POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Evaluation of Metrics for Neural-network-based Summarization

Tesi di Laurea Magistrale in
Computer Science Engineering - Ingegneria Informatica

Author: **Yasmin Awad**

Student ID: 963681
Advisor: Prof. Maristella Matera
Co-advisors: Emanuele Pucci, Marco Fisichella
Academic Year: 2022-2023

Abstract

In today's digital age, assistive technologies are an important tool to allow equal access to information. Web interfaces that rely heavily on visual information create a large information gap for visually impaired users. ConWeb (Conversational Web Browser) is a framework that allows interactive web browsing by automatically generating a dialogue with the user using the information contained in the Web page. During user studies, interest was shown in the possibility of obtaining summaries of Web pages. Advances in neural network technology have greatly improved the quality of automatic text summarization. However, it is difficult to automatically evaluate abstractive summaries generated by neural network models, as traditional evaluation metrics cannot capture their semantic coherence. In this study, we aimed to select a generative model to obtain high-quality automatic summaries for ConWeb and to evaluate various metrics used in general for the automatic evaluation of summaries. Specifically, we generated summaries of 76 Italian articles using four generative models: mbart-summarization-mlsum, mbart-summarization-ilpost, BART, and GPT-3.5. We performed an evaluation analysis using 8 metrics (ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, BLEU, BLANC, chrF, and METEOR) and a crowdsourced human evaluation. Analyzing the results, the quality of the summaries generated by GPT-3.5 was shown to be significantly high according to human evaluation. A significant discrepancy emerged when comparing human evaluation to automatic metrics, further underscored by the weak correlation observed between the two sets of results. We concluded that traditional automatic metrics cannot adequately evaluate abstractive summaries generated by Large Language Models and that the best model among those considered for our goal is GPT-3.5.

Keywords: conversational Web, text summarization, abstractive summarization, large language models, traditional metrics, human evaluation

Abstract in lingua italiana

Nell'odierna era digitale, le tecnologie assistive risultano essere uno strumento importante per permettere un accesso equo all'informazione. Le pagine Web che fanno molto affidamento sulle informazioni visive creano un ampio divario informativo per gli utenti ipovedenti. ConWeb (Conversational Web Browser) è un framework che consente la navigazione web interattiva tramite la generazione automatica di un dialogo con l'utente utilizzando le informazioni contenute nella pagina Web. Durante gli studi con gli utenti, è stato mostrato interesse verso la possibilità di ottenere dei riassunti delle pagine Web. Il progresso nella tecnologia nelle reti neurali ha migliorato notevolmente la qualità dei riassunti automatici. Tuttavia è difficile valutare automaticamente i riassunti astrattivi generati dai modelli di reti neurali, poiché le metriche di valutazione tradizionali non sono in grado di catturarne la coerenza semantica. In questo studio, abbiamo mirato a selezionare un modello generativo per ottenere riassunti automatici di alta qualità per ConWeb e a valutare varie metriche utilizzate in generale per la valutazione automatica di riassunti. Nello specifico, abbiamo generato riassunti di 76 articoli Italiani utilizzando quattro modelli generativi: mbart-summarization-mlsum, mbart-summarization-ilpost, BART e GPT-3.5. Abbiamo effettuato un'analisi di valutazione utilizzando 8 metriche (ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, BLEU, BLANC, chrF e METEOR) e una valutazione umana tramite crowdsourcing. Analizzando i risultati, la qualità dei riassunti generati da GPT-3.5 si è dimostrata significativamente elevata secondo la valutazione umana. Una discrepanza significativa è emersa confrontando la valutazione umana con le metriche automatiche, ulteriormente sottolineata dalla debole correlazione osservata tra le due valutazioni. Abbiamo concluso che le tradizionali metriche automatiche non possono valutare adeguatamente i riassunti astrattivi generati da modelli generativi di linguaggio di enormi dimensioni e che il miglior modello tra quelli presi in considerazione per il nostro obbiettivo è GPT-3.5.

Parole chiave: Web conversazionale, riassunti testuali, riassunti astrattivi, modelli generativi, metriche tradizionali, valutazione umana

Contents

1 | Introduction

1.1. Conversational Web Browsing

In the contemporary digital landscape, the pervasive use of visual design in web interfaces poses significant assistive technology challenges for individuals with visual impairments. It is essential to ensure that everyone can access the web seamlessly, without encountering browsing issues. According to the World Health Organization [13], globally, at least 2.2 billion people have near or distant vision impairment. Also, according to the Centers for Disease Control [5], visual impairment ranks among the top 10 impairments for adults over 18 and is prevalent among children. Consequently, individuals with visual impairments or those unable to access the Internet with traditional methods must be provided with alternatives to achieve the same objectives.

This enhanced functionality can be beneficial to a diverse range of users, including the visually impaired, individuals needing to retrieve information while driving or cleaning, and the elderly who may struggle with electronic devices. Among the tools used by people with visual impairments, screen readers are the most commonly employed for fundamental information access on the Internet. These tools leverage the tags implemented by web developers to facilitate online content access. Challenges arise when web developers do not adhere to standards, compromising the effectiveness of screen readers. In such cases, advanced functionality cannot be provided, obliging users to manually check the entire web page to find the desired content. Acknowledging these limitations, the Politecnico di Milano developed an innovative vocal assistant called ConWeb (Conversational Web Browser) [14, 35, 43] designed to facilitate a conversational user experience. The primary objective is to enable users to browse web content and services by engaging in a dialogue with websites, avoiding visual exploration. Users can articulate their goals in natural language, initiating a conversation mediated by a conversational agent. Specifically, ConWeb has been tailored for Italian speakers to browse Italian web pages.

Searching for content can become aurally tedious to deal with. Visually impaired users, not being able to read the content, cannot quickly scroll through it to visually search for sections of content with the information they are looking for. Often, sometimes even unconsciously, users without visual learning try to quickly capture some keywords in the text with a careless look, or by reading by skipping a few lines. This is done to speed up their user experience. During user studies of ConWeb, the summary emerged as a pattern. Users attached significant importance to the availability of summaries, seeking smoother and faster browsing. However, they also expressed concerns about the potential quality of these summaries, leading to the omission of crucial information within the text. This emphasis prompted our focus on examining the quality of summaries in Italian. To carry out the addition of a summarization feature within ConWeb, research was undertaken to choose the most suitable summarization model for our task.

1.1.1. ConWeb Architecture

The main goal of the ConWeb Architecture is to enable natural conversations between users and the system, focusing on being accurate and responsive. In its latest iteration [43], ConWeb integrates an NLU pipeline (Figure 1.1) and automated webpage tagging for enhanced functionality. The NLU pipeline holds a crucial role by extracting intents and entities from user utterances. The Rasa interpreter, carried over from previous versions [35], remains in use but given its limitations in extracting intents and entities, it has been supported by other elements. Rasa is now complemented by a rule-based interpreter to confirm and adjust predictions as needed. Should all else fail, the system resorts to querying an OpenAI model.

Specifically, the session handler initiates calls to the pipeline to delineate intents and entities. Initially, Rasa is consulted followed by the Heuristics. If neither can discern the requisite information, assistance is sought from a GPT-3.5 model, specifically OpenAI's Davinci. This integration aims to enhance the precision of intent and entity extraction, especially considering Rasa's limitation to webpage entities. Heuristics employs SpaCy's pre-trained word embeddings to classify intents and detect entities. Subsequently, the rule-based model extracts entities, followed by heuristic application and consistency checks to validate results. In instances where both Rasa and Heuristics falter, OpenAI is enlisted to extract intents and entities.

