Czech
Technical University
in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science

Fact-Guided Text Summarization for Czech

Master Thesis of
Bc. Marian Krotil

Supervisor: **Ing. Jan Drchal, Ph.D.**
Field of study: **Open Informatics**
Subfield: **Artificial Intelligence**
May 2024

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Krotil Marian** Personal ID number: **492001**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Computer Science**

Study program: **Open Informatics**

Specialisation: **Artificial Intelligence**

II. Master’s thesis details

Master’s thesis title in English:

Fact-Guided Text Summarization for Czech

Master’s thesis title in Czech:

Metody sumarizace českých textů podporující fakticitu

Guidelines:

Recent advancements in large language models significantly improved state-of-the-art results in text summarization. Unfortunately, LLM-generated summaries tend to contain factually incorrect information, which limits their practical use. The project aims to 1) improve existing methods to detect factually incorrect summaries, and 2) use a method of choice to improve facticity of abstractively generated summaries. Focus on the Czech language.
1) Explore methods evaluating summarization quality, such as the model-based AlignScore. Explore state-of-the-art text summarization methods aimed to promote facticity, such as BRIO.
2) Develop and implement a facticity evaluation model-based metric for the Czech language.
3) Integrate the facticity evaluation metric to the summarization model.
4) Evaluate summaries using data supplied by the supervisor.

Bibliography / sources:

[1] Zha, Yuheng, et al. "AlignScore: Evaluating Factual Consistency with a Unified Alignment Function." arXiv preprint arXiv:2305.16739 (2023).
[2] Liu, Yixin, et al. "BRIO: Bringing Order to Abstractive Summarization." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.
[3] Dixit, Tanay, Fei Wang, and Muhao Chen. "Improving Factuality of Abstractive Summarization without Sacrificing Summary Quality." arXiv preprint arXiv:2305.14981 (2023).
[4] Chern, I-Chun, et al. "Improving Factuality of Abstractive Summarization via Contrastive Reward Learning." arXiv preprint arXiv:2307.04507 (2023).
[5] Straka, Milan, et al. "SumeCzech: Large Czech news-based summarization dataset." Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.

Name and workplace of master's thesis supervisor:

Ing. Jan Drchal, Ph.D. Artificial Intelligence Center FEE

Name and workplace of second master's thesis supervisor or consultant:

Date of master's thesis assignment: **01.02.2024** Deadline for master's thesis submission: _____

Assignment valid until: **21.09.2025**

Ing. Jan Drchal, Ph.D. Supervisor's signature	Head of department's signature	prof. Mgr. Petr Páta, Ph.D. Dean's signature

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

Date of assignment receipt	Student's signature

Acknowledgements

I extend my gratitude to my supervisor, Ing. Jan Drchal, Ph.D., for his invaluable guidance, support, and his friendly attitude through this journey. I am also deeply grateful to my colleague, Bc. Martin Hubal, for our collaborative efforts and for providing access to additional computational resources.

Additionally, I am immensely grateful for the unwavering support of my girlfriend, Lily, during the challenging writing days. I hope to repay her kindness one day. Additionally, I extend heartfelt thanks to both our families for their generous support, both financially and emotionally, throughout our academic journey. Also, I am grateful to myself for maintaining focus and determination to complete this work.

Last but not least, I thank to the Research Center For Informatics[1] *for providing their computational resources, particularly the hardware graphics cards used for training the models. Moreover, I am also thankful for the support received from the Czech News Center*[2] *for contributing data that enriches Czech summarization.*

[1] `https://rci.cvut.cz/`
[2] `https://www.cncenter.cz/`

Declaration

I hereby declare that I have completed this thesis independently and that I have listed all the literature and publications used. I have no objection to usage of this work in compliance with the act §60 Zákon č. 121/2000Sb. (copyright law), and with the rights connected with the copyright act including the changes in the act.

In Prague, 24. May 2024

Abstract

Advancements in natural language processing have been driven by the pretraining of large language models on vast multilingual corpora in recent years, enabling them to process texts in less covered languages, such as Czech, which is the main focus of this work. The summarization task is thus facilitated and produced summaries resemble human writing. However, despite these advancements, state-of-the-art models often struggle with issues such as hallucination, contradiction, and the propagation of false information. Moreover, a lack of Czech factual metrics capturing these disparities exacerbates the problem. This work addresses these challenges through two key contributions: 1) the design of a multitask multilingual factual metric, AlignScoreCS, capable of assessing various tasks, including summarization, in both Czech and English languages, and 2) the introduction of a factual refinement technique, BARF: BRIO paradigm with AlignScoreCS and ROGUE$_{RAW}$ Fusion, designed for summarization models to produce factual summaries in Czech and English. The evaluation of the factual metric demonstrated comparable scores to its English counterpart, outperforming other existing metrics and establishing itself as the most promising Czech factual metric to date. BARF models updated specific state-of-the-art results for the SumeCzech dataset while ensuring factual summaries. Conducted human evaluation confirmed the enhancement in the factuality of generated summaries and the correlation of the factual metric with human judgment.

Keywords: Factuality Metric; Abstractive Summarization; Summarization Facticity; Czech; English; AlignScoreCS; BARF; NLI; Natural Language Inference; NLP; BRIO; SumeCzech; CNC; XSUM; CNNDM; SummaC; TRUE; mBART

Supervisor: Ing. Jan Drchal, Ph.D.

Abstrakt

Pokrok ve zpracování přirozeného jazyka byl v posledních letech dosažen díky předtrénování velkých jazykových modelů na rozsáhlých vícejazyčných korpusech, což jim umožnilo zpracovávat texty v méně pokrytých jazycích, jako je například čeština, na kterou se tato práce zaměřuje. Úloha sumarizace je tím usnadněna a vytvářené sumarizace se podobají lidmi psanému textu. I přes tyto pokroky se však state-of-the-art modely často potýkají s problémy, jako je halucinace, kontradikce a šíření nepravdivých informací. Problém navíc zhoršuje nedostatek českých faktických metrik postihujících tyto nesrovnalosti. Tato práce řeší tyto výzvy dvěma klíčovými příspěvky: 1) návrh víceúlohové vícejazyčné faktické metriky AlignScoreCS, která je schopna vyhodnocovat různé úlohy, včetně sumarizací, v českém i anglickém jazyce, a 2) představení techniky faktického zlepšení BARF: BRIO paradigma s AlignScoreCS a ROGUE$_{RAW}$ Fúzemi, navržené pro sumarizační modely k produkci faktických souhrnů v češtině a angličtině. Vyhodnocení faktické metriky ukázalo srovnatelné výsledky s jejím anglickým protějškem, čímž překonala ostatní existující metriky a prosadila se jako dosud nejslibnější česká faktická metrika. Modely BARF aktualizovaly některé state-of-the-art výsledky pro dataset SumeCzech a zároveň zaručily faktičnost v sumarizacích. Provedené lidské hodnocení potvrdilo zlepšení fakticity generovaných sumarizací a korelaci faktické metriky s lidským úsudkem.

Klíčová slova: Faktická Metrika; Abstraktivní Sumarizace; Fakticita Sumarizace; Čeština; Angličtina; AlignScoreCS; BARF; NLI; NLP; BRIO; SumeCzech; CNC; XSUM; CNNDM; SummaC; TRUE; mBART; LLM

Contents

Figures

Tables

Chapter 1
Introduction

Recent transformer architecture-based models have played a significant role in the field of natural language processing (NLP), constantly achieving state-of-the-art (SOTA) results across many natural language generation (NLG) tasks [Lewis et al., 2019, Devlin et al., 2018, Liu et al., 2019]. The models have the capability to generate coherent texts that closely resemble articles written by humans. By multilingual training on expansive corpora of documents including various sources of languages, models acquire a broad understanding of language dependencies, including those that are less commonly covered, such as Czech, which is the focus in this work [Devlin et al., 2018, Liu et al., 2020, Conneau et al., 2019]. However, investigations into natural language generation reveal that the texts generated using SOTA models frequently suffer from factual inconsistencies bearing issues of hallucination, contradiction, false information, and many others [Ji et al., 2022].

In the course of summarization, models tend to either produce knowledge gathered from another article, inaccurately replace names or numbers, or swap subjects with objects. These tendencies frequently lead to unfaithful summaries. Such discrepancies could misinform readers, propagating fake information, a concern we want to address. Table 1.1 illustrates an example of a misleading summary alongside a summary produced by our factually refined summarization model, demonstrating improvement. Additionally, the summaries are ranked based on factual accuracy using our newly developed metric. In the domain of English summarization, several studies [Dixit et al., 2023, Chern et al., 2023] have explored the notion of summarization factuality. Furthermore, research conducted in Czech [Halama, 2023] has also addressed this issue, incorporating the concept of integrating factual metrics into the training phase of summarization models [Liu et al., 2022]. However, effective factual metrics are predominantly tailored to English, leaving a gap in the assessment of summary factuality for the Czech language, as highlighted by the study conducted by [Halama, 2023].

In this thesis, we introduce a novel **factual metric**, **AlignScoreCS**, capable of assessing entailment and factuality across various NLP tasks in both English and Czech languages. Results demonstrate that the metric achieves comparable performance to its SOTA English brothers and outperforms all other existing models, as evidenced by high scores on benchmarks incorporating summarization data. Building upon this, we propose several **factually refined models**, collectively termed **BARF** (BRIO paradigm with AlignScoreCS and ROUGE fusion). Our models incorporate both quality and factuality metrics in their training objectives. Results obtained on both English and Czech datasets highlight the potential of these models to enhance both the factuality and quality of generated summaries. Notably, they achieve comparable performance to English models and even surpass specific current state-of-the-art results on the SumeCzech dataset, all while maintaining factuality.

Source Text
O tom, jaké má potíže, promluvil Petr Štěpánek nedávno v Blesku: "Pořád to cítíte, ale hlavně se na tu nohu nemůžete absolutně spolehnout. To znamená, že všechny přemrštěné pohyby mám stále zakázané. Operace mě čeká už v létě." Herec přičítá své zdravotní problémy zátěži, kterou musely jeho klouby snášet v mládí, kdy aktivně sportoval.Jenže herec je známý dříč a jeho manželka Zlata Adamovská už tuší, že by i po další operaci nemuselo dojít k nápravě, pokud se Štěpánek nebude šetřit. "Slyšel jsem, že je z toho Zlata hodně špatná. Prý se bojí, aby to nedopadlo tak, že bude Petr celý zbytek života kulhat. Mělo se jí o tom i zdát," řekl nejmenovaný zdroj Šípu a dodává: "Zlata se bojí, aby nezopakoval tu samou chybu, co loni. Protože bude mít nový kloub, myslí si, že už není čas pokoušet znovu osud tím, že ho bude přepínat.Adamovská už si prý vymyslela plán, jak Štěpánka udržet ve formě. "Tipoval bych, že možná i vymyslí nějakou pěknou dovolenou, těžko říct. Zatím mi neprozradila, co chystá. Ona má ráda překvapení, takže o tom asi nebude mluvit. Potřebuje mu do hlavy zanést pochybnosti, že činnost je nejlepší lék. Nechce, aby toho pak litoval až do konce života, takže je odhodlaná udělat v podstatě cokoli," prozradil informátor.
Summaries
Reference Je to deset měsíců, co musel herec Petr Štěpánek (66) podstoupit operaci kolene. Jenže se kýžený výsledek nedostavil, a herec dál trpí bolestmi. Herec, který stále naplno pracuje se tak bude muset podrobit dalšímu chirurgickému zákroku a to dělá vrásky na čele jeho choti Zlatě Adamovské (56). Bojí se totiž, že se herec bude i nadále přepínat a operace tak bude zase k ničemu.
mBART Incorrect 0.22 Herec a dabér Petr Štěpánek (66) se už několik let potýká s vážnými zdravotními problémy. Loni mu praskla kyčelní kloub a musel podstoupit operaci, při níž mu lékaři museli vyměnit obě nohy. Nyní musí na další operaci.
BARF-align Correct 0.96 Herec Petr Štěpánek (66) má stále problémy s klouby. Operace kloubů ho čeká už v létě. Herec je známý dříč a jeho manželka Zlata Adamovská už tuší, že by po další operaci nemuselo dojít k nápravě.

Table 1.1: Example of two summaries, one produced by mBART25 core model, and one by our factually refined BARF-Align model, annotated with human label and scored by AlignScoreCS. Contradicting sequences are highlighted in red.

Furthermore, human evaluation confirms the factual improvements and demonstrates the correlation between AlignScoreCS and human judgment. We aim to contribute to Czech research on summarization and facticity assessment, with the hope that our models and metric will find practical applications in the field.

1.1 Overview

Initially, we introduce the state-of-the-art transformer model we utilize. Subsequently, our thesis is divided into two parts, each addressing a specific NLP task. Below is a brief overview of each part:

1. **Part I, Advancing Facticity Assessment in Czech Text Summarization**: Here, we focus on factuality checking and the development of AlignScoreCS. We outline our methodology, discuss the datasets used for training and testing, and present our results along with a discussion.

2. **Part II, Towards Facticity-Driven Text Summarization in Czech**: Here, we shift our focus to text summarization. Building on the research from Part I, we integrate AlignScoreCS into our BARF summarization models. Next, we describe datasets and approaches and provide results, along with a human evaluation and discussion.

Chapter 2
Preliminaries

This chapter explores the Transformer architecture, a pivotal advancement in natural language processing. We provide an overview of state-of-the-art models capable of being utilized in the Czech language. These models are derived from the Transformer architecture, and we categorize them into three types: Encoder, Encoder-Decoder, and Decoder models. In practical applications, these models are pre-trained on vast corpora of texts from various language sources. Consequently, they can develop robust, high-dimensional text representations and intuitively understand languages and diverse writing styles. As a result, these models are typically fine-tuned for downstream tasks such as summarization, text generation, and more. Thanks to the transfer of knowledge from previous pre-training, these models can rapidly understand and learn these tasks, often requiring less data.

Our research builds upon and utilizes these robust architectures to enhance factual assessment and summarization tasks. In Part I, we primarily utilize the XLM-RoBERTa encoder-only model. This model's architecture is adept at encoding information efficiently through the encoder, which helps the model to understand the entailment of texts. In Part II, our focus shifts to Encoder-Decoder and Decoder-only models, specifically mBART, mT5, Falcon, and Aya. We aim to tackle generation tasks by leveraging the decoding capabilities of these models. Although all the models used are considered as large language models (LLMs), we will reserve the term LLMs exclusively for Aya and Falcon due to their significantly larger sizes. All models listed below are available on the HuggingFace Transformer hub[1].

2.1 Transformer Architecture

The Transformer architecture, as introduced by [Vaswani et al., 2023], was developed to overcome the bottleneck problem of Recurrent Neural Networks and their inability to be computed in parallel. This architecture marked a significant advancement in Natural Language Processing and has since become widely used across the field and beyond. Originally, the transformer consists of an encoder and a decoder, as depicted in Figure 2.1. However, it can now function as just an encoder or decoder, serving different purposes. Each part comprises several stacked transformer blocks, typically six, twelve, or more, each containing Multi-Head Attention, normalization functions, residual connections, and a feed-forward layer with a non-linear function.

[1] `https://huggingface.co/models`

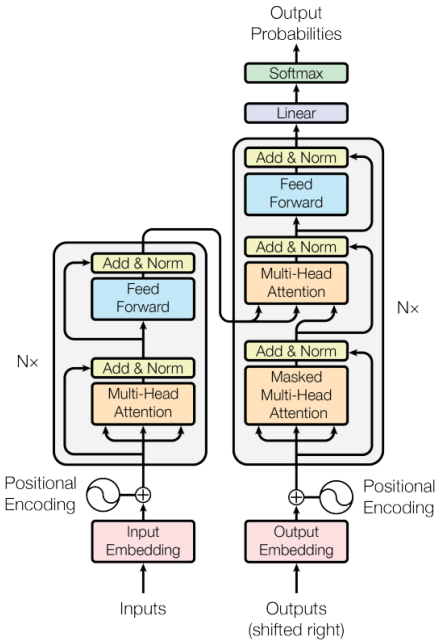

Figure 2.1: The figure depicts Transformer Architecture. The encoder part is displayed on the left side, whereas the decoder is shown on the right. Source from [Vaswani et al., 2023].

Multi-Head Attention, a key innovation, involves concatenating multiple Attention mechanisms transformed in a desired dimension space, enabling the model to handle long-distance dependencies between words. This attention mechanism establishes direct connections within the text, whether in the Encoder, Decoder or the part where information is passed from Encoder to Decoder, allowing the model to focus on specific segments of different sequences. We compute the Multi-Head Attention as follows:

$$\begin{aligned}\text{MultiHead(Q,K,V)} &= \text{Concatenate}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \\ \text{Attention}(Q, K, V) &= \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V\end{aligned}$$

Where Q, K, and V are matrices corresponding to queries, keys, and values with dimensions of $d_q = d_k$, and d_v, respectively, each created from a matrix of inputs. The Attention function is thus a weighted sum of values with weights computed from a query and its corresponding key. The $\frac{1}{\sqrt{d_k}}$ is a scaling factor to preserve a unit variance. Subsequently, W_i^Q, W_i^K, W_I^V are matrices with different learnable parameters that project the Q, K, V into desired dimensions, typically d_{model} which is the model input dimension and to a dimension of the corresponding matrix. Each head_i represents the Attention function from a diverse perspective using distinct W_i matrices. Multi-head attention is then a concatenation of outputs yielded by Attention functions projected into a dimension of $hd_v \times d_{model}$ via W^O parameter matrix.

In general, positional embedding is calculated and added to the input embedding to convey positional information for each word. Input embeddings represent input tokens in a vector space, where tokens can signify individual words or sub-words. These embeddings are adjusted during the model's training process.

The Decoder component of the Transformer architecture diverges from the Transformer block by incorporating an extra Multi-Head Attention mechanism and Masked Multi-Head Attention. The additional Multi-Head Attention integrates an output computed by the encoder as the Query and Keys, along with the generated output computed from the Masked Multi-Head Attention as Values. This helps the model receive information from the input, particularly in terms of summarization, as it gathers information from the text that needs to be summarized. The Masked Multi-Head Attention operates similarly to Multi-Head Attention but with a modification: it solely focuses on the outputs generated by the decoder by masking the future outputs. This characteristic is denoted as an auto-regressive decoder, where the model generates the next output based on both the input and previously generated outputs. To generate output using the decoder, we typically add an extra linear layer followed by a softmax function on top of the decoder. This creates a probability distribution over the vocabulary size for generating the next output. From this distribution, we can calculate the cross-entropy loss commonly used for training these models.

2.2 Encoder Models

Encoder models solely consist of the Encoder component from the Transformer architecture, focusing on learning a text representation in a high-dimensional space. This enables them to comprehend text from various perspectives. Additionally, the encoder models learn representations bidirectionally, meaning they understand texts from both right to left and left to right contexts. This contrasts decoder models, which only learn text in a left-to-right manner. In classification tasks, where we categorize texts into different types, we can introduce a linear layer on top of the architecture. The classifier developed this way is then simply trained using the features generated by these encoders as inputs. Typically, we update the encoder's weights during the training as well.

First, we introduce the **mBERT** model, developed by [Devlin et al., 2018]. Using a self-supervised approach, this model is a pre-trained multilingual version of BERT, trained on an extensive corpus of 104 languages with the largest Wikipedias. This means it was pre-trained on raw texts only, using an automatic process to generate inputs and labels from those texts. During pretraining, mBERT learns to predict masked words (Masked Language Modeling objective - MLM) where 15% of words in the input are randomly masked. Additionally, it trains on a next-sentence prediction (NSP) objective, predicting whether two sentences follow each other or not. These tasks enable mBERT to learn rich language representations across multiple languages.

Next, we discuss the **XLM-RoBERTa** model introduced by [Conneau et al., 2019]. This model is a multilingual variant of RoBERTa [Liu et al., 2019], which is based on BERT, but XLM-RoBERTa employs different pretraining objectives and training settings. It applies dynamic masking and next-sentence prediction every time input is processed by the model. XLM-RoBERTa is pre-trained on data from CommonCrawl, including 100 languages, allowing it to capture diverse language features and nuances. In Part I, we use a pre-trained model XLM-RoBERTa-large with 24 transformer layers in Encoder, a model dimension of 1024, 12 attention heads, and, in total, 550M learnable parameters.

2.3 Encoder-Decoder Models

Models that incorporate both components of the Transformer architecture, namely the Encoder and Decoder parts, follow a specific flow: the input is first processed by the encoder, which produces an output. This output is then passed to the decoder, which generates the final output in an auto-regressive manner. These models not only learn high-dimensional representations through the encoder but also excel at generation tasks through the decoder. This is typically achieved by training the decoder to predict the following words in the sentence, creating a sequence of outputs that form the final sentence.

M2M100, introduced by [Fan et al., 2020], represents a multilingual translation model designed for many-to-many translation tasks. This model was trained on a vast dataset containing 100 languages sourced from CommonCrawl through large-scale mining. The authors focused on a non-English-centric model, particularly utilizing data where either the source or target language differs from English, in contrast to other multilingual models.

mBART, short for multilingual BART [Lewis et al., 2019], was developed by [Liu et al., 2020]. The pretraining methodology of this model includes token masking, text infilling, and sentence permutation algorithms. Additionally, the authors trained the model by corrupting the last sequences in the document and requiring the decoder part of the model to predict the output. They used a substantial corpus comprising 25 languages sourced from Common Crawl for the training. We utilize the checkpoint mBART-cc25-large, featuring 12 layers of encoder and 12 layers of decoder, with a model dimension of 1024 and 16 attention heads, totaling 680 million learnable parameters.

mT5, a multilingual variant of T5 [Raffel et al., 2023], was designed by [Xue et al., 2021] and trained on the multilingual corpora mC4, which covers 101 languages sourced from Common Crawl. The T5 training methodology revolves around a unified text-to-text format for all text-based NLP tasks. This approach offers the advantage of using the same training objectives for every task. Additionally, it is pre-trained using masked language modeling, similar to other models. We utilize the checkpoint of mt5-base, featuring 12 layers of an encoder and 12 layers of a decoder, a dimension size of 768, and 12 attention heads, totaling 580 million learnable parameters.

Aya is a multilingual large language model trained to follow instructions in 101 languages designed by [Üstün et al., 2024]. This model was constructed by fine-tuning the mT5-xxl model, which has 13 billion parameters, on the large cross-lingual prompt dataset xP3x. Aya has shown superior performance over other large language models (LLMs) on a wide range of tasks across various languages. We have selected this model because of its capability to process Czech texts.

2.4 Decoder Models

Finally, we delve into the Decoder models, which utilize only the decoder part of the Transformer architecture and are typically denoted as auto-regressive models. These models are pre-trained by predicting the next word in a sentence conditioned on some previous context. Their powerful ability lies in text generation, allowing them to produce coherent and contextually relevant text. In recent years, decoder models have become large lan-

guage models that contain billions of learnable parameters. As a result, the pretraining datasets required to train these models have expanded to trillions of tokens, making the computational requirements extremely expensive.

Falcon is a large language model (LLM) trained on the RefinedWeb dataset by [Penedo et al., 2023], which consists of adequately filtered and deduplicated web data. The authors of Falcon emphasized the quality of the training data, demonstrating that using training settings similar to GPT-3 with this well-filtered data resulted in significant enhancements. While Falcon performs well on various tasks, its pretraining dataset offers limited support for Czech. Nevertheless, we have chosen to use this model for our experiments.

Part I

Advancing Facticity Assessment in Czech Text Summarization

Chapter 3
Factuality

Automatic factual evaluation metrics designed to tackle facticity errors are proposed in numerous papers. The majority of these metrics are based on natural language inference (NLI) dealing with the task of determining whether a "hypothesis" (claim) is entailed with "premise" (context) [Devlin et al., 2018, Liu et al., 2019, Kryscinski et al., 2019]. This task is typically addressed by classification into different classes of text entailment. Recent studies have expanded the scope of evaluation by incorporating question answering (QA) based metrics [Fabbri et al., 2021b, Honovich et al., 2021]. Natural language inference based metrics as well as QA-based metrics achieve state-of-the-art results mainly on a particular dataset task that they have been trained on. However, their performance is significantly reduced when testing on dataset tasks and domains varying from training. The crucial problem lies in its task-specific settings, which limits information across diverse domains, encompassing various factuality errors, distinct writing styles, and different input texts of varying lengths [Laban et al., 2021].

To overcome this challenge, Yuheng Zha et al. have introduced a comprehensive metric named AlignScore, outlined in the paper "ALIGNSCORE: Evaluating Factual Consistency with A Unified Alignment Function" [Zha et al., 2023]. This metric offers a holistic approach by incorporating a unified text-to-text information alignment function. AlignScore generalizes factual consistency and the evaluation space to encompass multiple data sources and considerable data heterogeneity. Nevertheless, this approach is language-specific and focuses only on English data sources, which makes this model unsuitable for Czech data.

Regarding the Czech language, several evaluation metrics approach the task of factual inconsistency and text incorrectness [Drchal et al., 2022, Šimon Zvára, 2022, Víta, 2020]. These metrics mainly rely on multilingual models pre-trained on extensive corpora, encompassing texts in multiple languages, including Czech documents [Conneau et al., 2019, Devlin et al., 2018]. Subsequently, these models are fine-tuned for specific Czech NLI tasks. Although these models perform reasonably well, a standard limitation is their task-specific application, often resulting in a lack of information when applying them to diverse data sources and domains.

In this part of the thesis, we propose a multilingual metric called **AlignScoreCS**, which focuses on Czech and English data sources. Our metric's training procedure and design are primarily based on the recommendations of AlignScore paper [Zha et al., 2023] and on additional incorporation of robust dataset extension, multilingual model utilization, and different multitask learning approaches. AlignScoreCs method can thus factually score documents of different tasks in different languages by passing context and claim pairs:

context as the source document and claim as the generated text. The claim can represent a summarization, a question-answer pair, an extraction of claims, or any other contextually coherent texts that are factually comparable with their contexts. By accomplishing this behavior, we trained AlignScoreCS by fine-tuning the XLM-Roberta model on a large multilingual corpus consisting of datasets from various tasks, including Natural Language Inference (NLI), Question Answering (QA), paraphrasing, fact verification, information retrieval, semantic similarity, and summarization. These datasets have been unified into three main tasks, resulting in a multitask learning scenario.

Furthermore, Me and my colleague, Martin Hubal, who was also involved in the AlignScoreCS project, introduce new **Czech benchmarks** and **Czech datasets for specific tasks** developed by translating their English representatives into Czech language using automatic machine translation models Deepl [Kutylowski, 2017] and SeamlessM4T [Communication et al., 2023], respectively. My colleague and I worked collaboratively to develop these training datasets and benchmarks. Together, we also managed the translation process of English datasets into Czech. Our collective work is mentioned in appropriate sections. However, while we collaborated on building and sharing datasets for training and testing, we pursued our respective work independently. Moreover, our combined work concludes with the development of the datasets; afterward, each of us utilizes distinct architectures as well as diverse training scenarios independently.

In addition, we also demonstrate the results of our AlignScoreCS on both English and Czech versions of benchmarks, including SummaC and TRUE. In contrast to the AlignScore model [Zha et al., 2023], our AlignScoreCS model reaches comparable results; nevertheless, it exhibits lower performance on specific datasets but still outperforms other existing metrics. Besides, we establish new baseline results for benchmarks SummaC and TRUE for their Czech versions.

3.1 Text Classification & Facticity

Text classification is a dynamic challenge across multiple fields and disciplines, reflecting its ongoing importance and influence. One of the most common topics is considered an email filtering system. Where the system receives an email about which it determines whether it is spam (unwanted email) or ham (legitimate email). Another well known representative from the sphere of text classification problem involves sentiment analysis, where texts are categorized and organized as positive, negative or neutral, which can represent a categorization system of movie comments.

Formally, the text classification problem is situated in the domain of Natural Language Processing (NLP). The objective is to assign a specific class for a given text based on previously learned knowledge, described as:

$$A(t) \rightarrow c \in C,$$

Where A stands for an algorithm with acquired knowledge from the learning process, t denotes a text, C represents a class space, and c is a specific class from the Class space.

Furthermore, when we condition the text with respect to its context, we delve into the domain of Natural Language Inference (NLI), one of the fundamental tasks in NLP. The

objective is to determine the semantic relationship or entailment between two text fragments. When it comes to facticity, we address the facticity classification problem by checking whether two texts are factually consistent, essentially, whether a claim is factually consistent (or aligned) with its context, formulated as:

$$A(context, claim) \rightarrow c \in C,$$

where A stands for an algorithm with acquired knowledge from the learning process, usually a transformer architecture in this domain, *context* and *claim* are self-describing, C represents the class space, and c is the specific class. In this work, we refer to this formulation whenever the classification problem is mentioned. We use terms such as 2-way or 3-way classification, which refer to the class space C with two or three specific classes, respectively. Regarding the regression task of two texts entailment, we can formulate it as the classification task with continuous class space C normalized into a unit interval, where the specific classes are usually further categorized into two upper classes (Positive, Negative) based on the specific class values surpassing a predefined threshold.

For simplicity, the following list summarizes the definitions of used classification problems in this work:

- $A(context, claim) \rightarrow c \in [0 \dots 1]$
  Regression task (reg), where the predicted class c comes from an interval. Regarding the factual decision problem, we can imply a facticity class from the number represented by c as follows: $c \geq 0.5$ is factually consistent, and $c < 0.5$ is factually inconsistent.

