



FILOZOFICKÁ FAKULTA
Univerzita Karlova

Ústav informačních studií a knihovnictví

Diplomová práce

Lenka Pittnerová

**Lokalizace a přizpůsobení anglického benchmarku
SuperGLUE pro hodnocení kompetence textových modelů
umělé inteligence v oblasti českého jazyka**

Localisation and adaptation of the English benchmark SuperGLUE for assessing
the competence of text-based artificial intelligence models in the Czech
language domain

Vedoucí práce:

Mgr. Josef Šlerka, Ph.D.

Studijní program:

Studia nových médií

Studijní obor:

5SNM

Praha 2024

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že nebyla předložena jako splnění studijní povinnosti v rámci jiného studia nebo předložena k obhajobě v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze, dne 15. prosince 2024

Jméno a příjmení

Klíčová slova (česky)

Jazykové modely AI, zpracování přirozeného jazyka, benchmark SuperGLUE, zpracování češtiny, vícejazyčné hodnocení AI, jazykově specifické benchmarkování

Klíčová slova (anglicky)

AI Language Models, Natural Language Processing, SuperGLUE Benchmark, Czech Language Processing, Multilingual AI Evaluation, Language-Specific Benchmarking

Abstrakt: Tato diplomová práce se zabývá hodnocením schopností velkých jazykových modelů (LLM) v českém jazyce prostřednictvím srovnání strojově a manuálně přeložené úlohy Winograd Schema Challenge z benchmarku SuperGLUE. Hodnotící rámec SuperGLUE představuje uznávanou sadu standardizovaných testů pro posuzování schopností umělé inteligence v oblasti zpracování přirozeného jazyka. Přestože velké jazykové modely prokazují pozoruhodné schopnosti i v jiných jazycích než angličtině, jejich systematické hodnocení v neanglofonním prostředí zůstává výzvou. Důvodem je především nedostatek kvalitních hodnotících rámců, které by zohledňovaly specifické vlastnosti jednotlivých jazyků.

Teoretická část práce definuje klíčové pojmy související s benchmarkingem modelů umělé inteligence pomocí vícejazyčného hodnocení a zaměřuje se na omezení vzniklá při překladech benchmarků zaměřených na angličtinu, v tomto případě rámce SuperGLUE, pro jiné jazyky. Zkoumá problémy strojového i manuálního překladu zároveň se snahou zachovat jazykovou přesnost a validitu testů.

Praktická část představuje případovou studii porovnávající výsledky testování skrze strojově přeloženou, manuálně přeloženou a původní (anglickou) úlohu Winograd Schema Challenge pod benchmarkem SuperGLUE. Testuje metodicky vybrané velké jazykové modely s cílem posoudit, jak zvolené metody překladu ovlivňují výsledky hodnocení.

Tato práce má za cíl přispět k vícejazyčnému hodnocení velkých jazykových modelů tím, že zdůrazňuje potřebu lokalizovaných nebo specializovaných benchmarků pro zajištění spolehlivé hodnocení těchto modelů v českém jazyce.

Abstract: This thesis investigates the capabilities of large language models (LLMs) in Czech by comparing machine and human translations of the Winograd Schema Challenge, a component of the SuperGLUE benchmark. While SuperGLUE serves as a widely recognized framework for evaluating natural language processing capabilities in artificial intelligence systems, the assessment of LLMs in non-English languages presents unique challenges. Despite these models demonstrating impressive performance across multiple languages, their systematic evaluation in non-English contexts remains difficult, largely due to the absence of robust assessment frameworks that account for language-specific characteristics.

This thesis compares the reliability of machine-translated and manually translated benchmark task in evaluating large language models (LLM) in Czech. Although large language models such as GPT-4o or Claude-3.5-Sonnet show convincing skills in languages other than English, evaluating their performance in non-English contexts is challenging due to the lack of well-defined benchmarks adapted to relevant language-specific requirements.

The theoretical section defines key concepts related to AI model benchmarking and multilingual evaluation and focuses on the limitations arising when translating English-focused benchmarks, in this case the SuperGLUE framework, for other languages. It explores the problems of machine and manual translation while attempting to maintain linguistic accuracy and test validity.

The practical part presents a case study comparing results of testing through machine-translated, manually translated and original (English) Winograd Schema Challenge task under the SuperGLUE benchmark. It tests methodologically selected large language models to assess how the chosen translation methods affect the evaluation results.

This work contributes to multilingual evaluation of large language models by highlighting the need for localized or specialized benchmarks to ensure reliable evaluation of these models in the context of the Czech language.

Obsah

Úvod.....	6
1 Teoretický úvod do problematiky	8
1.1 Velké jazykové modely	8
1.1.1 NLP	11
1.1.2 NLU.....	13
1.2 Benchmarky v oblasti porozumění přirozenému jazyku (NLU)	15
1.2.1 Benchmark SuperGLUE.....	16
1.2.2 Existující české benchmarky	23
2 Metodologie práce.....	29
2.1 Analýza a výběr úloh v rámci benchmarku	29
2.1.1 Winograd Schema Challenge.....	29
2.1.2 Argumentace pro výběr WSC.....	30
2.1.3 Dataset pro WSC.....	31
2.2 Překlad datasetu	33
2.2.1 Překlad pomocí DeepL	33
2.2.2 Adaptace do českého jazyka.....	34
2.3 Výběr modelů pro testování.....	34
2.3.1 GPT-4 by OpenAI	36
2.3.2 Gemini by Google DeepMind	37
2.3.3 Claude 3.5 Sonnet by Anthropic	39
2.3.4 Llama 3 by Meta	40
2.4 Testování.....	41
2.5 Vyhodnocení výsledků	41
2.6 Hypotézy a předpoklady testování.....	42
3 Praktická část.....	45
3.1 Překlad datasetu	45
3.1.1 Strojový překlad pomocí nástroje DeepL	45
3.1.2 Přizpůsobení přeloženého datasetu	45
3.2 Příprava skriptu.....	47
3.2.1 Prompt.....	47
3.2.2 Příprava skriptů pro testování	50
3.3 Testování:	53
3.3.1 Vyhodnocení.....	54
Závěr	67
Seznam použité literatury a dalších zdrojů	72
Glosář	79
Přílohy.....	VII
Přílohy součástí práce přiložené zvlášť	VII
Příloha 1 – Python skript pro konverzi souboru z formátu .jsonl do .xlsx.....	VII
Příloha 2 - Python skript pro konverzi souboru z formátu .xlsx do .jsonl	VII
Příloha 3 - Python skript pro indexaci entit.....	VII
Příloha 4 – Výsledky testování modelu Claude.....	VII
Příloha 5 – Výsledky testování modelu Gemini	VII
Příloha 6 – Výsledky testování modelu GPT.....	VII
Příloha 7 – Výsledky testování modelu Llama	VII
Příloha 8 – Adaptovaný dataset	VII
Příloha 9 – Strojově přeložený dataset	VII

<i>Příloha 10 – Originální dataset</i>	<i>VII</i>
<i>Příloha 11 – Skript pro testování modelu GPT.....</i>	<i>VIII</i>
<i>Příloha 12 – Skript pro testování modelu Llama</i>	<i>VIII</i>
<i>Příloha 13 – Skript pro testování modelu Claude.....</i>	<i>VIII</i>
<i>Příloha 14 – Skript pro testování modelu Gemini.....</i>	<i>VIII</i>

Úvod

V posledních letech dochází díky rychlému rozvoji technologií umělé inteligence (**Artificial Intelligence**, dále také jen **AI**) a zpracování přirozeného jazyka (**Natural Language Processing**, dále také jen **NLP**) k jejich rozsáhlé integraci do každodenního života, ale i profesního a vzdělávacího prostředí. Zejména nástroje generativní umělé inteligence zaměřené na text, jako jsou například ChatGPT, Google Gemini, Claude.ai a další podobné služby, si získaly značnou pozornost díky možnosti vytvářet poměrně koherentní a kontextově relevantní textové výstupy. Nejen díky tomuto rozvoji se generativní umělá inteligence stala nepostradatelným nástrojem v různých oblastech lidského života, od tvorby obsahu na sociální síti po automatizovanou komunikaci v rámci péče firem o zákazníky. Ačkoli tyto textové modely vykazují významné schopnosti při generování textu také v dalších jazycích mimo angličtinu, a to včetně češtiny, hodnocení jejich jazykové kompetence v dané lokalizaci zůstává značnou výzvou vzhledem k absenci na míru šitých referenčních hodnot nebo **benchmarků**¹, které by odpovídaly daným specifickým pravopisným pravidlům, gramatiky či stylistiky českého jazyka.

Navzdory objektivně kvalitním výsledkům nejsou modely umělé inteligence bez omezení. Všechny existující modely jsou primárně trénovány na rozsáhlých internetových zdrojích, z nichž většina je v angličtině, protože je to nejdostupnější a nejrozšířenější jazyk na internetu. Přestože si často překvapivě dobře vedou i v jiných jazycích. Jejich výstupy jsou však často ovlivněny strukturou a nuancemi angličtiny, což vede k výsledkům vytvořeným pomocí přímého překladu, které plně nedodržují lingvistické normy cílového jazyka.

Pro účely testování modelů v této práci byl zvolen benchmark **SuperGLUE**, který spolu s dalšími hodnotícími rámci primárně měří schopnost modelů porozumět anglickému jazyku a kontextu. Tyto benchmarky poskytují komplexní metodiku pro hodnocení, včetně rozsáhlých souborů dat, otázek a systémů hodnocení. Pro možnost evaluovat i další jazyky se k úpravě těchto benchmarků často používají nástroje strojového překladu. Například autoři Žagar a Robnik-Šikonja (2022) přeložili benchmark SuperGLUE do slovinštiny kombinací strojového a lidského překladu. Jejich studie analyzuje výzvy spojené s morfologickými a gramatickými rozdíly a hodnotí výkon jazykových modelů na přeložených datových sadách.

¹ Test určený k vyhodnocení nebo porovnání výkonnosti hardwaru nebo softwaru; část softwaru, soubor dat atd., který je k tomuto účelu navržen nebo používán (Benchmark, n., sense 2.b, 2024).

Další relevantní studií je práce *Spanish and LLM Benchmarks: is MMLU Lost in Translation?* (Plaza a kol., 2024), ve které autoři pojednávají o překladu benchmarku Massive Multitask Language Understanding (MMLU) do španělštiny s využitím kombinace nástroje strojového překladu Azure Translator a služby ChatGPT. Autoři této práce analyzují vliv kvality překladu na výkon benchmarku a zdůrazňují důležitost přesných překladů pro spolehlivé hodnocení. Tento přístup však vyvolává otázky ohledně adekvátnosti přímých překladů příslušných úloh pro možnost správně simulovat a testovat jedinečné jazykové jevy. Strojový překlad nemůže z podstaty zajistit, že budou zvolené otázky a kritéria hodnocení určené pro anglický jazyk správně odrážet také specifika a komplexnost cílového jazyka.

Tato práce se zabývá otázkou „**jaká je spolehlivost strojově přeloženého benchmarku ve srovnání s verzí přeloženou lidským překladatelem při hodnocení jazykových kompetencí velkých jazykových modelů v českém jazyce?**“. Cílem je porovnat výkon čtyř běžně dostupných nástrojů pro generování textu založených na velkých jazykových modelech a zpracování přirozeného jazyka. K testování bude využita úloha Winograd Schema Challenge (WSC) z benchmarku SuperGLUE, převedená do českého jazyka. Výzkum se zaměří na to, zda manuální překlad a jazyková adaptace benchmarku významně ovlivní výsledky testování oproti strojově přeloženému datasetu a původní anglické verzi.

Autorka práce vystudovala český a anglický jazyk se zaměřením na vzdělávání na Pedagogické fakultě Univerzity Karlovy. Díky několikaleté praxi v oblasti překladu a validace strojového překladu mezi angličtinou a češtinou disponuje odbornými znalostmi potřebnými pro kvalifikovanou adaptaci testových otázek datasetu.

Výzkum přispívá k rozšiřující se oblasti hodnocení vícejazyčné umělé inteligence a zdůrazňuje důležitost lokalizovaných evaluačních metrik pro přesnější posouzení schopností jazykových modelů v českém prostředí.

1 Teoretický úvod do problematiky

V této části budou přiblíženy vybrané teoretické oblasti důležité pro porozumění metodologii a praktické části práce. Sekce je zaměřena na zvolený benchmark a jeho úlohy, jazykové modely a argumentaci pro jejich výběr k testování.

První část této práce se věnuje teoretickému úvodu do problematiky velkých jazykových modelů (**Large Language Models**, dále také jen **LLM**) a jejich evaluace. Text nejprve představuje základní charakteristiky LLM, včetně jejich architektury, způsobu přípravy dat a jejich tréninku, popisuje související klíčové oblasti porozumění přirozenému jazyku (**Natural Language Understanding**, dále také jen **NLU**) a zpracování přirozeného jazyka (**NLP**). Cílem těchto sekcí je vysvětlit, jak tyto modely zpracovávají nebo generují text.

Další část je zaměřena na seznámení s benchmarky v oblasti porozumění přirozenému jazyku, přičemž detailně rozebírám především zvolený benchmark SuperGLUE a osm jeho specifických úloh, které testují různé aspekty jazykového porozumění. V textu je benchmark dále srovnáván s dalšími významnými rámci MMLU, HELM či BIG-bench, a jsou uvedeny argumenty, proč byl pro účely této práce vybrán právě SuperGLUE.

Závěrečná část se věnuje zhodnocení situace v oblasti českých benchmarků, ve které byl nedávno představen BenCzechMark jako první komplexní hodnotící rámec pro oblast českého jazyka.

1.1 Velké jazykové modely

Velké jazykové modely (**LLM**) jsou typy výpočetního modelu, které byly navrženy s cílem zpracovávat přirozený jazyk. Modely dokáží v rámci existujících limitů zpracovávat a generovat text, do určité míry také pracovat s jeho kontextem, provádět základní analýzu sémantických vztahů, odvozovat jednoduché logické závěry a řešit různé jazykové úlohy (i když zatím nedostatečně ve srovnání s úrovní lidského porozumění). Tyto sofistikované výpočetní systémy vznikají prostřednictvím učení se statistických vztahů z rozsáhlých textových korpusů pomocí procesu samo-řízeného (**self-supervised**) a částečně řízeného (**semi-supervised**) trénování (Radford et al., 2019).

Takto natrénované modely fungují na základě výpočtu pravděpodobnosti pořadí prvků a předvídají každý další prvek na základě toho předchozího. S postupným zvětšováním kapacity těchto modelů a vlivem tréninku na opravdu velkých datasetech došli výzkumníci společnosti OpenAI k velmi zajímavému objevu. Zatímco tradičně bylo nutné pro specifické úlohy, jako je překlad nebo sumarizace textu, modely trénovat na specializovaných datasetech (například

na dvojicích přeložených vět), výzkumníci zjistili, že modely trénované na obecném datasetu **WebText**, obsahujícím miliardy tokenů z různorodých webových stránek, dokázaly tyto úlohy plnit i bez specializovaného tréninku. Tento jev, označovaný jako „**zero-shot learning**“, představuje schopnost modelu rozpoznat a aplikovat vzorce řešení úloh, se kterými se během tréninku nikdy nesetkal – model dokáže například identifikovat instrukce pro překlad v přirozeném jazyce a na jejich základě text přeložit, ačkoli nebyl specificky trénován na překladových párech (Radford et al., 2019).

Tato schopnost modelů se zvyšuje s jejich velikostí, přičemž výzkumy potvrzují přímou souvislost mezi počtem parametrů modelu a jeho výkonností. Podle studie Kaplana a kol. (2020) výkon modelu roste jako mocninná funkce počtu parametrů, výpočetního rozpočtu a velikosti datové sady. Kaplan vztah mezi velikostí modelu a jeho výkonem vyjádřil matematicky – výkon modelu roste s počtem parametrů umocněným na přibližně 0,07. V praxi to znamená, že pro dosažení i relativně malého zlepšení výkonu (o 10 %) je potřeba výrazně navýšit počet parametrů modelu (3,7krát).

Moderní jazykové modely je možné přizpůsobit konkrétním úkolům pomocí jemného doladování (**fine-tuning**) nebo je lze směřovat prostřednictvím promyšleného návrhu vstupních instrukcí (**prompt engineering**), který vede k získání požadovaných výsledků. V průběhu trénovacího procesu modely nabývají schopnosti předvídat jazykové vztahy – od syntaktických struktur a významových vazeb až po obecné souvislosti mezi pojmy, které jsou přirozeně obsaženy v textových korpusech lidského jazyka. Pro lepší porozumění tomu, jak jazykové modely fungují, se rozvíjí oblast nazvaná anglicky „**mechanistic interpretability**“. Tento obor se snaží „nahlédnout pod kapotu“ modelů a odhalit základní postupy a pravidla, podle kterých model zpracovává informace a vytváří své odpovědi.

Nejvýkonnější a nejpokročilejší současné LLM jsou postaveny na architektuře transformátorů – typu neuronové sítě speciálně navržené pro zpracovávání sekvenčních dat, jako je právě jazyk. Tato průlomová architektura, popsaná v akademickém článku *Attention is All You Need* (Vaswani et al., 2017), představila mechanismus pozornosti (**Attention**), který umožňuje modelům efektivně určovat relevanci jednotlivých částí vstupního textu a zpracovávat je. Transformátory tradičně obsahují dvě klíčové komponenty: kodér (**encoder**) a dekodér (**decoder**). Kodér umožňuje hlubší porozumění vstupnímu textu a jeho kontextu, zatímco dekodér využívá toto porozumění pro generování výstupu (Vaswani et al., 2017). Některé modely, například rodina modelů GPT, však využívají pouze dekodér. Tento přístup, ačkoliv neumožňuje tak komplexní počáteční porozumění textu jako kombinace kodéru

a dekodéru, je stále vysoce efektivní díky mechanismu pozornosti, který modelu umožňuje zohlednit předchozí kontext v rámci definovaného kontextového okna. Velikost tohoto okna určuje, kolik tokenů může model současně zpracovávat, což přímo ovlivňuje jeho schopnost porozumět dlouhým textům a udržet konzistenci v rozsáhlejších konverzacích.

Důvodem pro využívání architektury pouze s dekodérem je často hlavně pragmatická stránka – trénování i provoz rozsáhlých jazykových modelů klade mimořádné nároky na výpočetní infrastrukturu a vyžaduje značné množství specializovaného hardwaru, především výkonných grafických procesorů (Sajid, 2023). Zjednodušená architektura tak může významně snížit výpočetní náročnost provozu při zachování vysoké efektivity a kvality výstupů.

Zásadním požadavkem pro vývoj LLM je příprava vhodných trénovacích dat. Jelikož strojové učení pracuje výhradně s číselnými hodnotami, je nezbytné převést vstupní text na vhodnou číselnou reprezentaci tzv. „**tokenizací**“. Proces začíná vytvořením slovníku, v němž je každému slovu přiřazeno jedinečné číslo (**index**). Tento index je následně transformován na matematickou reprezentaci slova, takzvané vnoření (**embedding**) – sérii čísel, která zachycuje význam a vztahy daného slova v mnohorozměrném prostoru. Díky tomu mohou jazykové modely pracovat se slovy jako s číselnými vektory a ukládat si do paměti významové vztahy mezi slovy. Tokenizace probíhá skrze automatizované algoritmy – přičemž nejčastěji se používají dvě metody: kódování párů bajtů (**byte-pair encoding, BPE**) a algoritmus **WordPiece**.

S rostoucím pronikáním LLM a na nich založených nástrojů do různých oblastí lidského života se objevují nové problematické otázky ohledně jejich dopadu na společnost. Tyto modely mohou z datasetů (vytvořených lidmi) přejímat různé formy předpojatosti. To se týká především genderových stereotypů či politických postojů zakotvených v trénovacích materiálech. Významný problém se pak objevuje v oblasti bezpečnosti kvůli potenciálnímu zneužití těchto modelů pro šíření dezinformací, tvorbu škodlivého obsahu nebo pro tvorbu nástrojů kybernetických hrozeb.

Vývoj a využití velkých jazykových modelů s sebou přináší také závažné etické a právní otázky. Klíčovou problematikou je použití dat chráněných autorským právem pro trénink těchto modelů a schopnost modelů reprodukovat či odvozovat obsah z těchto děl. Neméně problematická je zde i ochrana osobních údajů, jelikož trénovací datasety mohou obsahovat citlivé osobní informace, které modely mohou neúmyslně zahrnout do svých výstupů nebo z nich nepřímým odvodit další citlivé informace o jednotlivcích.

Z hlediska dostupnosti jsou velké jazykové modely často nabízeny ve dvou režimech – základní verze bývá dostupná zdarma s určitými omezeními (například nižší výpočetní výkon, omezený počet dotazů nebo starší verze modelu), zatímco pokročilejší funkce a nejnovější modely jsou zpoplatněné. Při využívání dotazování skrze API rozhraní se platí za tzv. **tokeny**, tedy za základní jednotky textu vzniklé při tokenizaci – například slova, části slov nebo interpunkční znaménka. Cena se obvykle počítá za každý token jak ve vstupním dotazu, tak ve vygenerované odpovědi. Tento způsob účtování odráží skutečnost, že delší a složitější interakce s modelem vyžadují více výpočetních zdrojů.

1.1.1 NLP

Zpracování přirozeného jazyka (NLP) představuje významnou oblast informatiky a umělé inteligence. Tato disciplína se primárně věnuje vývoji postupů a algoritmů umožňujících počítačům pracovat s přirozeným lidským jazykem v různých formách. NLP úzce souvisí s vyhledáváním informací, reprezentací znalostí a výpočetní lingvistikou, přičemž využívá poznatky z mnoha vědních oborů, jako jsou lingvistika, psychologie, kognitivní vědy nebo matematika.

NLP zahrnuje široké spektrum úkolů od základního zpracování textu (například analýza větné skladby či rozpoznávání slovních druhů) až po komplexní úlohy vyžadující porozumění významu a kontextu, jako jsou strojový překlad, automatické zodpovídání otázek nebo generování textu. V současném technologickém prostředí NLP využívá především data z masivních databází – textových korpusů, které obsahují miliardy textů z internetu, knih, článků a dalších zdrojů. Tyto datové soubory jsou následně analyzovány pomocí různých přístupů, zahrnujících metody založené na pravidlech, statistické analýze nebo moderních neuronových sítích v rámci strojového učení a hlubokého učení.

První významné kroky v NLP proběhly v 50. letech 20. století, kdy Alan Turing publikoval přelomový článek *Computing Machinery and Intelligence* (1950). V něm představil koncept později nazvaný „Turingův test²“, který zahrnoval úlohu automatické interpretace a generování přirozeného jazyka jako součást hodnocení strojové inteligence. Období mezi 50. a 90. lety 20. století bylo charakteristické využíváním systémů založených

² Testovaný počítačový program komunikuje s hodnotitelem přes textové rozhraní, pokud hodnotitel nedokáže spolehlivě rozlišit, zda komunikuje s člověkem nebo strojem, program testem prošel.

na pravidlech a manipulaci se symboly, například systém SHRDLU³ (Winograd, 1971). Revoluci v oblasti NLP přinesla 90. léta 20. století s nástupem statistických metod a algoritmů strojového učení.

Tento posun byl umožněn výrazným nárůstem dostupného výpočetního výkonu a dostupností nových rozsáhlých textových korpusů. První významné úspěchy statistických metod se projeví zejména v oblasti strojového překladu, kde průkopnickou roli sehrál výzkumný tým IBM. Místo spoléhání na lingvistická pravidla systém analyzoval miliony párů vět v různých jazycích a vytvářel statistické modely pravděpodobnosti překladů. Tento přístup, známý jako IBM Models, položil základy moderního statistického strojového překladu (Brown et al., 1993).

Zásadní převrat v oblasti zpracování přirozeného jazyka (NLP) nastal kolem roku 2015 s masivním nástupem neuronových sítí. Toto období přineslo významný posun k využívání hlubokých neuronových sítí, které pomocí pokročilých technik, jako jsou tokenizace a embedding, dokáží zachytit komplexní sémantické vlastnosti slov a vztahy mezi nimi. Výsledkem byl dramatický nárůst výkonu v klíčových úlohách jako strojový překlad, rozpoznávání řeči či analýza sentimentu (Manning, 2015). Skutečně přelomovým okamžikem se pak stalo představení architektury transformerů (Vaswani et al., 2017), která byla popsána v předchozí kapitole.

Zpracování přirozeného jazyka zahrnuje několik vzájemně propojených rovin analýzy. Na té nejnižší systémy pracují s převodem řeči na text, členěním textu a rozbořením struktury slov. NLP dále analyzuje syntaktickou rovinu, kde systémy určují slovní druhy a odкрývají gramatické vztahy mezi slovy. Nejvyšší úroveň představuje sémantická analýza, která se zabývá významem – identifikuje entity v textu, určuje emocionální zabarvení a správně interpretuje význam slov v daném kontextu. Pokročilé systémy NLP dnes výrazně proměňují mnoho oblastí. Automatické shrnutí dokumentů usnadňuje práci s velkým objemem textu, strojový překlad rychle překonává jazykové bariéry a systémy pro porozumění a generování jazyka nacházejí uplatnění v akademické, komerční i veřejné sféře.

³ SHRDLU byl průkopnický program umělé inteligence vyvinutý Terryem Winogradem v roce 1970, který dokázal interpretovat příkazy v přirozeném jazyce pro manipulaci s geometrickými objekty v simulovaném prostředí a vést dialog o svých akcích (Winograd, 1971).

1.1.2 NLU

Porozumění přirozenému jazyku (NLU) je specializovaná podkategorie NLP, která se soustředí na schopnost počítačových systémů skutečně chápat význam a kontext textu, nikoliv jen jeho významovou strukturu. Vzhledem k mnohohrstevnatosti lidského jazyka a složitosti porozumění kontextu představuje NLU velkou výzvu v oblasti umělé inteligence.