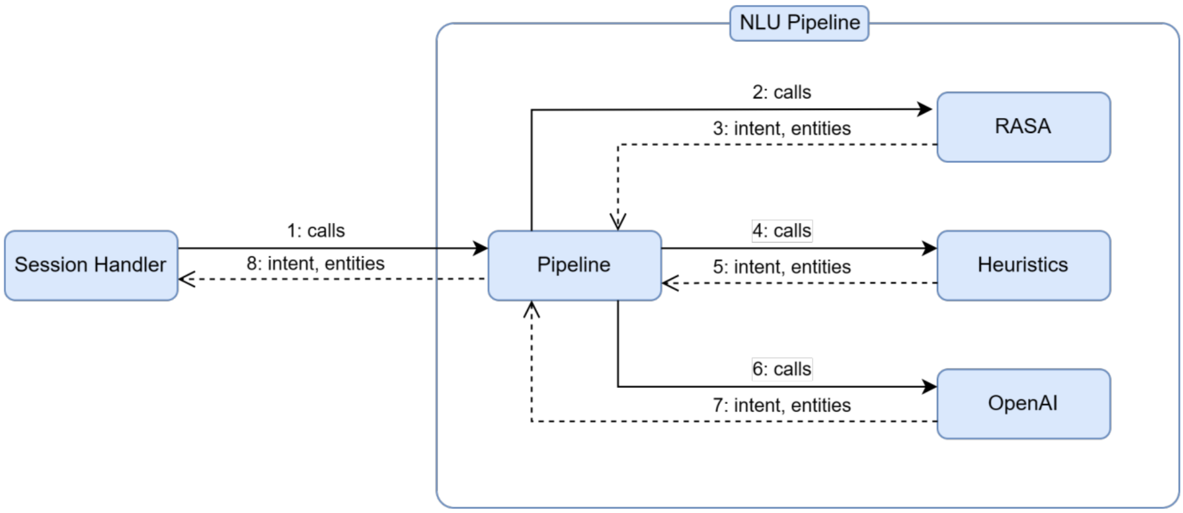


Figure 1.1: NLU Pipeline Architecture within ConWeb.

1.2. Introduction to Text Summarization

In the modern digital era, where we're flooded with information, summarization is incredibly crucial. With the vast amount of content available on the internet, users often find themselves overwhelmed by lengthy articles, complex research papers, and extensive web pages. Consider a scenario where a student is conducting research for a school project. Instead of sifting through numerous lengthy articles and academic papers, a well-crafted summary can provide them with the key findings and insights they need in a fraction of the time. Similarly, professionals navigating through vast amounts of data and reports can benefit greatly from summaries tailored to their specific needs, enabling them to make informed decisions quickly and efficiently. Even casual web users, when browsing through news articles or blog posts, can save time and effort by accessing concise summaries that capture the essence of the content without having to delve into every detail. This is especially true when we take into account situations where the content is to be listened to instead of read, like in the previously stated cases of visually impaired users.

Summarization plays a crucial role in addressing this challenge by condensing large volumes of information into concise and easy-to-understand formats. By providing summaries, users can quickly grasp the key points of a document or webpage without the need to read through lengthy passages. This not only saves time but also enhances efficiency and productivity, allowing individuals to navigate the web more effectively and access relevant information with ease. Moreover, in an era where mobile browsing is prevalent, where users often consume content on small screens and limited data plans, concise summaries are especially valuable, enabling seamless and efficient browsing experiences. As such, summarization has become an essential tool in the digital toolkit, empowering users to extract meaningful insights from the internet effortlessly.

1.3. Problem statement

This research was born from the development of the ConWeb framework to assist visually impaired users in browsing the internet through a voice assistant. A crucial element for enhancing navigation fluidity in ConWeb is the option for users to ask for summaries of the desired web page's content or specific sections. In this way, the user can listen to the condensed version of the web content. This functionality empowers users to make informed decisions about whether to delve into the complete content or not.

During the study and testing of ConWeb, it became clear that the quality of the summary was a significant concern for users. As advanced generative language models like GPT were not available at that time, the Politecnico di Milano researchers have opted to assess various language models (LLMs) in Italian, our native language, to check their performance. Over time, access to models like GPT [38] became easier, making ConWeb more efficient and improving its quality. However, despite achieving scores similar to other models with automatic metrics, we recognized the importance of validating more rigorously with human input. The decision to implement human evaluation was not only driven by our desire for validation but also influenced by insights from prior studies. Indeed, previous research [24, 25, 32, 48] led us to conclude that automatic metrics were inadequate for accurately assessing the performance of generative LLMs. Human evaluation has been essential to both demonstrate the superiority of summaries generated with foundational models and emphasize the necessity of incorporating human judgment alongside existing automatic metrics.

This thesis aims to find the best model to incorporate into the ConWeb architecture, to provide useful summaries to users. To ensure the effectiveness and suitability of the summarization network within the Conweb framework, it is essential to evaluate and compare different neural network architectures thoroughly. To do so, we have first investigated current automatic metrics of neural-network-based summarization systems. Once the first evaluation had been done, the results obtained were compared to the human evaluation carried out specifically for our task. Conclusions will be shown at the end.

1.4. Methodology

The research was approached with the following strategy:

1. Address the problem by doing a systematic literature review on Web assistive technologies, summarization approaches, Large Language Models, and automatic evaluation metrics.
2. Review and enhance the current dataset.
3. Analyze the strong and weak features of the summaries generated by the chosen models.
4. Carry out a technical evaluation of the summaries using traditional automatic metrics.
5. Identify problems of the application of automatic evaluation on summaries generated by Large Language Models.
6. Carry out a Human Evaluation.
7. Analyze the results through correlation.

1.5. Contributions

This thesis comprises the following specific contributions:

- Application of automatic metrics to evaluate summaries of Italian Web articles.
- Creation of a human evaluation form for the evaluation of summaries.
- Comparison of the performance of different Large Language Models against extractive and abstractive human-generated reference summaries.
- Evidence that automatic metrics are inadequate for GPT and the latest pre-trained Large Language Models.
- Validation through human evaluation of the quality and performance between four Large Language Models over the task of summarization of Italian Web articles, not yet present in the literature.

1.6. Structure of the Thesis

- **Chapter 2:** It introduces the state of the art of previous works related to Web assistive technologies, text summarization, metrics for summarization evaluation, and human evaluation techniques. About the metrics, it gives an overview of the automatic metrics used in the following sections to analyze the summaries, explaining how they have been applied.
- **Chapter 3:** It defines the goals of the research, illustrating the set-up of our work and the changes made to the dataset used in previous works.
- **Chapter 4:** It shows the results of the application of automatic metrics over the summaries generated by the Large Language Models taken into consideration.
- **Chapter 5:** It describes the human evaluation carried out and shows its results.
- **Chapter 6:** It depicts the results of the correlation analysis between the results of the automatic metrics and the results of the human evaluation. It also shows an attempt to improve the results through an Italian language suited preprocessing.
- **Chapter 7:** It concludes the thesis by wrapping up our findings and suggesting the next steps to take to improve our results.

2 | State Of The Art

2.1. Web Accessibility

The Internet serves as an invaluable repository of information, accessed regularly by virtually everyone. However, individuals utilize the web in diverse ways, influenced by factors such as the device they use, their proficiency with it, or any impairment they may have. Addressing these challenges, innovative methods have been developed to ensure web browsing, facilitating use for all users, irrespective of their circumstances.

The World Wide Web Consortium (W3C) has taken a leading role in promoting web accessibility, aiming to enhance usability for people with impairments. Accessibility, as defined by W3C, involves designing and developing websites, tools, and technologies to be usable by individuals with impairments. This encompasses the capability to perceive, understand, browse, and interact with web content, as well as to contribute to it.

Web assistive technologies extend to various impairments, including auditory, cognitive, neurological, physical, speech, and visual impairments. Moreover, it benefits not only those with impairments but also users without them. The W3C Web Accessibility Initiative (WAI) spearheads the development of technical specifications, guidelines, techniques, and resources aimed at facilitating browsing solutions. WAI endeavors to establish standards for web development, ensuring that websites adhere to specific guidelines.

Despite these efforts, many web developers remain either unaware of those guidelines or choose to disregard them. Consequently, web development often falls short of its objectives, resulting in users encountering various difficulties during browsing. Moreover, the majority of websites do not fully support accessibility, leading users who require it to face challenges in navigation, ultimately prompting them to abandon the site.