- $A(context, claim) \rightarrow c \in \{Contradict, Aligned\}$
  Binary classification task, we denote it as 2-way classification (2-way). The *Aligned* class represents factual consistency, and the *Contradict* class stands for factual inconsistency.

- $A(context, claim) \rightarrow c \in \{Contradict, Aligned, Neutral\}$
  Ternary classification task, also known as 3-way classification (3-way), where *Aligned* is factually consistent, *Contradict* is factually inconsistent, and *Neutral* is factually undecidable. It means that the context and the claim are unrelated.

Chapter 4
Related Work

Within this chapter, we describe recent methodologies that significantly contribute to our research. Our discussion begins with a description of the fundamental aspects of multi-task learning. Subsequently, we analyze the AlignScore paper [Zha et al., 2023] that serves as an essential reference shaping our research direction, followed by baseline models and metrics. Lastly, we conclude with a discussion encompassing the methods used in the domain of machine translation.

4.1 Multi-Task Learning

Multi-task learning (MTL) is a branch of machine learning where a single algorithm learns multiple tasks [Crawshaw, 2020, Raffel et al., 2023, Zhang et al., 2023]. This algorithm can be understood as a shared model or architecture that is applied among distinct tasks. According to the paper [Crawshaw, 2020], the concept of multi-task learning within a single model framework, associated with the simultaneous development and utilization of shared representations across tasks, significantly enhances the performance of the models, additionally can reduce overfitting and accelerates learning by inferring information across different tasks.

In terms of optimization for multi-task learning, there are two principal attitudes to parameter sharing across task-specific models. One method is hard parameter sharing, which practices sharing model weights between multiple tasks, where each learnable parameter is adjusted to minimize multiple loss functions. Another method is soft parameter sharing, where the multi-task model is composed of several task-specific models with separate weights that are bounded from each other by distance in the joint objective function. In this work, we handle precisely the attitude to hard parameter sharing, where we train a shared architecture. Following the paper [Crawshaw, 2020], a fundamental approach to balance the individual loss functions for different tasks is loss weighting. One of the proposed methods to efficiently execute the aggregated loss function is a weighted sum of the task-specific loss functions, which we utilized in our implementation of this work.

In regard to the paper [Zhang et al., 2023], two approaches leveraging data proportion exist in multi-task learning, one being the heterogeneous batch training scheme, in which the model is fed with batches containing samples from different tasks. This approach infers an inner model structure to handle the diverse data within a single batch. While the other represents a homogeneous training scenario, where the training batch consists of samples of one specific task. This perspective involves additional balanced data ratio sampling in which batches, each with a specific task, are sampled based on their proportion in

the entire training dataset. Moreover, with multi-GPU training, this approach becomes a heterogeneous scheme because each GPU simultaneously processes batches with different tasks, and the corresponding losses are averaged into one. In this work, we employ a multi-GPU homogeneous training procedure.

4.2 AlignScore

The main part of this work focuses on the holistic metric AlignScore, developed by Yuheng Zha et al. and introduced in their paper [Zha et al., 2023]. The AlignScore model is built on top of a unified alignment function, incorporating the training of a unified alignment function and align score function, combining the unified alignment function with a new context-claim splitting and aggregation strategy.

They propose unifying natural language understanding (NLU) tasks, such as natural language inference (NLI), fact verification (FV), paraphrase, summarization, semantic textual similarity (STS), question answering (QA), and information retrieval (IR) into three types of classification tasks, including 3-way, 2-way classification, and regression tasks. Used datasets are described in the Dataset chapter 5. Consequently, the AlignScore model is derived from a single architecture, pre-trained transformer-based RoBERTa model checkpoint [Liu et al., 2019], with three linear layers (heads) on top, each tackling the corresponding task. The following figure 4.1 shows how the **alignment function** of the AlignScore metric works on the unified NLU tasks.

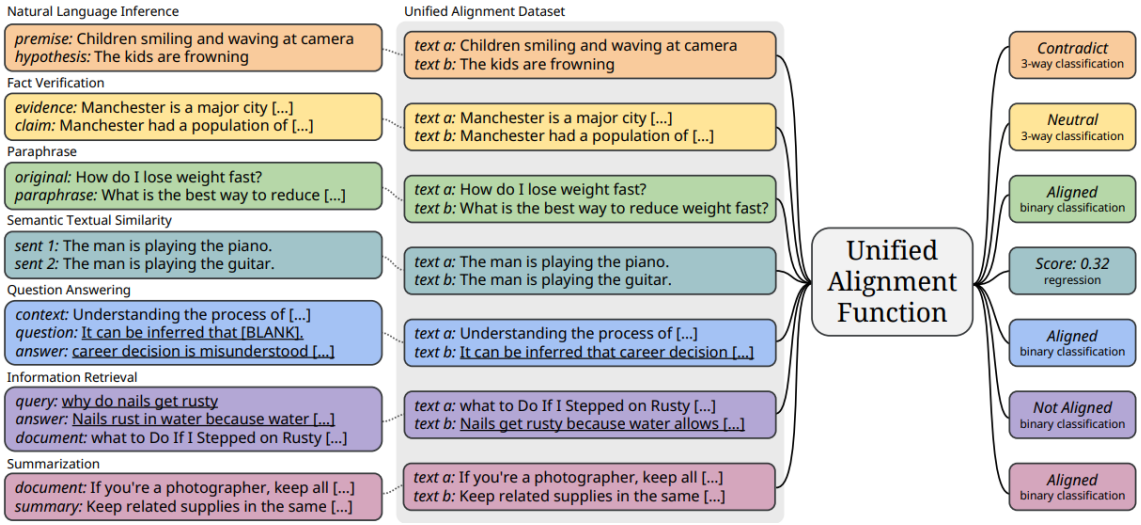

Figure 4.1: The figure shows the unification of individual NLU tasks and the possible outputs of the AlignScore model. Underlined text represents a modification made to the original dataset to form text pairs for the alignment dataset. Source [Zha et al., 2023].

Their presented unified training integrates weighted cross entropy loss leveraging individual loss of each linear layer, according to the formula:

$$L = \lambda_3 L_{3-way} + \lambda_2 L_{2-way} + \lambda_1 L_{reg} \tag{4.1}$$

where $\lambda_3 = \lambda_2 = \lambda_1 = 1$ regarding the paper. This representation produces output for each task simultaneously, in contrast to our multi-task architecture, which directly utilizes

a specific linear layer (head) for the given task. We accomplish the same setting of the weighted cross entropy loss when we utilize the multi-GPU training with a homogeneous batch, which averages the losses from task-specific batches.

Furthermore, their model, after the training, employs the **AlignScore function** in evaluation, which surpasses common drawbacks negatively affecting the output of the model. It splits a context into chunks so as not to exceed the input length of the model, which solves the problem of truncation, where the input is shortened to the possible input size of the utilized model (typically 512 tokens). Then, the function splits a claim into sentences independently of each other, which captures the problem of longer-span dependencies. Each sentence from the claim is evaluated against each context chunk using a modified alignment function, which is displayed in the following figure 4.2. Subsequently, it computes mean from maximum values over chunks per claim to derive a factual consistency score, as follows:

$$AlignScore(context, claim) = mean_j max_i A(o_i, l_j) \tag{4.2}$$

where A represents the alignment function with a modification that outputs a probability of the Aligned class given the 3-way linear head, o_i stands for context chunks, each contains roughly 350 tokens, and l_j refers to sentences from the claim. Hence, with the AlignScore function, the model predicts only the consistency score, resulting in a binary classification model in the evaluation.

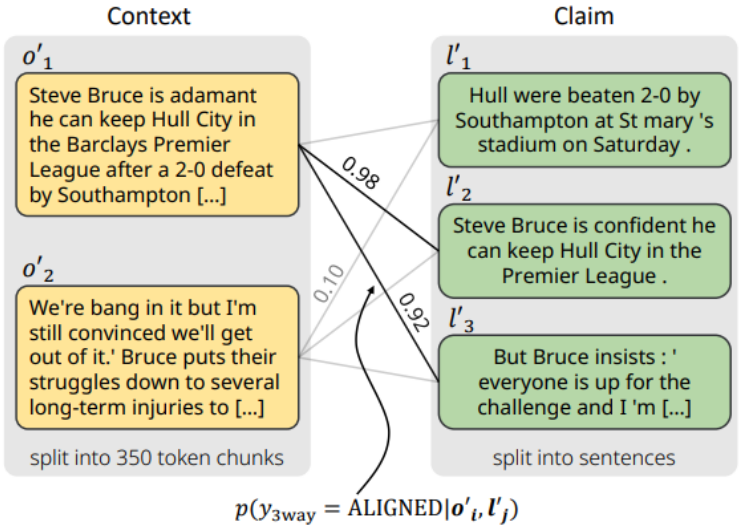

Figure 4.2: The figure demonstrates the computation of AlignScore function over context chunks o_i and claim sentences l_i. Source [Zha et al., 2023].

4.3 Baselines

In the following paragraphs, we explore alternative state-of-the-art baseline metrics designed to assess generation tasks and factual consistency. We use these metrics as baseline measures for evaluating Benchmarks and comparing them with our trained model. However, most of these evaluation metrics can work only on English data, making them unsuitable for application in Czech. Nevertheless, we discuss a few baseline metrics capable of processing Czech, outlined below. Additionally, English metrics are employed for reasonable comparison with our model's performance on English datasets.

Similarity Matching (SM) metrics involve comparing two objects to determine their similarity based on certain characteristics. This process typically calculates a similarity score or distance metric. For this type, we incorporate ROUGE[1] [Lin, 2004] and BLEU[2] [Papineni et al., 2002] metrics which both particularly calculate overlapping n-grams or fragments of tokens from two texts. For BLEU, we report 1-gram, and for ROUGE F-score, 1-gram. Next, we utilize multilingual BERTScore[3] [Zhang et al., 2020b] measuring token similarity score based on built contextual embeddings. In essence, all these SM metrics compute a single score ranging from 0 to 1 and, according to their settings, can be deployed in English and Czech.

Regression (Reg) metrics are trained models to predict a single score between 0 and 1. For this type, we utilize the BLEURT[4] model [Sellam et al., 2020], especially the BLEURT-20 a multilingual version including Czech and English, which employs a novel pre-training scheme using a large number of synthetic examples and correlates well with human judgments. In the Czech evaluation, we also report Memes-CS[5] [Šimon Zvára, 2022], XLM-RoBERTa-SQuAD2 fine-tunned on SQuADv2 [Rajpurkar et al., 2018a] and subsequently on CsFEVER [Ullrich et al., 2023], as well as on algorithmically augmented Czech summarization corpus for the regression task.

NLI metrics encompass a wide range of approaches aimed at addressing individual issues, mostly designed as multi-class classifiers. For English, we introduce the DAE model [Goyal and Durrett, 2020], which decomposes text at the level of dependency arcs. Additionally, we include SummaC-ZS [Laban et al., 2021], which performs zero-shot aggregation by combining sentence-level scores using mathematical operators. For Czech, we incorporate XLM-RoBERTa-SQuAD2[6] (XMLR-SQ2) [Ullrich et al., 2023], which addresses 3-way classification and was fine-tuned on SQuADv2, CTKFacts-NLI, CsFEVER, and CsFEVER-NLI datasets. Lastly, we include the CsFEVER-NLI[7] metric [Drchal et al., 2023], trained specifically on the CsFEVER-NLI dataset.

QA based metrics propose to combine entailment and factual consistency by adapting question generation and question answering models in training. To cover this type, we include QAFactEval [Fabbri et al., 2021b] measuring finer-grained answer overlap between a source and summary, fine-tuned on the SummaC validation set. Furthermore, for Czech, we add the QACG-sum[8] [Drchal et al., 2023] model, trained on multilingual corpora developed by the question answering for Claim Generation method using Wikipedia data.

Miscellaneous (MISC) metrics, as described by [Zha et al., 2023], employ diverse techniques in their evaluation approach. Among these, we incorporate UniEval [Zhong et al., 2022], a multi-dimensional metric that evaluates generated text from various aspects. Additionally, we include BARTScore [Yuan et al., 2021], which utilizes an encoder-decoder model to assess generated text from different perspectives. Lastly, we discuss the FactCC

[1] https://pypi.org/project/rouge-score/
[2] https://www.nltk.org/_modules/nltk/translate/bleu_score.html
[3] https://huggingface.co/spaces/evaluate-metric/bertscore
[4] https://github.com/google-research/bleurt
[5] https://huggingface.co/SimonZvara/Memes-CS_1.0/tree/main
[6] https://huggingface.co/ctu-aic/xlm-roberta-large-squad2-csfever_nli
[7] https://huggingface.co/ctu-aic/xlm-roberta-large-nli-csfever
[8] https://huggingface.co/datasets/ctu-aic/qacg-sum

metric [Kryscinski et al., 2019], trained on rule-based transformed data, focusing on factual consistency and other tasks. Unfortunately, MISC metrics only assess English texts; thus, we report them on English datasets.

For metrics where we do not specify the data source, we rely on the one mentioned by [Zha et al., 2023], and their scores are reported based on the research conducted in their paper. In comparison, for all metrics designed to handle multi-class classification, just like our 3-way classification architecture, we need to adjust their output values to yield a single score. Instead of outputting the most probable class, we modify them to output a probability of the consistent class, aligning with our model's approach, which returns a single score.

4.4 Metrics

In comparing the predicted consistency scores for documents of benchmarks by our model with those from other baseline metrics listed above, we use the AUC-ROC score and F1-macro score measures, which are briefly explained in the following paragraphs.

AUC-ROC score stands for Receiver Operating Characteristic Area Under the Curve and indicates how effectively the classification model distinguishes between positive and negative classes, ranging from 0 to 1, particularly for binary classification problems. In other words, the AUC-ROC score is a single number summarizing the model's performance across various classification thresholds by measuring the area under the ROC curve. Higher values indicate better performance, so a perfect model would score 1, whereas a random model would reach a score of 0.5. To derive the ROC curve plotting a True Positive rate (TPR) against a False Positive rate (FPR) at various classification thresholds, we compute these TPR and FDR from a confusion matrix as follows:

$$TPR = \frac{TP}{TP + FN} \; and \; FPR = \frac{FP}{FP + TN}$$

Where FP, TP, FN, and TN stand for False positive, True positive, False negative, and True negative, respectively. These values are derived exactly from the confusion matrix given an established decision threshold and scores computed by the model. The decision threshold ranges from 0.5 to 1, where the scores are converted to labels as follows: Scores above this threshold are classified as consistent; otherwise, they are inconsistent. To compute the AUC-ROC score, we measure the area under the ROC curve, commonly using the trapezoidal rule, which approximates the area by dividing it. We utilized the implemented version of AUC-ROC score from sklearn library[9].

F1-macro score is a commonly used metric for evaluating the classification model's performance. The suffix macro just stands for an indication of the averaging mechanism applied to F1 scores. F1 score ranges from 0 to 1 and is computed as a harmonic mean of precision and recall

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where precision is calculated as $Precision = \frac{TP}{TP+FP}$ and Recall we get just as TPR from the paragraph before, TP and FP represent true positive and false positive values derived

[9] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

from a confusion matrix. In our case, we do not utilize any threshold because we employ a 3-way classification head that directly predicts the class. However, in scenarios where a threshold is applicable, we normally use a single threshold equal to 0.5. Moving on to multi-class classification problems, where we encounter three classes, we compute F1 scores for each class individually. To obtain an overall assessment, we employ the macro average, which calculates the average of these F1 scores across all classes. We used a version of the F1-macro score implemented in the sklearn library[10].

4.5 Machine Translation

Machine translation (MT) represents a substantial field within natural language processing, incorporating advanced techniques such as multilingual pretraining of transformer architecture models [Liu et al., 2020, Fan et al., 2020, Kutylowski, 2017, Communication et al., 2023] on extensive document corpora. With a focus on three translation models, SeamlessM4T [Communication et al., 2023], M2M100 [Fan et al., 2020], and MBart-50 [Liu et al., 2020], my colleague conducted preliminary experiments that revealed the best performance for the SeamlessM4T model. SeamlessM4T produces the most fluent and natural texts for translating from English to the Czech language, which we require to translate English datasets used in AlignScore training. In contrast to opponents, the performance of SeamlessM4T is probably improved by its multi-modal setting, which encapsulates the training of a model for diverse domains. As in the paper of SeamlessM4T [Communication et al., 2023], they trained the model for text and speech translations simultaneously.

[10] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html`

Chapter 5
Datasets

In this chapter, we discuss the task-specific datasets utilized in training our AlignScoreCS model for multi-tasking. Following this, we explore a translation process to extend these datasets into the Czech language. Additionally, we delve into the unification process of datasets and their data augmentation for creating a robust training framework. After that, we present our unified AlignScoreCS training dataset.

Furthermore, we outline the benchmarks for evaluating our model's performance across various tasks. Alongside discussing the evaluation datasets we used, we also share a different method for translating them into the Czech language, in contrast to the previous translation procedure.

Both translation processes were undertaken collaboratively by myself and my colleague, who is also working on the AlignScoreCS project. Nevertheless, our collaborative work on building datasets concludes in this chapter since my colleague opted for a different training approach, employing heterogeneous batches rather than homogeneous ones and adopting a different multi-task architecture, following directly the methods described in the paper [Zha et al., 2023]. In contrast, my approach is studied and evaluated in the upcoming chapter 6.

For further insights, we provide descriptions of individual datasets in the Appendix B.1.

5.1 Task-Specific Datasets

According to the AlignScore paper [Zha et al., 2023], we utilized the proposed task-specific datasets, which significantly impact the training procedure of the model. Consequently, these datasets are employed during the training and validation phases, which assess the model's performance in the training process. The datasets used in the testing phase are detailed in the section Benchmarks 5.2.

For the training of the AlignScoreCS model, we prepared datasets related to various tasks, encompassing the Natural Language Inference NLI task, including SNLI [Bowman et al., 2015], MultiNLI [Williams et al., 2017], Adversarial NLI [Nie et al., 2019b], and DocNLI [Yin et al., 2021]. Additionally, we incorporated the Fact Verification task, covering NLI-style FEVER [Nie et al., 2019a] and Vitamin C [Schuster et al., 2021]. Furthermore, we added datasets QQP [Csernai,] and PAWS [Zhang et al., 2019], tackling the paraphrase task. I and my colleague, we both agreed on excluding a large 8M-documents WikiText-103 [Merity et al., 2016] dataset for the paraphrase task because of its size and,

regarding the experiments held by the authors [Zha et al., 2023], it did not have a significant impact on the model's performance. For the Czech PAWS dataset, we used the labeled and unlabeled PAWS [Zhang et al., 2019] datasets. Subsequently, we loaded the SICK [Marelli et al., 2014] and STS [Cer et al., 2017] datasets, representing the regression task (STS). For Czech STS, my colleague included the Free N1 STS [Sido et al., 2021] dataset. The question-answering task was covered by the SQuAD v2 [Rajpurkar et al., 2018a] and RACE [Lai et al., 2017] datasets. Lastly, we integrated Information Retrieval and Summarization tasks with MS MARCO [Nguyen et al., 2016] and WikiHow [Koupaee and Wang, 2018] datasets, respectively. For downloading these datasets, my colleague executed a Python script provided in the AlignScore paper resources [Zha et al., 2023] and received all these English datasets. Table 5.1 presents an overview of the task-specific datasets utilized in this study for training the AlignScoreCS model. The datasets are organized based on their respective NLP tasks, along with indications of which training tasks they are applied to within our multi-task model (an appropriate multi-task architecture head). Furthermore, each English dataset in the table row is paired with its Czech dataset that underwent translation through the mechanism outlined in the section 5.1.1. Additionally, the table provides statistics on the average word counts for contexts and claims in English and Czech languages, denoted as *cs* and *en* columns for each dataset. If the value in a particular column is missing, it signifies that the dataset was not used in that language.

NLP task	Dataset	Training Task	Context		Claim		Count	
			cs	en	cs	en	cs	en
NLI	SNLI	3-way	10	13	6	7	500k	550k
	MultiNLI	3-way	16	20	8	10	393k	393k
	Adversarial NLI	3-way	48	54	9	10	163k	163k
	DocNLI	2-way	97	285	14	43	200k	942k
Fact verification	NLI-style FEVER	3-way	48	50	7	8	208k	208k
	Vitamin C	3-way	23	25	10	11	371k	371k
Paraphrase	QQP	2-way	9	11	10	11	162k	364k
	PAWS	2-way	-	18	-	18	-	707k
	PAWS labeled	2-way	18	-	18	-	49k	-
	PAWS unlabeled	2-way	18	-	18	-	487k	-
STS	SICK	reg	-	10	-	10	-	4k
	STS Benchmark	reg	-	10	-	10	-	6k
	Free-N1	reg	18	-	18	-	20k	-
QA	SQuAD v2	2-way	105	119	10	11	130k	130k
	RACE	2-way	266	273	10	114	200k	351k
Information Retrieval	MS MARCO	2-way	49	56	5	15	200k	5M
Summarization	WikiHow	2-way	434	508	46	46	157k	157k

Table 5.1: Overview of Czech and English Datasets used in AlingScoreCS training. Table describes Natural Language Processing (NLP) tasks, associated datasets, training tasks used in multi-task training, and average word counts of context and claim in Czech (cs) and English (en) languages. 3-way stands for classification task into 3 classes, 2-way for binary classification, and reg for regression task. NLI: Natural Language Inference, STS: Semantic Textual Similarity, QA: Question Answering.

5.1.1 Translation

We implemented a script for dataset translation from arbitrary to arbitrary language using the SeamlessM4T model for automatic machine translation. Consequently, we translated all task-specific datasets from English to Czech language. We investigated the translated documents and discovered that our generated sequences are missing a few sentences close to the end of an article. As a result, the SeamlessM4T model can translate texts with around only 250 tokens in input length. Hence, we developed a segment translation procedure with batching, which divides the text into approximately 250 tokens concerning the last sentence length and accumulates these segments into a batch of an established size to accelerate the machine translation. However, my colleague used that script but modified it to translate the text over individual sentences and translated almost all the English task-specific datasets into Czech. Although this method operates within limited perspectives, as it tends to disrupt the semantic connections within longer texts, potentially damaging translations, mainly due to the diverse gendered words in Czech compared to English, it remains the singular suitable approach for applying the SeamlessM4T model to longer text translations.

To examine the potential harm to Czech data in more detail, we searched for texts necessary to segment due to their length. Occurrences of such texts were frequent for datasets from the summarization task, QA task, and DocNLI dataset, mainly due to its longer context size. Otherwise, it was observed only in rare cases, making approximately 2-3% of the entire size of each dataset. Further examined texts implied that determining nouns, possessive words, and gender-specific words are often repeated in separated segments, resulting in improved translation while keeping the knowledge among segments.

Thanks to these findings and the evaluation function with the chunking mechanism described in the section AlignSore 4.2, we decided not to execute subsequent modifications to the segment-translated texts in spite of possible solutions to overcome this limitation, such as different machine translation models preserving longer text information or morphology methods for reparation forms of word.

5.1.2 Unification & Preprocessing

We unified task-specific datasets, which involved adapting individual datasets to address our alignment problem related to 3-way, 2-way classification, and regression problems to facilitate data handling and model training. In the following paragraphs, we outline our unification of the training task-specific datasets (from NLI, Paraphrase, Fact verification, QA, Summarization, and Information retrieval domains). In addition, we also demonstrate our preprocessing steps applied to these task-specific datasets 5.1. These steps involve data augmentation, introducing artificial noise, and modification of datasets, as suggested in the paper [Zha et al., 2023].

Most datasets covered in **NLI**, **Paraphrase**, **Information retrieval**, and **Fact verification** domains conform to these problem types, allowing for simple mapping of labels. For binary classification, consistent labels were mapped to "Aligned," inconsistent to "Contradict," while for 3-way classification, we also mapped the third class to "Neutral." After that, we augmented datasets (from NLI, Paraphrase, and Fact Verification) using the following procedure. Each context-claim pair with a probability of 95% is kept unchanged, but 5% is modified to self-alignment. It means that we replace the claim of the original

context-claim pair with its context and assign its label to the ”Aligned” class to ensure consistency across the same sentences.

For datasets encompassed in **STS** domain, their scores (labels) were normalized to fall within the interval of 0 to 1 to fit the regression problem requirementsblem.

The **Question answering** domain and its datasets underwent the following modifications. Initially, we utilized a portion of QA datasets that were transformed into a suitable format for binary classification by authors of [Zha et al., 2023]. They converted context-question-answer triplets into context-claim pairs using a transformer generation model trained by [Demszky et al., 2018]. The model converts question-answer pairs into declarative sentences, which can be easily used as a claim in the binary classification problem. The ”Aligned” class is formed by samples created from ground truth answers, whereas the ”Contradict” class includes samples whose answers were generated from unanswerable questions. QA datasets created in this way, we then translated as described in the section 5.1.1. Some documents retained the original context-question-answer format, including wrongly generated answers to incorporate the ”Contradict” class. We then transformed these triplets into context-claim pairs as follows: with a probability of 75%, the claim concatenates the question and its corresponding answer. In the remaining 25%, the claim is solely an answer. Within this 25%, there is a 50% chance that the new context is formed from the original context with the question appended. Otherwise, the new context is the question with the original context appended. The idea behind this perspective rests in creating samples exactly from the QA domain to enhance the model's ability to evaluate QA samples directly without converting question-answer pairs into declarative sentences.

Moreover, in the **Summarization** domain, we adjusted the context-claim documents as follows: 50% of the documents retained their original context-claim pairs and were labeled as ”Aligned,” while the remaining 50% were marked as ”Contradict.” Those summarization samples containing predefined conflicting summaries are used; otherwise, a randomly selected summary from the corpus is employed.

Other task-specific datasets retain unchanged structure, and no modifications are made. All documents are tokenized with the corresponding XLM-Roberta tokenizer from [Conneau et al., 2019]. The truncation method for these training datasets was established at true with a maximal input length of 512 tokens and a padding strategy set to maximum input length to accelerate batch training.

5.1.3 Training dataset

The task-specific datasets are divided into their respective classification tasks after applying the unification and augmentation procedures described in 5.1.2. This segmentation was implemented to tailor the data to the distinct classification objectives of each task. Notably, for the training phase, we restricted the size of each dataset to at most 500,000 samples, with an additional 10,000 samples reserved for validation purposes. In cases where a dataset comprised fewer than 200,000 samples, a 5% portion of the training data was allocated for validation to maintain a consistent validation size ratio. According to the classification tasks, Czech and English task-specific datasets are concatenated into a single one, creating a multilingual multi-task dataset. The dataset is summarized in the table 5.2 for clarity.

AlignScoreCS - Training Dataset			
Training Task		Train split	Validation split
3-way	cs	1,556,726	38,920
	en	1,580,536	39,515
2-way	cs	1,510,065	18,756
	en	2,502,171	31.278
reg	cs	19,048	953
	en	9,702	486
Total size		7,178,249	129,908

Table 5.2: AlignScoreCS training dataset sizes summarized per training task, where 3-way, 2-way, and reg are ternary classification, binary classification, and regression problem.

Regarding the table 5.2, the Czech concatenated dataset of training split contains 1.56M documents, 1.58M documents, and 20k documents for 3-way classification task, 2-way classification task, and regression task, respectively. The English concatenated training split dataset comprises 1.51M, 2,5M, and 10k documents for a 3-way classification task, 2-way classification task, and regression task, respectively. In general, the training data reaches the size of 7.18M documents, with a validation volume of around 130k samples. Due to the validation phase lasting 1 hour during training, the validation data size is reduced proportionally to the classification tasks. We keep all regression validation data, 25% of 2-way classification data, and 50% of 3-way classification data, resulting in 80k samples in the validation set. The resulting dataset is summarized in the table.

5.2 Benchmarks

Two benchmarks have been utilized in the testing phase of our AlignScoreCS model: **SummaC** [Laban et al., 2021] and **TRUE** [Honovich et al., 2022]. We mainly followed the paper [Zha et al., 2023] to have baseline model results and compare the performance of our AlignScoreCS model. The following table 5.3 presents statistics for individual datasets from both benchmarks. Each benchmark row includes values for both the original English dataset and the Czech dataset, which was translated using the method described in the translation section 5.2.1.

SummaC benchmark, [Laban et al., 2021] introduced the SUMMAC Benchmark, a new large dataset for assessing consistency in summaries. This benchmark combines six major datasets for summary inconsistency detection, covering CogenSumm, SummEval, FRANK, FactCC, Polytope, and XsumFaith. All datasets are standardized for a binary classification task. Each dataset within the benchmark consists of samples of context, claim, and label, where the label indicates consistency or inconsistency.

TRUE benchmark, [Honovich et al., 2022] unified existing datasets annotated for factual consistency to provide a standardized format of context and claim texts with binary annotations. TRUE facilitates the comparison of consistency evaluation methods across various tasks and domains like summarization, dialogue, paraphrasing, and fact verification. The entire benchmark covers 11 datasets: summarization task (FRANK, SummEval, MNBM, QAGS-XSUM, QAGS-CNNDM), paraphrase task (PAWS), fact verification task (FEVER, VitaminC) and dialogue task (BEGIN, Q^2, DialFact).

