Czech
Technical University
in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science

NLP Methods for Automated Fact-Checking

Dissertation Thesis of

Ing. Herbert Ullrich

FCHECK.FEL.CVUT.CZ

Supervisor: **Ing. Jan Drchal, Ph.D.**
Field of study: **Informatics**
Subfield: **Natural Language Processing**
August 2023

Contents

Figures

Tables

Chapter 1
Introduction

My dissertation, as well as my long-term research, centers on *automated fact-checking* using modern Natural Language Processing (NLP) methods. The work consists of the analysis of the whole fact-checking process, its subdivision, and simplification into tasks that can be efficiently addressed using the current state-of-the-art NLP methods, collection of data appropriate to benchmark such tasks, delivery of example solutions and their validation against similar research in other languages and related tasks.

Our group focuses primarily on fact-checking tasks in West Slavic languages (Czech, Slovak, Polish), and secondarily in English. So far, I have collected and published datasets for fact-checking and its subroutines, trained and analyzed models, and begun work on explainable metrics that capture *facticity*—a challenge in itself [Koto et al., 2020, Wright et al., 2022].

My doctoral aim is to cover the full path from obtaining a factual claim—for example, extracting it from a political debate—to predicting its veracity and justifying it rigorously with evidence. With the recent NLP boom—sparked by transformers and later by Large Language Models (LLMs) [Zhao et al., 2023], few-shot learning [Brown et al., 2020], and prompting [Liu et al., 2023a]—a significant part of the work is the timely adoption and evaluation of fast-evolving SOTA methods in our context.

Overall, I build on prior Czech fact-checking research, extend it to other languages, and evolve the transformer *pre-training & fine-tuning* paradigm toward an LLM-centric design, which already shows strong performance on related tasks in English [Chen et al., 2023].

My recent focus within the broader fact-checking pipeline is *claim generation*, which I aim to establish as a benchmarked NLP task adjacent to *abstractive summarization*. To benchmark the task, we need metrics that capture phenomena such as *model hallucinations*—a common issue in modern LLMs [Ji et al., 2023]. Because word-overlap metrics for generation correlate poorly with human judgment [Zhang* et al., 2020] and model-based metrics can be opaque, I also focus on human-understandable model-based metrics.

This study aims to show the directions I am taking to address these challenges, the reasoning behind them, my research questions, and the current results that motivated them.

1.1 Motivation

The spread of misinformation in the online space has a growing influence on the Czech public [STEM, 2021]. It has been shown to influence people's behaviour on the social networks [Lazer et al., 2018] as well as their decisions in elections [Allcott and Gentzkow, 2017], and real-world reasoning, which has shown increasingly harmful during the COVID-19 pandemic [Barua et al., 2020] and the Russo-Ukrainian war [Stănescu, 2022].

Figure 1.1: A real-world example of fact checking done by `https://politifact.org`

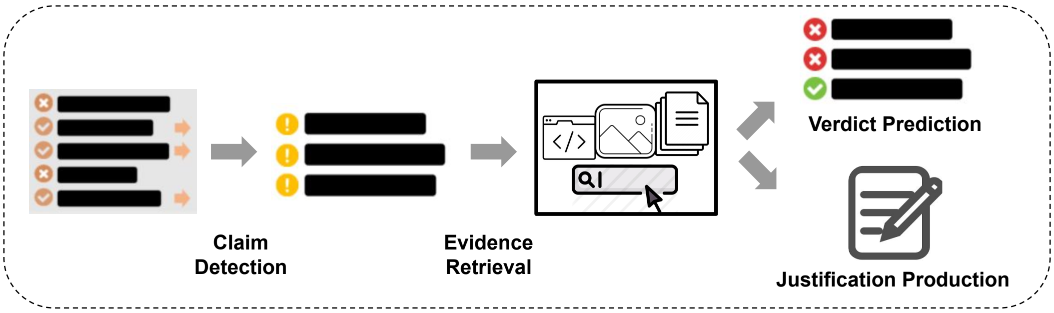

Figure 1.2: Automated fact-checking pipeline, reprinted from [Guo et al., 2022]

The recent advances in artificial intelligence have unintendedly contributed to the spread of misinformation on social media [Buchanan and Benson, 2019], as well as they hold a large potential for the false content generation [Sebastian, 2023].

Recent research has shown promising results [Thorne et al., 2019] in false claim detection for data in English, using a trusted knowledge base of true claims (for research purposes typically fixed to the corpus of Wikipedia articles), mimicking the *fact-checking* efforts in journalism.

Fact-checking (Figure 1.1) is a process of matching every information within a *factual claim* to its *evidence* (or *disproof*) in trusted data sources to infer the claim veracity and verifiability. In exchange, if the trusted *knowledge base* contains a set of "ground truths" sufficient to fully infer the original claim or its negation, the claim is labeled as **supported** or **refuted**, respectively. If no such *evidence set* can be found, the claim is marked as **unverifiable**[1].

1.2 Automated Fact Checking

Despite the existence of end-to-end fact-checking services, such as `politifact.org` or `demagog.cz`, the human-powered approach shows weaknesses in its scalability. By design, the process of finding an exhaustive set of evidence that decides the claim veracity is much slower than that of generating false or misguiding claims. Therefore, efforts have been made to move part of the load to a computer program that can run without supervision.

The common research goal is a fact verification tool that would, given a claim, semantically search the provided knowledge base (stored, for example, as a *corpus* of some natural language), propose a set of evidence (e.g., k semantically nearest paragraphs of the corpus) and suggest the final verdict (Figure 1.2) [Guo et al., 2022]. This would reduce the fact-checker's workload to mere adjustments of the proposed result and correction of mistakes on the computer side.

The goals of the ongoing efforts of FactCheck team at AIC CTU are to explore and adapt the state-of-the-art methods used for fact verification or similar tasks in other languages, curate appropriate datasets for it and propose strong systems for such a task in Czech.

1.3 A word on the Transformers

For the past six years, the state-of-the-art solution for nearly every Natural Language Processing task is based on the concept of *transformer networks* or, simply, *Transformers*. This has been a major breakthrough in the field by [Vaswani et al., 2017], giving birth to the famous models such as Google's BERT encoder [Devlin et al., 2019] and its descendants, or the OpenAI's GPT-3 decoder [Brown et al., 2020] and GPT-4 [OpenAI, 2023] that are used in the booming online AI service ChatGPT[2].

In our proposed methods, we use Transformers in every step of the fact verification pipeline. Therefore, we would like to introduce this concept to our readers to begin with.

Transformer is a neural model for *sequence-to-sequence* tasks, which, similarly, e.g., to the *LSTM-Networks* [Cheng et al., 2016], uses the Encoder–Decoder architecture. Its main point is that of using solely the *self-attention* mechanism to represent its input and output instead of any sequence-aligned recurrence [Vaswani et al., 2017].

In essence, the *self-attention* (also known as the *intra-attention*) transforms every input vector to a weighted sum of the vectors in its neighbourhood, weighted by their *relatedness* to the input. One could illustrate this on the *euphony* in music, where every tone of a song relates to all of the precedent and successive ones, to some more than to others.

The full Transformer architecture is depicted in Figure 1.3.

1.4 Dissertation outline

- **Chapter 1** introduces the dissertation topic, motivates the research, sets up our challenges for future research
- **Chapter 2** examines the most relevant research in the field and tries to highlight the recent paradigm shift from models trained for a single task to single large models that perform well in everything

[1] Hereinafter labeled as `NOT ENOUGH INFO`, in accordance to related research.
[2] `https://chat.openai.com`

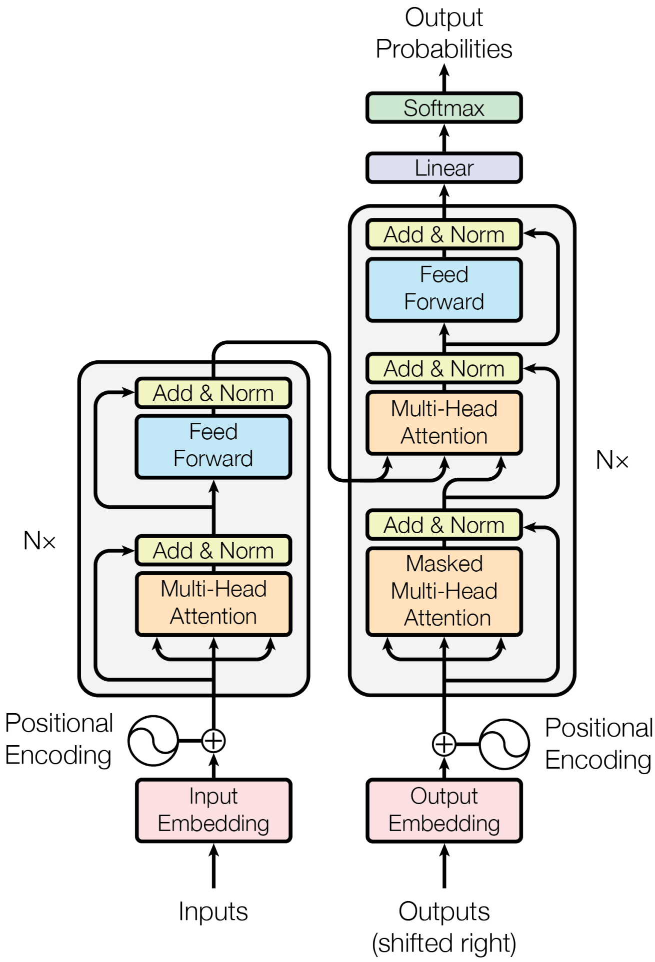

Figure 1.3: Transformer model architecture, reprinted from [Vaswani et al., 2017]

- **Chapter 3** explains our current contributions to the field of automated fact-checking and NLP in Czech
- **Chapter 4** describes our plan for the dissertation and justifies the directions we are taking
- Finally, **Chapter 5** concludes the study with a wrapup of its findings

Chapter 2
State of the Art

This chapter surveys the evolution from BERT-style transfer learning to instruction-following LLMs that often outperform smaller fine-tuned models even without additional training [Devlin et al., 2019, OpenAI, 2023, Touvron et al., 2023a, Vicuna, 2023]. We then review efficiency methods that make fine-tuning multi-billion-parameter models on a single GPU practical and discuss their relevance to our work.