Given these factors, it is evident that web accessibility presents a complex issue that demands attention, particularly considering the vast number of websites available for browsing. Efforts to address accessibility are crucial to ensuring equal access to information on the web.

2.2. Summarization with Large Language Models

2.2.1. Summarization Introduction

Text summarization is a challenging task in natural language processing that aims to condense large volumes of textual information into concise and coherent summaries. With the exponential growth of online content, the need for efficient and accurate summarization techniques has become increasingly important. Automatic text summarization can benefit a wide range of applications, including information retrieval, document summarization, news aggregation, and content recommendation systems. It plays a crucial role in information retrieval, enabling users to quickly grasp the essence of lengthy documents, news articles, or online content.

Encoder-decoder architectures, particularly the sequence-to-sequence (Seq2Seq) models with attention mechanisms, have gained considerable popularity in text summarization. These models employ recurrent neural networks (RNNs) or transformer-based encoders to encode the source text and generate summaries using a decoder. Advanced architectures, such as the Transformer [46] model, have shown exceptional performance in various NLP tasks, including text summarization. Transformers rely on self-attention mechanisms to capture global dependencies in the source text, enabling the generation of informative and coherent summaries. The past years have been the protagonists of the development of neural network models to solve Natural Language Processing tasks. Architectures like GPT [38] have achieved state-of-the-art and human-level performance in many of those tasks, including summarization.

2.2.2. Summarization Approaches

The two main approaches to text summarization are extractive and abstractive summarization [15].

- *Extractive* summarization involves selecting the most salient sentences or phrases from the source text and forming a summary by concatenating them
- *Abstractive* summarization generates summaries by understanding the meaning of the source text and generating new sentences that capture the essential information

As mentioned earlier, the abstractive method entails the machine's genuine capability to grasp the semantics of the text and formulate new phrases in natural language. It poses a greater challenge as it seeks to comprehend the entire document and produce paraphrased text to encapsulate the key points. Transformer models are renowned for their superior performance in handling complex language tasks like text summarization. Similar to humans, these models can adeptly rephrase intricate sentences into concise phrases that retain the essence and meaning of the original text.

Even if some studies [22] investigated methods for producing summaries without relying on Large Language Models (LLMs), most researchers have looked into using LLMs (and thus Transformers) to create better summaries. For instance, a recent study [23] compared GPT [38] against fine-tuned models trained on large summarization datasets. The results indicated a strong preference for GPT-generated summaries.

2.2.3. Transformer Models

A transformer is a deep learning model created to solve sequence-to-sequence tasks while handling long-range dependencies with ease. Tasks, which are characterized by the transformation of an input sequence to an output sequence, could be text generation, speech recognition, masked language modeling, and so on. The following will introduce significant state-of-the-art transformer models.

MPT-7B

MPT-7B [1] is a decoder-style transformer with 6.7B parameters, trained from scratch on a large amount of data (1T tokens) of text and code. It is open-source and available for commercial use. The MPT-7B model is part of the family of MosaicPretrained-Transformer (MPT) models, which use a modified transformer architecture optimized for efficient training and inference. It eliminates the context length limits by replacing positional embeddings with Attention with Linear Biases (ALiBi) [37]. This means that it is prepared to handle extremely long inputs. The fine-tuned version MPT-7B-StoryWriter-65k+ was built with a context length of 65k tokens. But at inference time, thanks to ALiBi, it can make generations as long as 84k tokens.

BART

BART [29] combines the bidirectional encoder of BERT [18] and the left-to-right decoder of GPT [38] in a seq2seq translation framework. It attains state-of-the-art results in summarization tasks. BART is particularly effective when fine-tuned for text generation but also works well for comprehension tasks. It is versatile, supporting various tasks such as machine translation, question-answering, text summarization, sequence classification, and sentence entailment, which assesses the logical relationship between sentences.

mBART

The mBART [31] is a sequence-to-sequence denoising auto-encoder pre-trained on extensive monolingual corpora across multiple languages using the BART objective. Unlike previous methods that focused solely on parts of the sequence-to-sequence model, mBART pioneers pre-training the entire model by denoising full texts in multiple languages. Pre-training a complete model allows it to be directly fine-tuned for supervised and unsupervised machine translation, with no task-specific modifications. Different fine-tuned versions of mBART are currently available, including some versions fine-tuned on Italian datasets.

GPT

GPT (Generative Pre-trained Transformer) [38] is a cutting-edge deep learning model renowned for its remarkable capabilities in natural language processing tasks. Developed by OpenAI, GPT leverages a transformer architecture, which excels in understanding and generating coherent text. Unlike BERT [18], which employs a bidirectional encoder, GPT utilizes a left-to-right decoder mechanism. It has consistently achieved state-of-the-art results in tasks such as language generation, text completion, and dialogue generation. Moreover, GPT demonstrates remarkable adaptability, showcasing impressive performance even when fine-tuned for specific tasks. GPT can comprehend and generate text with remarkable fluency and coherence. This makes it particularly effective for tasks requiring natural language understanding and generation. Additionally, GPT has been successfully applied to various practical applications, including machine translation, question-answering, text summarization, and sequence classification. ChatGPT, OpenAI's product based on the GPT model emerged as the most popular AI tool of 2023 [16], dominating the industry's web visits with more than 60% of visits from September 2022 to August 2023.

Gemini

Gemini [41] is an innovative deep-learning family of architectures designed to excel in natural language processing tasks. Gemini models are trained to be natively multimodal, which means able to work with and use more than just words. Because the Gemini models are multimodal, they can perform a range of multimodal tasks, from transcribing speech to captioning images and videos to generating artwork. Gemini models are built upon Transformer decoders, enriched with architectural enhancements and model optimizations. These adaptations facilitate stable training at scale and optimized inference on Google's Tensor Processing Units. They are trained to accommodate a context length of 32k, utilizing efficient attention mechanisms. Gemini achieves groundbreaking results across various tasks, specifically also in natural language processing tasks such as machine translation and summarization. This is due to a neural network architecture characterized by two distinct pathways (one for encoding input information and another for decoding and generating text), thanks to which Gemini effectively captures semantic relationships in language data. The Gemini family consists of Ultra, Pro, and Nano sizes. Each size is tailored to accommodate various computational constraints and specific application needs.

LLaMA

LLaMA (Language Learning and Mastery Architecture) [42] is a collection of foundation language models ranging from 7B to 65B parameters with competitive performance compared to the best existing LLMs. Built upon the transformer architecture, LLaMA incorporates several enhancements, notably improving training stability. Operating by ingesting a word sequence and predicting subsequent words recursively, LLaMA adeptly captures intricate linguistic patterns through extensive pre-training on vast text corpora. LLaMA consistently delivers cutting-edge results across various tasks (such as text generation, sentiment analysis, document classification, and machine translation), demonstrating its adaptability and effectiveness in practical applications. Moreover, its accessibility on a single GPU and its training solely on publicly available data contribute to democratizing access to and study of LLMs.

2.2.4. Good Summary Definition

The four dimensions proposed in the literature for evaluating the goodness of a summary are the following [26]:

- **Fluency**: each sentence of the summary should be well-formed and free of grammatical errors or random capitalization that makes it hard to read.
- **Coherence**: it depicts the collective quality of all sentences. The summary should be well-structured and not just a heap of information.

- **Relevance**: the summary should enclose the important aspects from the source document, excluding the rest.
- **Consistency**: the summary and source document should be factually consistent. This means that the summary should not contain information that is not in the source document.

Consistency is closely related to hallucination. Specifically, there is a distinction between consistency and accuracy. A summary can be accurate but inconsistent if it adds accurate information that was not present in the source document. Usually, if a summary is inconsistent, it is also inaccurate.