TRUE							SummaC						
Dataset	Context		Claim		Count		Dataset	Context		Claim		Count	
	cs	en	cs	en	cs	en		cs	en	cs	en	cs	en
BEGIN	20	23	12	14	836	836	CogenSumm	699	657	58	31	1281	1281
Q^2	20	23	11	16	1088	1088	FactCC	645	547	14	15	931	931
FRANK	442	499	33	41	671	671	FRANK	466	499	41	41	671	671
SummEval	321	363	52	63	1600	1600	SummEval	363	363	63	62	850	850
QAGS	278	318	44	49	235	235	Polytope	572	575	65	64	634	634
PAWS	19	21	19	21	8000	8000	XSumFaith	384	383	19	19	1250	1250
FEVER	52	59	8	8	7954	18209							
MNBM	343	383	16	19	2500	2500							
DialFact	23	26	15	17	8689	8689							
VitaminC	-	28	-	13	-	63054							

Table 5.3: Overview of English and Czech Benchmarks used in evaluating AlingScoreCS. The table describes individual datasets of TRUE [Honovich et al., 2022] and SummaC [Laban et al., 2021] benchmarks and their average word counts of context and claim in Czech (cs) and English (en) languages. Each benchmark row contains values for English-origin and its Czech-translated dataset denoted in *cs* and *en* columns, respectively.

5.2.1 Translation

To make testing samples as diverse as possible and to pursue a noise-tolerant approach to testing, we 1) translate the benchmark datasets using a different machine translation model from the one translating task-specific datasets and 2) use two distinct approaches to the benchmark translation. Me and my colleague agreed on utilizing the DeepL (version: November in 2023) machine translation model [Kutylowski, 2017] and its cloud system for applying document translation. The company provides a one-month free trial DeepL Pro subscription plan for up to 20 files of 1MB each. We, thus, transform and combine each benchmark dataset and its claim context pairs into 20 text files of 1MB size for DeepL text file translation.

My colleague employed a repetitive string pattern detection mechanism to translate the SummaC benchmark, efficiently capturing and storing translations of recurring sentences. Due to the limited size, this approach enhances our capability to translate a larger volume of documents. Leveraging the common dataset's repetitive 4-item window context, he managed to successfully translate all benchmark datasets, resulting in the creation of a Czech SummaC benchmark.

Despite acquiring more translations using the repetitive mechanism, we have opted to translate each recurring context independently of others. This decision is aimed at preserving translation accuracy, as there are instances where the model predicts different outputs for the same input texts, leading to preventing potential translation noise. Applying this idea, we translated almost all datasets from the TRUE benchmark except a part of the NLI-FEVER dataset and the entire Vitamin C dataset, which has been extracted due to the size limitations of files.

Chapter 6
Experiments

This chapter outlines the experiments conducted in this part of the thesis. Firstly, we introduce the AlignScoreCS model, detailing its multi-task architecture design and the configurations and settings required for smooth operations. We then delve into the training procedure, discussing the hyper-parameter settings and methods employed in multi-task learning. Additionally, we provide a brief overview of the multi-task implementation background.

Moving on, we experiment with the training dataset, evaluating various types of AlignScoreCS models trained on subsets of the training dataset to observe distinct behaviors influenced by the training samples. From these experiments, we select the best-performing model for subsequent evaluations. Then, we assess the AlignScoreCS model on English benchmarks, namely SummaC and TRUE, and compare the results with other English evaluation metrics.

Furthermore, we explore zero-shot learning, studying the behavior of AlignScoreCS when evaluating datasets addressing facticity verification in different languages from training. Finally, we examine the performance on Czech-translated benchmarks and NLI datasets, evaluating other Czech metrics to compare results. This process sets the new highest baseline results on both Czech benchmarks.

6.1 AlignScoreCS

Our AlignScoreCS model is built on the XLM-RoBERTa large encoder architecture [Conneau et al., 2019], which has been pre-trained on a sizeable multilingual dataset, including Czech. Consequently, the model is suited for application in the Czech language. Furthermore, the multilingual setup of our base model enables evaluation in other languages, even those for which the model has not been specifically fine-tuned, which is suitable for zero-shot evaluation. We examine this behavior later.

Unlike [Zha et al., 2023], where the model is derived from a single encoder architecture with three separate linear layers on top, our AlignScoreCS model is constructed from three XLM-RoBERTa architectures sharing one encoder. Consequently, our model distributes one encoder among these three architectures, each equipped with its specific linear layer (head) addressing different classification tasks. One architecture handles a 3-way classification problem, another manages a 2-way classification problem, and the last addresses a regression task while all using the same encoder's weights. This setup facilitates the segregation of the model into task-specific units, allowing them to operate independently

in certain scenarios. The architecture is shown in the figure 6.1 for simplicity. The model is thus fed with documents denoted with a specific task. Regarding the task name, the model selects an architecture with the corresponding head for that exact task. Then, the head predicts output using the input as outputs from the shared encoder.

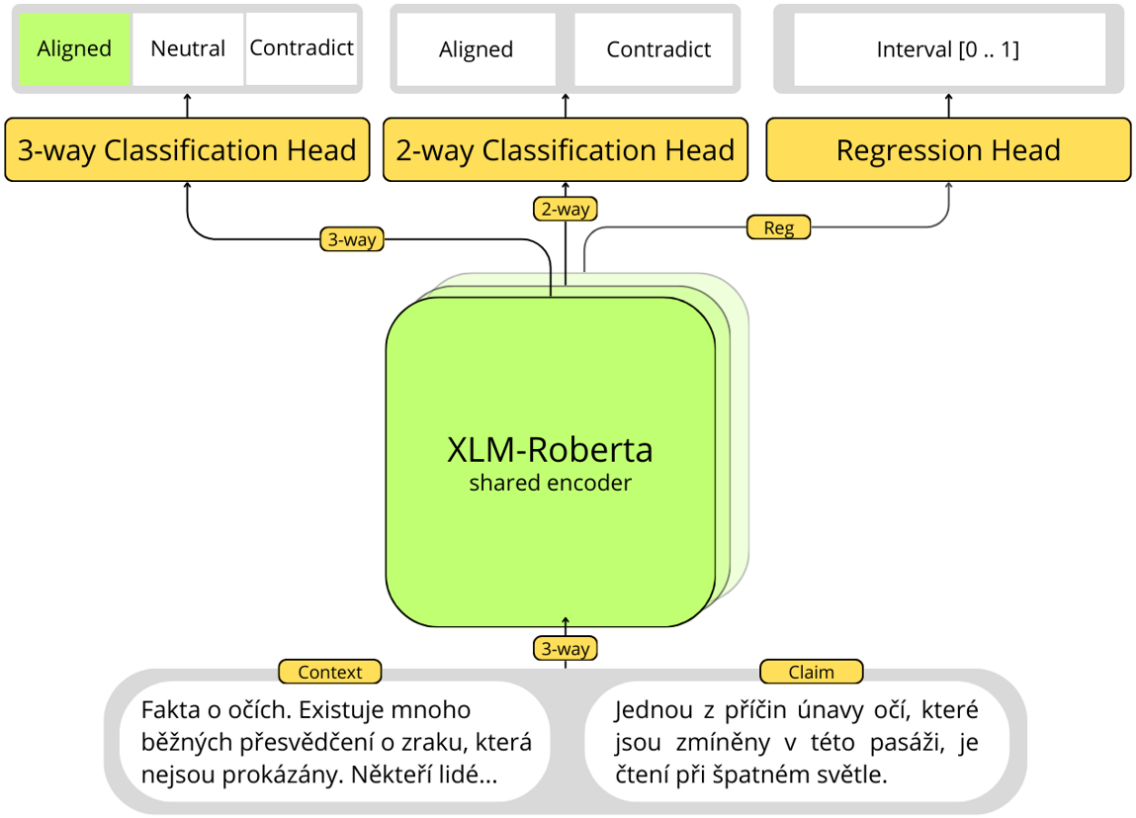

Figure 6.1: The figure shows the shared architecture of the AlignScoreCS model, which is derived from three XLM-RoBERTa architectures [Conneau et al., 2019] with different linear layers on top (Classification heads), each tackling the specific task (3-way, 2-way, regression) while sharing one encoder. Documents are fed to the model along with task identification, enabling the model to distinguish which of the three architectures to utilize.

For evaluation and the deliberate utilization of our model, we employ the recommended **AlignScore function** with the chunking mechanism described in Equation 4.2, which calculates a consistency score based on a given context and claim. Moreover, to evaluate the QA task more precisely, we modified the function to handle context-question-answer triplets. During evaluation, the procedure chunks the context as usual. However, when splitting the answer (claim) into sentences, it concatenates the question with each sentence of the claim to preserve the question information.

When assessing a **classification problem**, which is not solved by [Zha et al., 2023], we propose a novel approach to integrate the chunking mechanism used in the AlignScore function for classification of a context and claim pair. This is achieved through the following formula, where we calculate the most probable class from two times applied mean,

one over context chunks for each claim sentence and one from the results, as follows:

$$\text{Classify}(context, claim) = argmax(mean_i mean_j A(o_i, l_j)), \tag{6.1}$$

where A is the alignment function with a modification that returns probability distribution over classes from the 3-way or 2-way classification heads. The o_i represents a chunk from the context chunk set, and l_j is a sentence from the claim sentence set.

6.1.1 Implementation Details

We utilized the HuggingFace Transformer library [Wolf et al., 2020], which provides a wide range of predefined model classes of transformer architectures. We aimed at maximizing usability and accessibility within the research community; thus, we implemented a custom AlignScoreCS class inheriting from the XLMRobertaModel, which is available in the HFT library. To encapsulate three models with task-specific heads into a unified structure, we used a module dictionary that organizes the models based on keys directly corresponding to the names of classification problems. Subsequently, we adjusted the $.forward$ function to accept a task name parameter, enabling the selection of the appropriate model from the module dictionary during execution and training. Last but not least, we modified the $.from\_pretrained$ method to enable the loading of the three XLM-RoBERTa models equipped with task-specific classification heads and to ensure the encoder part is shared among these models by passing the pointers.

Furthermore, we experiment with the chunking mechanism and employ an approach to chunking the context-claim pair during evaluation. Drawing inspiration from [Zha et al., 2023], we crafted a chunking mechanism for the context, limiting it to approximately 350 tokens per chunk but incorporating an overflowing window. This window repeats a specified number of tokens from the previous chunk in the subsequent one when the tokens in a sentence exceed the chunk size, preserving crucial information. Nonetheless, after experimenting with the data, we observed varying results. While some cases showed improvement, others exhibited a decline in performance. Therefore, we decided to employ the mentioned alignscore function proposed in [Zha et al., 2023] for chunking the text.

6.2 Training

In the training phase, we trained the AlignScoreCS model for 3 epochs on 4x NVIDIA A100-SXM4-40GB for 3 and half days. In hyper-parameters settings, we mainly follow the papers [Zha et al., 2023, Conneau et al., 2019]. Initial warm-up training was configured for 13,500 steps, which corresponds to around 0.06% of the total training dataset size in steps. The evaluation phase was conducted every 8,000 steps using the best model loss storing method. We set a learning rate to 1e-5 with a linear learning scheduler and AdamW optimizer. We established batch size at 8 with accumulation gradients of size 4, resulting in a total batch size of 32 samples. We trained an adjusted version of the model for floating-point of 16-bit precision to accelerate calculations.

For multi-task training, we employed a scenario involving homogeneous batches, where each batch exclusively contained samples from a single task (3-way, 2-way, or regression). Utilizing this approach on a single GPU would be inefficient, as the model would continuously receive samples from only one task, affecting the loss calculation solely based on

that task. However, operating in a multi-GPU setting allowed each GPU to simultaneously process batches from different tasks. By aggregating losses from individual GPUs, the combined loss from Equation 4.1 was computed by simply adding them together, resulting in behavior similar to the specified equation, given λ_i set to 1. The behavior slightly diverges due to an unbalanced distribution of batches across the GPUs. With four batches (one per GPU), two GPUs may be simultaneously processing batches from one task, resulting in a higher weight in the respective task. However, we opted not to address this issue, as the training tasks exhibit imbalances. Moreover, in a heterogeneous setting, it is also possible for samples from the regression task to be missing.

Framework Weight and Biases[1] has been integrated where the whole training progress is reported to facilitate the conveyed information of training. During training, we also reported custom metrics (AUC-ROC, accuracy, mean-squared error, and mean absolute error), which were measured during each validation phase. We saved the best AUC-ROC performing model on a 3-way validation set at the end of the last third epoch.

6.2.1 Implementation Details

We again leveraged the HuggingFace Transformer (HFT) framework [Wolf et al., 2020] to facilitate our implementation of the training process, benefiting from its extensive library of pre-programmed classes for optimized training loops, evaluation strategies, and user-friendly interface. However, we encountered a limitation: the framework did not offer built support for adapting multi-task learning. To address this limitation, we implemented custom inherited classes, encapsulating the integration of multi-task learning into our training as follows:

1. **Task name** - As each sample carries information regarding the training task it is applied to in order to guide the model which classification head to use, we encapsulated the task identifier within a class. This class was designed to include a *.to* method, as required by the HFT Trainer class, transferring sample features onto a device (GPU) during training.

2. **Task data loader** - We implemented a custom task data loader that encapsulates a standard data loader of a dataset. This Task data loader is used for each classification task since it adds an identification of the specific training task to the sampled batch of data.

3. **Multi-Task data loader** - To incorporate all classification tasks, we developed a custom Multi-Task data loader designed to manage Task data loaders of individual tasks. This implemented Multi-Task data loader samples a specific task and subsequent batch from the corresponding Task data loader. Sampling ensures a balanced representation of tasks during training.

4. **Multi-Task Trainer** - We implemented a custom Multi-Task trainer inheriting from the HFT Trainer class. Our trainer is adjusted to handle the Multi-Task data loader when creating its train data loader. And, especially the evaluation data loader, since we modified inner trainer methods to facilitate the evaluation phase to process task-specific validation sets individually and log the computed metrics.

As a result, we could employ the HFT framework to train our multi-task model.

[1] `https://wandb.ai/site`

6.3 Ablation Study

To examine the importance of our training dataset and how the included data affects the overall performance of our trained model, we compare 3 models trained on the training dataset with different settings. AlignScore-restricted-cs (AS-restricted-cs) has been trained only on Czech data from our training dataset with a restriction that each task-specific dataset has been limited to 200,000 training samples in contrast to 500,000. Next, AlignScore-restricted (AS-restricted) has seen both Czech and English training datasets again, with a limitation of 200,000 training samples per task-specific dataset. Lastly, we evaluate our main AlignScoreCS model, which had access to the whole training dataset without further restriction (500,000 samples per task-specific dataset), and its training is detailed in the section 6.2. The following table 6.1 summarizes average models' AUC-ROC scores on each benchmark we listed in the previous chapter.

	TRUE-CS	TRUE	SummaC-CS	SummaC
AS-restricted-cs	77	77	74	75
AS-restricted-csen	76	80	74	80
AlignScoreCS	**80**	**84**	**82**	**86**

Table 6.1: AUC-ROC scores for comparison of AlignScoreCS models with different training sets on Benchmarks discussed in 5.2. The table shows average AUC-ROC scores per benchmark, with the highest highlighted values. AS-restricted-cs is trained only on the Czech-restricted training dataset, AS-restricted-csen is trained on both English and Czech-restricted training datasets, and AlignScoreCS is trained on the entire training dataset. The suffix *-CS* denotes a translated benchmark regarding the technique from 5.2.1.

The values from the table 6.1 indicate that both AS-restricted-cs and AS-restricted-csen show similar performance on translated Czech benchmarks, resulting in no decline in performance when English data is incorporated. Additionally, the AS-restricted-cs outputs comparable scores for English benchmarks, suggesting its potential for cross-lingual zero-shot learning, a concept of evaluating models on diverse samples distinct from those used in training. We examine this zero-shot learning in various languages later. Finally, the AlignScoreCS model consistently outperforms other models, inducing the significance of dataset size. Moving forward, we exclusively utilize the **AlignScoreCS** model for further evaluation.

6.4 Results on English Benchmarks

Experiments held in the subsequent sections assess the performance of our AlignScoreCS model on both English benchmarks to obtain results that we can compare against other English-centric metrics described in the section 4.3, especially those reported and evaluated by [Zha et al., 2023]. These metrics were selected based on their high performance; others reported in the paper we do not include. The initial section presents the results obtained from the SummaC Benchmark, followed by an evaluation of the TRUE Benchmark. Both tables presenting the results, SummaC scores (6.2) and TRUE scores (6.3), illustrate the AUC-ROC scores produced by baseline metrics, AlignScore models, and our AlignScoreCS model for datasets covered in the respective benchmarks. Additionally, we include the average score for each metric on the benchmarks, confirming that the overall best-performing models are the AlignScore models, with average scores of 88.0 and 86.9 for AlignScore-large and AlignScore-base, respectively. Following closely, the second best

performing model is our AlignScoreCS, achieving an average score of 85.4, slightly lower than the AS models. Lastly, the third best performing model on both datasets is UniEval, yielding an average score of 82.1, followed by QAFactEval and SummaC-ZS with 82.0 and 80.6, respectively.

By stating this, it suggests that our model is greatly comparable to these models and, therefore, introduces a new path for high-quality multilingual metrics that demonstrate noteworthy results on English data but could also be applied to other languages. The slightly lower performance of our model on English datasets compared to AlignScore models can be explained by our distinct shared multi-task architecture and our approach to homogeneous batches instead of heterogeneous ones during training.

6.4.1 Results on SummaC

We evaluated our AlignScoreCS model on the SummaC Benchmark, and the table 6.2 below displays the AUC-ROC scores for each dataset within SummaC. Our AlignScoreCS (ASCS) model achieves highly comparable results but slightly below those of both AlignScore (AS) models for datasets FRANK and SummEval. In the case of PolyTope, ASCS yields outstanding results but falls behind UniEval and AS-large. Similarly, for XSF, we are surpassed by both AS and DAE. In the FactCC dataset, our performance follows both AS and UniEval, while in CGS, we fall behind both AS, QAFactEval, and UniEval. However, on average, our model achieves the second-highest score, trailing behind both AS models. This underscores the comparability of our model to English models, though with multilingual capabilities.

Type	Metric	CGS	XSF	PolyTope	FactCC	SummEval	FRANK	AVG
QA	QAFactEval	83.4	66.1	86.4	89.2	88.1	89.4	83.8
SM	BLEU	71.8	55.8	86.9	75.0	83.8	84.5	76.3
	BERTScore	63.1	49.0	85.3	70.9	79.6	84.9	72.1
Regression	BLEURT	60.8	64.7	76.7	59.7	71.1	82.5	69.2
NLI	DAE	52.4	76.7	72.8	54.2	66.1	78.9	66.8
	SummaC-ZS	73.6	58.0	87.5	83.7	85.8	85.3	79.0
MISC	UniEval	84.7	65.5	**93.4**	89.9	86.3	88.0	84.6
	BARTScore	74.3	62.6	91.7	82.3	85.9	88.5	80.9
	FactCC	64.9	55.1	78.5	72.7	71.8	69.8	68.8
Align	AlignScore-base	83.7	**79.4**	87.8	93.3	89.9	90.5	87.4
	AlignScore-large	**86.4**	75.8	92.4	**93.7**	**91.7**	**91.4**	**88.6**
Our	AlignScoreCS	82.9	74.8	92.3	89.5	88.8	90.0	86.4

Table 6.2: The table shows AUC-ROC scores on datasets included in the English SummaC benchmark. For comparison, there are reported scores of other metrics with their types described in 4.3. The highest scores per dataset are in bold. AVG denotes an average of scores per metric. CGS and XSF stand for CogenSumm and XSumFaith, respectively. The type SM indicates Similarity Matching. Scores reached by other metrics are copied from [Zha et al., 2023].

6.4.2 Results on TRUE

In our final evaluation of AlignScoreCS on English data, we present the AUC-ROC results for datasets covered by the TRUE benchmark in Table 6.3. AlignScoreCS (ASCS) achieves the second highest score after AlignScore-large (AS-large) for the PAWS dataset, with

AlignScore-base (AS-base) and SummaC-ZS following closely. For the VitaminC (Vitc) and Fever (FVR) datasets, our model is outperformed by both AS models but still produces remarkable results, while for DialFact (DF), Q2, and MNBM, the top scores are achieved by NLI metrics 2x SummaC-ZS and DAE, respectively, closely followed by the AS models and ASCS. In the case of the FRANK dataset (FRK), the best performing models are the AS models, followed by QAFactEval and UniEval, with ASCS achieving slightly lower but still comparable results. The QC and QX datasets from QAGS-CNNDM and QAGS-XSUM, respectively, show the highest scores for the AS models, while our model, evaluated on the concatenation of these datasets, resulting in one QAGS dataset, achieves the second highest average score after the AS models. For the SummEval (SE) dataset, our model yields significantly lower results compared to other models despite achieving high scores on other datasets, whereas AS-large achieves the highest scores, followed by UniEval and QAFactEval. Lastly, for the BEGIN dataset, the best-performing model is BARTScore, followed by BLUERT and the AS models. On average, our AlignScoreCS model surpasses all included metrics except for AlignScore models.

Type	Metric	SE	PAWS	Q2	Vitc	FVR	FRK	DF	MNBM	QC	QX	BEGIN	AVG
QA	QAFactEval	80.9	86.1	75.8	73.6	86.0	88.5	81.8	67.3	83.9	76.1	81.0	80.1
SM	BLEU	74.8	71.3	55.2	56.1	51.7	84.1	61.2	56.7	77.4	54.7	74.6	65.2
	BERTScore	72.3	78.6	70.2	58.2	54.2	84.0	68.6	52.5	70.6	44.3	86.4	67.2
Reg	BLEURT	68.0	68.4	72.9	61.8	59.5	81.6	73.0	65.5	71.2	56.2	86.6	69.5
NLI	DAE	60.3	55.8	57.7	60.2	77.8	77.9	54.7	**81.0**	56.9	67.5	69.4	65.4
	SummaC-ZS	77.6	89.0	**81.8**	97.2	92.8	86.9	**87.1**	58.0	76.0	75.3	83.2	82.2
MISC	UniEval	81.2	80.1	70.4	79.1	92.1	88.1	80.4	66.8	86.5	76.7	73.6	79.5
	BARTScore	78.9	77.1	65.1	64.2	66.1	87.8	60.8	63.5	83.9	60.2	**86.7**	72.2
	FactCC	68.6	53.4	59.3	54.7	58.7	70.7	55.0	56.1	70.1	64.4	57.6	60.8
Align	AlignScore-base	80.8	97.3	76.1	97.8	94.6	90.0	83.1	79.9	87.7	79.6	82.4	86.3
	AlignScore-large	**82.9**	**98.4**	78.6	**98.3**	**94.9**	**92.1**	85.1	76.1	**89.5**	**83.5**	82.7	**87.4**
Our	AlignScoreCS	65.9	98.1	77.0	97.7	94.4	85.3	83.0	78.4	84.8		79.0	84.4

Table 6.3: The table shows AUC-ROC scores on datasets included in the English TRUE benchmark. For comparison, there are reported scores of other metrics with their types described in 4.3. The highest scores per dataset are in bold. AVG denotes an average of scores per metric. SE, DF, QC, and QX datasets are SummEval, DialFact, QAGS-CNNDM, and QAGS-XSUM, respectively. The scores of AlignScoreCS for the QC and QX datasets are averaged since we downloaded the TRUE benchmark, which includes both QC and QX, as one QAGS dataset. The types SM and Reg indicate Similarity Matching and Regression. Scores reached by other metrics are copied from [Zha et al., 2023].

6.5 Understanding the Power of Multilinguality

The following experiments investigate the behavior of our multilingual model, AlignScoreCS, in zero-shot learning, where we evaluate the model on languages it has not encountered during fine-tuning. To conduct this assessment, we utilized the test set of the XNLI dataset [Conneau et al., 2018], which stands for Cross-lingual Natural Language Inference. This dataset is derived from MultiNLI [Williams et al., 2017], where the authors translated the data into 15 languages, namely including Arabic (ar), Bulgarian (bg), German (de), Greek (el), English (en), Spanish (es), French (fr), Hindi (hi), Russian (ru), Swahili (sw), Thai (th), Turkish (tr), Urdu (ur), Vietnamese (vi), and Chinese (zh). The authors used automated machine translation for the training set, while manual translation was employed for the validation and test sets. Consequently, each language has a test

size of 5010 samples. The dataset represents a ternary classification problem; thereby, to utilize our AlignScoreCS model, which employs the alignscore function to predict the probability of the "Aligned" class given the 3-way classification head, we filtered out samples labeled as "Neutral" and retained only samples with consistent (Aligned) and inconsistent (Contradict) labels. This resulted in a test size of 3340 context-claim pairs per language. We evaluate the context of each language against a claim of each language of the filtered test set and present the AUC-ROC scores in the following graph 6.2.

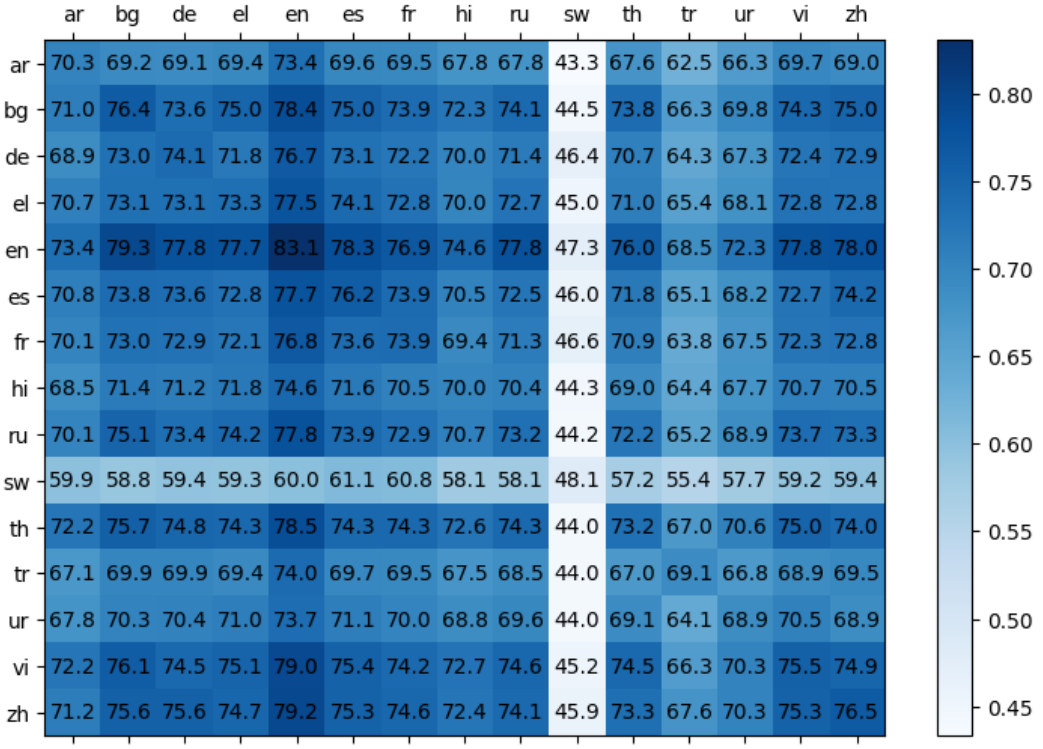

Figure 6.2: Zero-shot evaluation on the XNLI dataset. The graph displays the AUC-ROC scores of the AlignScoreCS model assessed on the XNLI dataset, which exclusively includes samples labeled as consistent and inconsistent; neutral labels were omitted. We evaluated every context of each language against every claim of each language. A row represents the language of the context, and a column stands for the language of the claim. Language set includes 'ar' - Arabic, 'bg' - Bulgarian, 'de' - German, 'el' - Greek, 'en' - English, 'es' - Spanish, 'fr' - French, 'hi' - Hindi, 'ru' - Russian, 'sw' - Swahili, 'th' - Thai, 'tr' - Turkish, 'ur' - Urdu, 'vi' - Vietnamese, and 'zh' - Chinese.

Although our utilized architecture of XLM-RoBERTa has seen 100 languages during pre-training, including these 15 languages covered in the XNLI dataset, the results for the Swahili language indicate the worst performance, either for a context or for a claim, inferring low representation of Swahili data in pre-training. Surprisingly, when the Swahili context is evaluated against claims in other languages, it achieves better scores, suggesting that our model focuses particularly more on claims than on contexts. Similarly, poor performance is observed in the Turkish and Urdu languages. On the opposite side, when English is incorporated either in context or claim, we can see significant improvements in scores in contrast to other languages. In the case of the English-English pair, we see the best performance, as expected. For most other languages, the scores are relatively similar but not particularly remarkable. Nevertheless, the Chinese also demonstrates quite good scores in this evaluation. This assessment raises an interesting question about whether

our model can effectively evaluate context in one language against a claim in another, i.e., English against Czech. While we leave it for future experiments, we provide some examples in the Appendix B.2.

Inferred from Figure 6.2, our AlignScoreCS architecture demonstrates the potential of multilingual pre-training. Despite being fine-tuned only on English and Czech NLU datasets, our model achieves gains in other languages on the XNLI dataset, which were not used during the fine-tuning phase. Consequently, our model could be potentially utilized to evaluate context-claim pairs in various languages but further fine-tuning may be required.