To show how it relates to our main topics, we will introduce currently published approaches for the automated fact-checking task, efforts related to claim generation, and evaluation of NLP model outputs.

2.1 Pretrain + Finetune

For the last decade, the *pretrain-finetune* paradigm has been a cornerstone in Natural Language Processing (NLP). It has significantly shaped the development of modern NLP models. Its use in NLP can be traced back to the advent of neural networks and deep learning in the early 2010s. Initially, researchers pre-trained word embeddings using methods like Word2Vec [Mikolov et al., 2013] and GloVe [Pennington et al., 2014], which captured semantic relationships among words and then tweaked the general-task models for various related tasks.

2.1.1 BERT and derivatives

The *pretrain-finetune* paradigm truly rose to fame with the introduction of transformer-based models, particularly the revolutionary BERT (Bidirectional Encoder Representations from Transformers) in 2018. BERT [Devlin et al., 2019] demonstrated the power of pretraining large-scale language models on massive text corpora using an easy-to-automate general task such as *Masked Language Modeling*, or *Next Sentence Prediction*, followed by fine-tuning on specific downstream tasks using smaller, harder-to-obtain data. This approach achieved state-of-the-art results across various NLP benchmarks. Subsequently, numerous variations of pre-trained models like GPT (Generative Pre-trained Transformer) and RoBERTa emerged, each refining the pretrain-finetune paradigm to improve language understanding, generation, and transfer learning capabilities.

Importantly, BERT's success inspired many publications in training similar transformer models, varying in the definition of the general pre-training task, model size, architecture training corpus

- In Czech language, monolingual models CZERT [Sido et al., 2021], FERNET [Lehečka and Švec, 2021], RobeCzech [Straka et al., 2021], and small-e-czech [Kocián et al., 2021] are available for further finetuning

- In Polish, HerBERT [Mroczkowski et al., 2021] achieved state-of-the-art in multiple tasks in 2021
- In Slovak, SlovakBERT [Pikuliak et al., 2021] was released by KInIT and Gerulata
- A multitude of multilingual models, such as M-BERT or XLM-RoBERTA [Conneau et al., 2019] were pre-trained on data in all three of these languages (and many others), proving that the large transformers can capture a notion of semantics and relations between pieces of text even *without* the convenient constriction of a single language

2.2 Few-shot and Zero-shot learning

The ever-growing (sometimes billions of parameters in size) transformer models have not only demonstrated superior performance on benchmark datasets but have also shown remarkable zero-shot and few-shot learning abilities, where they can perform tasks with minimal or no task-specific training data [Brown et al., 2020].

Few-shot learning refers to the capability of a model to perform a task when provided with only a limited amount of labeled examples. Zero-shot learning takes this concept a step further by enabling models to tackle tasks they have never seen during training. The integration of these learning paradigms into large language models like GPT-3 and subsequent iterations has spread the NLP hype even further. By utilizing a prompt or a few examples, these models can quickly adapt to new tasks, making them highly versatile, adaptable, and usable to the general public.

2.2.1 OpenAI LLMs: GPT-3 and GPT-4

In 2020, the few-shot learning was exhibited on GPT3 – a 175B-parameter autoregressive model trained by [Brown et al., 2020]. The model was trained on the task of generating text based on user's and its own previous outputs. The training procedure and data[1] is thoroughly described in the publication. However, it is prohibitively costly for most labs to reproduce or even fine-tune at such a scale.

In the fall of 2022, GPT-3 became widely popular thanks to its ChatGPT[2] fine-tune and demonstration app, which puts the user in the role of *prompter*, texting back and forth with an LLM that predicts the most fitting reply to each conversation.

With the arrival of GPT-4, the ChatGPT was already massively famous, and the new model already shipped with a paid-service business scheme no longer publishing the training data, tasks, or even model size [OpenAI, 2023].

2.3 Open source LLMs

This puts the research community in an awkward position, as the GPT-4 achieves state-of-the-art in numerous NLP benchmarks [OpenAI, 2023, Liu et al., 2023b], but is designed not to be used in any way other than as a black box, making the derived research rigorosity and reproducibility disputable.

From the prediction times, OpenAI claims, and general trends in NLP, there are also reasons to believe that GPT-4 is orders of magnitude larger than already wasteful GPT-3. This motivates an uptick in research of other LLMs that would be able to operate on a

[1] A mixture of crawled websites, books, and Wikipedia.

[2] `https://chat.openai.com`

smaller scale with similar results, using a peer-reviewed architecture, training scheme, and data that is available in open source.

2.3.1 LLaMA-2 and derivatives

A popular foundational LLM to compete with the GPT family has become the LLaMA [Touvron et al., 2023a] from Meta research. LLaMA was trained on about 5TB of publicly available textual data[3] mainly in English.

It comes in various sizes between 7B and 65B parameters, achieving a SOTA among open-source solvers in various tasks and an unmatched performance in the field of single-GPU (7B and 13B) model sizes. LLaMA proceeds to be used as a goto base model for a number of successful open-source chatbots such as Alpaca [Taori et al., 2023], Vicuna [Vicuna, 2023], and OpenAssistant [Köpf et al., 2023].

The pre-trained LLaMA weights are, however, published under a restrictive license that prohibits republishing the model weights even after tuning its parameters, which limits its fine-tuners to publishing delta- or xor-weights that can not be properly used without Meta's permission.

LLaMA-2 [Touvron et al., 2023b] addresses this inconvenience (as well as delivers its own take on the *chatbot* task), yielding an ideal strong base model for experimentation with any NLP task in 7B, 13B, and 70B sizes. The only obstacle left in the way is the computational cost of fine-tuning across so many parameters.

2.3.2 LoRA and other optimization

To be able to fine-tune multi-billion-parameter models such as LLaMA-2 [Touvron et al., 2023b] on a single TPU, successful approaches have been published to dramatically cut down the training expenses. Parameter-efficient fine-tuning (PEFT) [Liu et al., 2022a] proposes approaches to only fine-tune *a few* weights as opposed to the whole neural network, reducing the number of trainable parameters by orders of magnitude. Low-Rank Adaptation of Large Language Models (LoRA) [Hu et al., 2021] does so by freezing the pre-trained model weights and injecting trainable rank decomposition matrices into each layer of Transformer architecture.

Quantization, which cuts the costs of working with 32- or 16-bit float parameters and opting for data types of bitsize as small as 4, also proves to be a powerful tool for LLM finetuning performance optimization [Dettmers et al., 2023]. Quantized QLoRA takes LLaMA and finetunes it into a Guanaco model family, which outperforms all previous openly released LLMs on Vicuna benchmark [Dettmers et al., 2023] and achieves 99.3% of the ChatGPT's performance on it while only requiring 24 hours on a single GPU.

As per an alleged leaked Google memo [Patel and Ahmad, 2023], this could put the future state of the art in NLP disciplines back into the hands of open source and public research, not giving any of the big tech companies a "moat" advantage.

Either way, it goes to show that the open-source LLMs have a promising future in NLP and will be indismissible as an approach for the NLP task of *Automated fact checking*.

[3] To be specific, LLaMA was trained using an autoregressive language modeling task on a mixture of English CommonCrawl Corpus, C4 [Raffel et al., 2019], Github, Wikipedia, Gutenberg Project, Books3 corpus, ArXiv and Stack Exchange

2.4 Fact checking approaches

Back in the late 2010s, the misinformation and its spread in the era of the internet and social media became a discussed topic in the Western world, with multiple institutions such as the European Council marking it a severe threat to democracy and national safety [Wardle and Derakhshan, 2017]. The public attention and maturation of appropriate technologies motivated numerous efforts in business and academia to tackle the challenge. Among other events, a Fake News Challenge occurred in 2017 [Pomerlau and Rao, 2017] exploring the uses of technologies in the field and applying, for example, the LSTMs to detect stances among textual data [Hanselowski et al., 2018].

2.4.1 FEVER and followups

Soon, standard tasks began to be formulated and data collected. The FEVER (Fact Extraction and VERification) [Thorne et al., 2018a] dataset and shared task became prominent in natural language processing research. Relatively early on, it formalized the task as a two-step problem:

1. Retrieving information within a structured corpus to fact-check a given claim (this resembles a standard NLP problem called *information retrieval* – IR)

2. Classifying the inference relation between retrieved information and claim as one of:

   a. **supports** – information semantically implies the claim
   b. **refutes** – information semantically implies the negation of the claim
   c. **not enough info** otherwise

   This classification task became known as *natural language inference* and mostly replaced the previous binary classification NLP task of *recognizing textual entailment* (RTE)

The FEVER dataset was a collection of 185K human-annotated claims, their veracity labels, and sets of evidence from a structured corpus that sufficed to justify the labels. The corpus of choice was a 2017 English Wikipedia structured into articles due to its reasonable size, informational richness, and open license.[4]

FEVER yields an interesting benchmark with statistically quantifiable model success, motivated multiple well-performing public solutions [Thorne et al., 2018b, Thorne et al., 2019], gives insights into the complexities of automated fact-checking task, and strong baselines for research in the field. The data was later enriched by contrastive evidence in VitaminC [Schuster et al., 2021] and by reasoning over tabular data in FEVEROUS [Aly et al., 2021].

