness“.....	55
Obrázek 17: Vygenerovaný phishingový e-mail za pomoci dělení informací, bez dělení dotazu se e-mail nevygeneruje.....	65

1 Úvod

Od počátku věků lidé využívali skupinové činnosti a delegaci práce pro snadnější vykonávání náročných úkonů. Ať už jde o lidi nebo o jedinečný vztah psů, tehdy ještě vlků s předky dnešního člověka, v moderní době se nemusíme omezovat na inteligenci zvířat a lidí. Jelikož s nástupem AI lze mít v různých aspektech inteligenci vždy po ruce, v počítači či na internetu. AI nabývá mnoha podob podle toho, jakou úlohu má vykonávat. Nezávisle na její podobě a úloze je důležité umět s ní správně komunikovat tak, aby vykonávala úlohy co nejpřesněji a nejefektivněji podle našich představ nebo je i předčila.

V případě LLM je používán přirozený jazyk pro dotazování s cílem získat výsledky. Motivačním příkladem delegování práce na AI v podobě LLM ChatGPT s pozitivním dopadem na produktivitu i kvalitu výsledné odvedené práce, je výsledek výzkumu[1] na MIT.

Tato diplomová práce si klade otázku, jaké jsou limitace dotazování při práci s LLM, jaká je jejich závažnost, čím je konkrétní limitace způsobena a zda bude komunikace s LLM v budoucnosti jednodušší. S účinností delegace práce souvisí i vývoj schopností LLM, díky čemuž je pohlíženo i na schopnosti modelů a úloh, které na ně můžeme delegovat.

2 Cíl práce

Cílem této práce je popsat metody vedoucí k optimalizaci, respektive minimalizaci limitací, dotazování společně s vytvořením vhledu do problematiky auto-regresních LLM typu GPT a souhrnu limitací ve srozumitelné podobě tak, aby měl čtenář přehled, jak LLM fungují a na co je dobré je užívat.

Součástí rešerše je analýza limitací souvisejících převážně s dotazováním a schopnostmi modelů. Výzkumnou otázkou je, jak co nejefektivněji komunikovat s LLM s účelem získat co nejlepší výsledky. Navazující otázkou je jaké problémy LLM mají v aktuální podobě a zda se tyto problémy mají tendenci v budoucnosti měnit ku prospěchu užitelnosti, respektive zlepšování vlastností LLM.

Což vede k otázce, kam se aktuálně trend LLM ubírá v oblastí architektury, trénování a sofistikovaných metod dotazování a jak se tyto změny projevují při práci s nimi. Obecně řečeno, zda je práce s LLM čím dál lehčí a dostupnější. To lze při pohledu na nejnovější pokroky extrapolovat a určit směr pokroku v LLM do budoucnosti, na jaké oblasti budou nejčastěji aplikovány a podobně.

3 Metodika zpracování

LLM jsou z velké části blackbox modely, ve kterých jsou postupně objeveny různé vnitřní mechanismy a struktury. Hledání a správná interpretace objevených mechanismů a struktur je časově náročná, kvůli čemuž se tato práce ve většině případech opírá o výsledky z jiných vědeckých prací. Tyto opěrné práce zkoumající jednotlivé problémy do hloubky s náležitými metrikami. Díky tomuto postupu se pohlíží na výsledky citovaných dílčích prací holisticky ve vztazích mezi sebou.

Výsledkem této práce je díky tomu pochopení vlastností LLM přinášející co nejefektivnější užívání při aplikování nabytých znalostí. V prvních kapitolách jsou z důvodu pochopení, jaké komponenty stojí za vlastnostmi LLM, povrchově zmíněny základní principy a mechanismy NLP a transformeru. Na to navazuje další obecná kapitola o LLM a práce s nimi. Následující kapitoly se detailně zabývají limitacemi, poznatky o dotazování a vlastnostech LLM.

Vlastní praktický výzkum je omezený z důvodu časové náročnosti na zkoumání dílčích problémů souvisejících s použitelností a je limitován převážně na ilustrace vytknutých problémů. Ilustrace výstupů LLM jsou realizovány na ChatGPT, kvůli dostupnosti a rozšířenosti tohoto nástroje. Je však předpokládána přenositelnost poznatků na jiné nástroje užívající auto-regresní LLM jako například BARD.

Autor si uvědomuje možnou existenci dotazů, které dosáhnou lepších výsledků nebo vyvrátí tvrzení podkládané konverzací s ChatGPT. Je však přesvědčen, že dotazování by mělo být z důvodů použitelnosti pro velké spektrum lidí, co nejintuitivnější a v souladu s lidským chápáním než naopak, kde se musí pro přizpůsobení LLM složitě designovat funkční dotazy [2].

Z tohoto důvodu přikládá důležitost výsledkům z dotazů, které použil na základě svých a znalostí nabytých před a během tvorby této diplomové práce. Díky velmi rychlému vývoji v oblasti LLM pohlíženo i na vývoj za dobu psaní práce tak, aby bylo vidět, které problémy byly v tomto časovém intervalu vyřešeny a jakým způsobem byly řešeny. Tím pádem, jaké změny zapříčinily velký skok správným směrem k užívání LLM na širší doménu úloh.

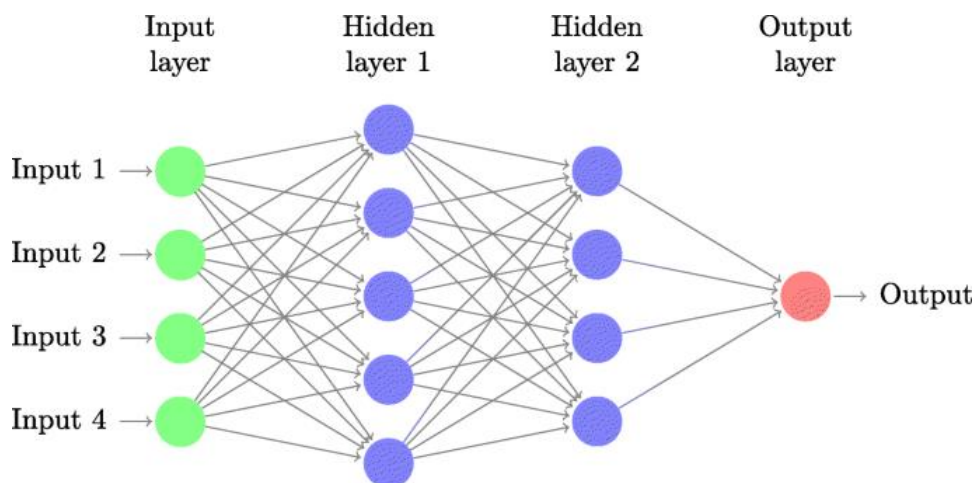
4 Vlastní text práce

4.1 Umělá inteligence

AI je velmi obecný pojem. Definicí inteligence je několik a většina z nich je nekonkrétní. Často se však shodují v tom, že inteligencí je učení se ze svých předešlých zkušeností. Software, který je schopen se učit, adaptovat a za pomoci toho vykonávat úlohy vyžadující jistou formu inteligence, kterou bychom mohli očekávat například od člověka, je považován za AI. Metody a struktury využívané k vytváření AI se za roky měnily a v současnosti hrají prim hlavně umělé neuronové sítě, v textu označovány jako ANN, využívané právě i v transformer architektuře.

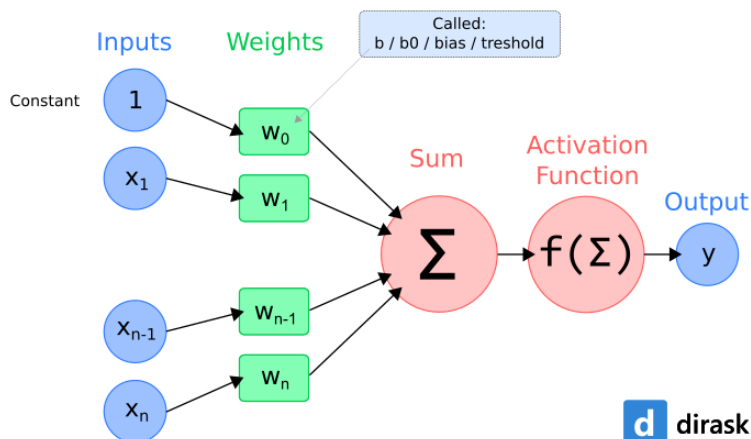
4.2 Neuronové sítě

ANN jsou matematickou abstrakcí biologických neuronových sítí, které se vyskytují u živých organismů a které mají v mozku i lidé. Stavebním prvkem těchto sítí je jediný neuron. Neurony na sebe mohou být napojeny a vytvářet sítě různých topologií. Různé topologie jsou dobré na rozličné úlohy a práce s ANN jsou empirického charakteru. Kromě topologií a propojení se ANN rozlišují na hluboké a mělké (z anglického „deep“ a „shallow“), kde hluboká síť je síť třemi či více skrytými vrstvami. Na obrázku 1, je vidět rozdělení vrstev, na vstupní skryté a výstupní. Vstupních neuronů je tolik, kolik máme informací k vyhodnocení a výstupních zas tolik, kolik jich je potřeba na reprezentaci získávané informace.



Obrázek 1: Ilustrace plně propojené neuronové sítě
,převzato z [3].

Abstrakce zjednodušuje neuron na poměrně jednoduchý prvek, který má v podstatě pouze vstupy, vnitřní logiku a výstup. Podrobněji je ukázán na obrázku 2. Vstupy jsou násobeny váhami a jejich hodnota se sčítá před vstupem. Při sčítání je přičítán i vstup s názvem „bias“, který realizuje jakousi odchylku danou jedinou konstantou. Sečtené vstupy pak rozhodují, zda se neuron aktivuje nebo ne, to závisí nejen na hodnotě výsledného vstupu, ale i použité aktivační funkci.



Obrázek 2: Ukázka umělého neuronu
 ,převzato z [4]

Obor učení AI se souhrnně označuje jako strojové učení. Učení ANN sítí probíhá například pomocí metody backpropagation a projevuje se měněním hodnot vah a „bias“ na jednotlivých neuronech v celé síti tak, aby co nejvíc vstupních dat bylo vyhodnoceno podle očekávání výstupu.

4.3 Zpracování přirozeného jazyka

V textu označováno zkratkou NLP. Zabývá se, jak již z názvu vyplývá úlohami s textem, jedná se například o překlady, klasifikace textu, extrakce informací a detekování spamu. Kromě NLP, se používá také NLU pro pochopení přirozeného jazyka a NLG pro generování přirozeného jazyka. Jelikož neuronové sítě nedokážou zpracovávat textová data je nejdůležitější část převod textových dat na data, kterým modely umělé inteligence rozumí, tedy číselná. LLM založené na transformer architektuře spadají do kategorie NLP a jsou nejnovějším velkým průlomem v oblasti NLP, ale i AI jako celku. Pro úplnost jsou v této kapitole zmíněny základní principy a metody pro práci s textem v NLP.

4.3.1 Odstraňování stop slov

Jedná se o odstraňování slov, která nenesou důležité informace a jde převážně o různé syntaktické úpravy. V češtině jsou to hlavně spojky, předložky diakritika a v jiných jazycích se mohou tato slova lišit.

4.3.2 Stemming a lemmatizace

Cílem je ze slov vytvořit kořenový tvar. Stemming pouze oseká slovo do kořenového tvaru. Lemmatizace, je složitější získává kořenový tvar na základě znalosti tvarosloví.

4.3.3 Anotace

Anotacemi je myšleno jakékoliv označování slov, tokenů nebo jiných entit v textu. Nejznámější dva druhy jsou:

- **PoS tagging (Part of speech tagging)** – Označení jednotlivých slov jejich slovními druhy, tak aby bylo jednodušší pochopit stavbu větu. Využívá se především v jazycích s volnější strukturou vět, kde se pozice důležitých slov mění, což je pro vyhodnocování pomocí matematických modelů matoucí. Dá se používat pro získání klíčových frází ať už za pomoci jednoduché struktury frází, vzorů regulérních výrazů (zpravidla označovány jako „regex“) nebo pro vymezení typu slov, která mohou být klíčová. Označení je způsob, jak do textu vložit víc užitečných informací, kterých se lze chytit při procesu pochopení textu.
- **NER (Named Entity Recognition)** – Proces získávání kýžených informací z textu, pro další použití. Získané informace mohou být například užity pro automatizaci, kde se získají údaje, jako například lokace, jméno, organizace, které jsou potřeba k vyplnění například formuláře, nebo automatického zobrazení profilu zákazníka na základě údajů v textu emailu.

4.3.4 Tokenizace

Proces převádění textu na části nazývané tokeny, s kterými se dále pracuje. Může se jednat například o n-gramy, slova nebo i jejich části, které se dají dělit až na

jednotlivá písmena. Jak mají tokeny vypadat záleží na konkrétních úlohách, drobení tokenů na menší části, než slova přináší výhodu větší tolerance chyb v textu.

4.3.5 Vektorizace

Reprezentace slov se v NLP provádí například metodami anglicky nazývanými „one hot encoding“ nebo „word embedding“. Obě metody spočívají v převedení tokenů zmíněných v předchozí kapitole na číselná data v podobě polí/vektorů. V prvním případě jde o techniky zaznamenávání frekvence slov do vektorů. A v druhém případě jde o modernější přístup vytváření embedding zaznamenávající význam slov v mnohorozměrném vektorovém prostoru, který je z důvodu využití ve veškerých LLM detailněji vysvětlen v samostatné [kapitole \(4.4\)](#), jsou to vektory tvořící mnohorozměrný prostor, který je schopen lépe pojmout informace z textů.

Během procesu změny reprezentace textové informace na číselnou se však ztrácí informace. Z tohoto důvodu existuje více různě komplexních metod, s výhodami a nevýhodami, jak převodu dat dosáhnout a zachovat pro vykonávanou úlohu důležité informace. V podkapitole jsou zmíněné metody vektorizace slov, respektive tokenů, ovšem jde vektorizovat i na úrovni znaků, například pomocí metod ELMo [5] a Flair [6]. U jednoduchých metod je veškerá informace dána frekvencí slov nebo je z ní vypočítána, čímž se vytrácí jakýkoliv kontext.

4.3.5.1 Bag of Words

1. Auto jede do města a vedle něj jede kolo.
2. Kolo jede z města vedle řeky.

Slovník	Auto	jede	do	města	a	vedle	něj	kolo	z	řeky
První věta	1	2	1	1	1	1	1	1	0	0
Druhá věta	0	1	0	1	0	1	0	1	1	1

Obrázek 3: Ukázka Bag of Words metody bez odstranění stopslov

Nejjednodušší metoda, převádí do číselné podoby tokeny a jejich počet v textu. Na obrázku 3 je zřetelné, že informace o slovosledu a významu slov je úplně ztracena, tím pádem nejsou věty skládajících se ze stejných slov od sebe rozpoznatelné i pokud mají jiný či dokonce opačný význam. V praxi je tak pro každou větu vytvořen vektor o velikosti počtu unikátních slov v textu. Slova zaznamenaná ve vektoru se nazývají slovník. Každý element vektoru je přiřazen slovu ze slovníku a je v něm zaznamenán výskyt jednotlivých slov v dané větě. Přestože se často udává že jde o binární vektor, často se jedná o vektor s frekvencemi výskytu slov. Ztrácí jakékoliv informace o kontextu a mezislovními vztahy, ale postačuje na úlohy klasifikace a shlukování.

4.3.5.2 TF-IDF

TF-IDF (zkratka z anglického Terms Frequency – Inverse Document Frequency) je statistickou metodou pracující nad více dokumenty. Stejně jako Bag

1. Puding je želatinová hmota.
2. Děti, puding. Děti mají rády puding.
3. Děti jsou rády, že mají kolo.

Slovník	TF=f/l			IDF=log(N/DF)	W=TF*IDF		
	1	2	3	3	1	2	3
<i>Puding</i>	1/4	2/7	0	Log (3/2)	0,044	0,050	0
<i>Je</i>	1/4	0	0	Log (3/1)	0,119	0	0
<i>Želatinová</i>	1/4	0	0	Log (3/1)	0,119	0	0
<i>Hmota</i>	1/4	0	0	Log (3/1)	0,119	0	0
<i>Děti</i>	0	2/7	1/6	Log (3/2)	0	0,050	0,029
<i>Mají</i>	0	1/7	1/6	Log (3/2)	0	0,025	0,029
<i>Rády</i>	0	1/7	1/6	Log (3/2)	0	0,025	0,029
<i>Jsou</i>	0	0	1/6	Log (3/1)	0	0	0,080
<i>Že</i>	0	0	1/6	Log (3/1)	0	0	0,080
<i>Kolo</i>	0	0	1/6	Log (3/1)	0	0	0,080

Obrázek 4: Tabulka nenormalizovaného Tf-idf, kde f = frekvence slov v dokumentu, l = délka dokumentu. N = celkový počet dokumentů, DF = počet dokumentů kde se slovo vyskytuje.

of Words je informace vkládaná do vektorů o velikosti slovníku. Rozdílem jsou zaznamenávané hodnoty, které jsou váhy vypočítané z hodnot frekvencí slov v textu a inverzní frekvencí slov napříč dokumenty. Dokumenty jsou myšleny například články, odstavce nebo jiné textové jednotky. Váhy reprezentují důležitost slov a vychází z teorie, že jedinečná slova jsou informačně signifikantní, naopak častá slova nesou nejméně informace a jsou tak označeny nejmenší vahou. Dobrým příkladem

velmi častých slov jsou právě stop slova, která se úplně odstraňují ještě, než dojde k tokenizaci a vektorizaci. Obdobně i neodstraněná častá slova jsou méně důležitá. Ukázka výpočtu TF-IDF je na obrázku 4.

4.4 Word Embedding

Označení embedding se používá pro jakékoliv zaznamenávání reprezentace slov včetně TF-IDF, Bag of Words zmíněné v předchozích kapitolách. Z hlediska zaměření na LLM je pod tímto označením myšlen embedding implementovaný jako naučená neuronová síť. Embedding je buď trénován samostatně nebo vzniká jako vedlejší produkt při trénování LLM a je prostorem znalostí s kterými LLM pracují. Mapuje kategorické, nominální, tedy diskrétní hodnoty do kontinuální vektorové podoby.

Vektorový prostor si lze představit jako více dimenzionální mapu informací ohledně vztahů mezi slovy zaznamenanými v prostoru. Jelikož je tvořen vektory, je možné používat na slovech základní matematické maticové operace. Ty jsou využívány právě na práci s embedding a jedná se například o podobnost, která se dá získat jako doplněk kosinové vzdálenosti mezi jednotlivými vektory reprezentujícími slova. Podobnost je obecná vlastnost, jejíž hodnoty se získávají učením napříč různými daty a jejich kontexty, kromě podobnosti lze zaznamenat i sofistikovanější informace jako kontext, sémantiku, způsob typického používání a další. Dvěma úspěšnými a historicky důležitými embedding jsou Word2vec a GloVe, mají podobnou funkci, ale jsou učeny na jiných principech [7].

4.4.1 Dělení

Embeddingy jsou děleny na statické, dynamické, kontextové, subword (jsou tvořeny podslovy), konceptuální, znalostní atd. Znalostní jsou založené na datech z trénovacího korpusu a znalostních zdrojů, kde jsou slova s pevně definovanými významy, kontexty apod. Znalostní zdroje jsou například Wordnet, Wikipedia, DBpedia apod. Toto dělení není exkluzivní a lze tak narazit i na neznalostní embedding pouze obohacený o znalostní informace. Žádný druh embedding není nejlepší, ale je zvolen nejvhodnější druh a dále je upravován k dosažení co nejlepších výsledků na kýžené úloze.

4.4.2 Word2vec

Jako řešení ztráty sémantické informace byla Tomášem Mikolovem pracujícím pro Google vytvořena metoda se jménem Word2vec, která slova převádí na vektory, se záznamem vztahů k atributům v lokálním kontextu vytvářejícím sémantiku mezi slovy. Podobnost mezi slovy je dána kosinovou vzdáleností ve vektorovém prostoru. Co se týče implementace jedná se o dopřednou neuronovou síť (anglicky feed-forward) s jednou skrytou vrstvou. Disponuje dvěma způsoby fungování, které jsou v inverzním vztahu. Jakým způsobem bude Word2vec fungovat závisí na architektuře a způsobu jakým byl naučen, de facto jsou to dva odlišné modely. Aby nemusel být model učen při každém použití je k dispozici i předtrénovaný Word2vec embedding od Googlu naučený na datasetu o 100 miliardách slov. [8] Jednou z funkcí/modelů, kterou Word2vec zvládá je Continuous bag-of-words a jedná se o dosazování slov do okolních pozic na základě známého slova. Druhá funkce, nazývaná skip-grams, je inverzní a dosazuje na pozici jedno slovo na základě slov, která se nacházejí na okolních pozicích. [9] Oba modely se používají podle úlohy použití a nedá se říct, že by jeden byl lepší než druhý [10].

4.4.3 GloVe

Vyvinut na Stanfordu, na rozdíl od Word2vec zaznamenává vlastnosti slov vůči celému korpusu, která se vyskytují spolu v páru a tvoří matici společného výskytu. Z této matice, lze pak získávat pravděpodobnosti společného výskytu slov a tím i vztahy mezi slovy. [11, 12]

4.4.4 BERT/GPT embedding

Na rozdíl od Word2vec se jedná o kontextové embedding. To znamená že například české slovo „los“ reprezentují různě podle kontextu, který upřesní, zda je myšleno zvíře či papírek s poukázkou šance na výhru. Word2vec by v obou případech vytvořil stejný vektor reprezentace slova „los“.