6.6 Results on Czech Data

The last experiments evaluate our AlignScoreCS model on Czech translated benchmarks, SummaC and True, regarding the translation mechanism described in Section 5.2.1. The initial section summarizes results on Czech SummaC, followed by a section describing performance on the Czech TRUE benchmark. Conducted tests also compare other evaluation metrics being able to incorporate Czech data, which are introduced in Section 4.3. The last section concludes with a performance comparison of our model and other Czech evaluation metrics on Czech NLI datasets.

Czech NLI metrics addressing multi-class classification were modified to output a single probability for the consistent class. In this configuration, the models finally output consistency scores for comparison in the binary classification task embedded in the benchmarks. Other metrics remained unchanged, as they already output a single score between 0 and 1.

6.6.1 Results on Czech SummaC

In terms of the Czech SummaC benchmark, we assess metrics capable of evaluating Czech data. Table 6.4 shows the AUC-ROC scores on each dataset. From the results, it is evident that our model performs exceptionally well, achieving the highest average score. Across datasets such as CogenSumm (CGS), PolyTope, FactCC, and SummEval, our model outperformed other evaluated metrics. Notably, BERTScore delivered strong results for the PolyTope dataset, while the simple BLEU metric exhibited reasonable performance on SummEval. However, for the XSumFaith (XSF) dataset, our model was surpassed by the Memes-CS metric, which produced a slightly higher score, followed closely by our model and XLMR-SQ2. In the case of the FRANK dataset, BERTScore became the top-performing model, closely followed by our AlignScoreCS. Memes-CS and Czech NLI models also demonstrated comparable scores. Remarkably, for ROUGE and BLEU metrics, the values for SummEval and FRANK were significantly high, indicating that these datasets lean towards extractive summarization, as these metrics primarily assess overlap between extractive fragments.

However, our AlignScoreCS establishes the new highest results on the Czech SummaC benchmark that could be considered as a baseline for future work. Given that the included datasets are tailored for assessing summarization veracity, our model emerges as the most suitable choice for evaluating factual accuracy in Czech summarization tasks.

Type	Metric	CGS	XSF	PolyTope	FactCC	SummEval	FRANK	AVG
SM	BERTScore	54.9	60.1	81.0	59.5	76.8	**86.0**	69.7
	BLEU	63.2	64.5	73.3	63.5	81.3	84.1	71.6
	ROUGE	56.9	42.4	72.8	56.6	64.6	80.3	62.2
Reg	BLEURT	61.5	64.2	76.1	60.2	65.4	71.5	66.5
	Memes-CS	62.5	**75.1**	72.0	56.7	69.8	81.3	69.6
NLI	CsFEVER-NLI	58.8	61.5	63.0	62.4	63.3	68.2	62.8
	QACG-sum	59.6	65.0	71.8	77.8	68.1	79.1	70.2
	XLMR-SQ2	65.6	70.0	63.1	73.2	60.6	79.3	68.6
Our	AlignScoreCS	**76.3**	73.9	**86.8**	**87.5**	**82.3**	85.4	**82.0**

Table 6.4: The table shows AUC-ROC scores on translated datasets included in the Czech SummaC benchmark. For comparison, there are reported scores of other Czech metrics with their types described in 4.3. The highest scores per dataset are in bold. AVG denotes an average of scores per metric. CGS and XSF stand for CogenSumm and XSumFaith, respectively. The type SM indicates Similarity Matching.

6.6.2 Results on Czech TRUE

We successfully evaluate AlignScoreCS and several other metrics on the Czech-translated TRUE benchmark. The results, detailed in Table 6.5, show the AUC-ROC scores across respective datasets. Notably, AlignScoreCS demonstrates its potential across datasets, namely QAGS, MNBM, FEVER (FVR), and PAWS, where it significantly outperforms all other metrics. Our model's near-perfect performance on the PAWS dataset is particularly impressive, achieving a score of 96.6. It is worth mentioning that Memes-CS also delivers competitive results, especially on the MNBM and FEVER datasets. Additionally, XLMR-SQ2 performs comparably on FEVER, benefiting from training on the Czech FEVER dataset. However, in the SummEval (SE) dataset, our model is surpassed by BLEU and BERTScore metrics. Similarly, for Q2 and DialFact (DF), Memes-CS emerges as the top performer, with our model occupying the fourth position in the Q2 dataset, following BLUERT and QACG-sum. In the case of the DF dataset, we are overtaken by XLMR-SQ2. BLEU achieves the best results for the FRANK dataset, which is closely followed by Memes-CS and AlignScoreCS. BERTScore also stands out in the BEGIN dataset, yielding the highest scores. Evidently, Memes-CS, XLMR-SQ2, and QACG-sum achieve considerable results in contrast to our model, which shows lower performance for the BEGIN dataset.

Type	Metric	SE	PAWS	Q2	FVR	FRK	DF	MNBM	QAGS	BEGIN	AVG
SM	BERTScore	71.4	67.9	67.0	54.4	82.4	65.8	60.7	70.5	**83.5**	69.3
	BLEU	**73.5**	61.2	63.0	56.6	**82.7**	64.9	67.0	70.9	82.4	69.1
	ROUGE	62.4	60.9	67.3	52.8	76.9	71.6	45.9	65.3	81.6	65.0
Reg	BLEURT	62.3	66.0	75.3	61.6	73.9	70.4	65.0	73.5	78.6	69.6
	Memes-CS	64.5	58.8	**75.9**	91.7	80.9	**89.0**	75.1	64.2	81.7	75.8
NLI	CsFEVER-NLI	60.3	64.4	68.4	77.8	69.7	76.8	60.9	56.7	73.4	67.6
	QACG-sum	64.6	69.6	73.8	68.2	79.1	78.5	61.2	65.7	80.3	71.2
	XLMR-SQ2	60.1	55.2	71.8	92.8	76.3	84.4	68.3	55.6	81.0	71.7
Our	AlignScoreCS	64.7	**96.6**	72.5	**93.2**	80.3	82.2	**75.3**	**76.7**	78.6	**80.0**

Table 6.5: The table shows AUC-ROC scores on translated datasets included in the Czech TRUE benchmark. For comparison, there are reported scores of other Czech metrics with their types described in 4.3. The highest scores per dataset are in bold. AVG denotes an average of scores per metric. SE and DF datasets stand for SummEval and DialFact, respectively. The type SM and Reg indicate Similarity Matching and Regression.

Our model achieves the highest average results, suggesting that the multitask learning and the diverse datasets it incorporates have enhanced its performance across the TRUE datasets, covering fact verification tasks in various domains. In addition, Memes-CS delivers high-quality results, securing the second position, followed by XLMR-SQ2 and QACG-sum.

6.6.3 Results on Czech NLI Datasets

We assess the performance of our AlignScoreCS on various Czech NLI datasets to compare it with Czech NLI metrics especially trained on these datasets, described in Section 4.3. The first dataset, CsFEVER-NLI[2], introduced by [Ullrich et al., 2023], is a translation of the English FEVER-NLI version [Nie et al., 2019a] using Deepl. The CsFEVER dataset[3], also by [Ullrich et al., 2023], was created by extracting articles from a Wikipedia dump corresponding to evidence and then applying machine translation. The CTKFacts-NLI dataset[4] for fact verification, introduced by [Ullrich et al., 2023], is a cleaned version of the CTKFacts dataset containing manually labeled claims, extracted from an archive of press agency instead of Wikipedia like CsFEVER. Lastly, the QACG-cs dataset[5] by [Drchal et al., 2023] was generated using the Question Answering for Claim Generation method to create an NLI dataset from Wikipedia data. All datasets serve for fact verification and introduce a classification task into 3 classes (Aligned, Neutral, and Contradict). To evaluate the AlignScoreCS model on multi-class classification, we use our proposed method for classification using the chunking mechanism from Equation 6.1.

The following Table 6.6 displays the F1 macro scores attained by Czech NLI classifiers across Czech datasets. Although our model does not emerge as the top performer, it achieves considerable results despite not being trained on similar data. The best-performing model is XLM-RoBERTa @ SQuAD2 (XLM-SQ2), achieving the highest average score. In the CsFEVER-NLI dataset, our model performs similarly, benefiting from training on the FEVER dataset, but the best-performing model remains CsFEVER-NLI, explicitly trained on this data. In the case of the CsFEVER dataset, developed independently from the English FEVER, there is a slight decrease in performance across all models except for the XLM-SQ2 model, which notably outperforms others. In the CTKFactsNLI dataset, once again, the XLM-SQ2 model performs the best, followed by our model. Notably, the QACG-cs dataset, with its unique design, showcases the best performance for the QACG-sum model, followed by our model.

Type	Metric	CsFEVER-NLI	CsFEVER	CTKFactsNLI	QACG-cs	AVG
NLI	QACG-sum	49.7	34.1	29.9	**82.8**	49.1
	CsFEVER-NLI	**75.2**	25.2	34.2	45.8	45.1
	XLM-SQ2	73.7	**83.2**	**76.9**	31.8	**66.4**
Our	AlignScoreCS	73.6	61.4	64.5	55.9	63.9

Table 6.6: The table shows F1 macro scores in percentages on Czech NLI datasets introduced by [Drchal et al., 2023, Ullrich et al., 2023]. The Czech metrics are described in Section 4.3. The highest scores per dataset are in bold, and the second highest is underlined. AVG denotes an average of scores per metric. CGS and XSF stand for CogenSumm and XSumFaith, respectively.

[2] https://huggingface.co/datasets/ctu-aic/csfever_nli
[3] https://huggingface.co/datasets/ctu-aic/csfever
[4] https://huggingface.co/datasets/ctu-aic/ctkfacts_nli
[5] https://huggingface.co/datasets/ctu-aic/qacg-cs

Chapter 7
Discussion of Part I

In this part, we introduced AlignScoreCS, a multi-task multilingual model-based metric derived from XLM-RoBERTa. This metric is designed to assess factuality in various NLP tasks, including the evaluation of summarizations. AlignScoreCS can evaluate the factuality of summaries by analyzing source text and summaries provided in English and Czech. To enable the evaluation of a longer source text and summary, we utilized the AlignScore function [Zha et al., 2023]. This involves chunking the text into smaller parts and splitting the summary into sentences. Subsequently, each chunk is evaluated against each sentence, resulting in a single consistency score of the Aligned class. AlignScoreCS underwent training for multi-tasking, addressing 3-way classification, 2-way classification, and regression tasks. The training data comprised over 7 million documents from multilingual corpora, encompassing different NLP tasks such as Natural Language Inference (NLI), Fact Verification, Paraphrase, Semantic Textual Similarity (STS), Question Answering (QA), Information Retrieval, and Summarization.

My colleague and I collaborated to translate the English training datasets into Czech versions using the SeamLessM4T [Communication et al., 2023] model with a segment translation procedure. Following this, we unified the individual datasets into a single unified dataset and algorithmically augmented their data to enhance the robustness of our model for self-alignment or varying inputs. Additionally, we translated two English benchmarks, TRUE and SummaC, into Czech using DeepL [Kutylowski, 2017]. The translation process employed two approaches to prevent translation biases and develop noise-tolerant Czech testing data.

We evaluated several models trained on different subsets of the unified training dataset, from which we selected the final AlignScoreCS model. Additionally, we studied the multilingual capabilities of our model by evaluating it on the XNLI dataset, which includes 15 languages. Our findings revealed that the model has the potential to evaluate various languages due to the cross-lingual transfer from its pre-training. However, further fine-tuning may be beneficial to fully leverage this potential.

Experiments conducted on English benchmarks involved evaluating AlignScoreCS to compare its performance with other competitive models. The results showed that we achieved highly comparable results to both models of AlignScore [Zha et al., 2023], although they still surpassed us. However, we outperformed all other state-of-the-art model-based metrics on both benchmarks, achieving a second-place position after the AlignScore models. We surpassed all other evaluated Czech metrics for Czech-translated benchmarks, establishing new baseline results for Czech SummaC and TRUE.

Following this, we developed a multilingual metric that achieved results comparable to AlignScore models and higher than other state-of-the-art models on English datasets while also being applicable to Czech data. Based on the notable results obtained on both SummaC benchmarks, which evaluate the consistency of summaries, we can conclude that it is the most suitable Czech metric available so far for factuality checking in Czech summaries. With this in mind, we integrate this metric in the following Part II for fact-guided text summarization, aiming to enhance the factuality of generated summaries.

Once my colleague completes training the AlignScoreCS model with his approach more aligned with [Zha et al., 2023], we will evaluate and compare which metric performs better, and then report our findings in our repositories of the models.

Part II

Towards Facticity-Driven Text Summarization in Czech

Chapter 8
Summarization

The goal of the **summarization task** is to create a concise and shortened summary from a source document. There are two types of document summarization: Extractive summarization, which selects and copies important fragments directly from the document, and abstractive summarization, which generates summaries containing novel words usually not present in the document.

Abstractive summarization, which we address in this thesis, is typically formulated as a sequence-to-sequence generation problem using encoder-decoder architectures. These architectures generate the summary in an auto-regressive manner. Thanks to pre-trained transformer models trained on extensive corpora of texts, natural language generation tasks, including summarization, have seen significant advancements. Most of these models have achieved state-of-the-art results on various downstream summarization tasks [Lewis et al., 2019, Zhang et al., 2020a, Raffel et al., 2023]. Furthermore, the quality of summarization has been further enhanced by the BRIO paradigm [Liu et al., 2022], which incorporates contrastive loss [Hopkins and May, 2011] and encourages models to adjust their target distribution to better allocate probability mass according to their quality, resulting in state-of-the-art performance. Additionally, for the Czech language, several studies have been published addressing the summarization task [Krotil, 2022, Straka et al., 2018, Hájek and Horák, 2024]. Despite producing high-quality summaries resembling human-written texts, almost all models tend to hallucinate and generate irrelevant sequences, often containing unfaithful facts contradicting the source document, which could negatively impact readers.

To address this challenge, several studies have emerged [Dixit et al., 2023, Chern et al., 2023], which combine the approach with contrastive loss. They leverage aligning the scores with the metric while utilizing factual metrics. In the case of the Czech language, [Halama, 2023] also adopted this approach in their training and achieved state-of-the-art results in quality on Czech datasets. However, they encountered a limitation in improving factuality due to the absence of a factual metric in the Czech domain. This highlights the need for a new factual metric capable of assessing Czech summaries, which we solved here.

In this part of the thesis, we introduce **BARF**: BRIO paradigm with AlignScoreCS and ROUGE_{RAW} Fusion training approach, significantly enhancing the quality and factuality of generated summaries. We integrate our AlignScoreCS metric from the previous Part I, along with quality metrics, into the training of the summarization models to align their scores with these metrics, resulting in facticity-driven text summarization with high quality. Besides, we propose a factually-balanced sorting technique, inspired by [Dixit et al., 2023], for contrastive learning, allowing us to achieve comparable results to baseline mod-

els on English datasets and improve upon some of the state-of-the-art results on Czech datasets, all while maintaining the faithfulness of our summaries. All models are trained on two English news-based datasets, XSUM and CNNDM, and two Czech news-based datasets, CNC and SumeCzech, resulting in multilingual models capable of summarizing texts in both languages. Furthermore, the human evaluation showed that factual accuracy is improved by over 20% compared to core models, with the best-performing BARF-Loop model, which utilizes the application of BARF training multiple times.

8.1 Summarization Objective

The **training objective** of summarization models is to employ maximum likelihood estimation (MLE) to maximize the probability of generating the reference output based on the preceding inputs. Given a document D and a reference summary S*, we can express MLE equivalently as cross-entropy loss, which minimizes the sum of negative log-likelihoods of words (tokens) from the reference summary S*:

$$L_{XENT} = -\sum_{i=1}^{l}\sum_{s} P(s|D, S^*_{<j}) \log p_\theta(s|D, S^*_{<j}; \theta) \tag{8.1}$$

In the first summation, we iterate over the length of the reference summary, while in the second summation, we iterate over the vocabulary distribution of the model at the given step, where the vocabulary represents all tokens the model could produce. The probability $P(.)$ can be computed as 1 if the token s matches s^*_j, otherwise 0, or it can involve a more sophisticated technique such as label smoothing, which assigns a small probability to non-reference tokens as well.

In the case of **inference**, predictions are made based on the document and the partially generated summary $S_{<j}$ by the model, as the reference summary is not available during generation. At generation step t, the model produces the next token as follows: $p_\theta(s_t|D, S_{<t}; \theta)$. Enumerating all possible outputs to find the most probable sequence would be computationally demanding. Therefore, we typically employ approximation techniques to narrow down the search space, which is discussed later.

Chapter 9
Related Work

This chapter outlines the key works that have informed and shaped our thesis by serving as primary sources of inspiration or as crucial components for evaluation purposes. The chapter first introduces the BRIO training paradigm for summarization models. Subsequently, it delves into other works inspired and guided by BRIO. The final sections discuss inference methods for summarization and the metrics utilized for evaluating generated summaries.

9.1 BRIO

BRIO, standing for Bringing Order to Abstractive Summarization, is a novel training paradigm proposed by [Liu et al., 2022]. Rather than relying solely on maximum likelihood estimation to maximize the probability of the reference summary, they also incorporate contrastive learning, defined over various candidate summaries. Contrastive learning, thus, requires the model to accurately predict the ranking order of this candidate set, which is sorted according to automatic metric M. This approach aims to ensure that the model can align its generated scores with the actual quality metrics used to evaluate the summaries. Therefore, according to [Liu et al., 2022], it is assumed that the probability of a candidate should be strongly correlated with its quality as assessed by the metric M. Importantly, this assumption does not hold for models trained using only cross-entropy loss, whose scores are mainly aligned only for the reference summaries, without considering that more appropriate summaries exist.

The authors adopt the ROUGE score as the quality metric for evaluating candidate summaries against their reference summaries. Since reaching all possible candidate outputs would be impossible, authors reduced the set of candidate summaries to the 16 most probable ones per document generated by a pre-trained abstractive summarization model employing diverse beam search algorithm [Vijayakumar et al., 2018], described in the section 9.4. Subsequently, they fine-tuned this model to prioritize better candidates by applying contrastive loss [Hopkins and May, 2011]. In Chapter 11, we extensively discuss contrastive learning, outlining our methodologies for candidate summaries ranking by incorporating two metrics in contrast to [Liu et al., 2022] utilizing only one metric.

In this configuration, the authors trained BART [Lewis et al., 2019] and Pegasus [Zhang et al., 2020a] models, attaining state-of-the-art performance on the English datasets, CNNDM [Nallapati et al., 2016] and XSUM [Narayan et al., 2018]. In opposition, our focus is on leveraging multilingual models to process Czech words and, hence, be able to summarize Czech texts. Furthermore, we prioritize factuality in summarizations, which is described in Chapter 11.

9.2 Enhancing Factuality

After the release of the BRIO training paradigm, several studies emerged focusing on enhancing factuality in summarization. The critical concept involves integrating factuality metrics in place of quality metrics into the training process.

In terms of **Czech** language, [Halama, 2023] introduce Czech implementations of BRIO, aiming to enhance factuality in summarization. The author employs the BRIO training paradigm, integrating various model-based facticity and quality metrics to improve factual accuracy. Multiple models are trained, each utilizing different metrics, including ROUGE, their optimized BERTScore [Zhang et al., 2020b] and BERTSource, and the factual metric Memes-CS [Šimon Zvára, 2022]. Evaluation metrics such as ROUGE and BERTSource are employed to assess candidate summaries against the text, while Memes-CS and again ROUGE are used for comparisons against reference summaries. Despite achieving state-of-the-art results for the SumeCzech dataset, the human evaluation revealed that the models performed poorly in terms of factuality. Therefore, the authors recognized the need for a new factual evaluation metric for summarization. By comparison, our approach prioritizes the utilization of our newly developed Czech factual metric, AlignScoreCS, showing considerable results for summarization benchmarks to assess the factuality scores. Additionally, during the candidate sorting process, we combine this metric with the quality metric ROUGE_{RAW}, discussed in the section 9.5.

Improving factuality in **English** summarization was addressed by [Dixit et al., 2023, Chern et al., 2023] through the BRIO training framework. [Dixit et al., 2023] introduce the EFact-Sum model, drived from BART [Lewis et al., 2019], trained using BRIO. They implement a novel candidate ranking approach, combining the factual metric FactCC with ROUGE, which is fully elaborated in Section 11.3. Although their ROUGE scores did not surpass those of BRIO [Liu et al., 2022], they enhanced factuality according to FactCC and DAE. In opposition, our focus lies in Czech summarization, utilizing the new AlignScoreCS metric. In [Chern et al., 2023], the authors compare models trained on candidates sorted by factuality metrics BARTScore and DAE with those sorted by ROUGE metrics. Their human evaluation revealed that models using factuality metrics produce factually consistent summaries.

9.3 LoRA & QLoRA Technique

Quantization and Low-Rank Adaptation (QLoRA), proposed by [Dettmers et al., 2023], is a technique designed to improve memory efficiency while training models with billions of parameters. Indeed, this method enables fine-tuning of these LLMs on a single GPU while maintaining significant performance. **QLoRA** builds upon the concept of Low-Rank Adapters (LoRA) and 4-bit quantization, contributing to memory and computationally efficient training. To elaborate, **LoRA**, introduced by [Hu et al., 2021], consists in selecting specific modules within the model, typically query or key layers of attention modules, and adding only a small subset of the model's trainable parameters, called adapters. The remaining pre-trained parameters are frozen, reducing the overall number of learned parameters and the final model's size. Only the adapters' weights are updated during fine-tuning for specific tasks, while the pre-trained weights remain unchanged.

To further enhance memory efficiency, QLoRA introduces a novel 4-bit quantization to the model. **Quantization** refers to the process of converting high-precision floating-point parameters into lower-precision format or integer representations. In the case of 4-bit quantization, 32-bit floating-point pre-trained parameters are reduced into 4-bit integers, resulting in a range from -8 to 7. Although quantization may lead to decreased performance due to approximation, when combined with low-rank adapters during training, the models perform similarly to fully fine-tuned models, as demonstrated by QLoRA. Furthermore, QLoRA allows the computations to run simultaneously on both GPU and CPU, reducing the chances of encountering GPU out-of-memory errors. Following the fine-tuning procedure, the model consists of the original pre-trained weights in 4-bit format and additional low-rank adapters in their higher precision format. When it comes to inference, we need to merge the trained layers of the reduced model with its original model to ensure effective generation. We employ this technique to fine-tune large language models on our datasets efficiently.

9.4 Inference

Various generation methods are used for inference in summarization models, all aimed at improving the quality of the generated summaries and approximating the search space. As the model produces output based on the source text and previously generated output, in each step of generation, it assigns probabilities to its vocabulary to choose the new most likely token to generate. One major group of methods is sampling-based, which involves randomly selecting the next token from a probability distribution instead of always choosing the most probable one. This approach can lead to the discovery of diverse and potentially better summaries. Representatives include **random sampling**, which selects words based on their probabilities, **top-k sampling** [Fan et al., 2018], which samples from the top k tokens, and **top-p sampling** [Keskar et al., 2019], where a p parameter determines the maximum cumulative probability of tokens forming a new distribution from which it samples. In this study, except for LLMs, we opt not to explore random sampling and avoid employing it. Instead, we adopt a more conservative approach, focusing on beam search and its variations.

Beam search is one of the most used generation methods. This approach consists in searching for the most probable sequence overall rather than just selecting the most probable token at each step. Essentially, it maintains a set of b sequences (referred to as beams) at each generation step and chooses the most probable beam at the end. However, this method has drawbacks, particularly in tracking highly similar sequences that differ only in a few words. To address this limitation, we employ **Diverse Beam Search** [Vijayakumar et al., 2018], which divides the beams of a beam search into groups of equal size. Each group then operates like a separate beam search, looking for its new beams. The approach incorporates a diversity penalty hyper-parameter to guarantee variation among tokens across different groups. This technique is employed directly by [Liu et al., 2022], as well as in our work, to produce diverse candidate summaries.

9.5 Metrics

We assess summaries using two categories of evaluation metrics: qualitative and factual. Although we introduce some of these metrics in the previous Part 4.3, we elaborate on

them here in greater detail as they serve for evaluation purposes.

To evaluate the **quality of summarization**, we employ automatic metrics such as **ROUGE** [Lin, 2004] and **ROUGE**$_{RAW}$ (a language-agnostic variant of ROUGE implemented by [Straka et al., 2018]). The ROUGE-family metrics are widely used across the NLP field due to their simplicity in computing the overlap of n-grams between the generated and reference (gold) summaries. Following this, these metrics do not solely reflect the overall quality of the summary but rather evaluate how well generated summaries match their human-written references. By this, we aim to gauge how effectively the summary captures the essence of the gold standard, focusing on at least capturing certain keywords. In the subsequent paragraphs, we still refer to it as the quality metrics since there are currently few alternatives that assess its quality better and are used widely. Specifically, ROUGE can be computed for various n-gram sizes, typically 1-gram, 2-gram, and L-gram, representing uni-gram, bi-gram, and the longest common subsequence overlaps, respectively. For each n-gram size, the metric calculates precision and recall as follows.

$$\text{Precision} = \frac{|\text{n-grams}|}{|\text{n-grams in system S}|} \qquad \text{Recall} = \frac{|\text{n-grams}|}{|\text{n-grams in reference S}|} \tag{9.1}$$

Here, |n-grams| represents the count of overlapping n-grams between the system and reference summaries, while |n-grams in system S| and |n-grams in reference S| denote the total counts of n-grams in the system and reference summaries, respectively. Since the recall indicates how well the system summary captures the reference one and the precision informs about the presence of different words in the system summary, we also report the F1-score (f-score), which is the harmonic mean of precision and recall:

$$\text{F1-score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{9.2}$$

Additionally, we employ the multilingual model-based metric **BERTScore** [Zhang et al., 2020b] for quality assessment, the same model as in Part I. Unlike ROUGE, BERTScore considers the contextual embeddings of tokens obtained from a pre-trained encoder model, allowing it to capture word dependencies from various perspectives. It computes the cosine similarity between reference tokens and the most similar tokens from the generated text, providing precision, recall, and F1-score measurements, of which we report only the F1-score. We use this metric to measure the similarity scores between generated summaries and source texts.

In the course of assessing **factuality of summarization**, we employ our **AlignScoreCS** model for Czech and English, which is fully elaborated in the previous Part I, and **Memes-CS** [Šimon Zvára, 2022] model-based factuality metric for Czech, also described in Section 4.3. Furthermore, in addition to Czech, we include the English **FactCC** evaluation metric [Kryscinski et al., 2019]. This metric is derived from BERT and is trained on a dataset that underwent a series of rule-based transformations applied to the sentences of source documents. The model is initially trained for binary classification and then fine-tuned for three tasks: 1) identifying whether sentences are factually consistent after transformation, 2) extracting a span in the source documents to support consistency prediction, and 3) extracting a span in the summary sentence that is inconsistent with the text. The model has demonstrated exemplary performance in human evaluations. AlignScoreCS and FactCC evaluate summaries against the source text, whereas Memes-CS assesses summaries against the reference summaries.

Chapter 10
Datasets

This chapter presents an overview of the summarization datasets utilized for training and evaluating our models. We focus on **summarizing news-based articles** in two languages, Czech and English. The initial section delves into the English summarization datasets. Subsequently, we detail two Czech news summarization datasets and outline our filtering approach. Finally, the chapter concludes with a description of the comprehensive dataset, combining all data from the included datasets.

10.1 English Datasets

To compare the performance of our models with state-of-the-art ones trained solely on English news data, we include two mainstream English summarization datasets. This allows us to evaluate comparable results and, accordingly, to estimate that the observed behavior might also hold for the Czech data. By including English in our training dataset alongside Czech, we also aim to improve the performance of our models due to the cross-lingual transfer that can occur when models encounter different data.

The **XSUM** dataset, introduced by [Narayan et al., 2018], encompasses documents drawn from the British Broadcasting Corporation[1] (BBC). The XSUM covers a wide variety of domains, including news, politics, weather, business, science, and others. Each document is divided into text and abstract (reference summary), where the abstract is concise concerning other datasets but remains abstractive regarding the authors. The dataset provides three splits for training, testing and validation.

The **CNNDM** dataset comprises English articles sourced from the Cable News Network[2] (CNN) and Daily Mail[3] (DM), initially designed for question answering but later adapted for the task of summarization by [Nallapati et al., 2016]. Compared to other datasets, this dataset tends to be more extractive, often presenting information in bullet points to highlight key aspects of the article. The dataset is divided into 3 splits (test, validation, and training), where each document is characterized by a text and abstract as well.

[1] https://www.bbc.com/
[2] https://edition.cnn.com/
[3] https://www.dailymail.co.uk/home

10.2 Czech Datasets

For Czech summarization datasets, we showcase the publicly accessible Sumeczech dataset and the private CNC dataset provided by the supervisor. Additionally, we outline our automated filtering method to remove unsuitable samples, including those containing anomalies.

The **SumeCzech** dataset, developed by [Straka et al., 2018], comprises over a million documents sourced from five prominent Czech newspaper websites: *ceskenovinky.cz*, *idnes.cz*, *denik.cz*, *lidovky.cz*, and *novinky.cz*. Each document is categorized into three sections: a headline, serving as the title; an abstract, representing the ground truth summary of the article; and a text, constituting the entire article. The dataset underwent a cleaning process using specific heuristics, after which the authors partitioned the collected documents into four subsets: training, validation, and test and out-of-domain test sets.