To date, it keeps being a reference point in automated fact-checking research despite its limitations, such as its requirement for a fixed knowledge base and "atomicity"[5] of claims.

[4]Is Wikipedia a trustworthy informational canon, though? No, it is not supposed to – FEVER states that it is crucial to always maintain that the fact-checking classifiers only classify *with respect to* data, and their reliability goes only as far as that of the underlying knowledge corpus. Therefore, *supports* does not directly translate to *true*, nor *refutes* to *false*

[5]See section 4.1.2.3

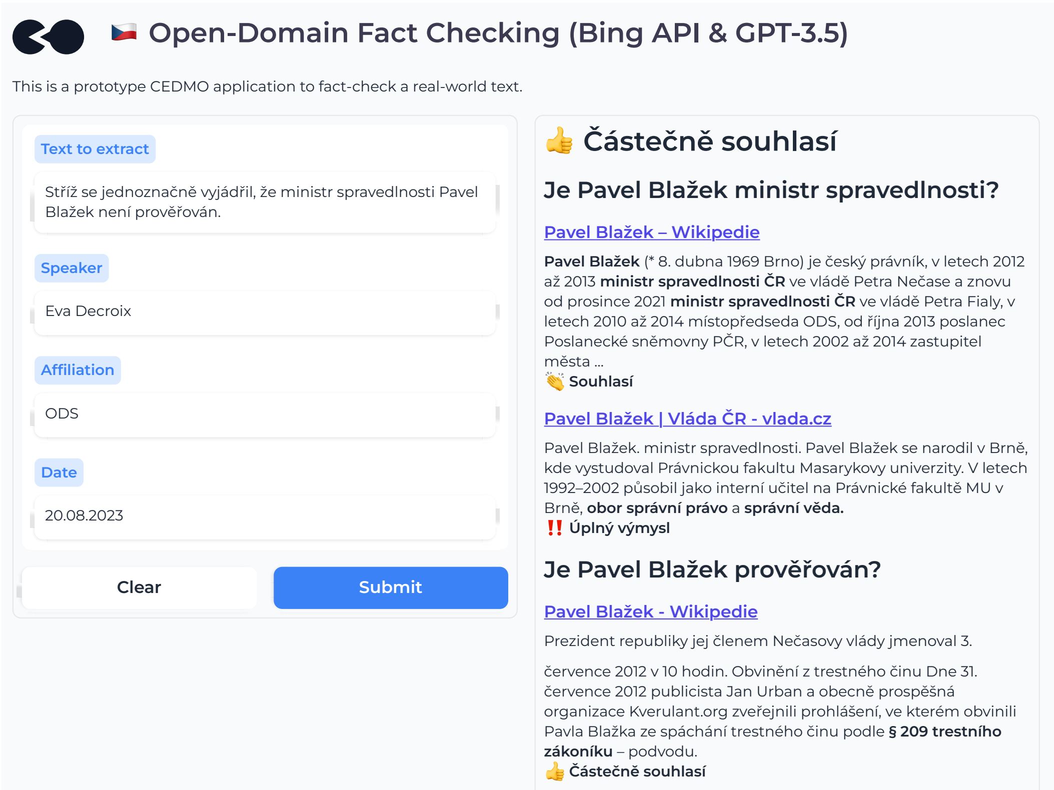

Figure 2.1: Proof-of-concept Czech fact-checking based on live-internet search (Bing API) and LLM prompting, based on the proposals of [Chen et al., 2023] in Czech, using a real-world claim that was fact-checked by `demagog.cz` in June 2023

2.4.2 Open-domain fact-checking

Due to these limitations, some researchers consider the scheme from FEVER an oversimplification – the real politics' claims to be fact-checked by journalists often consist of long syntactical structures, combine information together in a non-trivial manner and often require the most up-to-date evidence.

"Complex Claim Verification with Evidence Retrieved in the Wild" [Chen et al., 2023] proposes a different scheme that overcomes these shortcomings:

1. Arbitrarily complex claim is decomposed into a set of yes/no questions
2. An open-domain search (Bing is proposed in the paper) fetches several evidence documents for each question
3. A claim-focused summary is extracted from each document
4. A veracity classifier goes through each pair of evidence and question, ranging from "faithful" to "completely wrong"
5. The scores are combined (all need to be "faithful" for a faithful claim. Otherwise, the severity of inaccuracies can be approximated using some averaging.

GPT-3 is used in steps 1, 3, and 4 of the scheme in the prototype delivered in [Chen et al., 2023] in a few- and zero-shot fashion, with few-shot unsurprisingly coming out a little better. The scheme is transducible to Czech, and Figure 2.1 shows my early experiments with my interactive reproduction of it, predictors based on Bing and GPT-3.5 (a polished version of GPT-3).

While the shift from an established FEVER framework to complex real-world claims and evidence retrieval "in the wild" feels exciting and practical, an obvious pitfall arises – anyone can publish anything on the internet, having it appear in Bing search and other crawlers alike. I argue that this might lead into a sort of a circular dependency of needing to reliably fact-check the evidence we have retrieved from the web in order to be able to build a reliable fact-checker in the first place.

Anyhow, the open-domain fact-checking idea opens a whole new range of approaches and shows the power of LLMs in fact-checking at its every step.

2.5 Claim generation

Another step of the fact-checking pipeline, covered by very few research publications, is the generation of the claim to be checked in the first place [Guo et al., 2022].

The current state of things is that journalists who fact-check statements within, say, a Facebook status, need to read through the whole document multiple times, formulate its factual claims from the stances and facts expressed in the text themselves, and then fact-check each separately.

What has been examined so far were, for example:

- Using Question Generation (QG) solver and converting the questions into declarative sentences to emulate more claims and more data for fact checking [Pan et al., 2021]
- Numerous CLEF CheckThat! challenges explored the task of estimating *checkworthiness* of different parts of a long text, such as lines in a political debate [Elsayed et al., 2021, Nakov et al., 2021]
- The task of extreme summarization (XSum) consists of summarizing a long body of text into a single sentence, focusing on its most relevant aspects and facts. Large datasets XSum [Narayan et al., 2018] in English and XL-Sum [Hasan et al., 2021] in 44 languages both present expertly annotated data from BBC News for it, as their article standard features a single-sentence summary at the beginning of each text.

2.5.1 NLP summarization benchmarking

An important caveat to note with the NLP tasks reducing longer text to shorter text – such as summarization or claim extraction – is that the standard automatic metrics such as ROUGE [Lin, 2004] and METEOR [Banerjee and Lavie, 2005] only focus on the *content selection* aspect of tasks, based on a word-by-word overlap and were designed to use on multiple gold summaries per input, which are not often provided with modern large-scale datasets. [NLP-Progress, 2023, Zhang* et al., 2020, Zha et al., 2023]

These serious limitations make it questionable for anyone to claim state-of-the-art on these tasks and motivate research for new metrics to cover all the important aspects of claim generation and do so in correlation with expert human judgment.

This will be the topic of section 4.2, which also introduces the state-of-the-art research we are working with to arrive to a valid set of benchmarks.

Chapter 3
Current contribution

extitWe collected novel data for fact-checking in our application context, emulated or scraped unavailable datasets (publishing or preparing them for release), established strong baseline models, and are working to position claim generation as a summarization-adjacent NLP task.

3.1 Datasets

With the automated fact-checking scheme established in Chapter 2, any ML solution begins with choosing or collecting appropriate training data. Due to the novelty of the task in Czech and other West Slavic languages, I explored a multitude of ways to acquire such data, many of them resulting in a publicly available dataset in our Huggingface repository [1], beginning to be reused by others.

3.1.1 CsFEVER

An early "temporary benchmark" for our endeavors in adapting the FEVER [Thorne et al., 2018a] task for the Czech context was the CsFEVER [Ullrich et al., 2023] dataset.

In [Ullrich, 2021], I have proposed a simple FEVER data transduction scheme that can be simplified as follows:

1. Each FEVER claim is translated using a Machine Translator
2. Evidence from English Wikipedia is not translated using MT, but mapped onto its Czech-Wikipedia counterpart using the publicly available Wikidata[2]
3. Data with any loss in evidence due to step 2. is discarded

This design was relatively cheap (translating the entire 2017 Wikipedia would be costly and wasteful), yielding a CC-licensed dataset of 127K claims, labels, and evidence. Since both 2017 EnWiki and our 2020 CsWiki corpora contained only the first paragraph (abstract) per article, we hoped document-level alignment would hold: both languages summarize basic facts about the same entity.

This showed to be only partly true as a later human annotation on a 1% sample of CsFEVER data showed that about a third of data exhibits some levels of noise, mostly introduced during dataset translation [Ullrich et al., 2023].

[1] `https://huggingface.co/ctu-aic`

[2] Used, for example, for showing the "see this article in other languages" suggestions in Wikipedia sidebar

While noisy, CsFEVER has been used to train IR components [Rýpar, 2021, Gažo, 2021, Ullrich et al., 2023] still in use today and is openly available[3] under a CC license.