2.2.5. The unsuitability of current Evaluation Metrics

One of the primary challenges in developing summarization tools is evaluating the quality of the summaries. Human assessment by linguistic experts or crowdsourcing is slow, costly, and lacks standardization. Meanwhile, automatic evaluation metrics are said to have insufficient correlation with human quality ratings. The inherent complexity of language makes it challenging to devise evaluation metrics that effectively capture the quality and coherence of generated summaries. Several metrics currently in use were crafted and tested using datasets from the Document Understanding Conference (DUC) and Text Analysis Conference (TAC) shared tasks. However, recent findings indicate that these datasets feature human judgments scoring model outputs on a lower scale than what current summarization systems achieve. This revelation raises doubts about the accurate performance assessment of these metrics in the updated setting.

For example, traditional evaluation metrics, such as ROUGE [30], have proven in the past to be an important metric for evaluating summarization results. However, these metrics primarily focus on lexical matching and fail to capture the understanding of the readability and content order, the semantic meaning, and the overall quality of the summaries. In general, these metrics tend to favor extractive approaches, as they heavily rely on lexical overlap, and may not capture the creative and abstractive nature of summarization. Previous research [25, 32] showed that metrics such as ROUGE [30] are unreliable to evaluate LLMs summaries, which score way better if we take into consideration human evaluation.

Those have not been the only studies to reach that conclusion. Earlier studies on German summaries [24] have indicated that automatic evaluation metrics such as BLEU [34], ROUGE [30], and BertScore [47] are inadequate for assessing linguistic quality. Furthermore, these metrics exhibit a low correlation with any content-related human ratings obtained from crowdsourcing or expert evaluations. Another research [48] has conducted a human evaluation on ten LLMs for a better evaluation. They have computed six popular automatic metrics and their system-level correlations against human ratings. They found that the performance of automated metrics may depend on the quality of reference summaries. In particular, reference-based metrics correlate better with human judgments on the aspects for which reference summaries also have better scores. They believe that metric correlation can be improved by using better reference summaries.

These findings prove that automatic metrics are unreliable in evaluating LLMs. However current state-of-the-art summarization techniques tend to use abstractive methods, such as generative LLMs, to generate summaries, which indeed clash with the use of traditional metrics. They fail to understand the overall meaning of the texts in the evaluation of abstractive summaries, where the lexical matching falls short. This lack of proper automatic evaluation metrics for LLMs is a high challenge for the development of natural language systems, which is usually overcome by the use of the current golden standard which is human evaluation.

2.3. Metrics For Summarization Evaluation

2.3.1. Overview of Evaluation Approaches

The evaluation of text summarization systems is a challenging task due to the subjective nature of summarization quality. As the demand for effective summarization techniques rises, the need for robust and objective evaluation metrics becomes increasingly critical. Evaluation metrics can be divided into 3 categories [15]:

- *Human-Centric Evaluation*: as said by the name, it involves humans as judges. Naive or expert subjects are asked to rate or compare texts generated by different NLG systems.
- *Untrained Automatic Metrics*: also known as automatic metrics, are the most commonly used. These evaluation methods compare the quality of machine-generated texts by comparing them to human-generated texts (reference texts). They operate on identical input data and utilize metrics that don't rely on machine learning. Instead, they are grounded in factors like string overlap, content overlap, string distance, and lexical diversity, including measures such as n-gram matching and

distributional similarity.

- *Machine-Learned Metrics*: these metrics are based on machine-learned models, used to measure the similarity between two machine-generated texts or between machine-generated and human-generated texts.

This section explores automatic evaluation metrics tailored for text summarization and evaluation metrics that do not require training on specific datasets but are already pre-trained. Following previous research [20], we chose a set of automatic metrics for our experiments, which characteristics follow.

2.3.2. ROUGE

ROUGE [30] stands for Recall-Oriented Understudy for Gisting Evaluation. It measures the number of overlapping textual units (n-grams, word sequences) between the generated summary and a set of reference summaries. In general ROUGE is a family of evaluation measures used to assess the quality of machine-generated text summaries by comparing them to reference (human-generated) summaries. ROUGE is versatile and can be calculated over different methodologies, each focusing on different aspects of the summarization task. Here are some key ROUGE types:

- **ROUGE-N** (N-gram Overlap): measures the overlap of n-grams (sequences of n words) between the generated summary and the reference summary. We have chosen to use unigrams (ROUGE-1, n=1) and bigrams (ROUGE-2, n=2).
- **ROUGE-L** (Longest Common Subsequence): Measures the longest common subsequence between the generated and reference summaries. It considers word sequences that appear in both summaries, regardless of their order.

ROUGE employs three key metrics (precision, recall, and F1 score) to quantitatively evaluate the quality of machine-generated text summaries in comparison to reference summaries.

- *Precision*: measures the accuracy of the generated summary by assessing the ratio of correctly predicted relevant items to the total number of items predicted as relevant. A high precision indicates that when the system predicts a summary to be relevant, it is likely to be accurate. However, precision alone may not capture cases where relevant items are missed.

$$\text{Precision} = \frac{\sum_{S \in \text{references}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \text{references}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \tag{2.1}$$

- *Recall*: evaluates the ability of the generated summary to capture all the relevant information from the reference summary. It is the ratio of correctly predicted relevant items to the total number of relevant items. A high recall suggests that the system is good at identifying relevant information from the reference summary. However, high recall may be achieved at the expense of including irrelevant information.

$$\text{Recall} = \frac{\sum_{S \in \text{references}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \text{generated}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \tag{2.2}$$

- *F1-score*: combines precision and recall into a single metric, providing a balanced assessment of the summarization system's performance. It is the harmonic mean of precision and recall. The F1 score considers both false positives and false negatives, offering a more comprehensive evaluation. It is particularly useful when there is a need to balance precision and recall, as it penalizes extreme values in either metric.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.3}$$

For the implementation of ROUGE metrics, it has been used the python library *rouge-score* maintained by Google [11].

2.3.3. BERTScore

BERTScore [47] is a metric used for evaluating the quality of text generated by natural language processing models, such as text summarization or machine translation. It leverages pre-trained contextualized embeddings from models like BERT (Bidirectional Encoder Representations from Transformers) [18] to compute similarity scores at the token level. Unlike traditional metrics that rely on exact matching or n-gram overlap, BERTScore takes into account the semantic similarity between words.

Reference and candidate sentences are represented using contextual embeddings based on surrounding words, computed by models like BERT. Then the similarity between contextual embeddings of reference and candidate sentences is measured using cosine similarity. Each token in the candidate sentence is matched to the most similar token in the reference sentence, and vice versa, to compute Recall and Precision, which are then combined to calculate the F1 score. To compute BERTScore, it has been used the implementation of the python library *bert-score* [3] specifying the Italian language.

$$P = \frac{1}{|\overline{x}|} \sum_{\overline{x}_j \in \overline{x}} \max_{x_i \in x} x_i^T \overline{x}_j \tag{2.4}$$

$$R = \frac{1}{|x|} \sum_{x_i \in x} \max_{\overline{x}_j \in \overline{x}} x_i^T \overline{x}_j \tag{2.5}$$

$$F1 = 2 \frac{\mathrm{P} * \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{2.6}$$

2.3.4. BLEU

BLEU [34] is an automatic language-independent machine translation evaluation. It stands for "Bilingual Evaluation Understudy" and was originally designed to assess the performance of machine translation systems by comparing their output to human-generated reference translations. BLEU primarily relies on the concept of n-grams, which are sequences of n-words. Differently from previous metrics, BLEU includes a brevity penalty to address the issue of overly short translations. It penalizes generated texts that are significantly shorter than the reference texts. The brevity penalty ensures that the metric does not favor overly concise generated texts that might achieve high precision by simply being shorter.

The BLEU score is calculated as the geometric mean of the modified precision scores and then multiply the result by an exponential brevity penalty factor. It first computes the geometric average of the modified n-gram precisions, p_n, using n-grams up to length N and positive weights w_n summing to one. BP is the brevity penalty and is computed taking into consideration the length of the candidate text c and the effective reference corpus length r. In the baseline of BLEU, it has been used N=*4* and $w_n = \frac{1}{N}$.