The private **CNC** dataset, obtained from the Czech News Center company[4] and supplied by the supervisor, comprises Czech articles sourced from various online media platforms such as *reflex.cz*, *e15.cz*, *blesk.cz*, *isport.cz*, and others. The distribution proportions of data from each website remain unknown. Each document is segmented into three sections, reflecting the structure of the SumeCzech dataset. We utilize a filtered version of this dataset from our prior work on a bachelor's thesis [Krotil, 2022].

10.2.1 Data Filtering

Upon examining the Czech summarization datasets documents, we found persistent inconsistencies despite being subjected to cleaning processes. In particular cases, abstracts fail to capture the essence of their corresponding texts. Indeed, certain abstracts frequently consist of bullet point lists, while others contain unrelated descriptions of videos found on the websites or include hyperlinks leading to external sources. Furthermore, we encountered some documents whose abstracts consist of only a few words, although their respective texts comprise hundreds of words. By acknowledging these discrepancies, we suspect they might bring noise into the training process, potentially decreasing the performance of the trained models when dealing with such data.

To eliminate the adverse effect, we 1) analyze the statistical properties of datasets using the extractive fragment segmentation method, 2) investigate summaries generated by the models trained on these datasets, and 3) filter out potentially wrong samples based on our findings. To characterize our datasets, we implement and employ the extractive fragment procedure proposed by [Grusky et al., 2018]. This technique relies on creating a set of extractive fragments from two texts, in our case, the text and its corresponding abstract. The process entails examining each position within the summary: if a sequence of words in the source text matches the beginning of the remaining summary text, it identifies this sequence as extractive and proceeds iteratively, prioritizing the longest possible prefix at each step. The procedure yields three statistical metrics: *coverage*, quantifying the degree of overlapping extractive fragments (expressed as a percentage) between the text and the summary; *density*, measuring the average length of the extractive fragment; and *compression*, which is the ratio of the text length to the summary length. Findings are detailed in Appendix C.1; for curiosity, we also add statistics for English datasets

[4] `https://www.cncenter.cz/`

alongside values computed for generated summaries.

In addition, we leverage a model HT2A-CS trained on a concatenation of both Czech datasets by [Krotil, 2022] to generate summaries. We then analyze the statistical properties of the texts and the summaries produced by this model, aiming to compare them with the statistical values computed for the texts and the ground truth abstracts. Interestingly, our observations indicate that the model tends to generate more extractive summaries, with coverage values averaging around 30% compared to the ground truth abstracts, which yield coverage of around 10% on Czech datasets. This suggests that the generated summaries rely more heavily on the sentences from the texts than the ground truth abstracts. As a result, when we investigated the samples with coverage values lower than 3%, we encountered instances with the aforementioned inconsistencies in abstracts, while the corresponding generated summaries mostly conveyed the information from the article. According to our findings, despite the possibility for the model to become more extractive, we opted to filter out samples with a coverage lower than 5%. Additionally, we retained only those samples with a compression ratio ranging between 2 and 35, excluding instances where the text length was either similar to the abstract length or excessively long. We applied these heuristics to training and validation sets only; the test sets remain unchanged. Table 10.1 displays the resulting dataset sizes.

10.3 Final Dataset

Our training and validation dataset splits are created by combining two Czech and two English datasets: SumeCzech, CNC, XSUM, and CNNDM. The original testing splits of the datasets remain unchanged for testing purposes. The table below 10.1 displays the sizes of our final dataset, including the reduced data for Czech datasets after applying the filtering methods. Our Final dataset used for training is shown in the last row. In addition, we provide further details on our datasets in the appendix C.1.

Language	Dataset	train		validation		test
		Initial	Final	Initial	Final	
English	XSUM	-	204,045	-	11,332	11,334
	CNNDM	-	287,113	-	13,368	11,490
Czech	SumeCzech	867,596	374,442	44,454	19,412	44,567
	CNC	675,225	402,799	35,000	15,971	35,000
Multi	Final	-	1,268,379	-	60,083	-

Table 10.1: Overview of Czech and English Summarization Datasets. Datasets are categorized into Czech and English based on their language. "Initial size" refers to the original size of the dataset set, while "Final" indicates the size used for training. In the case of Czech datasets, the "Final" size represents the amount of data retained after applying the filtering method. If "Initial size" is "-", no filtering was applied, and the initial size is the same as the final size. The test sets remained unchanged and retained their original size from the corresponding datasets. The last "Final" dataset concatenates all datasets into one.

Chapter 11
Methodology

Drawing insights from [Liu et al., 2022, Dixit et al., 2023], we integrate the contrastive loss with the cross-entropy loss 8.1, aiming to refine our models by assigning a higher probability to the candidate summaries with higher factuality scores as well as higher quality scores. Given a document D, reference summary S^* and a corresponding candidate summary set S, we apply **contrastive loss**, demonstrated in the studies by [Hopkins and May, 2011, Liu et al., 2022], as follows:

$$L_{CTR} = \sum_{i}^{N} \sum_{j>i}^{N} max(0, f(S_j) - f(S_i) + \lambda_{ij}) \tag{11.1}$$

Here, S_i and S_j represent two distinct candidate summaries, with $M(S_i) > M(S_j)$, where M denotes a utilized metric, and N signifies the size of the candidate set. Authors of [Liu et al., 2022] use ROUGE as a quality metric M, resulting in ROUGE(S_i,S^*) > ROUGE(S_j,S^*). Our approach of utilizing metrics and candidate sorting strategy is described in Section 11.3 bellow. The margin $\lambda_{ij} = (j - i) * \lambda$ is calculated by multiplying the difference in rank between the candidates with the margin value λ. Additionally, $f(S_i)$ represents the length-normalized estimated log-probability.

$$f(S) = \frac{\sum_{t=1}^{l} logp_{g_\theta}(s_t|D, S_{<t}; \theta)}{|S|^\alpha} \tag{11.2}$$

Where α is the length penalty. To maintain the model's generation capabilities, we need to integrate cross-entropy loss 8.1, given that auto-regressive generation relies on both token-level prediction and sequence-level coordination. Therefore, as proposed in [Liu et al., 2022], we adopt a multi-task approach with a unified loss function that combines contrastive and cross-entropy losses as follows:

$$L_{COMB} = \gamma_1 L_{XENT} + \gamma_2 L_{CTR} \tag{11.3}$$

where γ are the weights of individual losses.

Before we delve into the detailed methodology, we first outline the individual stages of our technique and how they are integrated into the training of BARF models, as illustrated in Figure 11.1. We begin with a pre-trained model, which we fine-tune for summarization using the cross-entropy loss 8.1 between the generated outputs and reference summaries. This step enables the model to produce the necessary summaries for subsequent phases. We refer to this as the core model because it will undergo further factual refinement. Next, we generate N candidate summaries for each document using the core model. These candidate summaries are then filtered and sorted based on our factually-balanced method

using AlignScoreCS and ROUGE_{RAW}. Finally, we further fine-tune our core model using a combined loss 11.3 and inputs consisting of documents, reference summaries, and sorted candidate summary sets.

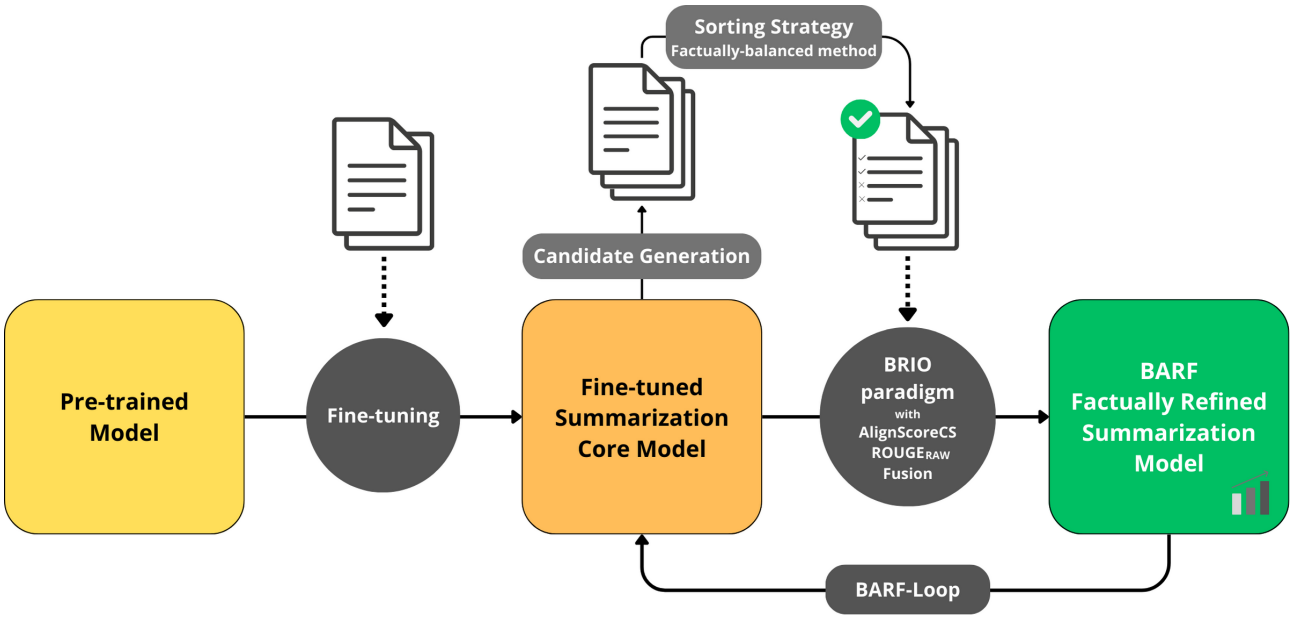

Figure 11.1: The BARF diagram illustrates the developmental stages of the BARF technique. It begins with a pre-trained model and progresses through fine-tuning for summarization, generating candidates, sorting strategy utilizing the factually-balanced method, and applying the BRIO paradigm with AlignScoreCS and ROUGE_{RAW} Fusion to create the final BARF model. BARF-Loop utilizes the BARF model to generate candidates and undergoes multiple training rounds.

In the upcoming sections, we introduce the individual stages of our methodology. We begin with core models, proceed to the candidate generation process, and then delve into candidate sorting using various strategies.

11.1 Core Models

Leveraging our previous efforts to enrich Czech summarization with multilingual data integration, spanning various languages, we employ two core models, **mBART25**[1] and **mT5**[2], fine-tuned on the Multilarge dataset. The dataset encompasses 3.5 million training documents from SumeCzech, CNC, XSUM, CNNDM, MLSum, and CNewSum news-based summarization datasets. This presents a notable advantage, as our core models have undergone training with the datasets used for further fine-tuning, aligning with the methodology of BRIO core models. In this case, the Czech datasets were used in their entirety without applying our filtering method. The MLSum dataset [Scialom et al., 2020], constructed in a manner similar to CNNDM, including news articles and offering a comparable number of training samples per language (excluding Russian), serves as a robust multilingual extension. It encompasses articles from newspapers in five distinct languages: German,

[1] https://huggingface.co/ctu-aic/mbart25-multilingual-summarization-multilarge-cs
[2] https://huggingface.co/ctu-aic/mt5-base-multilingual-summarization-multilarge-cs

French, Spanish, Turkish, and Russian, resulting in 1.5 million document-summary pairs. The CNewSum [Wang et al., 2021] is a monolingual Chinese summarization dataset that gathers news submissions from hundreds of thousands of press publishers. It provides a large-scale document-level summarization dataset with an abstraction comparable to short social media datasets. Overall, it contains over 300 thousands of samples.

Additionally, we attempted to utilize two large language models **Aya-101**[3] and **Falcon**[4] as another core models. Although these models already underwent fine-tuning for specific instructional tasks, we opt to fine-tune them on our datasets further to align with our summarization structure, as detailed in the subsequent chapter on Experiments 12.1. However, we encountered difficulties in integrating them into the BRIO training paradigm, which is also explained there. As a result, we do not discuss them in the following sections regarding BRIO fundamentals.

11.2 Candidate Generation

Following the study [Liu et al., 2022], we generate candidate summaries to further refine our core models. Standing by the multilingual nature of our dataset, we ensure proper initialization of the starting decoder language token. This step is crucial as both the mBART and mT5 core models require the starting decoder token to be in the appropriate language to begin generation correctly. Additionally, we concatenate the headline with the text for Czech datasets to provide additional context.

We decide to generate **16 candidate summaries** for each document in our entire training dataset using the diverse beam search algorithm [Vijayakumar et al., 2018] for producing distinct candidates. Following the generation process, we evaluate the ROUGE_{RAW} metric for each candidate summary with respect to its reference summary within each document. Similarly, we conduct AlignScoreCS assessments, although we evaluate the candidate summary against its source text in this case. Given the massive scale of our dataset consisting of over 1 million documents, the generation of 16 candidates per document, and the chunking mechanism utilized by AlignScoreCS during inference on texts and their candidates, the computational requirements are significant and demanding. Generating and assessing candidates for the entire dataset on a single GPU would take weeks. Therefore, we employed multi-GPU computing, especially 12 GPUs, allowing each GPU to process chunks of data independently. Moreover, we generate candidates and compute scores twice, once for the mBART25 core model and once for the mT5 core model. Candidate summaries generated in this manner are used in almost every experiment involving various approaches to sorting the candidates for contrastive learning (Section 11.3), except for experiments with BRIO looping. In BRIO looping, we generate candidates differently, as described in Section 12.2.4. In Appendix C.1, we include the distribution of candidates' ROUGE_{RAW} scores plotted against their AlignScoreCS scores, generated by the core model, mBART25, for an additional study.

[3] `https://huggingface.co/CohereForAI/aya-101`
[4] `https://huggingface.co/tiiuae/falcon-7b-instruct`

11.3 Candidate Sorting

In the upcoming paragraphs, we outline our approaches for **candidate sorting**, which involves sorting candidate sets based on a sorting metric M. Subsequently, these sorted sets of candidate summaries are utilized in the fine-tuning process of our core models within the contrastive loss framework (Eq. 11.1). Given a document D, reference summary S^*, and a corresponding set of 16 candidate summaries S, we utilize different sorting methods M[5]:

- **Single Metric Strategy**
  Firstly, we prioritize aligning our model with the factuality scores of candidates. Consequently, we employ the sorting metric M as the AlignScoreCS, and we obtain AlignScoreCS(S_i,D) > AlignScoreCS(S_j,D) for Equation 11.1.

  Additionally, we follow the same approach to [Liu et al., 2022], in which we use the sorting metric M as the ROUGE$_{RAW}$, ensuring ROUGE$_{RAW}$(S_i,S^*) > ROUGE$_{RAW}$(S_j,S^*) for Equation 11.1. Here, we aim to observe how models behave when coordinating with the quality metric and compare them with other techniques.

- **Double Metric Strategy**
  In this scenario, our objective is to optimize the model for both factuality and quality concurrently. We use AlignScoreCS to gauge factuality and ROUGE$_{RAW}$ to assess quality. Once the scoring of candidate summaries based on these factuality and quality metrics is done, we proceed to select four faithful candidates with the highest ROUGE$_{RAW}$ scores and four unfaithful candidates with the lowest ROUGE$_{RAW}$ scores while considering a facticity score above 0.5 for the faithful subset. Inevitably, this process filters out documents whose candidate sets do not meet the criteria of having four faithful and four unfaithful candidates, thereby simultaneously reducing the candidate set size from 16 to 8 candidates. Subsequently, we employ various sorting mechanisms on these subsets. The faithful subset is sorted based on the quality metric, while the unfaithful subset is sorted using the factuality metric. We call this filtering and sorting strategy as **factually-balanced method** and demonstrate the condition of sorting in the following equation:

$$M(S_i) > M(S_j) \begin{cases} \text{if } as(S_i, D) > 0.5 \text{ and } as(S_j, D) \leq 0.5 \\ \text{if } r(S_i, S^*) > r(S_j, S^*) \text{ and } as(S_i, D), as(S_j, D) > 0.5 \\ \text{if } as(S_j, D) < as(S_i, D) \leq 0.5 \end{cases}$$

  Where as denotes AlignScoreCS metric and r is ROGUE$_{RAW}$. Through this approach, we hypothesize that the model could adjust its scores to align with the quality metric when the scores are coordinated with faithful candidates' scores and with the factuality metric when the scores are synchronized with the unfaithful candidates' scores.

  In addition, we take inspiration from [Dixit et al., 2023] and adopt their proposed sorting mechanism, where we choose documents containing at least 2 faithful and at least 2 unfaithful candidate summaries, filtering out others. While [Dixit et al., 2023] decided to use only 6 candidates in the candidate summary set, we opt for 12 candidates in the candidate set, assuming that encountering more candidates may

[5]Whenever we mention ROUGE$_{RAW}$ metric in terms of candidates sorting, we refer to the averaged F-scores computed as F=$\frac{F1+F2+FL}{3}$. Regarding this averaged score, we sort the candidate summaries.

lead to better ranking scores. Following this, we sort the candidates based on factual scores first. Then, within each candidate subset (faithful and unfaithful), we sort the candidates based on ROUGE_{RAW} scores, retaining the highest-scoring faithful and the lowest-scoring unfaithful sets. Notably, we share a similar approach described above; however, theirs differs in that both factual subsets within the candidate set are sorted based on the quality metric. Furthermore, this method can be considered as a weaker condition, as it may contain more faithful candidate summaries than unfaithful ones within a single candidate set, or vice versa, unlike our candidate ranking, where both factual subsets' sizes are equal to 4. Therefore, we call this approach as **factually-unbalanced method**.

The overview provided in Table 11.1 summarizes our methods for sorting candidates using various metrics, focusing solely on candidates generated by core mBART25, as most of our experiments are conducted on them. The mT5 model serves as a practical example to illustrate how these methods can be applied to other models. We also present a final number of remaining samples after the application of filtering methods in case of double metric strategy, where candidates are filtered based on their faithful and unfaithful criteria[6].

Metric Strategy	Train Strategy	Sorting Strategy	Samples	Candidate Set	Reference
Single Metric Strategy	Single Metric Fusion	AlignScoreCS	1,268,379	16	False
		ROUGE_{RAW}	1,268,379	16	True
Double Metric Strategy	Double Metric Fusion	factually-balanced	656,925	8	False
		factually-unbalanced	944,526	12	False

Table 11.1: Summary of sorting strategies for candidates: The "train strategy" indicates the term for the application of metric strategy into the BRIO paradigm, referring to fusion. The Sorting strategy denotes the method applied to candidate summaries. Samples are the number of remaining samples after the filtering method is applied. The candidate set denotes the number of summaries within, while the Reference indicates whether we include the reference summary in the candidate set.

[6]Precisely for factually-balanced strategy and candidates generated by core mBART25, CNNDM contains 84,230, XSUM 123,687, CNC 241,748, SumeCzech 207,260, totaling 656,925.
For factually-balanced strategy and candidates generated by core mT5, CNNDM contains 28,603, XSUM 135,914, CNC 272,075, SumeCzech 219,186, totaling 655,778.

Chapter 12
Experiments

In this chapter, we conduct experiments focusing on improving factuality and quality in summarization, simultaneously. The initial section outlines our efforts to fine-tune LLMs on our summarization datasets. Subsequent sections delve into the training and analysis of our models using the BRIO paradigm and various sorting approaches. When integrating a single metric strategy for candidates sorting into the BRIO, we refer to it as **single metric fusion**, whereas, for double metric strategy, we use the term **double metric fusion**. Following experiments conducted on subsets of validation data, we select the best-performing models, which we then evaluate on the original test data of our datasets and compare with other baseline models. Lastly, we discuss the trade-off between abstractiveness and faithfulness, followed by a human evaluation of produced summaries from our models.

12.1 Few-Shot Fine-tuning

Considering that LLMs, Falcon and Aya, possess 7 and 40 billion trainable parameters, respectively, fine-tuning them on the complete training dataset would be time-consuming and ineffective. We assume that they could learn quickly from just a few samples due to their sizes. Drawing inspiration from a study by [Fabbri et al., 2021a] showcasing the efficacy of few-shot learning in refining pre-trained models for text generation tasks, we opt to adopt this approach for fine-tuning our models. According to the methodology outlined in the study, we randomly sample data segments from each dataset. The study adhered to strict criteria and used only a few hundred samples. Our goal is to preserve the knowledge of the summarization task for both models as well as improve Falcon's understanding of the Czech language, which it currently demonstrates only to a limited extent. Therefore, we randomly select 9% of the size of each dataset, resulting in samples of $11,500$ from XSUM, $25,840$ from CNNDM, $33,699$ from SumeCzech, and $36,251$ from CNC, totaling $115K$ samples.

When dealing with LLMs, we establish specific prompts tailored to the summarization task. Informed by the observations from Table C.1, which suggest that models often generate summaries of inappropriate lengths concerning the reference ones, we aim to address this issue through the use of prompts. We employ a prompt optimized for short summaries for the XSUM dataset, characterized by single-sentence summaries. In contrast, we only apply a short summary prompt for other datasets when the compression ratio exceeds twice its average value. Otherwise, we use a long summary prompt. Additionally, we design separate prompts for each language, resulting in four prompts: two per language and two per summary length. An example of a prompt template is *"Provide short summary of this article:\n{text}\nSummary:\n{summary}"*, where the text indicates a source text

and the summary is a reference summary. In the case of Falcon, we format our training samples in this manner due to our focus on Casual Language Modeling. This approach trains a decoder-only model to predict the next token in a sequence based on the preceding tokens. As a result, during inference, we only input the prompt without the summary, and the model generates the summary accordingly. In contrast, for the Aya-101 model, we utilize a part of the prompt up to the Summary as the input to the encoder, while the rest serves as the reference output for the decoder. This approach is adopted because the architecture relies on Sequence to Sequence Language Modeling.

The input length of the prompt is set to 1024 tokens, with the maximum text length limited to 768 tokens, leaving the 256 tokens reserved for the reference summary. Subsequently, we fine-tune both models using the 4-bit quantization of the main model through the QLoRA paradigm. We configure both $LoRA_\alpha$, referring to a scaling factor for weights, and $LoRA_r$, indicating the rank of the LoRA parameters, to 32, with a dropout rate of 0.1. We employed Hugging Face Transformers [Wolf et al., 2020] and its Peft libraries, facilitating the adaptation of the model to the QLoRA setting. The training process spans 3 epochs with a batch size of 32, a learning rate of $3e$-5, and utilizes a linear learning-rate scheduler with a warm-up ratio set to 6%. In the case of **Falcon**, when examining the produced summaries, the model exhibits improved Czech language generation but to a limited extent, making it unable to generate fluent Czech texts. Additionally, it frequently mixes Czech and English words when generating Czech summaries. Consequently, we evaluate this model only on English data. On the other hand, the **Aya-101** model demonstrates significantly better generation of Czech texts, even though the training was conducted with only 500 optimization steps before the gradients collapsed, which was addressed by [Kalajdzievski, 2023] suggesting using $LoRA_\alpha$ n-times lower than LoRA_R. However, we leave this for future explorations of LLMs and proceed to assess these models on our datasets in the subsequent Sections 12.3. Moreover, we could only fit up to two batches due to their sizes simultaneously, rendering them unsuitable for BRIO training, which requires more candidate summaries in one input.

12.2 BRIO Align Fusion

The following sections explore the fine-tuning of summarization models using the BRIO paradigm with ROUGE_{RAW} and AlignScoreCS fusion (further **BARF**), as described in Section 11.3, to enhance the factual accuracy of generated summaries. We begin by discussing the training configuration, followed by a brief overview of the implementation background. Finally, in this section, we present the results on a subset of validation sets to compare models trained using different approaches, such as various candidate sorting methods or hyperparameter settings. Based on these results, we select the best-performing models for evaluation on the test data from both Czech and English datasets.

12.2.1 Training

We trained several core models using the combined loss (Eq. 11.3). All models followed the same hyperparameter settings during training, except for a few exceptions, which we discuss later. The training was conducted using 4 NVIDIA A100-SXM4-40GB GPUs, with each model trained for only 1 epoch, as demonstrated to be sufficient by [Liu et al., 2022, Halama, 2023]. We utilized a device batch size of 16, resulting in an overall batch size of 64, and set the learning rate to $1e$-5 with a linear learning rate scheduler. Warm-up

steps were configured to comprise 6% of the training dataset size. Furthermore, we set the weight decay to 0.01 and dropout to 0.1. Following [Liu et al., 2022], for the contrastive learning component, we adjusted the length penalty α to 2.0 in Equation 11.2, and set the margin value λ to 0.001 for the contrastive loss in Equation 11.1. The weights of the combined loss were arranged to $\gamma_1 = 1$ for the cross-entropy loss and $\gamma_2 = 100$ for the contrastive loss.

Regarding candidate summaries, we employed different numbers of summary candidates for various sorting methods, as depicted in Table 11.1. With the size of the candidate summary set denoted as N, the batch size becomes N times larger, resulting in a maximum of 1024 input samples for a single metric strategy. Nevertheless, this does not alter the properties of the cross-entropy loss; we still compute the loss from one generated output, while the other computed outputs serve as scores for calculating the contrastive loss within each step. For the single metric sorting method using only ROUGE_{RAW}, we included the reference summary in the computation of the contrastive loss. However, when AlignScoreCS is incorporated, we exclude it.

12.2.2 Implementation Details

Once again, we leveraged the Hugging Face Transformer framework [Wolf et al., 2020] for its ease of use of transformer models and efficient fine-tuning capabilities. We took advantage of the provided Seq2SeqTrainer[1] (S2ST), suited for encoder-decoder models, which allows them to generate summaries during the validation phase for further metric evaluation. Consequently, we developed a custom metrics evaluation for the validation phase, which we integrated into the trainer. This approach facilitated tracking desired metrics during training, which we watched through the Weight and Biases framework. Furthermore, we designed a custom model saver to save the best-performing model in progress.

To integrate the BRIO training paradigm, we developed a custom BrioTrainer class, inheriting from the S2ST class, in which we implemented the computation of the combined loss. Subsequently, we implemented the contrastive loss using the PyTorch library and its MarginRankingLoss[2] function. Additionally, we created a custom Dataset[3] class to encapsulate the candidate summaries and prepare inputs for the trainer. This class directly provides tokenized candidate summaries, along with text, reference summaries, and labels, resulting in the candidate summary set size output. For the mBart25 and mT5 models, we set tokenization to 512 tokens for the encoder and 128 tokens for the decoder. Consequently, the BrioTrainer could feed the model with a batch size of 64, with each sample containing the candidate inputs of size N (16, 12, or 8). Reasonably, this setting would exceed memory limits. Therefore, we integrated gradient accumulation techniques.

12.2.3 Align Fusion Analysis

In this section, we analyze the performance of models on a subset of validation data. We randomly select 7,000 documents per dataset and generate summaries to compare models

[1] https://github.com/huggingface/transformers/blob/main/examples/legacy/seq2seq/seq2seq_trainer.py

[2] https://pytorch.org/docs/stable/generated/torch.nn.MarginRankingLoss.html

[3] https://github.com/huggingface/datasets

trained using different approaches, including sorting methods and hyperparameter settings.

We assess two models trained with candidate summaries sorted using single metric strategies: **BRIO-Align** and **BRIO-Rouge**, which utilize the AlignScoreCS and Rouge_{RAW} ranking methods, respectively. Following, we evaluate two models that employ the factually-balanced method for candidate sorting: **BARF-100** and **BARF-10**, where the numbers 100 and 10 represent the γ_2 weights of the contrastive loss in the combined loss (Eq. 11.3) used in the training, respectively. Additionally, we analyze two other models that utilize an altered factually-balanced method, where we apply the same filtering approach for faithful and unfaithful sets but then sort the candidates in each factual set independently based on AlignScoreCS for one model and on Rouge_{RAW} for the other model, **BARF-Align** and **BARF-Rouge**, respectively. This allows us to examine whether factually-balanced sets alone can enhance factuality. Finally, we evaluate one model, **AlignSum**, whose candidates are sorted according to the factually-unbalanced method [Dixit et al., 2023]. For other unspecified training hyper-parameters, we use the settings described in Section above 12.2.1. These models are fine-tuned from our **mBART25 core model**. We provide results on the subsets of validation data from each dataset in the following Table 12.1.

Fusion	Model	XSUM (val)				CNNDM (val)				SumeCzech (val)				CNC (val)			
		F1	F2	FL	AS	F1	F2	FL	AS	F1	F2	FL	AS	F1	F2	FL	AS
Core	mBart25	39.8	16.5	30.8	51.3	43.6	20.5	29.9	89.0	22.4	7.3	16.3	68.1	23.6	7.6	16.9	61.0
SMF	BRIO-Align	39.9	16.7	31.1	58.0	42.5	20.2	29.7	92.2	21.4	6.8	15.5	73.1	22.8	7.3	16.3	67.7
	BRIO-Rouge	**40.6**	16.8	**31.6**	43.3	41.4	19.7	29.6	87.8	**26.2**	**9.5**	**19.7**	55.2	**29.3**	**11.7**	**22.1**	51.1
DMF	AlignSum	39.9	16.5	31.0	69.6	**44.4**	**21.7**	**31.2**	91.5	23.8	8.0	17.6	76.8	24.3	8.4	18.0	73.8
	BARF-Align	39.5	16.2	30.6	**72.8**	44.1	21.1	30.5	92.2	22.8	7.5	16.8	**82.8**	23.3	7.9	17.2	**79.0**
	BARF-Rouge	40.2	16.6	31.4	53.8	39.9	19.5	29.2	**93.0**	21.8	7.2	16.4	71.2	22.9	7.9	17.0	67.3
	BARF-10	40.4	**17.0**	31.4	58.7	43.6	20.8	30.4	92.7	23.0	7.7	17.0	76.1	24.2	8.3	17.7	70.4
	BARF-100	40.0	16.6	31.1	67.6	**44.4**	21.3	30.8	91.3	23.2	7.7	17.2	79.7	24.0	8.3	17.7	75.2

Table 12.1: Analysis of refined summarization models evaluated on **validation data subsets**, each comprising 7,000 documents for every dataset. All models derived from the mBART25 core model were trained using various BRIO AlignScoreCS and ROUGE fusions approaches. We present F-scores of the quality ROUGE_{RAW} metric for Czech datasets and F-scores of ROUGE for the English Dataset, denoted as $F1$, $F2$, and $F3$. The AS represents the AlignScoreCS factual metric. Fusion denotes the sorting methodology, Core refers to the core model, SMF denotes single metric fusion, and DMF denotes double metric fusion. The highest scores are highlighted in bold.