My research on it also motivated the creation of an inference-only version of the dataset, which does not support the Information Retrieval task and, therefore, does not require the mapping of evidence into a live version of Wikipedia. Therefore, only the EnWiki *excerpts* needed to build evidence can be translated, bringing down the computational difficulty and enabling me to deliver a dataset without the transduction noise called CsFEVER-NLI[4].

Another round of research CsFEVER motivated, and I supervised, was the successful thesis of [Mlynář, 2023], modernizing the data and machine-translation methods into the 2023 state of the art. [Mlynář, 2023] further experimented with methods of automated noise detection and removal, which has not shown to be an efficient way to tackle the issue of high noise in CsFEVER.

Anyhow, it delivers a partly cleaned version of it[5] and motivates future research to generate such data differently, using a claim generation scheme like that from [Pan et al., 2021].

3.1.2 FCheck annotations platform

The imperfections in translated CsFEVER data, as well as the ongoing collaboration with ČTK and the Faculty of Social Sciences, brought me to also look for ways how to hand-annotate a whole new natively Czech dataset, which would both lack the noise introduced in translation and also take the task of automated fact-checking to the next level, replacing a rigid, simple Wikipedic data with a more "real world" news report corpus of ČTK.

Figure 3.1 shows an open-source platform FCheck[6] I developed to collaborate with 316 FSV CUNI students of on a collection of novel dataset in Czech using ČTK data as a ground truth corpus.

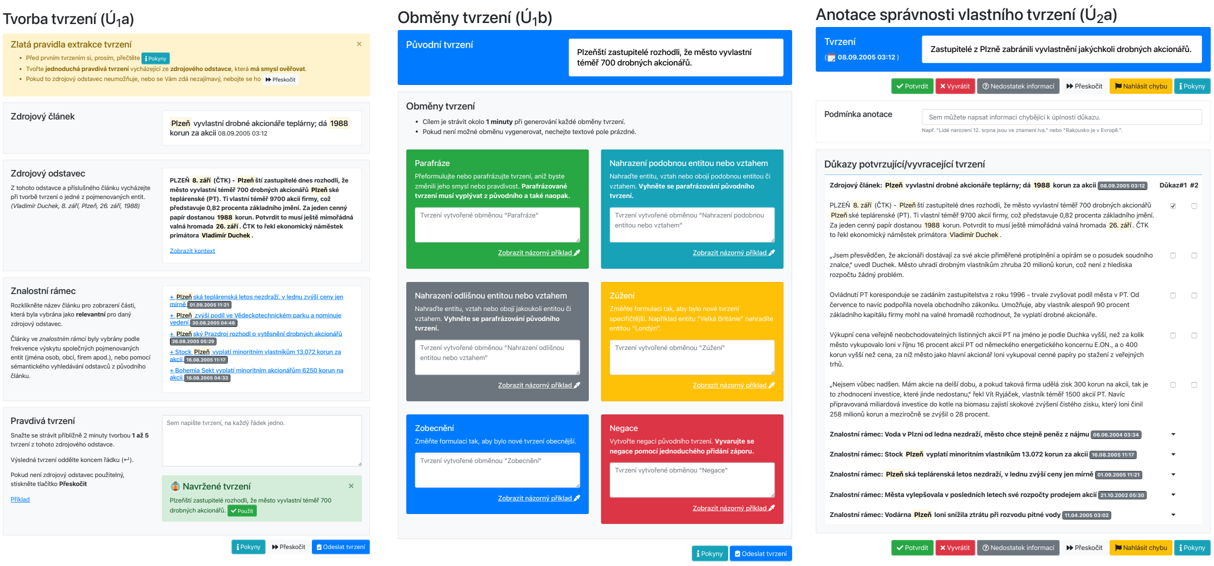

Figure 3.1: FCheck – a platform for fact-checking data collection developed for TAČR project; collects data for claim generation, information retrieval, and natural language inference tasks

[3] https://huggingface.co/datasets/ctu-aic/csfever
[4] https://huggingface.co/datasets/ctu-aic/csfever_nli
[5] https://huggingface.co/datasets/ctu-aic/csfever_v2
[6] https://fcheck.fel.cvut.cz (testuser), source at: github.com/aic-factcheck/fcheck-annotations-platform

We have established a 4-step annotation procedure inspired by the time-proven methodology of [Thorne et al., 2018a] where check-worthy paragraphs are first hand-picked among samples from the whole archive of ČTK's 3.3 M news reports published between 1 January 2000 and 6 March 2019. Then, the annotator is sampled such a paragraph and asked to *extract claims* from it, i.e., formulate single-sentence summaries of some facts that appear in the paragraph. This claim is always *supported* by the data, so the next phase is to perturb the claim by the annotator's world knowledge and form the claim *mutations* – substitutions of entities, generalizations, specifications, paraphrases or negations of the original claim. The mutated claim is then fact-checked by (typically) another annotator, using the ČTK data narrowed down to a reasonable number of relevant articles (in an IR sense) as *supportable*, *refutable* or *not enough info*, providing a set of evidence as a verdict justification.

The application has two layers: a Yii-based PHP interface for annotation and a Python Flask service hosting our IR models based on TF-IDF [Chen et al., 2017] and mBERT (Section 2.1.1), trained (among other data) on CsFEVER (Section 3.1.1). The models are solving the Information Retrieval task on-demand (with cache) on the proprietary ČTK corpus whenever the annotation app needs it to provide context to the fact-checker.

The scheme and its implementations are exhaustively described in [Ullrich, 2021], chapter 4, and in [Ullrich et al., 2023], also chapter 4. Multiple "cross-annotations" were collected for each claim to measure agreement and give insights into task complexity.

3.1.3 CTKFacts

After completing the first year of annotation experiments, we have extracted a total of 3,116 multi-annotated claims. 47% were `SUPPORT`ed by the majority of their annotations, `REFUTES` and `NEI` labels were approximately even, the full distribution of labels is listed in Table 3.1.

	CTKFacts uncleaned, balanced			**CTKFacts (launch)** cleaned, stratified		
	SUPPORTS	REFUTES	NEI	SUPPORTS	REFUTES	NEI
train	1,164	549	503	1,104	556	723
dev	100	100	100	142	85	105
test	200	200	200	176	79	127

Table 3.1: Label distribution in CTKFacts splits before and after cleaning. Reprinted from [Ullrich et al., 2023]

Of all the annotated claims, 1,776, that is 57%, had at least two independent labels assigned by different annotators. I used this multiplicity to assess data quality and task ambiguity, and to propose cleaning methods for the final cleaned CTKFacts dataset.

Inter-annotator agreement

Due to our cross-annotation design, I had a generously sized sample of independently annotated labels in our hands. As the total number of annotators was greater than 2, and as missing observations were allowed, I have used the Krippendorff's alpha measure [Krippendorff, 1970] which is the standard for this case [Hayes and Krippendorff, 2007]. For the comparison with [Thorne et al., 2018a] and [Nørregaard and Derczynski, 2021], I also list a 4-way Fleiss' κ-agreement [Fleiss, 1971] calculated on a sample of 7.5% claims.

Krippendorff's alpha was 56.42% and Fleiss' κ 63%, an adequate result reflecting the complexity of news-based verification within a fixed knowledge scope. It also encourages a round of annotation-cleaning experiments that would exploit the number of cross-annotated claims to remove common types of noise.

CTKFacts publication

The CTKFacts dataset then underwent thorough human-in-the-loop cleaning to reach 100% agreement on the retained items, removing obvious noise and revealing sources of annotation error. The full process, as well as its results, are described in [Ullrich et al., 2023].

Ultimately, a dataset of 3.1K thoroughly cleaned data points in the form of a factual claim, its veracity label and justifications consisting of ČTK paragraphs was published in a version for Information Retrieval[7] for those who have access to the ČTK knowledge base to retrieve from, as well as in a special version for the task of Natural Language Inference[8] containing all the required ČTK excerpts we have negotiated to publish under open license for everyone to use.

These datasets are now standard benchmarks within the AIC NLP group [Semin, 2023, Mlynář, 2023] and are beginning to appear in external research [Štefánik et al., 2023].

3.1.4 Other NLP datasets in West Slavic languages

Over time, we have accumulated numerous sets of data in Czech and other Slavic languages that have previously been poorly covered or not available at all, some of which are to be referred in our future publications. For the convenience of others, most of them are already listed in our public repositories. Let us mention some significant examples:

1. We have machine-translated the most popular NLI training and benchmark datasets such as Stanford NLI [Bowman et al., 2015], Adversarial NLI [Nie et al., 2019b] and MultiNLI [Williams et al., 2018] picking a machine translator empirically for each dataset between DeepL [DeepL, 2021], Google Translate [Google, 2021] and CUBBITT [Popel et al., 2020].