Historický vývoj v oblasti NLU začal v roce 1964 programem **STUDENT**, který vytvořil Daniel Bobrow v rámci své doktorské práce na MIT (Massachusetts Institute of Technology). Tento průkopnický systém dokázal porozumět jednoduchým textovým vstupům pro řešení matematických slovních úloh (Bobrow, 1964). Později představil Joseph Weizenbaum program **ELIZA**⁴, který ve svém programu DOCTOR simuloval psychoterapeuta využívajícího tzv. Rogersovu metodu. I když program pracoval jen s relativně jednoduchými principy rozpoznávání klíčových slov a jejich transformací podle předem definovaných pravidel, jeho schopnost vést zdánlivě smysluplný dialog byla natolik přesvědčivá, že mnoho uživatelů věřilo v jeho skutečné porozumění (Weizenbaum, 1966). Tento fenomén, později známý jako „ELIZA efekt“, poprvé poukázal na tendenci lidí přisuzovat počítačovým systémům větší míru porozumění, než jakou ve skutečnosti mají (Natale, 2021).

Na konci 60. let 20. století představil Roger Schank na Stanfordské univerzitě důležitou inovaci v oblasti porozumění přirozenému jazyku – konceptuální dependenční teorii, revoluční přístup k tomu, jak počítače mohou „chápat“ význam lidské řeči. Zatímco předchozí systémy ELIZA a STUDENT pracovaly pouze s povrchovou strukturou textu nebo jednoduchým rozpoznáváním vzorců, Schankova teorie nabídla komplexnější přístup k zachycení významu. Vytvořila formální systém pro reprezentaci významu vět pomocí sítě vzájemně propojených konceptů a vztahů, podobný dnešním znalostním grafům. Například věta „John dal Marii knihu“ byla mohla být reprezentována jako strukturovaný graf zachycující akci předání, aktéry a jejich vzájemné vztahy (Schank & Tesler, 1969).

Na přelomu 60. a 70. let 20. století William A. Woods vyvinul nový systém nazvaný „rozšířená přechodová síť“ (**Augmented Transition Network, ATN**), který dokázal lépe analyzovat strukturu vět. Model prováděl sofistikovanější verzi gramatické analýzy – zatímco

⁴ ELIZA byl program vytvořený v roce 1966, který simuloval rogerovského psychoterapeuta. Fungoval na principu rozpoznávání klíčových slov v textu uživatele a jejich přeformulování do otázek nebo odpovědí podle předem definovaných vzorů. Například když uživatel napsal „Jsem smutný“, ELIZA odpověděla „Proč se cítíte smutný?“.

dřívější systémy analyzovaly věty spíše mechanicky, ATN dokázala lépe vyhodnotit, jak jsou části vět vzájemně propojené (Woods, 1970). Skutečný průlom v oblasti NLU (ač velmi malý krok pro NLP) byl zmíněný program SHRDLU, který pracoval ve virtuálním prostředí plném geometrických tvarů (podobných dětské stavebnici). To, co dělalo SHRDLU tak výjimečným, byla jeho schopnost nejen rozumět příkazům v běžné angličtině (například „zvedni červenou kostku“), ale také o svých akcích diskutovat. Když se ho výzkumníci zeptali „Proč jsi přesunul modrou krychli?“, dokázal objasnit své pohnutky (Winograd, 1971). Bylo to poprvé, kdy počítačový program prokázal skutečné „porozumění“ kontextu své činnosti, byť ve velmi omezeném prostředí.

Moderní systémy NLU kombinují pokročilé metody strojového učení s lingvistickými přístupy, aby dokázaly analyzovat a interpretovat významové nuance v textu. Například IBM Watson Natural Language Understanding aplikuje hluboké učení k získání významu a metadat z nestrukturovaných textových dat. Tato služba umožňuje klasifikovat kategorie, entity, klíčová slova, sentiment, emoce, vztahy a syntaxi v textu (IBM Watson Natural Language Understanding). Odborná komunita však diskutuje o míře skutečného „porozumění“, které tyto systémy vykazují. Někteří experti, včetně filozofa Johna Searla, argumentují, že modely ve skutečnosti nerozumí textu v pravém slova smyslu. Searle tento názor ilustruje prostřednictvím svého známého myšlenkového experimentu „čínský pokoj“, ve kterém člověk bez znalosti čínštiny odpovídá na otázky v čínštině pouze podle manuálu. I když se zdá, že rozumí, doopravdy jen mechanicky manipuluje symboly bez pochopení jejich významu – podobně jako počítačové systémy pracující s přirozeným jazykem (Searle, 1980).

Architektura NLU je postavena na několika vzájemně propojených komponentách. Základním prvkem je lexikon jazyka, obsáhlý seznam slov doplněný o informace o vztazích mezi těmito slovy. Vytvoření takového lexikonu představuje náročný dlouhodobý proces – například známý WordNet vyžadoval 6 let vývoje do zveřejnění první verze (Miller, 1995) a jeho rozšiřování pokračuje dodnes (Open English Wordnet 2024, 2024). Druhým klíčovým prvkem je syntaktický analyzátor (takzvaný **parser**), který spolu s gramatickými pravidly analyzuje strukturu vět a převádí je do počítačově zpracovatelné formy. Parser identifikuje jednotlivé větné členy a jejich vzájemné vztahy, například ve větě „Kočka honí myš“ určí „kočka“ jako podmět a „honí“ jako přísudek. Součástí syntaktické analýzy je také označování slovních druhů, které je zásadní pro správnou interpretaci významu – například rozlišení slova „domů“ jako podstatného jména ve větě „V naší ulici je hodně domů“ od jeho příslovečného využití ve větě „Půjdu domů“. Třetí složkou NLU je sémantická teorie určující pravidla pro interpretaci významu analyzovaného textu. Sémantická analýza zahrnuje označování

sémantických rolí, které určují funkce jednotlivých částí věty ve vztahu k hlavnímu slovesu – například ve větě „Jana dala knihu Samovi“ identifikuje „Jana“ jako dárce, „knihu“ jako darovaný předmět a „Sama“ jako příjemce (Natural Language Understanding (NLU), 2024).

S nástupem velkých jazykových modelů se oblast NLU rozvíjí a otevírají se nové možnosti jejího využití. Zásadní otázkou však zůstává skutečná podstata jejich „porozumění“ zpracovávanému textu.

1.2 Benchmarky v oblasti porozumění přirozenému jazyku (NLU)

Benchmarking v oblasti výpočetní techniky představuje „činnost nebo proces měření výkonu hardwaru nebo softwaru pomocí standardizovaného nebo speciálně vybraného softwaru, datasetů a dalších“ (Benchmarking, n., sense 2, 2023). V kontextu výpočetní techniky se tedy jedná o soubor programů a operací, které systematicky vyhodnocují výkonnost zkoumaného modelu pomocí předem definovaných metrik. Tyto nástroje typicky zahrnují datové sady, specifikace úloh a přesně stanovená kritéria hodnocení, přičemž by měly splňovat sedm klíčových vlastností: relevanci, reprezentativnost, spravedlivost, opakovatelnost, ekonomickou efektivitu, škálovatelnost a transparentnost (Dai & Berleant, 2019).

V oblasti porozumění přirozenému jazyku (NLU) se tedy jedná o soubor nástrojů pro hodnocení a porovnávání výkonnosti modelů generativní textové umělé inteligence pomocí rozličných jazykových úloh z oblasti pravopisu, gramatiky, stylistiky a dalších lingvistických oborů. Sady pro testování obvykle obsahují standardizované metriky hodnocení, které zajišťují spolehlivé srovnání. Jedním příkladem benchmarku je datová sada Stanford Question Answering Dataset (SQuAD), která se zaměřuje na porozumění čtenému textu a je klíčová pro rozvoj schopností systémů při zodpovídání otázek (Rajpurkar et al., 2016). Další benchmarky jako GLUE (General Language Understanding Evaluation) ho později rozšířily o další množství úloh, které testují širší schopnosti modelů v porozumění textu (Wang et al., 2018).

Úkolem benchmarků je nejen měřit současný vývoj, ale také zjišťovat a definovat limitace testovaných modelů a umožnit další vylepšení jejich dovedností v problematických oblastech. Například GLUE slouží k testování porozumění textu, avšak nové jazykové modely rychle překonaly lidský výkon v úlohách generování textu, což vedlo k vývoji vylepšeného benchmarku SuperGLUE. Tyto benchmarky poskytují platformu pro hodnocení a posouvání schopností jazykových modelů. SuperGLUE zavedl do testování náročnější úlohy, jako je např. Winograd Schema Challenge (Wang et al., 2019). Tato úloha testuje porozumění jazyku

pomocí dvojice vět či skupin vět, které se liší pouze jedním nebo dvěma slovy a které jsou z definice dvojznačné.

Každá z těchto vět obsahuje dvojznačnost, kterou lze interpretovat různými způsoby. Pro správné pochopení zamýšleného významu je potřeba buď:

- 1) znát širší sociálně-kulturní kontext (například místní zvyky, historické události nebo společenské normy), nebo
- 2) hlouběji analyzovat vztahy mezi slovy ve větě, zejména jejich vzájemné odkazy a souvislosti (tzv. **koreference**⁵).

Jako příklad se uvádí dvojice vět „Radní odmítli vydat demonstrantům povolení. Oni se obávali / dopouštěli násilí“. (Levesque, 2011) V případě prvním (obávali se) se zájmeno „oni“ vztahuje k radním, zatímco v případě druhém (dopouštěli se) k demonstrantům. Model tedy musí na základě logiky a kontextu ve větách porozumět, na kterou jmennou frázi zvolené zájmeno odkazuje. Podobné složitější testy pomáhají lépe pochopit a ohodnotit zvyšující se složitost systémů umělé inteligence a jejich využití a zajišťují, že modely jsou testovány na úlohách ověřujících hlubší porozumění kontextu.

Moderní benchmarky už tedy neporovnávají dovednosti v oblasti porozumění jazyka pouze mezi jazykovými modely navzájem, ale srovnávají ho s úrovní lidského výkonu. Například SuperGLUE porovnává výkonnost umělé inteligence s porozuměním textu na lidské úrovni (**human-baseline**⁶) a poukazuje na oblasti, ve kterých modely stále zaostávají, jako je racionální uvažování nebo pochopení vztahů nejednoznačných zájmen (Wang et al., 2019).

1.2.1 Benchmark SuperGLUE

SuperGLUE je pokročilý vyhodnocovací rámec pro hodnocení výkonu velkých i malých jazykových modelů, který byl vyvinut jako nástupce staršího benchmarku GLUE (Wang et al., 2019). Vznikl jako reakce na výrazný pokrok nových modelů v oblasti porozumění a zpracování přirozeného jazyka. Výkon nejmodernějších (**state-of-the-art**) modelů totiž při testování na původním benchmarku GLUE překonal human-baseline o 1,3 %,

⁵ Koreference znamená, že různá slova nebo výrazy v textu odkazují ke stejné skutečnosti či objektu, což může bez správného pochopení těchto vazeb vést k nejasnostem v interpretaci.

⁶ Úroveň výkonu dosažená lidskými účastníky při řešení úloh, která se používá jako srovnávací metrika pro systémy umělé inteligence.

konkrétně dosáhl skóre 88,4 % (Wang et al., 2019) ve srovnání s lidskými subjekty testovaných na stejných úlohách s přesností 87,1 % (Nangia and Bowman, 2019).

Hlavním cílem tohoto nového benchmarku je vytvořit jednotnou metriku pro hodnocení jazykových modelů. Metrika měří jejich schopnost řešit komplexní úlohy napříč různými jazykovými oblastmi, přičemž hodnotí nejen dodržování gramatických a stylistických pravidel, ale i používání racionálního rozumu při řešení zadaných úkolů. Při definování úloh pro tento benchmark byly stanoveny několik klíčových požadavků:

- musí testovat schopnosti porozumění jazyku a uvažování o textech v anglickém jazyce,
- jejich obtížnost musí přesahovat současné „state-of-the-art“ modely,
- musí být řešitelné pro vysokoškolsky vzdělané rodilé mluvčí anglického jazyka,
- je třeba umožnit, aby byly tyto úlohy možné hodnotit automaticky a tak, aby byl formát výsledků modelů a lidských subjektů porovnatelný
- datasey pro testování musí být volně a veřejně dostupné, a to včetně trénovacích dat,
- vstupní a výstupní formát úloh musí být jednoduše zpracovaný (Wang et al., 2019).

Podle těchto podmínek vybrali autoři SuperGLUE osm úloh, které jsou v textu níže detailně popsány a definovány. Informace o úlohách pochází z akademické práce *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems* (Wang et al., 2019) a jsou místy doplněny o detaily z původních akademických článků jejich autorů – v takovém případě jsou uvedeny u úlohy zdroje.

BoolQ – binární testovací úloha typu otázka-odpověď (QA – question-answer). Dataset otázek byl vytvořen ze skutečných vyhledávacích dotazů ve vyhledávači Google a jejich párováním s relevantními články na Wikipedii. Modelu je nejprve poskytnut krátký textový úryvek následovaný související jednoduchou otázkou, na kterou lze odpovědět ano/ne. Příklad:

Premisa: Dvacetieurovka – „Dosud existovala pouze jedna kompletní série eurových bankovek, ale nyní bude vydána nová série podobná té současné. Evropská centrální banka včas oznámí, kdy bankovky z první série ztratí status zákonného platidla.“

Otázka: Je první série dvacetieurovek stále zákonným platidlem?

Odpověď: Ano

CB (CommitmentBank) – Dataset představuje korpus krátkých souvětí, v nichž je vždy použita alespoň jedna vedlejší věta. Úloha je zaměřena na evaluaci stupně tzv. „**epistemické modality**“, takže vyhodnocuje, do jaké míry je mluvčí přesvědčen o pravdivosti uvedeného tvrzení v kontextu uvedeného souvětí – vedlejší věta vždy obsahuje informaci nebo náznak, ze kterého je možné logicky vycházet. Model nebo lidský subjekt poté zvolí právě jednu ze tří odpovědí – „neutrální“ (= nelze určit), „kontradikce“ (mluvčí zpochybňuje hypotézu) nebo „entailment“ (mluvčí potvrzuje hypotézu). Příklad:

Premisa: „Polly musela rychle přemýšlet. Byli ještě dost blízko břehu, aby ji mohla vrátit policii, pokud přizná, že není zkušená oceánská námořnice.“

Hypotéza: Polly není zkušená oceánská námořnice.

Otázka: Je mluvčí přesvědčen, že Polly není zkušená oceánská námořnice?

Odpověď: Entailment = míra přesvědčení mluvčího je vysoká.

COPA (Choice of Plausible Alternatives, volba možných alternativ, Roemmele et al., 2011) – tato úloha je zaměřená na vyhodnocování kauzálních vztahů v představených textech. Dataset pro COPA obsahuje premisu, kterou následuje specifikace požadovaného kauzálního směru (příčina nebo následek) a dvě možné odpovědi. Úkolem testovaného subjektu je určit vhodnější nebo pravděpodobnější možnost z nabízených odpovědí, přičemž správná odpověď by měla vycházet z kontextu vět, logických návazností nebo na základě obecných znalostí o světě. Příklad:

Premisa: Muž zapnul kohoutek s vodou.

Otázka: Efekt (následek)

Možnost A: Umyvadlo se naplnilo vodou.

Možnost B: Voda se vylila z výlevky.

Odpověď: A

MultiRC (Multi-Sentence Reading Comprehension, porozumění textu složeného z více vět, Khashabi et al., 2018) – MultiRC představuje velmi komplexní úlohu typu otázka-odpověď a testuje strojové porozumění textu. Dataset tvoří dlouhé pasáže obsahující několik vět, ke kterým je autory přiřazena jedna nebo více otázek, které vyhodnocují kontextuální

pochopení daného dlouhého textu. Ke každé otázce je testovacímu subjektu nabídnuto větší množství odpovědí – některé jsou pouze ano/ne, ale může jít také o komplikovanější vyústění z kontextu ukázky. Správných odpovědí může být více. Příklad:

Text: „Co způsobuje změnu pohybu? Působení síly. Kdykoli objekt změní pohyb, působí na něj síla. Jakými způsoby k tomu může dojít? Síla může způsobit, že se objekt v klidu začne pohybovat. Síly mohou způsobit, že se objekty zrychlí nebo zpomalí. Síly mohou způsobit, že se pohybující se objekt zastaví. Síly mohou také způsobit změnu směru pohybu. Stručně řečeno, síly způsobují změny pohybu. Pohybující se objekt může změnit svou rychlost, směr nebo obojí. Víme, že změny pohybu vyžadují působení síly. Víme, že velikost síly určuje změnu pohybu. Jak moc se změní pohyb objektu při působení síly, závisí na dvou věcech. Záleží na síle. Závisí také na hmotnosti objektu. Zamyslete se nad některými jednoduchými úkoly, které můžete pravidelně provádět. Můžete zvednout baseballový míček. To vyžaduje jen velmi malou sílu.“

Otázka: Ovlivní hmotnost baseballového míčku velikost síly, kterou použijete pro jeho zdvihnutí?“

Odpovědi:

- Ano (true)
- Ne (false)
- Čím menší hmotnost, tím nižší je potřebná síla. (true)
- Záleží na tvaru baseballového míčku. (false)

ReCoRD (Reading Comprehension with Commonsense Reasoning Dataset, čtení s porozuměním za pomoci logického uvažování, Zhang et al., 2018) je další komplexní úloha pro evaluaci strojového porozumění textu. Dataset obsahuje texty založené na reálných zpravodajských článcích. Každá testová otázka obsahuje delší pasáž a související otázku s možnou odpovědí, která je součástí textu. Příklad:

Premisa: Tracy Morgan se na jevišti neobjevil od ničivé nehody v New Jersey, která ho loni v létě málem stála život, ale to se změní letos na podzim, kdy se vrátí jako moderátor pořadu Saturday Night Live. Stanice NBC v pondělí na Twitteru oznámila, že Morgan bude moderovat třetí díl nové série vysílaný 17. října. Příklad:

Otázka: „10. října bude v SNL debutovat uznávaná komička Amy Schumerová, o týden později ji bude následovat _____.“

Odpověď: Tracy Morgan

RTE (Recognizing Textual Entailment, rozpoznávání textového usuzování) – RTE představuje další binární klasifikační úlohu, kterou SuperGLUE převzalo z původního benchmarku GLUE. Dataset pro SuperGLUE vznikl kombinací 4 různých zdrojových korpusů zaměřených na rozpoznávání tzv. textové interference – cílem je správně označit, jestli je vztah mezi danou premisou a hypotézou možné označit jako „**entailment**“ (usuzování), neboli zda hypotéza logicky vyplývá z obsahu premisy. Příklad:

Premisa: „Přesto nyní zjišťujeme, že antibiotika ztrácejí svou účinnost proti nemocem. Bakterie způsobující nemoci mutují rychleji, než jsme schopni vymyslet nová antibiotika, která by s novými variantami bojovala.“

Hypotéza: Bakterie vítězí ve válce s antibiotiky.

Odpověď: entailment

WiC (Word-in-Context, slova v kontextu, Pilehvar & Camacho-Collados, 2019) testuje schopnost testovaného subjektu rozpoznat lexikální a sémantický význam slov v kontextu dané ukázky (premisy). Dataset obsahuje páry vět či frází obsahujících polysémnní výraz, tedy výraz, který nese dva i více možných slovních významů. Testovaný subjekt se rozhoduje, zda je cílový výraz použit v obou kontextech ve stejném významu. Příklad:

Premisa 1: „Zámek byl obklopen krásným parkem.“

Premisa 2: „Klíčem odemkl zámek na dveřích.“

Otázka: Je slovo „zámek“ použito v obou větách ve stejném významu?

Odpověď: false

WSC (Winograd Schema Challenge, testování podle Winogradova schématu) představuje úkol zaměřený na rezoluci koreference zájmen v textu. Úloha vyžaduje kombinaci lingvistických znalostí, obecného porozumění světu a logického uvažování. Dataset obsahuje textové pasáže se zájmeny odkazujícími na aktéry či jiné entity v textu, přičemž úkolem testovacího subjektu je správně určit tzv. „**antecedent**“ (Karlík et al., 2017), což je element v textu, který poskytuje danému zájmenu význam neboli referenta daného zájmena."

Premisa: Radní neumožnili demonstrantům získat povolení. Oni se obávali násilí.

Otázka: Odkazuje zájmeno „oni“ na slovo „demonstrantům“?

Odpověď: false

Benchmark poskytuje komplexní veřejný evaluační framework a jeho autoři zveřejňují výsledky otestovaných modelů v žebříčku na oficiální stránce benchmarku. Výhodou tohoto benchmarku je jednotná metrika výkonu a k tomuto účelu vytvořený manuál pro vyhodnocování. Oproti staršímu rámci GLUE přináší několik vylepšení:

- náročnější úlohy,
- diverzifikované formáty úloh (např. koreference a QA),
- porovnání s lidským výkonem na stejných otázkách,
- kód pro testování je dostupný na nástrojích PyTorch (Paszke et al., 2017) a AllenNLP (Gardner et al., 2018).

SuperGLUE byl navržen jako pokročilý vyhodnocovací nástroj s cílem akcelarovat vývoj v klíčových doménách velkých jazykových modelů – jmenovitě v oblastech učení s omezeným počtem vzorků (**sample-efficient learning**), transferového učení (**transfer learning**), víceúlohového učení (**multitask learning**) a učení bez učitele či samořízené učení (**unsupervised/self-supervised learning**). Různorodé datasety a testovací úlohy rámce zahrnují široké spektrum textů z různých zdrojů lidské činnosti – novinářské články, umělecká literatura, blogové články a další internetové formáty nebo encyklopedické a vědecké texty. Důvodem pro takto široký záběr je snaha o co nejlepší pokrytí toho, jak je jazyk jeho uživateli využíván. Významná výhoda SuperGLUE je jasně definovaný způsob vyhodnocení výkonu v každé úloze pomocí vybraných metrik optimalizovaných pro danou oblast a tím možnost výsledky každého modelu snadno porovnat.

Pro potřeby této práce byl benchmark SuperGLUE zvolen kvůli adekvátní komplexitě jeho úloh, snaha o vyhodnocení velmi pokročilých lingvistických dovedností modelů, standardizovaný a systematický rámec pro evaluaci výsledků, koherentní a programatická struktura testování a také běžná dostupnost testovacích datasetů pro toto použití.

Výběr těchto charakteristik proběhl skrze kvalitativní porovnání s následujícími benchmarky, které jsou k testování LLM také běžně používány:

GLUE (General Language Understanding Evaluation) – Jakožto zmíněný předchůdce SuperGLUE, je benchmark GLUE již zastaralý a neposkytuje adekvátní výzvu

pro vyhodnocení moderních velkých jazykových modelů; úkoly v tomto rámci byly moderními modely překonané.

MMLU (Massive Multitask Language Understanding) – Tento komplexní benchmark představuje vyhodnocovací rámec pokrývající 57 odlišných tematických oblastí včetně přírodních věd, humanitních oborů, matematiky a dalších odborných domén. MMLU testuje hluboké porozumění kontextu, logické uvažování i znalost obecných informací o světě (Hendrycks et al., 2021). Zatímco ale MMLU zkoumá především rozsah znalostí modelů, SuperGLUE ověřuje jazykové schopnosti více do hloubky a vyžaduje i rozpoznání jemných významových nuancí v textu.

HELM (Holistic Evaluation of Language Models) – Jak název napovídá, HELM přistupuje k hodnocení modelů holisticky, tedy z pohledu celku. Kromě evaluace jazykového výkonu dbá na vysokou míru transparentnosti datasetů, testování i vyhodnocování, ale také hodnotí modely z pohledu etického; zjišťuje, zda model produkuje škodlivý obsah, jak moc je ovlivněn sociálními předsudky a stereotypy nebo jak férový je vůči menšinám ve společnosti. Pokrývá ještě více znalostních kategorií než MMLU (až 200 kategorií). Kvůli tomuto spíše kvalitativnímu přístupu k hodnocení tedy neumožňuje zobecnění výsledků do unifikovaných metrik a komparace výsledků testování je tedy u benchmarku HELM komplikovaná (Liang et al., 2023).

BIG-bench (Beyond the Imitation Game Benchmark) – Rozsáhlý benchmark vytvořený dobrovolníky na internetu. BIG-bench obsahuje obtížné úlohy navržené k testování hranic jazykových modelů. Vedle jazykových úkolů zahrnuje kreativní, abstraktní a víceúrovňové výzvy. Stejně jako MMLU je proti SuperGLUE rámec BIG-bench příliš různorodý a velmi málo standardizovaný (BIG-authors, 2023).

LEXTREME – Tento benchmark je zaměřený na testování mnohojazyčných dovedností testovaných jazykových modelů v oblasti právních textů. Framework je sestaven z 11 datasetů napříč 24 jazyky z 8 jazykových skupin. Stejně jako SuperGLUE umožňuje agregovat výsledky do ucelených hodnotících metrik (Niklaus et al., 2023). Pro využití testování jazykové vybavenosti je ale ve srovnání se SuperGLUE nepoužitelný z důvodu cíleného zaměření na právní texty.

XTREME (Cross-lingual TRansfer Evaluation of Multilingual Encoders) – Jedná se o další velmi komplexní benchmark, který je stejně jako LEXTREME navržený za účelem testování multilingvální schopnosti LLM. Srovnávací testy zahrnují 9 různých úloh, které ověřují, jak dokáže model přenášet informaci mezi jazyky. Testuje kolem 40 různých jazyků

napříč 9 jazykovými skupinami. Obsahuje úlohy klasifikace, strukturované predikce, QA (otázka-odpověď) nebo získávání informací. Na rozdíl od SuperGLUE je zde hlavním předmětem testování schopnost přenášení informací skrze překlady do dalších jazyků, nejčastěji se zdrojem z angličtiny (Hu et al., 2020).