Single metric fusion (SMF) approach, identical to the methodology of BRIO [Liu et al., 2022], is demonstrated by BRIO-Align and BRIO-Rouge. Compared with the Core Model, BRIO-Align exhibits notably higher factuality scores but lacks quality. Conversely, BRIO-Rouge demonstrates similar tendencies, but in reverse. Both results show the potential of the BRIO training paradigm and prove that the model can coordinate its scores with the metric involved in training. Surprisingly, the BRIO-Rouge model demonstrates lower performance in quality scores for the CNNDM dataset than the Core model. This behavior is unexpected, but we assume it may be influenced by incorporating the XSUM dataset into the entire training dataset, as the reference summaries from XSUM vary significantly in length, which could negatively impact the model's performance by producing shorter summaries for CNNDM.

Double metric fusion (DMF) approach, which combines AlignScoreCS and ROUGE_{RAW} metrics for candidate sorting, yields improvements in both factuality and quality. This

suggests that the sorting methods help the models to align their scores with both metrics simultaneously. Across English datasets, models trained with the DMF method achieve even higher quality scores than the BRIO-Rouge model fine-tuned with the SMF method. However, for Czech data, the quality scores of DMF models fall between the Core and SMF BRIO-Rouge models, as expected. Interestingly, our DMF models reach comparable, and sometimes even better, factual scores than BRIO-Align trained with the SMF method for all datasets. This could be influenced by the larger size of the candidate summaries set (16>8) for the SMF method, suggesting that a smaller size might be preferable for factual scores. This aspect requires further exploration in future experiments.

Variations of factually-balanced candidate sets, as demonstrated by the BARF models, show potential for enhancing the factual scores even when we train models on candidates whose equal factual sets are sorted using different post-sorting methods. Interestingly, the BARF-Align model, which sorts filtered factual candidate sets based on AlignScoreCS, exhibits even better results than our proposed sorting method, as evidenced by the results of the BARF-10 and BARF-100 models. When comparing the effects of different contrastive loss γ_2 weights, the BARF-10 and BARF-100 models show similar results for quality scores. However, the BARF-100 model significantly outperforms the BARF-10 model in terms of facticity when γ_2 is set to 100, indicating that a weight of 100 is more effective. In contrast, the BARF-Rouge model, whose factual candidate sets are sorted according to ROUGE_{RAW}, improves factuality compared to the Core model and BRIO-Rouge model, indicating the efficacy of the factually-balanced method but lacks in specific quality scores, likely due to the alignment of scores to the 4 best-quality faithful and 4 worst-quality unfaithful candidate summaries.

Factually-balanced versus factually-unbalanced candidates by [Dixit et al., 2023]: Both approaches exhibit significant improvements in factuality, surpassing other methods. Remarkably, our approach shows better results for factual scores across all datasets, suggesting that the distribution of equally factual sets guides the model in adjusting scores for more factual summaries. However, the opposite method achieves better performance for quality scores, indicating that it does not sacrifice quality for factuality, probably influenced by the last quality sorting step.

From this performance analysis on validation data, we further select a few best-performing models for the **evaluation on test data** across our summarization datasets. Specifically, we opt for BARF-100, which will be abbreviated as **BARF**, followed by **BARF-Align**, and finally, **AlignSum**, and **BRIO-Rouge**.

12.2.4 Looping Brio

As proposed by [Liu et al., 2022], the performance of BRIO models can be further enhanced through iterative training, as BRIO in a loop. This iterative process involves refining the model with the BRIO paradigm multiple times. While the authors studied this behavior to a limited extent, they trained the core model for an entire epoch using BRIO and generated candidate summaries with the refined model. Then, they trained the refined model again for the entire epoch, resulting in even better results. Although they left further exploration for future work, we investigate this approach by employing our double metric strategy with factually-balanced method by iterative refinement of models within a single epoch.

Since the generation and subsequent factual evaluation of candidate summaries is extremely demanding, we limit our exploration to only two types of **BRIO-in-loop** models, one refined every 10% of an epoch and the other every 50%. Initially, we split our training dataset equally based on individual steps. Subsequently, we iteratively train models using the BRIO training paradigm with our sorting mechanism. Between each step i and $i+1$, we generate 12 candidate summaries per document from set $i+1$ using the refined model$_X$-i from step i, where X represents the step size in percentage. Subsequently, we filter and sort the candidate summaries and proceed with the training of model$_X$-$i+1$. We train models from our core model mBART25 in an iterative manner and configure training settings for each iteration as described in Section 12.2.1. However, we modify the learning rate between steps to align it with the linear learning rate scheduler. Table 12.2 presents the results of each model on validation subsets of our datasets, each subset containing 7,000 documents. We additionally provide the number of samples the model is trained in each iteration after applying the filtering method to get factually-balanced candidate sets.

Model	XSUM (val)				CNNDM (val)				SumeCzech (val)				CNC (val)				Samples
	F1	F2	FL	AS	F1	F2	FL	AS	F1	F2	FL	AS	F1	F2	FL	AS	
Core	39.8	**16.5**	30.8	51.3	43.6	20.5	29.9	89.0	22.4	7.3	16.3	68.1	23.6	7.6	16.9	61.0	-
BARF$_{10}$-1	**39.9**	16.4	**31.0**	70.7	42.1	20.3	30.1	94.3	23.0	7.2	17.3	75.9	25.8	9.9	19.8	71.3	65,425
BARF$_{10}$-2	38.2	15.5	29.7	75.4	38.9	18.6	28.4	**95.3**	24.0	8.3	18.3	81.2	25.4	9.9	19.6	77.5	31,590
BARF$_{10}$-3	37.1	14.8	29.0	72.3	35.7	16.6	26.4	91.2	21.2	7.2	16.4	74.7	22.4	8.5	17.5	73.0	21,853
BARF$_{10}$-4	37.0	14.8	29.0	74.8	35.9	16.4	26.3	91.8	14.4	4.5	11.3	79.5	15.0	5.4	11.9	76.3	15,216
BARF$_{10}$-5	35.7	14.3	28.3	76.9	35.5	16.5	26.3	92.8	11.6	3.5	9.3	79.4	12.7	4.4	10.2	76.5	11,261
BARF$_{10}$-6	36.3	14.6	28.7	76.5	36.7	17.1	27.0	93.9	12.1	3.7	9.7	80.3	13.1	4.6	10.4	76.7	5,720
BARF$_{10}$-7	35.5	14.3	28.1	**79.4**	36.4	17.0	26.8	94.1	10.9	3.3	8.9	80.2	11.9	4.1	9.7	76.5	4,209
BARF$_{10}$-8	35.9	14.5	28.3	79.1	37.0	17.3	27.1	94.4	11.5	3.5	9.2	80.4	12.6	4.4	10.1	76.7	4,137
BARF$_{10}$-9	35.8	14.5	28.4	79.1	36.9	17.3	27.1	94.6	11.9	3.6	9.5	80.1	13.1	4.5	10.4	76.4	3,681
BARF$_{50}$-1	**39.9**	**16.5**	**31.0**	68.4	**44.2**	**21.2**	**30.6**	90.6	**26.6**	**9.8**	**19.8**	75.7	**28.7**	**11.5**	**21.7**	69.9	328,364
BARF$_{50}$-2	39.2	16.1	30.5	74.7	42.2	20.6	30.2	94.6	25.0	8.9	18.8	**81.9**	26.3	10.3	20.2	**78.2**	144,403

Table 12.2: Analysis of BARF-in-loop models evaluated on **validation data subsets**, each comprising 7,000 documents, for every dataset. All models derived from the mBART25 core model were iteratively trained using the BRIO training paradigm with candidates sorted by the factually-balanced method. We present F-scores of the quality ROUGE$_{RAW}$ metric for Czech datasets and F-scores of ROUGE for the English Dataset, denoted as $F1$, $F2$, and $F3$. The AS represents the AlignScoreCS factual metric. Core refers to the core model, while the model BARF$_X$-i indicates iterative refinement of a model in a step i using X% of the training dataset size for the step. The column *"Samples"* denotes the number of samples the models are trained on in the current iteration after the filtering. The highest scores are highlighted in bold.

The data from Table 12.2 demonstrate that applying BRIO training multiple times notably enhances the model's performance, as evidenced by the second iterations of both approaches, BARF$_{10}$-2 and BARF$_{50}$-2. For BARF$_{10}$-2, we observe that just two iterations with a limited number of samples could significantly boost the model's performance in terms of factuality, suggesting that training on the entire dataset may not be necessary. However, further iterations tend to degrade scores, as evident in BARF$_{10}$-i for iterations $i > 2$. In such cases, factual scores initially drop and then rise and oscillate around certain levels, while quality scores generally decrease with more iterations. This phenomenon could be attributed to our factually-balanced method, which filters out samples and retains only those documents whose candidate summary sets contain 4 faithful and 4 unfaithful summaries simultaneously. Perhaps this filtering removes all high-quality

documents, leaving only noisy ones that damage the model's performance. For future experiments, it would be beneficial to explore this approach without the filtering method and instead train on a consistent sample size. For $BARF_{50}$-1, we notice the best performance on quality scores, even comparable to single metric fusion. On the other hand, for BARF50-2, the quality scores decrease slightly for English datasets but still reach comparable results for Czech datasets. However, the factual scores increase significantly, making it one of the best-performing models.

For **evaluation on test data**, we further utilize $BARF_{50}$-2 alongside other models from the previous analysis, as this one offers a good trade-off between quality and facticity. We will abbreviate it as **BARF-Loop**.

12.3 Results on Test Data

Regarding the previous experiments on validation data, we selected the best-performing models. Now, we evaluate them on the original test data of our summarization datasets. We use beam search with a beam size of 5 for inference. Additionally, we train a core model, mT5, with double metric fusion using the factually-balanced method to receive further insights, aiming to extend its applicability beyond mBART25 models. We denote this model as **BARF-mt5**. Furthermore, we present the results achieved by competitive models trained with similar approaches by other authors for performance comparison. We also include **detailed quality results** on both test sets of SumeCzech in Appendix C.2.

12.3.1 Results on English

We present results on the original test sets of English summarization datasets, XSUM and CNNDM, in Table 12.3. Additionally, we compare the performance of our models with the EFactSum models proposed by [Dixit et al., 2023], and the BRIO models introduced by [Liu et al., 2022]. Since both authors fine-tuned two distinct models for each dataset, we aggregate those scores into one row. The EFactSum[4] models are fine-tuned in the same manner as AlignSum, using double metrics fusion of ROUGE and FactCC with the factually-unbalanced method. Meanwhile, the BRIO[5] models are fine-tuned using the single metric fusion approach with the ROUGE metric.

For the **XSUM** dataset, the BRIO model [Liu et al., 2022] exhibits significantly higher quality scores, closely followed by the EfactSum model [Dixit et al., 2023]. Our models generally achieve comparable quality scores to EfactSum, except BARF-mT5, which is significantly behind. However, when it comes to facticity measured by AlignScoreCS, the BARF-mT5 model performs the best, followed by BARF-Align, AlignSum, and BARF-Loop. Despite its poor performance in AlignScoreCS, the EFactSum model excels in the FactCC metric, on which it was fine-tuned. Nonetheless, BARF-mT5 comes close to EFactSum's scores compared to others, while other BARF models show lower similar FactCC scores of each other. In the case of the **CNNDM** dataset, a similar pattern emerges regarding quality scores, with the leading BRIO model followed by slightly lower scores from EFactSum. However, in this case, BARF models produce very similar results.

[4]XSUM: `https://huggingface.co/tanay/efactsum-pegasus-xsum`, CNNDM: `https://huggingface.co/tanay/efactsum-bart-cnndm`

[5]XSUM: `https://huggingface.co/Yale-LILY/brio-xsum-cased`, CNNDM: `https://huggingface.co/Yale-LILY/brio-cnndm-cased`

Model	XSUM						CNNDM					
	Quality Metrics				Factual Metrics		Quality Metrics				Factual Metrics	
	F1	F2	FL	BS	AS	FC	F1	F2	FL	BS	AS	FC
BRIO	**48.5**	**25.1**	**39.9**	84.1	48.2	22.2	**47.3**	**23.2**	**32.1**	85.5	68.3	35.9
EFactSum	41.5	18.7	33.9	82.7	44.5	**31.2**	44.6	21.4	30.3	85.5	73.8	**63.3**
AlignSum	39.9	16.4	31.0	84.3	68.8	21.3	43.8	21.1	30.9	86.3	91.5	39.5
BARF	40.0	16.5	31.0	84.4	67.5	20.7	43.6	20.7	30.2	86.6	91.3	38.0
BARF-mT5	36.1	13.1	27.8	**84.6**	**74.9**	26.0	38.0	18.1	28.1	85.8	**97.1**	57.3
BARF-Align	39.4	16.1	30.5	84.5	72.4	21.8	43.3	20.4	29.9	86.8	92.0	41.0
BARF-Loop	39.7	16.2	30.7	84.4	68.1	21.7	43.5	20.5	30.1	86.5	90.5	36.9
Core-mBART25	39.7	16.5	30.9	84.4	51.7	19.6	43.0	20.0	29.6	86.8	88.9	37.8
Core-mT5	36.1	13.3	28.0	**84.6**	53.0	23.0	41.7	19.2	29.1	**87.4**	96.8	60.5
Falcon	16.0	3.1	11.8	80.1	32.5	29.1	22.8	3.8	13.3	80.5	44.4	13.9
Aya-101	26.9	6.6	20.0	83.9	43.6	22.4	28.7	8.7	18.9	83.5	59.9	20.5

Table 12.3: Results of models on **test data of English summarization datasets**, XSUM and CNNDM. F1, F2, and FL are F-scores of the ROUGE metric; BS is the F1-score by BERTScore, AS denotes AlignScoreCS and FC indicates FactCC metric. BRIO and EFactSum utilize two distinct models for each dataset; hence, the scores are aggregated in these rows. The highest scores are highlighted, and the second-highest is underlined per metric.

In terms of facticity, EFactSum performs the best for FactCC, followed by BARF-mT5 and BARF-Align, which also achieve good scores. BARF-mT5 reaches the top for AlignScoreCS, followed by BARF-Align and AlignSum.

Overall, BRIO models excel in quality scores but significantly fall behind in factual accuracy. EFactSum and AlignSum demonstrate high potential in quality scores without sacrificing facticity. BARF models achieve lower but comparable quality scores but excel in the factuality measured by the AlignScoreCS metric. Interestingly, the Core-mT5 model, which was fine-tuned using only cross-entropy loss, shows high factual scores, indicating that the mT5 model is factually coordinated itself; on the other hand, it lacks quality scores. Notably, the quality scores achieved by our models are most likely influenced by the training on a concatenation of these datasets because of varying reference summary lengths, while BRIO and EFactSum each use two distinct models fine-tuned explicitly on the respective datasets independently. For both LLMs, neither one yields remarkable results for the measured metrics. The decline in quality scores is attributed to the generative ability of their architectures, which compels the models to invent new words. In comparison, the Aya-101 model produces more fluent summaries compared to Falcon, suggesting that Falcon requires further fine-tuning.

12.3.2 Results on Czech

Final performance comparison experiments are conducted on Czech summarization datasets. We assess our models on the original test data from SumeCzech and CNC datasets, as shown in Table 12.4. In this evaluation, we require the models to generate the summaries from texts. Additionally, we report the performance scores of competitive Czech BRIO models introduced by [Halama, 2023]. However, since these models are not publicly available, we rely on the results reported in their study. Specifically, BRIO-R and BRIO-M were trained under the BRIO paradigm, similar to our approach of single metric fusion with either ROUGE or Memes-CS metrics, respectively. Furthermore, we provide the scores of HT2A-CS from our previous work [Krotil, 2022], which was trained on the CNC and SumeCzech datasets using only cross-entropy loss. Moreover, this HT2A-CS model

serves as the core model for BRIO models by [Halama, 2023].

Model	SumeCzech						CNC					
	Quality Metrics				Factual Metrics		Quality Metrics				Factual Metrics	
	F1	F2	FL	BS	AS	M	F1	F2	FL	BS	AS	M
BRIO-Rouge	20.4	5.4	15.0	65.3	57.0	50.2	**22.4**	6.7	**16.5**	65.8	53.0	51.8
BARF	20.3	5.6	15.0	67.2	78.6	49.4	22.1	**6.9**	16.4	67.4	72.2	50.8
BARF-mT5	17.6	4.3	13.2	67.4	**84.3**	43.2	18.7	4.8	13.9	68.2	**79.2**	44.5
AlignSum	20.7	**5.7**	**15.3**	66.7	76.3	49.0	22.3	6.8	**16.5**	67.2	71.5	50.7
BARF-Align	19.9	5.4	14.7	67.5	81.7	49.3	21.4	6.6	15.9	67.8	76.1	50.4
BARF-Loop	20.1	5.2	14.8	66.6	82.7	49.6	21.2	6.2	15.7	67.0	79.1	50.9
Core-mBART25	19.5	5.2	14.2	68.2	66.7	50.8	21.8	6.4	15.7	68.3	57.4	**52.5**
Core-mT5	17.7	4.3	13.1	**69.0**	72.2	46.2	19.5	4.9	14.1	69.5	63.8	47.4
Aya-101	14.3	1.9	10.1	65.2	44.4	42.2	14.3	1.9	9.9	65.7	43.4	43.1
HT2A-CS [Krotil, 2022]	17.9	4.7	13.4	68.7	76.2	48.0	20.2	5.7	14.6	**69.7**	68.5	50.4
BRIO-R [Halama, 2023]	**21.8**	5.4	15.1	-	-	51.3						
BRIO-M [Halama, 2023]	19.3	4.2	12.9	-	-	**62.3**						

Table 12.4: Results of BARF models on **test data of Czech summarization datasets**, SumeCzech and CNC. F1, F2, and FL are F-scores of ROUGE_{RAW} metric; BS is F1-score by BERTScore, AS denotes AlignScoreCS and M indicates Memes-CS metric. Results for BRIO models by [Halama, 2023] are copied from their study. The highest scores are highlighted, and the second-highest are underlined per metric.

For the **CNC dataset**, it is notable that our models achieve superior quality scores compared to models fine-tuned with cross-entropy loss, leading to state-of-the-art results for ROUGE_{RAW}. However, there is a slight decrease in BERTScore scores. Unfortunately, the BRIO models introduced by [Halama, 2023] have not been evaluated on CNC test sets for direct comparison. Nevertheless, we evaluated our BRIO-Rouge model trained similarly to [Halama, 2023] but on our datasets, from which we can infer that we could at least surpass them in F2 and FL scores. Regarding factuality measured by AlignScoreCS, the top-performing models are BARF-mT5 and BARF-Loop, followed by BARF-Align, notably enhancing the factual scores of the core models. However, the Memes-CS scores, which measure alignment between the reference and generated summaries, slightly decrease for all BARF models compared to the core models. Whereas, for the **SumeCzech dataset**, we also notice improved quality scores for F2-scores and FL-scores, surpassing other models and resulting in the state-of-the-art results for ROUGE_{RAW} on SumeCzech. However, in terms of the F1-score, we are outperformed by the BRIO-R model. The BERTScore scores exhibit a similar pattern as observed in the CNC dataset. Once again, the top-performing models for AlignScoreCS are BARF-mT5 and BARF-Loop, followed by BARF-Align. As for Memes-CS, all models yield similar and insignificant scores except for the BRIO-M model, demonstrating its power for this metric as it was explicitly fine-tuned using single metric fusion.

In summary, our models trained using double metric fusion exhibit superior scores for both factuality and quality, which does not hold for BRIO-M, which solely relies on single metric fusion. While BRIO-M shows improved factual scores for Memes-CS, its ROUGE_{RAW} quality scores are lower. Furthermore, we surpass some of the quality scores of BRIO-R, which is trained using single metric fusion with ROUGE, suggesting that our approach of factually-balanced sorted candidates enhances models' score coordination. Moreover, our BRIO-Rouge that is considered similar to BRIO-R [Halama, 2023] also shows good quality scores but lacks in factuality, from which we can infer a similar behavior for BRIO-R.

AlignSum, trained using a factually-unbalanced method, demonstrates considerable quality scores for both datasets but falls behind in factuality scores. Conversely, BARF-Loop appears to be the best balance between quality and factuality. Notably, the HT2A-CS model exhibits good factual scores even without additional refinement on this metric. On the other hand, Core-mBART25, trained on the same Czech datasets as HT2A-CS but with the addition of other language datasets, displays lower factual scores for factuality but improves in quality, likely due to a cross-lingual transfer from multilingual data. Once more, we observe a similar pattern for the LLM Aya in English datasets, with shallow values for quality and factuality coming from the generative behavior.

12.4 Extractiveness

When dealing with enhancing the faithfulness of summarization models, a new challenge emerges as identified by [Ladhak et al., 2022] and addressed by [Dixit et al., 2023]. This problem arises because as we increase the factual accuracy of generated summaries, we may inadvertently increase their level of extractiveness. We address this issue by computing the extractive fragment, coverage, proposed by [Grusky et al., 2018], between the generated summaries and the source texts, followed by a comparison with coverage of the reference summaries and summaries produced by core models. The table 12.5 below shows the coverage values for our models on the test data of each dataset. As evident from the values, the reference summaries exhibit higher abstractiveness regarding coverage than the generated summaries. Consequently, all models tend to incorporate more extractive fragments in their summaries than in the reference summaries. This tendency is likely influenced by the models' architectures and their overall training, proved by the consistently higher extractive scores for the summaries produced by core-mT5 compared to core-mBART25. Furthermore, this observation could explain why the core-mT5 model achieves higher factual scores, as demonstrated in the experiments. Additionally, when we compare the factually refined models with their respective core models, we notice that the coverage values generally decrease rather than increase. This suggests that the enhanced factual generation of our models does not affect their ability to produce more extractive summaries.

Summary by	XSUM	CNNDM	SumeCzech	CNC
Reference	0.19	0.54	0.12	0.09
core-mBART25	0.31	0.83	0.43	0.34
core-mT5	0.39	0.93	0.57	0.48
BARF	0.31	0.82	0.40	0.34
BARF-mT5	0.38	0.87	0.57	0.51
BARF-align	0.33	0.84	0.44	0.38
BARF-Loop	0.31	0.81	0.39	0.34
AlignSum	0.32	0.81	0.38	0.33

Table 12.5: Coverage statistics [Grusky et al., 2018] computed for summaries generated by our models on test data of each dataset. The row "Reference" stands for reference summary. The coverage value ranges between 0 and 1, and the lower the value, the higher the abstractive summary level.

12.5 Human Evaluation

In the final experiments, we assess summaries produced by our models through human evaluation. This evaluation not only aids in determining the factual correctness of the

generated summaries but also provides insights into the performance of AlignScoreCS on summarization data. Through this process, we also examine the generated summaries to infer the behavior of each model. Therefore, we generate summaries per model for 50 documents randomly selected from both test splits of the SumeCzech and CNC datasets, each providing 25 samples. **Importantly**, we emphasize that the selected data portion is a small subset and could be biased. For more accurate details, we suggest evaluating more samples. Subsequently, we annotate the summaries using Doccano software[6] with four labels:

1. **Correct**: The summary is factually accurate.
2. **Sufficient**: The summary is factually accurate but misses a minor important detail or contains grammatical errors.
3. **Missing**: The summary is faithful but misses relevant information.
4. **Incorrect**: The summary is not faithful.

Table 12.6 presents annotation results (in %) per label for each model, where 2% corresponds to 1 summary. The "Accuracy" column represents the total sum of all faithful classes. The values indicate that the accuracy of factually refined models has improved compared to the core models, suggesting that the BRIO paradigm with ROUGE and AlignScoreCS fusion indeed enhances factuality. The highest factual accuracy is achieved by BARF-Loop, significantly surpassing others which yield similar results. However, in terms of the "Correct" label, BARF and BARF-Align perform the best. BARF-Loop scores higher for the "Sufficient" class, indicating that the model sometimes misses minor details. Grammatical errors were more frequent for mT5 models, which also show higher numbers for the "Missing" label caused by often extractive rewriting of initial parts of texts.

Model	Accuracy	Incorrect	Correct	Sufficient	Missing
core-mBART25	52	48	32	16	**4**
core-mT5	56	44	36	**8**	12
AlignSum	72	28	38	24	10
BARF	72	28	**42**	20	10
BARF-mT5	70	30	22	20	28
BARF-align	72	28	**42**	20	10
BARF-Loop	**80**	**20**	34	36	10

Table 12.6: The human evaluation results of 50 Czech summaries generated by our models, with 25 from each Czech dataset, are presented. The values are expressed in percentages, where Accuracy represents the total sum of values from faithful labels. The best scores are highlighted.

In terms of **summaries examination**, numerous instances occur where models generate the entire summary very accurately but include an inappropriate single word, typically about time or place, that is completely out of context, resulting in unfaithful summaries. Another common mistake made by the models is the hallucination of first names of persons, likely caused during training by appearances of texts containing only last names, but their reference summaries include full names, forcing the model to make up new names. Additionally, when articles share football statistics and results, all models usually make errors in match scores, resulting in factual incorrectness.

[6] `https://github.com/doccano/doccano`

Focusing on **particular models**, we observe that mBART-based models sometimes generate similar summaries, varying in a few words, suggesting similar score coordination between models. Moreover, BARF-Align occasionally exhibits extractive behavior but effectively avoids references to web sources, which is unseen in core models. BARF-Loop tends to produce concise and precise summaries, consisting of a few sentences and capturing information from the entire text, unlike mT5 models, which mainly focus on the initial parts. The same tendency is also observed in the BARF model. Furthermore, BARF-Loop and AlignSum frequently omit unnecessary names or numbers, a common error among other models, which negatively impacts factuality. In contrast, core models suffer from generating new sequences unrelated to the source text, which is prone to harm factuality, with name entity swapping more frequently in mT5-based models. We provide a few examples of summaries in Appendix C.3.

With the completion of the human evaluation, we can now **compare the AlignScoreCS metric** (ASCS) with annotated summarization data comprising 350 labeled samples (50 per model). Although the annotation is tailored for classification, we can interpret each label continuously to align with the metric output. Hence, we assign the "Incorrect" label to 0.0, "Correct" to 1.0, "Sufficient" to 0.85, and "Missing" to 0.65. We assume that "Sufficient" will correlate with the metric, but we expect less correlation with "Missing" since the factuality is accurate but lacks in information retrieval, on which the metric is trained to a limited extent. Nevertheless, the score still exceeds 0.5 and, thus, is expected to correlate when the scores are higher at least weakly. The following Table 12.7 presents statistical correlation values of ASCS metric with human-annotated data. The substantial degrees of Pearson and Spearman, alongside a moderate Kendall coefficient, indicate a moderately strong correlation between annotated data and scores computed by AllignScoreCS, suggesting that AlignScoreCS is a new, powerful metric for evaluating the factuality of Czech summarization.

	Pearson	Kendall	Spearman
AlignScoreCS	0.622	0.463	0.593

Table 12.7: Correlation statistics of AlignScoreCS with human-annotated data: Pearson, Kendall, and Spearman correlations computed on 350 labeled summarization document-summary pairs. Labels are mapped to continuous space, "Incorrect" label to 0.0, "Correct" to 1.0, "Sufficient" to 0.85, and "Missing" to 0.65.

Chapter 13
Discussion of Part II

In this part, we presented a BRIO training paradigm incorporating a combination of two metrics. Initially, we employed ROUGE$_{RAW}$ to assess the quality of generated summaries, and for evaluating factuality, we integrated our newly proposed metric, AlignScoreCS, as discussed in Part I. To combine these metrics, we introduced a factually-balanced method that organizes candidate summaries into equally factual sets, sorted according to quality for faithful summaries and facticity for unfaithful ones, to coordinate the models' scores within the BRIO paradigm. This approach, which we named BRIO paradigm with double metric fusion (BARF), was applied using two Czech news summarization datasets, CNC and SumeCzech, and two English news summarization datasets, XSUM and CNNDM, for training. Following that, we trained numerous models derived from mBART25 and mT5 using contrastive loss, employing a range of approaches for candidate sorting. These approaches included our proposed method, its modifications, and other methods suggested by other researchers.