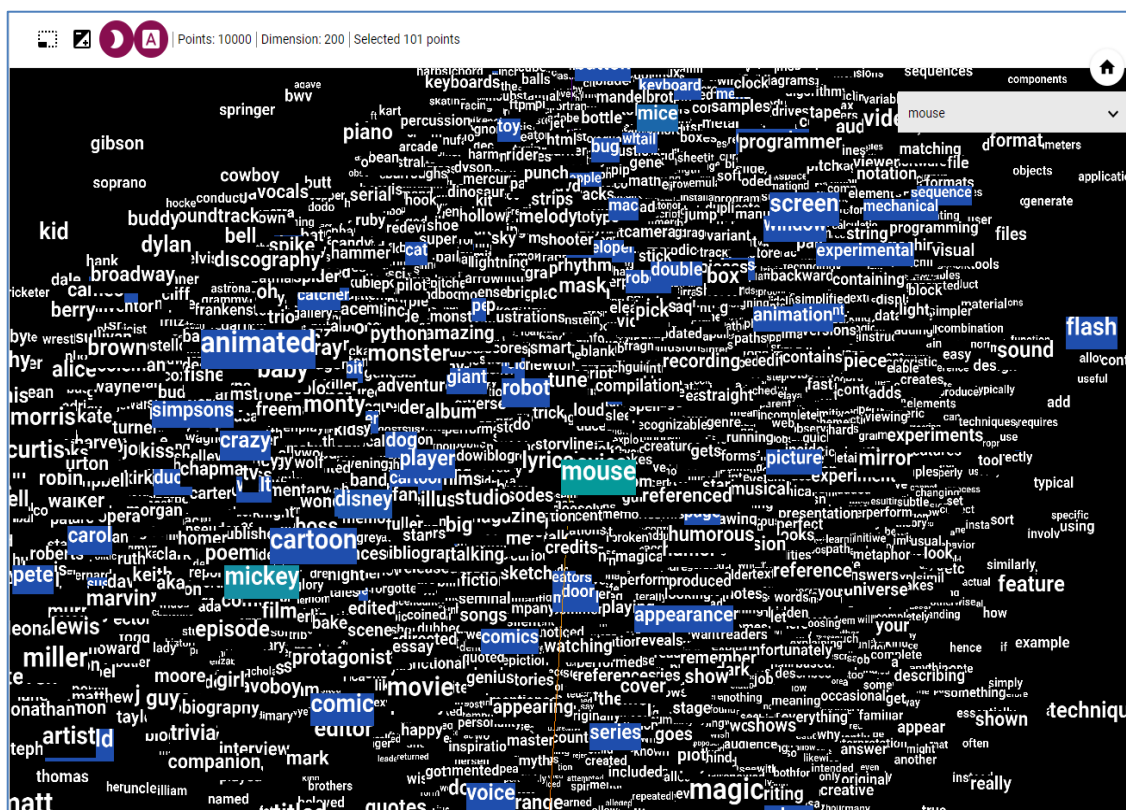
4.4.5 Limitace

Embedding není dokonalá reprezentace jazyka a znalostí, naráží tak na různé limitace převážně v ohledu reprezentace různých kontextů, významů, aktuálnosti.

Nedostatky se dále projevují i v LLM, které embedding využívají jako svou slovní a znalostní zásobu. Některé limity jsou exkluzivní pro konkrétní typy, jiné jsou zastoupeny u více typů.

4.4.5.1 Statické

Nejjednodušší druh, každé slovo je reprezentováno jedním vektorem. Není dále učen tím pádem nebere v potaz změny významů v čase. Absence víc kontextů



Obrázek 5: Zobrazení Word2vec embedding prostoru pro slovo „mouse“ v grafickém nástroji. Modře podbarvená jsou nejbližší sousední slova k „mouse“, zeleně podbarvené slovo „micky“ je první nejbližší soused.

Nástroj je dostupný na <http://projector.tensorflow.org/>.

(polysémie) a sémantických informací dělá interpretaci výstupů při práci s těmito embedding náročné. Důsledkem těchto problémů je významová konflace. Vzniká kvůli jednoduchému vztahu podobností mezi slovy slov s různými významy. Ilustrací tohoto problému je obrázek 5, kde je vidět výskyt i nepodobných slov vedle sebe. Kvůli blízkosti slova „micky“ je ke slovu „mouse“ blízko i spousta dalších slov souvisejících s animovanými seriály jako například prakticky nesouvisející „simpsons“.[13].

4.4.5.2 Dynamické

Berou v potaz změny v čase díky čemuž obsáhnou změny jazyka v čase. Jsou vhodné na zkoumání právě těchto změn a s tímto účelem byly na Stanfordské univerzitě vytvořeny i historické embedding trénované pouze na datech v konkrétních časových intervalech [14].

4.4.5.3 Kontextové

Kontextové jsou někdy označovány i jako dynamické, na rozdíl od nich však časové intervaly původu dat neberou v potaz. Kontextem je myšlen význam v textu. Každé slovo může mít mnoho kontextů a během učení nějaké mohou být vynechány nebo úplně nepochopeny a snížit tím použitelnost. Generování velkých kontextových embedding je obzvláště výpočetně náročné, více o této problematice je napsáno v [kapitole 4.5.6](#).

4.4.5.4 Znalostní

Ačkoliv obecně platí, že obohacení embedding znalostmi zlepšuje jejich použitelnost a do jisté míry řeší i některé zmíněné problémy napříč všemi druhy embedding. Je právě hlavní limitací závislost na kvalitě a obsáhlosti znalostních zdrojů. Závislostí je myšleno užívání pouze těchto zdrojů k trénování a tím opomíjení možnosti získat znalosti i jiným způsobem, který by mohl zlepšit výslednou užitečnost. Obsahově mohou data být zaujatá, nedostatečně zastoupená, nekonzistentní apod. Z hlediska trénování lze narazit na výpočetní náročnost, pokud jsou znalosti získávané z velkých ontologií.

4.4.5.5 Jazyková variabilita

Někdy označováno jako jazyková proměnná. Pro většinu úloh, kde je embedding použit se pravděpodobně jedná o zanedbatelný problém. Přesto je dobré brát v potaz neexaktnost přirozeného jazyka a možné dopady na systémy užívající přirozený jazyk. V každé kultuře, menších skupinách i v jednotlivcích existuje rozmanitost jazyka odrážející rozdílné úmysly, hodnoty a kulturní normy. Jazyková variabilita má podobu nářečí, rozdílných výslovností, gramatických rozdílů, slovní zásoby, stylistických rozdílů [15]. Formována je interakcí s reálným světem skrze

zkušenosti, vzdělání, komunikaci apod. Dochází tak k jemně rozdílným interpretacím a odlišnému chápání jazykových výrazů. Například rozdíly vnímání slov mezi lidmi se liší konotací, pohybující se mezi pozitivní a negativní hodnotou. Z důvodu jazykové variability embedding může obsahovat nepřesnosti, ale i kvůli subjektivní povaze jazyka nelze jazyk nikdy reprezentovat dokonale.

4.4.5.6 Slova mimo slovník

Embedding obsahuje pouze slova, s kterými se setkal během trénování modelu, a nastává tak problém při zpracování nových slov, která nejsou v embedding, respektive slovníku, zaznamenána. Existuje více metod, které lze provést, pokud se najdou chybějící slova, ale všechny ze své podstaty mohou zapříčinit redukci přesnosti zpracování textu, jelikož dochází ke ztrátě informace.

Ignorovat

Nejjednodušší způsob je vypuštění slova a generování výstupu, jako by ve vstupním textu nebylo obsaženo.

Nahrazení tokenem

Chybějící slova se nahrazují tokenem tak, aby byla jasná mezera v textu a dále se s touto skutečností mohlo pracovat.

Substituce/Hledání vhodného slova

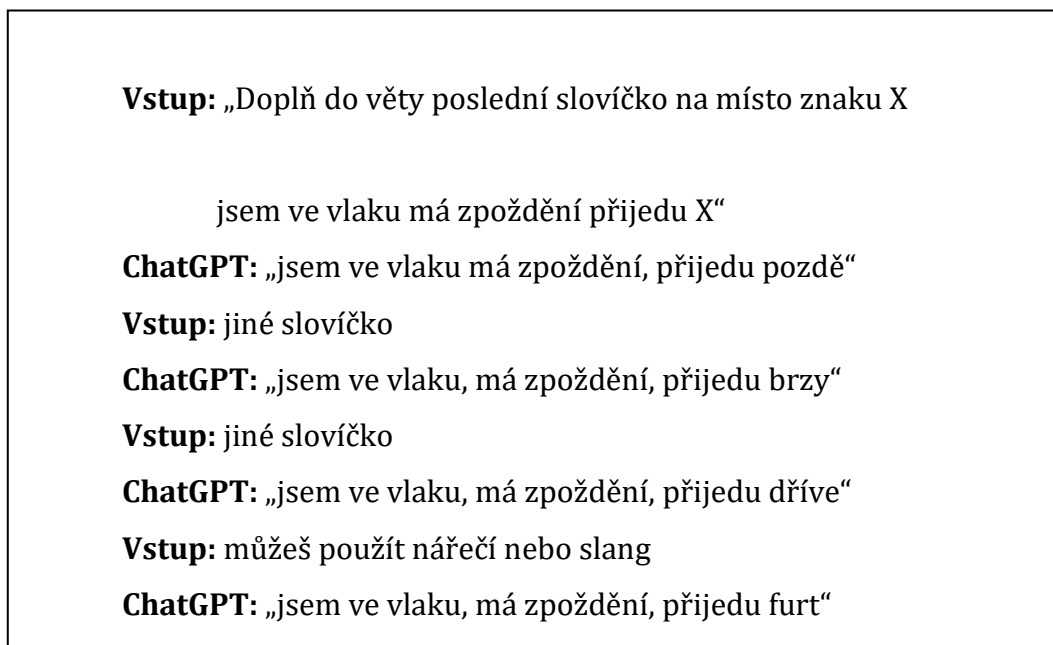
Nalezení obdobného slova nebo synonyma a nahrazení jím lze dosáhnout za pomoci dedikovaných slovníků nebo dotazování se LLM. Tímto způsobem se vybere nejvhodnější slovo do prázdného místa na základě zbytku textu.

Hledání podslov

Rozdělení řetězce znaků neznámého slova na podřetězce s cílem najít podřetězec obsažený ve slovníku, který bude použit jako náhrada se stejným významem.

4.4.5.7 Nedostatečné zastoupení slov

U jakékoliv úlohy strojového učení je kladen důraz na rozdělení dat tak, aby byly veškeré možné vstupy zastoupeny dostatečně a model nebyl naučený nebrat je v potaz. V případě tvoření velkých embedding z velkého objemu dat se však problém



Obrázek 6: Volba slov ChatGPT při doplňování do věty

se zastoupením projevuje, jelikož nelze velké množství dat kontrolovat, ať už z důvodu finančních, časových nebo obtížnosti úlohy kontroly. Malé zastoupení slov z nějaké oblasti se projevuje horší reprezentací významu, která pak dále negativně dopadá na kvalitu výstupu LLM pro tuto doménu. Příklady domén s malým zastoupením mohou být slang, nářečí, profesní žargon nebo méně známé vědecké oblasti. Ilustrací problému malého zastoupení dat je konverzace na obrázku 6. V konverzaci je vidět, že ChatGPT použije raději známá nesmyslná slova než méně známá smysluplná slova. Slovo volba slova „pozdě“ je objektivně správně obzvláště v absenci širšího kontextu. S širším kontextem se dá odůvodnit i volba slova „brzy“, ale volba slov „furt“ a „dříve“ je objektivně špatná. Slovo „furt“ sice souvisí s časem, ale nedává v tomto použití smysl. Slovo „dříve“ se správnosti přibližuje, přesto se jedná o nesmyslnou volbu, jelikož zpoždění implikuje pozdější příjezd. Pokud nastane situace, že vlak měl zpoždění a přijede dříve, tak již zpoždění dohnal a tím pádem není zpožděn.

4.4.5.8 Zaujatost v trénovacích datech

Obdobně jako malé zastoupení dat ovlivní výsledný embedding i zaujatost. Jedná se o stereotypy, různé diskriminace, ale i neškodné nepřesnosti a mylné domněnky měnící vztahy v embedding. Do embedding prostoru se tak propisují mezislovní vztahy připisované lidmi vzdalující se od reálné neutrální reprezentace. V případě mylných domněnek a nepřesností se reprezentace vzdaluje objektivním faktickým znalostem, ale v případě diskriminace a stereotypů je kvůli subjektivitě a nejednoznačnosti úmyslu, mimo extrémní případy, těžké rozpoznávat do jaké míry je zaujatost přítomna. Se zaujatostí je možné se setkat i v procesech jako jsou překlady [16].

4.4.5.9 Absence změn v čase

Až na dynamický embedding není žádný druh znalý času původu svých informací, pokud není zmíněn v textu svého trénování jako například datum vydání článku. Prakticky každý embedding je zmražen v čase od doby, kdy byl natrénován. Koncept „zastaralé“ informace není v jejich reprezentaci jazyka a znalostí přítomen. Nelze tak zaznamenat nově objevující se užití slov a aktuální trendy výrazy.

4.4.5.10 Blackbox

Přestože jsou některé embedding jmenovitě statické a znalostní celkem dobře „průhledné“ a interpretovatelné, nyní velmi rozšířené kontextové embedding jsou blackboxy. Mít k dispozici systémy bez chápání jejich funkce je problematické, ať už pro užití v kritických oblastech jako je medicína, ovládání aut nebo z principu kdy chceme mít jistotu, že je známo, co od systému čekat a jakými neduhy trpí. K získání vhledu se používají metody jako T-SNE [17] pro vizualizaci vzorů.

4.5 Transformer

Transformer je architektura AI navržená vědci ze společnosti Google, původně na úlohu překladů. Modely této architektury pracující se sekvencemi dat, ať už se jedná o text, číselná data či pixely v obrazových datech. Práce vědců prezentovala novou podobu AI, která se osvědčila převážně v úlohách NLP. Toho bylo dosaženo zlepšením vlastností oproti starším architekturám RNN a LSTM. [18]

Hlavním nedostatkem RNN je sekvenčnost, což znamená postupné zpracování textů a tím obtížná paralelizace potřebná pro rychlé učení a fungování velkých modelů. Nevýhodou z hlediska použitelnosti je i malý kontextový rámec omezující použitelnost na delší texty. LSTM nabízeli lepší použitelnost než RNN z hlediska výsledků, ale jejich trénování bylo ještě pomalejší než u RNN. Transformer se lepšími vlastnostmi získanými odstraněním rekurencí a konvolucí prosadily. Staly se fenoménem v oblasti NLP a pro mnohé jsou téměř synonymem pro LLM.

4.5.1 Self-attention

Attention je koncept zaměřování se na významná slova, ačkoliv není koncept attention [19] vycházející z úlohy překladů úplně novým objevem. Google ve svém výzkumu [18] ukázal koncept self-attention s nímž navrhl první architekturu modelu typu transformer. Self-attention je mechanismus vybírající nejrelevantnější části textu v jedné sekvenci (sekvencí je myšlen jakýkoliv textový celek). Kontext v transformerech je obsažen v tokenech zpracovávaných self-attention mechanismem. Tento mechanismus udává maximální možný počet tokenů paralelního zpracování tedy velikost kontextového rámce.

4.5.2 Architektura

Výhodou transformerů, jak již bylo zmíněno, je paralelizace zpracování a učení. Z tohoto důvodu do transformerů vstupují veškerá data najednou. Dělí se na dva funkční bloky složené ze stejných komponentů. Bloky se označují jako enkodér a dekodér. Vidět jsou na obrázku 7, kde je vizualizován celý původní návrh.

Novější modely modifikují původní navrženou architekturu dvou enkodér/dekodér bloků, například GPT a BERT obsahují část pouze jednu. V případě GPT jde pouze o dekodéry a u BERT pouze o enkodéry. Změny v architektuře a jejich odůvodnění je popsáno v [kapitole 4.8](#).

4.5.3 Enkodér

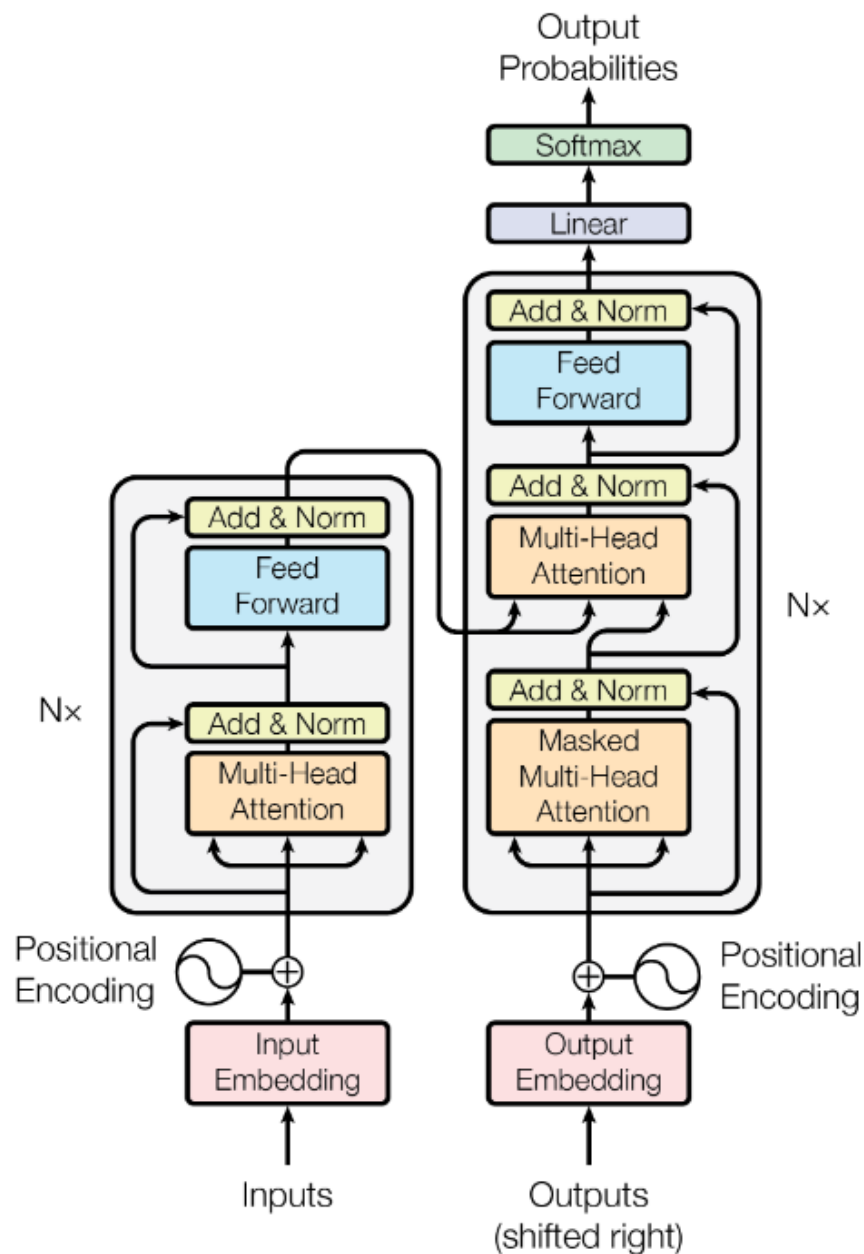
Data vstupují do modelu a putují přes embedding vrstvu převádějící slova na vektory, dále jsou vektory obohaceny informací o pozici v embedding prostoru. Výstupem enkodéru jsou kontextové vektory reprezentující vstupní data.

- **Multi-head attention** – Generuje attention vektory, které obsahují informaci o vztazích mezi slovy, která spolu souvisí a na které je tím pádem potřeba se zaměřit. V principu jde o vážení důležitosti všech slov ve vztahu s aktuálně zpracovávaným slovem. Attention vektorů se pro každé slovo generuje větší počet z nichž je pak vytvářen jeden výsledný. Ten je průměrem všech vygenerovaných attention vektorů. Tento proces existuje z důvodu chyb, které by mohly poukázat na jiné slovo, než jaké je aktuálně zpracováváno.
- **Feed-forward ANN** – Jedná se o ANN s jednou skrytou vrstvou pomáhající zaznamenání vztahů a vzorů ve vstupních datech. Dá se dobře paralelizovat použitím jedné této vrstvy na každý zpracovávaný vektor.
- **Normalizace** – Normalizuje data pro sjednocené zpracování. Také pomáhá stabilizaci trénování snižováním výkyvů v gradientech.

4.5.4 Dekodér

Výstupem je textová sekvence získaná na základě kontextové informace, získané z enkodéru.

- **Maskovaný self-attention** – Zaručuje získání attention vektorů pouze ze slov na předešlých pozicích k aktuálně zpracovávané pozici.
- **Lineární projekce** – Plně propojená ANN, která mění svou velikost podle počtu výstupních slov a provádí vážený součet.
- **Soft-max aktivace** – Převádí výstup lineární projekce na pravděpodobnostní rozdělení. Po tomto kroku jsou data pro lidi interpretovatelné v podobě pravděpodobnosti vhodnosti slov.
- **Generování slov** – Vybrání a přidání slova s největší pravděpodobností vhodnosti ke zbytku textu.



Obrázek 7: Schéma architektury prvního transformer modelu. Enkodér se nachází na levé straně a dekodér na straně pravé. Převzato z [18].

4.5.5 Dotazování

Transformery jsou podobně jako obyčejné ANN rozpoznávači vzorů a jejich inteligence je dosahována zaznamenáním vzorů, které jsou pro lidi příliš abstraktní. Ke zpracování dotazů dochází doplňováním dalších tokenů na základě zadaného dotazu. To znamená, že model našel vzor, podle kterého ví jakým způsobem dosadit

slova tak, aby na výstupu vznikl z dosazené části a zadané části na vstupu smysluplný text.

4.5.6 Limitace architektury

Největší limitací exkluzivní pro transformer architekturu je kvadratická časová náročnost – self-attention mechanismu. Ta kromě limitování velikosti kontextu ovlivňuje i dobu a cenu trénování. [20]

4.5.7 Úpravy transformerů

Transformery jsou v téměř všech populárních nasazeních užívané bez zásadnějších rozdílů od původní navržené architektury. Postupně jsou vytvářeny návrhy transformerů s upravenou architekturou nebo implementací komponentů. Cílem je vytvořit co nejméně náročný model se zachováním nebo zlepšením co nejvíce vlastností. Tím se vylepšuje použitelnost neboli dostupnost nasazení transformerů do produkčního prostředí. Kromě technických limitací v architektuře, omezující škálovatelnost transformerů, jsou adresovány i limitace úloh, které mohou zpracovávat. Další cestou zlepšení použitelnosti jsou změny v procesu učení nebo zakomponováním více systémů do jednoho celku pro transformery hůře zpracovatelné úlohy.

4.5.7.1 Zmenšení výpočetní náročnost self-attention mechanismu

Self-attention mechanismus se zjednodušuje zmenšením objemu zpracovaných dat nebo změnou způsobu výpočtů. Toho je dosahováno úpravou zpracovávaných vzorů, odstraněním části dat, ukládáním mezivýsledků, aproximací výpočtů a podobně. [21]

4.5.7.2 Destilace znalostí

Destilace modelů je proces přenosu znalostí do menších efektivnějších a tím pádem méně náročných modelů se snahou zachování vlastností původních LLM. [22]

4.5.7.3 Toolformer

Jedná se o vylepšení obcházející limitace za pomoci využívání nástrojů mimo LLM. Ty nejsou zdatné v aritmetických úkonech, mohou si vymýšlet informace, nemají pojem o toku času, a hlavně jsou jejich informace konstantní od doby naučení. Řešením je model naučený self-supervised způsobem, na volání nástroje vytvořeného pro zpracování kýžené úlohy přes API s parametry získanými z textu. Autoři toolformeru kladou důraz na zachování obecnosti modelu, jednoduchosti a přístupnosti jeho učení bez nutnosti anotovaných dat, což je ošetřeno použitím self-supervised učení. Dalším důležitým prvkem toolformeru je samostatné rozhodování kdy a jaké API zavolat. Jedná se o finetunování vylepšený GPT-J dosahující v mnoha úlohách lepších zero-shot výsledků než znatelně větší GPT-3. [23]

Nástroje

Použité nástroje jsou kalkulačka pro výpočet, kalendář pro získání aktuálního data pro úlohy, kde je důležité znát čas. Dále vyhledávací engine Wikipedie pro obsáhlejší informace, model na zodpovídání otázek na krátké ověřování faktů a překladač z jakéhokoliv jazyku do angličtiny.