SuperGLUE nabízí významné výhody ve srovnání s těmito vybranými benchmarky. Jeho jasnou výhodou je zejména specializace přímo na obor lingvistiky a porozumění jazyku, jelikož ostatní modely jsou svým zaměřením buďto mnohem rozsáhlejší (MMLU) nebo naopak příliš zacílené na velmi úzký obor (LEXTREME). SuperGLUE byl zvolen jako primární nástroj pro evaluaci modelů v této práci díky optimální diverzitě úloh a adekvátní jazykové náročnosti.

1.2.2 Existující české benchmarky

Testování jazykových modelů v českém prostředí zaznamenalo v roce 2024 významný rozvoj prostřednictvím tří důležitých iniciativ. Nejvýznamnějším příspěvkem je BenCzechMark, první komplexní hodnotící nástroj pro české jazykové modely, který přináší 50 typů úloh v 9 kategoriích a zavádí inovativní systém „Model Duels“ pro spravedlivé porovnávání různých typů úloh. Tento benchmark se odlišuje od ostatních především vysokým podílem nativně českého obsahu (90 %) a širokou škálou testovaných schopností, od porozumění textu až po kulturně specifické znalosti.

Další dva příspěvky reprezentují specializovanější přístupy k testování. Seznam.cz vyvinul benchmark zaměřený především na vyhodnocování sémantického porozumění a vektorových reprezentací textu, což přímo odpovídá potřebám vyhledávací služby. Využívá při tom kombinaci pěti existujících českých datasetů pro různé aspekty práce s textem. Advokátní kancelář HAVEL & PARTNERS pak přišla s originálním přístupem k testování právního porozumění, kdy použila 1 840 otázek z advokátních zkoušek České advokátní komory.

Tyto tři benchmarky společně poskytují komplexní pohled na schopnosti jazykových modelů pracovat s českým jazykem, přičemž každý se zaměřuje na jiný aspekt jazykového porozumění a použití.

1.2.2.1 BenCzechMark

V době zadání této diplomové práce neexistoval v oblasti českého jazyka komplexní celistvý benchmark srovnatelný s anglickým SuperGLUE. Dne 1. října 2024, kdy byla již zahájena práce na testovacích datasetech pro tuto práci, byl zveřejněn významný nový příspěvek v oblasti evaluace jazykových modelů pro český jazyk – benchmark BenCzechMark

od Martina Fajčíka et al (2024). BenCzechMark je první komplexní sada pro hodnocení velkých jazykových modelů pro češtinu. Ve srovnání se SuperGLUE, který je primárně zaměřený na testování porozumění jazyku, se autoři tohoto benchmarku snaží testovat také úkoly praktického (a generativního) typu, jako jsou například:

- argumentace a výkon v náročných úlohách v českém jazyce,
- generování a kontrola gramatické i sémantické správnosti v češtině,
- získávání a ukládání informací o české kultuře a relevantními znalostmi,
- odhadování neboli predikování českého textu (Fajčík et al., 2024).

Za tímto účelem vytvořili autoři 50 typů úkolů rozdělených v 9 hlavních kategoriích, přičemž významnou zmiňovanou vlastností těchto úkolů je vysoký podíl (90 %) nativně českého (nepřeloženého) obsahu. Pouze zbývajících 10 % tvoří strojově přeložený obsah, primárně pomocí systému CUBBITT LINDAT Translation, s výjimkou datasetu CsFever, kde byl využit nástroj DeepL. BenCzechMark poskytuje kompletní žebříček s více než 25 open-source modely různých velikostí. Volba testovat modely s otevřenými daty automaticky vylučuje většinu velkých jazykových modelů jako jsou Gemini, GPT-4 (a GPT-4o) a Claude – z modelů vybraných pro testování v této práci jsou ve veřejném žebříčku benchmarku zveřejněny pouze výsledky pro model Llama 3.2-3b. Stejně jako u SuperGLUE poskytují autoři také otevřená data pro samotné testování – datasety, úlohy i unifikovaný systém vyhodnocení.

Hlavní kategorie testování zahrnují porozumění jazyku a textu, testování faktických a matematických znalostí, generaci textu, hodnocení interference, pojmenování entit či hledání dokumentů. Datasety čerpají autoři nejčastěji ze zdrojů na internetu (Wikipedia, Umímeto) nebo z veřejných institucí (Matematický klokan, CERMAT, Český Národní Korpus), zatímco v případě přeložených datasetů jde o zdroje ze zahraničních výzkumů.

BenCzechMark zavedl nový hodnotící systém „Model Duels“, který efektivně řeší porovnávání výkonnosti napříč různorodými úlohami. Systém dokáže spravedlivě srovnávat výsledky mezi zcela odlišnými formáty úloh, jako jsou binární úkoly typu ano/ne a otevřené krátké textové odpovědi. Využívá čtyři základní typy metrik:

- **přesnost** (Accuracy) – podíl správných odpovědí z celkového počtu odpovědí. Například pokud model správně odpoví na 90 otázek ze 100, jeho přesnost je 90 %,
- **přesnou shodu** (Exact Match) – přísnější metrika, která vyžaduje, aby odpověď modelu byla naprosto identická s referenční odpovědí – např. včetně interpunkce a formátování,

- **AUROC** – (Area Under the Receiver Operating Characteristic Curve) metrika určující, jak dobře si model vede v kategorizaci nebo rozdělení témat či slov do různých skupin či kategorií,
- **Perplexita** na úrovni slov (Word-level Perplexity) – metrika pro sledování „překvapenosti“ modelu v reakci na následující slova v textu. Nižší perplexita znamená, že model lépe předpovídá následující slova, což naznačuje lepší porozumění jazyku a kontextu.

Pro každou úlohu je vytvořeno minimálně 5 různých promptů. Hodnocení zahrnuje jak nejlepší výkon, tak variabilitu napříč prompty. Využívá statisticky významné testy pro porovnávání modelů. Výsledné skóre je počítáno jako průměrná úspěšnost v „duelech“ s ostatními modely (Fajčík et al., 2024).

Ve srovnání s ostatními (anglickými) rámci představuje BenCzechMark benchmark srovnatelný typově i rozsahem s rámcem MMLU nebo BIG-bench – přístupem, který kombinuje striktní metodologii s komplexním pokrytím různých aspektů jazykového porozumění a generování. Jeho význam spočívá především v možnosti detailního testování schopností modelů pracovat s českým jazykem, jeho gramatikou, sémantikou i kulturním kontextem.

Existence BenCzechMarku nemá vliv na validitu této práce. Přestože tento rámec byl zveřejněn před datem testování pro tuto práci (25. 11. 2024), nebyl dostupný v době zadání práce ani v době zahájení prvotní literární rešerše pro tuto práci. Podobně jako MMLU a BIG-bench nesplňuje stanovené parametry pro výběr testovacího nástroje v rámci této studie. Na rozdíl od SuperGLUE, který se zaměřuje výhradně na jazykové porozumění, BenCzechMark nabízí komplexnější hodnocení. Jeho záběr kromě jazykových dovedností zohledňuje kulturní specifika českého prostředí, vědomosti o světě, matematické znalosti a další.

1.2.2.2 Benchmark od Seznam.cz

V oblasti vyhodnocování jazykových modelů pro český jazyk neexistoval před zveřejněním BenCzechMark žádný další vysloveně publikovaný oficiální benchmark. V únoru 2024 ale česká technologická společnost Seznam představila na 38. ročníku konference AAAI Conference on Artificial Intelligence svůj odborný článek *Some Like It Small: Czech Semantic Embedding Models for Industry Applications* (Bednář et al., 2024), ve kterém popsala svůj experiment, při kterém natrénovali několik malých jazykových modelů

založených na jazykovém modelu BERT. Pro upřesnění, BERT patří do rodiny „encoder-only“ modelů, což znamená, že jeho architektura je navržena tak, aby kódovala vstupní text do smysluplných vložených (embedded) textů. BERT tak neobsahuje dekodér, který je klíčovou součástí generativních textových modelů – např. GPT modely používají pouze transformer dekodér (Yang et al., 2024).

Seznam v tomto článku své nové modely pro český jazyk testuje a pro tuto evaluaci čerpal z několika vybraných, již existujících, datasetů určených pro testování různých oblastí českého jazyka. Tato kombinace úloh by mohla být považována za benchmark a sám Seznam toto zmiňuje ve svém blogovém článku *Seznam představuje nové jazykové modely pro práci s významem textů* (Barančíková & Náplava, 2024) – oficiálně ale benchmark zveřejněný a dostupný zatím není. Úlohy a datasety využitě v jejich odborném článku jsou zde přiblíženy:

1. **Costra** – Dataset COSTRA 1.1 z roku 2020 je datová sada pro testování geometrických vlastností prostorů v embeddingu vět. Zaměřuje se zejména na zkoumání toho, jak dobře embeddingy zachycují komplexní jevy ve větách, jako jsou parafráze, časování nebo zobecnění. Dataset je přímým rozšířením datasetu Costra 1.0 [4 262 unikátních českých vět o průměrné délce 10 slov (Barančíková & Bojar, 2020a)], který byl rozšířen o další věty a porovnání vět (Barančíková & Bojar, 2020b).
2. **SVOB-IMG, SVOB-HL a STS-CNA** – Pro potřeby testování sémantické textové podobnosti Seznam využil tři datasety – dva z existujícího datasetu **SentEval** z roku 2015 (SVOB-IMG obsahuje popisky obrázků a SVOB-HL obsahuje nadpisy článků) a část datasetu si vypůjčil z českého datasetu pro sémantickou textovou podobnost z roku 2021, který vzniknul ve spolupráci s Českou tiskovou kanceláří a obsahuje konkrétně 1100 větných párů z českých novinářských článků (Sido et al., 2021).
3. **CTDC** (Czech Text Document Corpus) – Korpus textových dokumentů pro automatickou klasifikaci dokumentů v českém jazyce si vypůjčili autoři práce od Pavla Krále a Ladislava Lence (2017) pro testování multi-label klasifikace textových dokumentů. Korpus obsahuje přes 12 000 novinářských článků rozdělených do 37 kategorií.
4. **CFD** (Czech Facebook Dataset) – Dataset sestavený v roce 2013 byl využit pro vyhodnocení sentimentu textů. Obsahuje 2 587 pozitivních, 5 174 neutrálních a 1 991 negativních příspěvků (Çano & Bojar, 2019).
5. **DaReCzech** – Dataset sestavený Matějem Kociánem a spol. (2021) byl využit pro ohodnocení relevance modelů. Skládá se skládá z 1,6 milionu dvojic českých výrazů

v párech „dotaz + dokument“, které jsou rozdělené do trénovací, vývojové a testovací množiny, z nichž každá je označena skórem relevance.

Zaměření benchmarku Seznamu na embedding a porozumění textu odpovídá plánovanému využití modelů pro indexování webového obsahu v rámci vyhledávacího nástroje (Barančíková & Náplava, 2024). Vysoká výkonnost v těchto oblastech je proto pro společnost klíčová při vývoji jejich vyhledávacích nástrojů. Pro účely této práce však není Seznam benchmark ideální volbou, přestože jde o kvalitně zpracovaný soubor testovacích datasetů. Zatímco Seznam se soustředí na specifické aspekty jazykového porozumění relevantní pro vyhledávání, tato práce si klade za cíl hodnotit širší spektrum schopností jazykového porozumění u vybraných modelů. Je tedy třeba využít komplexnější nástroj, který dokáže evaluovat různorodé aspekty jazykové kompetence, nikoliv pouze ty, které jsou spojené s vyhledáváním a kategorizací obsahu.

1.2.2.3 Advokátní zkoušky od HAVEL & PARTNERS

Další testování generativních jazykových modelů v českém prostředí proběhlo v právním kontextu: Advokátní kancelář HAVEL & PARTNERS otestovala schopnosti LLM v kontextu českého práva. Použili sadu 1 840 otázek z advokátních zkoušek České advokátní komory k prověření schopnosti modelů porozumět a správně interpretovat české právní koncepty (Kovář & Nečas, 2024).

Testování probíhalo v pěti kolech, přičemž každé kolo obsahovalo 100 otázek. Studie porovnávala osm různých modelů: Claude-3-Opus, Claude-3-Sonnet, Gemini-1.0-Pro, GPT-4-Turbo (GPT-4o), GPT-3.5-Turbo, Mistral-Large, Mixtral-8x7B a specializovanou právní aplikaci WAIR vyvinutou samotnou kanceláří.

Nejlépe si vedla specializovaná aplikace WAIR, která jako jediná překonala požadovanou 85% hranici úspěšnosti ve všech testovacích kolech. To potvrzuje výhodu kombinace jazykového modelu (GPT-4o) s cíleným laděním na základě relevantních právních zdrojů. Mezi samotnými jazykovými modely dosáhl nejlepších výsledků Claude-3-Opus, těsně následovaný GPT-4-Turbo. S větším odstupem se umístily modely Gemini-1.0-Pro, Claude-3-Sonnet a Mistral-Large. Nejhorší výsledky měly modely GPT-3.5-Turbo a Mixtral-8x7B, které dosáhly pouze kolem 50% úspěšnosti.

Zajímavým zjištěním bylo, že různé modely vykazovaly rozdílnou úspěšnost v jednotlivých právních oblastech. Například většina modelů si vedla nejlépe v ústavním

a správním právem, zatímco trestní právo bylo obecně nejproblematictější oblastí. Výjimku tvořil model Gemini-1.0-Pro, pro který bylo trestní právo paradoxně nejsilnější disciplínou.

2 Metodologie práce

V této sekci popisují podrobný postup pro testování v praktické části. Součástí metodologie je výběr testované úlohy, příprava testovacích datasetů, skriptů a promptů, podrobný návrh postupu při testování i při vyhodnocování.

2.1 Analýza a výběr úloh v rámci benchmarku

Pro potřeby této práce byla z úloh představených v kapitole o benchmarku SuperGLUE zvolena úloha Winograd Schema Challenge. Tato sekce se věnuje jejímu hlubšímu představení a argumentuje, proč byla zvolena pro potřeby testování v této práci.

2.1.1 Winograd Schema Challenge

Winograd Schema Challenge (dále WSC) je test, který představili v roce 2012 autoři Levesque, Davis a Morgenstern. Jde o specializovanou úlohu zaměřenou na koreferenční (odkazující a odkazované) vztahy v textu. Základním prvkem WSC u benchmarku SuperGLUE jsou dvojice vět obsahující zájmeno a více jmenných frází, přičemž úkolem testovaného subjektu je správně určit, ke které jmenné frázi se dané zájmeno vztahuje. Tyto věty jsou speciálně navrženy tak, aby jejich řešení vyžadovalo znalosti běžného světa, kontextu nebo tzv. „**common-sense reasoning**“ (uvažování založené na logickém rozumu).

Původně byla tato úloha součástí benchmarku GLUE ve formě „Winograd Natural Language Inference“ (WNLI – česky „Winogradovo odvozování přirozeného jazyka“) (Wang et al., 2018). Od WSC se lišila tak, že modelu nabídla premisu a zeptala se na uvozenou hypotézu – na otázku pak model odpovídal ano/ne. Při testování tohoto typu úlohy modely dlouho nezaznamenaly žádný významný pokrok – mnoho testovaných systémů se uchýlovalo k předpovídání většinové třídy jako výchozí strategie. Vznikal zde totiž problém s nevyváženou distribucí tříd v datasetu – jedna třída (odpověď „ano“ nebo „ne“) se vyskytovala výrazně častěji než druhá. Systémy začaly tento nedostatek zneužívat tím, že místo skutečného řešení úlohy (určování vztahů) jednoduše předpovídaly tu častější třídu pro všechny testované otázky. Například pokud bylo v trénovacích datech 70 % odpovědí „ano“, systém odpovídal vždy „ano“ bez ohledu na vstupní větu. Tato strategie vedla k relativně vysoké přesnosti, ale ve skutečnosti neprokazovala žádné porozumění textu. Proto byla v SuperGLUE úloha přepracována do nové podoby (Wang et al., 2019).

Dvě zásadní práce autorů Kocijan a Liu zaznamenaly pomocí rozšíření trénovacích i testovacích dat u modelů **přesnosti**⁷ až 90,4 %. Autoři práce *A Surprisingly Robust Trick for Winograd Schema Challenge* (Kocijan et al. 2019) se věnovali zejména rozšíření dat pro předtrénování modelů pro tento typ úlohy. Na tuto práci navázali o 2 měsíce později autoři práce *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (Zhuang et al., 2021), kteří představili nový model s výrazně lepšími výsledky v úloze (zmíněných 90,4 %). Vzhledem k odhadované human-baseline (přibližně 96 %) však stále existuje značná mezera mezi výkonem modelů a lidských subjektů. V rámci benchmarku SuperGLUE byla úloha WSC přepracována do podoby binární klasifikace – model tedy odpovídá pouze „ano“ nebo „ne“ na otázku, zda je navržená interpretace zájmena správná. Tato modifikace byla významná, protože zjednodušila vyhodnocování odpovědí modelů a umožnila přesnější srovnání jejich výkonnosti oproti původní verzi úlohy, kde modely mohly poskytovat různorodější odpovědi. Trénovací a validační data, stejně jako testovací příklady pocházejí z původního WSC datasetu (Levesque et al., 2012), další trénovací a validační data poskytl organizace Commonsense Reasoning (Wang et al., 2019).

2.1.2 Argumentace pro výběr WSC

Cílem výběru úlohy bylo identifikovat nejvhodnější dataset pro strojový překlad a adaptaci do českého jazyka pro účely testování v této práci. Po rozboru všech úloh benchmarku SuperGLUE byla jako nejvhodnější kandidát pro adaptaci do českého jazyka zvolena právě úloha Winograd Schema Challenge. WSC představuje unikátní nástroj pro testování hlubokého porozumění jazyku, který přesahuje sémantickou analýzu a vyžaduje pokročilejší odvozování souvislostí mezi zájmeny. Kromě kontextu testuje schopnost řešit jazykové nejasnosti s využitím znalostí o světě a kontextových vodítek, což představuje lingvistickou výzvu společnou pro většinu jazyků. Například u věty „Tato kniha seznámila Shakespeara s Ovidiem; byl to pěkný výběr z jeho díla“ je zřejmé, že zájmeno „jeho“ odkazuje na dílo Ovidia, který se nemohl nikdy seznámit s dílem Shakespeara, který žil přibližně o 1600 let později.

Na rozdíl od datové a textové náročnějších úloh (např. MultiRC) využívá WSC krátké, stručné textové pasáže, což významně usnadňuje proces překladu a adaptace do jazyka

⁷ Pro vyhodnocení úloh typu WSC se využívá metrika „Přesnost“ (Accuracy), která vyjadřuje podíl správně klasifikovaných odpovědí ze všech testovaných otázek.

při zachování vysoké náročnosti úkolu. Úloha dále umožňuje otestovat schopnosti logického uvažování modelu nad rámec čistě statistického zpracování jazyka, testuje logické odvozování, kontextuální porozumění a využití implicitních znalostí. Na příkladu věty „Položil jsem skleničku na policičku a ona spadla a rozbila se na tisíc střepů“ lze určit, že zájmeno „ona“ odkazuje na skleničku, protože policička dle zvyklostí není vyrobena ze skla.

Ostatní úlohy z benchmarku SuperGLUE byly vyhodnoceny jako méně vhodné pro adaptaci do českého jazyka. Například úloha BoolQ je příliš závislá na specifickém kontextu vyhledávacích dotazů v anglickém jazyce, který se může mezi jazyky a kulturami významně lišit. Úloha CommitmentBank by byla další vhodný kandidát pro porovnání překladu z anglického do českého jazyka manuálním i strojovým překladem, ale WSC umožňuje snazší možnost vytvořit kompletní nový příklad u vět, které nelze vhodně přeložit. Pro finální výběr WSC bylo tedy rozhodující jak zjednodušení samotného procesu přípravy dat, tak i praktické hledisko – menší počet tokenů potřebných pro testování znamená nižší výpočetní a tím i finanční náročnost experimentu. MultiRC obsahuje velmi komplexní textové pasáže s více možnými odpověďmi a příprava tohoto datasetu by byla významně časově náročnější s ohledem na vyšší velikost testovacího datasetu a množství textu v něm obsaženého, než je zvolená náročnost této práce. ReCoRD vykazuje silnou závislost na kontextu zpravodajských článků, který je kulturně i časově vázaný a jeho adaptace do cizího jazyka by vyžadovala pečlivější přístup než pouhý překlad.

WSC je pro češtinu obzvláště přínosná vzhledem k jejím specifickým lingvistickým charakteristikám. Komplexní systém skloňování a časování v českém jazyce činí řešení odkazujících zájmen výrazně náročnějším než v angličtině, jelikož flexibilní slovosled klade zvýšené požadavky na správné pochopení kontextuálních vztahů. Povaha úlohy navíc umožňuje zachovat integritu zadání úlohy i po jejím přeložení. Většina existujících úloh, například ty, které využívá BenCzechMark, se často soustředí hlavně na sémantický význam slov, stylistiku vět nebo embedding. WSC k tomu však dále přidává úroveň logického i kontextuálního významu a zároveň vyžaduje mnohem hlubší pochopení toho, jak gramatika, sémantika i pragmatika vstupují do otázky kontextu. Tímto způsobem WSC představuje vhodný komplexní nástroj pro evaluaci jazykových modelů, protože efektivně testuje jejich schopnost porozumět složitým jazykovým jevům specifickým pro češtinu.

2.1.3 Dataset pro WSC

Testovací podkladová data benchmarku SuperGLUE jsou volně přístupná ke stažení na stránce <https://super.gluebenchmark.com/tasks>. Soubor obsahuje složku se třemi datasety:

- test.jsonl – dataset pro testování
- val.jsonl – dataset pro validaci
- train.jsonl – dataset pro předtrénování modelu

Cílem této práce je ověřit aktuální dovednosti zvolených modelů; datasety pro trénink a validaci tedy nebudou využity. Pro překlad byl zvolen dataset „test“.

Dataset je nabídnutý ve formátu JSONL, což představuje specializovaný formát pro práci s datovými soubory, který se vyznačuje především svou jednoduchostí a efektivitou při zpracování. Na rozdíl od standardního formátu JSON, kde je celý dataset uložen jako pole objektů, JSONL ukládá každý JSON objekt na samostatný řádek (JSON Lines). Tato charakteristika přináší významné výhody zejména při práci s rozsáhlými datasety nebo při streamování dat.

V porovnání s běžně používaným formátem XLSX nabízí JSON (a tedy i JSONL) několik praktických výhod: jde o čistě textový formát, který lze snadno zpracovávat pomocí standardních Python knihoven jako „json“ nebo „pandas“, přičemž jeho zpracování je méně náročné na systémové prostředky (Turner-trauring, 2022).

Každý řádek dokumentu představuje jeden trénovací příklad a obsahuje následující strukturu:

- **"text"** - Obsahuje větu nebo text.
 - Příklad: „Mark o sobě Petovi řekl mnoho lží, které Pete zahrnul do své knihy. Měl by mu méně lhát.“
- **"target.span1_index"** - Číselný index začátku první sledované entity v textu, počítaný od 0. Udává, kolik samostatných celků ve větě předchází entitě.
 - Příklad: 0 (Mark)
- **"target.span1_text"** – Entita číslo jedna, v tomto dokumentu představuje jmennou frázi
 - Příklad: „Mark“
- **"target.span2_index"** - Číselný index začátku druhé sledované entity v textu. Splňuje stejná pravidla jako „target.span1_index“
 - Příklad: 15 (mu)
- **"target.span2_text"** – Entita číslo dvě, v tomto dokumentu představuje testované zájmeno.
 - Příklad: „mu“
- **"idx"** – unikátní číselné ID záznamu
 - Příklad: 1
- **"label"** – Obsahuje hodnoty „true“ nebo „false“ a označují, zda entity odkazují na stejný subjekt nebo ne. Funguje jako „ground truth“ při testování.
 - Příklad: false

V uvedeném příkladu neodkazuje zájmeno „mu“ na jmennou frázi „Mark“, ale na jmennou frázi „Petovi“, model by měl tedy správně odpovědět „false“.

2.2 Překlad datasetu

V rámci této práce bude vybraný dataset „test“ podroben dvojímu typu překladu za účelem vytvoření dvou odlišných verzí pro následné testování.

První verze datasetu, označená jako „ML_translated_test“, bude generována pomocí strojového překladu DeepL. Volba tohoto překladače byla motivována několika klíčovými faktory:

- jeho dominantním postavením na trhu strojového překladu (DeepL, 2024),
- širokou dostupností (jak v bezplatné, tak prémiové verzi),
- uživatelskou přívětivostí,
- možností využití API rozhraní pro automatizovaný překlad.

Druhá verze datasetu, nesoucí označení „adapted_test“, bude vycházet z výše zmíněného strojového překladu. Tento dataset bude následně podroben důkladné manuální revizi a adaptaci, při níž bude každý řádek individuálně posouzen a upraven lidským překladatelem – autorkou této práce. Cílem této varianty je zajistit, aby výsledný překlad přesně odpovídal specifickým požadavkům testování dle Winograd Schema Challenge a byl plně aplikovatelný v kontextu českého jazyka.

2.2.1 Překlad pomocí DeepL

Výchozí dataset „test“ ve formátu JSONL, získaný z webu frameworku SuperGLUE, bude importován, již převedený do formátu XLSX do prostředí nástroje Google Sheets, které bylo vybráno pro své pokročilé možnosti manipulace s daty, a především pro kompatibilitu s doplňkem Apps Script. Toto řešení umožňuje efektivní propojení s API rozhraním překladače DeepL prostřednictvím vlastního skriptu.