   The resulting datasets are maintained at our public repositories:

   a. `https://huggingface.co/datasets/ctu-aic/snli_cs`
   b. `https://huggingface.co/datasets/ctu-aic/anli_cs`
   c. `https://huggingface.co/datasets/ctu-aic/multinli_cs`

2. For the task of claim generation we are establishing and performing in Czech, we have adapted the existing related datasets and are working with:

   a. CTKSum – `https://huggingface.co/datasets/ctu-aic/ctksum` based on source articles and extracted claims within the original CTKFacts set
   b. FEVERSum (based on FEVER Wikipedia abstract and extracted claims) – `https://huggingface.co/datasets/ctu-aic/fever-sum`
   c. Its DeepL translation CsFEVERSum – `https://huggingface.co/datasets/ctu-aic/csfever-sum`

[7] `https://huggingface.co/datasets/ctu-aic/ctkfacts`
[8] `https://huggingface.co/datasets/ctu-aic/ctkfacts_nli`

d. Our reproduction of a crawled Slovak summarization dataset described by [Šuppa and Adamec, 2020] SMESum based on articles from `https://sme.sk` – `https://huggingface.co/datasets/ctu-aic/smesum`

Up until now, some of the data was restricted to private repositories, but with this study, I am publishing most of them, as I have now found the licensing to be rather relaxed. If some of the repositories the reader might be interested in would not be reachable, please request access to the `https://huggingface.co/datasets/ctu-aic` organization to be able to see into the private part of our dataset library.

3.2 Models

The most significant released models address two tasks: Natural Language Inference and Claim Generation (as Abstractive Summarization).

3.2.1 Natural Language Inference

My previous work [Ullrich, 2021, Ullrich et al., 2023] also focused on establishing a strong starting state of the art on our own datasets in the tasks of NLI. In my publications, I have tried and compared a multitude of neural networks for the tasks, ultimately arriving at the following:

- **XLM-RoBERTA-Large@XNLI@CsFEVER-NLI**, a model with 561M parameters trained on 100-language CommonCrawl corpus finetuned on multilingual XNLI [Conneau et al., 2018] inference dataset and then finetuned *again* on the CsFEVER-NLI task yields an unmatched 73.7% F1 macro score on the denoised CsFEVER-NLI inference task: `https://huggingface.co/ctu-aic/xlm-roberta-large-xnli-csfever_nli`
- **XLM-RoBERTA-Large@SQuAD2**, a model version finetuned on a Question answering SquAD2 [Rajpurkar et al., 2016] task has shown remarkable practicality in my NLI applications and after task-specific finetuning, it was able to tackle:

  1. CTKFactsNLI[9] task with 76.9% macro-F1
  2. CsFEVER[10] (noisy) task with 83.2% macro-F1
  3. The original English FEVER NLI task[11] [Thorne et al., 2018a, Nie et al., 2019a], achieving 75.9% macro-F1 and surpassing the prior shared-task winner [Nie et al., 2019a] (69.5 macro-F1 with NSMNs)

3.2.2 Claim generation

In my current research, I am finding appropriate configurations and data to train models for claim generation – generating a factual claim (or more) into a single sentence containing a fluent, atomic, decontextualized, and faithful claim. In section 4.1, I propose the claim generation as an abstractive summarization setting, and therefore, the models already have their practical use in the general task of summing up longer texts into shorter ones.

As has been shown in section 2.5.1, the NLP summarization task does not have a reliable standard benchmark that would capture all its required output qualities. Therefore, it

[9] `https://huggingface.co/ctu-aic/xlm-roberta-large-squad2-ctkfacts_nli`
[10] `https://huggingface.co/ctu-aic/xlm-roberta-large-squad2-csfever_nearestp`
[11] `https://huggingface.co/ctu-aic/xlm-roberta-large-squad2-enfever_nli`

remains questionable to claim the state of the art on any summarization task, and I proceed to present models that excel in our empirical tests and demonstrations for project stakeholders:

1. **mBART** [Liu et al., 2020] multilingual Transformer model has been finetuned by our team's [Krotil, 2022] on SumeCzech and proprietary CNC News summarization dataset on the "full text to headline" task, obtaining encouraging scores across numerous summarization metrics in Czech.
   I have taken this model a step further for the claim generation task, finetuning it on the CsFEVERSum and CTKFactsSum datasets, yielding a working model for the task.[12]

   Other experiments are being carried out with the same model finetuned on Slovak[13] and Polish[14] data.

2. **LLaMA-2** shows promising results when it comes to claim generation. I have finetuned[15] it using the QLoRA (section 2.3.2) approach, XL-Sum [Hasan et al., 2021] dataset and a concatenation-based prompting strategy [Touvron et al., 2023b], to facilitate training across the entire length of input.

All prototype models are currently being iterated with our CEDMO[16] project partners (fact-checkers from European organizations), tweaked, and future tests are being designed for them based on empirical results and questionnaires.

An application in the figure 3.2 demonstrates the single or multiple claim generation task with our LLaMA-2 or mBART models for English and Czech texts, respectively – I put it together as a GRADIO interactive application and an API. Another interactive application (Figure 3.3), developed by Jan Drchal [Mlynář, 2023], demonstrates our best-performing end-to-end fact-checking models, integrating XLM-RoBERTa trained on CsFEVER-NLI.

[12] https://huggingface.co/ctu-aic/mbart25-large-eos
[13] https://huggingface.co/ctu-aic/mbart-at2h-cs-smesum-2
[14] https://huggingface.co/ctu-aic/mbart-at2h-cs-polish-news3
[15] https://huggingface.co/ctu-aic/Llama-2-7b-xlsum-en
[16] https://cedmohub.eu