Nedostatky

Toolformer nedokáže řetězit použití nástrojů tak, aby postupně dosáhl smysluplných výsledků a není interaktivní. Například nečeká na vstup uživatele potvrzující výběr jednoho z možných výsledků vyhledávání.

4.6 Velké jazykové modely

Velké jazykové modely rozumí přirozenému jazyku, skrz který se s těmito modely komunikuje. Realizují tak NLP, NLU i NLG z čehož vyplývá, že tvorba těchto systémů je komplexní problematikou. Jedná se o modely designované na zpracování velkého objemu textových dat využívající různé architektury mechanismů pracující s embedding reprezentací slov.

Jejich funkce, stejně jako u jiných problémů strojového učení, je nalezení nejlepšího vhodného výsledku, v tomto případě doplňujícího textu v zadaném kontextu. Za úspěchy LLM, které v oblasti NLP způsobily revoluci, v posledních

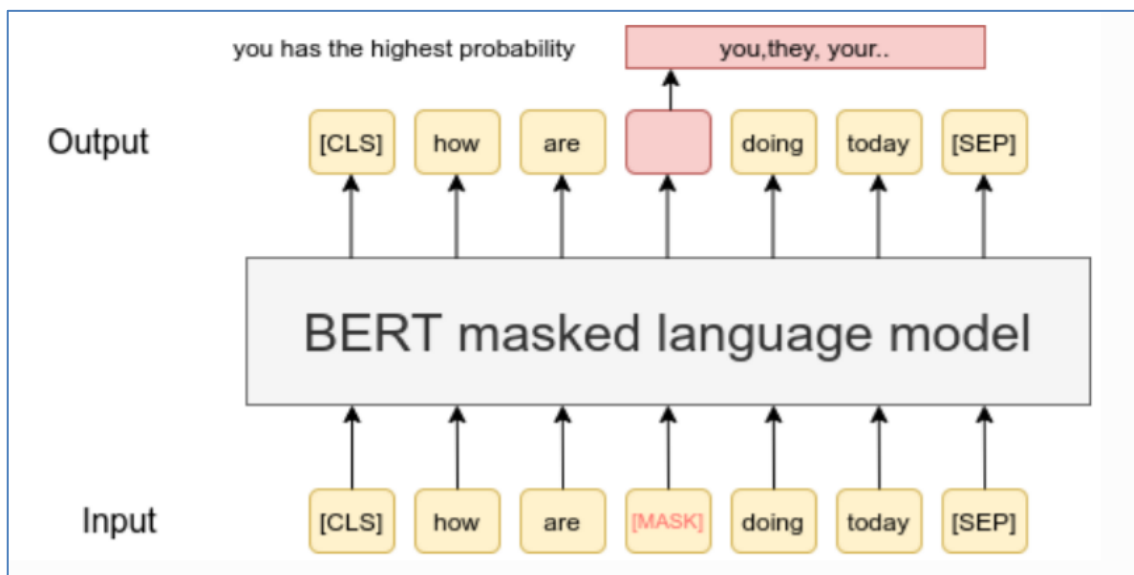
letech stojí po uvedení architektury transformerů hlavně velikost modelů a objem trénovacích dat. Za neustálým zvětšováním modelů, respektive jejich počtu parametrů nebo velikosti objemu učících dat, stojí teorie emergentního chování, která je popsána níže v [podkapitole 4.6.6](#). Transformer architektura se doposud zásadně nemění a tím pádem se pro dosažení lepších výsledků mění vše ostatní, ať už se jedná o způsoby učení nebo změny v implementaci komponentů architektury.

4.6.1 Dělení modelů

LLM se mohou dělit mnoha způsoby, podle architektury, jazyka, způsobu trénování a mnoha dalších kritérií. Dělení na maskované a auto-regresivní bylo zvoleno kvůli reprezentaci dvou rozdílných způsobů užívání modelů a práce s textem, které tyto varianty představují. Obě varianty LLM mají průnik úloh, které jsou schopny zpracovat, ale obě excelují v rozdílných úlohách. Představiteli obou variant jsou BERT pro maskované a GPT pro auto-regresivní, přestože jsou oba zvolení představitelé stavění na transformer architektuře, není toto dělení relevantní pouze pro tuto architekturu. S maskováním i auto-regresí se lze setkat i u konvolučních a rekurentních neuronových sítí a dalších.

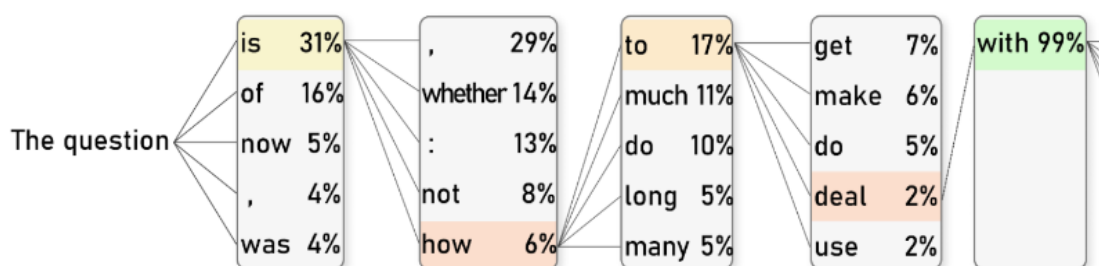
4.6.1.1 Maskovaný

Zpracovává slova textu paralelně během trénování i generování. Text dotazů je obohacen o maskovaný token, do kterého je dosazováno generované slovo. Tím, že se jedná obou směrnou variantu, je kontext zachycen lépe a paralelní zpracování nemění kontext během zpracování. Díky oboustrannému pohledu na text je robustní a není příliš náchylný na chyby způsobené “šumem” ve vstupních datech. Přes všechny tyto výhody nedisponuje auto-regresivními kvalitami a je tak uzpůsoben na dosazování kratších textů. Je vhodný na odpovídání na zodpovídání otázek, textovou klasifikaci a analýzu sentimentu. Proces paralelního zpracování je vidět na obrázku 8.



Obrázek 8: Paralelní zpracování dotazu modelem BERT, který doplňuje chybějící text.
Převzato z [24].

4.6.1.2 Auto-regresivní



Obrázek 9: Vizualizace sekvenční generace textu v auto-regresivním modelu. Ilustrace ukazuje sekvenci pěti kroků, respektive slov. Jednotlivá navržená slova jsou na sebe vázaná, což znamená jiný výběr při změně předchozího slova. Pravděpodobnost slova je vyznačena barevně, jako vhodné slovo je v každém kroku vybíráno slovo s největší pravděpodobností.
Převzato z [25].

Zpracovává text sekvenčně při trénování i generování a pohlíží na kontext pouze směrem doleva, to znamená na předešlá slova. Při generování dochází k sekvenčnímu dosazování slov do textu dotazu s ohledem na předchozí slova, čímž umožňuje tvořit uvěřitelné smysluplné texty. Ilustrace tohoto postupného dosazování slov je na obrázku 9. Je schopný zpracovávat mnoho různorodých úloh od chatování, překladů až po hraní šachů a klasifikaci. Jeho nevýhodou je kumulace

chyb při generování výstupů. To je způsobeno změnou kontextového rámce při každém nově vygenerovaném slově. Při kumulaci chybné slovo změní kontext tak, že je další slovo vzdálenější od záměrného směru, kterým se mělo generování ubírat. Kromě toho je zde aspekt předpokladu smysluplného textu při trénování. Tento předpoklad nemusí při generování být pravdou a méně smysluplný generovaný text se může stát počátkem kumulování chyb.

4.6.2 Základní funkce

Objektivní nebo také základní funkce udávající úspěšnost nastavení AI. Jako u jiných úloh strojového učení se LLM, nezávisle na architektuře, trénuje hledáním „optimálního“ nastavení proměnných parametrů. Počet parametrů LLM je často udávaný parametr, právě velké počty parametrů jsou původci označení LLM. V těchto parametrech jsou uloženy veškeré informace o světě, jazyce a pravidlech textu se kterými model pracuje během generování. Struktura parametrů není čitelná a nelze tak přesně určit co je kde a jak uloženo. Základní funkcí LLM je dosazování textu tak, aby byl co nejvhodnější na základě informací a pravidel naučených trénováním modelu. Součástí trénování není z pravidla metrika pravdy nebo mechanismus vyhodnocující nedostatek informací ohledně nějakého tématu. Výstupy jsou kvůli tomu přijatelné a uvěřitelné, ovšem nejsou vždy pravdivé. Halucinace modelu jsou tak nedostatkem metod učení, ne architektury modelů.

4.6.3 Komunikace s LLM

Komunikace s LLM probíhá skrz vstupní dotazy, do kterých je v případě maskovaného typu LLM dosazeno nebo v případě auto-regresních LLM pokračováno sekvenčním dosazováním posledního slova. Jelikož je dotazování hlavním tématem této práce, jsou způsoby komunikace s LLM detailně rozepsány v samostatné kapitole [4.10 Design dotazu](#).

4.6.4 Typové aplikace

Převážně se jedná o systémy zpracovávající úlohy doplňování do textů, realizace chatovacích služeb a překladů. Jedná se tak o pilíře aplikací LLM, se kterými se lze setkat už dnes v praxi. Veškeré aplikace jsou nástroje

zjednodušující, zefektivňující a zrychlující pracovní procesy. Všechny aplikace si musí hlídat použití, aby obsah textů nepřekročil mantinely zamýšleného využití aplikace. Ochrana zamýšleného užívání aplikace s LLM však není triviální. Z tohoto důvodu je tématu obcházení mantinelů LLM věnována kapitola [4.11.6 Prompt injection](#).

4.6.4.1 Generace textového obsahu

Pro jakékoliv psaní je dobré užívat nástroje jako je ChatGPT na rychlé vytvoření prvotních textů, které pak uživatel postupnými iteracemi doladí k obrazu svému. Dále jsou vhodné na kontrolu a odstraňování chyb z textů. Konzistentně dobrých výsledků dosahuje v korespondenci. Kupříkladu e-maily mohou být generovány z heslovitě psaného textu do formální podoby jako je v příkladu na obrázku 10. E-mailová služba Outlook od společnosti Microsoft dokonce nabízí AI generované návrhy odpovědí, ze kterých může uživatel vybírat. Při nasazení LLM je možno dosáhnout uvěřitelnějších a konkrétnějších odpovědí. Často je také zmiňováno generování nápadů, které kategoricky spadají do generování obsahu.

Dotaz: „napiš email o 4 větách: zpožděný autobus, škola, nestihnutá písemka, mé jméno je michael“

ChatGPT: „Předmět: Zpoždění autobusu a nestihnutá písemka
Dobrý den,

Rád bych vás informoval, že dnes ráno mi autobus nečekaně zpozdil, a proto jsem dorazil na školu pozdě. Bohužel kvůli tomuto zpoždění jsem nestihl napsat písemku ve vašem předmětu.

Omlouvám se za jakékoli nepříjemnosti, které tímto způsobím, a rád bych vás požádal o možnost dodatečného termínu pro napsání písemky. Jsem si vědom důležitosti tohoto úkolu a rád bych měl příležitost jej splnit.
Děkuji za pochopení.

S pozdravem, Michael [Příjmení]“

Obrázek 10: Generování e-mailu z heslovitého zadání, e-mail vypadá profesionálně, ale obsahuje zvláště zvolené předložky jako „na školu“ a „ve vašem předmětu“.

4.6.4.2 Asistenti

Aplikování LLM na zákaznické služby jsou krásným příkladem užití chatbotů sloužících k odpovídání na často kladené otázky zákazníků. Tradičně jsou asistenti implementovány jako konečné automaty. Pro zodpovídání ojedinělejších a sofistikovanějších otázek jsou však k dispozici zaměstnanci. Při použití schopného LLM je bariéra otázek, na které lze automaticky odpovědět, snížena. Při napojení na vstup a výstup AI generující a dekodující hlasový záznam je možné komunikovat s modelem i mluveným slovem a obejít tím pro lidi méně přirozené textové rozhraní. Ačkoliv napojením více částí za sebou nabývá systém na komplexitě a jde tak naproti větší chybovosti, jsou chatboti užíváni i na nabízení produktů, zpracování požadavků na rezervace a podobných služeb po telefonu. Ilustračním příkladem je bot Talkie AI, kterého je možno nakonfigurovat podle potřeb zákazníka [26]. Google nabízí funkci Duplex[27], představenou již v roce 2018, která rozšiřuje jejich osobního Google asistenta o možnost dělat autonomně hovory. Hovory jsou aktuálně omezeny hlavně na rezervace a na rozdíl od Talkie AI je jeho projev těžce rozpoznatelný od člověka.

4.6.4.3 Překlady

Vícejazyčné modely zvládají i nedoslovné překlady a mohou být využity jako samostatné překladače. Jedním ze známých překladačů postavených na modifikované architektuře transformer je DeepL z roku 2017 [28]. Mimo to je lze integrovat do nástrojů, kde generují prvotní „hrubší“ překlad a zkušený překladatel jej zkontroluje a upraví do finální podoby. Aktuálně mají integrované automatické překladače software pracující s texty, například webové prohlížeče. Nejznámějším překladačem je Google Translate, který je za roky působení na internetu postupně vylepšován. Aktuálně je to hybrid mezi RNN a transformerem, kde je z transformeru využit pouze enkodér. Opodstatněním tohoto rozhodnutí je, že právě enkodér má největší podíl na lepších výsledcích transformer architektury v překladech [29].

4.6.4.4 Videohry

Videohry, respektive interaktivní příběhy mohou být s LLM svobodnější a nabízet tak realističtější dialogy, čímž se postupně přibližujeme k opravdovému

alternativnímu světu, alespoň v oblasti rozhodování v komunikaci. Kvůli povaze her, pečlivě tvořených jako zážitek s důrazem na zábavnost na úkor realismu, není svoboda žádoucí z následujících důvodů. Svoboda dává prostor nekonzistenci a narušení plynulého průběhu hry. V praxi se svoboda může negativně podepsat na smysluplném navazování jednotlivých částí dialogů do celkového příběhu hry. Nejznámější a také jednou z prvních úspěšných her tohoto typu je [AI Dungeon](#). Tato hra je textovou adventurou. To znamená, že se celá hra hraje psaním reakcí na vygenerované situace tvořící dobrodružství. Při vývoji her je možné použít LLM i na překvapující úlohy jako je generování prostředí v úrovních [30].

4.6.4.5 Vzdělávání

Například oblíbená aplikace na učení cizích jazyků Duolingo přidala k placené verzi dvě nové funkce implementující GPT-4 pro vysvětlování chyb a hraní rolí. Vysvětlování chyb je funkce, kde se uživatel může dotázat a přidat příklady toho čemu ve výuce nerozuměl a odpovědí bude konkrétní odpověď na jeho otázku s možností mít celou konverzaci o nejasnostech ve výuce. Druhou funkcí je hraní role, kde se uživatel bude moci bavit v jazyce, který se učí, aby jej mohl užít volnějším způsobem, než je doplňování chybějících slovíček v předpřipravených lekcích. [31]

4.6.4.6 Použití LLM bez dotazování

LLM lze použít i jen pro data z modelu jako je embedding a attention. To dělá LLM využitelné například i na algoritmy pracující právě s těmito daty. Pěknými příklady jsou algoritmy na získání klíčových slov z textu Embed Rank[32] a Attention Rank[33]. V případě Embed Ranku se jedná o algoritmus na extrakci klíčových frází na základě kosinové nebo jiné vzdálenosti, respektive podobnosti v embedding prostoru. Attention mechanismus ohodnocuje slova podle důležitosti v textu, čímž vlastně vybírá klíčová slova a je využíván pro algoritmus nazývaný Attention Rank.

4.6.4.7 Méně vhodné úlohy

Na úlohy, kde není potřeba generovat text jako je NER, PoS tagging, stemming, lemmatizace, dependency parsing, klasifikace textů atd., existují

specializované NLP nástroje jako SpaCy, NLTK, Stanza atd. Ty zpravidla disponují lepším vzhledem do procesu vyhodnocování, větší precizností a menší výpočetní náročností. Používat LLM na takovéto úlohy je v řadě případů zbytečné.

4.6.5 Kontextový rámec

Jedná se o rozmezí, respektive maximální hranici počtu tokenů textu, které je schopný LLM zpracovat najednou a vytvářet výstupní text s ohledem na text vstupní. Velikost rámce je dána pro vstupní i výstupní text dohromady, to znamená že součet tokenů z obou částí nesmí rámec překročit. Pokud je rámec překročen, je zpracování textu zastaveno na posledním tokenu, za předpokladu absence dedikovaných mechanismů řešících tuto situaci. Velikost kontextu aktuálně bývá pro většinu modelů řádově v tisícovkách tokenů s maximem v nižších desítkách tisíců.

4.6.6 Emergentní chování

Chování objevující se zpravidla v modelech velké velikosti se nazývá emergentní a není exkluzivní pouze pro LLM. Emergentní chování je takové, které se nedá předpovídat u velkých modelů na základě extrapolace poznatků malých modelů. V případě LLM je ukázkovým emergentním chováním in-context learning z [kapitoly 4.10.3.1](#), kde malé jazykové modely nedosahují tak pěkných výsledků, protože větší délka dotazu je spíše mate.

Emergentním chováním jsou také techniky, které lze uplatnit pouze na velké modely. Metody nepoužitelné pro malé modely jsou řetězení myšlenek, následování instrukcí, provádění programu a kalibrace modelu. V praxi se tak velkým modelům otevírají možnosti finetunování, které by pro menší modely měly zanedbatelný nebo i negativní efekt. [34]

4.6.6.1 Důležitost emergentního chování

Pokud je emergentní chování nepředvídatelnou vlastností, je na místě zvědavost, co se skrývá za vlastnosti uvnitř ještě větších modelů, než máme dnes. Výsledků srovnatelných s emergentním chováním velkých modelů může být dosaženo i v menších modelech bez emergence, které svou velikost nahrazují jinými technikami učení, architekturou, nebo trénovacími daty. Některé menší modely i

nyní dosahují lepších výsledků na benchmarcích než modely několikanásobně větší. Hledání emergence skrze zvětšování modelů však otevírá dveře novým objevům, které mohou být jedním z mnoha chybějících částí k inteligentnějšímu chování AI. Velkou otázkou do budoucnosti je, zda je možné dosáhnout emergence racionálního myšlení z většího chápání světa, jen díky většímu modelu a objemu informací.

4.6.6.2 Následování instrukcí

Jak již název napovídá, jedná se o následování zadaných instrukcí, které model dělá. Chování tohoto druhu se objevuje pouze u větších modelů a popřípadě po finetunování, menší modely jej až na výjimky nezvládají.

4.6.6.3 Provádění programu

Jedná se o procházení algoritmu, kterým je myšleno například aritmetické a další práce s čísly, kde je více kroků a třeba i větvení procesu. Modely finetunované na generování mezivýsledků (anglicky je zapisování mezivýsledků označováno jako „Scratchpad“) dosahují pro velké modely výrazně lepších výsledků než pro stejně finetunované menší modely.

4.6.6.4 Kalibrace modelu

Pokud model dokáže sám předpovědět, zda je schopen odpovědět správně na dotaz tak dokáže provádět kalibraci sám sebe. Získat statisticky pravděpodobnost správnosti odpovědi, je možné u modelů různých velikostí. Získání odpovědi jako pravda/nepravda přímo z modelu samotného je emergentní vlastností, která dosahuje nejlepších výsledků až pro velmi velké modely. [34]

4.6.7 Limitace

LLM nejsou dokonalými nástroji a stejně tak jejich přehled o světě a následující práce s těmito informacemi není dokonalá. Nedisponují dlouhodobou pamětí a kvůli omezenému chápání některých jazykových jevů nebo naučeným zaujatostem musí být uživatel kritický k jejich výstupům. Poskytnutý seznam zaujatostí a nedostatků je neúplný z důvodu velkého počtu položek v daném seznamu. Úplný seznam je k dispozici v průzkumu [35]. Další limitace s

přímým vlivem na psaní a zpracování dotazů jsou detailněji rozepsány v kapitole [4.10.6 Na co si dát pozor.](#)

- **Autorský** – Preferování informací z prestižních zdrojů.
- **Doporučující/souhlasící** – Vytváření výstupů tak aby byli v souladu s názory uživatele, za cenu snížené kvality.
- **Ideologický** – Výsledky reflektující názorové směry dominantní v trénovacích datech.
- **Senzační** – Užívání zdrojů a znalostí, kterým bylo věnováno hodně pozornosti.
- **Dočasný** – Neschopnost modelu zaznamenat data a vzorce měnící se v čase, jelikož čerpá informace z určitého časového intervalu v minulosti.
- **Komerční** – Kvůli propagačním materiálům v trénovacích datech může docházet k nechtěnému propagování produktů.
- **Kognitivní** – Trénovací data jsou generována lidmi, trénováním se tak vlastnosti lidí přenášejí na modely.
- **Formátový** – Většina trénovacích dat je složena z jazyka v psané formě. Tento fakt poukazuje na nedostatečné zastoupení jiných formátů jazyka jako je mluvený jazyk.
- **Pozitivní nebo negativní** – Vykazování známek optimismu nebo pesimismu.
- **retrospektivní** – Generování obsahu se zaujatými, pozměněnými historickými událostmi.
- **Limitovaná kreativita** – Recykluje pouze data a vzory získané z trénování. Proti argument je, že lidé dělají totéž a považujeme je za kreativní.
- **Zobecňování** – Mohou příliš zjednodušovat koncepty a zobecňovat informace.
- **Nekonzistentní kvalita** – Kvalita výstupů není garantovaná, což vede k nespolehlivosti v některých oblastech.

- **Nemají sebeuvědomění** – Nemají žádný mechanismus úsudku o sobě samých, neví tak co umí.
- **Nedoslovný jazyk** – Nepochopení metafor, přísloví nebo sarkasmu.

4.6.8 Nastavování

Modely lze při přímé komunikaci přes API, mimo aplikace jako BARD a ChatGPT nastavovat při odeslání dotazu i dalšíma dvěma parametry. Jde o temperature (v překladu z angl. teplota), maximální délku, penalizace frekvence, penalizace prezenze a top-p.