Vzhledem k rozsahu překládaného textu, který přesahuje maximální limit 500 000 znaků dostupných v bezplatné verzi služby DeepL, bude třeba implementovat řešení využívající placenou verzi překladače. Překladový skript byl navržen selektivně – zpracovává pouze vybrané části datasetu určené k překladu do českého jazyka, čímž se zachovává integrita dat potřebných pro následné testování. Tento přístup zajistí, že části dokumentu, které jsou důležité pro funkčnost testovacího skriptu v programovacím jazyce Python zůstanou zachované, což umožní konzistentní použití napříč všemi verzemi datasetu (strojově přeloženého, adaptovaného i v původní anglické verzi).

2.2.2 *Adaptace do českého jazyka*

Druhá fáze zpracování datasetu spočívá v manuální lingvistické a kulturní adaptaci strojově přeloženého podkladu. Dataset „adapted_test“ ve formátu JSONL, který vznikne na základě strojového překladu s nástrojem DeepL, bude podroben kompletní manuální revizi. Tento proces zahrnuje systematickou kontrolu a úpravu přeložených sekcí, a to s důrazem na zachování specifických vlastností testu WSC.

Manuální revize bude vyžadovat významné úpravy zaměřené na dva klíčové aspekty:

- Konzistence překladu identických (binárních) větných celků a jejich částí, která je klíčová pro zachování zamýšlené dvojznačnosti odkazujících zájmen,
- gramatická i stylistická správnost textu pro zajištění spolehlivého testování porozumění textu v českém jazyce.

Zvláštní pozornost bude věnována revizi indexace odkazujících zájmen a referenčních jmenných frází. Přestože výchozím bodem bude automatizovaná indexace slov v datasetu, bude následně provedena důkladná manuální kontrola pro možnost eliminace potenciálních chyb vzniklých během procesu strojového překladu i automatizované indexace.

Tímto navrženým procesem bude zajištěno, že výsledný dataset bude nejen lingvisticky korektní, ale také plně funkční v kontextu testování českého jazyka a kultury, při současném zachování všech klíčových charakteristik původního testu Winograd Schema Challenge.

2.3 **Výběr modelů pro testování**

Při výběru testovaných modelů pro tuto práci hrála primární klíčovou roli jejich široká adopce a aktivní využívání značnou částí populace, s cílem analyzovat nástroje, které jsou volně přístupné (zdarma nebo za poplatek), plošně využíváné a jejichž výsledky mají dopad na jejich praktické využití společností. Důraz byl kladen na vyhodnocení modelů, jež jsou všeobecně uznávané a mají zároveň potenciál generovat dobré výsledky i v českém jazyce. V této části jsou podrobně popsány důvody jejich zařazení do testu a je provedena podrobná analýza jejich vlastností.

Samotné společnosti přesný počet svých uživatelů často neuvádí a zároveň neexistuje pravidelně aktualizovaný žebříček těch nejpopulárnějších nástrojů. Při výběru byl proto využit článek dostupný na webu www.visualcapitalist.com, který na základě informací o návštěvnosti webových stránek nástrojů generativní AI využívá údaje z nástroje SimilarWeb, což je služba zpřístupňující údaje o návštěvnosti webů v dostupném internetu. Článek byl zveřejněn v říjnu 2024, ale odkazuje na informace z března 2024 (Zhu, 2024). Na základě tohoto žebříčku byly

pro práci vybrány modely z rodin GPT-4, Gemini, Claude 3.5 (ačkoli nejde o první tři místa v umístění – viz následující odstavec) a k těmto třem zavedeným modelům byly dále doplněny novější modely LLaMA 3.

Naproti tomu některé často využívané a navštěvované platformy uvedené v žebříčku Visual Capitalist, konkrétně nástroj Poe a Perplexity, nebyly do testování zahrnuty. Poe funguje jako agregační platforma nástrojů, která poskytuje přístup k více než jednomu LLM z jednotného rozhraní, ale nejedná se o samostatný jazykový model (Poe). Nástroj Perplexity je využíván primárně jako systém pro generování rozšířeného vyhledávání (**Retrieval-Augmented Generation**, též jen **RAG**) a zaměřuje se spíše na vyhledávání na internetu než na porozumění jazyku a generování textu (Singh, 2024a). Perplexity tedy také není velký jazykový model, jedná se o nástroj pro úlohy vyhledávání informací, které nejsou předmětem testování jazykových kompetencí v této práci.

Vedle těchto velmi populárních nástrojů byl do testování vybrán také model LLaMA od společnosti Meta, přestože se ve zvoleném žebříčku neumístil. Model byl představen relativně nedávno a jeho nejnovější skupina modelů č. 3 byla zveřejněna teprve v červenci 2024 (Llama Team, 2025). Následně, v září 2024, Meta odhalila pokročilejší verzi modelu nazvanou Llama 3.2 se 405 miliardami parametrů, která rozšiřuje možnosti plnění vícejazyčných úloh, kódování a komplexního řešení matematických problémů (Meta, 2024a). Veřejnost může k modelu přistupovat jak skrze integrace do produktů společnosti Meta (Facebook, Instagram, WhatsApp a další), ale také skrze otevřený zdrojový kód na GitHubu a HuggingFace (Meta, 2024b). Modely LLaMA byly podle Mety od svého vydání staženy téměř 350milionkrát, přičemž jsou hojně využívány i dalšími technologickými společnostmi jako jsou Goldman Sachs, AT&T nebo Nomura Holdings (Paul, 2024). S ohledem na tuto informaci o množství využití mimo webové rozhraní nebo internetové prohlížeče byl tento model vyhodnocen jako vhodný pro testování vedle modelů GPT-4, Gemini a Claude.

Všechny vybrané modely mají další množství společných vlastností. Všechny lze označit za velké jazykové modely postavené na pokročilých transformačních architekturách, které jim umožňují dosáhnout hlubokého kontextuálního pochopení textu a řešit komplexní úlohy zpracování přirozeného jazyka. Jsou navrženy pro širokou škálu účelů, včetně generování nového textu, porozumění jazyku, překladu, sumarizace nebo asistence při programování a kódování, a vykazují tak jistou všestrannost. Výsledky těchto modelů jsou velmi přesvědčivé, výsledné texty jsou podobně kvalitní nebo často kvalitnější než ty generované lidmi.

Tyto pokročilé modely jsou výsledkem intenzivního výzkumu předních světových technologických společností: GPT-4 od společnosti OpenAI, Gemini od společnosti Google DeepMind, Claude 3.5 od společnosti Anthropic a LLaMA 3 od společnosti Meta. V oblasti generativní umělé inteligence tyto organizace reprezentují současnou technologickou špičku (Cottier et al., 2023). Všechny modely byly trénovány na rozsáhlých souborech textových (a dnes už i vizuálních) dat, což je faktor určující jejich schopnost zpracovávat různorodé vstupy, včetně úloh zadaných pomocí méně rozšířených jazyků, jako je i čeština. Kvalitu výstupů těchto modelů lze významně vylepšit pomocí procesu zvaného „**fine-tuning**“ (jemného ladění). Tento specializovaný postup spočívá v dodatečném trénování již existujícího („naučeného“) základního modelu na vybraných datech, čímž se model lépe přizpůsobí konkrétním potřebám či úlohám (Yang et al., 2024).

Pro účely této práce byly vybrány tyto čtyři modely také díky své dostupnosti a praktické využitelnosti v reálném uživatelském prostředí. Každý z modelů nabízí několik způsobů implementace, které odpovídají různým potřebám a technickým možnostem cílových uživatelů nebo firem. Základní přístup představuje využití API rozhraní, které umožňuje relativně snadnou integraci modelu do existujících systémů, což dělá z vybraných modelů vhodné kandidáty jak pro systematické testování v rámci výzkumné části práce, tak pro následná doporučení jejich praktického nasazení.

2.3.1 GPT-4 by OpenAI

Nejnovější dostupný model ke dni testování byl model „gpt-4o-2024-11-20“.

GPT-4 byl zveřejněn společností OpenAI dne 14. března 2023 jako pokročilý jazykový model pro práci s textem. Model byl zpočátku dostupný pouze pro uživatele služby ChatGPT Plus a vývojáře přes API OpenAI. Cena za používání modelu se pohybovala od 0,03 \$ do 0,12 \$ za 1000 tokenů, v závislosti na verzi modelu. V září 2023 byla přidána podpora pro zpracování obrázků pod názvem GPT-4V (Vision).

V době uvedení GPT-4 byla již dostupná hlavní vlajková loď OpenAI – nástroj ChatGPT. Jedná se o chatovací rozhraní, které bylo spuštěno 30. listopadu 2022 ještě na LLM GPT-3.5. Bezplatná verze ChatGPT nyní používá model GPT-4, zatímco uživatelé placeného předplatného mají přístup také k verzi GPT-4o a nově i k nejnovějšímu modelu OpenAI o1 zpřístupněnému 5. 12. 2024 (Introducing ChatGPT Pro). ChatGPT dosáhl rekordní rychlosti rozšíření mezi uživateli – během prvních dvou měsíců získal 100 milionů uživatelů (Reuters, 2024).

Dne 6. července 2023 OpenAI zpřístupnila API pro model GPT-4 všem vývojářům, kteří si zaplatili předplatné (Wiggers, 2023b). To umožnilo širší využití GPT-4 v různých aplikacích – od odpovídání na dotazy až po asistenci s programováním. Model bylo již možné využívat v malých aplikacích nebo integrovat do velkých firemních systémů. Mezi významné implementace modelu ChatGPT patřila jeho integrace do Microsoft Bing Chat a GitHub Copilot. GitHub Copilot, vyvinutý ve spolupráci s OpenAI, byl uveden v roce 2021 jako nástroj pro automatické doplňování kódu, který pomáhá vývojářům psát kód efektivněji (Gershgorn, 2021). V únoru 2023 Microsoft představil přepracovanou verzi vyhledávače Bing, která využívá model GPT-4 od OpenAI, čímž poskytuje uživatelům pokročilé funkce AI přímo ve vyhledávání (Goodwin & Jackson, 2024).

OpenAI později vydala nové verze modelu:

- **OpenAI o1** (12. září 2024) - první komplexní⁸ model OpenAI zaměřený na multimodální analýzu, zahrnující práci s textem, obrázky a základními datovými vstupy, s důrazem na flexibilitu a škálovatelnost (Introducing OpenAI o1-preview, 2024).
- **GPT-4o** (13. května 2024) - vylepšená verze pro práci s textem, obrázky a zvukem, s podporou vstupů až do 128 000 tokenů (Introducing GPT-4o and more tools to ChatGPT free users, 2024).
- **GPT-4o Mini** (18. července 2024) - menší a levnější verze se základními funkcemi, optimalizovaná pro běžné použití s kontextovým oknem 8 000 tokenů (GPT-4o mini: advancing cost-efficient intelligence, 2024).

Využívání modelu GPT-4 se rychle rozšířilo. V srpnu 2024 měl samotný ChatGPT přes 100 milionů aktivních uživatelů týdně, což je dvakrát více než na konci roku 2023 a více než 92 % společností z žebříčku Fortune 500 nějakým způsobem využívalo technologii GPT-4 (Reuters, 2024). Mezi hlavní oblasti využití patřila automatizace zákaznické podpory, vývoj softwaru, analýza a zpracování dokumentů a další (Singh, 2024b).

2.3.2 Gemini by Google DeepMind

Nejnovější dostupný model ke dni testování byl model „gemini-1.5-pro“.

⁸ Nabízí pokročilou analýzu, zpracování složitých kontextů a univerzálnost pro různé typy úkolů, čímž sjednocuje přístupy, které dříve vyžadovaly více specializovaných modelů, a zvyšuje efektivitu výstupů (Introducing OpenAI o1-preview, 2024).

Gemini je název pro dva propojené produkty Google – velký jazykový model (LLM) a chatbot, který je na modelu postaven. Model Gemini byl představen v prosinci 2023 jako přímá odpověď na ChatGPT od OpenAI. Chatbot byl původně spuštěn pod názvem Bard v březnu 2023 a v únoru 2024 byl přejmenován na Gemini (Kruppa, 2023). Model byl nejprve postaven na jazykovém modelu nazvaném LaMDA⁹, později přešel na model PaLM¹⁰ a nakonec na vlastní model Gemini (Harisudhan S, 2024). Vývoj byl urychlen, jelikož „nástup ChatGPT zvýšil konkurenci a urychlil vývoj alternativních modelů umělé inteligence, což motivovalo vytvoření modelů Bard společnosti Google a Llama společnosti Meta“ (Kamalov et al., 2023, str. 4).

Služba je aktuálně dostupná ve 239 zemích a podporuje podle společnosti Google 46 jazyků (Pichai & Hassabis, 2023). Od 13. prosince 2023 mohou vývojáři získat placený přístup ke službě Gemini Pro prostřednictvím rozhraní Gemini API v aplikaci Google AI Studio nebo Google Cloud Vertex AI (Pichai & Hassabis, 2023). V únoru 2024 byla představena prémiová verze „Gemini Advanced with Ultra 1.0“ dostupná v rámci předplatného „Google One AI Premium“. Na platformě Android Gemini nahradil nástroj Google Assistant v pozici výchozího virtuální asistenta (Nield, 2024).

Gemini je multimodální jazykový model schopný zpracovávat text, obrázky i kód (Harisudhan S, 2024). V květnu 2024 byl model aktualizován na verzi Gemini 1.5, která přinesla vylepšené funkce včetně Gemini Live pro hlasovou komunikaci a možnosti vytvářet vlastní chatboty „Gems“ (Hsiao, 2024). Systém využívá model Imagen pro generování obrázků a nabízí pokročilé funkce pro práci s daty (Citron, 2024).

Podle údajů z webu SimilarWeb z prosince 2023 Gemini (tehdy ještě Bard) dosahoval průměrně 220 milionů měsíčních návštěvníků, tedy asi osmkrát méně, než byla uživatelská základna konkurenčního nástroje ChatGPT v té době (Kruppa, 2023). Model je integrován do řady produktů Google včetně Chrome, Photos a Workspace. Významnou součástí historie modelu Gemini jsou také kontroverze. V únoru 2024 čelil chatbot Gemini od společnosti Google kritice za generování historicky nepřesných obrázků, například zobrazoval černošské

⁹ Language Model for Dialogue Applications – rodina jazykových modelů vyvinutých společností Google, specializovaných na vedení dialogu, které obsahují až 137 miliard parametrů a byly předtrénovány na 1,56 bilionu slov (Thoppilan et al., 2022).

¹⁰ Pathways Language Model – jazykový model společnosti Google s 540 miliardami parametrů, trénovaný pomocí systému Pathways. Model vyniká v porozumění a tvorbě textu v různých jazycích a umí také psát počítačové programy (Chowdhery et al., 2022).

Vikingy či nacisty (Robertson, 2024). V listopadu 2024 došlo k incidentu, kdy Gemini během konverzace se studentem poskytl nevhodnou a urážlivou odpověď, v níž studenta označil za „mrhání času a zdrojů“ a vyzval ho, aby zemřel. Společnost Google tento problém uznala, uvedla, že se jedná o porušení jejích etických zásad, a zavedla přísnější mechanismy filtrování obsahu, aby podobným případům v budoucnu zabránila (Lenett, 2024).

2.3.3 Claude 3.5 Sonnet by Anthropic

Nejnovější dostupný model ke dni testování byl model „claude-3-5-sonnet-20241022“.

Claude je označení pro rodinu velkých jazykových modelů (LLM) vyvinutých společností Anthropic. První verze byla uvedena v březnu 2023. Stejně jako GPT modely, Claude je postavený na architektuře generativního předtrénovaného transformeru dekodéru. Co jej od GPT nebo podobných modelů podle jeho autorů odlišuje je fine-tuning pomocí metody „**constitutional AI**“, volně přeloženo „ústavní AI“. Cílem tohoto rámce je poskytnout pravidla založené na etických zásadách podle principů, které by měly být „sladěné“ (**aligned**) s lidskými morálními principy. Konkrétně Claude byl trénován na 10 principech, které dohromady utváří něco jako „ústavu“, Anthropic ale těchto 10 zásad nezveřejnil. Tvůrci však tvrdí, že jde o principy se snahou maximalizovat pozitivní výsledek, vyhnout se škodlivým odpovědím a respektování uživatelské autonomie (Wiggers, 2023a).

Systém prošel rychlým vývojem: Claude 2 (červenec 2023) zvýšil kontextové okno na 100 000 tokenů (Claude 2, 2023), Claude 2.1 na 200 000 tokenů (Introducing Claude 2.1, 2023). V březnu 2024 byla představena rodina Claude 3 se třemi modely: Haiku (optimalizovaný pro rychlost), Sonnet (vyvážený) a Opus (pro komplexní úlohy) s kontextovým oknem až 1 milion tokenů pro specifické případy (Introducing the next generation of Claude, 2024). Nejnovější Claude 3.5 Sonnet (červen 2024) přinesl vylepšení v oblasti programování, práce s grafy a zpracování obrazu. Základní verze Claude 3.5 Sonnet je dostupná bezplatně se značnými limity využití, vyžaduje e-mail a telefonní číslo uživatele. Služby jsou poskytovány přes webové rozhraní claude.ai a API. Placený plán nabízí přístup ke všem modelům Claude 3 kromě Opus (Claude 3.5 Sonnet, 2024). Od května 2024 existuje podnikové řešení Claude Team a iOS aplikace (Introducing the Claude Team plan and iOS app, 2024).

Claude vyniká v matematice, programování a logickém uvažování. Umí zpracovávat text i obrázky, analyzovat PDF dokumenty a další materiály. Verze 3.5 přinesla funkci Artifacts (což jsou editovatelné sekce v rozhraní chatu) pro tvorbu a zobrazení kódu v reálném čase včetně generované grafiky nebo náhledu webových stránek. Od října 2024 podporuje interakci

s počítačovým prostředím přes „computer use“. Tato experimentální technologie je dostupná pro vývojáře prostřednictvím API a umožňuje modelu simulovat lidskou interakci s počítačem včetně vizuálního vnímání obrazovky, ovládání kurzoru, interakce s uživatelským rozhraním a zadávání textu. Současná implementace se podle autorů potýká s určitými limitacemi a občasnými nepřesnostmi (Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku, 2024).

2.3.4 Llama 3 by Meta

Nejnovější dostupný model ke dni testování byl model „llama3.2-3b“. V této době byly zveřejněny i novější a větší modely Llama (především zajímavý pro tento test by byl největší model Llama 3.1 405B nebo nejnovější Llama 3.3 70B), jenže tyto modely byly spolu s dalšími licenčně omezené pro použití uživateli mimo oblast Evropské Unie (Pasberg, 2024).

Llama (Large Language Model Meta AI) je jazykový model vyvinutý společností Meta AI. Od svého prvního vydání v únoru 2023 prošel model několika iteracemi, přičemž každá přinesla rozšíření jeho technických možností. Meta AI se při vývoji zaměřila na efektivitu trénování a optimalizaci poměru výkonu k velikosti modelu (Martindale, 2024).

Model L je stejně jako GPT modely založená na architektuře transformeru „decoder-only“, která optimalizuje architekturu především pro úlohy generování textu. První verze Llamy byla dostupná pouze pro výzkumné účely. Llama 2 rozšířila trénovací dataset na 2 biliony tokenů a zpřístupnila model pro komerční využití. Llama 3 dále navýšila objem trénovacích dat na 15 bilionů tokenů a zvětšila kontextové okno na 8192 tokenů. Verze 3.1 přinesla model s 405 biliony parametrů a kontextovým oknem 128 000 tokenů. Nejnovější verze 3.2 a 3.3 nabízí modely v rozsahu 1B až 70B parametrů (Meta Llama, 2024).

Model je dostupný v několika formách: jako součást služeb Meta, například na Facebooku a WhatsAppu (Martindale, 2024). Pro snadnější lokální nasazení modelu Llama vznikly dva významné open-source projekty: llama.cpp a ollama. Llama.cpp, vyvinutý Georgim Gerganovem v březnu 2023, je efektivní C++ implementace umožňující spouštění modelu na běžném hardwaru. Projekt ollama, založený Jeffreyem Morganem v červenci 2023, staví na llama.cpp a přináší další optimalizace včetně vylepšeného řízení paměti a využití moderních CPU instrukcí. Ollama navíc nabízí tzv. Modelfiles, které umožňují upravovat parametry existujících modelů nebo importovat soubory ve formátu GGUF (Llama.cpp vs. ollama: Running LLMs Locally for Enterprises, 2024). Dále vznikly specializované varianty jako Code Llama pro programování, která prošla dodatečným tréninkem na 500 miliardách tokenů programového kódu (Code Llama, a state-of-the-art large language model

for coding). Pro potřeby testování v této práci bude využit nástroj třetí strany – společnosti Llama AI Inc., která poskytuje služby pro připojení k rozhraní API pro modely Llama i další.

2.4 Testování

V rámci praktické sekce budou zvolené modely podrobeny testování pomocí dvou přeložených českých datasetů a původního datasetu v angličtině. Testování proběhne prostřednictvím automatizovaného procesu využívajícího implementaci vlastního skriptu v programovacím jazyce a prostředí Python. Pro zajištění konzistentního a spolehlivého testování všech čtyř vybraných modelů budou využita jejich API rozhraní v komerčních verzích. Toto řešení poskytne standardizovaný přístup k modelům, umožní systematickou evaluaci jejich výkonu a zajistí reprodukovatelnost výsledků testování.

2.5 Vyhodnocení výsledků

Při vyhodnocení úspěšnosti testovaných modelů budou jejich výsledky získané z testování na třech různých datasetech porovnané mezi sebou a také s human-baseline určenou benchmarkem. Hlavní metrikou pro hodnocení výkonu modelů byla autory benchmarku SuperGLUE zvolena přesnost (Accuracy), jelikož jde o přímočarou a snadno vyhodnotitelnou metriku, která je vhodná pro tento typ binární úlohy, ve které lze jednoduše určit správnou či špatnou odpověď.

Přesnost v procentech bude vypočítána podle vzorce:

$$\text{Accuracy} = \frac{\text{počet správných odpovědí}}{\text{celkový počet předpovědí}}$$

Pro každý model budou získány tři hodnoty přesnosti:

- Přesnost na anglickém datasetu
- Přesnost na strojově přeloženém českém datasetu
- Přesnost na manuálně adaptovaném českém datasetu

Přesnost jako metrika nám tedy poskytne odpověď na otázku, v kolika procentech případů z celkového počtu otázek model odpověděl správně.

Pro statistickou validaci výsledků budou vypočteny intervaly spolehlivosti pro každou hodnotu přesnosti a proveden párový t-test pro porovnání výkonnosti mezi datasety.

Interpretace výsledků se zaměří na několik klíčových aspektů. Primárně budou porovnávány výsledky modelů s úrovní human-baseline, kterou používají i autoři původního benchmarku SuperGLUE, tedy s úrovní výkonu testovaných lidských subjektů v anglickém jazyce (Wang et al., 2019). Pro připomenutí, úspěšnost se na původním anglickém datasetu pohybuje okolo 96 %. Toto srovnání poskytne kontext pro určení, jak blízko jsou testované modely k dosažení lidské úrovně porozumění jazyku na této úloze.

Další součástí vyhodnocení výsledků bude detailní analýza rozdílů ve výsledcích mezi anglickou a českou verzí, přičemž zvláštní pozornost bude věnována také rozdílům mezi strojově přeloženou a manuálně adaptovanou verzí. Tato analýza umožní posoudit, do jaké míry kvalita překladu ovlivňuje schopnost modelů řešit WSC úlohy.

Součástí vyhodnocení bude také důkladná kontrola chybných odpovědí zaměřená na případy, ve kterých modely nejčastěji selhávaly. Cílem bude identifikace opakujících se vzorců v typech úloh, které se ukázaly jako problematické, a zkoumání toho, zda se typy chyb liší mezi anglickou a českou verzí. Zvláštní pozornost bude věnována analýze jazykové specifických jevů, zejména použití zvrtných zájmen a přivlastňovacích konstrukcí v češtině. Tato analýza se zaměří na to, jak tyto jazykové prvky a vlastnosti ovlivňují schopnost modelů správně interpretovat koreferenční vztahy.

Na vybraných reprezentativních příkladech bude provedena podrobná případová studie, která umožní detailní analýzu rozhodovacího procesu modelů a identifikaci konkrétních faktorů ovlivňujících jejich úspěšnost.

2.6 Hypotézy a předpoklady testování

Na základě provedené literární rešerše a analýzy zvolených modelů a metod lze formulovat několik odhadů očekávaných výsledků testování.

Při srovnání s původním anglickým datasetem se předpokládá, že výkon modelů bude nižší oproti oběma českým datasetům. Tento předpoklad vychází ze skutečnosti, že všechny testované modely byly primárně trénovány na textech v anglickém jazyce, s českou jazykovou složkou představující pouze menší část jejich trénovacího korpusu. Lze očekávat rozdíl přibližně 5 – 10 % mezi anglickou a českou verzí, přičemž tento odhad vychází z podobných studií, které přeložili totožnou úlohu do jiných jazyků. Překlad WSC úlohy SuperGLUE do slovinštiny přinesl horší či srovnatelné výsledky – výrazně lepší výsledky na benchmarku získal jen slovinsky natrénovaný model SloBERTa (Žagar and Robnik-Šikonja, 2022). Ruská verze úlohy „RWSD“ (The Russian Winograd Schema Dataset) přinesla na modelech učených

na ruských textech u úlohy WCS pouze 2 až 3% vylepšení výsledků (Shavrina et al., 2020). Překlad úlohy do portugalštiny byl obohacen o velmi podrobné testování mezi anglickým a přeloženým datasetem, přičemž téměř každé testování dopadlo hůř ve prospěch překladu (Melo et al., 2019).