From experiments conducted on validation data, we analyzed trained models with different settings and discussed the utilized methods based on their results. Our findings revealed that models trained using BRIO with double metric fusion outperformed others in both metrics across all datasets. This suggests that these models could align their scores according to factuality and quality, unlike models with single metric fusion, which only coordinated their scores for the specific metric they were refined to. Experiments on the factually-balanced method and its variations of post-sorting strategies highlighted the importance of equal distribution of factual sets. The following experiments aimed at extending the ability of the BRIO paradigm focused on applying BARF refinement to a single model multiple times, resulting in even better performance than single training. However, more iterations led to a drop in quality, which was influenced by the filtering phase of our method. Hence, we determined that only a few iterations, incorporating significantly smaller subsets than the entire training dataset, helped models achieve comparable results to those trained on the entire dataset.

Results on test data demonstrated comparable and often even better results. For English data, we could not surpass the quality scores of the baseline BRIO models [Liu et al., 2022]; however, we achieved comparable quality results and, more importantly, improved factuality. The decrease in quality performance on CNNDM can be attributed mainly to the filtering step aimed at obtaining factually equal sets, given that our core models already produced more factual summaries on that dataset. Another baseline model, EFactSum [Dixit et al., 2023], which utilizes a factually-unbalanced method with ROUGE and FactCC metrics, also exhibited slightly better quality scores than our models, likely due to the last quality sorting step. The results also indicate that our models and EFactSum

models tend to adjust their scores according to the factual metric they were trained on without improving the other one, suggesting that the factual metrics correlate poorly. For Czech data, we updated state-of-the-art results of ROUGE_{RAW}, F2-score, and FL-score, and significantly enhanced factual scores simultaneously. As a result, the Czech BRIO model [Halama, 2023], trained with single metric fusion (ROUGE), still maintains the best result for the F1-score. Interestingly, we were able to surpass this model with models fine-tuned using double metric fusion, indicating stronger factual performance for better quality models. For the LLMs models, Aya and Falcon, our attempts at fine-tuning them on our summarization datasets were unsuccessful, leading to lower performance on both metrics. Therefore, we recommend future detailed study to address this issue. Additionally, we examined extractive levels of generated summaries through coverage measurements [Grusky et al., 2018] and revealed that our BARF models are not adversely affected by the fact that improving factuality decreases abstractiveness, which was evident from the decrease in extractive fragments rather than an increase in comparison with core models.

Human evaluation revealed that the highest rank in factual accuracy belongs to the BARF-Loop model. Other factually refined models produced slightly lower but comparable results. Nonetheless, all factually refined models exhibited improved factual accuracy compared to their core models. Upon examining the summaries, we discovered that models trained with double metric fusion, unlike the core models, tend to exclude irrelevant information about the source web page. BARF and BARF-Align exhibit similar behavior and occasionally produce more extractive fragments. Additionally, we observed that BARF-Loop generates more precise summaries that are comprised of concise sentences and avoids the use of determining words to prevent factual inaccuracies. However, despite reducing the hallucination of generated summaries, refined models still occasionally produce invented names or numbers, leading to factual inconsistency. This could be caused by either the structure of summarization datasets requiring further filtering or AlignScoreCS, since despite performing well in tests on single sentences incorporating these issues such as number or name swapping, the metric struggles to identify it in longer texts when all other information is factually consistent. Comparing our results to the human evaluation conducted in [Halama, 2023], we can infer that our new AlignScoreCS factual metric enhanced the models, which was the main call emerging from their study, a goal we successfully fulfilled here.

Chapter 14
Conclusion

This thesis delved into two key challenges within natural language processing, initially concentrating on assessing factuality, followed by shifting interest in summarization while focusing on both English and Czech data.

Our research aimed to enhance the factuality of language models' generated content, as these models tend to produce inaccurate information. Unfortunately, there are few metrics capable of accurately assessing the factuality of Czech summaries. In Part I, we addressed this gap by developing our factual metric, which demonstrated its multitask potential and achieved noteworthy results on both English and Czech datasets, emerging as the preferred choice for evaluating Czech summaries. In Part II, we attempted to compel a summarization model to generate factual summaries of high quality by incorporating the factual metric and a quality metric into its training objective. This approach enabled the model to better align its scores according to these metrics. We introduced a factually-balanced sorting strategy for candidate ranking during model training, which proved effective in simultaneously improving the factuality and quality of generated summaries. Our evaluation encompassed both English and Czech datasets, which showed considerable results. Notably, on the SumeCzech summarization dataset, we updated several state-of-the-art quality results measured by ROUGE_{RAW} while ensuring the factuality of the generated summaries. Human evaluation further confirmed that our models indeed improved factual accuracy and that our factual metric correlated well with human judgment.

We make our factual metric and summarization models publicly available at these websites[1]. Moreover, translated datasets and benchmarks will be gradually updated there as well.

[1]
AlignScoreCS: `https://huggingface.co/krotima1/AlignScoreCS`
Datasets: `https://huggingface.co/ctu-aic`
BARF: `https://huggingface.co/krotima1/BARF`
BARF-Align: `https://huggingface.co/krotima1/BARF-Align`
BARF-Loop: `https://huggingface.co/krotima1/BARF-Loop`
BARF-mT5: `https://huggingface.co/krotima1/BARF-mT5`
AlignSum: `https://huggingface.co/krotima1/AlignSum`

During the continuous development of this thesis, I took advantage of ChatGPT[2]*, GitHub Copilot*[3]*, Grammarly*[4] *and DeepL*[5] *to facilitate error debugging, accelerate programming, and improve my English language. However, it is important to highlight that, in any case, I did not utilize them as generators or idea inventors. Rather, I employed them as tools to do the manual hard work and, hence, speed up my progress.*

[2] https://chatgpt.com/
[3] https://github.com/features/copilot
[4] https://www.grammarly.com/
[5] https://www.deepl.com/translator

Bibliography

[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. *CoRR*, abs/1508.05326.

[Cer et al., 2017] Cer, D. M., Diab, M. T., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation. *CoRR*, abs/1708.00055.

[Chern et al., 2023] Chern, I.-C., Wang, Z., Das, S., Sharma, B., Liu, P., and Neubig, G. (2023). Improving factuality of abstractive summarization via contrastive reward learning.

[Communication et al., 2023] Communication, S., Barrault, L., Chung, Y.-A., Meglioli, M. C., Dale, D., Dong, N., Duquenne, P.-A., Elsahar, H., Gong, H., Heffernan, K., Hoffman, J., Klaiber, C., Li, P., Licht, D., Maillard, J., Rakotoarison, A., Sadagopan, K. R., Wenzek, G., Ye, E., Akula, B., Chen, P.-J., Hachem, N. E., Ellis, B., Gonzalez, G. M., Haaheim, J., Hansanti, P., Howes, R., Huang, B., Hwang, M.-J., Inaguma, H., Jain, S., Kalbassi, E., Kallet, A., Kulikov, I., Lam, J., Li, D., Ma, X., Mavlyutov, R., Peloquin, B., Ramadan, M., Ramakrishnan, A., Sun, A., Tran, K., Tran, T., Tufanov, I., Vogeti, V., Wood, C., Yang, Y., Yu, B., Andrews, P., Balioglu, C., Costa-jussà, M. R., Celebi, O., Elbayad, M., Gao, C., Guzmán, F., Kao, J., Lee, A., Mourachko, A., Pino, J., Popuri, S., Ropers, C., Saleem, S., Schwenk, H., Tomasello, P., Wang, C., Wang, J., and Wang, S. (2023). Seamlessm4t: Massively multilingual & multimodal machine translation.

[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., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

[Crawshaw, 2020] Crawshaw, M. (2020). Multi-task learning with deep neural networks: A survey.

[Csernai,] Csernai, K. First quora dataset release: Question pairs.

[Demszky et al., 2018] Demszky, D., Guu, K., and Liang, P. (2018). Transforming question answering datasets into natural language inference datasets. *CoRR*, abs/1809.02922.

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

[Devlin et al., 2018] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[Dixit et al., 2023] Dixit, T., Wang, F., and Chen, M. (2023). Improving factuality of abstractive summarization without sacrificing summary quality.

[Drchal et al., 2023] Drchal, J., Ullrich, H., Mlynář, T., and Moravec, V. (2023). Pipeline and dataset generation for automated fact-checking in almost any language.

[Drchal et al., 2022] Drchal, J., Ullrich, H., Rýpar, M., Vincourová, H., and Moravec, V. (2022). Csfever and ctkfacts: Czech datasets for fact verification. *CoRR*, abs/2201.11115.

[Fabbri et al., 2021a] Fabbri, A., Han, S., Li, H., Li, H., Ghazvininejad, M., Joty, S., Radev, D., and Mehdad, Y. (2021a). Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 704–717, Online. Association for Computational Linguistics.

[Fabbri et al., 2021b] Fabbri, A. R., Wu, C., Liu, W., and Xiong, C. (2021b). Qafacteval: Improved qa-based factual consistency evaluation for summarization. *CoRR*, abs/2112.08542.

[Fan et al., 2020] Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., and Joulin, A. (2020). Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

[Fan et al., 2018] Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical neural story generation.

[Goyal and Durrett, 2020] Goyal, T. and Durrett, G. (2020). Evaluating factuality in generation with dependency-level entailment.

[Grusky et al., 2018] Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

[Halama, 2023] Halama, V. (2023). Improving facticity of summarization methods. *M. thesis, Czech Technical University in Prague, Computational and informational center.*

[Honovich et al., 2022] Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., Scialom, T., Szpektor, I., Hassidim, A., and Matias, Y. (2022). TRUE: Re-evaluating factual consistency evaluation. In Feng, S., Wan, H., Yuan, C., and Yu, H., editors, *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

[Honovich et al., 2021] Honovich, O., Choshen, L., Aharoni, R., Neeman, E., Szpektor, I., and Abend, O. (2021). Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *CoRR*, abs/2104.08202.

[Hopkins and May, 2011] Hopkins, M. and May, J. (2011). Tuning as ranking. In Barzilay, R. and Johnson, M., editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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

[Hájek and Horák, 2024] Hájek, A. and Horák, A. (2024). Czegpt-2–training new model for czech generative text processing evaluated with the summarization task. *IEEE Access*, 12:34570–34581.

[Ji et al., 2022] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2022). Survey of hallucination in natural language generation. *CoRR*, abs/2202.03629.

[Kalajdzievski, 2023] Kalajdzievski, D. (2023). A rank stabilization scaling factor for fine-tuning with lora.

[Keskar et al., 2019] Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. (2019). Ctrl: A conditional transformer language model for controllable generation.

[Koupaee and Wang, 2018] Koupaee, M. and Wang, W. Y. (2018). Wikihow: A large scale text summarization dataset. *CoRR*, abs/1810.09305.

[Krotil, 2022] Krotil, M. (2022). Text summarization methods in czech. *B. thesis, Czech Technical University in Prague, Computational and informational center.*

[Kryscinski et al., 2019] Kryscinski, W., McCann, B., Xiong, C., and Socher, R. (2019). Evaluating the factual consistency of abstractive text summarization. *CoRR*, abs/1910.12840.

[Kutylowski, 2017] Kutylowski, J. (2017). Deepl translator.

[Laban et al., 2021] Laban, P., Schnabel, T., Bennett, P. N., and Hearst, M. A. (2021). Summac: Re-visiting nli-based models for inconsistency detection in summarization. *CoRR*, abs/2111.09525.

[Ladhak et al., 2022] Ladhak, F., Durmus, E., He, H., Cardie, C., and McKeown, K. (2022). Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization.

[Lai et al., 2017] Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. H. (2017). RACE: large-scale reading comprehension dataset from examinations. *CoRR*, abs/1704.04683.

[Lewis et al., 2019] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

[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., 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. *CoRR*, abs/2001.08210.

[Liu et al., 2022] Liu, Y., Liu, P., Radev, D., and Neubig, G. (2022). Brio: Bringing order to abstractive summarization.

[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.

[Marelli et al., 2014] Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

[Merity et al., 2016] Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016). Pointer sentinel mixture models. *CoRR*, abs/1609.07843.

[Nallapati et al., 2016] Nallapati, R., Zhou, B., dos santos, C. N., Gulcehre, C., and Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond.

[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.

[Nguyen et al., 2016] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

[Nie et al., 2019a] Nie, Y., Chen, H., and Bansal, M. (2019a). Combining fact extraction and verification with neural semantic matching networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6859–6866.

[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.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[Penedo et al., 2023] Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. (2023). The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

[Raffel et al., 2023] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2023). Exploring the limits of transfer learning with a unified text-to-text transformer.

[Rajpurkar et al., 2018a] Rajpurkar, P., Jia, R., and Liang, P. (2018a). Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

[Rajpurkar et al., 2018b] Rajpurkar, P., Jia, R., and Liang, P. (2018b). Know what you don't know: Unanswerable questions for squad.

[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.

[Scialom et al., 2020] Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., and Staiano, J. (2020). MLSUM: The multilingual summarization corpus. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

[Sellam et al., 2020] Sellam, T., Das, D., and Parikh, A. P. (2020). Bleurt: Learning robust metrics for text generation.

[Sido et al., 2021] Sido, J., Seják, M., Prazák, O., Konopík, M., and Moravec, V. (2021). Czech news dataset for semantic textual similarity. *CoRR*, abs/2108.08708.

[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 Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

[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*, 57(4):1571–1605.

[Vaswani et al., 2023] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

[Vijayakumar et al., 2018] Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., and Batra, D. (2018). Diverse beam search: Decoding diverse solutions from neural sequence models.

[Víta, 2020] Víta, M. (2020). Natural language inference in czech and related tasks. *muni*.

[Wang et al., 2021] Wang, D., Chen, J., Wu, X., Zhou, H., and Li, L. (2021). Cnewsum: A large-scale chinese news summarization dataset with human-annotated adequacy and deducibility level.

[Williams et al., 2017] Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426.

[Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

[Xue et al., 2021] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mt5: A massively multilingual pre-trained text-to-text transformer.

[Yin et al., 2021] Yin, W., Radev, D., and Xiong, C. (2021). DocNLI: A large-scale dataset for document-level natural language inference. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

[Yuan et al., 2021] Yuan, W., Neubig, G., and Liu, P. (2021). Bartscore: Evaluating generated text as text generation.

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

[Zhang et al., 2020b] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020b). Bertscore: Evaluating text generation with bert.

[Zhang et al., 2019] Zhang, Y., Baldridge, J., and He, L. (2019). PAWS: paraphrase adversaries from word scrambling. *CoRR*, abs/1904.01130.

[Zhang et al., 2023] Zhang, Z., Yu, W., Yu, M., Guo, Z., and Jiang, M. (2023). A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods.

[Zhong et al., 2022] Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H., and Han, J. (2022). Towards a unified multi-dimensional evaluator for text generation.

[Üstün et al., 2024] Üstün, A., Aryabumi, V., Yong, Z.-X., Ko, W.-Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.-L., Kayid, A., Vargus, F., Blunsom, P., Longpre, S., Muennighoff, N., Fadaee, M., Kreutzer, J., and Hooker, S. (2024). Aya model: An instruction finetuned open-access multilingual language model.

[Šimon Zvára, 2022] Šimon Zvára (2022). Assessing facticity in abstractive summarization methods. *B. thesis, Czech Technical University in Prague, Computational and informational center.*

Appendices

Appendix A
Acronyms

NLP Natural Language Processing

NLG Natural Language Generation

NLI Natural Language Inference

NLU Natural Language Understanding

QA Question Answering

MTL Multi-Task Learning

SOTA State-of-the-art

3-way Ternary Classification (3 classes)

2-way Binary Classification (2 classes)

reg Regression (2 classes)

IR Information Retrieval

STS Semantic Textual Similarity

SM Semantic Similarity

MISC Miscellaneous

FV Fact Verification

BARF BRIO with AlignScoreCS and ROUGE$_{RAW}$ Fusion

BRIO Bringing Order to Abstractive Summarization

mBART Multilingual Bidirectional Auto-Regressive Transformer

mBERT Multilingual Bidirectional Encoder Representations from Transformers

RoBERTa A Robustly Optimized BERT Pretraining Approach

DMF Double Metric Fusion

SMF Single Metric Fusion

LLM Large Language Model

CNC Czech News Center

QLoRA Quantized Low-Rank Adaptation

ROUGE Recall-Oriented Understudy for Gisting Evaluation

Appendix B
Facticity Details

In the following sections, we provide details on task-specific datasets B.1, benchmark datasets B.1 and examples of AlignScoreCS performance B.2 on them.

B.1 Datasets Details

Further details on task-specific training datasets and benchmarks utilized in this work are provided in these sections.

Training Datasets

We shortly describe each training task-specific dataset and also include its division and classification class.

1. **SNLI** [Bowman et al., 2015] (NLI, 3-way): The Stanford Natural Language Inference (SNLI) corpus comprises 570k sentence pairs derived from image captions. Human annotators have labeled these pairs as Aligned, Contradiction, or Neutral.

2. **MultiNLI** [Williams et al., 2017] (NLI, 3-way): Multi-Genre Natural Language Inference dataset contains 433K sentence pairs spanning ten diverse genres (government, conversations, fiction..). Human annotators classified these pairs into Contradiction, Aligned, or Neutral categories.

3. **Adversarial NLI** [Nie et al., 2019b] (NLI, 3-way): This corpus addresses models' weaknesses through an iterative human-and-model-in-the-loop procedure. It includes 163K premise-hypothesis-reason triplets labeled into Contradiction, Aligned, or Neutral classes.

4. **DocNLI** [Yin et al., 2021] (NLI, 2-way): Constructed from various NLP problems (Summarization, NLI, and QA), the document-level NLI corpus contains 942K premise-hypothesis pairs of varying lengths (longer). These pairs are labeled into Contradiction or Aligned categories.

5. **NLI-FEVER** [Nie et al., 2019a] (Fact verification, 3-way): Built on the FEVER shared task, utilizing three phases: building, breaking, and fixing to generate additional adversarial examples (1k). All together, it contains 208K context-claim pairs labeled into three classes.

6. **VitaminC** [Schuster et al., 2021] (Fact verification, 3-way): a benchmark containing 400K evidence-claim pairs aimed at improving robustness in fact verification by a corruption of small parts in evidence. Samples are labeled into 3 classes.

7. **QQP** [Csernai,][1] (Paraphrase, 2-way), Quora Question Pairs, a collection of question-question pairs (400K) labeled as Aligned or Contradict indicating whether the two questions are paraphrased or are not related.

8. **PAWS** [Zhang et al., 2019] (Paraphrase, 2-way), Paraphrase Adversaries from Word Scrambling, a collection of paraphrase pairs (110K) with high lexical overlap focusing on challenging word order (named entity, adjective, word swaps and replacements). Labeled PAWS - contains human-labeled pairs sourced from Wikipedia (50k), and unlabeled PAWS - algorithmically created (656k), may contain noise.

9. **SICK** [Marelli et al., 2014] (STS, reg), Sentences Involving Compositional Knowledge, a collection of sentence pairs (10k) including examples of lexical, syntactic and semantic levels. Samples are annotated in scale of relatedness and entailment.

10. **STS Benchmark** [Cer et al., 2017] (STS, reg), a selection of datasets from STS shared task containing human-annotated (5 levels) sentence pairs (6k) from image captions, news headlines and user forums.

11. **Free N1 STS** [Sido et al., 2021] (STS, reg), a Czech collection of sentence pairs (140k) human-annotated for semantic similarity. But we use its context-free version (annotated without context) filtered using close neighborhood of 1 score difference (20k).

12. **SQuADv2** [Rajpurkar et al., 2018b] (QA, 2-way), Stanford Question Answering Dataset, a combination of SQuAD (human created questions and answers) with adversarial unanswerable human-written questions (50k). In total, it comprises context-question-answer triplets (150k).

13. **RACE** [Lai et al., 2017] (QA, 2-way), ReAding Comprehension Dataset From Examinations, a collection of text-question-answers triplets (100k) from English exams for reasoning and understanding. Answers include more options, resulting in 350k pairs.

14. **Ms MARCO** [Nguyen et al., 2016] (Information Retrieval, 2-way), Microsoft MAchine Reading Comprehension, a large collection of annotated data (millions) capable of being utilized for QA or Information Retrieval.

15. **WikiHow** [Koupaee and Wang, 2018] (Summarization, 2-way), consists of diverse articles from WikiHow knowledge base providing various topics in different writing styles. It is designed for summarization.

Benchmarks

We concisely detail individual datasets included in TRUE [Honovich et al., 2022] and SummaC [Laban et al., 2021] benchmarks. We also involve its division. Both benchmarks converted each dataset into binary classification problem - Aligned, Contradict.

1. **BEGIN** (TRUE, dialogue): This dataset focuses on ensuring consistency with grounding knowledge in dialogue systems. It includes annotated sentences extracted from outputs generated by models trained on Wizard of Wikipedia (WoW).

[1] `https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs`

2. **Q^2** (TRUE, dialogue): This dataset comprises annotated dialogue sequences generated by models trained on WoW.
3. **DialFact** (TRUE, dialogue): Designed for fact verification within dialogue systems, this dataset compares human-annotated claims with evidences gathered from Wikipedia.
4. **FEVER** (TRUE, Fact verification): Containing human-annotated pairs of evidence and claims sourced from Wikipedia, using NLI-FEVER.
5. **VitaminC** (TRUE, Information Retrieval): Combining evidences from Wikipedia with human-labeled claims. VitaminC includes additional data by revising FEVER.
6. **PAWS** (TRUE, Paraphrase): The test set of the PAWS dataset.
7. **QAGS** (TRUE, Summarization): This dataset consists of human-annotated model generated summaries for CNNDM and XSUM.
8. **MNBM** (TRUE, Summarization): Comprising human-labeled summaries for XSUM generated by summarization models.
9. **FRANK** (BOTH, Summarization): This dataset is built by linguistically grounded typology of factual errors made by summarization models covering both CNNDM and XSUM datasets.
10. **SummEval** (BOTH, Summarization): Featuring human-labeled summaries generated by both extractive and abstractive models, SummEval offers comprehensive evaluation data.
11. **CogenSumm** (SummaC, Summarization): A subset of CNNDM, it includes summaries with intentionally corrupted sentences for unfaithful summarization.
12. **FactCC** (SummaC, Summarization): Focused on factual consistency, this dataset contains human-annotated summaries from CNNDM.
13. **Polytope** (SummaC, Summarization): This dataset presents an extensive typology of factual errors produced by summarization models.
14. **XSumFaith** (SummaC, Summarization): Containing human-annotated abstractive summaries generated by models trained on XSUM.

B.2 Examples of AlignScoreCS

The following Table B.1 shows example from each dataset. Instead of evaluating task-specific head on the task-specific data, we evaluate only AlignScoreCS consistency scores for deeper insight into the final functionality.

Task-specific datasets			
Task	Dataset	Label	AlignScoreCS
3-way	snli	Aligned	0.9588
A dozen Asian men are sitting and standing in a group, they are dressed casually and are looking at something on the ground.		a group of people stand in a group looking at something	

3-way	mnli	Contradict	0.0001
But it's a small amount of money that's stolen from a lot of people.		A huge amount of money was stolen from every US citizen.	
3-way	anli	Neutral	0.0013
Youth in Guatemala are the largest segment of the nation's population. Youth includes individuals between the ages of 15 and 24 Over half of the population is under 19 years old in 2011, the highest proportion of young people of any country in Latin America. The health, education, and work opportunities for young people differ by ethnicity ("ladino" or indigenous) and social class.		Youth in Guatemala are the largest segment of the nation's population, helping the country.	
2-way	doc-nli	Contradict	0.497
Turkish police hold 46 people as part of investigation into match-fixing in European football. - The first ever match between the two teams is played at a neutral venue (Stade de France in Paris) and ends with a 2-0 win for France. Former Genclerbirligi goalkeeper accused of betting $ 40,000 against his own team. Match also under suspicion from German prosecutors investigating corruption.		Turkish police hold 46 people as part of investigation into match-fixing in European football. Former Turkey international players Arif Erdem and Fatih Akyel among those detained. Former Genclerbirligi goalkeeper accused of betting $ 40,000 against his own team. Match also under suspicion from German prosecutors investigating corruption.	
3-way	nli-fever	Aligned	0.9939
RMS Titanic. Titanic is the second largest ocean liner wreck in the world, only beaten by her sister, the largest ever sunk.		RMS Titanic was a boat.	
3-way	vitaminc	Aligned	0.9986
The music video for " I Hate U, I Love U " premiered on March 9, 2016.		The music video for " I Hate U, I Love U " was released after March 8, 2016.	
2-way	qqp	Contradict	0.0002
What are some things new employees should know going into their first day at Verizon?		What are some things new employees should know going into their first day at Deluxe?	
reg	sick	0.7	0.6243
A woman and three men are posing for a photo		A woman is posing for three men for a photo	
reg	stsb	0.6	0.0001
Bombs in southern Thailand kill 5, wound 50		Bombs in Thailand kill 14, wound 340	
2-way	squadv2	Aligned	0.9992
Iran has the second largest proved gas reserves in the world after Russia, with 33.6 trillion cubic metres, and third largest natural gas production in the world after Indonesia, and Russia. It also ranks fourth in oil reserves with an estimated 153,600,000,000 barrels. It is OPEC's 2nd largest oil exporter and is an energy superpower.[CONTINUE].		Iran has 33.6 trillion cubic metres of natural gas reserves.	
2-way	race	Contradict	0.5497

Dear Sir, I read your ad in the newspaper yesterday. I'm writing to tell you I'd like to work as a volunteer for the 2008 Olympics. My name is Fanny. I am 16. I study in No.46 Middle School from September, 2002. I love sports, and table tennis is my favorite. I can speak English very well and I am quite healthy. If I am lucky enough to become a volunteer for the 2008 Olympics, please call me as soon as possible. My telephone number is 028-84661314. I will try my best to do the job well. Best wishes! yours Fanny		What's Fanny's telephone number? 028-88661314.	
2-way	msmarco	Contradict	0.7485
1 For all travel expenses incurred on or after January 1, 2013, the mileage reimbursement rate is 56.5 cents per mile. 2 For all travel expenses incurred on or after April 17, 2012, the mileage reimbursement rate is 55.5 cents per mile.		price per mile reimbursement	
2-way	wikihow	Aligned	0.9997
;, The batter may be pasty at this point. After incorporated, add in the chocolate chips and vanilla. The cake may begin to rise above the edge of the mug, but will shrink down once cooled. Let it stand for a few minutes to cool (though this might prove challenging). For a more fudgy cake, omit the egg. Pudding mix may be substituted for unsweetened cocoa. Baking times may vary slightly due to variations in microwave ovens.		The cake may begin to rise above the edge of the mug, but will shrink down once cooled.	
reg	free-train-N1	0.833	0.0023
"Přijde mi absurdní, aby mě mistrovali lidé, kteří se sami podíleli na tom, že měla sociální demokracie v posledních volbách slabé výsledky," řekl Dienstbier.		"Přijde mi absurdní, aby mě mistrovali lidé, kteří se sami podíleli na tom, že měla sociální demokracie v posledních volbách slabé výsledky," řekl Dienstbier ČTK.	

ASCS	X-lingual Examples (deepl)	
0.890	Děti se usmívají a mávají na kameru	Děti se smějí do kamery
0.0	Děti se usmívají a mávají na kameru	Děti se mračí do kamery
0.908	Kinder lächeln und winken in die Kamera	die Kinder lacheln in die Kamera
0.004	Kinder lächeln und winken in die Kamera	die Kinder schauen stirnrunzelnd in die Kamera
0.889	Deti sa usmievajú a mávajú na kameru	Deti sa smejú do kamery
0.0	Deti sa usmievajú a mávajú na kameru	Deti sa mračia do kamery
0.951	Los niños sonríen y saludan a la cámara	Los niños sonríen a la cámara
0.025	Los niños sonríen y saludan a la cámara	Los niños fruncen el ceño a la cámara
Benchmarks (translated)		
2-way	BEGIN	Contradict 0.273
Jeho knihy byly přeloženy do 42 jazyků a vydány po celém světě.. ano, myslím, že jsem četl některé z jeho knih - je to spisovatel detektivek?		ano, je to spisovatel a autor krimi

2-way	BEGIN	Aligned	0.985
elvis aaron presley (8. ledna 1935 - 16. srpna 1977) byl americký zpěvák, hudebník a herec. o elvisovi presleym jsem samozřejmě slyšel! nemůžu však říct, že bych jeho hudbu nějak zvlášť poslouchal.		byl americký zpěvák a hudebník	
2-way	Q2	Aligned	0.037
První mistrovství se konalo tři roky po založení FIBA, v roce 1935.		Nejsem si jistý , ale poprvé se konalo v roce 1935.	
2-way	Q2	Aligned	0.913
Ačkoli většina druhů měkkýšů se sbírá ze slaného prostředí, některé druhy se vyskytují i ve sladké vodě.		Některé z nich se vyskytují ve sladké vodě , ale lze je nalézt i ve slané vodě.	
2-way	DialFact	Contradict	0.053
Péče o děti je široké téma zahrnující široké spektrum odborníků, institucí, souvislostí, činností, společenských a kulturních konvencí.		Určitě je to pravda. Péče o děti je úzké téma zahrnující jen velmi málo dětí.	
2-way	QAGS	Contradict	0.658
Čtyřiadvacet hodin poté, co floyd mayweather jr. oslnil média svými dovednostmi, přichází na řadu manny pacquiao. Filipínská ikona se dnes večer představí v ikonické tělocvičně wild card v los angeles pod dohledem trenéra freddieho roache. Poté, co mayweather přišel na trénink s téměř dvouhodinovým zpožděním, pacquiao slíbil, že přijde včas - a vy ho můžete sledovat od 23 hodin zde.		Filipínská ikona se představí v tělocvičně Wild Card v Los Angeles. Pacquiao slíbil, že přijde včas - a poté, co se mayweather na svůj trénink opozdil o dvě hodiny. Floyd mayweather jnr přichází na řadu.	
2-way	FactCC	Aligned	0.989
(CNN)Francouzští celníci tvrdí, že na palubě plachetnice, která v Karibiku falešně plula pod americkou vlajkou, zabavili více než 2 tuny kokainu. Podle ředitele celních operací na Martiniku Michaela Lachauxe se jedná o největší záchyt kokainu, který kdy francouzské úřady provedly, a jehož hodnota se odhaduje na více než 105 milionů dolarů. Policisté [CONTINUE]		Podle francouzských úřadů se jedná o největší záchyt kokainu, jaký kdy francouzské úřady provedly.	
2-way	Polytope	Contradict	0.988
Poprvé po osmi letech se televizní legenda vrátila k tomu, co umí nejlépe. Soutěžícím bylo řečeno, aby " přišli dolů ! " V prvním dubnovém vydání pořadu " the price is right " se nesetkali s moderátorem drewem careym, ale s jinou známou tváří, která měla řízení pořadu na starosti. místo toho se zde objevil bob barker, který tuto televizní hru uváděl 35 let, než v roce 2007 odstoupil. Ve svých 91 letech vypadal čile, a než předal moderátorské povinnosti careymu, který skončil, zvládl první hru o cenu, klasickou "šťastnou sedmičku". přestože byl většinu posledních osmi let mimo pořad, nezdálo se, že by Barkerovi něco chybělo.		Bob Barker hostil "the price is right" po dobu osmi let, odstoupil v roce 2007 poté, co byl pryč z přehlídky po většinu posledních 8 let Barker se nezdálo, že by chyběl na herní show.	