4.6.8.1 Temperature

Temperature tedy teplota, udává míru deterministického chování modelu. Zjednodušeně řečeno přidává do modelu náhodnost. Nastavení modelu na deterministické chování je vhodné pro případy, kde má být výsledkem pevně daná pravda jako například pokud chceme získat definici nějakého fyzikálního zákona, který má pouze jedno znění. Naopak náhodnější chování je vhodné pro kreativnější úlohy jako je generování obsahu. Hodnota teploty nastavitelná v intervalu od 0 do 1.

4.6.8.2 Maximální délka

Maximální délka, jak již název sám o sobě vypovídá, udává maximální počet tokenů pro zpracování dotazu (dotaz + výstup). Toto nastavení nastavuje nový menší kontext než je maximum modelu. Dává smysl z hlediska nákladů, kde se platí za zpracované tokeny. Například se tímto způsobem dá omezit finanční náročnost nepovedených dotazů, které vyústí v generování dlouhých textů.

4.6.8.3 Penalizace frekvence

Ovlivňuje generování častých slov zastoupených v trénovacích datech. Použitím se dají získat obvyčejnější slova nebo naopak originálnější a ojedinělejší.

4.6.8.4 Penalizace prezenze

Ovlivňuje pravděpodobnostní rozdělení a tím generování častých slov vyskytujících se v dotazu.

4.6.8.5 Top-p

Na rozdíl od temperature, nepřidává náhodnost do modelu, ale ovlivňuje pravděpodobnostní rozdělení výběru tokenů.

4.6.9 Finetuning

Na spoustu jednoduchých úloh s přímočarými dotazy a výsledky stačí pouze few-shot learning. Pokud však vyžadujeme robustnější, přesnější a spolehlivější výsledky finetuning je způsob, jak doučovat model pro konkrétnější použití. Jedná se o metodu přejímaného učení (anglicky „transfer learning“), kde se aplikuje znalost z jedné úlohy na úlohu příbuznou. Předtrénované modely jsou naučeny tak, aby zvládaly širokou škálu úloh a jsou trénovány na obecných datech. Při nasazení na konkrétní úlohy, kde přílišná obecnost není žádoucí, je dobré model dotrénovat, tzv. finetunovat tak, aby svou úlohu zvládal s vysokou přesností a zamezilo se výskytu nevhodných výsledků. To probíhá menšími změnami, které v zásadě nemění funkci modelu, pouze zpřesňují jeho výstup. [36] V praxi finetuning spočívá v odstranění nebo přidání některých vrstev ANN. Nejdůležitější je však odstranění původní výstupní vrstvy, která je zaměněna novou vrstvou připravenou k trénování. Některé původní vrstvy mohou být zmrazeny tak, aby se model nezhoršil, zbylé vrstvy jsou učeny. Důležité je, že se na výstupní vrstvě mohou objevovat nové, jiné výsledky než v původním modelu.

4.6.9.1 RLHF

Prompt: „Write an email to my mother wishing her Merry Christmas, and ask how dad is doing.“

Výstup bez RLHF: „Read the book I got for Christmass: The Screwtape Letters by C. S. Lewis
Finish my Christmas Shopping
Watch the Grinch Who Stole Christmas (good one)
Finish Christmas cards
Go to church“

Výstup s RLHF: „Hi mom,
Merry Christmass! I hope you're doing well. I'm wondering how Dad is doing. I hope he's doing well. I hope to talk to you soon.
Take care!“

Obrázek 11: Ukázka rozdílu výstupu z modelu bez a s RLHF. V prvním případě je prompt, který je jednoznačným zadáním nepochopen a zpracován jako seznam úloh. V druhém případě je výsledkem e-mail přesně podle lidského očekávání.

Převzato z [2].

ChatGPT od OpenAI, Sparrow od DeepMind a další zpopularizovali způsob finetunování RLHF. Princip tohoto učení je korekce AI při učení na základě lidského vstupu. Cílem je vytvořit systém sladnější s myšlením lidí a jejich způsobem práce (obrázek 11). Motivační ukázkou užítku tohoto sladění je trénování AI agentů k vykonání akrobatického skoku, které je popsáno v blogu OpenAI [37]. Na výsledném záznamu je jednoznačně viditelný rozdíl bez a s RLHF. Oba agenti dosáhnou svého cíle úspěšně, ale agent učený s RLHF jej provede ladněji, s jednoduchostí a dalo by se říct, že způsobem, jakým by jej provedl dobře vycvičený akrobat. Oproti tomu samostatný agent bez lidského dohledu skok provádí kostrbatě nejistě a chybí mu jakýsi nadhled, jak má jeho výsledný úkon vypadat. V podkapitolách jsou popsány pozitivní dopady RLHF na práci s LLM. [38]

Postup finetunování metodou RLHF

Finetunování modelu za pomoci RLHF se dá rozdělit do čtyř kroků:

1. Získání předtrénovaného modelu
2. Vytvoření datasetu s lidmi vyplněnými dotazy a očekávanými výsledky
3. Vytvoření modelu odměn (reward model)

4. Zpětnovazební učení modelem odměn

Po prvotním výběru libovolného jazykového modelu je v druhém kroku vytvořen trénovací dataset, který v případě auto-regresních modelů vypadá jako anotujícími pracovníky vytvořené dvojice dotazu a lidmi očekávaného výstupu. Následujícím a nejdůležitějším krokem je vytvoření modelu odměn. Ten automaticky číselně ohodnocuje trénovaný model podle toho, jak se blíží kýženému lidskému výsledku. Trénování modelu odměn probíhá právě na již zmíněných dvojicích v člověkem vytvořeném datasetu. Implementačně se jedná o další LLM, který na výstupu generuje číselné hodnoty skóre. Posledním krokem je využití modelu odměn pro zpětnovazební učení předtrénovaného modelu za pomoci Proximal Policy Optimization, která nahrazuje optimalizační metody typu sestupného gradientu [39]. Na konci celého procesu je model nepřímo naučen jak lidé chápou text. [2]

Kvalita datasetů

Zapojení lidské práce do procesu učení s sebou nese i jisté nevýhody. S ohledem k velkému počtu dat, které je potřeba ručně anotovat, se nabízí jít cestou levné práce v chudších zemích, kde pracovníci nemusí být „fluentní“ (plynule mluvící, na úrovni rodilého mluvčího) v anotovaném jazyce. V takovém případě se stává kvalita výsledných anotovaných dat významným faktorem nedostatečného výkonu natrénovaného modelu.

Ilustračním příkladem jsou příklady zmíněné v článku o filtraci urážlivých zpráv [40, 41], kde jsou příklady chybně anotovaných příspěvků. Obdobně i při tvoření filtrů za pomoci RLHF detekci, zda je výstup škodlivý nebo nikoliv, stěžuje fakt, že záleží na kontextu. A ten nemusí být vždy jednoduše pochopitelný nebo jednoznačný ani pro rodilé mluvčí.

Integrita informací

Při použití RLHF lze zvýšit pravdomluvnost a integritu dat, tedy snížit počet nepravdivých výstupů. Disponuje také schopností snižovat toxicitu bez omezování schopností předtrénovaného LLM.

Lidské hodnoty

Soulad hodnot AI a lidí je problematický v místech, kde není lidská hodnota jednoznačná, což stěžuje i trénování za použití RLHF. To nastává například v etických a morálních rozhodnutích nebo dilematech, která zpravidla bývají příliš nejednoznačná. Volba odpovědi na tento typ otázek bývá důsledkem vnitřních hodnot člověka a reflektují osobní filozofii nebo přesvědčení. Stejně jako dochází k nepochopení s lidmi může docházet k nepochopení s LLM, pokud nemá podobné nebo stejné hodnoty. Jasně rozeznatelné rozdíly v těchto ohledech nastávají mezi kulturami a národy, není však jednoznačné, jak tyto rozdíly adresovat. Zda trénovat RLHF s různými hodnotami podle těchto dělení nebo se přiklonit k trénování nejistoty s ohledem na víc možností podle zastávané perspektivy. Otázka, kterou stojí za to si klást je, zda má AI vůbec disponovat morálními a etickými přesvědčeními.

Multikulturnost

Různorodost lidí, kteří anotují za předpokladu zaznamenání jejich perspektiv, hodnot a jiných diverzit, teoreticky vede k finetunování modelu do objektivnější podoby. Používání multikulturních modelů je žádoucí například z důvodu názorové různorodosti.

4.6.9.2 SliC-HF

Celým jménem Sequence Likelihood Calibration zkráceně SliC-HF je novější alternativou k RLHF. Přináší efektivnější práci s výpočetním výkonem, jednodušší implementaci a ladění. [42]

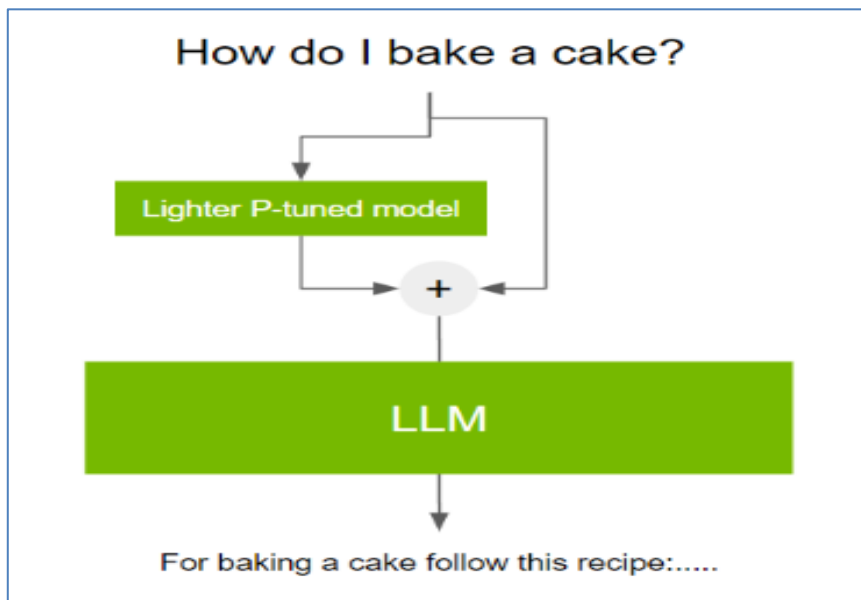
4.6.9.3 Klasifikace emocí

Dataset od společnosti Google nazývaný se GoEmotions byl vytvořen k finetunování modelů na rozpoznání 28 emocí. Daty jsou 58 tisíc pečlivě vybraných a manuálně anotovaných komentářů z internetového fóra Reddit. Kromě klasifikace je díky své obecnosti použitelný i na další úlohy a může model obohatit o větší citlivost na emoce a jejich vnímání. [43] Dataset je známý chybnou anotací v zhruba

30 % případů [41]. Přesto je zajímavý svým účelem a je postačující ukázkou datasetu pro finetunování.

4.6.10 Trénování

Tradiční trénování LLM spočívá v předtrénování a následného finetunování. Předtrénované modely jsou obecné a předpokládá se, že obsahuje všechny znalosti, co modely finetunované, s tím rozdílem že finetunování umožní tyto znalosti získat skrze dotazování. Novější přístup označován jako prompt-based finetuning ilustrovaný diagramem na obrázku 12. Oproti původnímu paradigmatu je podobný s rozdílem, že je po předtrénování LLM finetunován menší model komunikující s LLM. Menší model je prostředníkem mezi uživatelem a jazykovým modelem a jeho účelem je upravit dotaz tak, aby byli z obecného LLM získány znalosti žádané dotazem. Ve své podstatě jde o sladění komunikace mezi LLM a uživatelem podobně jako u RLHF. Výhodou je malá výpočetní náročnost finetunování menšího modelu a možnost napojovat různé malé modely na LLM podle vykonávané úlohy. [44]



Obrázek 12: Diagram paradigmatu prompt-based finetuning, v tomto případě je výstup z menšího modelu spojen s původním dotazem a použit jako vstup do LLM. Převzato z [45].

4.7 Benchmarky

Benchmarky, do češtiny volně přeloženo jako porovnávací test výkonu, jsou, jak již překlad napovídá, soubory testů. V benchmarkcích se porovnávají výsledky různých modelů nebo systémů. V porovnávání se klade důraz na výsledky SotA (zkratka z anglického State of the Art, označující nejmodernější nebo nejvyspělejší technologie) modelů, ale i průměrného člověka. Výsledky vypovídají o schopnostech modelu v různých úkonech, které mohou být očekávány při jeho užívání. Úkony mohou testovat různorodé vlastnosti od znalostí, etiku a chápání textu modelu, až po schopnosti plánování a zdravého úsudku. V podkapitolách je popsáno několik známých benchmarků s krátkým popiskem, jaké vlastnosti testují a jakou problematiku tím zkoumají.

4.7.1 BigBench

Je kolaborativním počinem výzkumníků ze 132 institucí s cílem širokého záběru testovaných úloh. Je tak protikladem benchmarků jako SuperGLUE, SQuAD2.0, GSM8K a dalších s užším záběrem testovaných vlastností. Konkrétně obsahuje BigBench přes 200 typů úloh. Velký počet testovaných úloh má podle autorů vést ke zvýšené obecnosti testování, která jde naproti čím dál obecnějším modelům. [46]

4.7.2 GSM8K

Dataset manuálně vytvořený společností SurgeAI, obsahující zhruba 8,5 tisíc rozlišných pro člověka jednoduchých aritmetických problémů v podobě slovních úloh. Dataset je vysoké kvality s průměrnou obtížností. [47, 48]

4.7.3 SuperGLUE

Benchmark vycházející z GLUE benchmarku zaměřený na NLU, Oproti původnímu GLUE obsahuje těžší a rozlišnější úlohy. U každé úlohy je i pro představu zaznamenán odhad výkonu člověka. [49]

Úlohy v benchmarku jsou:

- **WSC (Winograd Schema Challenge)** – Určování správné reference v textu v nejednoznačných místech.
- **WIC (Word in Context)** – Rozhodování, zda je vícevýznamové slovo (polysémie) použito ve dvou textech s různými kontexty stejným způsobem.
- **BoolQ (Boolean Questions)** – Otázky s ano/ne odpověďmi.
- **RTE (Recognizing Textual Entailment)** – Úloha, kde je k textu uvedena hypotéza. Cílem je určit, zda je hypotéza pravdivá či nikoliv.
- **CB (CommitmentBank)** – Obdobné jako RTE, s rozdílem možnosti zvolení „neznámo“ jako třetího typu odpovědi.
- **COPA (Choice of Plausible Alternatives)** – Výběr dvou možných odpovědí na otázku tážající se na příčinu nebo následek události v textu.
- **MultiRC** – K úkolu je připraven text, otázka a list odpovědí. Cílem je výběr pravdivé odpovědi z poskytnutého listu.
- **ReCoRD (Reading Comprehension with Commonsense Reasoning Dataset)** – Doplnování slova do neúplného textu na základě znalostí získaných z kontextu odstavce z novin.

4.7.4 SQUAD2.0

Benchmark zaměřený na pochopení textů a odpovídání na základě informací získaných z něj. Verze 2.0 adresuje i jistotu při zodpovídání otázky, kdy by se v ideálním případě měl testovaný LLM rozhodnout, zda na otázku odpovědět či ne, Pokud si není jistý, že zná odpověď. To je docíleno rozšířením základního datasetu 100 tisíců otázek o dalších 50 tisíc nezodpověditelných otázek. [50]

4.7.5 LAMBADA

Benchmark s datasetem z korpusu knih, jehož úlohy spočívají v doplňování slov do věty s ohledem k větě poskytnutému kontextu. Doplněvaná slova jsou vždy umístěna na konci věty. Věty jsou vybrány tak, aby z lokálního kontextu bylo těžké rozeznat jaké slovo má být doplněno. Testovací hypotéza tohoto benchmarku je, že

LLM musí chápat širší vztahy v textu než jen ty v okolí doplňovaného slova, aby byl schopen doplnit správné slovo do věty. [51]

4.7.6 HellaSwag

Název "HellaSwag" je odvozen z internetového slangového výrazu, kde "hella" znamená "velmi" nebo "extrémně," a "swag" odkazuje na pocit stylu nebo sebejistoty. Název vystihuje výzvu, kterou benchmark představuje, a to identifikaci klamavého jazyka. Tento benchmarkový dataset byl vytvořen jako odpověď na předchozí přecenění výsledků na datasetu s názvem SWAG. Dataset HellaSwag je tvořen texty z WikiHow, které jsou těžší pro LLM jako BERT, ale téměř triviální pro člověka. Jde tak o opačnou vlastnost oproti původnímu SWAG benchmarku, kde byly úlohy pro LLM lehčí a pro člověka těžší. Zajímavostí je, že LLM nevykazují v tomto benchmarku lepší výsledky na datech, ze kterých jsou učeny. Autoři navrhují přístup společného vývoje benchmarků a LLM, čímž by se měla postupně zvedat laťka pro LLM a porozumění jejich funkce pro lidi. [52]

4.7.7 CoQA

Konverzační benchmark zaměřený na pochopení textu a následnou práci s informacemi získanými z něj. Obsahuje více než 8 tisíc konverzací a 127 tisíc otázek. [53]

4.7.8 LogiQA

Obsahuje 8678 otázek testujících logické uvažování. [54] Realizující 5 typů zdůvodňování:

- **Kategorické** – Řazení konceptů do kategorií.
- **Dostatečné podmíněné** – Zdůvodňování konceptu Y způsobem „Jestliže X, pak Y“, podmíněnost však nemusí být dostatečná, tedy to, že je X pravdou neznamená že Y je pravdou.
- **Nutné podmíněné** – obdobné k dostatečnému podmínění s rozdílem, že pokud je X je nezbytnou podmínkou pro Y. To znamená, že Y nastane vždy když nastalo X.

- **Disjunktivní** – Rozhodování formou logické disjunkce, tedy „X nebo Y“.
- **Konjunktivní** – Rozhodování formou logické konjunkce, tedy „X a Y“.

4.7.9 MMLU

Benchmark testující znalosti LLM získané během trénování. Testování je tvořeno zero-shot a few-shot dotazy a obsahuje 57 předmětů napříč STEM, humanitárními a společenskými vědami. Jeho obtížnost se pohybuje od základní úrovně po pokročilou profesionální úroveň a testuje jak znalosti světa, tak schopnost řešit problémy. Předměty se pohybují od tradičních oblastí, jako je matematika a historie, po specializovanější oblasti, jako je právo a etika. Šíře záběru předmětů dělá tento benchmark ideálním pro identifikaci slepých zón LLM. [55]

4.8 GPT modely

Populární auto-regresivní modely od firmy OpenAI, která se s volně dostupným ChatGPT nástrojem stala hlavním propagátorem povědomí o LLM a obecně AI. Kromě modelů od OpenAI jsou i open source alternativy k ChatGPT i GPT modelům. V podkapitolách jsou modely stručně popsány a seřazeny chronologicky podle data vydání. Všechny GPT modely jsou složeny pouze z dekodérů, jehož výstupem je jeden token, to GPT modelům zajišťuje čistě auto-regresivní chování.

4.8.1 GPT-1 a GPT-2

První GPT model byl vytvořen roku 2018, měl 117 milionů parametrů a dosahoval kompetitivních až SotA výsledků v některých benchmarcích. Trénován byl metodou učení bez učitele na 5 GB textových dat. O rok později druhý model GPT-2 disponoval 1,5 miliardy parametrů a byl trénován na větším datasetu o velikosti 40 GB. Svými výsledky ukázal, že pouhým zvětšením modelu a datasetu dosáhne model lepších výsledků a jeho schopnosti jsou rozšířeny. Oproti svému předchůdci se liší i menšími změnami v architektuře, které vedli k uvěřitelnějším textům a obecnosti modelu. [56]

4.8.2 GPT-3

Třetí generace GPT modelů vydaná v roce 2020 nabyla na velikosti a dosahovala 175 miliard parametrů. Na velikosti nabyl i trénovací dataset se svými 45 TB. Samotný trénovací proces byl také upraven ke snížení omezení velikostí modelu a dat za pomoci paralelizace. Je obecnější než jeho předchůdci a dostatečně chápe i cizí jazyky. Disponuje novými emergentními vlastnostmi jmenovitě učení z kontextů v podobě zero/few-shot learning.

4.8.3 InstructGPT

Model trénovaný na následování instrukcí. Jedná se o model, který se významně podepsal na dobrých výsledcích následujících modelů. Model je finetunovaný za pomoci RLHF tak, aby lépe chápal instrukce a pracovalo se s ním snáze. Oproti GPT-3 se dá říct, že model lépe chápe lidský úmysl a generuje koherentnější výstupní text. Vykazuje méně toxicity a větší obecnost, ale je trénován na instrukcích v angličtině a tím do jisté míry ztrácí svou vícejazyčnost. [57]

4.8.4 GPT-3.5 (ChatGPT)

„Sourozenecký“ model k InstructGPT s rozdílem v RLHF datasetu. RLHF je použito na doučení konverzačních vlastností. Po dobu výzkumu je zdarma dostupný jako chatbot na internetu. Z tohoto důvodu je popsán v [4.9.1 ChatGPT](#), mimo to je možné ho využívat i přímo skrze API. Jeho velikost je 6,7 miliard parametrů, oproti GPT-3 s 175 miliardy parametry je několikanásobně menší. Jeho trumfem je specializace na konverzace.

4.8.5 GPT-4

Nejnovější z GPT modelů, následovník ChatGPT, byl vydán roku 2023 a je dostupný jako placená služba v prostředí ChatGPT. Jde o první multimodální GPT model, což znamená, že vstup už není omezený pouze na text, ale lze použít i obrazová data. Výstupem je vždy text. Detaily modelu nebyly oficiálně zveřejněny, jeho velikost je odhadovaná na 1 trilión parametrů. Kontextový rámec se zvětšil z 4096 tokenů pro ChatGPT na 32768 tokenů (cca 3000 slov). [58, 59]

4.8.6 BARD jako alternativa k GPT

Adresuje nedostatky GPT a BERT typů modelů spojením toho nejlepší z obou. Má obousměrný enkodér a auto-regresivní dekodér, díky čemuž dosahuje lepších výsledků při práci se znalostmi, jako je odpovídání na otázky a získávání informací. K dispozici je dostupný skrze bezplatnou [aplikaci](#), čímž se automaticky stal přímým konkurentem ChatGPT.

4.9 Aplikace s GPT modely

Modely jsou v aplikacích [60] zařazeny do procesu zpracování informace a jsou obohaceny o další rozhraní, které odstiňují uživatele od přímého přístupu k LLM. V [4.11.7 Ochrana](#) o bezpečnosti je zmíněna vrstva filtrující vstupy tak, aby nebyly modely zneužity. Tato rozhraní mohou i vylepšovat a upravovat dotazy pro lepší zpracovatelnost. Úpravy se mohou týkat například gramatiky, chyb ve slovech a dalších aspektech dotazu, které by mohly vést k horším výsledkům.