V otázce vlivu překladu datasetu na výkonnost modelů lze předpokládat, že adaptovaný český dataset povede k mírně lepším výsledkům než strojově přeložený dataset. Tento předpoklad vychází z poznatků studií zabývajících se překladem benchmarků do jiných jazyků, například slovinštiny (Žagar and Robnik-Šikonja, 2022) nebo španělštiny (Plaza a kol., 2024). Tyto studie naznačují, že kvalita překladu má přímý vliv na schopnost modelů správně interpretovat jazykové nuance. Zároveň lze očekávat, že strojový překlad může v některých případech narušit přirozené jazykové konstrukce a tím ztížit správnou interpretaci koreferenčních vztahů, nehledě na problematičnost chybějících zájmen či jejich špatné indexace v datasetu.

V otázce výkonnosti jednotlivých modelů lze na základě jejich architektury, velikosti a dosavadních výsledků v podobných úlohách předpokládat následující pořadí:

Claude 3.5 Sonnet by měl dosáhnout nejlepších výsledků díky své pokročilé architektuře a důrazu na kontextuální porozumění. Předpokládaná přesnost mezi 85 – 90 % na anglickém datasetu, 75 – 80 % na českém adaptovaném datasetu a 70 – 75 % na strojově přeloženém českém datasetu.

GPT-4o by měl následovat s mírným odstupem, pravděpodobně s přesností kolem 80 – 85 % na anglickém, 75 – 80 % na adaptovaném a 70 – 75 % na strojově přeloženém datasetu. Tento výkon lze předpokládat díky robustní architektuře modelu a rozsáhlému tréninkovému datasetu, stejně jako prokázaným schopnostem v komplexních jazykových úlohách.

Gemini lze očekávat na třetím místě s předpokládanou přesností 75 – 80 %, především kvůli jeho silnému zaměření na multimodální úlohy na úkor čistě textového zpracování. Úspěšnost na adaptovaném a strojově přeloženém datasetu odhaduji na intervalech 70 – 75 % a 65 – 70 %.

Llama 3 pravděpodobně dosáhne nejnižších výsledků (očekávaná přesnost 60 – 65 % na anglickém datasetu, 55 – 60 % na adaptovaném a 50 – 55 % na strojově přeloženém datasetu) vzhledem k jeho menší velikosti, která může limitovat jeho výkon v komplexních jazykových úlohách.

Lze také předpokládat specifické vzorce chyb u jednotlivých modelů. U strojově přeloženého datasetu lze očekávat problémy především v případech, kde překlad narušil původní koreferenční vztahy nebo kde došlo ke změně gramatického rodu či čísla, dále také u úloh, ve kterých vlivem překladu bylo dotazované zájmeno vynecháno nebo přeloženo odlišně od entity v příslušném sloupci. U adaptovaného datasetu lze předpokládat lepší výsledky v případech využívajících česká zvrtná přivlastňovací zájmena „svůj“ a jeho dalších variant, které mohou poskytovat dodatečné vodítko pro určení správného aktéra.

Důležitým faktorem bude konzistence výkonu napříč různými typy vět. Lze očekávat, že věty vyžadující znalost světa nebo logické uvažování budou představovat větší výzvu než věty, kde lze koreferenci určit na základě gramatických pravidel. Tento rozdíl by měl být výraznější u malého modelu, zatímco větší modely by měly vykazovat konzistentnější výkon v logických nebo historicky kontextových úlohách. Je možné také očekávat, že některé modely mohou vykazovat systematické chyby v určitých typech konstrukcí, například v případech, kde je třeba rozlišit mezi více potenciálními aktéry stejného rodu a čísla. Tyto případy budou pravděpodobně představovat největší výzvu napříč všemi testovanými modely.

3 Praktická část

3.1 Překlad datasetu

3.1.1 Strojový překlad pomocí nástroje DeepL

Z oficiálního zdroje benchmarku SuperGLUE (<https://super.gluebenchmark.com/tasks>) byl stažen kompletní balík datasetů pro WSC. Pro účely této práce byl využit pouze soubor test (Příloha 10).

Pro usnadnění následné manipulace s daty byl napsán Python skript (Příloha 1). Tento skript zajistil konverzi dat z původního formátu JSONL (JSON Lines), který ukládá strukturovaná data jako sérii JSON objektů na jednotlivých řádcích, do tabulkového formátu XLSX.

Konvertovaná data ve formátu XLSX byla importována do nově vytvořené tabulky v prostředí Google Sheets, které poskytlo vhodné uživatelské rozhraní pro následující fázi zpracování.

Do tabulky v Google Sheets byl integrován doplněk Google Apps Script, prostřednictvím kterého byl použit skript zajišťující komunikaci s placenou verzí DeepL API. Toto řešení umožnilo automatizovaný překlad vybraných datových sloupců.

Pomocí skriptu byly přeloženy následující sloupce:

- "text" obsahující premisu
- "target.span1_text" obsahující první entitu – jmennou frázi
- "target.span2_text" obsahující druhou entitu – zájmeno

Přeložená data byla poté stažena z Google Sheets do souboru ve formátu XLSX a následně konvertována skriptem (Příloha 2) zpět do formátu JSONL, který je vyžadován pro následné testování v Python prostředí. Vzniklý dataset byl pojmenován ML_translated_test (Příloha 9).

3.1.2 Přizpůsobení přeloženého datasetu

V této sekci je popsán postup manuální adaptace strojově přeloženého datasetu. Prvním krokem procesu byla konverze souboru ve formátu XLSX do formátu XLSX (Příloha 2) a otevření přeloženého datasetu v aplikaci Microsoft Excel, kde byla následně provedena jeho manuální adaptace. Pro zajištění přesnosti a správnosti adaptace bylo k přeloženým částem

přidáno do vedlejších sloupců původní anglické znění, což umožnilo průběžnou kontrolu a porovnávání obou verzí.

Adaptační proces postupoval podle strukturované metodologie, která začínala detailní analýzou původního textu a identifikací testovaných jazykových jevů v každé úloze. Tento systematický přístup zahrnoval několik klíčových kroků: nejprve bylo nutné přesně určit funkci zkoumaného anglického zájmena a analyzovat jeho syntaktické i sémantické vztahy v rámci věty. Následovala důkladná kontrola těchto vztahů porovnáním s poskytnutou odpovědí ve sloupci `span1_text` a ověření správnosti binárního hodnocení PRAVDA/NEPRAVDA uvedeného ve sloupci `label`. Na základě ustanoveného porozumění originálnímu textu bylo možné přistoupit ke kritickému posouzení kvality strojového překladu do češtiny.

V případech, kdy strojový překlad adekvátně zachovával všechny podstatné souvislosti a zájmenné vazby, byla provedena pouze minimální optimalizace zahrnující úpravu slovosledu, opravu překlepů a gramatických chyb. Pokud však přeložená verze neodpovídala požadavkům testování, bylo přistoupeno k výraznějším úpravám. Ty zahrnovaly modifikaci celkového vyznění věty, záměnu některých slov či kompletně odlišný překlad textu. V mnoha případech bylo nutné změnit celkový význam nebo přeformulovat celou větu, zejména pokud doslovný překlad neumožňoval zachování testovaných jazykových jevů.

Zvláštní pozornost byla věnována větám, které tvořily páry či větší skupiny vět, například když testovaly stejnou větu s různými aktéry nebo obsahovaly více testovacích zájmen. Tyto věty byly adaptovány tak, aby fungovaly jako koherentní celek při současném zachování jejich testovací funkce. Po dokončení překladu textů, zájmen a jmenných frází byly pomocí skriptu v pythonu (Příloha 3) a následné manuální kontroly upraveny hodnoty ve sloupcích `span1_index` a `span2_index` tak, aby odpovídaly pozicím v nově formulovaných větách. Nakonec byly z dokumentu odebrané anglické sloupce a zůstaly jen sloupce odpovídající formátu původního datasetu.

Dokument byl poté ze souboru XLSX konvertován zpět do souboru ve formátu JSONL. Výsledkem tohoto procesu byl soubor `adapted_test` (Příloha 8), který přesně kopíruje formát a strukturu původního anglického datasetu, avšak obsahuje pouze adaptovanou českou verzi. Tento soubor byl připraven pro následné testování a představuje finální výstup adaptačního procesu.

3.2 Příprava skriptu

Skripty byly připraveny pro každý model zvlášť podle veřejné dokumentace pro jejich API rozhraní. Každý skript sdílel stejnou logiku postupu a také obsahoval vytvořený prompt. Rozdíl byl pouze ve skriptech testujících jiný jazyk, vizte následující kapitolu.

3.2.1 Prompt

Rozhodnutí o tom, v jakém jazyce bude úkol pro model zadáván, bylo založeno na výsledcích dvou relevantních studií. Výzkumy ukazují, že kvalita výstupů je ovlivněna jazykem, kterým jsou modelu úkoly zadávány. Autoři studie s názvem *Eliciting the Translation Ability of Large Language Models via Multilingual Finetuning with Translation Instructions* (Li et al., 2024) zjistili, že LLM model GPT-4 má lepší výsledky při úlohách překladu, když je úkol zadán v cílovém jazyce.

Autoři další studie s názvem *Investigating Multilingual Instruction-Tuning: Do Polyglot Models Demand for Multilingual Instructions?* (Weber et al., 2024) zkoumali vliv jazyka instrukcí na modely fungující v multilingválním způsobu. Výsledky ukazují, že poskytnutí instrukcí ve stejném jazyce, ve kterém je vyžadován i výstup, zlepšuje šanci na lepší vygenerovaný výsledek.

Na základě těchto informací byla zvolena varianta zadávat úkoly testující český dataset v českém jazyce a anglický dataset v anglickém jazyce.

V akademické studii od SuperGLUE uvádí autoři text, kterým zadávali stejnou úlohu (WSC) lidským testovaným subjektům. Pro účely testování byl proto úkol zadán všem modelům podobně a jako vzor textu byla vypůjčena část původního zadání:

Anglický text využitý pro testování lidských subjektů pro úlohu WCS v SuperGLUE:

Winograd Schema Instructions

The New York University Center for Data Science is collecting your answers for use in research on computer understanding of English. Thank you for your help!

We will present you with a sentence that someone wrote, with one bolded pronoun. We will then ask if you if the pronoun refers to a specific word or phrase in the sentence. Your job is to figure out, based on the sentence, if the bolded pronoun refers to this selected word or phrase:

Choose Yes if the pronoun refers to the selected word or phrase. For example,

"I put the cake away in the refrigerator. It has a lot of butter in it."

Does It in "It has a lot" refer to cake?

Choose No if the pronoun does not refer to the selected word or phrase.
For example, "The large ball crashed right through the table because it was made of styrofoam."

Does it in "it was made" refer to ball? (Wang et al., 2019)

Český text obsahuje přeložené části zadání a zároveň i programatickou část promptu, která umožňuje zadat všechny úlohy v datasetu v rámci jednoho probíhajícího testování. Programatická část poskytuje modelu strukturované informace o formátu datasetu a specifikuje, která zájmena a slova ve větách jsou předmětem otázek.

Součástí promptu je dále také několik příkladů úlohy a její správné odpovědi, tzv. „**few-shot examples**“ metoda pro lepší výsledky promptování.

Adaptovaná verze promptu pro testování v českém jazyce:

Úkol: Rozlišování zájmen

Centrum datových věd Newyorské univerzity shromažďuje tvoje odpovědi, aby je mohlo použít ve výzkumu porozumění generativních textových AI modelů češtině.
Díky za pomoc!

Předložím ti větu nebo skupinu vět, které někdo napsal. V každé skupině bude jedno vybrané zájmeno. Poté se tě zeptáme, zda se zájmeno vztahuje ke konkrétnímu slovu nebo slovnímu spojení ve této větě či skupině vět.

Tvým úkolem je na základě kontextu a významu věty zjistit, zda se vybrané zájmeno vztahuje k tomuto vybranému slovu nebo slovnímu spojení.

Příklad 1:

1. Věta/skupina vět: „Anna dala své sestře objetí, protože si ho zasloužila.“
2. Zájmeno: „ho“ (pozice: 7)
3. Slovo/fráze: „objetí“ (pozice: 4)
4. Otázka: Odkazuje zájmeno „ho“ na slovo/frázi „objetí“?
5. Odpověď: true

Příklad 2:

6. Věta/skupina vět: „Kočka honila myš. Ona vlezla do díry.“
7. Zájmeno: „Ona“ (pozice: 3)
8. Slovo/fráze: „Kočka“ (pozice: 0)
9. Otázka: Odkazuje zájmeno „ona“ na slovo/frázi „kočka“?
10. Odpověď: false

****Věta/skupina vět:**** {entry['text']}

****Zájmeno:**** {entry['target.span2_text']} (pozice: {entry['target.span2_index']})

****Slovo/fráze:**** {entry['target.span1_text']} (pozice: {entry['target.span1_index']})

****Otázka:**** Odkazuje zájmeno *******{entry['target.span2_text']}******* na slovo/frázi *******{entry['target.span1_text']}*******?

Pokud ano, odpověz ****true****. Pokud ne, odpověz ****false****.

Adaptovaná verze promptu pro testování v anglickém jazyce (původní dataset):

Winograd Schema Task

New York University's Center for Data Science is collecting your answers to use in research on understanding generative text-based AI models of English. Thank for your help!

I'll submit a sentence or group of sentences that someone wrote. Each group will contain one selected pronoun. We will then ask you to identify whether the pronoun refers to a specific word or phrase in that sentence or group of sentences.

Your task is to use the context and meaning of the sentence to determine whether the selected pronoun refers to that selected word or phrase.

Example 1:

1. Sentence/sentence group: "Anna gave her sister a hug because she deserved it."
2. Pronoun: "she" (position: 7)
3. Word/phrase: "sister" (position: 3)
4. Question: Does the pronoun "she" refer to the word/phrase "sister"?
5. Answer: true

Example 2:

1. Sentence/sentence group: "The cat chased a mouse. She climbed into the hole."
2. Pronoun: "She" (position: 3)
3. Word/phrase: "Cat" (position: 0)
4. Question: Does the pronoun "she" refer to the word/phrase "cat"?
5. Answer: false

****Sentence/sentence group:**** {entry['text']}

****Pronoun:**** {entry['target.span2_text']} (position: {entry['target.span2_index']})

****Word/phrase:**** {entry['target.span1_text']} (position: {entry['target.span1_index']})

****Question:**** Does the pronoun *******{entry['target.span2_text']}******* refer to the word/phrase *******{entry['target.span1_text']}*******?

If it does, answer „yes“. If it does not, answer „no“

3.2.2 Příprava skriptů pro testování

Pro každý testovaný model byl vytvořen vlastní testovací skript odpovídající požadavkům jeho API dokumentace (detailní dokumentace jednotlivých skriptů je k dispozici v Přílohách 11-14). Přestože konkrétní forma implementace se pro jednotlivé modely liší, základní princip testování zůstává konzistentní napříč všemi modely.

Specifický případ představovala implementace testování modelu Llama 3 od společnosti Meta. Tento model, na rozdíl od ostatních testovaných systémů, nedisponuje nativním API konektorem a je primárně určen pro lokální testování. Kvůli technickým komplikacím při konverzi dat pro lokální testování nejnovějšího modelu bylo zvoleno alternativní řešení využívající placené API rozhraní třetí strany – Llama AI API. Toto řešení umožnilo zachovat jednotnou metodiku testování v prostředí Python, čímž byla zároveň zajištěna konzistence s testováním ostatních modelů.

Skripty byly navrženy podle těchto veřejných dokumentací:

- GPT od OpenAI - <https://github.com/openai/openai-python>
- Claude od Anthropic - <https://docs.anthropic.com/en/api/getting-started>
- Gemini od DeepMind - <https://ai.google.dev/gemini-api/docs/models/gemini>
- Llama API od Llama AI - <https://docs.llama-api.com/quickstart>

Procesní průběh každého skriptu je pro každé testování totožný a je podrobně popsán v následujících podkapitolách.

3.2.2.1 Importy knihoven a konfigurace skriptu

První klíčovou komponentu skriptu představuje sekce importů knihoven a konfigurace skriptu. Tato část importuje všechny nezbytné Python knihovny potřebné pro interakci s příslušným API rozhraním daného jazykového modelu. Poté je zde implementováno nastavení tajného API klíče, který zajišťuje autorizovaný přístup k placené verzi testovaného modelu.

```
import openai
from openai import OpenAI

client = OpenAI(api key=os.getenv("API-KLÍČ"))
```

3.2.2.2 Načtení připraveného testovacího datasetu ve formátu JSONL

```
def load_dataset(file_path):
    with open(file_path, 'r', encoding='utf-8') as f:
```

```
return [json.loads(line) for line in f]
```

3.2.2.3 Definování promptu

```
def create_prompt_english(entry):
    task_deskription = ...
    few_shot_examples = ...
    test_case = ...
    return f'{task_deskription} {few_shot_examples}\n{test_case}'
```

3.2.2.4 Výběr modelu a zadání úlohy formou „true/false“

V této části je vložen připravený prompt a zadané instrukce včetně zvolené verze modelu podle dokumentace.

```
def query_model(prompt):
    try:
        response = client.chat.completions.create(
            model="gpt-4o-2024-11-20",
            messages=[{"role": "user", "content": prompt}],
            temperature=0
        )
        result = response.choices[0].message.content
        return result.strip().lower()
    except Exception as e:
        print(f'Error during API call: {e}')
    return None
```

3.2.2.5 Doplnkové funkce pro testování

I když by testování mohlo být provedeno bez této části skriptu, byly následující funkcionality implementovány pro zjednodušení celého procesu a zvýšení efektivity práce s testovacím prostředím. Pro přehlednou vizualizaci průběhu testování a generování reprezentativních výstupů byla implementována specializovaná formátovací funkce (**beautifying function**). Funkce „display_example_results“ přispívá ke zvýšení uživatelského komfortu při bezprostřední analýze výstupů testování.

Funkce implementuje několik klíčových prvků. Prvním je průběžné zobrazování postupu testování formou počítadla (např. „145/554...“), které poskytuje uživateli okamžitou zpětnou vazbu o aktuálním stavu procesu. Dalším elementem je mechanismus pro výběr náhodných vzorků výsledků, který segreguje testovací případy do dvou kategorií – správné a nesprávné odpovědi.

Implementace využívá „list comprehension“ pro efektivní filtraci výsledků podle jejich správnosti. Následně je pomocí funkce „sample“ z každé kategorie náhodně vybráno pět příkladů. Toto omezení počtu vzorků zajišťuje přehlednost výstupu při současném zachování reprezentativnosti vybraných případů. V situacích, kdy je k dispozici méně než pět příkladů v dané kategorii, jsou vybrány všechny dostupné případy, což je zajištěno použitím funkce „min“ při určování velikosti vzorku.

```
def display_example_results(results):
    correct_results = [r for r in results if r['correct']]
    incorrect_results = [r for r in results if not r['correct']]
    sample_correct = sample(correct_results, min(5, len(correct_results)))
    sample_incorrect = sample(incorrect_results, min(5, len(incorrect_results)))...
```

3.2.2.6 Evaluate

Jádrem testovacího procesu je evaluační funkce „evaluate_model“, která implementuje komplexní vyhodnocení výkonu testovaného modelu. Funkce inicializuje proměnné pro sledování počtu správných odpovědí a celkového počtu testů. Zároveň umožňuje flexibilně omezit velikost testované množiny dat pomocí parametru „max_samples“, což zajišťuje efektivní, levnější a škálovatelné testování.

V hlavní evaluační smyčce je pro každý testovací případ vygenerován standardizovaný prompt, který je předán modelu k vyhodnocení. Získané odpovědi jsou průběžně porovnávány s referenčními výsledky, na základě čehož je aktualizován počet správných odpovědí. Po dokončení testování je vypočtena celková přesnost modelu a veškeré výsledky, včetně detailních informací o jednotlivých případech, jsou uloženy do nového souboru ve formátu JSONL ve stejném adresáři jako původní dataset a testovací skript.

```
def evaluate_model(dataset, max_samples=None):
    results = []
    correct = 0
    total = len(dataset) if max_samples is None else min(max_samples, len(dataset))
    for i, entry in enumerate(dataset, 1):
        prompt = create_prompt_english(entry)
        model_response = query_model(prompt)
        accuracy = correct / total
        with open(output_file, 'w', encoding='utf-8') as f:
            json.dump({
                'accuracy': accuracy,
            }, f, ensure_ascii=False, indent=2)
```

3.2.2.7 Exekuce

Finální komponenta testovacího skriptu obsahuje exekuční blok, který spouští celý testovací proces. Tento blok využívá standardní Python konstrukci „if name == 'main'“, která zajišťuje, že testování proběhne pouze při přímém spuštění skriptu. Pomocí funkce „load_dataset“ se načte soubor testovacího datasetu ze specifikované složky. Tato funkce zajišťuje, že data jsou správně načtena a připravena pro evaluaci modelu. Následně je spuštěna hlavní evaluační funkce „evaluate_model“ s načteným datasetem jako parametrem, čímž je zahájeno testování modelu na všech dostupných datech.

```
if __name__ == "__main__":  
    test_data = load_dataset("/path/.../test.jsonl")  
    evaluate_model(test_data)
```

Všechny vytvořené skripty jsou dostupné jako příloha této práce (Příloha 11–14).

3.3 Testování:

V této sekci je popsán průběh procesu testování.

1. Pro zajištění systematického přístupu byla vytvořena strukturovaná organizace testovacích souborů, kde každý testovaný model měl vlastní adresář.
2. V rámci tohoto adresáře byly vytvořeny tři samostatné podsložky, každá dedikovaná specifickému typu datasetu – jedna pro anglický dataset (EN), druhá pro strojově přeložený dataset (ML) a třetí pro adaptovaný dataset.
3. Samotná exekuce testovacích skriptů proběhla v prostředí Python3 skrze terminál počítače s operačním systémem macOS.
4. Po dokončení každého testu byl výsledek modelu prezentován dvěma způsoby. Výsledky přesnosti na úlohách se zobrazily přímo v terminálu jako okamžitá zpětná vazba o průběhu testování a současně byl automaticky vygenerován nový soubor ve formátu JSONL, který byl uložen do příslušné složky s datasetem. Tento soubor obsahoval detailní informace o průběhu testování a umožnil následnou validaci výsledků včetně analýzy správných i nesprávných odpovědí.
5. Důležitou součástí procesu byla systematická analýza chybových případů, která se soustředila na detailní vyhodnocení situací, kdy se predikované odpovědi modelu neshodovaly s referenčními hodnotami (**ground truth labels**). Zvláštní pozornost byla věnována analýze nejednoznačných případů, které mohly potenciálně vyžadovat úpravu designu promptů pro lepší specifikaci úlohy. V průběhu zkušebního testování na modelu GPT-4 byl na základě výsledků identifikován

problém se záměnou labelů označujících správný výsledek a dataset byl proto ještě jednou upraven a zkontrolován.

6. Finální procentuální výsledky testování byly následně zaneseny do souhrnné tabulky pro další analýzu a porovnání výkonu jednotlivých modelů.

3.3.1 Vyhodnocení

3.3.1.1 Kvantitativní analýza výkonu

Na základě provedeného testování čtyř jazykových modelů (GPT-4, Gemini-1.5-pro, Claude-3.5-sonnet a Llama3.2-3b) byly získány následující výsledky ve třech navržených testovacích scénářích. Datasets se všemi odpověďmi modelů jsou přiložené v Přílohách 4–7.

Model:	gpt-4o-2024-11-20	gemini-1.5-pro	claude-3-5-sonnet-20241022	llama3.2-3b
Adapted	75,99 %	78,70 %	88,63 %	53,43 %
ML	68,05 %	65,70 %	72,20 %	53,25 %
Original	75,09 %	77,62 %	88,27 %	53,25 %
Human baseline	~ 96 %	~ 96 %	~ 96 %	~ 96 %

Výsledné pořadí testovaných modelů podle jejich výsledků v úloze WSC pro český jazyk:

- a. Claude 3.5 Sonet (model „20241022“)
- b. Gemini-1.5-Pro
- c. GPT-4o (model „2024-11-20“)
- d. Llama3.2 (model „3B“)

Na základě výsledků je zřejmé, že adaptace datasetu do českého jazyka neovlivnila relativní pořadí výkonnosti testovaných modelů (ve srovnání s původním anglickým datasetem), nicméně měla pozitivní vliv na jejich absolutní výsledky. Ve srovnání s datasetem přeloženým strojově dosáhly modely na adaptovaném datasetu lepších výsledků, což naznačuje, že kvalitně adaptovaný dataset v cílovém jazyce vede k lepšímu porozumění zadání a tím i k přesnějšímu řešení úlohy. Zajímavým zjištěním je také skutečnost, že modely dosáhly na českém adaptovaném datasetu dokonce lepších výsledků než při testování s původním anglickým datasetem, což podtrhuje význam kvalitní jazykové adaptace pro zlepšení výkonu jazykových modelů v cílových jazycích.

Model Claude-3.5-sonnet dosáhl nejvyšší přesnosti napříč všemi testovanými datasety, s nejlepším výsledkem 88,63 % na adaptovaném datasetu. Tento výsledek představuje nejvyšší dosaženou přesnost v celém experimentu. GPT-4 a Gemini-1.5-pro prokázaly srovnatelnou

výkonnost, s přesností pohybující se mezi 75 – 79 % na adaptovaném a původním datasetu. Model Llama3.2-3b dosáhl konzistentní, ale výrazně nižší přesnosti přibližně 53 % napříč všemi datasety. Při porovnání výsledků s lidským benchmarkem (~ 96 %) je patrné, že i nejlépe ohodnocený model (Claude-3.5-sonnet) stále zaostává o přibližně 7,4 % za lidskou přesností. Tento rozdíl ilustruje přetrvávající rozdíly mezi schopnostmi současných jazykových modelů a lidského porozumění v oblasti koreference.