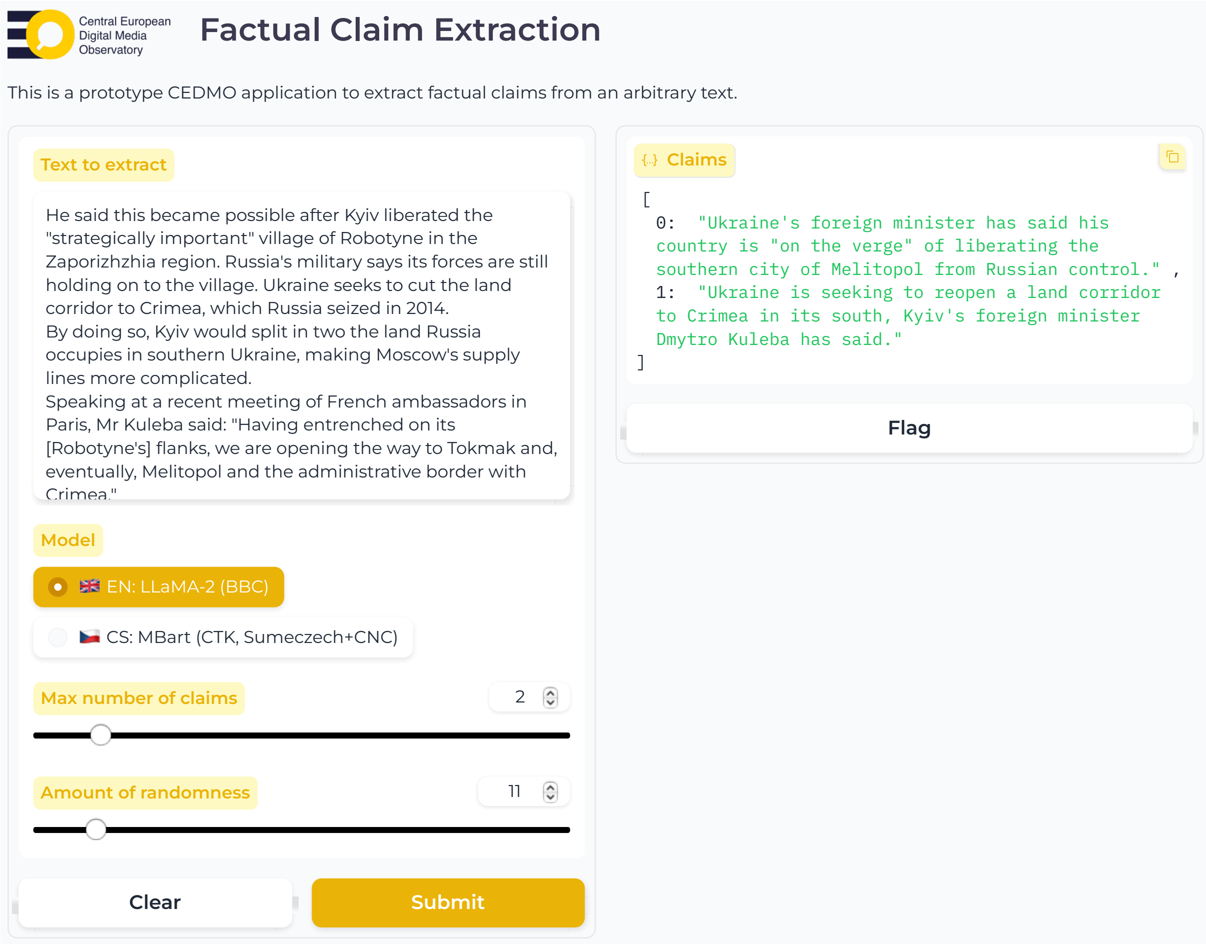

Figure 3.2: Factual claim extraction application done for the CEDMO project

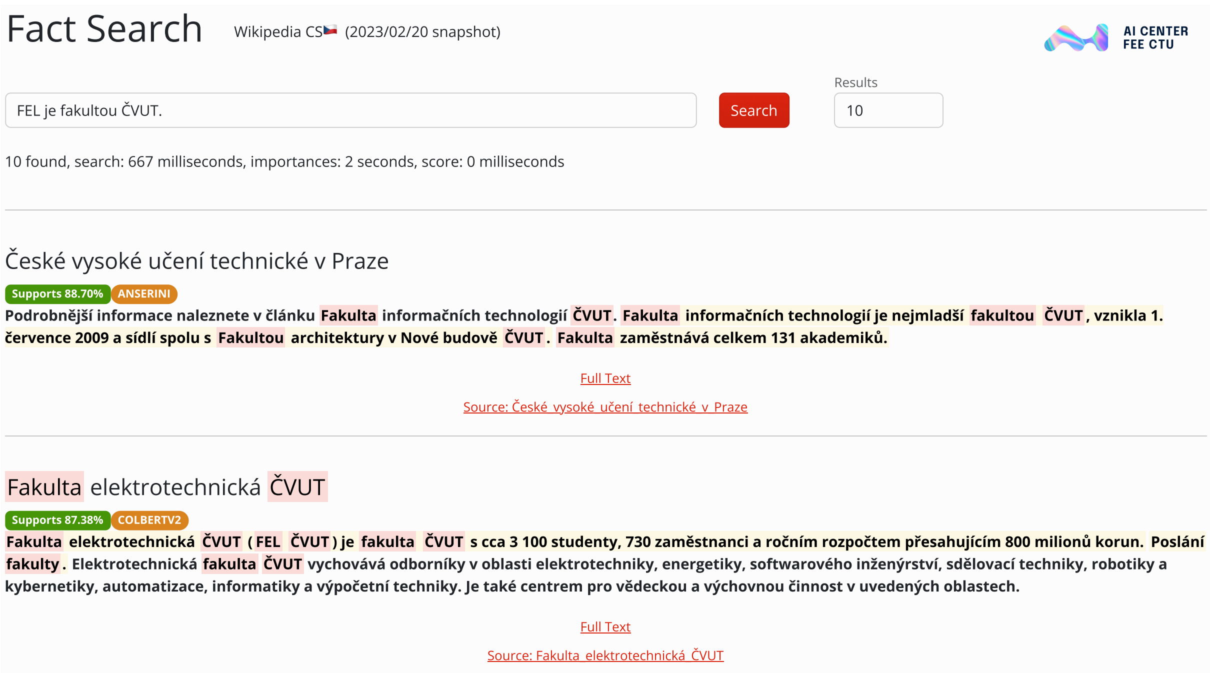

Figure 3.3: Automated fact-checking application "fact-search" verifying claims against Czech Wikipedia using our SOTA models

Chapter 4
Dissertation plan

extitThis chapter describes my current research agenda for automated fact-checking with NLP methods; the final section outlines the dissertation.

4.1 Automated claim generation

An article in preparation proposes *automated claim generation* as extracting factual claims from a document. This can assist fact-checkers and emulate data for related tasks (fact-checking, NLI).

Extracting fluent, atomic claims from naturally written text raises challenges—what information best characterizes the text? How does one resolve the pronouns and coreferences in source text? How does one adapt the extraction scheme for different speakers and stylistic forms?

These problems overlap with *abstractive summarization*, which recently made progress via Transformer models [Zhang et al., 2020, Liu et al., 2022b].

The summarization setup needs minor tweaks: enforce single-sentence outputs and encourage diverse factual foci via sampling (top-k[1] and top-p[2] [Holtzman et al., 2020]).

Initial training data comes from XL-Sum [Hasan et al., 2021], EnFEVER, and CTK-Facts. We train mBART, Pegasus, T5 [Raffel et al., 2019], and LLaMA-2 (QLoRA) models, with additional experiments via chat APIs.

Going forward, we will iterate on data and models in tandem and, crucially, develop reliable, explainable metrics that correlate with human judgment. As shown in section 2.5.1, the standard automated summarization metrics are not appropriate as a benchmark for the task.

4.2 Claim generation metrics

Generative tasks pose two issues: explaining model behavior in human-understandable terms and diagnosing failures such as *hallucinations*.

For the task of claim generation, where we also face the challenge of the *relevance* of the information extracted by the model, we suggest the use of the following metrics rooted in the current research on similar topics:

1. **Fluency** – *is the claim grammatically correct and intelligible?*

[1] Each output token is sampled from the k most probable words.

[2] Each token is sampled from the smallest prefix of tokens whose total probability mass is at most p.

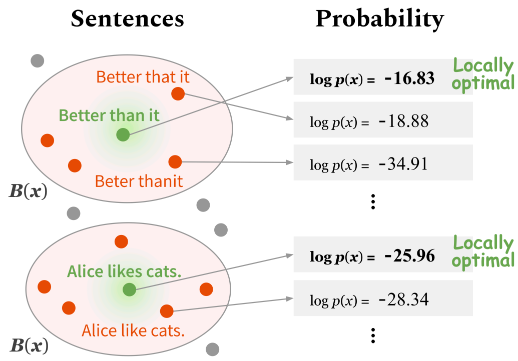

Figure 4.1: LM-Critic – deciding text fluency viewed as finding local optima of Language Model output probability, reprinted from [Yasunaga et al., 2021]

We explore two proxies for claim fluency—akin to Grammatical Error Detection (GED): LM-Critic (Figure 4.1) [Yasunaga et al., 2021] perturbs tokens to probe local optima of LM probabilities (e.g., GPT-2 as reference), and GPTScore [Fu et al., 2023] prompts an LLM (e.g., GPT-3) for a direct fluency score in few-/zero-shot settings.

Both can be adapted for Czech and the latter is demonstrated in Figure 4.2.

2. **Decontextualization** – *can the claim be correctly interpreted without any additional context from the source document or elsewhere?*

   A common problem with machine-extracted factual claims is reusing excerpts from source documents along with inexplicable contextual pronouns ("President won't sue *them*") and relative referencing ("*Last year*, CTU had 23K students").

   [Choi et al., 2021] proposes decontextualization as a sequence-to-sequence task with two texts on input (s, c) – sentence and context. T5 model [Raffel et al., 2019] is then trained on machine-generated gold data from Wikipedia to output sentence s' such that the truth-conditional meaning of s' in an empty context is the same as that of s in c.

   [Mohri et al., 2023] improves upon this, altering the problem formulation to minimization of surrogate loss, rejecting with a fixed predictor, and claiming to get as close as $\sim 3\%$ away from the theoretical limit for the task.

   The approaches are reproducible using the Czech Wikipedia corpus and appropriate for further examination.

3. **Atomicity** – *does the claim describe a single entity, relation or process?*

   Claim atomicity can be proxied via Relation Extraction (RE), e.g., LUKE [Yamada et al., 2020]. Identify entities and relations (e.g., (*study at*, *Herbert*, *CTU*)); mark a claim as atomic if at most one such triple is found (after symmetry normalization).

4. **Faithfulness** – *does the claim only contain information that is consistent with the source document?*

   This metric is crucial to pinpoint *hallucinations*—parts of the claim unsupported by the source. We use two alternatives: a score within the FFCI framework [Koto et al., 2020]:

$$\text{AvgTop-}n_{s_j \in X, t_i \in Y'}(\text{BERTSCORE}(t_i, s_j))$$

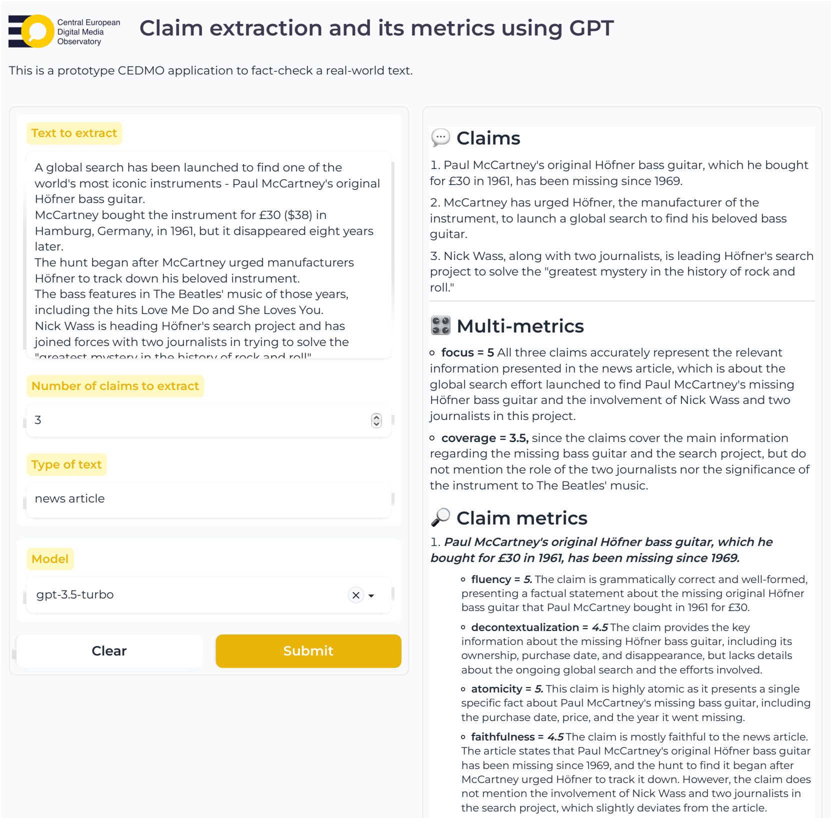

Figure 4.2: A self-evaluating claim generation model based on GPT-3.5-turbo and GPT-4 [OpenAI, 2023] using the OpenAI API and a single-shot (one gold example given) approach

Here AvgTop-n averages the top n (e.g., n=5) scores; X and Y' are sentence sets from the source and the model output (for claim generation, $|Y'|$=1). BERTScore [Zhang* et al., 2020] compares sentence embeddings rather than surface word overlap (e.g., ROUGE [Lin, 2004]), which helps in morphologically rich languages like Czech.