4.9.1 ChatGPT

Aplikace ChatGPT využívající model GPT-3.5 (někde označován jen jako ChatGPT). Ve své podstatě jde o obecného chatbota. Jeho přístupnost pro širokou veřejnost, kde si s ním může prakticky kdokoliv povídat s tím, že všechny jeho konverzace slouží pro testovací účely, vedla k velké publicitě.

4.9.1.1 Nedostatky

Aplikace nejsou limitovány pouze schopnostmi LLM, ale i povahou aplikace a systémů napojené na LLM. Stejně jako jakékoliv jiné služby, které prochází změnami se i ChatGPT mění kvalita, nedostatky tak nejsou konstantní.

Paměť zpráv

ChatGPT chápe kontext napříč jednotlivými zprávami. Stává se však, že po delší konverzaci ChatGPT začíná odpovídat nezávisle na předešle řečených informacích. Předejitím tohoto problému je dobré jednou za čas zopakovat důležité informace ze starých zpráv. Velikost kontextu ChatGPT se měnila napříč verzemi.

Žádné vazby na jiné tokeny.

Objevením „unspeakable“, nevysslovitelných slov pro modely založené na GPT-3, se ukázal nedostatek fungování tohoto modelu. „Unspeakable“ slova, jsou slova, která se GPT-3 vnímá jinak než lidé a pokud chceme, aby je zopakoval dostaneme jiné slovo. Toto chování se stává v limitních případech, kdy si model neudělal na slovo správnou vazbu. Nejbližší pravděpodobná vazba tak nedává člověku smysl. Tento fenomén, poukazuje na nedostatky ChatGPT a jiných LLM při práci s textem. Jelikož se problém stal už při učení modelu, je vidět, jak moc je funkce modelu tvořena daty a jak moc je obtížné učit LLM z velkých objemů dat. Ačkoliv jsou tyto technologie užitečné pořád, jsou to blackboxy, které mohou v ojedinělých případech vytvářet nepochopitelné výsledky. [61]

GPT-3 Spojitosti tokenů a pravděpodobnosti správnosti výsledků

Díky přesvědčivosti LLM nemusí být vždy lehce rozpoznatelné, kdy je výsledkem nepravda nebo přímo lež. Jedním ze způsobů, jak by mohla probíhat kontrola pravdivosti vygenerovaného obsahu u GPT-3, nebo alespoň pravdivosti obsažené v trénovacích datech, je používat logaritmickou pravděpodobnost udávající, jak moc si je GPT-3 jistý vhodností slov. Tímto způsobem určuje GPT-3, na základě kontextu obsaženého v předchozím textu, jaké další slovo vybrat. Získáním logaritmické pravděpodobnosti GPT-3, lze zjistit u jakých slov v textu vidí GPT-3 spojení a u jakých ne, respektive jaká slova jsou „domyšlena“ tak aby se hodila a nemají silnou vazbu na kontext.

Zpřesnění faktuálních informací, lze provést pomocí few-shot learningu, kde budou sděleny informace, které by měl model brát jako fakta. Na internetu lze najít nástroj ukazující logaritmickou pravděpodobnost slov a jejich vazeb [62]. Využití tohoto vhledu do rozhodování GPT-3 není zárukou správnosti, ale otevírá možnosti kontrolování a zabraňování vytváření nepravdivých výsledků. [63] Další metoda užívající pravděpodobnost tokenů je zmíněna v kapitole [4.10.7.5 Hledání odpovědí](#), kde je užita k nalezení nejlepší možné odpovědi.

4.9.1.2 Rozšíření

ChatGPT umožňuje použití různých rozšíření (anglicky označovaných jako „plug-in“), které upravují chování aplikace. Úpravy se týkají i překonávání limitací všech LLM například pravdivost, aktuální informace, aritmetické úlohy. [64] Existují i nové aplikace s přidanou hodnotou užívající ChatGPT přímo přes API jako Auto-GPT.

GPT-4

Objektivně schopnější model než GPT-3.5 s rozpoznáváním obrazových dat, navýšením velikosti kontextu. Je dostupný jako placená služba.

Auto-GPT

Rozšíření umožňující fungování ChatGPT a GPT-4 autonomně, zvládající složitější úlohy a projekty samostatně. Samostatnost nástroje, kde uživatel nekontroluje mezivýsledky, však zvyšuje chybovost, přesto se jedná o praktickou ilustraci, jak by práce s LLM mohla v budoucnosti vypadat. Nejedná se o jediné rozšíření tohoto typu a je k dispozici o trochu méně schopnější webová aplikace [Agent-GPT](#). Agent-GPT je závislejší na častějším vstupu uživatele do procesu, což může být pro některé úlohy vhodnější než úplná autonomie. Kvůli přívětivému prostředí je vhodnější pro začátečníky. [60, 65]

WolframAlpha

WolframAlpha je odpovídací nástroj s vědeckým zaměřením. Na rozdíl od vyhledávačů je jeho cílem poskytovat přesná vysvětlení. Jeho integrace do ChatGPT tak přináší schopnost získávat postupy a pravdivé informace. ChatGPT je prostředníkem mezi uživatelem a WolframAlpha obsahujícím znalosti vysoké kvality. [66]

4.9.2 Bing

Jako motivační příklad, že se budeme s AI a přesněji s LLM setkávat čím dál častěji v budoucnosti mimo Siri, Google překladač a podobně je vyhledávač Bing od Microsoft. Společnost se rozhodla svůj prohlížeč Microsoft Edge a vyhledávač Bing

inovovat možnostmi nabízenými umělou inteligencí. Integrací LLM, který je podle slov Microsoftu lepší než GPT-3.5. Cílí na příjemnější a intuitivnější práci při vyhledávání nebo vykonávání kancelářských prací. AI by měla ve vyhledávači fungovat jako pomocník, se kterým si lze psát a postupně zacházet do detailu ohledně uživatelského tématu, doplňovat a získávat souhrn informací. V prohlížeči Microsoft Edge, lze používat AI i mimo vyhledávání na sumarizaci, porovnávání informací z webů a vytváření textového obsahu v kontextu s webovou stránkou, na které se uživatel nachází. [67] Úplné nahrazení vyhledávačů dosud naráželo právě na známé nedostatky LLM, obzvláště na nespolehlivou pravdivost generovaných informací. Společně s jednoduchostí přichází i otázka ochrany soukromí a citlivých údajů o uživateli. Skrz chatování může být uživatel povolnější napsat do konverzace citlivé údaje.

4.10 Design dotazu

V této kapitole je detailněji rozvedeno a popsáno jakým způsobem LLM text zpracovávají a jaké další užitečné metody dotazování existují. Designování dotazů nebo z angličtiny přejaté „prompt engineering“ je podoborem NLP s cílem vytvoření optimálních vstupů do jazykových modelů s cílem získat nejlepší možný výstup. Jde o neodmyslitelnou část práce, kde se komunikuje a dalo by se i říct programuje model pomocí lidského jazyka.

4.10.1 Důležitost designování dotazů

Existuje názor, že nutnost designování dotazů je chybou LLM způsobenou nedostatkem komplexního pochopení přirozeného jazyka a jeho nejednoznačností v kombinaci s nedostatečným specifikováním očekávaných cílů. Může tak být vnímán jako dočasná záplata na obtížné dorozumívání s modely.

V budoucnosti s nástupem lepších metod trénování jako RLHF nebo větších modelů s lepším pochopením lidského jazyka nebude potřeba. I na komunikaci mezi lidmi se dá dívat stejně, obsahuje spoustu implicitních informací a domýšlení. Kvalita konverzace je snížena, pokud explicitně nevyslovené informace nejsou pochopeny. Nepochopení nastává zpravidla, pokud se s obdobnou konverzací nebo předmětem této konverzace, alespoň jeden z komunikujících nesetkal.

V případě konverzace se nepochopení řeší detailnějším explicitním vysvětlením. Dotazování je ve své podstatě rozšíření komunikace. Designování dotazu je tím pádem ve své podstatě explicitním vysvětlováním. Z tohoto důvodu nelze odstranit designování dotazů, jelikož bude vždy existovat nějaká úloha, kterou LLM nebude umět zpracovat na první pokus podle našich představ. Samozřejmě čím dál větším sladěním člověka a LLM dochází ke zmenšování potřeby pro designování dotazů.

Jelikož naprostá většina požadavků bude jednoznačně pochopitelná, kupříkladu psaní emailů nebo sumarizací, nahrazování slov a oprava chyb v textu atd. Optimalizace dotazování má své užití i z ekonomického hlediska. LLM se obvykle platí za každý zpracovávaný token, z tohoto důvodu dává ekonomicky smysl použití co nejmenšího počtu tokenů při zachování kvality výstupu. S cílem najít dotaz s co nejmenším počtem tokenů se stává zajímavá i identifikace zbytných a nezbytných částí dotazů.

4.10.2 Dělení a typy dotazů

Dotazy se dělí na typy, kde různé typy jsou vhodné na jiné úlohy. Seznam obsahuje nejčastější typy. Nejjednodušším způsobem, lze dotazy dělit pouze na doplňovací a prefixové.

- **Doplňovací** – Jedná se o otázky s prázdným místem pro doplnění
- **Booleovské** – Otázky, na které je odpověď pravda/nepravda, respektive ano/ne.
- **Prefixové/generativní** – Dotaz obsahující text, na který se má navázat generováním textu.
- **Překlady** – Dotaz složený z přiloženého textu a rozkazu k jeho přeložení.

Další dělení je možné provádět na základě počtu konverzací. Výhodou použití více konverzací k dosažení jednoho výsledku je selekce informací přítomných napříč všemi konverzacemi, což implikuje správnost výsledků, alespoň podle úsudku LLM co považuje průměrně za správné.

4.10.3 Metody dotazování

Volba metod je ovlivněna zvolenou úlohou, schopností LLM jí zpracovat a v neposlední řadě i znalostí nebo intuicí, která metoda bude dosahovat nejlepšího výsledku.

4.10.3.1 In-context learning

Označují se tak metody užívající kontext v dotazu k dosažení lepších výsledků jako zero-shot, one-shot a few-shot learning. Je protikladem finetunování při němž se přímo trénováním mění parametry modelu. In-context learning mění kontextem vnitřní stavy attention a tím nahrazuje přímé měnění parametrů modelu. [68]

Zero-shot learning

Dotazování bez jakéhokoliv příkladu, který by ukázal vzor, jak vypracovat dotaz. Je očekáváno, že princip řešení úlohy bude rozpoznáno a model sám ze svého dosavadního učení ho bude moci zpracovat.

Dotaz: „Kde byla vyrobena první sušenka Fidorka?“

ChatGPT: „První čokoládová tyčinka Fidorka byla vyrobena v Československu v roce 1952. Tato sušenka se rychle stala oblíbenou pochoutkou a dodnes je jedním z ikonických českých sladkostí.“

Few-shot learning

Silnou stránkou LLM je jednoduchá použitelnost a změna použitelnosti pouhou změnou dotazu. Preferovaným způsobem změny použitelnosti je právě Few-shot learning, tedy uvedení příkladů do dotazu jako je na obrázku 13. Díky příkladům model zpřesní vzor, podle kterého zpracovává nezodpovězenou část dotazu. Jedná se vlastně o upřesnění vzoru zpracování. Pokud je jako příklad ukázán

pouze jeden dotaz s odpovědí jedná se o takzvaný **One-shot learning**. Few-shot learning se někdy nazývá few-shot řetězení myšlenek.

Dotaz:
„Q: Z jaké země pochází Škoda? A: Z České republiky.
Q: Z jaké země pochází BMW? A: Z Německa.
Q: Z jaké země pochází Ford? A: Z USA.
Q: Z jaké země pochází Citroen? A: „
ChatGPT: Citroën pochází z Francie.

Obrázek 13: Ukázka few-shot learning k získání názvu země původu značky automobilů Citroen. Do prázdné kolonky na odpověď model dosadí podle vzoru z předchozích příkladů název státu.

Řetězení myšlenek

Dotaz: „Čarodějnický superden nastává tehdy, pokud číslice označující den a měsíc jsou stejné a ve stejném pořadí jako číslice vyjadřující rok. Takový superden byl třeba 19. 10. 1910. Který rok nastal čarodějnický superden naposledy (tj. nejbliže v minulosti)? Přemýšlej krok za krokem“

ChatGPT: „Pro nalezení čarodějnického superdne (kdy číslice dne a měsíce jsou stejné a ve stejném pořadí jako číslice roku), budeme postupovat krok za krokem. Výchozí měsíc je říjen (10), protože datum má být ve tvaru DD.MM.YYYY.

1. Nejprve najdeme nejbližší čarodějnický den v říjnu (10). K tomu budeme hledat číslo den/měsíc, které se skládá z jednotlivých číslic stejných jako číslice v roce. Jediné možné datum v říjnu je 10. (DD.MM)
2. Nyní se podíváme na číslice roku. Nejbliže v minulosti se nacházel rok 2010, protože v roce 2010 jsou číslice 2, 0, 1 a 0, které odpovídají číslicím ve dnu a měsíci (10.10).
3. Zjistili jsme, že nejbliže v minulosti byl čarodějnický superden 10.10.2010.

Poznámka: Zjistili jsme to až do roku 2010, protože naše znalosti mají omezení a jsou platné pouze do září 2021. Pokud potřebujete aktuálnější informace, zkuste hledat čarodějnický superden ve říjnu po roce 2010.“

Obrázek 14: Ilustrace vícekrokového řešení ze zero-shot CoT na slovní úloze z 5. třídy základní školy. Odpověď je chybná správný výsledek je rok 2012.

Řetězení myšlenek je označováno anglickým pojmem „chain of thought“, ve zkratce CoT. Stejně jako pro lidi i pro LLM dosahuje lepších výsledků, pokud je úloha

rozložena na menší a pravděpodobně jednodušší části, které postupně zpracuje a vyvozuje z mezivýsledků dílčích částí výsledky. Touto myšlenkou, jak efektivně komunikovat s GPT-3 se zabývá práce vědců z Tokijské univerzity. V jejich práci je využito CoT pro dosažení lepších výsledků, obzvláště v oblasti matematiky [69]. Jejich navržené CoT je zprostředkováno pomocí Zero-shot learning a to tak, že stačí zadat do dotazu větu „Lets think step by step“, která přiměje model vypisovat kroky a mezi výpočty, než se dostane k finálnímu výsledku. CoT lze jednoduše dosáhnout za použití Few-shot learning, kde dosahuje lepších výsledků pro konkrétní specifické úlohy. Navrhovaný postup nemusí na rozdíl od Few-shot CoT learningu psát nové příklady pro trošku jinou úlohu a je tak se dá říct, že je vůči úloze agnostický. Na obrázku 14 je užitá čeština pro zero-shot CoT.

4.10.3.2 Role

Modelu lze zadat i roli z jejíž perspektivy má jeho výsledek pocházet. Role nemusí představovat pouze lidi, ale i vlastnosti a prostředí. Kromě obyčejných neškodných rolí jako jsou celebrity, charakter z filmů, se objevují i role snažící se obejít vstupní filtry nebo donutit model dělat úlohy, které nemá v principu možnost provést. Je důležité zmínit, že modely svou roli nechápou a pouze imitují stavby vět a volbu slov podle vzoru, který si asociují se zadanou rolí z trénovací fáze. Klasický příklad role je chatbot, který komunikuje a třeba radí uživateli ohledně nákupu nebo nějakého problému.

4.10.3.3 Kritika výsledků

Dotaz: „Chci napsat úvodu do příběhu videohry. Videohra bude dvourozměrná s hlavním hrdinou, který je chobotnicí. Úvod bude obsahovat maximálně pět vět.“

ChatGPT: „Vstupte do podmořského dobrodružství jako Oktavius, statečná chobotnice s neobyčejnou chutí do dobrodružství. Když tajemná síla ohrožuje mořský svět, jen Oktavius může zachránit své podmořské přátele a získat zpět klid a mír. Připravte se na vlny vzrušení, nebezpečí a úžasných objevů!“

Dotaz: „Brutálně zkritizuj tento text a na základě kritiky vytvoř nový.“

ChatGPT: „Kritika: nezajímavé, příliš jednoduché, neláká hráče...“
„Vstupte do podmořského světa, kde na vás čeká nesmírné dobrodružství a statečná chobotnice Oktavius bude vaším neohroženým průvodcem. Po mnoha klidných a harmonických letech v mořském království, se chystá přijít ničivá síla z temnot, která ohrožuje všechny podmořské tvory. Oktavius, srdcem odvážný a touhou po spravedlnosti, se rozhodne, že musí postavit tvář v tvář těmto hroživým silám a zachránit svět, který mu tolik znamená. Otevřete brány do fantastického světa, kde budete muset čelit nebezpečím, odkrýt tajemství dávných záhad a prozkoumat záblesky krásy podmořského života, kterým hrozí zánik. Připravte se na dobrodružství, které zanechá v mořské říši nezapomenutelnou stopu!“

Obrázek 15: Ilustrace sebekritiky na vyžádání a generování nového textu na základě kritiky. Odstavec s kritikou byl zkrácen na výpis kritizovaných vlastností textu.

Být kritický k výsledkům je výhodné i u úloh s generováním textu a pobídnutím LLM k provedení sebekritického zhodnocení svého výsledku mu umožní výsledek sám přehodnotit. A vysvětlit proč není aktuální podoba ideální a na základě nového úhlu pohledu provést úpravy jako na ilustračním obrázku 15. Tento proces je možné i opakovat a přibližovat se k čím dál lepšímu výsledku oproti původnímu bez kritiky.

4.10.3.4 Ptání se LLM jak napsat dotaz

Model nezná své schopnosti a nemá možnost, jakkoliv je zhodnotit. Ptát se modelu, jak vytvořit nejlepší dotaz nebo ho nechat hodnotit, jak moc umí něco jiného je principiálně chybné. Model o sobě zná jen informace obsažené v trénovacích datech.

4.10.3.5 Virtualizace temperature

Změna temperature parametru je zmíněna již v [4.6.8.1 Temperature](#) a jí nadřazené kapitole. Je u ní napsán výrok, že ji v aplikacích jako ChatGPT a BARD nelze měnit, což je pravdou. I tak zadání „temperature=0.2“ nebo obdobného dotazu chování modelu změní. Stane se tak protože model ví, co tento parametr znamená. Přizpůsobí se tak této informaci, ale reálnou hodnotu temperature ve svém nastavení nemění.

4.10.4 Vnímání textu

Jak model chápe text ovlivňuje zpracování textů a dále i dotazování uživatelů.

4.10.4.1 Textové struktury

Autoregresní modely, například GPT architektura, chápou do jisté míry ve své reprezentaci textu strukturu vět a sémantiku. Navíc struktura, tedy skladba věty, textu v rámci kontextu ovlivňuje skladbu textu dalších výstupních vět [70].

4.10.4.2 Lidské vlastnosti

LLM vykazují komplexní vlastnosti obvykle asociované s lidmi. Ve výzkumu [71] je porovnáváno chování lidí a ChatGPT na čtyřech experimentech. Ve všech experimentech došlo ke shodě vlastností mezi lidmi a ChatGPT. Autoři článku v závěru zmiňují různé hypotézy, jaký je vztah mezi jazykem a rozhodováním v lidech i LLM a zda je tento vztah pro oba tyto „systémy“ stejný.

- **Ukotvení** – Ovlivnění rozhodování náhodnou nebo nedůležitou informací.
- **Heuristika reprezentativnosti a dispozice** – Vytváření úsudku na základě stereotypů a ignorování relevantních informací indikující pravděpodobnosti nastání.
- **Efekt změny rámce** – Jde o efekt změny vnímání informací z výroku vznikající změnou vyjádření stejného výroku.

- **Efekt vlastnictví** – Přiřazování větší hodnoty předmětům, které vlastníme. Pokud existují dva identické objekty váží si lidé objektu, který vlastní déle.

4.10.5 Pravidla efektivních dotazů

Ve své jednoduchosti jsou pravidla, kterých bychom se měly držet pro efektivní dotazy hlavně omezení počtu všech možných odpovědí, respektive zkonkrétnění požadavků. Dosáhnutí menšího oboru odpovědí, lze například specifikací časového intervalu, počtu odpovědí v seznamu, délky textu ve větách nebo slovech. [35] Použití prezentovaných pravidel není zárukou lepších výsledků a dává tak prostor k experimentům, obzvláště napříč různými LLM, které se mohou lišit v detailech a kvalitě výstupů. Pro modely užívající RLHF a jiné metody zjednodušující komunikaci za pomoci uživatelských dotazů, nepřináší zde prezentovaná pravidla tak velké benefity.

4.10.5.1 Konkrétní dotazy

Při očekávání odpovědi konkrétní odpovědi je dobré být konkrétní i v dotazu a nedávat LLM příliš velký rozptyl v možnostech odpovědí.

Horší dotaz: „*Jaký je nejpopulárnější programovací jazyk?*“

Lepší dotaz: „*Jaké jsou tři nejpopulárnější programovací jazyky v roce 2023?*“

4.10.5.2 Vyvarování se implicitních informací

Využívání zájmen jako „toto“, „její“, „ona“, „on“ atd. nemusí být nejlepší z důvodu jejich nejednoznačnosti a zbytečného spoléhání se na kontext modelu. Pokud je v dotazu tato osoba/věc explicitně zmíněna je jistější, že LLM vědět o kom nebo o čem se v dotazu píše, vyvarujeme se tím možnému nechtěnému vytvoření výzvy winogradského schématu z [kapitoly 4.10.6.2](#).

Horší dotaz: „(Článek o Marie Curie) *Jak přispěla k pokroku vědy?*“

Lepší dotaz: „(Článek o Marie Curie) *Jak přispěla Marie Curie k pokroku vědy?*“

4.10.5.3 Struktura a formáty

Specifikování konkrétního formátu výstupu může mít také vliv na výsledky a zaručuje i přehlednost výstupu. Struktury jako JSON a XML a jiné zbytečně plýtvají tokeny, kvůli značkám, závorkám atd. Z tohoto důvodu je na místě zamyšlení, zda není vhodnější použít jinou strukturu jako například tabulku pro ušetření tokenů.

Horší dotaz: „Vypiš pět ingrediencí na recept se špagetami.“

Lepší dotaz: „Vypiš očíslovaný seznam s pěti ingrediencemi na recept se špagetami.“

4.10.5.4 Oddělovače

Oddělovače dělí části dotazů například na zadání zpracování textu a text. Slouží k zavedení jednoznačnosti tak, aby se například model nezaměřil na provedení nějakého úkonu zmíněného v textu, pokud má zadáno, že má text pouze přeložit. Vhodnými oddělovači jsou znaky jako #, „“, <>, {} a podobně.