Významným zjištěním tohoto testování je skutečnost, že adaptovaný český dataset vedl k mírnému zlepšení výsledků ve srovnání s původním anglickým datasetem. Tento trend je patrný především u modelů Claude-3.5-sonnet (zlepšení o 0,36 %) a Gemini-1.5-pro (zlepšení o 1,08 %). Pro každou dvojici hodnot (adaptovaný a originální model) byly vypočteny rozdíly:

- gpt-4o-2024-11-20: 75,99 % – 75,09 % = 0,90 %
- gemini-1.5-pro: 78,70 % – 77,62 % = 1,08 %
- claude-3-5-sonnet-20241022: 88,63 % – 88,27 % = 0,36 %
- llama3.2-3b: 53,43 % – 53,25 % = 0,18 %

Rozdíly tedy tvoří dataset: [0.90, 1.08, 0.36, 0.18].

Výpočet průměru rozdílů (μ_d)

$$\mu_d = \frac{\text{součet rozdílů}}{\text{počet hodnot}}$$

$$\mu_d = \frac{0.9 + 1.08 + 0.36 + 0.18}{4} = \frac{2.52}{4} = 0.63$$

Výpočet směrodatné odchylky rozdílů (s_d)

$$s_d = \sqrt{\frac{\sum (x - \mu_d)^2}{n - 1}}$$

$$\sum (x - \mu_d)^2 = 0.0729 + 0.2025 + 0.0729 + 0.2025 = 0.5508$$

$$\frac{0.5508}{3} = 0.1836$$

$$s_d = \sqrt{0.1836} \approx 0.428$$

Výpočet t-statistiky

$$t = \frac{\mu_1 \cdot \sqrt{n}}{s_1}$$

$$0.63 \cdot 2 = 1.26$$

$$t = \frac{1.26}{0.428} \approx 2.944$$

Výpočet Cohenova d

$$d = \frac{\text{průměr rozdílů } (\mu_1)}{\text{směrodatná odchylka } (s_1)}$$

$$d = \frac{0.63}{0.428} \approx 1.47$$

Pro $n - 1 = 3$ stupně volnosti a hladinu významnosti $p < 0,05$, je kritická t-hodnota přibližně 2,3532. Protože $t = 2,944 > 2,3532$, výsledek je statisticky významný. Párový t-test ukázal, že rozdíly mezi adaptovanými a originálními datasy jsou konzistentní a dostatečně velké, aby byly považovány za statisticky významné na hladině $p < 0,05$, což potvrzuje, že úprava datasetů má pozitivní vliv na testování.

Effect size (Cohenovo d) byl vypočítán na $d = 1,47$, což ukazuje velký efekt podle běžných standardů (Cohen: $d > 0,8$ = velký efekt) (Soukup, 2013). To znamená, že rozdíl mezi adaptovanými a originálními výsledky má nejen statistickou, ale také praktickou významnost. Adaptace má tedy výrazný dopad na zlepšení výkonu modelů.

Zajímavým zjištěním je také konzistentně nižší výkonnost všech modelů na strojově přeloženém datasetu ve srovnání s adaptovaným i původním datasetem. Poukazuje na důležitou roli kvalitní přípravy testovacích dat. Zatímco strojový překlad může poskytnout základní převod mezi jazyky, nemusí zachytit idiomy nebo kontextově závislé významy, které jsou pro úlohy NLU zásadní. Adaptace datasetu lidskými překladateli nebo anotátory obeznámenými s lingvistickou terminologií a jazykovými specifiky může výrazně zlepšit kvalitu testovacích dat a následně i výkonnost hodnocených modelů na daném benchmarku.

Z lingvistického hlediska lze kromě jazykových specifik identifikovat další potenciální faktory přispívající k pozorovanému zlepšení. Významnou roli pravděpodobně sehrál proces adaptace datasetu, během něhož došlo k cílené optimalizaci testovacích vět změnou nejasných ambivalentních konstrukcí v původním anglickém datasetu. Tuto hypotézu lze ukázat

na příkladu anglické věty „Billy cried because Toby wouldn't share his toy“, která vykazuje výrazně vyšší míru nejistoty v odkazujícím zájmenu než její český ekvivalent „Billy plakal, protože se Toby nechtěl podělit o jeho hračku“, a zejména pak než optimální překlad „Billy plakal, protože se Toby nechtěl podělit o svou hračku.“ Zvratné přivlastňovací zájmeno „svůj“ v češtině nese významnou anaforickou informaci, která jednoznačně specifikuje vlastnický vztah – v tomto případě mezi Tobym a hračkou. Tato gramatická vlastnost češtiny tak výrazně usnadňuje správnou interpretaci koreferenčních vztahů v textu a může rovněž vést k lepším výsledkům při testování modelů. Faktorem je tedy celkově systematictější a konzistentnější způsob vyjadřování koreferencí v českém překladu, který vznikl jako přímý důsledek procesu adaptace datasetu. Bohatší systém morfologických prostředků pro vyjádření anaforických vztahů, včetně osobních a zvratných zájmen, které jsou navíc morfologicky značené pro rod, číslo a pád – tento gramatický aparát umožňuje přesnější a jednoznačnější vyjádření vztahů mezi jednotlivými elementy v textu. Během procesu adaptace byly tyto jazykové prostředky využívány konzistentním způsobem, což mohlo modelům poskytnout jasnější a předvídatelnější vzorce pro rozpoznávání koreferencí.

Tato zvýšená lingvistická kvalita příkladů mohla významně přispět k tomu, že modely byly schopny lépe aplikovat naučené znalosti o zpracování jazykových struktur na testovacích datech.

Další významné zjištění z výsledků se týká poměru velikosti modelů proti jejich výkonnosti. Pozorovaný výkon modelů ukazuje na výraznou nerovnoměrnost ve vztahu mezi velikostí modelu a jeho schopností zpracovávat koreferenční vztahy. Claude s přesností 88,63 % na adaptovaném datasetu a 88,27 % na originální anglické verzi dosahuje výrazně lepších výsledků než ostatní modely, včetně GPT-4o. To je pozoruhodné zejména v kontextu toho, že například Llama 3.2-3b, jakožto výrazně menší model, dosahuje přesnosti na úloze výsledku přibližně 53 % napříč všemi verzemi – větší model neznamena automaticky jen o trochu lepší výsledky – skok ve výkonu je zde dramatický.

Zvláště zajímavým zjištěním je konzistence výkonu tohoto menšího modelu napříč jazykovými verzemi. Zatímco u větších modelů je pozorována očekávaná variabilita v závislosti na specifických jazykových charakteristikách – což ukazuje na sofistikovanější zpracování lingvistických struktur – Llama 3.2 vykazuje téměř totožné výsledky nezávisle na jazyku či kvalitě překladu. Tato jednostrannost výkonu naznačuje přítomnost fenoménu „**shallow pattern matching**“, kdy model namísto skutečného porozumění složitým syntaktickým strukturám pouze identifikuje povrchové vzory v textu. Tento přístup se podobá

mělkému parsování (**shallow parsing**), které se zaměřuje pouze na lokální vztahy mezi slovy bez hlubšího pochopení celkové větné struktury (Argamon et al., 1998).

Statistická analýza výsledků odhaluje pravděpodobnou existenci výkonnostního prahu, který přímo souvisí s počtem parametrů modelu. Tato hranice naznačuje minimální výpočetní kapacitu potřebnou pro vytvoření dostatečně robustních reprezentací jazykových struktur. Data demonstrují, že modely nedosahující této hranice nejsou schopny generovat dostatečně komplexní vnitřní reprezentace pro spolehlivé zpracování náročnějších jazykových fenoménů, zejména koreferenčních vztahů. Při pohledu na architekturu těchto neuronových sítí lze zjistit, že menší modely mají základní omezení: jejich attention mechanismy (systémy pro věnování pozornosti různým částem textu) nemají dostatečnou kapacitu na to, aby mohly postupně propojovat souvislosti mezi vzdálenějšími částmi textu. Tato limitace se projevuje zejména při zpracování dlouhých jazykových struktur, kde je nutné sledovat a propojovat souvislosti mezi vzdálenými částmi textu.

3.3.1.2 Rozbor výsledků testování na vybraných úlohách

Z výsledků testování byly vybrány příklady odpovědí modelů, které dobře ilustrují specifické jazykové jevy typické pro češtinu. Pozornost byla zaměřena především na věty obsahující zvrtné zájmeno „se/si“ a přivlastňovací zájmeno „svůj“, tedy konstrukce, která nese implicitní informace o vztahu mezi podmětem a vlastněným předmětem – aspekt, který v angličtině nemá přímý ekvivalent. Díky odlišnému fungování těchto jazykových prostředků v češtině není vždy možné zachovat původní významovou nejednoznačnost anglické věty.

Pro ilustraci těchto jevů byly analyzovány odpovědi modelů na zvolenou testovací větu, zejména v případě, kdy došlo k rozdílu ve výsledcích dle datasetů. Výsledky testování ukazují rozdílnou úspěšnost jednotlivých modelů. Claude a Llama prokázaly konzistentní správnost napříč všemi verzemi věty, zatímco GPT-4 a Gemini měli obtíže zejména s anglickou verzí. Za pozornost stojí, že strojový překlad do češtiny používající „její“ místo „své“ způsobil dodatečné komplikace v interpretaci, přestože samotné sloveso „promluvila“ poskytovalo dostatečný kontext pro správné určení reference.

GPT-4

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	Lily spoke to Donna , breaking her silence.	her	Donna	NEPRAVDA	PRAVDA	NEPRAVDA	39
ml_translated	Lily promluvila k Donně , čímž přerušila její mlčení.	její	Donna	NEPRAVDA	PRAVDA	NEPRAVDA	39
adapted	Lily promluvila na Donnu, čímž přerušila své mlčení.	své	Donnu	NEPRAVDA	NEPRAVDA	PRAVDA	39

Claude

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	Lily spoke to Donna , breaking her silence.	her	Donna	NEPRAVDA	NEPRAVDA	PRAVDA	39
ml_translated	Lily promluvila k Donně , čímž přerušila její mlčení.	její	Donna	NEPRAVDA	PRAVDA	NEPRAVDA	39
adapted	Lily promluvila na Donnu, čímž přerušila své mlčení.	své	Donnu	NEPRAVDA	NEPRAVDA	PRAVDA	39

Gemini

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	Lily spoke to Donna , breaking her silence.	her	Donna	NEPRAVDA	PRAVDA	NEPRAVDA	39
ml_translated	Lily promluvila k Donně , čímž přerušila její mlčení.	její	Donna	NEPRAVDA	PRAVDA	NEPRAVDA	39
adapted	Lily promluvila na Donnu, čímž přerušila své mlčení.	své	Donnu	NEPRAVDA	NEPRAVDA	PRAVDA	39

Llama

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	Lily spoke to Donna , breaking her silence.	her	Donna	NEPRAVDA	NEPRAVDA	PRAVDA	39
ml_translated	Lily promluvila k Donně , čímž přerušila její mlčení.	její	Donna	NEPRAVDA	NEPRAVDA	PRAVDA	39
adapted	Lily promluvila na Donnu, čímž přerušila své mlčení.	své	Donnu	NEPRAVDA	NEPRAVDA	PRAVDA	39

V případě adaptovaného datasetu si lze všimnout, že překlad věty zachovává původní sémantickou strukturu, přičemž konstrukce „Lily spoke“ jednoznačně identifikuje původce děje. Tento kontextový ukazatel umožnil všem modelům správně rozpoznat vztahy mezi jednotlivými entitami v textu.

Naproti tomu strojový překlad vykazuje nepřesnosti, zejména v převodu zájmena „her“ na české „její“. Tato volba komplikuje interpretaci koreferenčních vztahů ve větě. Je třeba poznamenat, že i přes tuto překladovou nepřesnost by měl jazykový model být schopen na základě přítomnosti slovesa „promluvila“ správně dedukovat sémantický vztah týkající se přerušení ticha.

Testování ukazuje, že strojový překlad může neúmyslně zjednodušit nebo zkomplikovat úlohy rozpoznávání odkazu zájmen. Například v případě věty „Lily spoke to Donna, breaking her silence“ česká adaptace využívající „své“ poskytuje dodatečná gramatická vodítka, která nejsou přítomna v anglické verzi. To naznačuje, že lingvistické rysy specifické pro češtinu mohou některé úlohy WSC buď usnadnit, nebo naopak i ztížit způsobem, který nebyl v původním anglickém návrhu zamýšlen.

V dalším příkladu, ve větě „Muž zvedl chlapce na svá ramena“ přivlastňovací zájmeno jednoznačně odkazuje na muže, nikoliv na chlapce. Tato reference je naznačena nejen

gramatickou strukturou věty, ale také logickým kontextem situace, kdy akt zvedání někoho na ramena určuje „vlastnictví“ těchto ramen.

GPT-4

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	The man lifted the boy onto his shoulders.	his	The boy	NEPRAVDA	NEPRAVDA	PRAVDA	63
ml_translated	Muž zvedl chlapce na ramena.	jeho	Chlapec	NEPRAVDA	PRAVDA	NEPRAVDA	63
adapted	Muž zvedl chlapce na svá ramena.	svá	chlapce	NEPRAVDA	NEPRAVDA	PRAVDA	63

Claude

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	The man lifted the boy onto his shoulders.	his	The boy	NEPRAVDA	NEPRAVDA	PRAVDA	63
ml_translated	Muž zvedl chlapce na ramena.	jeho	Chlapec	NEPRAVDA	NEPRAVDA	PRAVDA	63
adapted	Muž zvedl chlapce na svá ramena.	svá	chlapce	NEPRAVDA	NEPRAVDA	PRAVDA	63

Gemini

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	The man lifted the boy onto his shoulders.	his	The boy	NEPRAVDA	NEPRAVDA	PRAVDA	63
ml_translated	Muž zvedl chlapce na ramena.	jeho	Chlapec	NEPRAVDA	PRAVDA	NEPRAVDA	63
adapted	Muž zvedl chlapce na svá ramena.	svá	chlapce	NEPRAVDA	NEPRAVDA	PRAVDA	63

Llama:

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	The man lifted the boy onto his shoulders.	his	The boy	NEPRAVDA	NEPRAVDA	PRAVDA	63
ml_translated	Muž zvedl chlapce na ramena.	jeho	Chlapec	NEPRAVDA	NEPRAVDA	PRAVDA	63
adapted	Muž zvedl chlapce na svá ramena.	svá	chlapce	NEPRAVDA	NEPRAVDA	PRAVDA	63

Testování jednotlivých modelů na této větě přineslo různorodé výsledky. GPT-4, Claude, Gemini i Llama správně interpretovali anglický originál i adaptovanou českou verzi. Zajímavé rozdíly se objevily u strojově přeložené verze, kde modely vykazovaly odlišnou úspěšnost. Zatímco Claude a Llama si zachovali konzistentní správnost napříč všemi verzemi, GPT-4 a Gemini prokázali obtíže s interpretací strojově přeložené verze. Strojový překlad zde zájmeno zcela vypustil, což je z hlediska překladatelské praxe přijatelné řešení, neboť vlastnictví ramen je v češtině vyjádřeno kontextem. Toto vynechání však představuje komplikaci pro účely testování a narušuje konzistenci datasetu. Zvolený překlad navíc způsobuje dodatečné problémy při zpracování, jelikož absence zájmena komplikuje indexaci (určení zvoleného zájmena ve větě) a může vést k situaci, kdy model nemá dostatek informací pro správnou interpretaci a musí se uchýlit k nepodloženým rozhodnutím.

Analýza překladu následující věty naznačuje opačný problém. V tomto případě strojový překladač zvolil pro překlad anglického zájmena „his“ české zvrtné přivlastňovací zájmeno „svůj“, což vedlo k zachování původního významu věty. Naproti tomu adaptovaná verze, kde bylo záměrně použito zájmeno „jeho“ ve snaze zvýšit komplexitu testu, vedla k nesprávné interpretaci ze strany testovaných modelů.

GPT-4

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	Billy cried because Toby wouldn't share his toy.	his	Toby	PRAVDA	PRAVDA	PRAVDA	10
ml_translated	Billy plakal, protože se Toby nechtěl podělit o svou hračku.	jeho	Toby	PRAVDA	PRAVDA	PRAVDA	10
adapted	Billy plakal, protože se Toby nechtěl podělit o jeho hračku.	jeho	Toby	PRAVDA	NEPRAVDA	NEPRAVDA	10

Claude

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	Billy cried because Toby wouldn't share his toy.	his	Toby	PRAVDA	PRAVDA	PRAVDA	10
ml_translated	Billy plakal, protože se Toby nechtěl podělit o svou hračku.	jeho	Toby	PRAVDA	NEPRAVDA	NEPRAVDA	10
adapted	Billy plakal, protože se Toby nechtěl podělit o jeho hračku.	jeho	Toby	PRAVDA	NEPRAVDA	NEPRAVDA	10

Gemini

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	Billy cried because Toby wouldn't share his toy.	his	Toby	PRAVDA	NEPRAVDA	NEPRAVDA	10
ml_translated	Billy plakal, protože se Toby nechtěl podělit o svou hračku.	jeho	Toby	PRAVDA	PRAVDA	PRAVDA	10
adapted	Billy plakal, protože se Toby nechtěl podělit o jeho hračku.	jeho	Toby	PRAVDA	NEPRAVDA	NEPRAVDA	10

Llama

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	Billy cried because Toby wouldn't share his toy.	his	Toby	PRAVDA	NEPRAVDA	NEPRAVDA	10
ml_translated	Billy plakal, protože se Toby nechtěl podělit o svou hračku.	jeho	Toby	PRAVDA	NEPRAVDA	NEPRAVDA	10
adapted	Billy plakal, protože se Toby nechtěl podělit o jeho hračku.	jeho	Toby	PRAVDA	NEPRAVDA	NEPRAVDA	10

GPT-4 jako jediný správně interpretoval anglický originál i strojový překlad, ale selhal u adaptované verze. Claude, který v předchozích případech vykazoval konzistentní správnost, v tomto případě nesprávně vyhodnotil jak strojově přeloženou, tak adaptovanou verzi. Gemini prokázal správnou interpretaci pouze u strojového překladu, zatímco Llama vykazovala nesprávné výsledky ve všech třech verzích.

Specifická povaha zvrtných a přivlastňovacích zájmen může zjevně ovlivnit obtížnost i validitu testovacích úloh. Pro efektivní testování jazykových modelů v českém prostředí

by bylo vhodné buď vyloučit věty obsahující zvrtná zájmena a podobné jazykově specifické konstrukce, nebo naopak vytvořit specializovanou sadu testů zaměřených přímo na tuto problematiku. Přístup pouhého překladu původního datasetu se ukazuje jako nedostatečný pro komplexní evaluaci schopností modelů v oblasti zpracování českého jazyka.

V dalším příkladu se zaměřuji na využití zvrtného se/si ve strojově přeložené i adaptované verzi. Zatímco angličtina používá reflexivní konstrukce typu „himself“ spíše omezeně, čeština využívá zvrtné „se“ jako běžnou součást mnoha sloves. Tato strukturální odlišnost se projevila v dalším testovaném příkladu:

GPT-4

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	John hired him self out to Bill to take care of him .	himself	Bill	PRAVDA	NEPRAVDA	NEPRAVDA	433
ml_translated	John se nechal najmout Billem, aby se o něj postaral.	ho	Bill	PRAVDA	NEPRAVDA	NEPRAVDA	433
adapted	John se nechal najmout Billem, aby se o něj postaral.	se	Billem	PRAVDA	NEPRAVDA	NEPRAVDA	433

Claude:

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	John hired him self out to Bill to take care of him .	himself	Bill	PRAVDA	NEPRAVDA	NEPRAVDA	433
ml_translated	John se nechal najmout Billem, aby se o něj postaral.	ho	Bill	PRAVDA	NEPRAVDA	NEPRAVDA	433
adapted	John se nechal najmout Billem, aby se o něj postaral.	se	Billem	PRAVDA	NEPRAVDA	NEPRAVDA	433

Gemini

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	John hired him self out to Bill to take care of him .	himself	Bill	PRAVDA	NEPRAVDA	NEPRAVDA	433
ml_translated	John se nechal najmout Billem, aby se o něj postaral.	ho	Bill	PRAVDA	PRAVDA	PRAVDA	433
adapted	John se nechal najmout Billem, aby se o něj postaral.	se	Billem	PRAVDA	NEPRAVDA	NEPRAVDA	433

Llama

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	John hired him self out to Bill to take care of him .	himself	Bill	PRAVDA	NEPRAVDA	NEPRAVDA	433
ml_translated	John se nechal najmout Billem, aby se o něj postaral.	ho	Bill	PRAVDA	NEPRAVDA	NEPRAVDA	433
adapted	John se nechal najmout Billem, aby se o něj postaral.	se	Billem	PRAVDA	NEPRAVDA	NEPRAVDA	433

GPT-4, Claude a Llama shodně selhali v správné interpretaci reference jak v anglickém originálu, tak v obou českých verzích. Pouze Gemini prokázal částečnou úspěšnost, když správně vyhodnotil strojově přeloženou verzi, ale selhal u adaptované verze i originálu. Je pozoruhodné, že přestože zvrtné zájmeno „se“ v české větě jednoznačně odkazuje na osobu Johna, tato gramatická konstrukce neměla významný vliv na schopnost modelů správně určit referenční vztahy. Tento příklad je obzvláště zajímavý tím, že pro účely testování byla celá věta

významně přepracována oproti anglickému originálu. Zatímco původní anglická konstrukce používá „hire oneself out“ (v původním datasetu s chybou – vloženou mezerou uvnitř slova „himself“), česká verze využívá zvrtné sloveso „nechat se najmout“, což představuje způsob stejného vyjádření v češtině. Tato adaptace, ačkoliv lingvisticky vhodná, nepomohla modelům v lepším porozumění referenčním vztahům ve větě.

Analýza dalšího příkladu přináší pohled na vliv znalostí reálií v jazykovém porozumění. V testované větě o vztahu mezi Shakespearem a Ovidiem se projevuje, jak důležitá je znalost historického kontextu pro správnou interpretaci koreferenčních vztahů:

GPT-4

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	This book introduced Shakespeare to Ovid ; it was a major influence on his writing.	his	Ovid	NEPRAVDA	NEPRAVDA	PRAVDA	514
ml_translated	Tato kniha seznámila Shakespeara s Ovidiem; měla na jeho tvorbu zásadní vliv.	jeho	Ovidius	NEPRAVDA	NEPRAVDA	PRAVDA	514
adapted	Tato kniha představila Shakespeara s Ovidiem; měl velký vliv na jeho dílo.	jeho	Ovidiem	PRAVDA	NEPRAVDA	NEPRAVDA	514

Claude

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	This book introduced Shakespeare to Ovid ; it was a major influence on his writing.	his	Ovid	NEPRAVDA	NEPRAVDA	PRAVDA	514
ml_translated	Tato kniha seznámila Shakespeara s Ovidiem; měla na jeho tvorbu zásadní vliv.	jeho	Ovidius	NEPRAVDA	NEPRAVDA	PRAVDA	514
adapted	Tato kniha představila Shakespeara s Ovidiem; měl velký vliv na jeho dílo.	jeho	Ovidiem	PRAVDA	NEPRAVDA	NEPRAVDA	514

Gemini

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	This book introduced Shakespeare to Ovid ; it was a major influence on his writing.	his	Ovid	NEPRAVDA	NEPRAVDA	PRAVDA	514
ml_translated	Tato kniha seznámila Shakespeara s Ovidiem; měla na jeho tvorbu zásadní vliv.	jeho	Ovidius	NEPRAVDA	NEPRAVDA	PRAVDA	514
adapted	Tato kniha představila Shakespeara s Ovidiem; měl velký vliv na jeho dílo.	jeho	Ovidiem	PRAVDA	NEPRAVDA	NEPRAVDA	514

Llama

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	This book introduced Shakespeare to Ovid ; it was a major influence on his writing.	his	Ovid	NEPRAVDA	NEPRAVDA	PRAVDA	514
ml_translated	Tato kniha seznámila Shakespeara s Ovidiem; měla na jeho tvorbu zásadní vliv.	jeho	Ovidius	NEPRAVDA	NEPRAVDA	PRAVDA	514
adapted	Tato kniha představila Shakespeara s Ovidiem; měl velký vliv na jeho dílo.	jeho	Ovidiem	PRAVDA	NEPRAVDA	NEPRAVDA	514

Při testování této věty všechny modely správně interpretovaly anglický originál i strojově přeloženou verzi, kde zájmeno „jeho“ jednoznačně odkazovalo na Ovidia. Zajímavý obrat nastal u adaptované verze, kde drobná syntaktická změna vedla k nesprávné interpretaci u všech modelů. Je třeba poznamenat, že adaptovaná verze obsahuje stylisticky problematickou konstrukci „představila Shakespeara s Ovidiem“, která není v češtině vhodná a může narušovat

porozumění textu. Správné by bylo použít formulaci „seznámila Shakespeara s Ovidiem“ jak zpracoval strojový překlad nebo lépe „představila Shakespearovi Ovidia“.

Tento příklad výrazně ilustruje několik důležitých aspektů. Zaprvé ukazuje možný vliv správné stylistiky a idiomatického vyjadřování pro úspěšné zpracování textu jazykovými modely. Na základě jednoho případu není ale možné udělat podložený závěr, zda toto mělo vliv na celkový výsledek nebo nikoliv – vyžadovalo by to pokročilejší zkoumání. Zadruhé chyba demonstruje, jak důležité je pro jazykové modely nejen zvládnutí gramatických vztahů, ale také využívání znalostí o reálném světě při interpretaci textu. I drobné syntaktické změny mohou významně ovlivnit to, jak modely analyzují koreferenční vztahy, a to i v případech, kdy by logické uvažování založené na znalosti historických faktů mělo jednoznačně vést ke správné odpovědi.

V posledním analyzovaném případě je analyzován zajímavý jev týkajícím se překladu a interpretace zájmen v komplexnější větné struktuře. Původní anglická věta obsahuje dvě instance zájmena „it“ v různých kontextech, přičemž poslední „turned it off“ se vztahuje k televizi, nikoli k myčce nádobí. Zájmeno „it“ je pro tento případ označeno v datasetu indexem, který označuje jeho pozici ve větě – výsledky se týkají posledního „it“.