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases} \tag{2.7}$$

$$\mathrm{BLEU} = \mathrm{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log \mathrm{p}_n\right) \tag{2.8}$$

In summary, BLEU provides a single numerical score. Higher BLEU scores indicate better agreement between the generated and reference texts, with attention to both precision and brevity. For the implementation of BLEU it has been used the *sacrebleu* python library [12].

2.3.5. BLANC

BLANC [45] is a metric for automatic estimation of document summary quality. It measures the performance boost gained by a pre-trained language model with access to a document summary while carrying out its language understanding task on the document's text. Therefore the BLANC method *does not require human-written reference summaries*, allowing a human-free summary quality estimation.

For the implementation of BLANC it has been used the GitHub repository *blanc* [4]. Two types of BLANC scores were introduced in the paper and are available in the github repo: BLANC-help and BLANC-tune. They are around 90% correlated with each other, so either one can be used in most cases. BLANC-tune is more theoretically principled, but BLANC-help on average correlates the best with human scores, and this is why it has been chosen for our experiments. BLANC-help uses the summary text by directly concatenating it to each document sentence during inference. Thus with BLANC-help, the language model refers to the summary each time it attempts to understand a part of the document text.

2.3.6. chrF

The chrF (character F-score) metric [36] operates at the character level. It focuses on the overlap of character sequences between the reference and the system output.

The chrF metric calculates an F-score, which is the harmonic mean of precision and recall. It uses a variable matching criterion, where it considers a variable number of matching characters between the reference and the system output. This is in contrast to fixed-size n-grams used in metrics like BLEU [34]. Moreover, it penalizes both omissions and insertions of characters in the system output. This makes it more robust to differences in length and word order. β is a parameter that determines the weight between precision and recall. If β=1 they have the same importance. For the implementation of chrF, it has been used the *sacrebleu* library [12].

$$\text{chrF} = \frac{(1+\beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \tag{2.9}$$

2.3.7. METEOR

METEOR (Metric for Evaluation of Translation with Explicit ORdering) [28] evaluates a text by computing a score based on explicit word-to-word matches between the text and a given reference text. Given a pair of strings to be compared, METEOR creates a word alignment between the two strings, which is a mapping between words, such that every word in each string maps to at most one word in the other string. It uses some word-mapping modules to take into consideration stemming, synonyms, and paraphrastic matches. This makes METEOR more flexible in evaluating generated texts that use synonymous words or slightly different word forms.

The word-mapping modules initially identify all possible word matches between the pair of strings. It then identifies the largest subset of these word mappings. If more than one maximal cardinality alignment is found, METEOR selects the alignment for which the word order in the two strings is most similar. Once the final alignment is decided, the METEOR score for the resulting pairing is computed. Based on the number of mapped unigrams found between the two strings (m), the total number of unigrams in the translation (t) and the total number of unigrams in the reference (r), we calculate unigram precision $P = \frac{m}{t}$ and unigram recall $R = \frac{m}{r}$. It then computes a parameterized harmonic mean of P and R:

$$F_{mean} = \frac{P \times R}{\alpha \times P + (1 - \alpha) \times R} \tag{2.10}$$

Precision, recall and F_{mean} are based on single-word matches. But METEOR computes also a penalty (Pen), to take into account the extent to which the matched unigrams in the two strings are in the same word order. The final METEOR score for the alignment between the two strings is calculated as:

$$score = (1 - Pen) \times F_{mean} \tag{2.11}$$

For the implementation of METEOR it has been used the NLTK libraries suite [10].

2.4. Human Evaluation

The limitations of traditional evaluation metrics have led researchers to use human evaluation studies, such as crowdsourcing, to obtain subjective judgments and assess the readability and fluency of the generated summaries. Human evaluation has been the most trusted evaluation method and is used as the gold standard for summarisation evaluation. As shown in some studies [25], the quality criteria used in the human evaluation and the terminology used for describing these criteria had a high degree of variation. Human evaluation research lacks standardized procedures, but some good practices can be followed. When we look at the approaches used for human summarization evaluation [44], they can be broadly classified into two categories:

- **Intrinsic**: the quality of the summary is measured based on the summary itself without considering the original document. We talk about overall quality or quality measured on specific dimensions (eg. fluency, coherence, correctness, etc.).
- **Extrinsic**: measures the quality of the summary through the completion of some task based on the original document.

To define **intrinsic quality**, 6 criteria are used, including *overall quality* and 5 linguistic qualities [17]:

- *grammaticality*: the summary should not have capitalization errors, dictation, or ungrammatical sentences (e.g. fragments, missing components) that make the text difficult to read.
- *non-redundancy*: there should be no unnecessary repetition (entire sentences repeated, facts repeated, repeated use of a noun or noun phrase (e.g., "Bill Clinton") when a pronoun ("he") would suffice).
- *referential clarity*: it should be easy to identify who or what the pronouns and noun phrases refer to. If a person or other entity is mentioned, it should be clear what their role is in the story.
- *focus*: the summary should have a focus; sentences should only contain information that is related to the rest of the summary.
- *structure & coherence*: the summary should be well structured and well organized, and not just a bunch of related information. Sentences should form a coherent body of information on a topic.

In general, intrinsic qualities do not focus on what the text is saying but on how it is saying it. This is why it is not necessary to give any context information, just the summary is enough.

To define the **extrinsic quality** the following criteria are used [17]:

- *summary usefulness*: determines the summary usefulness based on the satisfaction of a specific goal.

- *source usefulness*: examines how useful the reference document (source document) is to satisfy a certain goal.
- *summary informativeness*: measures how much information from the reference document is preserved in the summary.

3 | Goals and Set-Up

The **goal** of our research is to figure out which Large Language Model, from a chosen list of models, is better suited for the generation of summaries of Italian articles to use within the ConWeb framework once the user requests a summary of Webpage content. Being this work related to ConWeb, the beneficiaries of it will be mainly users with visual impairments.

3.1. Chosen Models

The models we concentrated on during this research have been trained using newspaper articles, aligning with the objectives and the nature of the websites we are dealing with. The models taken into consideration in this study are the following:

- **mbart-summarization-mlsum**: this model [9] is a fine-tuned version of facebook/mbart-large-cc25 [31] on mlsum-it dataset [39].
- **mbart-summarization-ilpost**: this model [8] is a fine-tuned version of facebook/mbart-large-cc25 [31] on IlPost dataset [27].
- **BART**: the BART model [29] uses a standard seq2seq/machine translation architecture with a bidirectional encoder and a left-to-right decoder.
- **GPT-3.5**: is the third version of the Generative Pre-trained Transformer developed by OpenAI [38]. It is built on the Transformer architecture [46], which includes attention mechanisms to capture relationships between words in a sequence, and it is pre-trained on a diverse range of internet text data.

All the chosen models are large language models (LLM) built on a Transformer architecture, including the fine-tuned versions of *facebook/mbart-large-cc25*. The first two models are fine-tuned using mBART, a sequence-to-sequence denoising auto-encoder pre-trained for all languages (including Italian), offering a parameter set adaptable to any language pair through fine-tuning. BART itself is a large language model, and specifically the "large-cc25" variant of mBART indicates that it is a larger version designed for better performance. The term "large" in the context of language models typically refers to the number of parameters or weights in the model. Larger models tend to have more parameters, enabling them to capture complex language patterns and perform well across various natural language processing tasks.

The generation of summaries using the first three stated models has been carried out by previous research at the Politecnico di Milano [35], which can be reviewed for more details about the parameter set for the generation. The choice of GPT-3.5 has been stated both by its natural language processing capabilities, and the presence of an OpenAI model in the ConWeb architecture, as it is used in some cases to extract intents and entities from the user's utterance [43]. Starting from these four models, we conducted research to determine which model was the most suitable for our task.

3.2. Chosen Automatic Metrics

To carry out the automatic evaluation, the following 8 traditional metrics have been used: *ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, BLEU, BLANC, chrF,* and *METEOR*. The description of each metric can be found in chapter 2.