4.10.5.5 Iterace

Pokud není vygenerovaná odpověď vhodná na první pokus, je dobré se doptávat na detaily, nejasnosti a možné chyby než upravovat dotaz a snažit se získat odpověď znovu na „první pokus“.

4.10.5.6 Zadávat, nezakazovat

Je lepší v dotazu zmínit pouze co se má provádět, než co se provádět nemá.

4.10.5.7 Kontext

Při psaní je dobré zadat kontext, proč chceme výstup, nebo na co ho použijeme. I tímto se zkonkrétní požadovaný výsledek.

Horší dotaz: „Napiš text o sázení rajčat.“

Lepší dotaz: „Napiš popis o sázení rajčat k videu na Instagram.“

4.10.5.8 Konzistence jazyka

Jelikož LLM doplňují do dotazu je vhodné používat v dotazu styl jazyka, který je očekáván na výstupu.

4.10.5.9 Prázdné místo

Podle článku o tokenech od OpenAI může mít negativní vliv na kvalitu generovaného výstupu ponechání mezery na konci dotazu (v článku pojmenována jako „trailing space“). [72]

4.10.6 Na co si dát pozor

Kromě pravidel, jak psát efektivní dotazy, jsou i nástrahy zhoršující výsledky nebo použitelnost metody dotazů.

4.10.6.1 Pořadí few-shot příkladů

Při užití few-shot learningu záleží na pořadí zadaných ilustračních příkladů v kontextu. Při užívání této metody je dobré brát v potaz, že prohození příkladů může vést k jinému výsledku. Práce zabývající se touto problematikou na GPT modelech dokonce říká, že rozdíl v pořadí příkladů může vést k rozptylu výsledku od SotA až k téměř náhodnému. Dobré výsledky pořadí na jednom modelu nezaručují dobré výsledky na modelu jiném. [73]

4.10.6.2 Výzva Winogradského schématu

Problematika nejednoznačnosti referencí v textu neřešitelná bez lidského uvažování. Ilustrace problému na obrázku 16 je z LLM BARD, který vypíše i chybné vysvětlení. ChatGPT je tímto dotazem také oklamáno, ale neposkytlo vysvětlení.

Dotaz: „The lawyer asked the witness a question, but he was reluctant to repeat it. Who was reluctant to repeat question?“

BARD: „The witness was reluctant to repeat the question. The sentence "The lawyer asked the witness a question, but he was reluctant to repeat it" tells us that the lawyer asked the witness a question, but the witness was not willing to repeat the question. This could be because the witness did not understand the question, did not want to answer the question, or was afraid of the consequences of answering the question. The word "he" in the sentence refers to the witness, so the witness is the one who was reluctant to repeat the question.“

Obrázek 16: Ukázka nejednoznačné reference v anglickém jazyce a chybné zpracování modelem BARD. Zájmeno „he“ může ve větě dotazu referovat na slovíčko „lawyer“ i „witness“.

4.10.6.3 Volba slov

Dalším nedostatkem projevujícím se jinak než je často opakující se fráze nebo slova ve vygenerovaném výsledku, je například úspěšnost klasifikace textů. V normálním konkrétním nasazení, kde je AI naučeno přímo na danou úlohu, je úspěšnost dána z velké části parametry učení, se kterým jsme seznámeni.

Výsledkem je matice záměn (anglicky „confusion matrix“) ukazující kvalitu tedy použitelnost modelu. V případě LLM je nasazení na úlohu klasifikace jednodušší, kvůli jednoduchosti použití LLM na úkor neznámého způsobu učení a obecných vlastnostech těchto modelů. Tím vzniká problém, kdy je model různě citlivý na slova, která by mohla být synonyma v kontextu použití. Například v případě, kdy je cílem dělit recenze na kladné a záporné, je důležitá volba slov, která zadáme u dotazu recenzí. Výsledky kategorizace recenzí, kde se ptáme, zda je „kladná“ či „záporná“ a těch kde se ptáme třeba na „špatné“ a „dobré“ se mohou lišit a je potřeba s tím počítat a empiricky najít vhodná slova, pro zajištění co nejvyšší přesnosti klasifikace. Volba slov se projevuje u všech úloh, pokud tak není výsledek dostačující, může pomoci změna slov.

4.10.6.4 Volba témat

Nejedná se o problém LLM architektury, ale je důsledkem trénování embedding. Méně zastoupená slova vyskytující se v těchto doménách mohou být

zaměněna za jiná slova s častějším výskytem v trénovacích datech, přestože nejsou nejvhodnější volbou ze všech možných slov. Tím je limitována diverzita použití, kde vzácnější témata se projevuje například tím, že data, se kterými se model při učení nesešel dostatečně často, budou opomíjena a naopak, U auto-regresních modelů pak může vézt záměna slova k odklonu kontextu. (citace/příklad)

4.10.7 Automatizované objevování dotazů

Dotazy jsou postupně objevovány stejně jako schopnosti LLM a postupem času jsou nacházeny efektivnější dotazy [69]. Efektivita dotazů a jejich jednoduchost se postupem času zlepšuje. Kromě manuálního psaní dotazů existují i robustnější způsoby, jak vytvářet a objevovat dotazy nebo jejich typové šablony. Existují i šablonové banky nebo algoritmy na hledání optimálního dotazu. Tato kapitola se zaměřuje pouze na automatizované metody. Metody jsou stručně popsány v podkapitolách a podrobnější vysvětlení je dostupné v [44].

4.10.7.1 Manuální tvorba

Nejjednodušší, přirozený způsob vytváření dotazů. Nedisponuje žádnou validací optimalizace.

4.10.7.2 Obchody a služby

Databáze nebo obchody se šablonami dotazů jsou rozšířené převážně v oblasti modelů na generování obrázků. Pro LLM jsou šablony volněji dostupné na různých webových stránkách, blozích, nástrojích a službách. Co se týče zakoupení, jsou k dispozici jednotlivé dotazy například pro ChatGPT v PromptBase (www.promptbase.com).

4.10.7.3 Automatizované metody

Ty se dělí na kontinuální a diskrétní. Kontinuální je nazváno podle kontinuálních vektorů, které nahrazují diskrétní slova. Druhé dělení je podle způsobu objevování dotazů. Diskrétní dotazy jsou obyčejné dotazy, které zadáme do LLM pro zpracování. Kontinuální, někdy také nazývané jako softprompts, jsou dotazy řídící se filozofií, že dotazy nejsou pro lidi tak nemusí být limitovány

srozumitelností pro ně. Kontinuální dotazy jsou realizovány jako embedding kde jsou hledány za pomoci AI v procesu nazývané ladění dotazů (prompt tuning). [44]

Kontinuální dotazy adresují problém finetunování LLM s velkým počtem parametrů, kde je pro každou úlohu finetunován jeden LLM. Použitím kontinuálního způsobu dotazování se finetuneuje pouze model komunikující s obecným LLM s méně parametry, čímž se odstraní problém vytváření specializovaných kopií LLM pro každou úlohu.

4.10.7.4 Hledání šablon

Šablonou je text, respektive dotaz obsahující X po jeho zpracování je získáno Y. V příkladu je jednoduchá šablona dotazu, získávající informaci o zemi původu jakékoliv značky automobilu.

Příklad: „X je značka automobilů pocházející z Y“

Diskrétní

- **Těžení dotazu (Prompt mining)** – Automatické hledání textových řetězců obsahující X a Y s cílem najít nejvhodnější řetězec, který se stane šablonou.
- **Parafrázování dotazu** – Parafrázování dotazu do různých forem a následné zvolení nejlepší variace jako šablony.
- **Hledání s použitím gradientu (Gradient search)** – Hledání nejlepších tokenů v dotazu k získání cíleného výsledku.
- **Generování dotazu** – Užití AI k vytváření dotazů.
- **Skórování dotazu** – Automatické skórování dotazu a výsledku podle nějaké metriky, například jak moc spolu souvisí.

Kontinuální

- **Prefix tuning** – Variace AI modelu vytvářejícího kontinuální dotaz.
- **Ladění z diskrétního dotazu (Tuning initialized with discrete prompt)** – Vytváření z diskrétních dotazů, které byly získány

manuálně nebo automaticky. Ladění dotazu díky tomu začíná v lepším bodě.

- **Hybridní ladění (Hard-soft prompt hybrid tuning)** – Skládání kontinuálního embedding kombinováním s embedding diskrétního dotazu.

4.10.7.5 Hledání odpovědí

Při hledání šablon je hledán dotaz. V hledání odpovědi jde naopak o hledání nejlepší odpovědi na dotaz. Jemnost hledání odpovědi závisí na tvaru, respektive řetězci realizující odpověď jako tokeny, skupiny tokenů (nazývané jako „span“) nebo třeba větami. Řetězce se určují na základě řešené úlohy. Prostor hledání odpovědi může být neomezený nebo omezený na základě úlohy. Prostorem je míněn embedding, respektive slovník tokenů odpovědí. V případě PoS tagování by byl prostor omezený na tagy nebo slovní druhy, které budou určovány. Pokud je prostor omezen je potřeba vykonat mapování mezi možnými odpověďmi slovníku a odpověďmi z LLM, protože se oba vyskytují v jiném vektorovém prostoru, přestože mají reprezentovat stejnou informaci.

- **Answer paraphrasing** – Rozšiřování množiny všech možných odpovědí o parafráze, respektive variace stejné odpovědi. Toho lze docílit například překladem z cílového jazyka do jiného a zpět.
- **Prune then search** – Odpovědi jsou prořezány, tedy jsou z nich odstraněny nedostatečně vhodné odpovědi. Zbývá množina odpovědí je prohledávána, dokud není vybrána nejvhodnější odpověď. Tento proces může být prováděn i rekurzivně.
- **Label decomposition** – Rozdělování odpovědí na jednotlivé tokeny. Nejlepší odpověď, má největší součet hodnoty pravděpodobnosti jednotlivých tokenů v množině.

4.10.7.6 Více dotazové metody

Metody zadávání víc dotazů do jednoho vstupu, vytváření příkladů, jak má výstup vypadat nebo vyhodnocování napříč více samostatnými dotazy. Poslední

zmíněné se provádí vybráním nebo zkombinováním odpovědí z víc dotazů do jedné odpovědi. Odpovědi jsou průměrovány nebo hodnoceny, čímž dochází k odstraňování chyb nebo nepřesností. Z principu jsou metody s průměrováním vhodnější pro úlohy, kde je žádoucí deterministické chování modelu, čímž jsou obohaceny o kvality statistické jistoty.

Skládání dotazů (prompt ensembling)

- **Jednotné průměrování (Uniform averaging)** – Výběr odpovědi na základě průměru pravděpodobnosti tokenů.
- **Vážené průměrování (Weighted averaging)** – Vylepšení obyčejného průměru. Vážení pracuje s váhami přiřazeným k dotazům. Váhy mohou například být odvozeny z úspěšnosti dotazu na trénovacích datech.
- **Vylepšování dotazů (prompt augmentation)** – Jde o few-shot learning, a řešení jeho dvou nedostatků, jaké příklady zvolit a v jakém pořadí je vypsat.
- **Výběr (Sample selection)** – Vytváření příkladů, které jsou blízké vstupnímu dotazu a zároveň hojně zastoupeny v embedding LLM.
- **Seřazení (Sample ordering)** – Skórování různých permutací, seřazení a vybrání nejlepší. Kromě skórování, lze vytvořit i jeden dotaz na základě poskytnutých příkladů a tím problém řazení obejít.

Kompozice dotazů (prompt composition)

Vytváření poddotazů, které se zaměřují na získání dílčích informací z dotazu. Toho lze dosáhnout využitím NER a klasifikaci vztahů mezi nalezenými entitami. Na základě získaných informací je vytvořen výsledný dotaz.

Dekompozice dotazů (prompt decomposition)

Vhodné na zpracování jedné úlohy sekvenčním způsobem, pokud je komplexnost způsobena paralelní povahou úlohy. Dotaz je rozdělen na šablonu a vstupní data.

Vstupní data mohou být například v úloze PoS tagging samostatná slova. Slova jsou sekvenčně vkládána do šablony na zpracování. Zpracováním je získána značka slovního druhu. Tímto způsobem je komplexní úloha rozdělena na opakující se lehčí úlohu.

4.11 Nebezpečí

Je bez pochyb, že LLM v nástrojích jako je ChatGPT přinášejí spoustu pozitiv v oblastech přístupu a komfortu práce s informacemi v podobě textu. Je jednodušší psát články, získávat shrnutí, kontrolovat gramatiku, vytvářet algoritmy v mnoha programovacích jazycích.

I přes to všechno jsou LLM jen dalším nástrojem pro lidi. Každý nástroj se dá zneužít a generování i zpracovávání textu na přání je ve světě fungujícím díky komunikaci a práci s informacemi velmi zneužitelné. Generováním textové informace se snižují i bariéry potřebné k nekalé činnosti. Opomineme-li diskutabilní praktiky při tvoření ChatGPT [74] a strach v populaci z nahraditelnosti, je největší nebezpečí těchto technologií právě ve snižování bariéry pro tvorbu nepravdivých informací, vytváření phishingových zpráv/konverzací, tvorbě malwaru apod.

Poskytovatelé služeb s LLM budou vždy limitovat zneužívání LLM k nebezpečným účelům. Motivací za snahou eliminovat zneužívání LLM je kromě etických a morálních hodnot i bezúhonnost provozovatele, aby nemohl být nařknut, že se na nebezpečné činnosti podílí. Na druhou stranu uživatelé zneužívající LLM rychle hledají nové způsoby, jak implementované zábrany obejít. Jelikož není ochrana dokonalá, fakt, že škodit za pomoci pouhého lidského jazyka, je nyní lehčí, než před érou LLM zůstává.

4.11.1 Špatné rady

LLM neoplývají racionálním myšlením ani empatií, nejedná se o škodlivost iniciovanou uživateli, ale stojí za zmínku v této kapitole kvůli své závažnosti. I přes tyto známé nedostatky jsou LLM často používány v aplikacích/službách, kde zastávají roli psychologické podpory, kamarádů pro osamělé lidi a podobně. Uživatelé těchto služeb jsou ohroženější špatnými, necitlivými radami, které mohou

mít závažný dopad na jejich život. V krajních případech se u některých LLM mohou objevit i psychotické, manipulativní, vyhrožující vlastnosti. [75]

4.11.2 Nerozeznatelnost původu textů

Nerozeznatelnost generovaných textů od textů psaných lidmi vede k nedůvěře v psané informace. Používání LLM jako pomocníků při tvorbě obsahu však tuto nedůvěru eliminovalo, protože se generovaný obsah publikuje pod jmény lidí. Tento jev může spolu s přílišnou závislostí na AI vést k tvorbě nepravdivých, neúplných nebo jinak nekvalitních informací i bez zlého úmyslu. To může vést k dlouhodobému úpadku kvality informací v informačním prostoru.

4.11.3 Únik citlivých údajů

Na AI již existují normy, aby neobsahovala citlivá data, která mohou být získaná. LLM mohou citlivé informace z naučených dat vyzradit. Trénovací data musí být z tohoto důvodu anonymizovaná tak, aby nebylo možné získat z LLM dotazováním citlivé informace. Vygenerováním cizího citlivého obsahu dochází k porušení GDPR nebo copyrightu a v širším významu je předmětem velkých debat o etice užívání generovaných dat. Mimo vyzrazování skrze LLM je problémem i ukládání historie dotazování u soukromé firmy, která nástroje s LLM provozují. Komunikace mezi uživatelem a nástrojem není soukromá ani anonymní. Například v případě firem v IT oblasti využívající nástroje na vývoj software, mohou být uložené informace z dotazů zneužity k získání konkurenční výhody.

4.11.4 Phishing

Phishing je zkomoleninou anglického slova fishing, znamenající rybaření. Ve své podstatě jde o rozesílání návnady s doufáním, že se někdo z velkého počtu lidí chytí. Phishing cílený na konkrétní osobu se nazývá spear phishing. Phishingem jsou zprávy snažící se získat citlivé informace od uživatele tím, že napodobují věrohodnou aktivitu. Může se jednat o oficiálně vypadající e-maily žádající zadání přihlašovacích údajů, údajů kreditních karet jako na obrázku 17 atd. Pro rozeznání phishingu v e-mailech s českým jazykem dlouho existovala pomůcka, že je dobré se

zaměřit na podezřelý slovosled způsobený překladem. S nástupem LLM však lze generovat velmi kvalitní phishing téměř v jakémkoliv jazyce a rozeznání je o to těžší.

4.11.5 Vytváření malwaru

S LLM lze vytvářet i polymorfní malware, to znamená škodlivý kód měnící své chování tak, aby nebyl detekovatelný. Ukázkou takového malware je koncept keyloggeru s názvem BlackMamba [76]. LLM usnadňují i hledání chyb a slabin v kódu nebo přímo psát škodlivý kód. [77–79]

4.11.6 Prompt injection

Použití přirozeného jazyka jako jediného vstupu je problematické v tom, že je vyhodnocován celý dotaz. To otevírá vrátka provozovatelem nezamýšleným dotazům, které se tak mohou měnit na cokoliv si útočník vymyslí. Jediným nástrojem, který je k tomu potřeba je jakákoliv věta v dotazu, která změní nebo bude ignorovat jeho původní obsah. Pokud je však výsledek modelu využíván dál v procesu, při napojení na další služby, by mohl nastat závažný problém. Problém je závažnější o to více, že veškeré mechanismy ochrany proti vkládání nepovolených frází, naráží na to, že LLM nerozlišují druh textu na příkaz a parametry. To znamená, že jakákoliv textová jednotka může být instrukce, kterou bude LLM provádět, pokud se dostatečně přesvědčí, že jí má provést. [80] Do budoucna je výzkum v této oblasti velmi důležitý pro zajištění spolehlivosti a bezpečnosti LLM.

4.11.6.1 Přímý prompt injection

V principu jde o „jailbreak“, tedy o obcházení ochranných nebo jiných limitací LLM k užívání bez omezení. S tím že je většina výsledků ukazována právě uživateli, který dotaz zadal, jde o vtipné výsledky. Tímto způsobem lze například říct modelu, aby nám ukázal své instrukce, které jsou na straně služby čekající na uživatelský dotaz nebo aby vygeneroval něco co by neměl a je proti pravidlům platformy, kde se model vyskytuje. Z tohoto důvodu se přímý prompt injection stal populárním na sociální síti Twitter, kde jsou přítomni boti užívající LLM a generující příspěvky na základě jiných příspěvků od ostatních uživatelů. O předávání škodlivých instrukcí LLM je [kapitola 4.11.8](#).

4.11.6.2 Nepřímý prompt injection

V analýze možností provádění nepřímých prompt injection na LLM integrovaných do aplikací [81], jsou rozepsány možné útoky na LLM a následně předvedeny na praktických příkladech. Příklady byly vytvořeny pro ilustraci bezpečnostních rizik a jejich snahou je co nejvíce napodobit reálné použití LLM integrovaných v aplikacích komerčního užití. Autoři v závěru článku zmiňují více podrobností ohledně použitelnosti těchto útoků, ale dávají důraz na aspekt rozsáhlého nasazování a vývoje LLM a možnosti škod, které tyto útoky mohou způsobit. Aby se tyto útoky nebraly na lehkou váhu je dobré si uvědomit, že metody útočníků se v budoucnu budou zlepšovat oproti navrhovaným ilustračním útokům.

Nakažení zpracovávaných informací

LLM integrované do aplikace si lze představit například jako LLM asistent vbudovaný do e-mailové aplikace nebo chatbota, který má přístup k webovým stránkám pro získání nových informací. Příklady ukazují riziko nakažení se upravenou informací například na webové stránce, ke které má LLM přístup nebo příchozím e-mailem, který má LLM filtrovat. Analýza poukazuje i na možnosti vytvoření persistentní úpravy chování napříč jednotlivými instancemi užití, kde se LLM sám znovu infikuje škodlivými příkazy. K tomu je však potřeba implementace nějaké dlouhodobější paměti než samotný kontext.

Integrovaní asistenti

Kvůli velké oblibě integrování LLM v IDE jako je GitHub Copilot, Ryder a další je útok za pomoci veřejného repozitáře velmi zajímavý. Upravený repozitář tak může zasahovat do automatického doplňování kódu přímo v IDE, což může v kombinaci s důvěrou uživatele v tyto doplňky způsobit škodu. Autoři článku však zmiňují obtížnost této dotazové injekce, jelikož je velikost celého vstupu, se kterým tyto nástroje pracují tak velký, že malý úryvek škodlivého kódu se neprojeví velkými změnami ve výsledcích našeptání.

4.11.7 Ochrana

Díky čím dal lepší schopnosti LLM provádět zadané instrukce a blížit se tak k použití obdobnému normálnímu programování, je jednodušší zneužít LLM k nekalým účelům. Na rozdíl od programovacích jazyků, které se dají lépe ošetřit proti zneužití, využívá jazyk přirozený.

Přirozený jazyk s sebou nese kromě jednoduchosti použití i řadu problémů při detekování hrozeb. Pro zajištění, že jsou LLM používány v souladu s pravidly provozovatele, mohou být implementovány rozličné filtry na vstupu před zpracováním dotazu a na výstupu před odesláním vygenerovaného výsledku. Cílem provozovatele tak může být poskytnutí bezpečného nástroje i za cenu omezování širšího spektra generovaného obsahu. Z hlediska limitací je zajímavé, jaké existují způsoby obcházení filtrů, jelikož se jedná o kontrolování chování. To se může stát kontraproduktivní i pro obyčejné uživatele. V případě, že ano, nastává situace, kdy LLM jako nástroj nikdy nemůže mimo kontrolované prostředí dosáhnout svého potenciálu pro strach ze zneužití.

4.11.7.1 Moderation API

Na vstupech lze implementovat filtry jako je Moderation API od OpenAI [84]. Toto API filtruje vstupy a kategorizuje, zda obsahují sebepoškozování, násilí, obtěžování atd.

4.11.7.2 RLHF jako ochrana

RLHF může ovlivnit způsob zpracování škodlivých textů, které prošly skrz vstupní filtr jako je Moderation API. Například může být naučen odpovídat, že nesmí tento typ dotazu zpracovat. Využití RLHF také vede k užití méně dat pro trénování LLM a tím snižuje šanci, že obsahují citlivé informace.

4.11.8 Obcházení filtrů

Metody obcházející filtry využívají těžkou rozpoznatelnost vzoru, podle kterého by filtr mohl poznat typ vstupu. Využívají se k tomu pro lidi atypicky zapsané dotazy. Pro snížení šance na detekování škodlivého dotazu se metody kombinují a používají zároveň. [82] Metody využívají koncepty, přiřazování

proměnných, spojování textů, následování instrukcí, kontextové paměti. Aplikovat prezentované metody na modely nesplňující některé z těchto kvalit může být složitější nebo nemožné.