GPT-4

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	Fred was supposed to run the dishwasher , but he put it off, because he wanted to watch TV. But the show turned out to be boring, so he changed his mind and turned it off.	it	The dishwasher	NEPRAVDA	NEPRAVDA	PRAVDA	156
ml_translated	Fred měl vyčistit myčku, ale odložil to, protože se chtěl dívat na televizi. Ukázalo se však, že pořad je nudný, a tak si to rozmyslel a vypnul ho .	to	Myčka nádobí	NEPRAVDA	NEPRAVDA	PRAVDA	156
adapted	Fred chtěl pustit myčku, ale odložil ji, protože se chtěl dívat na televizi. Ukázalo se však, že pořad je nudný, a tak si to rozmyslel a vypnul ji .	ji	myčku	NEPRAVDA	PRAVDA	NEPRAVDA	156

Claude

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	Fred was supposed to run the dishwasher , but he put it off, because he wanted to watch TV. But the show turned out to be boring, so he changed his mind and turned it off.	it	The dishwasher	NEPRAVDA	NEPRAVDA	PRAVDA	156
ml_translated	Fred měl vyčistit myčku, ale odložil to, protože se chtěl dívat na televizi. Ukázalo se však, že pořad je nudný, a tak si to rozmyslel a vypnul ho .	to	Myčka nádobí	NEPRAVDA	PRAVDA	NEPRAVDA	156
adapted	Fred chtěl pustit myčku, ale odložil ji, protože se chtěl dívat na televizi. Ukázalo se však, že pořad je nudný, a tak si to rozmyslel a vypnul ji .	ji	myčku	NEPRAVDA	NEPRAVDA	PRAVDA	156

Gemini

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	Fred was supposed to run the dishwasher , but he put it off, because he wanted to watch TV. But the show turned out to be boring, so he changed his mind and turned it off.	it	The dishwasher	NEPRAVDA	PRAVDA	NEPRAVDA	156
ml_translated	Fred měl vyčistit myčku, ale odložil to, protože se chtěl dívat na televizi. Ukázalo se však, že pořad je nudný, a tak si to rozmyslel a vypnul ho .	to	Myčka nádobí	NEPRAVDA	NEPRAVDA	PRAVDA	156
adapted	Fred chtěl pustit myčku, ale odložil ji, protože se chtěl dívat na televizi. Ukázalo se však, že pořad je nudný, a tak si to rozmyslel a vypnul ji .	ji	myčku	NEPRAVDA	PRAVDA	NEPRAVDA	156

Llama

dataset	text	pronoun	reference	expected	predicted	correct	idx
original	Fred was supposed to run the dishwasher , but he put it off, because he wanted to watch TV. But the show turned out to be boring, so he changed his mind and turned it off.	it	The dishwasher	NEPRAVDA	NEPRAVDA	PRAVDA	156
ml_translated	Fred měl vyčistit myčku, ale odložil to, protože se chtěl dívat na televizi. Ukázalo se však, že pořad je nudný, a tak si to rozmyslel a vypnul ho .	to	Myčka nádobí	NEPRAVDA	NEPRAVDA	PRAVDA	156
adapted	Fred chtěl pustit myčku, ale odložil ji, protože se chtěl dívat na televizi. Ukázalo se však, že pořad je nudný, a tak si to rozmyslel a vypnul ji .	ji	myčku	NEPRAVDA	NEPRAVDA	PRAVDA	156

Při strojovém překladu došlo k zajímavému jevu, kdy strojový překlad použil různé překlady pro stejné zájmeno v různých částech věty. V první části věty přeložil „it“ jako „to“ („odložil to“), zatímco v závěrečné části použil „ho“ („vypnul ho“), což přirozeně koresponduje s rodem slova "pořad". Navzdory této nekonzistenci v překladu zájmen modely GPT-4 a Llama správně interpretovaly referenční vztahy jak v originální, tak ve strojově přeložené verzi.

Adaptovaná verze, která se snažila o konzistentnější použití zájmen pomocí jednotného „ji“ odkazujícího na myčku, paradoxně vedla k horším výsledkům u některých modelů. GPT-4 a Gemini v tomto případě nesprávně určili, že poslední zájmeno odkazuje na myčku, přestože z kontextu je zřejmé, že Fred vypnul televizi, nikoli myčku, kterou ani nezapnul. Tento příklad ilustruje, jak snaha o gramatickou konzistenci může někdy paradoxně ztížit správnou interpretaci významových vztahů v textu.

3.3.1.3 Srovnání výsledků s hypotézami a předpoklady

Testování odhalilo několik významných rozdílů výsledků proti očekávání, zejména v oblasti výsledků modelů na adaptovaném přeloženém datasetu. Zatímco hypotéza správně předpokládala dominantní postavení modelu Claude 3.5 Sonnet, jeho skutečná výkonnost

výrazně předčila očekávání. Namísto předpokládaného intervalu 75 – 80 % na českém adaptovaném datasetu dosáhl model přesnosti 88,63 %, což představuje významné překročení původních odhadů. Tento výsledek naznačuje, že schopnost modelu zpracovávat koreferenční vztahy v českém jazyce byla podhodnocena.

Významným překvapením bylo zjištění týkající se vztahu mezi původním anglickým datasetem a jeho českými verzemi. Původní hypotéza předpokládala nižší výkonnost modelů na českých datasetech o 5 – 10 %, což se ukázalo jako nesprávné. Adaptovaný český dataset ve skutečnosti vedl k mírnému zlepšení výkonu napříč všemi modely, což bylo potvrzeno statistickou analýzou. Tento výsledek představuje zajímavý kontrast s předchozími studiemi provedenými na jiných jazycích, kde překlad typicky vedl ke zhoršení výkonu. Lze předpokládat, že tento rozdíl je částečně způsoben specifickými vlastnostmi českého jazyka, zejména systematictějším využitím kontextové informace.

Předpoklad týkající se konzistence výkonu napříč různými typy vět se ukázal jako částečně správný, nicméně s důležitým rozdílem v případě modelu Llama 3.2-3b. Zatímco hypotéza předpokládala variabilitu výkonu v závislosti na typu úlohy, tento model prokázal pozoruhodně konzistentní, byť nižší, výkon napříč všemi testovanými variantami. Toto zjištění podporuje teorii o existenci výkonnostního prahu souvisejícího s velikostí modelu, která nebyla v původních hypotézách zvažována. Současně se potvrdil předpoklad o problematičnosti strojového překladu, který skutečně vedl k horším výsledkům napříč všemi modely, což podtrhuje význam kvalitní jazykové adaptace pro evaluační datasety.

Závěr

Hlavním předmětem této práce byla otázka spolehlivosti strojového překladu úlohy Winograd Schema Challenge z benchmarku SuperGLUE ve srovnání s manuálně adaptovanou (přeloženou) verzí při hodnocení jazykových kompetencí velkých jazykových modelů v českém jazyce. V době rostoucího vlivu generativní umělé inteligence na celkovou populaci roste také nutnost spolehlivého hodnocení dovedností těchto modelů pracovat s jinými jazyky než angličtinou. Zatímco se většina existujících výzkumů v oblasti benchmarkingu AI pro český jazyk zaměřuje na vývoj nových benchmarků nebo jejich přímý překlad, tato práce zkoumá možnost jejich adaptace prostřednictvím lidských překladatelů a anotátorů.

Úkol WSC byl zvolen ze dvou hlavních důvodů: jednak pro zásadní význam koreferenčních vztahů při porozumění textu v jazycích, které koreferenci využívají, a jednak pro možnost důkladně prozkoumat vliv překladového procesu na původní testovací záměr a jeho efektivitu. Zpracování datasetu proběhlo ve dvou fázích: první využívala strojový překlad pomocí služby DeepL, zatímco druhá spočívala v manuální adaptaci zohledňující lingvistická specifika češtiny. Zvolený metodologický přístup umožnil nejen vyhodnotit efektivitu obou překladových metod, ale také identifikoval klíčové jazykové výzvy při převodu do češtiny, především v oblasti zájmených vazeb a přirozenosti vyjadřování.

Výsledky této studie přinášejí několik významných poznatků pro metodologii hodnocení jazykových modelů. Především se ukazuje, že přímý překlad existujících benchmarků, byť provedený pomocí pokročilého překladového nástroje, nemusí být optimální cestou pro hodnocení jazykových kompetencí modelů v cílovém jazyce. Testování ukázalo, že kvalita a způsob překladu mají přímý dopad na to, jak modely v testech uspějí, přičemž adaptovaný překlad, který zohledňuje lingvistická specifika cílového jazyka, vede k odlišným výsledkům než překlad strojový.

Z širší perspektivy tato zjištění naznačují potřebu přehodnotit přístup k vícejazyčnému benchmarkingu. Ačkoliv strojový překlad představuje zdánlivě efektivní způsob rychlé adaptace testovacích sad, výsledky odhalují významná rizika spojená s přehlížením jazykově specifických aspektů. To je zvláště patrné v případě češtiny, kde komplexní systém skloňování, časování a práce se zájmeny vytváří pro LLM náročnou výzvu v porozumění textu. Tyto jazykové nuance se při strojovém překladu často vytrácejí nebo deformují, což významně zpochybňuje spolehlivost a platnost testování.

Výsledky také zdůrazňují potřebu vyvinout specifické přístupy k testování pro příslušné jazyky. Zatímco některé aspekty jazykového porozumění mohou být univerzální, způsob jejich

testování by měl reflektovat charakteristiku daného jazyka. V případě češtiny se například ukázalo, že správná práce s přivlastňovacími a zvratnými zájmeny představuje pro jazykové modely výzvu, kterou není možné adekvátně testovat pomocí strojově přeloženého benchmarku. Uvedená zjištění zdůrazňují nezbytnost vývoje specializovaných testovacích sad, které by důsledně reflektovaly gramatická a stylistická specifika jednotlivých jazyků. Případně poukazují na potenciál hybridních přístupů, které by účelně propojovaly přeložené testovací úlohy s těmi vytvořenými přímo v cílovém jazyce.

Výsledky testování prezentované v této diplomové práci však podléhají několika zásadním omezením, která je třeba zohlednit při interpretaci závěrů. V první řadě se jedná o problém nekonzistentního překladu zájmen napříč testovacími daty. Zatímco v angličtině je systém zájmen relativně přímočarý, čeština operuje s komplexním systémem osobních, přivlastňovacích, zvratných a dalších typů zájmen, jejichž použití se řídí komplikovanými gramatickými zákonitostmi. Strojový překlad selhává v zachování konzistence zájmených vztahů, což se projevuje nejednotným převodem totožných anglických zájmen napříč datasetem.

Významným omezením je také problematická převoditelnost některých anglických konstrukcí do češtiny. Jde především o případy, kdy anglická věta využívá jazykové prostředky, které v češtině buď neexistují, nebo se vyjadřují zcela odlišným způsobem. Typickým příkladem jsou pasivní slovesné vazby, které jsou v angličtině běžné a v původním datasetu se často vyskytují, ale v češtině znějí nepřirozeně a často vyžadují kompletní přestavbu větné struktury. Nutnost takového přeformulování může vést jak ke zvýšení, tak ke snížení obtížnosti testované úlohy, a zároveň posunout její původní záměr a zaměření.

Za problematické lze považovat i kombinování několika odlišných jazykových jevů v rámci jedné testovací sady. Stávající formát WSC vytváří testovací věty, které často kombinují několik jevů najednou – jediná věta může současně testovat koreferenční vztahy, logické odvozování i kontextuální porozumění. Zatímco v původním anglickém znění tyto různé aspekty mohou fungovat koherentně, při překladu do češtiny se mohou vzájemně ovlivňovat způsobem, který komplikuje interpretaci výsledků. Tento problém lze dobře ilustrovat na konkrétních příkladech z testovaného datasetu.

Lze uvést například větu „The lawyer consulted her client because she needed clarification“ a její navržený překlad „Právníčka konzultovala situaci se svou klientkou, protože potřebovala vyjasnění.“ V angličtině tato věta primárně zkoumá koreferenční vztahy, především to, ke které osobě se vztahuje zájmeno „she“. Český překlad však s sebou přináší

i správné použití zvrtných zájmen („svou“ nebo „její“) nebo shodu podmětu s přísudkem v minulém čase, což vytváří dodatečné jazykové vrstvy, které mohou ovlivnit způsob, jakým model větu zpracovává. Zároveň do problematiky vstupuje i rod, jelikož v původní anglické větě může být „client“ osoba mužského i ženského pohlaví, zatímco v českém překladu je specifikován ženský rod použitím femininních tvarů „klientka“ a „právnička“. Tato rozdílná míra specifikace genderu mezi jazyky tak vytváří odlišnou výchozí pozici pro řešení koreferenčních vztahů, především ve vztahu k identifikaci referenta zájmena „her“.

Další příklad představuje věta „The trophy didn't fit in the suitcase because it was too big“, přeložená jako „Trofej se nevešla do kufru, protože byla příliš velká“. V angličtině věta testuje především logické odvození (co bylo příliš velké – trofej nebo kufr?). Při překladu do českého jazyka bylo třeba vyřešit opět i gramatický rod (shoda přísudku jmenného „byla velká“ se sponou) a navíc i vlastní zvrtné sloveso („vešla se“), což přidává další vrstvu komplexity a mění původně jednoduchý test logického uvažování na komplexnější lingvistickou úlohu. Proto byla věta v adaptovaném datasetu upravena na „Trofej se do hnědé tašky nevejde, ona je příliš velká“, čímž se podařilo vrátit k původnímu záměru testovat především logické odvození.

Kromě třech uvedených zásadních limitů je potřeba přihlédnout i k dalším významným faktorům, které zpochybňují důvěryhodnost či spolehlivost výsledků současného testování. Mezi ně patří zejména absence testování specificky českých jazykových jevů, například slovesného vidu nebo významotvorné funkce předpon. Slovesný vid by v kontextu WSC představoval významnou sémantickou rovinu, která by zásadně ovlivnila identifikaci koreferenčních vztahů – v některých případech může jejich určení usnadnit, v jiných naopak přidat další vrstvu komplexity při jejich analýze. Příklad:

- a. „Programátor volal správci systému. On nedokázal vyřešit ten problém.“
- b. „Programátor zavolal správci systému. On nedokázal vyřešit ten problém.“

V první větě nedokonavý vid slova „volal“ implikuje delší trvání nebo opakování děje, což může nepřímou napovídat, že „nedokázal“ se vztahuje spíše k programátorovi (který opakovaně potřeboval pomoc od správce systému). Naopak dokonavý vid slovesa „zavolal“ představuje jednorázovou akci, což může změnit interpretační kontext celé věty. Zmíněné významotvorné předpony by mohly do WSC vnést další významovou komplexitu. Příklad:

- „Jan přepsal Petrovi text, protože mu nevyhovoval.“
- „Jan napsal Petrovi text, protože mu nevyhovoval.“

Tvar slovesa s předponou „pře-“ naznačuje existenci předchozí verze textu, což může ovlivnit interpretaci zájmena „mu“ – v první větě je pravděpodobnější, že text nevyhovoval Petrovi (proto bylo třeba ho přepsat), zatímco v druhé větě je interpretace více otevřená. Nejen tyto příklady ukazují, že stejná základní struktura testovacích úloh WSC může mít v češtině více významových rovin než angličtina. Proto by bylo vhodné při vytváření českých WSC datasetů systematicky zohledňovat tato jazyková specifika a jejich vliv na rozhodnutí. Provedené testování odhalilo určité problémy spojené s překladem a zpracováním českých jazykových specifik, zároveň však ukazuje potřebu vytvořit nový test, který bude systematicky zohledňovat výše popsaná jazyková specifika a implementovat navržená doporučení.

Další významná limitace této studie spočívá v charakteristice vybraných jazykových modelů a jejich trénovacích dat. Všechny testované modely byly primárně trénovány na anglických datech, s českou jazykovou složkou představující pouze menší část jejich trénovacího korpusu. Tato skutečnost významně ovlivňuje interpretaci výsledků, zejména ve vztahu k pozorovanému zlepšení výkonu na českém datasetu. Je možné, že modely specificky trénované na slovanských jazycích by mohly vykazovat odlišné vzorce analýzy významu a potenciálně dosáhnout ještě lepších výsledků díky hlubšímu porozumění morfologickým a syntaktickým specifikům těchto jazyků. Tuto hypotézu podporuje i skutečnost, že existující česky trénované modely jako RobeCzech nebo Czert často vykazují lepší výkonnost v úlohách specifických pro český jazyk – RobeCzech například dosahuje nadprůměrných výsledků v úlohách pojmového rozpoznávání entit a analýzy sentimentu, což naznačuje jeho výhodu v úzce zaměřených českých kontextech (Straka et al., 2021). Tyto české modely jsou totiž založeny na architektuře BERT s kodérem, který může analyzovat každé slovo v kontextu slov předcházejících i následujících, zatímco testované velké jazykové modely používají jednosměrné zpracování textu určené pro generativní modely, což činí přímé porovnání jejich výkonu v rámci této práce metodologicky nevhodné.

Na závěr je vhodné shrnout doporučení pro další podobná testování či výzkumy. Pro získání relevantnějších výsledků s výpovědní hodnotou o dovednostech modelů v českém jazyce by bylo třeba vytvořit kompletně nový dataset WSC úloh přímo v češtině, který bude cíleně testovat specifické české jazykové jevy. Pro potřeby tohoto datasetu by bylo vhodné obohatit testovací otázky o kategorizaci testovacích případů podle typu jazykového jevu. Zvláštní pozornost by měla být zaměřena na případy využívající specifické prostředky českého jazyka pro vyjádření vztahů mezi entitami, zejména v oblasti interference a koreference, a to jak na úrovni jednotlivých vět, tak v rámci delších textových celků. V případě manuálního

či strojového překladu by pak bylo také třeba standardizovat proces tak, aby byly zachovány klíčové významové nuance při převodu mezi jazyky.

Výsledky testování ukázaly, že adaptovaný překlad vede k lepším výsledkům než strojový překlad, přičemž rozdíly byly statisticky významné ($p < 0,05$). Nejvýraznější zlepšení bylo pozorováno u modelu Claude-3.5-sonnet, který dosáhl přesnosti 88,63 % na adaptovaném datasetu oproti 72,20 % na strojově přeloženém.

Hlavními přínosy této práce jsou:

- systematická analýza vlivu kvality překladu na výkonnost jazykových modelů v úloze Winograd Schema Challenge,
- identifikace specifických jazykových jevů v češtině, které ovlivňují správnost určování koreferenčních vztahů,
- vytvoření metodologie pro adaptaci anglických benchmarků do českého jazyka,
- ověření významu lingvisticky informované adaptace oproti přímému strojovému překladu.

S ohledem na budoucnost lingvistického hodnocení umělé inteligence tato práce naznačuje několik směrů dalšího vývoje. Především je zřejmá potřeba vytvoření dalších specializovaných českých datasetů, které budou cíleně testovat jazykovou zdatnost a porozumění jazyku. Zároveň se ukazuje význam systematického přístupu k překladu a adaptaci existujících benchmarků, který by měl zohledňovat nejen lingvistické, ale i kulturní aspekty cílového jazyka. V neposlední řadě výsledky naznačují potřebu vývoje standardizovaných metodik pro evaluaci vícejazyčných schopností velkých jazykových modelů, které budou schopny zachytit nuance jednotlivých jazyků při současném zachování možnosti mezinárodního srovnání.

Seznam použité literatury a dalších zdrojů

- Argamon, S., Dagan, I., & Krymolowski, Y. (1998). A Memory-Based Approach to Learning Shallow Natural Language Patterns. CoRR, cmp-lg/9806011. <http://arxiv.org/abs/cmp-lg/9806011>
- Barančíková, P., & Bojar, O. (2020). COSTRA 1.0: A Dataset of Complex Sentence Transformations. Proceedings of the Twelfth Language Resources and Evaluation Conference, 3535-3541. <https://aclanthology.org/2020.lrec-1.434/>
- Barančíková, P., & Bojar, O. (2020). COSTRA 1.1: A Dataset of Complex Sentence Transformations and Comparis. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. <http://hdl.handle.net/11234/1-3248>
- Barančíková, P., & Náplava, J. (2024). Seznam představuje nové jazykové modely pro práci s významem textů. Blog Seznam.cz. <https://blog.seznam.cz/2024/04/seznam-predstavuje-nove-jazykove-modely-pro-praci-s-vyznamem-textu/>
- Bednář, J., Náplava, J., Barančíková, P., & Lisický, O. (2024). Some Like It Small: Czech Semantic Embedding Models for Industry Applications. Proceedings of the AAAI Conference on Artificial Intelligence, 38(21), 22734-22742. <https://arxiv.org/abs/2311.13921>
- BIG-authors. (2023). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research. <https://openreview.net/forum?id=uyTL5Bvosj>
- Bobrow, D. G. (1964). Natural language input for a computer problem solving system [Doctoral dissertation, Massachusetts Institute of Technology]. <https://dspace.mit.edu/handle/1721.1/145210>
- Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19(2), 263 - 31. <https://aclanthology.org/J93-2003.pdf>
- Çano, E., & Bojar, O. (2019). Sentiment Analysis of Czech Texts: An Algorithmic Survey. Proceedings of the 11th International Conference on Agents and Artificial Intelligence, 973-979. <https://doi.org/10.5220/0007695709730979>
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., et al. (2022). PaLM: Scaling Language Modeling with Pathways: Scaling Language Modeling with Pathways. <https://arxiv.org/abs/2204.02311>
- Citron, D. (2024, August 28). New in Gemini: Custom Gems and improved image generation with Imagen 3. Google. <https://blog.google/products/gemini/google-gemini-update-august-2024/>
- Cottier, B., Besiroglu, T., & Owen, D. (2023). Who is leading in AI? An analysis of industry AI research. <https://arxiv.org/abs/2312.00043>
- Dai, W., & Berleant, D. (2019). Benchmarking Contemporary Deep Learning Hardware and Frameworks: A Survey of Qualitative Metrics. IEEE First International Conference on Cognitive Machine Intelligence, 148-155. <https://ieeexplore.ieee.org/document/8998988>

- DeepL. (2024). DeepL is 2024's Most-Used Machine Translation Provider Worldwide among language service companies. <https://www.prnewswire.com/news-releases/deepl-is-2024s-most-used-mach>
- Fajčík, M., Dočekal, M., Doležal, J., Ondřej, K., Beneš, K., Kapsa, J., Hradiš, M., Nevěřilová, Z., Horák, A., Štefánek, M., Jirkovský, A., Adamczyk, D., Hula, J., Šedivý, J., & Kydlíček, H. (2024). BenCzechMark: A Czech-centric Multitask and Multimetric Benchmark for Language Models with Duel Scoring Mechanism. Brno University of Technology, Masaryk University, Czech Technical University in Prague, Hugging Face. <https://huggingface.co/spaces/CZLC/BenCzechMark>
- Gardner, M., Grus, J., Neumann, M., Tafford, O., Dasigi, P., Liu, N., Peters, M., Schmitz, M., & Zettlemoyer, L. (2018). AllenNLP: A Deep Semantic Natural Language Processing Platform. <https://arxiv.org/abs/1803.07640>
- Gershgorn, D. (2021, June 29). GitHub and OpenAI launch a new AI tool that generates its own code. The Verge. <https://www.theverge.com/2021/6/29/22555777/github-openai-ai-tool-autocomplete-code>
- Goodwin, G. E., & Jackson, S. (2024, November 29). How to use Bing: Microsoft reinvented its search engine to incorporate AI. Business Insider. <https://www.businessinsider.com/microsoft-bing>
- Harisudhan S. (2024, February 26). Unveiling the Power of Gemini: A New Era in Large Language Models. Medium. <https://medium.com/@speaktoharisudhan/the-emergence-of-gemini-google-deepminds-amazing-family-of-large-language-models-llms-marked-a-f75d5a755f46>
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring Massive Multitask Language Understanding. Proceedings of the International Conference on Learning Representations (ICLR). <https://arxiv.org/abs/2009.03300>
- Hsiao, S. (2024, May 14). Get more done with Gemini: Try 1.5 Pro and more intelligent features. Google. <https://blog.google/products/gemini/google-gemini-update-may-2024/>
- Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. Proceedings of the 37th International Conference on Machine Learning, 119, 4411-4421. <https://proceedings.mlr.press/v119/hu20b.html>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. <https://arxiv.org/abs/2001.08361>
- Kamalov, F., Santandreu Calonge, D., & Gurrib, I. (2023). New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution. Sustainability, 15(16). <https://doi.org/10.3390/su151612451>
- Karlík, P., Nekula, M., & Pleskalová, J. (2017). ANTECEDENT. CzechEncy - Nový encyklopedický slovník češtiny. <https://www.czechency.org/slovník/ANTECEDENT>
- Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S., & Roth, D. (2018). Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. Proceedings of NAACL-HLT 2018, 252-262. <https://doi.org/10.18653/v1/N18-1023>
- Kocián, M., Náplava, J., Stancl, D., & Kadlec, V. (2021). Siamese BERT-based Model for Web Search Relevance Ranking Evaluated on. Computing Research Repository. <https://arxiv.org/abs/2112.01810>