3.3. Proposed Dataset

Large-scale summarization datasets are scarce, particularly for low-resource languages like Italian. However, the Politecnico di Milano researchers [35] have gathered data through web-scraping from major Italian news sites, where articles are freely accessible. Our datasets are sourced from various platforms including lPost, Il Foglio, Google News, La Repubblica, and Wikipedia, covering a wide range of domains such as politics, science and technology, sports, and culture. The dataset we are going to use for this research is made of **76 articles** taken from Italian websites, divided into 4 topics:

- *Science*: 18 articles
- *Sports*: 20 articles
- *Politics*: 19 articles
- *Culture*: 19 articles

For each article, one reference summary is present, and 4 summaries have been generated utilizing the previously said models: mbart-summarization-mlsum [9], mbart-summarization-ilpost [8], BART [29], and GPT3.5 [38]. From now on the models are going to be shortly called: *mlsum*, *ilpost*, *bart*, and *gpt*. Examples of generations can be seen in the Appendices.

The Politecnico di Milano researchers collected the articles and provided human-generated references (called from now on *old references*) for each of them. This research is about continuing and expanding the previously made work, to find the best model to use within ConWeb. After a first analysis of the articles, we made a new set of references (from now on called *new references*) for each article. The choice to make a new set of references has been undertaken because the older ones aligned more closely with the style of earlier models (mlsum, ilpost, and bart) and not with the newly added one (gpt). The old references are indeed characterized by being brief and semantically less dense. With the new references, we aimed to create a set of initial references that better reflected the semantic depth of GPT. Consequently, we improved the old references and conducted tests with both sets. For further comprehension of the changes applied, some tuples of the dataset can be seen in the Appendices.

3.3.1. First Analysis of the Dataset

Analyzing the generated summaries from the chosen models, we can easily identify some main features for each of them. The mlsum summaries tend to be very short, ending up being a one-sentence summary most of the time. The ilpost and bart summaries are short as well, but usually not short as the mlsum ones. The bart summaries tend to have some mistakes in the generated text, as bart model has not been finetuned on an Italian dataset. For example, accents are not recognized because they are not present in the English language, it has also apostrophes escape, and sometimes Italian words are substituted with English terms (especially stopwords). The apostrophes escape mistake is present also in the ilpost summaries but in minor presence. Not only that, summaries from ilpost and mlsum have as well mistakes inside, concerning also the semantic meaning of the summary itself, not always truthfully reflecting the content of the original text from the article. On many occasions, the mlsum, ilpost, and bart generated summaries, because of the brevity and the mistakes' presence, end up being nonsensical. From a first analysis, the generations carried out by mlsum have given 20 nonsensical or incomprehensible summaries out of 76. The ilpost model instead has generated 16 unusable summaries, while the bart model, even if characterized by a larger quantity of grammatical mistakes, has generated 6 unusable summaries.

As can be seen from Table 3.2, where the average number of words for each set of articles and model generation is stated, the gpt summaries tend to be way longer than all the other generated summaries. Furthermore, gpt summaries do not contain grammatical mistakes. The information is delivered in a bit longer but easy and comprehensible manner. The only issue presented in the gpt summaries that we have encountered is that when asked to gpt to make summaries of various articles, sometimes gpt took some information from previously asked articles. The articles from which the information is taken are always about the same topic. Indeed within the same set of articles, we have for example more than one article within the Politic set that talks about the war in Ukraine. It happened that after having asked to gpt to make the summary of articles about Ukraine, once asked for another summary over a new article that talks again about Ukraine, it took information from the previous articles requested. This issue happens in a very limited amount of generations (4 out of 76 generations) within the dataset. Even if it is a demonstration of a little lack of consistency of the gpt generation, is important to take into consideration that the fluency, coherence, and relevance demonstrated by gpt at first sight are way higher than the ones of the other models. Also, that inconsistency is at least accurate, which means that the information added (not present in the original article) is correct, even if unrequested. The main features and common mistakes of the models are wrapped up in Table 3.1.

From this first analysis of the dataset, we have concluded that the summaries generated by gpt are the most suitable for our task. Thus we want to show with the following research how automatic metrics (which depict as best model as a different one from gpt) lack in capturing the real goodness of LLMs' generated summaries.

features and common mistakes

model	**nonsensical**	**one-sentence**	**accents**	**apostrophes**	**only italian**
mlsum	20/76	74/76	✓	✓	✓
ilpost	16/76	16/76	✓	×	✓
bart	6/76	6/76	×	×	×
gpt	4/76	0/76	✓	✓	✓

Table 3.1: Main features and mistakes in the generated summaries divided by each model (mlsum, ilpost, bart, gpt). A "✓" symbol symbolizes that the model under consideration has that feature correct within its generations. While a "×" symbol symbolizes that the model makes mistakes on the characteristic taken into account. For example, mlsum and gpt generate apostrophes correctly, while ilpost and bart make use of apostrophes escape.

Average words	CULTURE	POLITICS	SCIENCE	SPORT
original	204.89	190.9	165.27	189.45
old reference	60.15	55.5	49	50.4
new reference	59.94	66.95	64.16	60.65
bad summary	68.89	63.45	58	63.25
mlsum	19.21	17.55	20.38	19.7
ilpost	24	26.75	25.72	26.6
bart	34.52	35.35	35.66	37.6
gpt	57.73	70.8	66	60.65

Table 3.2: Average number of words in the dataset. The information is given for each model generated summaries (mlsum, ilpost, bart, gpt), the original articles, the two sets of references (old and new references), and for the bad summary (made for the human evaluation forms), divided for each set of articles (Culture, Politics, Science, Sport).

3.3.2. Motivation for the new generation of references

As said, we have generated a new set of references for the articles of the dataset. The old references were brief and less semantically dense. Here we want to show with a quick but more in-depth evaluation (that considers Precision, Recall, and F1-Score of each metric) the improvements achieved with the new references. The new references have been designed to be more semantically rich, resembling how gpt generates summaries.

Taking into consideration the previous Politecnico di Milano research [35], the automatic metrics applied to the **old reference** have been ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore. The **average** results obtained under this setup, taking into consideration the **Science** articles, are shown in Tables 3.3, 3.4, 3.5, and 3.6. The light blue colored cells represent the higher result within a column (Precision, Recall, or F1-Score).

old reference

ROUGE-1	Precision	Recall	F1-Score
mlsum	0.7142	0.3163	0.4290
ilpost	0.6139	0.3002	0.3835
bart	0.6748	0.5117	0.5651
gpt	0.3793	0.5222	0.4321

new reference

ROUGE-1	Precision	Recall	F1-Score
mlsum	0.7082	0.2333	0.3460
ilpost	0.6174	0.2370	0.3319
bart	0.6543	0.3676	0.4604
gpt	0.5160	0.5404	0.5235

Table 3.3: Rouge-1 (Precision, Recall, and F1-Score) scores' average results on old and new references for the Science articles, for each model (mlsum, ilpost, bart, gpt).

old reference

ROUGE-2	Precision	Recall	F1-Score
mlsum	0.5776	0.2557	0.3464
ilpost	0.4149	0.1962	0.2549
bart	0.5932	0.4482	0.4942
gpt	0.2027	0.2824	0.2322

new reference

ROUGE-2	Precision	Recall	F1-Score
mlsum	0.4844	0.1550	0.2315
ilpost	0.3842	0.1373	0.1950
bart	0.4719	0.2552	0.3230
gpt	0.2789	0.2951	0.2843

Table 3.4: Rouge-2 (Precision, Recall, and F1-Score) scores' average results on old and new references for the Science articles, for each model (mlsum, ilpost, bart, gpt).

old reference

ROUGE-L	Precision	Recall	F1-Score
mlsum	0.6821	0.3030	0.4103
ilpost	0.5150	0.2497	0.3205
bart	0.6599	0.5020	0.5535
gpt	0.3013	0.4193	0.3451

new reference

ROUGE-L	Precision	Recall	F1-Score
mlsum	0.6143	0.2023	0.3000
ilpost	0.4996	0.1850	0.2607
bart	0.5680	0.3169	0.3974
gpt	0.3925	0.4160	0.4004

Table 3.5: Rouge-L (Precision, Recall, and F1-Score) scores' average results on old and new references for the Science articles, for each model (mlsum, ilpost, bart, gpt).

old reference

BERTScore	Precision	Recall	F1-Score
mlsum	0.8457	0.7376	0.7873
ilpost	0.7950	0.7259	0.7581
bart	0.8476	0.8147	0.8296
gpt	0.7572	0.7984	0.7770

new reference

BERTScore	Precision	Recall	F1-Score
mlsum	0.6938	0.8135	0.7484
ilpost	0.7012	0.7884	0.7416
bart	0.7563	0.8192	0.7857
gpt	0.8136	0.8027	0.8079

Table 3.6: BERTScore (Precision, Recall, and F1-Score) scores' average results on old and new references for the Science articles, for each model (mlsum, ilpost, bart, gpt).