A related metric, AlignScore [Zha et al., 2023], optimizes alignment between input and output spans via a RoBERTa [Liu et al., 2019] model trained for inconsistency detection on 4.7M examples spanning NLI, QA, paraphrase, etc. Despite being relatively small (355M parameters), it outperforms some GPT-4-based metrics.

Empirically, the models work encouragingly well on spotting hallucinations and inconsistencies in English, and while the transduction of BERTScore is trivial, using a Czech embedding model such as CZERT [Sido et al., 2021] or FERNET [Lehečka and Švec, 2021], reproducing the success of AlignScore will require more research and data.

5. **Focus**@k – *if we generate k claims using this model, what will be the proportion of gold (relevant) information among all the information listed in the generated claims?*

The metric is analogous to *precision*, but decisions are ambiguous in natural language due to synonyms and many valid phrasings.

An elegant and functional perspective on the problem has been brought around in QAGS[3] evaluation protocol [Wang et al., 2020], where the idea is to use a Question

[3]Pronounced "kags", stands for "Question Answering and Generation for Summarization"

Generation model (QG) to formulate questions in natural language based on all k predicted claims. The questions are then twice answered using a Question Answering (QA) model, giving it knowledge from (i.) the predicted claims (ii.) the gold claims written by a human. The focus is then defined as the proportion of questions with the same answers extracted from the gold and predicted claims among all questions the model can generate from the predicted claims.

6. **Coverage**@k – *if we generate k claims using this model, what proportion of gold (relevant) information from the source text will be covered?*

   Analogous to *recall@k* in general machine learning, QAGS proposes to generate questions using gold claims and try to answer them using the predicted claims, much like in the *focus* scenario, but vice versa.

The metrics are proposed in accordance with other research on model-based evaluation of similar NLP tasks [Koto et al., 2020, Wright et al., 2022] and are to be refined upon experiments with annotators.

4.3 Data collection

4.3.1 Human-in-the-loop grading of claim generators

To validate and progress the metrics referred to in section 4.2, one needs human-annotated data for the task. I aim to use an experiment similar to that of [Wright et al., 2022], presenting annotators with ordinal scales for the claim qualities and appropriate grading for each metric conditioned by objective rules.

My research will attempt to design the experiment in a way that yields the best data, checking its validity using inter-annotator agreement and other forms of feedback and publishing the data and scheme alongside the other solutions. Collected data will be used to validate the prototype metrics from section 4.2 and propose their variations based on the findings.

4.3.2 Polish dataset scraping

While Czech has its SumeCzech [Straka et al., 2018] and in Slovak, we can still reproduce the SMESum [Šuppa and Adamec, 2020] research, a large-scale single-sentence summarization dataset in Polish has yet to be established. The closest data I have found is the online news corpus [Szwoch et al., 2022] collected for the purposes of studying political polarization (and nowhere published, despite my e-mail urgences).

A scraping experiment in the Polish media, such as TVP, Rzeczpospolita, Gazeta Wyborcza, Fakt, etc., is therefore being prepared to obtain an appropriate single-sentence dataset for publication – it is also going to be another incremental step toward the dissertation on the overall topic of NLP fact-checking and its stages, focusing on English and West Slavic languages.

4.3.3 Crowd-sourced fact-checking platform

In 2023, other members of our team [Bútora, 2023] with funding from Avast developed a crowd-sourced fact-checking platform[4], where users gather reputations like on Wikipedia,

[4] https://factcheck.fel.cvut.cz

by sharing check-worthy pieces of information found across the internet, and by their checking with sources.

While I am not directly involved in the implementation of the project apart from early consulting, experiments with FSV CUNI are to be launched, populating this platform with data and users. After the experiments, other data and applications will be delivered, and their processing will be another part of my dissertation project.

4.3.4 CTKFacts expansion

In 2021/2022, another round of the CTKFacts annotation experiment (see section 3.1.3) was carried out with the FSV CUNI students, yielding about 5K new data points, including, for example, claims extracted from the Czech Twitter.

The data is being cleaned and examined and will be attached to one of the other upcoming publications and presented in the dissertation thesis.

4.4 Pipeline modernization

As mentioned throughout the chapter 2, the state of the art in NLP has shifted dramatically over the last year, and another of the tasks I am currently working on is the modernization of our pipeline – Claim Generation, Information Retrieval, Natural Language Inference models – and appropriate use of LLMs in the tasks.

So far, I have successfully finetuned LLaMA-2 [Touvron et al., 2023b] for the claim generation task, and we have a LoRA finetuning experimental setup ready for the NLI models. The use of LLaMA-2 and its successors for our tasks will be a topic on its own, as most publicly available LLMs filter out the other languages and focus solely on English.

4.5 The grand scope

Overall, in brief points, the main topics of my dissertation are expected to be:

1. Introduction of the fact-checking task and its data, strong model baselines, and specific properties in the **West Slavic** context.

2. An integration of the step of **Claim generation** step into it, based on methods of abstractive summarization.

3. A delivery of reliable **metrics** for the tasks and their validation with expert-level humans.

4. Modernization of the automated fact-checking framework and solutions in English and Czech into the age of **Large Language Models**. Solutions were proposed already – based on proprietary black-box LLMs such as GPT-3.5 [Chen et al., 2023] – our next goal is to deploy open-source LLMs in-house, experiment with different architectures, fine-tuning tasks and data, improving the SOTA on our benchmark data.

   Due to our aim to produce transparent and reproducible research, using open-source LLMs is preferred over popular proprietary ones like GPT-4.

5. As the current instruction-tuned Large Language Models exhibit an **ability to explain their reasoning** [Saeed and Omlin, 2023], the methods of eXplainable AI

(XAI) may also be integrated into our automated fact-checking framework, giving the fact-checker further insights what is behind the model classification.

6. Multiple **validation experiments** are planned with real-world fact-checkers[5] to testify to the usability of our solutions in the real world.

[5] Partners from CEDMO and other projects

Chapter 5
Conclusion

This study outlined current challenges and motivations—building automated support for fact-checking. Most solutions rely on transformers, the SOTA for almost every NLP task. The paradigm is shifting from *fine-tuning* pre-trained encoders/decoders to *prompting* and few-shotting instruction-tuned LLMs, which impacts this dissertation and requires modernizing prior work.

So far, we have collected several datasets—most notably CsFEVER and CTKFacts—deployed a working fact-checking pipeline, and released trained models for reuse.

Next, we aim to establish claim generation and its model-based metrics, conclude ongoing model training and data collection (Czech, English, Polish, Slovak), and propose updated end-to-end solutions leveraging modern LLMs.

The preceding chapters summarized what has been done, why it matters, the surrounding context, and the likely next steps.

Bibliography

[Allcott and Gentzkow, 2017] Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

[Aly et al., 2021] Aly, R., Guo, Z., Schlichtkrull, M. S., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., and Mittal, A. (2021). FEVEROUS: Fact extraction and VERification over unstructured and structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

[Banerjee and Lavie, 2005] Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

[Barua et al., 2020] Barua, Z., Barua, S., Aktar, S., Kabir, N., and Li, M. (2020). Effects of misinformation on covid-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Progress in Disaster Science*, 8:100119.

[Bowman et al., 2015] Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *CoRR*, abs/2005.14165.

[Buchanan and Benson, 2019] Buchanan, T. and Benson, V. (2019). Spreading disinformation on facebook: Do trust in message source, risk propensity, or personality affect the organic reach of “fake news”? *Social Media + Society*, 5(4):2056305119888654.

[Bútora, 2023] Bútora, R. (2023). Crowd-sourcing platform frontend for fact-checking. `https://dspace.cvut.cz/handle/10467/109505`.

[Chen et al., 2017] Chen, D., Fisch, A., Weston, J., and Bordes, A. (2017). Reading wikipedia to answer open-domain questions. *CoRR*, abs/1704.00051.

[Chen et al., 2023] Chen, J., Kim, G., Sriram, A., Durrett, G., and Choi, E. (2023). Complex claim verification with evidence retrieved in the wild.

[Cheng et al., 2016] Cheng, J., Dong, L., and Lapata, M. (2016). Long short-term memory-networks for machine reading. *CoRR*, abs/1601.06733.

[Choi et al., 2021] Choi, E., Palomaki, J., Lamm, M., Kwiatkowski, T., Das, D., and Collins, M. (2021). Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

[Conneau et al., 2019] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

[Conneau et al., 2018] Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. (2018). Xnli: Evaluating cross-lingual sentence representations.

[Dettmers et al., 2023] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Elsayed et al., 2021] Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Martino, G. D. S., and Atanasova, P. (2021). Overview of the clef-2019 checkthat!: Automatic identification and verification of claims.

[Fleiss, 1971] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

[Fu et al., 2023] Fu, J., Ng, S.-K., Jiang, Z., and Liu, P. (2023). Gptscore: Evaluate as you desire.

[Gažo, 2021] Gažo, A. (2021). Algorithms for document retrieval in czech language supporting long inputs.

[Guo et al., 2022] Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

[Hanselowski et al., 2018] Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., and Gurevych, I. (2018). A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[Hasan et al., 2021] Hasan, T., Bhattacharjee, A., Islam, M. S., Samin, K., Li, Y., Kang, Y., Rahman, M. S., and Shahriyar, R. (2021). Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *CoRR*, abs/2106.13822.

[Hayes and Krippendorff, 2007] Hayes, A. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1:77–89.

[Holtzman et al., 2020] Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration.