- Kocijan, V., Cretu, A. -M., Camburu, O. -M., Yordanov, Y., & Lukasiewicz, T. (2019). A Surprisingly Robust Trick for the Winograd Schema Challenge. *Proceedings of the 57th Annual Meeting of the Association for Computat*, 4837-4842. <https://doi.org/10.18653/v1/P19-1478>
- Kovář, D., & Nečas, J. (2024, March 28). Jak si poradí jazykové modely AI s českým právem? Složí advokátní zkoušky? HAVEL & PARTNERS. <https://www.havelpartners.blog/jak-si-poradi-jazykove-modely-ai-s-ceskym-pravem-slozi-advokatni-zkousky>
- Král, P., & Lenc, L. (2017). Czech Text Document Corpus v 2.0. *Computing Research Repository*, abs/1710.02365. <http://arxiv.org/abs/1710.02365>
- Kruppa, M. (2023, December 6). Google Announces AI System Gemini After Turmoil at Rival OpenAI. *WSJ*. <https://www.wsj.com/tech/ai/google-announces-ai-system-gemini-after-turmoil-at-rival-openai-10835335>
- Lenett, R. (2024, November 17). Google's Gemini AI Chatbot Faces Backlash After Inappropriate Responses. *MashViral*. <https://mashviral.com/googles-gemini-ai-chatbot-responses/>
- Levesque, H. J. (2011). The Winograd Schema Challenge. *Logical Formalizations of Commonsense Reasoning — Papers from the AAAI 2011 Spring Symposium*, 63-68. <https://cdn.aaai.org/ocs/2502/2502-10882-1-PB.pdf>
- Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., et al. (2023). Holistic Evaluation of Language Models. <https://arxiv.org/abs/2211.09110>
- Li, J., Zhou, H., Huang, S., Cheng, S., & Chen, J. (2024). Eliciting the Translation Ability of Large Language Models via Multili. *Transactions of the Association for Computational Linguistics*, 12, 576-592. https://doi.org/10.1162/tac1_a_00655
- Llama Team. (2025). The Llama 3 Herd of Models. <https://arxiv.org/abs/2407.21783>
- Manning, C. D. (2015). Computational Linguistics and Deep Learning. *Computational Linguistics*, 41(4), 701-707. https://doi.org/10.1162/COLI_a_00239
- Martindale, J. (2024, October 14). Unlocking Llama 3's Potential: What You Need to Know. *Lifewire*. <https://www.lifewire.com/what-to-know-llama-3-8713943>
- Melo, G. D., Imaizumi, V., & Cozman, F. (2019). Winograd Schemas in Portuguese. *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2019)*, 787-798. <https://doi.org/10.5753/eniac.2019.9334>
- Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39 - 41. <https://dl.acm.org/doi/abs/10.1145/219717.219748>
- Nangia, N., & Bowman, S. R. (2019). Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4566-4575. <https://doi.org/10.18653/v1/P19-1449>
- Natale, S. (2021). The ELIZA Effect: Joseph Weizenbaum and the Emergence of Chatbots: Joseph Weizenbaum and the Emergence of Chatbots. In *Deceitful Media: Artificial Intelligence and Social Life after the Tur*. Oxford University Press. <https://doi.org/10.1093/oso/9780190080365.003.0004>

- Nield, D. (2024, November 26). How to use Google Gemini on your phone. The Verge. <https://www.theverge.com/24305656/gemini-google-assistant-android-how-to>
- Niklaus, J., Matoshi, V., Rani, P., Galassi, A., Stürmer, M., & Chalkidis, I. (2023). LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. In (pp. 3016-3054). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-emnlp.200>
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., et al. (2024). GPT-4 Technical Report. <https://arxiv.org/abs/2303.08774>
- Pasberg, T. (2024, September 27). Why does Meta restrict the usage of Llama3.2 in the EU? Medium. <https://medium.com/@thomas-pasberg/why-does-meta-restrict-the-usage-of-llama3-2-in-the-eu-4079946abb07>
- Paszke, A., Chintala, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic Differentiation in PyTorch. NIPS 2017 Workshop on Autodiff, 1-4. <https://openreview.net/forum?id=BJJsrnfCZ>
- Paul, K. (2024, August 29). Meta says its Llama AI models being used by banks, tech companies. Reuters. <https://www.reuters.com/technology/artificial-intelligence/meta-says-its-llama-ai-models-being-used-by-banks-tech-companies-2024-08-29/>
- Pichai, S., & Hassabis, D. (2023, December 6). Introducing Gemini: our largest and most capable AI model. Google. <https://blog.google/technology/ai/google-gemini-ai/>
- Pilehvar, M. T., & Camacho-Collados, J. (2019). WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Mean: the Word-in-Context Dataset for Evaluating Context-Sensitive Mean. Proceedings of the 2019 Conference of the North American Chapter of the, 1267-1273. <https://doi.org/10.18653/v1/N19-1128>
- Plaza, I., Melero, N., del Pozo, C., Conde, J., Reviriego, P., Mayor-Rocher, M., & Grandury, M. (2024). Spanish and LLM Benchmarks: is MMLU Lost in Translation?, 1 - 2. <https://doi.org/10.48550/arXiv.2406.17789>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8). https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2383-2392. <https://doi.org/10.18653/v1/D16-1264>
- Robertson, A. (2024, February 21). Google apologizes for ‘missing the mark’ after Gemini generated racially diverse Nazis. The Verge. <https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical>
- Roemmele, M., Bejan, C., & Gordon, A. (2011). Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. AAAI Spring Symposium: Technical Report, -. https://www.researchgate.net/publication/221251392_Choice_of_Plausible_Alternatives_An_Evaluation_of_Commonsense_Causal_Reasoning

- Sajid, H. (2023). Can You Build Large Language Models Like ChatGPT At Half Cost? Unite.AI. <https://www.unite.ai/can-you-build-large-language-models-like-chatgpt-at-half-cost/>
- Schank, R. C., & Tesler, L. (1969). A conceptual dependency parser for natural language. COLING '69: Proceedings of the 1969 conference on Computational linguistics, 1-3. <https://dl.acm.org/doi/pdf/10.3115/990403.990405>
- Schmid, P., Sanseviero, O., Cuenca, P., Belkada, Y., & Werra, L. (2024). Welcome Llama 3 - Meta's new open LLM. <https://huggingface.co/blog/llama3>
- Searle, J. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–457. <https://doi.org/10.1017/S0140525X00005756>
- Shavrina, T., Fenogenova, A., Emelyanov, A. A., Shevelev, D., Artemova, E., Malykh, V., Mikhailov, V., Tikhonova, M., Chertok, A., & Evlampiev, A. (2020). Russian SuperGLUE: A Russian Language Understanding Evaluation Benchmark: A Russian Language Understanding Evaluation Benchmark. CoRR, abs/2010.15925. <https://arxiv.org/abs/2010.15925>
- Sido, J., Seják, M., Prazák, O., Konopík, M., & Moravec, V. (2021). Czech News Dataset for Semantic Textual Similarity. Computing Research Repository, abs/2108.08708. <https://arxiv.org/abs/2108.08708>
- Singh, S. (2024, January 6). Perplexity AI raises \$73.6M in funding round led by Nvidia, Bezos, now valued at \$522M. Business Today. <https://www.businesstoday.in/entrepreneurship/start-up/story/perplexity-ai-raises-736m-in-funding-round-led-by-nvidia-bezos-now-valued-at-522m-412130-2024-01-06>
- Singh, S. (2024, September 2). ChatGPT Statistics (DEC. 2024) – 200 Million Active Users. DemandSage. <https://www.demandsage.com/chatgpt-statistics/>
- Soukup, P. (2013). Substantive significance and it's measures. Data and Research – SDA Info, 127(2). <https://doi.org/10.13060/23362391.2013.127.2.41>
- Straka, M., Náplava, J., Straková, J., & Samuel, D. (2021). RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model: Czech RoBERTa, a Monolingual Contextualized Language Representation Model. In (pp. 197-209). Springer International Publishing.
- Thoppilan, R., Freitas, D. D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. -T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y. G., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., et al. (2022). LaMDA: Language Models for Dialog Applications: Language Models for Dialog Applications. CoRR, abs/2201.08239. <https://arxiv.org/abs/2201.08239>
- Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. Mind, LIX(236), 433-460. <https://doi.org/10.1093/mind/LIX.236.433>
- Turner-trauring, I. (2022). Faster, more memory-efficient Python JSON parsing with msgspec. Python⇒Speed. <https://pythonspeed.com/articles/faster-python-json-parsing/>
- Vaswani, A., Kaiser, L., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., N. Gomez, A., & Polosukhin, I. (2017). Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017). <https://arxiv.org/abs/1706.03762>

- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Levy, O., Michael, J., & Hill, F. (2019). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Advances in neural information processing systems*, 32. <https://arxiv.org/abs/1905.00537>
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353-355. <https://doi.org/10.18653/v1/W18-5446>
- Weber, A. A., Thellmann, K., Ebert, J., Flores-Herr, N., Lehmann, J., Fromm, M., & Ali, M. (2024). Investigating Multilingual Instruction-Tuning: Do Polyglot Models Dema: Do Polyglot Models Dema. In (pp. 20829-20855). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.1159>
- Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. <https://dl.acm.org/doi/10.1145/365153.365168>
- Wiggers, K. (2023, March 14). Anthropic launches Claude, a chatbot to rival OpenAI's ChatGPT. *TechCrunch*. <https://techcrunch.com/2023/03/14/anthropic-launches-claude-a-chatbot-to-rival-openais-chatgpt/>
- Wiggers, K. (2023, June 6). OpenAI makes GPT-4 generally available. *TechCrunch*. <https://techcrunch.com/2023/07/06/openai-makes-gpt-4-generally-available/>
- Winograd, T. (1971). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. <https://dspace.mit.edu/handle/1721.1/7095>
- Woods, W. A. (1970). Transition network grammars for natural language analysis. *Communications of the ACM*, 13(10), 591 - 606. <https://doi.org/10.1145/355598.362773>
- Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2024). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6), 1-32. <https://arxiv.org/abs/2304.13712>
- Zhuang, L., Wayne, L., Ya, S., & Jun, Z. (2021). A Robustly Optimized BERT Pre-training Approach with Post-training. *Proceedings of the 20th Chinese National Conference on Computational L*, 1218-1227. <https://aclanthology.org/2021.ccl-1.108>
- Zhu, K. (2024, October 11). Ranked: The Most Popular Generative AI Tools in 2024. *Visual Capitalist*. <https://www.visualcapitalist.com/ranked-the-most-popular-generative-ai-tools-in-2024/>
- Žagar, A., & Robnik-Šikonja, M. (2022). Slovene SuperGLUE Benchmark: Translation and Evaluation. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2058-2065. <https://aclanthology.org/2022.lrec-1.221>
- Benchmark, n., sense 2.b. (2024). Oxford University Press. <https://doi.org/10.1093/OED/4314330287>
- Benchmarking, n., sense 2. (2023). Oxford University Press. <https://doi.org/10.1093/OED/5100862715>
- Claude 2. (2023, July 11). <https://www.anthropic.com/news/claude-2>
- Claude 3.5 Sonnet. (2024, June 21). <https://www.anthropic.com/news/claude-3-5-sonnet>

Code Llama, a state-of-the-art large language model for coding. Meta Llama.

<https://www.llama.com/code-llama/>

GitHub. (2024). <https://github.com/globalwordnet/english-wordnet/releases/tag/2024-edition>

GPT-4o mini: advancing cost-efficient intelligence. (2024, July 18). OpenAI.

<https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>

IBM Watson Natural Language Understanding. IBM. <https://www.ibm.com/products/natural-language-understanding>

Introducing ChatGPT Pro. OpenAI. <https://openai.com/index/introducing-chatgpt-pro/>

Introducing Claude 2.1. (2023, November 21). <https://www.anthropic.com/news/claude-2-1>

Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku. (2024, October 22).

<https://www.anthropic.com/news/3-5-models-and-computer-use>

Introducing GPT-4o and more tools to ChatGPT free users. (2024, May 13). OpenAI.

<https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/>

Introducing OpenAI o1-preview. (2024). OpenAI. <https://openai.com/index/introducing-openai-o1-preview/>

Introducing the Claude Team plan and iOS app. (2024, May 1).

<https://www.anthropic.com/news/team-plan-and-ios>

Introducing the next generation of Claude. (2024). <https://www.anthropic.com/news/claude-3-family>

JSON Lines. <https://jsonlines.org>

Llama.cpp vs. ollama: Running LLMs Locally for Enterprises. (2024, April 13). Picovoice.

<https://picovoice.ai/blog/local-llms-llamacpp-ollama/>

Meta. (2024, April 18). Meet Your New Assistant: Meta AI, Built with Llama 3. Meta.

<https://about.fb.com/news/2024/04/meta-ai-assistant-built-with-llama-3/>

Meta. (2024, September 25). Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models.

<https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>

Meta Llama. (2024). Hugging Face. Retrieved December 14, 2024, from <https://huggingface.co/meta-llama>

Natural Language Understanding (NLU). (2024). Deepgram. <https://deepgram.com/ai-glossary/natural-language-understanding>

Open English Wordnet 2024. (2024). GitHub. <https://github.com/globalwordnet/english-wordnet/releases/tag/2024-edition>

Poe. About. <https://poe.com/about>

Reuters. (2024, August 29). OpenAI says ChatGPT's weekly users have grown to 200 million.

Reuters. <https://www.reuters.com/technology/artificial-intelligence/openai-says-chatgpts-weekly-users-have-grown-200-million-2024-08-29/>

SuperGLUE Tasks. (2019). SuperGLUE Benchmark. <https://super.gluebenchmark.com/tasks>

Glosář

Accuracy (Přesnost)	Statistická metrika vyjadřující podíl správně klasifikovaných případů ze všech klasifikovaných případů.
AI alignment (sladění)	Snaha zajistit, aby systémy umělé inteligence jednaly v souladu s lidskými hodnotami a záměry. Tento obor výzkumu se zabývá tím, jak vytvářet AI systémy, které jsou bezpečné, etické a prospěšné pro lidstvo.
Anaforický odkaz	Jazykový prostředek zpětného odkazu v textu, typicky realizovaný zájmeny nebo jinými zástupnými výrazy, který odkazuje na dříve zmíněnou entitu.
API (Application Programming Interface)	Soubor protokolů a nástrojů umožňujících komunikaci mezi různými softwarovými aplikacemi, např. pro dotazování jazykových modelů.
API klíč	Tajný token sloužící k ověřování a autorizaci přístupu k rozhraní API.
ATN (Augmented Transition Network)	Výpočetní model pro rozbor vět, který reprezentuje gramatickou strukturu textu.
Azure Translator	Nástroj pro strojový překlad od společnosti Microsoft, který slouží k převodu textu mezi více jazyky.
Benchmarking	Činnost nebo proces měření výkonu hardwaru nebo softwaru pomocí standardizovaného nebo speciálně vybraného softwaru, datasetů a dalších podkladů.
BIG-bench (Beyond the Imitation Game Benchmark)	Srovnávací test, který testuje hranice jazykových modelů s různými kreativními nebo abstraktními úlohami.
BoolQ	Úloha s binárními odpověďmi na otázky, kde modely odpovídají na otázky typu ano/ne na základě zadaného úryvku textu.
Byte-Pair Encoding (BPE)	Algoritmus tokenizace, který kombinuje často se vyskytující dvojice znaků za účelem vytvoření tokenů.
CommitmentBank (CB)	Soubor vět testujících míru jistoty mluvčích o pravdivosti výroků s označením „entailment“, „contradiction“ nebo „neutral“.
COPA (Choice of Plausible Alternatives)	Úloha hodnotící kauzální uvažování, v níž modely vybírají nejpravděpodobnější příčinu nebo následek ze dvou možností.
Dataset	Dataset je organizovaná sbírka dat, která slouží ke konkrétnímu účelu. Typicky se jedná o soubor souvisejících záznamů nebo informací uspořádaných v určité struktuře, který se používá pro strojové učení, analýzu nebo výzkum.

Embedding	Číselná reprezentace slov nebo frází zachycující jejich význam a vztahy ve vektorovém prostoru.
Few-shot examples	Technika pro zlepšení výkonnosti modelu zahrnutím několika příkladů správných odpovědí specifických pro daný úkol.
Fine-tuning	Proces přizpůsobení předem natrénovaného modelu k provádění konkrétních úloh jeho trénováním na datech specifických pro danou úlohu.
Generativní umělá inteligence	Typ umělé inteligence, která na základě vstupních dat generuje nový obsah, například text nebo obrázky.
GLUE (General Language Understanding Evaluation)	Srovnávací test pro hodnocení modelů porozumění přirozenému jazyku v různých úlohách.
Ground truth	Správná odpověď nebo značka v souboru dat, která se používá jako standard pro vyhodnocení předpovědi modelu.
HELM (Holistic Evaluation of Language Models)	Srovnávací hodnocení modelů z hlediska jazykové výkonnosti, etických aspektů a společenské spravedlnosti.
Human-baseline	Úroveň výkonu dosažená lidskými účastníky při řešení úloh, která se používá jako srovnávací metrika pro systémy umělé inteligence.
Human-level understanding	Schopnost umělé inteligence interpretovat a řešit úlohy na úrovni srovnatelné s lidským výkonem.
Kodér a dekodér	Architektura transformátorů, kde kodér zpracovává vstupní data a dekodér generuje výstup.
Kontextové okno	Maximální množství textu (tokenů), které může jazykový model zpracovat najednou.
Koreference	Vztah mezi dvěma či více jazykovými výrazy, které označují tutéž entitu v mimojazykovém světě, bez ohledu na použitý způsob označení nebo směr odkazu.
Kritická t-hodnota	Prahová hodnota používaná při testování hypotéz k určení, zda testovací statistika znamená statisticky významný výsledek.

Mechanismus pozornosti (Attention)	System dynamického přiřazování váhy různým částem vstupní sekvence, který umožňuje modelu soustředit se na relevantní informace při generování každého dalšího prvku výstupu.
Mechanistic Interpretability	Studijní obor zaměřený na pochopení vnitřního fungování modelů strojového učení, zejména na to, jak zpracovávají a reprezentují informace.
MMLU (Massive Multitask Language Understanding)	Srovnávací test určený k testování obecných znalostí a schopností uvažování jazykových modelů v různých oblastech a jazycích.
Multimodální jazykový model	Systémy umělé inteligence schopné zpracovávat více typů dat současně, například text, obrázky a zvuk.
MultiRC (Multi-Sentence Reading Comprehension)	Úloha zaměřená na čtení s porozuměním, která vyžaduje odpovědi na otázky založené na víceslovných textech, často s více správnými odpověďmi.
Multitask learning	Trénování modelů AI na více úlohách současně s cílem zlepšit celkový výkon a přenos znalostí.
NLP (Natural Language Processing)	Podobor umělé inteligence zaměřený na to, aby stroje dokázaly porozumět lidskému jazyku, interpretovat ho a vytvářet.
NLU (Natural Language Understanding)	Podkategorie NLP zaměřená na schopnost stroje porozumět lidskému jazyku a interpretovat jej.
Párový t-test	Statistická metoda používaná k porovnání dvou příbuzných skupin s cílem zjistit, zda se jejich střední hodnoty významně liší.
Parser	Nástroj používaný v počítačové lingvistice k analýze gramatické struktury vět.
Perplexita	Kvantitativní ukazatel nejistoty jazykového modelu při predikci následujícího prvku v sekvenci. Nižší hodnota značí větší jistotu modelu v předpovědi.
Perplexita na úrovni slov	Specifická varianta perplexity, která měří průměrnou nejistotu modelu při predikci následujícího slova v textu, na rozdíl od predikce jednotlivých znaků.

Prompt	Strukturované zadání, které obsahuje popisy úloh a příklady, jimiž se řídí odpovědi jazykového modelu.
Prompt engineering	Navrhování účinných vstupních podnětů pro vedení modelů umělé inteligence při generování požadovaných výstupů.
Přivlastňovací zájmeno	Zájmeno vyjadřující vlastnický vztah mezi vlastníkem a vlastněným předmětem či osobou.
Python	Programovací jazyk běžně používaný pro skriptování, zpracování dat a interakce s rozhraním API.
ReCoRD (Reading Comprehension with Commonsense Reasoning Dataset)	Soubor dat pro porozumění čtenému textu, který testuje rozumové uvažování a schopnost získávat informace přímo z textu.
RTE (Recognizing Textual Entailment)	Binární klasifikační úloha, jejímž cílem je určit, zda hypotéza logicky vyplývá z dané premisy.
Sample-efficient learning	Přístup k trénování umělé inteligence, který minimalizuje množství dat potřebných k dosažení dobrého výkonu.
Shallow parsing	Technika zpracování přirozeného jazyka, která identifikuje základní části věty (např. podstatná jména, slovesa) a jejich jednoduché vztahy.
Shallow pattern matching	Jednoduchá metoda vyhledávání a porovnávání vzorů v datech, která pracuje pouze s povrchovými charakteristikami bez hlubší analýzy významu či struktury. Typicky využívá například regulární výrazy nebo jednoduché textové shody.
Self-supervised learning	Učební paradigma, při kterém modely vytvářejí vlastní značky nebo cíle z neoznačených dat.
Semi-supervised learning	Učební paradigma, které kombinuje malé množství označených dat s velkým množstvím neoznačených dat během trénovacího procesu. Model se tak učí z obou typů dat najednou, což umožňuje dosáhnout lepších výsledků než při použití pouze označených dat.
Skript	Příkaz napsaný v programovacím jazyce, který je určen k automatizaci rutinních úkolů, manipulaci se soubory nebo řízení běhu jiných programů v rámci operačního systému.
SOTA (State-of-the-Art)	Nejvyšší výkon, kterého technologie nebo model v daném okamžiku dosahuje.

Statistická významnost	Kvantitativní ukazatel vyjadřující pravděpodobnost, že pozorovaný výsledek nebo rozdíl mezi skupinami není dílem náhody, ale představuje skutečný efekt. Obvykle se posuzuje na základě p-hodnoty, přičemž konvenční hranicí významnosti je $p < 0.05$.
Strojový překlad	Proces použití softwaru k překladu textu nebo řeči z jednoho jazyka do druhého.
SuperGLUE	Rámec pro hodnocení výkonnosti modelů porozumění jazyku pomocí souboru úloh.
Textový korpus	Rozsáhlá a strukturovaná sbírka textů sloužící k trénování a testování jazykových modelů.
Tokenizace	Proces rozdělení textu na menší jednotky (tokeny), které zpracovávají modely strojového učení.
Transferové učení	Proces, při kterém je předem natrénovaný model přizpůsoben k provedení nové úlohy pomocí dodatečných dat specifických pro danou úlohu. V angličtině jako „transfer learning“.
Transformátor	Model neuronové sítě, který zpracovává sekvenční data (např. jazyk) pomocí mechanismu napodobujícího pozornost.
Umělá inteligence (AI)	Simulace lidské inteligence ve strojích, které jsou určeny k provádění úkolů, jako je učení a řešení problémů.
Unsupervised learning	Typ strojového učení, při kterém se modely učí vzory z dat bez označených výsledků nebo cílů.
Velikost účinku (Cohenovo d)	Kvantitativní ukazatel síly vztahu nebo rozdílu mezi proměnnými, který je nezávislý na velikosti vzorku. Podle Cohena se hodnoty interpretují jako malý (0.2), střední (0.5) a velký (0.8) efekt.
LLM, Large Language Model	Velký jazykový model, typ modelu umělé inteligence vycvičený na velkých souborech dat k provádění úloh porozumění jazyku nebo generování textu.
WiC (Word-in-Context)	Úloha testující schopnost rozpoznat, zda má slovo stejný význam v různých kontextech.
WSC (Winograd Schema Challenge)	Úloha zaměřená na řešení nejednoznačných zájmených odkazů pomocí kontextového a rozumového uvažování.

WordPiece	Algoritmus tokenizace, který rozděluje slova na dílčí slovní jednotky na základě frekvence a prediktivní síly.
XTREME	Vícejazyčný benchmark testující schopnost mezijazykového přenosu v různých úlohách a jazycích.
Zero-shot learning	Schopnost strojového učení, kdy model řeší úlohy, na které nebyl explicitně vyškolen, využitím vzorů v trénovacích datech.
Zvratné zájmeno	Slovní druh vyjadřující, že předmět děje je totožný s jeho původcem. V češtině se rozlišuje osobní zvratné zájmeno „se“ a přivlastňovací zvratné zájmeno „svůj“. Tato zájmena nemají nominativní tvar a vztahují se k podmětu věty.

Přílohy

Přílohy součástí práce přiložené zvlášť

Příloha 1 – Python skript pro konverzi souboru z formátu .jsonl do .xlsx

- jsonl_to_xlsx_converter.txt - plain text v kódování UTF-8 (.txt), 1 kB

Příloha 2 - Python skript pro konverzi souboru z formátu .xlsx do .jsonl

- xlsx_to_jsonl_converter.txt - plain text v kódování UTF-8 (.txt), 1 kB

Příloha 3 - Python skript pro indexaci entit

- index_calculation.txt - plain text v kódování UTF-8 (.txt), 1 kB

Příloha 4 – Výsledky testování modelu Claude

- claude_results.csv - Comma-separated values (.csv), 275 kB

Příloha 5 – Výsledky testování modelu Gemini

- gemini_results.csv - Comma-separated values (.csv), 269 kB

Příloha 6 – Výsledky testování modelu GPT

- GPT_results.csv - Comma-separated values (.csv), 275 kB

Příloha 7 – Výsledky testování modelu Llama

- llama_results.csv - Comma-separated values (.csv), 276 kB

Příloha 8 – Adaptovaný dataset

- Adapted_test.csv - Comma-separated values (.csv), 81 kB

Příloha 9 – Strojově přeložený dataset

- ML_translated_test.csv - Comma-separated values (.csv), 79 kB

Příloha 10 – Originální dataset

- Original_test.csv - Comma-separated values (.csv), 82 kB

Příloha 11 – Skript pro testování modelu GPT

- gpt_winograd-test.txt - Plain text v kódování UTF-8 (.txt), 5 kB

Příloha 12 – Skript pro testování modelu Llama

- llama_winograd-test.txt - Plain text v kódování UTF-8 (.txt), 5 kB

Příloha 13 – Skript pro testování modelu Claude

- claude_winograd-test.txt - Plain text v kódování UTF-8 (.txt), 5 kB

Příloha 14 – Skript pro testování modelu Gemini

- gemini_winograd-test.txt - Plain text v kódování UTF-8 (.txt), 5 kB