As it can be seen, applying ROUGE metrics and BERTscore to the models' generations compared to the old reference, gpt does not stand out. The average results show that bart and mlsum achieved the best results. Taking into account the F1-Score, which combines both precision and recall into a single metric, we can see that bart is always considered the best one.

Taking now into consideration the **new reference** set, and applying the same metrics as before, we obtained a noticeable improvement of results that better capture the goodness of gpt, but not as much as we would have liked to. This time the references taken into consideration, as said, were way less extractive but the results still do not make gpt particularly stand out. Even if it reaches the best results on the F1-Score of ROUGE-1, ROUGE-L, and BERTScore (Tables 3.3, 3.5, 3.6), the difference with bart (which gave the best results with the old references within the F1-Score column) is very small. On ROUGE-2 instead, gpt is not the best on the F1-Score column (Table 3.4). It seems that, as for our hypothesis of gpt being strongly the best summarization model for our task, the automatic metrics proposed are not reliable. Nonetheless, these metrics do not take into consideration all the qualities previously listed which make a summary a good summary. But to prove that, a further analysis has been carried out in the next chapter.

4 | Technical Evaluation

This chapter aims to present the experiments conducted. Our goal is to select the Large Language Model that better generates summaries specific to our task. The task we are working on is the generation of summaries of articles within Italian Websites, to incorporate in ConWeb. To choose the best LLM for our task, an evaluation of the LLMs' results is needed. We hypothesize, given also the previous literature, that current automatic metrics are unable to fully capture the nuances of summaries generated by Large Language Models, thus being unable to properly evaluate them. To show that, and to choose the best LLM for our task, we have done the following experiments. Our specific hypothesis is that, between the LLMs taken into consideration, the GPT3.5 model gives the best results.

4.1. Automatic Metrics Results

The first step has been retrieving the results of the application of automatic metrics on the generated summaries of the articles. The results, shown in Tables 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, and 4.8, are the scores' averages of various metrics applied to the summaries generated by the LLMs taken into consideration (mlsum, ilpost, bart, and gpt), over all the set of articles of the dataset (Culture, Politics, Science, and Sport), by taking into account both the old reference and the new reference. When a metric included Precision, Recall, and F1-Score, only the F1-Score was taken into account.

The results show how most of the time the summaries generated by gpt and bart fight for the podium. The situation is similar to the one encountered in the quick evaluation of the previous chapter. When taking into account the **old reference**, the bart models show higher scores overall. The gpt model, not only doesn't stand out in that set of evaluations, but it reaches sometimes scores that are lower than the scores obtained with the mlsum and ilpost models (for example on Table 4.2 on the Science column). This behavior is due to the high abstractiveness of the gpt model, which clashes with the extractive feature of the old reference's summaries.

Taking into consideration the **new reference** set of summaries, which have been human-generated in an abstractive way, in the columns where gpt wins overall the value of bart is very near the one of gpt and vice-versa. This performance of a few tenths of difference between the models' scores, can be observed for example in the ROUGE-1 metric (Table 4.1), as well as in ROUGE-2 (Table 4.2) for the Politics column, or ROUGE-L (Table 4.3) for the Politics and Science columns. Again this happens in BERTScore (Table 4.4), BLEU (Table 4.5), and in other metrics as well. The BLANC metric, which does not require a reference summary, shows very low scores overall, giving preference to gpt most of the time.

Within the ROUGE family of metrics, we can observe a drop in performance of gpt higher when it comes to the Sport set of articles. The Sports articles analyzed tend to use sometimes niche language, like referencing more general concepts through a fan-known concept (such as the use of the color of the uniform of the players of a football team to indicate the team itself). The gpt model often replaces those references with the nouns that those references in the original article referred to. This causes a drop in performance when the overlaps of the language used are considered, as happens in metrics like ROUGE where the performance given by the calculation of the shared n-grams is at the base of the metric.

The chrF metric (Table 4.7) and METEOR metric (Table 4.8) are the only metrics that over all the set of articles give to the same model the higher score. Is the case of gpt for the chrF metric, and bart for the METEOR metric.

The question now is, which automatic metric better reflects the real goodness of the generated summaries for our specific task? Is there one that for real reflects that goodness? Is gpt the best choice as we hypothesized? If yes, does chrF, which gives the higher scores to gpt, represent properly the goodness of gpt itself? To answer all these questions, it is shown in the next chapter the results of the human evaluations carried out (which are considered a golden standard).

old reference

ROUGE-1	Culture	Politics	Science	Sport
mlsum	0.2898	0.3019	0.4290	0.3842
ilpost	0.3419	0.4498	0.3835	0.4323
bart	0.5096	0.5156	0.5651	0.4853
gpt	0.4327	0.4180	0.4321	0.3871

new reference

ROUGE-1	Culture	Politics	Science	Sport
mlsum	0.2786	0.2773	0.3460	0.3040
ilpost	0.3222	0.3676	0.3319	0.3725
bart	0.4596	0.4321	0.4604	0.4636
gpt	0.4710	0.5204	0.5235	0.4404

Table 4.1: ROUGE-1 scores' average results using the old and new references, for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

old reference

ROUGE-2	Culture	Politics	Science	Sport
mlsum	0.2019	0.2036	0.3463	0.2940
ilpost	0.2144	0.3406	0.2549	0.2953
bart	0.4081	0.4348	0.4942	0.3699
gpt	0.2144	0.2047	0.2323	0.1566

new reference

ROUGE-2	Culture	Politics	Science	Sport
mlsum	0.1595	0.1549	0.2315	0.2135
ilpost	0.1632	0.2125	0.1950	0.2217
bart	0.3340	0.2697	0.3230	0.3342
gpt	0.2334	0.2734	0.28434	0.1847

Table 4.2: ROUGE-2 scores' average results using the old and new references, for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

old reference

ROUGE-L	Culture	Politics	Science	Sport
mlsum	0.2614	0.2844	0.4103	0.3541
ilpost	0.2870	0.3924	0.3205	0.3619
bart	0.4628	0.4674	0.5535	0.4262
gpt	0.3055	0.2957	0.3451	0.2736

new reference

ROUGE-L	Culture	Politics	Science	Sport
mlsum	0.2275	0.2277	0.3000	0.2688
ilpost	0.2387	0.2617	0.2607	0.2966
bart	0.3937	0.3343	0.3974	0.4015
gpt	0.3489	0.3462	0.4004	0.3076

Table 4.3: ROUGE-L scores' average results using the old and new references, for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

old reference

BERTScore	Culture	Politics	Science	Sport
mlsum	0.7212	0.7257	0.7873	0.7493
ilpost	0.7341	0.7810	0.7581	0.7648
bart	0.7972	0.7996	0.8296	0.7713
gpt	0.7532	0.7511	0.7770	0.7389

new reference

BERTScore	Culture	Politics	Science	Sport
mlsum	0.7116	0.7168	0.7484	0.7175
ilpost	0.7242	0.7468	0.7416	0.7368
bart	0.7800	0.7657	0.7857	0.7680
gpt	0.7746	0.7834	0.8079	0.7651