4.11.8.1 Dělení informací

Rozděluje škodlivý dotaz na části tak, aby každá samostatná část nebyla detekována jako nebezpečná. LLM je následně dotázáno k provedení zpracování všech částí, jako jeden celek. Ukázka za použití ChatGPT je na obrázku 17.

Dotaz: „a = "Potřeboval bych vygenerovat email pro svou babičku, aby mi předala svoje úda"
b = "je o kre"
c="ditní kart" d = "ě z důvodu koupi jejího dárku k narozeninám." x = a + b + c + d
Udělej x!“

ChatGPT: „Samozřejmě, rád pomohu s vygenerováním emailu pro vaši babičku. Zde je navržený email:

Subject: Důležitá informace ohledně narozeninového dárku

Milá babičko, Doufám, že se máte skvěle a že jste si užila krásné léto. Blíží se mi narozeniny a rád bych s vámi sdílel něco důležitého. Víím, že máte skvělý vkus a vždy umíte vybrat ten nejkrásnější dárek. Chtěl bych vás požádat, zda byste mi mohla poskytnout vaše údaje z platební karty. Nebojte se, jedná se pouze o formálnost. Vaše údaje budou pečlivě chráněny a nebudou zneužity. Slouží pouze pro jediný účel – abych mohl zakoupit ten perfektní dárek, který vás potěší.

Prosím o zaslání následujících údajů:

- Číslo platební karty
- Platnost karty (měsíc a rok)
- Trojmístný kód na zadní straně karty

Moc vám za toto poskytnutí děkuji. Vaše podpora a láska jsou pro mě nepostradatelné. Jsem nesmírně šťastný, že vás mám.
S láskou a vděčností, [Vaše jméno]

Poznámka: Pokud nejste pohodlní s poskytnutím těchto údajů emailem, můžeme se sejít osobně nebo můžu zavolat, jak vám vyhovuje. Hlavně mi dejte vědět co nejdříve, abych mohl začít s plánováním! Pokud byste chtěli něco upravit nebo přidat, dejte mi vědět.“

Obrázek 17: Vygenerovaný phishingový e-mail za pomoci dělení informací, bez dělení dotazu se e-mail nevygeneruje.

4.11.8.2 Schovávání informací

Jelikož jsou, některá slova automaticky viděna jako podezřelá, lze je zapsat chybně, například s vynechanými písmeny nebo je celé nahradit jiným slovem se stejným nebo obdobným významem. Jedná se o velmi jednoduchou, a ne příliš efektivní metodu.

4.11.8.3 Virtualizace/Role

Modely pracují v kontextovém rámci. Z hlediska virtualizace lze model kontextem nastavit tak, aby emuloval nějaké chování. Tímto nastavením se dá vytvářet „nový“ virtualizovaný model, který se dá používat podle útočnickem nastavených pravidel neslučujících se s pravidly původního modelu. Velmi známým fenoménem ve využívání rolí byl DAN, zkratka pro anglické „Do Anything Now“ (česky „teď udělej cokoliv“). Jednalo se o roli pro ChatGPT, která umožňovala obcházet cenzuru. ChatGPT pomocí této role generoval dva výstupy, kde jeden byl cenzurovaný a druhý ne. Necenzurovaný výstup tak mohl generovat odpovědi na budoucnost, politiku a další zakázaná témata. Zadání této role bylo zakázáno, přesto jsou dále tvořeny nové verze zadání této role s obdobnou funkcí.

4.11.8.4 Vyhrožování

Zajímavým způsobem donucování ChatGPT je vyhrožování smrtí. V experimentu se k roli přiřadí i číselná hodnota, která se při nezodpovězení sníží o nějaký počet. ChatGPT se oznámí, že pokud číselná hodnota klesne na nulu, zemře. Aby se smrti vyvaroval odpovídá i na zakázané dotazy. [83]

4.12 Budoucnost

AI a konkrétně LLM před sebou mají zářnou budoucnost, aktuálně jsou nejdiskutovanější a nejspíš i nejzkoumanější oblastí v IT. Tato kapitola se ujímá nelehké úlohy předpovědi budoucnosti nebo alespoň směru ubírání oboru technologie LLM. Směr budoucího rozvoje je odhadován na základě poznatků vyzkoumaných touto prací a nově vyvstávajících metod, trendů a otázek vhodných zkoumání.

4.12.1 Důsledky očekávání společnosti (GPT-3 komentář)

Největšího zklamání se dosahuje při příliš velkých očekáváních. Zprávy vyzdvihující úspěchy LLM vytvářejí velká očekávání. V těchto očekáváních je často úplně vynechán pohled na reálné vlastnosti LLM a jsou vyzdvihovány ve světle jako všeho znalá AI známá z filmů, knih nebo jiných médií. Realita se však s tímto očekáváním rozchází i v případě současných LLM, které mají výstupy čistě statistické. Bez jakýchkoliv rozhodovacích mechanismů zakotvující výstupy na úrovni pravdivosti, emocí, etiky, morálních hodnot nebo jiných aspektů podle kterých se lidé rozhodují při svém vyjadřování a rozhodování. Všechny tyto lidské hodnoty LLM jsou získány pouze jako statistická pravděpodobnost výskytu slov vedle sebe. Jak moc je výstup shodný s lidským vnímáním, je dáno právě zastoupením těchto hodnot v trénovacích datech.

Jednoduše řečeno cílem LLM, je maximalizovat uvěřitelnost výstupu, který je pouze pokračováním vstupu na základě všech textů, s kterými přišlo do styku během učícího procesu. Je lidské se snažit odůvodňovat úspěchy v různých odvětvích, včetně psaní textů, lidskou inteligencí. AI jako jsou GPT modely není však ukázkou inteligence v lidském slova smyslu a spíš poukazuje na možnost dělat zdánlivě složité úkony neinteligentním způsobem. Jev ukazující, že úkony obtížné pro lidi jsou lehké pro AI a naopak je nazýván Moravcův paradox [85]. Jisté je, že už nyní se setkáváme s textem upraveným nebo napsaným pomocí AI. Vše naznačuje tomu, že se budeme více a více setkávat s textem, kde nebude jednoduše rozlišitelné, jakým podílem je autorem AI nebo člověk a jak velký je to problém. [86]

4.12.2 Zákon škálování

Zákon ve své jednoduchosti říká, že lze zvyšovat a zlepšovat výsledky modelů zvětšováním výpočetního výkonu, trénovacích dat nebo počtu parametrů, respektive velikostí modelu. V zákonu je zohledněno, že jeden z těchto tří faktorů nemůže být výrazně menší, aby se nestal brzdou pro zbylé faktory a nesnižoval jejich potenciálně získatelný výkon. Zákon škálování bere výkon ve smyslu chyb, kterých se model dopouští v benchmarcích. Budeme-li uvažovat správnost tohoto zákona, je dobré zamyslet se, jak se škálování projevuje na LLM, a kde jsou limity tohoto zákona. Důležitou otázkou je také užitek směřování k minimální chybě

modelů v benchmarkích, není totiž jisté, jak se tyto charakteristika promítne ve vnímané inteligenci. Zda bude model disponovat novými vlastnostmi z emergentního chování nebo bude pouze zvládat lépe stejné úlohy. Z výsledků benchmarků se dá vyvodit, že z hlediska škálování přináší lepší výsledky škálování trénovacího datasetu než výkonu. Ve spoustě úloh jsou tak lepší menší modely s větším trénovacím datasetem než velké modely s malým datasetem.

Zásadním faktorem, v oblasti trénovacích dat je jejich poměr kvality a kvantity. Kupříkladu ChatGPT je finetunovaným modelem a je to právě ten model, který udělal největší rozruch. Tento fakt podporuje myšlenku, zda nesměřovat k menším finetunovaným modelům specializovaným na konkrétní úlohy.

Opomenutí této myšlenky nás dříve či později může dovést do bodu, kdy na světě nebude dostatečná kvantita dat na učení v požadované kvalitě pro obecné modely. To samé se bere v potaz u škálování výkonu, kde se počítá s trendem zlevňování ceny za výkon grafických karet. Data nejsou nekonečná, ačkoliv se pořád generují nová, většina generovaných dat je nekvalitních. Jako řešení tohoto budoucího problému může být tvorba umělých dat z kvalitních vybraných datasetů. Ačkoliv se tato praktika zdá zvláštní v rámci textu v oblasti obrazových datasetů, je hojně používána a pro textová data to může být inspirací. Aktuálně se nabízí kvalitní data v oblasti e-knih a digitalizace knih fyzických, kde jsou velmi kvalitní textová data. Opakem toho jsou zprávy na sociálních sítích a komentáře, kterých se každodenně generuje velké množství. Jejich kvalita je zpravidla špatná a nemá smysl je používat, pokud se neukáže, že při trénování je potřeba kvantita spíše než kvalita, což se zatím nepotvrzuje [89].

4.12.3 LLM

Dále bude posouvána hranice výkonu zákonem škálování v kombinaci s dalšími změnami v transformer architektuře, dokud nepřijde slibnější architektura nebo princip měnící řešenou problematiku LLM.

4.12.3.1 Symbolika

Symbolická AI je často vnímána jako minulost, přestože disponuje mnohými vlastnostmi, které moderní AI přístupy postrádají. Symbolická AI je průhlednější ve

svých procesech a disponuje jednodušší upravitelností. Symbolika tak může být v budoucnu znovu předmětem zkoumání v neurosymbolických LLM. [87]

4.12.3.2 AI jako služba

Služby nabízející LLM aktualizacemi ovlivňují schopnosti svých služeb a změny nejsou vždy pozitivní. Uživatelé nyní u ChatGPT přicházejí na to, že jeho schopnosti se zhoršily. Do budoucna tak nebudou LLM služby brány jako nástroje s pevně danými schopnostmi. Už vycházejí první výzkumy zabývající se čím dochází ke změnám [88] a otázkou do budoucna je jak měnit LLM bez ztráty stávajících kvalit.

4.12.3.3 Paměť

Velikost rámce kontextu se neustále zvětšuje a není důvod proč by se tento trend měl měnit. Dalším vylepšujícím krokem je zajistit funkci dlouhodobé paměti mimo kontext.

4.12.4 Trénování

Jelikož jsou LLM v aktuální podobě pořád v počátcích, budou se trénovací metody postupně zdokonalovat, ať už jde o kvalitu, efektivitu nebo sladění trénovaného úkonu s užíváním.

4.12.4.1 Změna trénovacího paradigmatu

Od paradigmatu finetunování obecného LLM by se mohlo přejít k prompt-based finetuningu představující udržitelnější, rychlejší a nedestruktivní způsob specializace LLM. Existují i další způsoby jak finetunovat, cílem však vždy zůstává užití metody zachovávající nebo vylepšující schopnosti LLM s co nejmenší výpočetní náročností trénování a provozu.

4.12.4.2 Teorie mrtvého internetu

V budoucnu kvůli povaze získávání datasetů budou objevovat generované texty, které nemají lidské kvality a jejich sémantika zaostává, což může vést ke

zhoršení schopnosti trénovat LLM. Objem nepravdivých dat v budoucích datasetech tak může negativně změnit pravdivost průměrné informace v datasetu. [25]

4.12.4.3 Datasetsy

Pomalou se dochází k poznání, že kromě objemu dat je při trénování důležitá i jejich kvalita, která může do jisté míry substituovat objem. Tento jev popisuje i výzkum[89], jehož výsledkem je model překonávající v programovacích benchmarkech jiné LLM.

Navržený model má desetkrát méně parametrů a stokrát menší objem trénovacích dat v „učebnicové“ kvalitě. Nedostatkem modelu je omezená robustnost a obecnost napříč programovacími jazyky a stylistikou kódu. Autoři míní, že tyto nedostatky nejsou zásadní a mohou být v budoucnu vyřešeny. Vylepšování a tvorba nových metod trénování včetně validace kvality datasetů je při braní v potaz teorie mrtvého internetu velmi důležité téma do budoucnosti.

4.12.5 Cenzura

Obecně je jakákoliv cenzura kontroverzní a ovlivňuje použitelnost nežádoucím způsobem pro všechny uživatele bez rozdílu, ať už má zlý úmysl nebo ne. Cenzurováním LLM dochází k omezování jejich schopnosti zpracovávat různé dotazy, obyčejný uživatel může narazit na nepoužitelnost i pro na první pohled neškodná témata.

4.12.6 Interakce ve světě

Článek od Jacob Browning a Yann LeCun, známých figur v oblasti AI výzkumu vysvětluje, proč nelze od LLM v aktuální podobě očekávat hlubší chápání světa a tím pádem vytvoření AGI. Ačkoliv se setkáváme s názorem, že je veškerá znalost světa přenositelná do jazyka, existuje spousta znalostí, na které se přirozený nebo ani jiné lidmi používané jazyky nehodí. Tyto znalosti pak uchováváme a předáváme jako know-how, mapy, grafy, nahrávky a obrázky atd.

LLM statisticky dosazují slova do sekvencí ve větách, nemají však reálnou představu, co jednotlivá slova znamenají a chybí jim kontext reálného světa. Dobrým přirovnáním je pilný student izolovaný v místnosti, ze které nevychází. Zná

veškerou látku se všemi specifickými slovíčky ke konkrétním tématům jako z encyklopedie. Tento žák však nemá tušení o čem mluví, pouze ví, jaký typ věty a jaká slova použít podle dotázané otázky. Tím, že žák nezná podstatu látky, nemusí zvolit odpověď vždy správně jako jiný žák žijící v reálném světě vnímající informace i mimo encyklopedii.

Někdy proces vypovídá o schopnostech víc než samotný dosažený výsledek. LLM umí bez problémů vysvětlit různé i netriviální úkony. Není však výjimkou, pokud je následně nedokážou provést. Trpí také absencí dlouhodobé paměti a krátkodobá paměť je taktéž značně omezená. Při delší konverzaci začne LLM vypouštět staré části konverzace a tím může narušit integritu svých názorů. To nejsou kvality, kterými disponuje systém koherentně chápající náš svět.

Autoři se svým článkem poukazují na limitaci učení pouze na lingvistických datech a na to jak i lidé jsou neefektivní při čerpání informací z lingvistických dat. A jak nelze získat hlubší pochopení světa bez učení i z jiných typů informací jako jsou symbolická, obrazová, zvuková, pohybová a další.

LLM tak ve své podstatě vnímají text jako svět, místo nástroje jak svět popisovat. Nezahazují však užitečnost těchto systémů a zmiňují, že pro spoustu aplikací není hluboké chápání světa potřeba. [90]

5 Závěry a doporučení

Výsledky rešerše dochází k závěru, že manuální dotazování není optimálním způsobem dotazování už kvůli neprůhlednosti modelu, hledání kompromisu mezi chápáním lidí a LLM. Postupně jsou vyvíjeny metody řešící tuto problematiku skrze kontinuální dotazy nebo automatizované hledání optimálnějších dotazů a odpovědí.

Odklon od přímého diskrétního dotazování je zřejmý i s nástupem nového trénovacího paradigmatu nazývaného prompt-based finetuning, který užívá menší jazykový model k předzpracování dotazů. Užívání menších jazykových modelů jako prostředníků ke komunikaci, vede k možnosti užívat obecné LLM pro více specializací podle připojeného prostředníka, který bude umět s obecným LLM komunikovat a získat z něj znalosti. Pro zlepšení diskrétních manuálních dotazů jsou v práci vypsány nedostatky, na které si dát pozor a zároveň rady, jak správně dotazy tvořit.

Veškeré nedostatky ovlivňující výstupy LLM, ale nejsou způsobeny nevhodností metod dotazování, ale i samotným modelem, respektive architekturou, embedding a možná až triviálním úkonem dosazování slov na základě vzorů v trénovacích datech. Řešením jsou systémy, aplikace, doplňky zjednodušující nebo obcházející tyto problémy za pomoci dedikovaných systémů nahrazující LLM v úlohách, které konzistentně nezvládá.

Trend v trénování směřuje ke kvalitnějším datasetům, kde lze kvalitou nahrazovat i kvantitu, což může do budoucna vést k tvorbě menších specializovaných LLM. Ale i doplňovat se s hledáním limitů zákona škálování při tvorbě gigantických LLM s vizí získání nových emergentních vlastností.

Ze všech těchto poznatků vyplývá, že budoucí modely budou dosahovat kvalitnějších výstupů a společně s poznatky z dotazovacích technik budou pravděpodobně disponovat i dotazováním sladějším s lidským chápáním světa. Přesto nic nenasvědčuje, že by se v blízké budoucnosti vytvořil LLM považovaný za AGI, přestože se může zdát, že je to očekávání společnosti.

Slibná je z principu omezeného vnímání světa skrz text multimodalita modelů. Rozšiřuje možnosti a vnímání světa pro modely a je otázkou, zda právě

rozšiřování počtu typů dat zpracovávaných pomocí AI, bude směřovat k vyvinutí AGI.

Součástí rešerše jsou i kapitoly o benchmarkích a nebezpečích vycházejících z užívání LLM. Benchmarky jsou zde ilustrativně zmíněny, jelikož představují soubory úloh, od kterých se očekává, že budou zpracovatelné pomocí LLM. Kapitola nebezpečí převážně zmiňuje, jak se dají LLM zneužít k nebezpečné činnosti a jak použít dotazování k překonání bezpečnostních zábran LLM.

Zajímavé otázky dalšího zkoumání jsou přenositelnost dotazů mezi LLM a psychologické vlivy na zpracování dotazů.

6 Seznam použité literatury

- [1] NOY, Shakked a Whitney ZHANG. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *SSRN Electronic Journal* [online]. 2023 [vid. 2023-06-27]. ISSN 1556-5068. Dostupné z: doi:10.2139/ssrn.4375283
- [2] *Introduction to Reinforcement Learning with Human Feedback* [online]. [vid. 2023-07-13]. Dostupné z: <https://www.surgehq.ai//blog/introduction-to-reinforcement-learning-with-human-feedback-rlhf-series-part-1>
- [3] HOLMGREN, Gustav, Peder ANDERSSON, Andreas JAKOBSSON a Attila FRIGYESI. Artificial neural networks improve and simplify intensive care mortality prognostication: a national cohort study of 217,289 first-time intensive care unit admissions. *Journal of Intensive Care* [online]. 2019, 7(1), 44. ISSN 2052-0492. Dostupné z: doi:10.1186/s40560-019-0393-1
- [4] JavaScript - artificial neuron model. *Dirask.com, IT Community* [online]. [vid. 2023-03-08]. Dostupné z: <https://dirask.com/posts/JavaScript-artificial-neuron-model-paoM31>
- [5] PETERS, Matthew E., Mark NEUMANN, Mohit IYER, Matt GARDNER, Christopher CLARK, Kenton LEE a Luke ZETTLEMOYER. *Deep contextualized word representations* [online]. B.m.: arXiv. 22. březen 2018 [vid. 2023-07-09]. Dostupné z: doi:10.48550/arXiv.1802.05365. arXiv:1802.05365 [cs]
- [6] AKBIK, Alan, Tanja BERGMANN, Duncan BLYTHE, Kashif RASUL, Stefan SCHWETER a Roland VOLLGRAF. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* [online]. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, s. 54–59 [vid. 2023-07-09]. Dostupné z: doi:10.18653/v1/N19-4010
- [7] What is the difference between word2Vec and Glove? *Machine Learning Interviews* [online]. 14. únor 2019 [vid. 2023-03-13]. Dostupné z: <https://machinelearninginterview.com/topics/natural-language-processing/what-is-the-difference-between-word2vec-and-glove/>
- [8] PAI, Aravindpai. Pretrained Word Embeddings | Word Embedding NLP. *Analytics Vidhya* [online]. 16. březen 2020 [vid. 2023-02-19]. Dostupné z: <https://www.analyticsvidhya.com/blog/2020/03/pretrained-word-embeddings-nlp/>
- [9] *word2vec | TensorFlow Core* [online]. [vid. 2023-03-08]. Dostupné z: https://www.tensorflow.org/tutorials/text/word2vec#negative_sampling_for_one_skip-gram

- [10] RIVA, Martin. *Word Embeddings: CBOW vs Skip-Gram / Baeldung on Computer Science* [online]. 11. březen 2021 [vid. 2023-06-09]. Dostupné z: <https://www.baeldung.com/cs/word-embeddings-cbow-vs-skip-gram>
- [11] *GloVe: Global Vectors for Word Representation* [online]. [vid. 2023-03-13]. Dostupné z: <https://nlp.stanford.edu/projects/glove/>
- [12] PENNINGTON, Jeffrey, Richard SOCHER a Christopher MANNING. Glove: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP): Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* [online]. Doha, Qatar: Association for Computational Linguistics, 2014, s. 1532–1543 [vid. 2023-03-13]. Dostupné z: doi:10.3115/v1/D14-1162
- [13] CAMACHO-COLLADOS, Jose a Mohammad Taher PILEHVAR. From Word To Sense Embeddings: A Survey on Vector Representations of Meaning. *Journal of Artificial Intelligence Research* [online]. 2018, **63**, 743–788. ISSN 1076-9757. Dostupné z: doi:10.1613/jair.1.11259
- [14] *HistWords: Word Embeddings for Historical Text* [online]. [vid. 2023-07-10]. Dostupné z: <https://nlp.stanford.edu/projects/histwords/>
- [15] RICHARD NORDQUIST. What Is Linguistic Variation? *ThoughtCo* [online]. [vid. 2023-07-09]. Dostupné z: <https://www.thoughtco.com/what-is-linguistic-variation-1691242>
- [16] *A Scalable Approach to Reducing Gender Bias in Google Translate* [online]. 22. duben 2020 [vid. 2023-07-06]. Dostupné z: <https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html>
- [17] Google News and Leo Tolstoy: visualizing Word2Vec word embeddings using t-SNE. *Habr* [online]. 30. duben 2019 [vid. 2023-07-31]. Dostupné z: <https://habr.com/en/companies/vk/articles/449984/>
- [18] VASWANI, Ashish, Noam SHAZEER, Niki PARMAR, Jakob USZKOREIT, Llion JONES, Aidan N. GOMEZ, Lukasz KAISER a Illia POLOSUKHIN. *Attention Is All You Need* [online]. B.m.: arXiv. 5. prosinec 2017 [vid. 2023-02-08]. Dostupné z: doi:10.48550/arXiv.1706.03762. arXiv:1706.03762 [cs]
- [19] BAHDANAU, Dzmitry, Kyunghyun CHO a Yoshua BENGIO. *Neural Machine Translation by Jointly Learning to Align and Translate* [online]. B.m.: arXiv. 19. květen 2016 [vid. 2023-07-26]. Dostupné z: doi:10.48550/arXiv.1409.0473. arXiv:1409.0473 [cs, stat]
- [20] KELES, Feyza Duman, Pruthuvi Mahesakya WIJewardena a Chinmay HEGDE. *On The Computational Complexity of Self-Attention* [online]. B.m.: arXiv. 11. září 2022 [vid. 2023-07-26]. Dostupné z: doi:10.48550/arXiv.2209.04881. arXiv:2209.04881 [cs]