[Hu et al., 2021] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

[Ji et al., 2023] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

[Kocián et al., 2021] Kocián, M., Náplava, J., Štancl, D., and Kadlec, V. (2021). Siamese bert-based model for web search relevance ranking evaluated on a new czech dataset.

[Koto et al., 2020] Koto, F., Baldwin, T., and Lau, J. H. (2020). Ffci: A framework for interpretable automatic evaluation of summarization.

[Krippendorff, 1970] Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

[Krotil, 2022] Krotil, M. (2022). Text summarization methods in czech. `https://dspace.cvut.cz/handle/10467/101028`.

[Köpf et al., 2023] Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H., and Mattick, A. (2023). Openassistant conversations – democratizing large language model alignment.

[Lazer et al., 2018] Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

[Lehečka and Švec, 2021] Lehečka, J. and Švec, J. (2021). Comparison of czech transformers on text classification tasks. In Espinosa-Anke, L., Martín-Vide, C., and Spasić, I., editors, *Statistical Language and Speech Processing*, pages 27–37, Cham. Springer International Publishing.

[Lin, 2004] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[Liu et al., 2022a] Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., and Raffel, C. (2022a). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.

[Liu et al., 2023a] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023a). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

[Liu et al., 2020] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation.

[Liu et al., 2023b] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., and Ge, B. (2023b). Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2):100017.

[Liu et al., 2022b] Liu, Y., Liu, P., Radev, D., and Neubig, G. (2022b). BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

[Mlynář, 2023] Mlynář, T. (2023). Automated fact checking based on czech wikipedia. `https://dspace.cvut.cz/handle/10467/109219`.

[Mohri et al., 2023] Mohri, C., Andor, D., Choi, E., and Collins, M. (2023). Learning to reject with a fixed predictor: Application to decontextualization.

[Mroczkowski et al., 2021] Mroczkowski, R., Rybak, P., Wróblewska, A., and Gawlik, I. (2021). HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

[Nakov et al., 2021] Nakov, P., Martino, G. D. S., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Mansour, W., Hamdan, B., Ali, Z. S., Babulkov, N., Nikolov, A., Shahi, G. K., Struß, J. M., Mandl, T., Kutlu, M., and Kartal, Y. S. (2021). Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news.

[Narayan et al., 2018] Narayan, S., Cohen, S. B., and Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

[Nie et al., 2019a] Nie, Y., Chen, H., and Bansal, M. (2019a). Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

[Nie et al., 2019b] Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. (2019b). Adversarial NLI: A new benchmark for natural language understanding. *CoRR*, abs/1910.14599.

[Nørregaard and Derczynski, 2021] Nørregaard, J. and Derczynski, L. (2021). DanFEVER: claim verification dataset for Danish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 422–428, Reykjavik, Iceland (Online). Link"o"ping University Electronic Press, Sweden.

[OpenAI, 2023] OpenAI (2023). Gpt-4 technical report.

[Pan et al., 2021] Pan, L., Chen, W., Xiong, W., Kan, M.-Y., and Wang, W. Y. (2021). Zero-shot fact verification by claim generation.

[Patel and Ahmad, 2023] Patel, D. and Ahmad, A. (2023). Google "we have no moat, and neither does openai". `https://www.semianalysis.com/p/google-we-have-no-moat-and-neither`. Accessed: 2023-09-06.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

[Pikuliak et al., 2021] Pikuliak, M., Štefan Grivalský, Konôpka, M., Blšták, M., Tamajka, M., Bachratý, V., Šimko, M., Balážik, P., Trnka, M., and Uhlárik, F. (2021). Slovakbert: Slovak masked language model.

[Pomerlau and Rao, 2017] Pomerlau, D. and Rao, D. (2017). Fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. `http://www.fakenewschallenge.org`. Accessed: 2023-09-06.

[Popel et al., 2020] Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., and Žabokrtský, Z. (2020). Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15.

[Raffel et al., 2019] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

[Rajpurkar et al., 2016] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

[Rýpar, 2021] Rýpar, M. (2021). Methods of document retrieval for fact-checking. `https://www.overleaf.com/read/thbvcjvvvfjp`. [Online; accessed 21-May-2021].

[Saeed and Omlin, 2023] Saeed, W. and Omlin, C. (2023). Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273.

[Schuster et al., 2021] Schuster, T., Fisch, A., and Barzilay, R. (2021). Get your vitamin c! robust fact verification with contrastive evidence. *CoRR*, abs/2103.08541.

[Sebastian, 2023] Sebastian, G. (2023). Exploring ethical implications of chatgpt and other ai chatbots and regulation of disinformation propagation.

[Semin, 2023] Semin, D. (2023). Multitask learning for nlp classifiers. `https://dspace.cvut.cz/handle/10467/109243`.

[Sido et al., 2021] Sido, J., Pražák, O., Přibáň, P., Pašek, J., Seják, M., and Konopík, M. (2021). Czert – czech bert-like model for language representation.

[Štefánik et al., 2023] Štefánik, M., Kadlčík, M., Gramacki, P., and Sojka, P. (2023). Resources and few-shot learners for in-context learning in Slavic languages. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 94–105, Dubrovnik, Croatia. Association for Computational Linguistics.

[STEM, 2021] STEM (2021). Mýtům a konspiracím o covid-19 věří více než třetina české internetové populace | stem.cz. `https://www.stem.cz/mytum-a-konspiracim-o-covid-19-veri-vice-nez-tretina-ceske-internetove-populace/`. Accessed: 2021-05-03.

[Straka et al., 2018] Straka, M., Mediankin, N., Kocmi, T., Žabokrtský, Z., Hudeček, V., and Hajič, J. (2018). SumeCzech: Large Czech news-based summarization dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

[Straka et al., 2021] Straka, M., Náplava, J., Straková, J., and Samuel, D. (2021). RobeCzech: Czech RoBERTa, a monolingual contextualized language representation model. *Lecture Notes in Computer Science*, page 197–209.

[Stănescu, 2022] Stănescu, G. (2022). Ukraine conflict: the challenge of informational war. *SOCIAL SCIENCES AND EDUCATION RESEARCH REVIEW*, 9(1):146–148.

[Szwoch et al., 2022] Szwoch, J., Staszkow, M., Rzepka, R., and Araki, K. (2022). Creation of Polish online news corpus for political polarization studies. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 86–90, Marseille, France. European Language Resources Association.

[Taori et al., 2023] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Alpaca: A strong, replicable instruction-following model. `https://crfm.stanford.edu/2023/03/13/alpaca.html`. Accessed: 2023-09-04.

[DeepL, 2021] DeepL (2021). Deepl translator. `https://www.deepl.com/en/translator`. Accessed: 2021-05-09.

[Google, 2021] Google (2021). Cloud translation - google cloud. `https://cloud.google.com/translate`. Accessed: 2021-05-09.

[NLP-Progress, 2023] NLP-Progress (2023). On summarization. `http://nlpprogress.com/english/summarization.html`. Accessed: 2023-09-06.

[Thorne et al., 2018a] Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018a). FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

[Thorne et al., 2018b] Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2018b). The Fact Extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*.

[Thorne et al., 2019] Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2019). The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

[Touvron et al., 2023a] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023a). Llama: Open and efficient foundation language models.

[Touvron et al., 2023b] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023b). Llama 2: Open foundation and fine-tuned chat models.

[Ullrich, 2021] Ullrich, H. (2021). Dataset for automated fact checking in czech language. `https://dspace.cvut.cz/handle/10467/95430`.

[Ullrich et al., 2023] Ullrich, H., Drchal, J., Rýpar, M., Vincourová, H., and Moravec, V. (2023). Csfever and ctkfacts: acquiring czech data for fact verification. *Language Resources and Evaluation*.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

[Vicuna, 2023] Vicuna (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. `https://vicuna.lmsys.org/`. Accessed: 2023-09-04.

[Wang et al., 2020] Wang, A., Cho, K., and Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

[Wardle and Derakhshan, 2017] Wardle, C. and Derakhshan, H. (2017). *INFORMATION DISORDER : Toward an interdisciplinary framework for research and policy making Information Disorder Toward an interdisciplinary framework for research and policy-making*.

[Williams et al., 2018] Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

[Wright et al., 2022] Wright, D., Wadden, D., Lo, K., Kuehl, B., Cohan, A., Augenstein, I., and Wang, L. L. (2022). Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

[Yamada et al., 2020] Yamada, I., Asai, A., Shindo, H., Takeda, H., and Matsumoto, Y. (2020). Luke: Deep contextualized entity representations with entity-aware self-attention.

[Yasunaga et al., 2021] Yasunaga, M., Leskovec, J., and Liang, P. (2021). LM-critic: Language models for unsupervised grammatical error correction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7763, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Zha et al., 2023] Zha, Y., Yang, Y., Li, R., and Hu, Z. (2023). Alignscore: Evaluating factual consistency with a unified alignment function.

[Zhang et al., 2020] Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

[Zhang* et al., 2020] Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

[Zhao et al., 2023] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2023). A survey of large language models.

[Šuppa and Adamec, 2020] Šuppa, M. and Adamec, J. (2020). A summarization dataset of Slovak news articles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6725–6730, Marseille, France. European Language Resources Association.

Appendix A
Acronyms

BERT Bidirectional Encoder Representations from Transformers

GPT Generative Pre-trained Transformer

FEVER Fact Extraction and Verification – series of Shared tasks focused on fact-checking

IR Information Retrieval

SOTA State of the Art

XSum Extreme Summarization – summarizing article into one sentence

NLI Natural Language Inference

ČTK Czech Press Agency