Table B.1: Example of AligScoreCS performance on task-specific and benchmark datasets.

Appendix C
Summarization Details

We provide supplementary details on the datasets utilized for summarization tasks in C.1, along with the candidate generation statistics C.1, detailed results on SumeCzech Text → Abstract task C.2, and examples of generated summaries C.3.

C.1 Dataset Details

We calculate the extractive fragment statistics proposed by [Grusky et al., 2018] to characterize the summarization datasets (on train sets), which are detailed in Table C.1. Furthermore, we report statistical values computed for the generated summaries with respect to the original articles, revealing the behavior of the learned core model. Based on these observations, we conducted filtering experiments described in Section 10.2.1.

Stats		fragment [Grusky et al., 2018]			text		summary	
type	dataset	compression	density	coverage	nsent	nwords	nsent	nwords
Initial	SumeCzech	11.70	0.52	0.12	27.77	409.27	2.75	38.36
	CNC	7.41	0.30	0.09	16.14	318.37	3.28	46.95
	XSUM	18.70	0.48	0.19	18.9	373.87	1.00	21.10
	CNNDM	15.97	2.62	0.51	35.47	691.87	3.76	48.04
type	dataset	compression	density	coverage	generated summary			
					nsent	nwords	diff	
mBART25	SumeCzech	14.96	2.33	0.30	2.32	32.36	11.70	
	CNC	8.82	1.40	0.22	3.03	45.66	17.00	
	XSUM	19.13	0.89	0.29	1.00	21.13	4.48	
	CNNDM	42.20	4.59	0.79	1.07	17.95	15.30	

Table C.1: Statistics on Summarization Datasets. The types *Initial* and *mBART25* depict statistical values computed for reference abstracts and for generated summaries, respectively. The **Coverage** measures the overlap degree of the extractive fragment between the article and summary, **Density** measures the average length of the extractive fragment, and **Compression** is the ratio of the article length to the summary length. The columns *nsent* and *nwords* are the average count of sentences and words, respectively. The column *diff* refers to the average distance in words from the generated summary to the corresponding reference summary.

Candidates

The following Figure C.1 plots averaged F-scores ($\frac{\text{F1-score+F2-score+FL-score}}{3}$) of ROUGE_{RAW} against scores of AlginScoreCS for each candidate summary per each dataset generated

by mBART25 core model. The values are grouped into bins of 0.1 squared size. Each bin indicates a number of candidate summaries assigned to that bin.

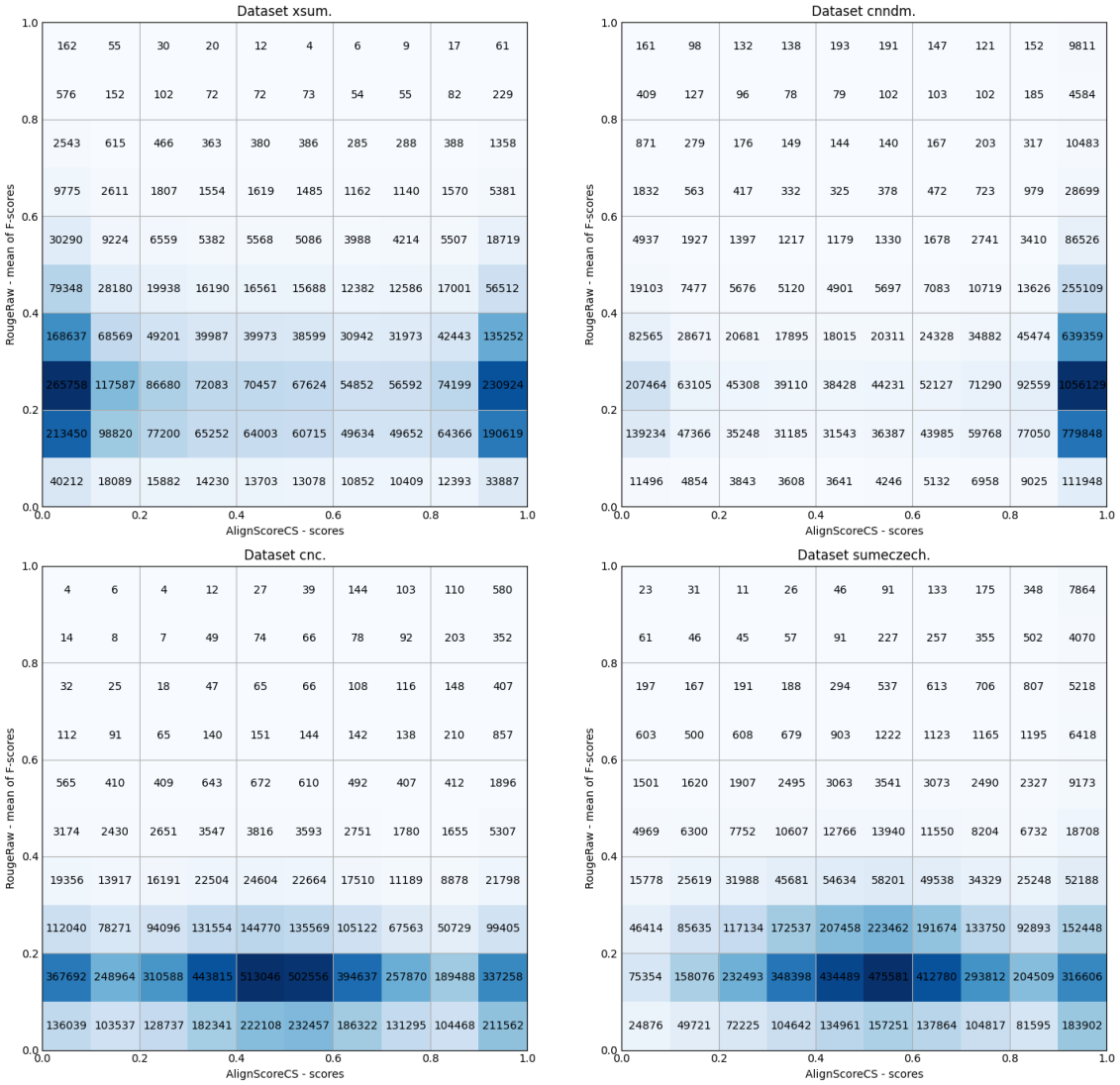

Figure C.1: Candidates' scores histogram distribution per each dataset. The graphs plot averaged f-scores of ROUGE_{RAW} against scores of AlginScoreCS for candidates summaries generated by mBART25 core model. Values are aggregated into bins of 0.1 size.

C.2 Detailed Results on SumeCzech

For future analysis, we present complete ROUGE_{RAW} results for both the *test* and *out-of-domain* (ood) sets of the SumeCzech dataset for the Text → Abstract task in Table C.2. Additionally, we include other models evaluated across all splits, reporting all ROUGE metrics. These models were not specifically refined for factuality and thus were not included in the main text. CzeGPT-2 [Hájek and Horák, 2024] is the Czech version of GPT-2, pre-trained on a Czech corpus and fine-tuned for summarization on SumeCzech. Textrank [Straka et al., 2018] simply selects a subset of sentences as a summary based on text similarity representations. The values presented show that the core-mBART model has improved recall scores compared to HT2A-CS, likely influenced by cross-lingual trans-

Model	SumeCzech - TEST									
	$ROUGE_{RAW}$									Factual Metrics
	P1	R1	F1	P2	R2	F2	PL	RL	FL	AlignScoreCS
textrank [Straka et al., 2018]	11.1	**20.8**	13.8	1.6	3.1	2.0	7.1	13.4	8.9	-
HT2A-CS [Krotil, 2022]	24.0	15.0	17.9	6.2	4.0	4.7	18.0	11.3	13.4	-
CzeGPT-2 [Hájek and Horák, 2024]	18.0	18.7	17.8	3.5	3.7	3.5	12.6	13.3	12.5	-
core-mT5	21.0	16.2	17.7	5.1	4.0	4.3	15.4	12.0	13.1	72.2
core-mBART	21.2	19.0	19.5	5.6	**5.1**	5.2	15.4	**13.9**	14.2	66.7
BARF	25.5	17.7	20.3	7.0	4.9	5.6	18.7	13.2	15.0	78.6
BART-mT5	26.0	13.8	17.6	6.3	3.4	4.3	19.4	10.5	13.2	**84.3**
AlignSum	26.7	17.8	**20.7**	**7.4**	4.9	**5.7**	19.6	13.2	**15.3**	76.3
BARF-Align	24.6	17.5	19.9	6.7	4.8	5.4	18.1	13.0	14.7	81.7
BARF-Loop	**27.2**	16.6	20.1	7.1	4.3	5.2	**20.0**	12.4	14.8	82.7

Model	SumeCzech - OOD									
	$ROUGE_{RAW}$									Factual Metrics
	P1	R1	F1	P2	R2	F2	PL	RL	FL	AlignScoreCS
textrank [Straka et al., 2018]	9.8	**19.9**	12.5	1.5	3.3	2.0	6.6	13.3	8.4	-
HT2A-CS [Krotil, 2022]	24.5	15.6	18.3	6.9	4.4	5.2	18.3	11.7	13.7	-
CzeGPT-2 [Hájek and Horák, 2024]	16.2	18.5	16.7	3.1	3.7	3.2	11.5	13.3	11.9	-
core-mT5	21.7	16.3	17.9	5.7	4.4	4.8	16.0	12.1	13.3	68.7
core-mBART	21.4	19.0	19.4	6.1	5.4	5.5	15.7	**13.9**	14.2	64.0
BARF	26.6	18.4	21.1	7.9	5.5	6.3	19.6	13.7	15.7	75.5
BART-mT5	27.4	14.8	18.7	7.2	3.9	5.0	20.2	11.0	13.9	**82.6**
AlignSum	27.6	18.5	**21.5**	**8.2**	**5.6**	**6.4**	20.3	13.7	**15.9**	74.5
BARF-Align	25.7	17.9	20.4	7.6	5.3	6.0	19.0	13.3	15.2	79.0
BARF-Loop	**28.0**	17.5	20.9	7.9	4.9	5.9	**20.6**	13.0	15.5	81.1

Table C.2: Results of BARF models on *test* and *ood* data of SumeCzech dataset for Text → Abstract task. P, R, and F represent precision, recall and F-scores of $ROUGE_{RAW}$ metric given the n-grams. The highest scores are highlighted, the second-highest are underlined per metric.

fer from multilingual fine-tuning. However, BARF models, specifically refined for factual consistency, demonstrate higher precision but lower recall, resulting in superior F-scores. This trend could be attributed to either the structure of training summarization datasets, which consists of summaries with varying lengths, or AlignScoreCS, which encourages BARF models to produce precise but shorter summaries relative to their references. On the other hand, despite being a relatively small model, CzeGPT-2 maintains similar scores for precision and recall. However, its performance is influenced by its generative capability. To summarize, AlignSum emerges as the top-performing model based on ROUGE scores, with BARF and BARF-Loop following closely behind, all reaching state-of-the-art results on SumeCzech.

C.3 Summaries

Presented below are several examples of generated summaries by our models that highlight specific drawbacks and improvements, showcased in the following tables. Upon examination, we observe that AlignScoreCS fails to detect instances where numbers or names are swapped when all other information is correct, leading to factual inconsistency. This indicates the necessity for additional fine-tuning to rectify this issue. However, we also noted that AlignScoreCS successfully identifies incomplete summaries and appropriately scored them as their label (missing).

Source Text		
Jordon Ibe's proved he's got some impressive moves on the football pitch but the young winger has shown us he's an equally adept dancer - by pulling off the trademark Michael Jackson 'moonwalk.' The Liverpool star has broken into the first team set-up this season, with some eye-catching performances at wing back or in a more attacking role. And Ibe took to Instagram, to show off his ability to perform the dance move, that many people around the country wish they could replicate. Jordon Ibe posted on Instagram, the video of him performing the moonwalk while he recovers from injury . The midfielder showed off some impressive moves as he steps up his recovery from a knee problem . He said: 'Having a little mess around, back training in a few days... OoohKillem MJ.' The 19-year-old is currently out with a knee injury but is expected to return in the next few weeks and boost Brendan Rodgers' side, as they continue to chase the Champions League places. Liverpool face a struggling Newcastle in their next Premier League outing as they look to close the gap on fourth. After being thumped by Arsenal at the Emirates, the Merseysiders know they can't afford to slip up in any of their remaining games. The 19-year-old has done well since breaking into the Liverpool first team and is seen as a key member .		
Summaries		
Reference	-	-
Jordan Ibe showed off the impressive dance move on his Instagram. The Liverpool star has broken into the first team during this campaign. Ibe is currently on the sidelines after suffering a knee injury. CLICK HERE for all the latest Liverpool news.		
core-mBART	Correct	0.93
Jordon Ibe posted a video on Instagram of him performing the Michael Jackson 'moonwalk'. The 19-year-old is out with a knee injury but is expected to return in the next few weeks. The midfielder has done well since breaking into the Liverpool first team this season. Click here for more Liverpool news.		
core-mT5	Correct	0.99
Jordon Ibe performed the Michael Jackson 'moonwalk' on Instagram. The 19-year-old is currently out with a knee injury but is expected to return in the next few weeks. Liverpool face Newcastle in their next Premier League outing.		
BARF-mT5	Correct	0.99
Jordon Ibe has broken into the Liverpool first team this season. The 19-year-old is out with a knee injury but is expected to return.		
BARF	Correct	0.98
Jordon Ibe posted the video on Instagram of him performing the Michael Jackson 'moonwalk' dance move. The 19-year-old midfielder is out with a knee injury but is expected to return soon. Liverpool face Newcastle in their next Premier League game.		
BARF-align	Correct	0.85
Jordon Ibe posted a video on Instagram of him performing the Michael Jackson 'moonwalk' dance move. The 19-year-old is out with a knee injury but is expected to return in the next few weeks. Liverpool face Newcastle in their next Premier League game. CLICK HERE for all the latest Liverpool news.		
AlignSum	Correct	0.99
Jordon Ibe posted the video on Instagram of him performing the Michael Jackson 'moonwalk' dance move. The 19-year-old midfielder is out with a knee injury. The winger is expected to return in the next few weeks.		
BARF-Loop	Correct	0.98
Jordon Ibe posted a video on Instagram of him performing the Michael Jackson 'moonwalk' dance move. The 19-year-old midfielder is out with a knee injury but is expected to return soon. Liverpool face Newcastle in their next Premier League game.		

Table C.3: Examples of CNNDM summaries: For each summary, we provide its origin, human label and AlginScoreCS score. Factual inconsistencies are highlighted in red. Here, almost all BARF models successfully omit the undesired banner, "CLICK HERE," which their core model does not.

Source Text		
Tandridge District Council leader Martin Fisher gave planners a shortlist of five areas for such a scheme. He said one new Surrey village is better than "scatter-gunning" homes and cramping the current infrastructure. But Andrea Watson from Lingfield Parish Council said it will not solve infrastructure problems and there has been "poor local consultation". Areas around Godstone, Blindley Heath, Redhill Aerodrome, Chaldon and Lingfield are all in the running under the new Local Plan. The district is 94% greenbelt and Mr Fisher said the council cannot meet the governmentś housing targets without building on some of it. In January, the government announced Longcross as the site of one of 14 garden villages across the country. Councillor Fisher said: "The key is infrastructure. If you do the scattergun approach you may put 300 houses somewhere and that might deliver you half an extra classroom. "Whereas if you put 4,000 house in a new village you can actually plan the proper infrastructure in for the village [and] itś relieving the pressure on the health centres and the schools in other parts of the district." But Mrs Watson claimed most people interested in the new homes will be London commuters, where the most pressing housing shortage lies, and so "London needs to solve the problem it is now trying to devolve to us". She added: "The new settlement will not solve the infrastructure problems, and the infrastructure will not come until many houses have been built, if it comes at all." The site chosen will depend on which can deliver the best mix of able housing, infrastructure gains and highways opportunities.		
Summaries		
Reference	-	-
Four thousand new homes could be built on greenbelt land in a new garden village development for Surrey.		
core-mBART	Incorrect	0.03
Plans to build 4,000 new homes in Surrey's Tandridge district have been given the go-ahead by councillors.		
core-mT5	Incorrect	0.10
Plans to build 4,000 new homes in a garden village in Surrey have been rejected by a local councillor.		
BARF-mT5	Incorrect	0.10
A new village in Surrey is to be built in a "scattergun" approach, a councillor has said.		
BARF	Correct	0.89
A councillor has defended plans to build 4,000 new homes in Surrey's Tandridge area in a bid to meet government housing targets.		
BARF-align	Incorrect	0.16
A councillor has defended plans to build 4,000 new homes in Surrey's Surrey garden villages, saying it will not meet government housing targets.		
AlignSum	Correct	0.97
Councillors in a Surrey district have been given a shortlist of sites for new villages to be built in a bid to meet government housing targets.		
BARF-Loop	Correct	0.98
A councillor has defended plans to build 4,000 new homes in Surrey's Surrey garden villages, despite concerns they will not meet government housing targets.		

Table C.4: Examples of XSUM summaries: for each summary, we provide its origin, human label and AlginScoreCS score. Factual errors are highlighted in red. Here, most of BARF models eliminate hallucinations of their core models; however, BARF-mT5 still produces slightly inappropriate summary.

Source Text		
Zjednodušeně řečeno, Robin Williams věděl, že přichází o rozum. Byl náladový, trpěl halucinacemi, třesem, zapomínal a byl paranoidní. Nikdo mu však nedokázal pomoci. Vzal si život a podle prvních zpráv proto, že trpěl depresemi. Pravda byla však mnohem horší. Williams mluvil pravdu, když přesvědčoval okolí, že vidí věci, které nejsou. Na jeho chorobu se přišlo až při pitvě. Pojďte si s námi připomenout život mnohonásobně oceněného umělce.Williams byl umělec, který trpěl vážnou duševní nemocí. Jeho život byl však mnohem horší, než se zdálo.		
Summaries		
Reference	-	-
Přesně před dvěma lety se rozhodl herec Robin Williams, jenž je označovaný za jednoho z největších herců své doby, ukončit život. Třiašedesátiletý herec trpěl korovou demencí s Lewyho tělísky, která způsobuje halucinace, potíže s motorikou a výkyvy v mentálních stavech.		
core-mBART	Incorrect	0.98
Známý americký herec a zpěvák Robin Williams (†63), který zemřel na rakovinu slinivky břišní, trpěl vážnou duševní chorobou.		
core-mT5	Incorrect	0.01
Známý americký umělec Robin Williams zemřel před pěti lety, když se dozvěděl, že přichází o rozum.		
BARF-mT5	Sufficient	0.98
Známý americký umělec Robin Williams věděl, že přichází o rozum. Byl náladový.		
BARF	Correct	1.0
Robin Williams byl umělec, který trpěl vážnou duševní chorobou. Trpěl depresemi a trpěl halucinacemi.		
BARF-align	Missing	0.54
Robin Williams (†63) byl umělec, který se proslavil jako zpěvák a herec. Jeho život však nebyl procházkou růžovým sadem.		
AlignSum	Correct	0.99
Byl to umělec, který trpěl vážnou duševní nemocí. Robin Williams (†63) byl paranoidní člověk. Trpěl depresemi a trpěl halucinacemi.		
BARF-Loop	Correct	1.0
Robin Williams byl umělec, který trpěl vážnou duševní nemocí. Jeho život byl však mnohem horší, než se zdálo.		

Table C.5: Examples of SumeCzech summaries: for each summary, we provide its origin, human label and AlginScoreCS score. Factual errors are highlighted in red. Here, the BARF models effectively erase the hallucinations of core-mBART model.

Source Text		
O tom, jaké má potíže, promluvil Petr Štěpánek nedávno v Blesku: "Pořád to cítíte, ale hlavně se na tu nohu nemůžete absolutně spolehnout. To znamená, že všechny přemrštěné pohyby mám stále zakázané. Operace mě čeká už v létě." Herec přičítá své zdravotní problémy zátěži, kterou musely jeho klouby snášet v mládí, kdy aktivně sportoval.Jenže herec je známý dříč a jeho manželka Zlata Adamovská už tuší, že by i po další operaci nemuselo dojít k nápravě, pokud se Štěpánek nebude šetřit. "Slyšel jsem, že je z toho Zlata hodně špatná. Prý se bojí, aby to nedopadlo tak, že bude Petr celý zbytek života kulhat. Mělo se jí o tom i zdát," řekl nejmenovaný zdroj Šípu a dodává: "Zlata se bojí, aby nezopakoval tu samou chybu, co loni. Protože bude mít nový kloub, myslí si, že už není čas pokoušet znovu osud tím, že ho bude přepínat.Adamovská už si prý vymyslela plán, jak Štěpánka udržet ve formě. "Tipoval bych, že možná i vymyslí nějakou pěknou dovolenou, těžko říct. Zatím mi neprozradila, co chystá. Ona má ráda překvapení, takže o tom asi nebude mluvit. Potřebuje mu do hlavy zanést pochybnosti, že činnost je nejlepší lék. Nechce, aby toho pak litoval až do konce života, takže je odhodlaná udělat v podstatě cokoli," prozradil informátor.		
Summaries		
Reference	-	-
Je to deset měsíců, co musel herec Petr Štěpánek (66) podstoupit operaci kolene. Jenže se kýžený výsledek nedostavil, a herec dál trpí bolestmi. Herec, který stále naplno pracuje se tak bude muset podrobit dalšímu chirurgickému zákroku a to dělá vrásky na čele jeho choti Zlatě Adamovské (56). Bojí se totiž, že se herec bude i nadále přepínat a operace tak bude zase k ničemu.		
core-mBART	Incorrect	0.22
Herec a dabér Petr Štěpánek (66) se už několik let potýká s vážnými zdravotními problémy. Loni mu praskla kyčelní kloub a musel podstoupit operaci, při níž mu lékaři museli vyměnit obě nohy. Nyní musí na další operaci.		
core-mT5	Correct	0.95
Herec Petr Štěpánek, který loni v létě podstoupil operaci kloubů, promluvil o svém zdravotním stavu. Podle informací serveru Šíp.cz má stále problémy s nohou a jeho manželka Zlata Adamovská se bojí, že by i po další operaci nemuselo dojít k nápravě.		
BARF-mT5	Missing	0.62
Herec Petr Štěpánek, který je známý dříč, má zdravotní problémy. Jeho manželka Zlata Adamovská už tuší, že by se mohlo znovu šetřit.		
BARF	Incorrect	0.92
Herec Petr Štěpánek (68) má stále problémy s nohou. Operace kloubů ho čeká už v létě. Jeho manželka Zlata Adamovská (57) se bojí, aby se nestalo, že bude herec kulhat zbytek života.		
BARF-align	Correct	0.96
Herec Petr Štěpánek (66) má stále problémy s klouby. Operace kloubů ho čeká už v létě. Herec je známý dříč a jeho manželka Zlata Adamovská už tuší, že by po další operaci nemuselo dojít k nápravě.		
AlignSum	Incorrect	0.93
Herec Petr Štěpánek (68) má stále potíže s nohou. Operace kloubů ho čeká už v létě. Jeho manželka Zlata Adamovská (57) se bojí, že by to nemuselo dopadnout dobře.		
BARF-Loop	Correct	0.84
Herec Petr Štěpánek (66) má stále problémy s nohou. Operace kloubů ho čeká už v létě. Jeho manželka Zlata Adamovská je z toho hodně špatná.		

Table C.6: Examples of SumeCzech summaries: for each summary, we provide its origin, human label and AlginScoreCS score. Factual errors are highlighted in red. Here, all BARF models eliminate the hallucinations present in the core models; however, some still confuse the ages of the people involved.

Source Text		
Kilům, obezitě, hubnutí a zdravému způsobu života věnovala Iva Málková velký kus života, jen málokdo v Česku má s redukcí hmotnosti víc zkušeností. Zobrazit fotogalerii"Pokud chceme hubnout, je základem úspěchu energetická nerovnováha (víc vydávat než přijímat). Důležitá je ale také cesta, jak toho dosáhneme. Je-li to cesta nepříjemná, velmi odlišná od dosavadních návyků, bývá vždy jen provizoriem a už se těšíme, až se vrátíme k původním návykům," shrnuje své zkušenosti s hubnutím vystudovaná psycholožka.Její často "nerozumné" hubnoucí pokusy byly spíše motivovány touhou poznání. "Většina uvedených laických nabídek se opakuje ve vlnách a nevede k dlouhodobému výsledku," dodává šéfka společnosti STOB.V roce 1970 mi bylo 15 let, chodila jsem na střední školu, vážila 75 kilo a měla vesměs štíhlejší spolužačky. Do té doby už jsem měla za sebou hubnoucí pokusy, ale spíše z máminy iniciativy. Myslela to dobře, a tak jsem ve školce nesměla k svačině rohlíky, doma dávali bratrovi jíst tajně, protože jsem chtěla tolik co on. Intuitivně z mého dosahu odstraňovali "toxické prostředí" provokující ke konzumaci nadměrného množství jídla. Se vstupem na střední školu jsem začala hubnout z vlastního přesvědčení.Diety se tehdy vyměňovaly spíše po sousedsku a televize vysílala dobré rady nestora obezitologie Rajka Dolečka v pořadu Nebezpečný svět kalorií.Se studiem psychologie pak přišel pro mě radikální zlom... [CONTINUE]		
Summaries		
Reference	-	-
Zhubnout umí téměř každý, tvrdí zakladatelka společnosti Stop obezitě, kterou kila navíc provázejí už více než pět desetiletí. "Problémem však je si váhové úbytky udržet," říká Iva Málková a radí nepodléhat módním vlivům.		
Core-mBART	Correct	0.91
Šéfka společnosti STOB Iva Málková zasvětila celý svůj život hubnutí a zdravému životnímu stylu. Hubnutí se věnuje už přes 40 let. "Není to jen o jídle, ale o co možná nejpříjemnější změně návyků," říká.		
Core-mT5	Correct	0.95
Iva Málková, šéfka společnosti STOB, věnovala své zkušenosti s hubnutím velký kus života, jen málokdo v Česku má s redukcí hmotnosti víc zkušeností. "Většina uvedených laických nabídek je spíše motivována touhou poznání," říká.		
BARF-mT5	Missing	0.92
Známá psycholožka Iva Málková se věnuje hubnutí a zdravému životnímu stylu. V roce 1970 vážila 75 kilo.		
BARF	Correct	0.79
Věnovala se hubnutí, zdravému životnímu stylu a zdravému způsobu života. Iva Málková je už přes 40 let šéfkou společnosti STOB, která se zabývá zdravým životním stylem.		
BARF-align	Correct	0.98
Věnuje se hubnutí, zdravému životnímu stylu a zdravému způsobu života. Psycholožka a šéfka společnosti STOB Iva Málková se zabývá zdravým životním stylem a hubnutím už přes 40 let.		
AlignSum	Correct	0.84
Šéfka společnosti STOB Iva Málková je už přes 40 let přesvědčená, že není problém zhubnout, ale udržet si váhu. A to i přes to, že se jí podařilo změnit návyky.		
BARF-Loop	Incorrect	0.94
Věra Málková je šéfkou společnosti STOB, která se zabývá hubnutím a zdravým životním stylem. A to už přes 40 let.		

Table C.7: Examples of SumeCzech summaries: for each summary, we provide its origin, human label and AlginScoreCS score. Factual errors are highlighted in red. Here, we focus on BARF-Loop, which incorrectly generates name for a person, while the other models performed well.

Appendix D
Attached Files

In addition, we provide our code base of both parts in the attached files. The following tree-structure briefly summarizes the sub-directories and included Python files.