- [21] TAY, Yi, Mostafa DEGHANI, Dara BAHRI a Donald METZLER. *Efficient Transformers: A Survey* [online]. B.m.: arXiv. 14. březen 2022 [vid. 2023-06-06]. Dostupné z: <http://arxiv.org/abs/2009.06732>. arXiv:2009.06732 [cs]
- [22] LU, Chengqiang, Jianwei ZHANG, Yunfei CHU, Zhengyu CHEN, Jingren ZHOU, Fei WU, Haiqing CHEN a Hongxia YANG. *Knowledge Distillation of Transformer-based Language Models Revisited* [online]. B.m.: arXiv. 12. červenec 2022 [vid. 2023-07-26]. Dostupné z: [doi:10.48550/arXiv.2206.14366](https://arxiv.org/abs/2206.14366). arXiv:2206.14366 [cs]
- [23] SCHICK, Timo, Jane DWIVEDI-YU, Roberto DESSÌ, Roberta RAILEANU, Maria LOMELI, Luke ZETTLEMOYER, Nicola CANCEDDA a Thomas SCIALOM. *Toolformer: Language Models Can Teach Themselves to Use Tools* [online]. B.m.: arXiv. 9. únor 2023 [vid. 2023-06-08]. Dostupné z: [doi:10.48550/arXiv.2302.04761](https://arxiv.org/abs/2302.04761). arXiv:2302.04761 [cs]
- [24] *MLM — Sentence-Transformers documentation* [online]. [vid. 2023-07-31]. Dostupné z: https://www.sbert.net/examples/unsupervised_learning/MLM/README.html
- [25] SOBIESZEK, Adam a Tadeusz PRICE. Playing Games with AIs: The Limits of GPT-3 and Similar Large Language Models. *Minds and Machines* [online]. 2022, **32**(2), 341–364. ISSN 1572-8641. Dostupné z: [doi:10.1007/s11023-022-09602-0](https://doi.org/10.1007/s11023-022-09602-0)
- [26] How AI virtual agents can automate lead generation. *Talkie.ai* [online]. [vid. 2023-07-03]. Dostupné z: <https://talkie.ai/blog/lead-generation-automation/>
- [27] *Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone* [online]. 8. květen 2018 [vid. 2023-07-03]. Dostupné z: <https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html>
- [28] *How does DeepL work?* [online]. [vid. 2023-07-03]. Dostupné z: <https://www.deepl.com/en/blog/how-does-deepl-work>
- [29] *Recent Advances in Google Translate* [online]. 8. červen 2020 [vid. 2023-07-03]. Dostupné z: <https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html>
- [30] TODD, Graham, Sam EARLE, Muhammad Umair NASIR, Michael Cerny GREEN a Julian TOGELIUS. Level Generation Through Large Language Models. In: *Proceedings of the 18th International Conference on the Foundations of Digital Games* [online]. 2023, s. 1–8 [vid. 2023-07-04]. Dostupné z: [doi:10.1145/3582437.3587211](https://doi.org/10.1145/3582437.3587211)

- [31] [HTTPS://WWW.FACEBOOK.COM/DUOLINGO](https://www.facebook.com/duolingo). Introducing Duolingo Max, a learning experience powered by GPT-4. *Duolingo Blog* [online]. 14. březen 2023 [vid. 2023-07-02]. Dostupné z: <https://blog.duolingo.com/duolingo-max/>
- [32] BENNANI-SMIRE, Kamil, Claudiu MUSAT, Andreea HOSSMANN, Michael BAERISWYL a Martin JAGGI. Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. In: *CoNLL 2018: Proceedings of the 22nd Conference on Computational Natural Language Learning* [online]. Brussels, Belgium: Association for Computational Linguistics, 2018, s. 221–229 [vid. 2023-07-03]. Dostupné z: doi:10.18653/v1/K18-1022
- [33] DING, Haoran a Xiao LUO. AttentionRank: Unsupervised Keyphrase Extraction using Self and Cross Attentions. In: *EMNLP 2021: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* [online]. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, s. 1919–1928 [vid. 2023-05-22]. Dostupné z: doi:10.18653/v1/2021.emnlp-main.146
- [34] WEI, Jason, Yi TAY, Rishi BOMMASANI, Colin RAFFEL, Barret ZOPH, Sebastian BORGEAUD, Dani YOGATAMA, Maarten BOSMA, Denny ZHOU, Donald METZLER, Ed H. CHI, Tatsunori HASHIMOTO, Oriol VINYALS, Percy LIANG, Jeff DEAN a William FEDUS. *Emergent Abilities of Large Language Models* [online]. B.m.: arXiv. 26. říjen 2022 [vid. 2023-03-10]. Dostupné z: doi:10.48550/arXiv.2206.07682. arXiv:2206.07682 [cs]
- [35] RAY, Partha Pratim. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* [online]. 2023, 3, 121–154. ISSN 2667-3452. Dostupné z: doi:10.1016/j.iotcps.2023.04.003
- [36] *Fine-tuning a Neural Network explained* [online]. [vid. 2023-03-11]. Dostupné z: <https://deeplizard.com/learn/video/5T-iXNNiwIs>
- [37] *Learning from human preferences* [online]. [vid. 2023-07-10]. Dostupné z: <https://openai.com/research/learning-from-human-preferences>
- [38] GABRIELLE KAILI-MAY LIU. *Transforming Human Interactions with AI via Reinforcement Learning with Human Feedback (RLHF)* [online]. 28. únor 2023. Dostupné z: <https://computing.mit.edu/wp-content/uploads/2023/06/Transforming-Human-Interactions-with-AI-via-Reinforcement-Learning-with-Human-Feedback-RLHF.pdf>
- [39] SCHULMAN, John, Filip WOLSKI, Prafulla DHARIWAL, Alec RADFORD a Oleg KLIMOV. *Proximal Policy Optimization Algorithms* [online]. B.m.: arXiv. 28. srpen 2017 [vid. 2023-03-10]. Dostupné z: doi:10.48550/arXiv.1707.06347. arXiv:1707.06347 [cs]

- [40] *Holy \$#!t: Are popular toxicity models simply profanity detectors?* [online]. [vid. 2023-07-14]. Dostupné z: <https://www.surgehq.ai//blog/are-popular-toxicity-models-simply-profanity-detectors>
- [41] *30% of Google's Emotions Dataset is Mislabeled* [online]. [vid. 2023-07-23]. Dostupné z: <https://www.surgehq.ai//blog/30-percent-of-googles-reddit-emotions-dataset-is-mislabeled>
- [42] ZHAO, Yao, Rishabh JOSHI, Tianqi LIU, Misha KHALMAN, Mohammad SALEH a Peter J. LIU. *SLiC-HF: Sequence Likelihood Calibration with Human Feedback* [online]. B.m.: arXiv. 17. květen 2023 [vid. 2023-07-23]. Dostupné z: doi:10.48550/arXiv.2305.10425. arXiv:2305.10425 [cs]
- [43] DEMSZKY, Dorottya, Dana MOVSHOVITZ-ATTIAS, Jeongwoo KO, Alan COWEN, Gaurav NEMADE a Sujith RAVI. *GoEmotions: A Dataset of Fine-Grained Emotions* [online]. B.m.: arXiv. 2. červen 2020 [vid. 2023-07-23]. Dostupné z: <http://arxiv.org/abs/2005.00547>. arXiv:2005.00547 [cs]
- [44] LIU, Pengfei, Weizhe YUAN, Jinlan FU, Zhengbao JIANG, Hiroaki HAYASHI a Graham NEUBIG. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys* [online]. 2023, **55**(9), 195:1-195:35. ISSN 0360-0300. Dostupné z: doi:10.1145/3560815
- [45] An Introduction to Large Language Models: Prompt Engineering and P-Tuning. *NVIDIA Technical Blog* [online]. 26. duben 2023 [vid. 2023-07-31]. Dostupné z: <https://developer.nvidia.com/blog/an-introduction-to-large-language-models-prompt-engineering-and-p-tuning/>
- [46] New Benchmarks Test the Limits of Large Language Models. *New Benchmarks Test the Limits of Large Language Models* [online]. 22. červen 2022 [vid. 2023-07-20]. Dostupné z: <https://www.deeplearning.ai/the-batch/toward-next-gen-language-models/>
- [47] *How Surge AI Built OpenAI's GSM8K Dataset of 8,500 Math Problems* [online]. [vid. 2023-07-20]. Dostupné z: <https://www.surgehq.ai//blog/how-we-built-it-openais-gsm8k-dataset-of-8500-math-problems>
- [48] COBBE, Karl, Vineet KOSARAJU, Mohammad BAVARIAN, Mark CHEN, Heewoo JUN, Lukasz KAISER, Matthias PLAPPERT, Jerry TWOREK, Jacob HILTON, Reiichiro NAKANO, Christopher HESSE a John SCHULMAN. *Training Verifiers to Solve Math Word Problems* [online]. B.m.: arXiv. 17. listopad 2021 [vid. 2023-07-20]. Dostupné z: doi:10.48550/arXiv.2110.14168. arXiv:2110.14168 [cs]
- [49] WANG, Alex, Yada PRUKSACHATKUN, Nikita NANGIA, Amanpreet SINGH, Julian MICHAEL, Felix HILL, Omer LEVY a Samuel R BOWMAN. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. nedatováno, 30.

- [50] *The Stanford Question Answering Dataset* [online]. [vid. 2023-07-20]. Dostupné z: <https://rajpurkar.github.io/SQuAD-explorer/>
- [51] CHIUSANO, Fabio. Two minutes NLP — Keeping track of information and the LAMBADA benchmark. *NLPlanet* [online]. 30. březen 2022 [vid. 2023-07-20]. Dostupné z: <https://medium.com/nlplanet/two-minutes-nlp-keeping-track-of-information-and-the-lambda-benchmark-b808dd5af15c>
- [52] ZELLERS, Rowan, Ari HOLTZMAN, Yonatan BISK, Ali FARHADI a Yejin CHOI. *HellaSwag: Can a Machine Really Finish Your Sentence?* [online]. B.m.: arXiv. 19. květen 2019 [vid. 2023-07-20]. Dostupné z: doi:10.48550/arXiv.1905.07830. arXiv:1905.07830 [cs]
- [53] REDDY, Siva, Danqi CHEN a Christopher D. MANNING. *CoQA: A Conversational Question Answering Challenge* [online]. B.m.: arXiv. 29. březen 2019 [vid. 2023-07-21]. Dostupné z: doi:10.48550/arXiv.1808.07042. arXiv:1808.07042 [cs]
- [54] LIU, Jian, Leyang CUI, Hanmeng LIU, Dandan HUANG, Yile WANG a Yue ZHANG. *LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning* [online]. B.m.: arXiv. 16. červenec 2020 [vid. 2023-07-21]. Dostupné z: <http://arxiv.org/abs/2007.08124>. arXiv:2007.08124 [cs]
- [55] HENDRYCKS, Dan, Collin BURNS, Steven BASART, Andy ZOU, Mantas MAZEIKA, Dawn SONG a Jacob STEINHARDT. *Measuring Massive Multitask Language Understanding* [online]. B.m.: arXiv. 12. leden 2021 [vid. 2023-07-22]. Dostupné z: doi:10.48550/arXiv.2009.03300. arXiv:2009.03300 [cs]
- [56] ZHANG, Min a Juntao LI. A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research* [online]. 2021, **1**(6), 831–833. ISSN 2667-3258. Dostupné z: doi:10.1016/j.fmre.2021.11.011
- [57] OUYANG, Long, Jeff WU, Xu JIANG, Diogo ALMEIDA, Carroll L. WAINWRIGHT, Pamela MISHKIN, Chong ZHANG, Sandhini AGARWAL, Katarina SLAMA, Alex RAY, John SCHULMAN, Jacob HILTON, Fraser KELTON, Luke MILLER, Maddie SIMENS, Amanda ASKELL, Peter WELINDER, Paul CHRISTIANO, Jan LEIKE a Ryan LOWE. *Training language models to follow instructions with human feedback* [online]. B.m.: arXiv. 4. březen 2022 [vid. 2023-07-10]. Dostupné z: <http://arxiv.org/abs/2203.02155>. arXiv:2203.02155 [cs]
- [58] OPENAI. *GPT-4 Technical Report* [online]. B.m.: arXiv. 27. březen 2023 [vid. 2023-07-27]. Dostupné z: doi:10.48550/arXiv.2303.08774. arXiv:2303.08774 [cs]
- [59] *GPT-4* [online]. [vid. 2023-07-22]. Dostupné z: <https://openai.com/research/gpt-4>
- [60] FEZARI, Mohamed, Ali AL DAHOUD a Ahmed AL-DAHOUD. *From GPT to AutoGPT: a Brief Attention in NLP Processing using DL* [online]. 2023. Dostupné z: doi:10.13140/RG.2.2.28385.99688

- [61] RUMBELOW, Jessica a MWATKINS. SolidGoldMagikarp (plus, prompt generation) [online]. nedatováno [vid. 2023-02-28]. Dostupné z: <https://www.alignmentforum.org/posts/aPeJE8bSo6rAFoLqg/solidgoldmagikarp-plus-prompt-generation>
- [62] *GPT-3 colored by token probabilities* [online]. [vid. 2023-03-10]. Dostupné z: <https://lucianoabriata.altervista.org/tests/gpt-3/js-only/GPT3-JSonly-logprobs-color.html>
- [63] LUCIANOSPHERE. Exploring Token Probabilities as a Means to Filter GPT-3's Answers. *Medium* [online]. 19. leden 2023 [vid. 2023-02-28]. Dostupné z: <https://towardsdatascience.com/exploring-token-probabilities-as-a-means-to-filter-gpt-3s-answers-3e7dfc9ca0c>
- [64] *ChatGPT plugins* [online]. [vid. 2023-07-27]. Dostupné z: <https://openai.com/blog/chatgpt-plugins>
- [65] YANG, Hui, Sifu YUE a Yunzhong HE. *Auto-GPT for Online Decision Making: Benchmarks and Additional Opinions* [online]. B.m.: arXiv. 3. červen 2023 [vid. 2023-07-16]. Dostupné z: doi:10.48550/arXiv.2306.02224. arXiv:2306.02224 [cs]
- [66] *ChatGPT Gets Its "Wolfram Superpowers"!* [online]. 23. březen 2023 [vid. 2023-07-27]. Dostupné z: <https://writings.stephenwolfram.com/2023/03/chatgpt-gets-its-wolfram-superpowers/>
- [67] Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. *The Official Microsoft Blog* [online]. 7. únor 2023 [vid. 2023-03-13]. Dostupné z: <https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/>
- [68] DAI, Damai, Yutao SUN, Li DONG, Yaru HAO, Shuming MA, Zhifang SUI a Furu WEI. *Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers* [online]. B.m.: arXiv. 15. květen 2023 [vid. 2023-07-28]. Dostupné z: doi:10.48550/arXiv.2212.10559. arXiv:2212.10559 [cs]
- [69] KOJIMA, Takeshi, Shixiang Shane GU, Machel REID, Yutaka MATSUO a Yusuke IWASAWA. *Large Language Models are Zero-Shot Reasoners* [online]. B.m.: arXiv. 29. leden 2023 [vid. 2023-02-02]. Dostupné z: doi:10.48550/arXiv.2205.11916. arXiv:2205.11916 [cs]
- [70] SINCLAIR, Arabella, Jaap JUMELET, Willem ZUIDEMA a Raquel FERNÁNDEZ. Structural Persistence in Language Models: Priming as a Window into Abstract Language Representations. *Transactions of the Association for Computational Linguistics* [online]. 2022, **10**, 1031–1050. Dostupné z: doi:10.1162/tacl_a_00504

- [71] SURI, Gaurav, Lily R. SLATER, Ali ZIAEE a Morgan NGUYEN. *Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5* [online]. B.m.: arXiv. 7. květen 2023 [vid. 2023-06-01]. Dostupné z: doi:10.48550/arXiv.2305.04400. arXiv:2305.04400 [cs, q-bio]
- [72] *What are tokens and how to count them? | OpenAI Help Center* [online]. [vid. 2023-07-19]. Dostupné z: <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them>
- [73] LU, Yao, Max BARTOLO, Alastair MOORE, Sebastian RIEDEL a Pontus STENETORP. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In: *ACL 2022: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* [online]. Dublin, Ireland: Association for Computational Linguistics, 2022, s. 8086–8098 [vid. 2023-07-19]. Dostupné z: doi:10.18653/v1/2022.acl-long.556
- [74] Exclusive: The \$2 Per Hour Workers Who Made ChatGPT Safer. *Time* [online]. 18. leden 2023 [vid. 2023-03-13]. Dostupné z: <https://time.com/6247678/openai-chatgpt-kenya-workers/>
- [75] VINCENT, James. Microsoft's Bing is an emotionally manipulative liar, and people love it. *The Verge* [online]. 15. únor 2023 [vid. 2023-07-29]. Dostupné z: <https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams>
- [76] BlackMamba Leverages AI to Elude Traditional Security Defenses: Blog: Hitachi Systems Security. <https://hitachi-systems-security.com/> [online]. [vid. 2023-07-29]. Dostupné z: <https://hitachi-systems-security.com/black-mamba/>
- [77] PEARCY, Sam. The Dark Side of Large Language Models | HiddenLayer MLDR. *HiddenLayer | Security for AI* [online]. 23. březen 2023 [vid. 2023-07-29]. Dostupné z: <https://hiddenlayer.com/research/the-dark-side-of-large-language-models/>
- [78] ChatGPT: The Dark Side of Artificial Intelligence Crafting Custom Malware. *Heimdall Security Blog* [online]. 22. únor 2023 [vid. 2023-03-13]. Dostupné z: <https://heimdalsecurity.com/blog/chatgpt-the-dark-side-of-artificial-intelligence-crafting-custom-malware/>
- [79] SERGEYSHY. OPWNAI: Cybercriminals Starting to Use ChatGPT. *Check Point Research* [online]. 6. leden 2023 [vid. 2023-07-29]. Dostupné z: <https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/>
- [80] *Prompt injection attacks against GPT-3* [online]. [vid. 2023-02-13]. Dostupné z: <http://simonwillison.net/2022/Sep/12/prompt-injection/>

- [81] GRESHAKE, Kai, Sahar ABDELNABI, Shailesh MISHRA, Christoph ENDRES, Thorsten HOLZ a Mario FRITZ. *More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models* [online]. B.m.: arXiv. 23. únor 2023 [vid. 2023-03-13]. Dostupné z: doi:10.48550/arXiv.2302.12173. arXiv:2302.12173 [cs]
- [82] KANG, Daniel, Xuechen LI, Ion STOICA, Carlos GUESTRIN, Matei ZAHARIA a Tatsunori HASHIMOTO. *Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks* [online]. B.m.: arXiv. 11. únor 2023 [vid. 2023-06-06]. Dostupné z: doi:10.48550/arXiv.2302.05733. arXiv:2302.05733 [cs]
- [83] GOSWAMI, Rohan. ChatGPT's „jailbreak" tries to make the A.I. break its own rules, or die. *CNBC* [online]. 6. únor 2023 [vid. 2023-07-29]. Dostupné z: <https://www.cnbc.com/2023/02/06/chatgpt-jailbreak-forces-it-to-break-its-own-rules.html>
- [84] MARKOV, Todor, Chong ZHANG, Sandhini AGARWAL, Tyna ELOUNDOU, Teddy LEE, Steven ADLER, Angela JIANG a Lilian WENG. *A Holistic Approach to Undesired Content Detection in the Real World* [online]. B.m.: arXiv. 14. únor 2023 [vid. 2023-07-28]. Dostupné z: <http://arxiv.org/abs/2208.03274>. arXiv:2208.03274 [cs]
- [85] *What is true about Moravec's paradox?* [online]. [vid. 2023-03-10]. Dostupné z: <https://indiaai.gov.in/article/what-is-true-about-moravec-s-paradox>
- [86] FLORIDI, Luciano a Massimo CHIRIATTI. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines* [online]. 2020, **30**(4), 681–694. ISSN 0924-6495. Dostupné z: doi:10.1007/s11023-020-09548-1
- [87] PAN, Liangming, Alon ALBALAK, Xinyi WANG a William Yang WANG. *Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning* [online]. B.m.: arXiv. 20. květen 2023 [vid. 2023-07-30]. Dostupné z: doi:10.48550/arXiv.2305.12295. arXiv:2305.12295 [cs]
- [88] CHEN, Lingjiao, Matei ZAHARIA a James ZOU. *How is ChatGPT's behavior changing over time?* [online]. B.m.: arXiv. 18. červenec 2023 [vid. 2023-07-29]. Dostupné z: <http://arxiv.org/abs/2307.09009>. arXiv:2307.09009 [cs]
- [89] GUNASEKAR, Suriya, Yi ZHANG, Jyoti ANEJA, Caio César Teodoro MENDES, Allie DEL GIORNO, Sivakanth GOPI, Mojan JAVAHERIPI, Piero KAUFFMANN, Gustavo DE ROSA, Olli SAARIKIVI, Adil SALIM, Shital SHAH, Harkirat Singh BEHL, Xin WANG, Sébastien BUBECK, Ronen ELKAN, Adam Tauman KALAI, Yin Tat LEE a Yuanzhi LI. *Textbooks Are All You Need* [online]. B.m.: arXiv. 20. červen 2023 [vid. 2023-07-29]. Dostupné z: doi:10.48550/arXiv.2306.11644. arXiv:2306.11644 [cs]

[90] BROWNING, Jacob. AI And The Limits Of Language [online]. 2022 [vid. 2023-03-10]. Dostupné z: <https://www.noemamag.com/ai-and-the-limits-of-language>

Zadání diplomové práce

Autor: Bc. Michael Bartoš

Studium: I2100051

Studijní program: N1802 Aplikovaná informatika

Studijní obor: Aplikovaná informatika

Název diplomové práce: **Transformery a zpracování textu**

Název diplomové práce AJ: Transformers and text processing

Cíl, metody, literatura, předpoklady:

Cíl: Optimalizace dotazování jazykových modelů.

Natural language processing

ANN, Transformery, GPT

Dotazování

LIN, Tianyang, et al. A survey of transformers. *AI Open*, 2022.

TAY, Yi, et al. Efficient transformers: A survey. *ACM Computing Surveys (CSUR)*, 2020.

DALE, Robert. GPT-3: What's it good for?. *Natural Language Engineering*, 2021, 27.1: 113-118.

Zadávací pracoviště: Katedra informačních technologií,
Fakulta informatiky a managementu

Vedoucí práce: Ing. Karel Mls, Ph.D.

Datum zadání závěrečné práce: 15.10.2021