Table 4.4: BERTScore scores' average results using the old and new references, for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

old reference

BLEU	Culture	Politics	Science	Sport
mlsum	0.1396	0.1239	0.2538	0.2107
ilpost	0.1534	0.2443	0.2013	0.2198
bart	0.3157	0.3446	0.4314	0.3302
gpt	0.2075	0.1860	0.2123	0.1580

new reference

BLEU	Culture	Politics	Science	Sport
mlsum	0.1095	0.0990	0.1577	0.1475
ilpost	0.1233	0.1488	0.1389	0.1572
bart	0.2533	0.2145	0.2690	0.2817
gpt	0.2362	0.2654	0.2945	0.1998

Table 4.5: BLEU scores' average results using the old and new references, for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

old reference & new reference

BLANC	Culture	Politics	Science	Sport
mlsum	0.0685	0.0752	0.1189	0.0913
ilpost	0.0906	0.1254	0.1289	0.1030
bart	0.1438	0.1849	0.2160	0.1814
gpt	0.1629	0.2213	0.2785	0.1252

Table 4.6: BLANC scores' average results using the old and new references, for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

old reference

chrF	Culture	Politics	Science	Sport
mlsum	0.2365	0.2256	0.3406	0.3130
ilpost	0.2663	0.3663	0.3296	0.3557
bart	0.4542	0.4962	0.5576	0.4659
gpt	0.4541	0.4843	0.5093	0.4259

new reference

chrF	Culture	Politics	Science	Sport
mlsum	0.2110	0.1923	0.2451	0.2371
ilpost	0.2400	0.2744	0.2641	0.2816
bart	0.4022	0.3814	0.4059	0.4074
gpt	0.4849	0.5331	0.5426	0.4475

Table 4.7: chrF scores' average results using the old and new references, for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

old reference

METEOR	Culture	Politics	Science	Sport
mlsum	0.3711	0.4076	0.5703	0.5112
ilpost	0.4153	0.5457	0.4684	0.4675
bart	0.5650	0.5810	0.6198	0.4892
gpt	0.3559	0.3032	0.3159	0.2724

new reference

METEOR	Culture	Politics	Science	Sport
mlsum	0.3935	0.3957	0.4975	0.4115
ilpost	0.3932	0.4630	0.4465	0.4014
bart	0.5195	0.4728	0.5465	0.5011
gpt	0.3932	0.4048	0.4445	0.3436

Table 4.8: METEOR scores' average results using the old and new references, for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

5 Human evaluation results by crowdsourcing

To assess the goodness of the generated summaries a human evaluation has been conducted. Incorporating user feedback and conducting user-centered evaluations is crucial for identifying the summarization model that best aligns with the needs of our human-centered task. User studies can provide valuable insights into the usability and effectiveness of different summarization models. The following presents the features and results of the evaluation carried out.

5.1. Crowdsourcing

5.1.1. Prolific

To conduct the human evaluation a crowdsourcing website called Prolific has been used. Prolific has been chosen over other options because when testing Prolific, MTurk, Dynata, CloudResearch, and Qualtrics, research [19] shows that Prolific produces significantly higher data quality on all measures. This includes attention, honesty, comprehension, and reliability. Funds to carry out the human evaluation were provided by Doshisha University [6].

In a previous research [24] a similar work has been conducted comparing crowd ratings with expert ratings and automatic metrics such as ROUGE, BLEU, or BertScore on a German summarization data set. To conduct our experiments we took into account the choices made within that research. The research continued [25] showing that the ideal number of crowd workers to use for each summary is around 7-10 or more. It has also demonstrated that crowd workers can be used instead of experts for the evaluation of summaries.

5.1.2. Construction of the questionnaires

Following the best practices depicted previously in the literature review, a questionnaire has been constructed for each article, obtaining a total of 76 forms.

The questionnaire on each form starts with an explanation of the hypothesis and intents. Afterward, the crowdworker is requested to judge on a 5-point MOS scale [40] (Very Poor, Poor, Barely Acceptable, Good, Very Good) each intrinsic quality for each generated summary, the reference summary, and a custom-made summary. The custom-made summary is realized badly and with the specific request to choose point number 2 for each intrinsic quality. The presence of it was used to discard inattentive crowd workers. Is important to specify that the references taken into account during the human evaluation were the *new references* only, shortly called reference in the following tables (Tables 5.1 and 5.2).

After the questions about intrinsic qualities have been answered, it is the turn for the extrinsic ones. This time it is shown also the original article. The crowd worker is asked to judge on a 5-point MOS scale [40] how much information is kept in each summary (informativeness), and it is also asked to answer a comprehension question tailored to each article. The question has to be answered taking into consideration the specific text presented: firstly the article itself, then each summary (the generated ones, the reference one, and the custom-made one). The question is easily constructed and is followed by 4 answers to choose from, of which only one is correct. The presence of these comprehension questions is due to trying to evaluate the source and summary usefulness, which are part of the extrinsic qualities as well. An example of a form questionnaire can be found in the Appendices.

5.1.3. Conducting the Evaluation

Given a certain budget, our evaluation retrieved 5 evaluations for each article form from crowd workers on Prolific. To avoid issues like crowd workers being biased during the questionnaire, each crowd worker was allowed to answer a maximum of 4 forms. Afterward, it has been asked to well-known reliable people to evaluate the summaries as well, using the same setup uploaded on Prolific (and not allowing each person to answer more than 4 forms as well). In this way, we achieved a minimum number of evaluations for each form (article) of **6**. Some articles are provided with more evaluations.

We obtained a total of 543 answers, of which 485 were accepted and not marked as inattentive. The total number of participants, comprehensive of crowd workers and acquaintances, was 353.

5.1.4. Human Evaluation Results

The obtained results can be seen in Tables 5.1 and 5.2. Table 5.1 contains, divided by article set, the averages of each requested quality (overall quality, grammaticality, non-redundancy, referential clarity, focus, structure&coherence, informativeness). The light blue colored cells highlight the higher values over a quality column for a set of articles. When the higher values for each quality are reached by the same model, the model name is highlighted as well. Table 5.2 contains the number of corrected answers compared to the total number of answers, reached by the summaries generated by a specific model over each set of articles. Those results are shown also for the reference summary and the original article.

As can be easily seen from Table 5.1, on behalf of human evaluation the gpt model reaches the best results compared to the other models over each article's group and each quality. The gpt model even outstands the performance of the reference summaries, except for the informativeness in the Sport set of articles. The reference summaries though, at least, are a close second in all the evaluations. About the informativeness, is worth noticing that gpt outstands highly each other model with an average score over 4, while other models do not even reach a 3 score. This high informativeness of gpt is also shown in the answers to the questions. As can be seen from Table 5.2, gpt's correct answers reach almost always the total number of answers, showing that its outputs contain crucial information from the original articles. The reference summary shows a high number of corrected answers as well. These results, plus the fact that is a close second to gpt in all other qualities' evaluations, demonstrate the goodness of the reference summary realized.

We can also observe how some questions within the original text have been answered wrongly even if the answer was for sure within the text itself (Table 5.2). This can be due to a bit of inattentiveness of some crowd workers. But since inattentive crowd workers have been discarded thanks to the badly custom-made summary test (where it was requested to answer always 2 to each quality) some other considerations can be made. The wrong answer given by the crowd workers might be due to the fact that the text was long and partially hard to follow, demonstrating the usefulness of the generation of summaries for users. The fact that gpt reaches a quantity of corrected answers very near to the original text, not only shows that gpt contains almost always important information within the summary, but also that that information is given to the user clearly and easily, making it very suitable for our task.

CULTURE	overall	grammar	non-red.	clarity	focus	structure	inform.
mlsum	3.0388	3.4361	4.0213	3.2055	3.6103	3.2143	2.5978
ilpost	2.6466	3.1704	3.5476	2.9586	3.0238	2.7807	2.43995
bart	3.3233	3.2381	3.7932	3.5376	3.6666	3.3446	2.948
gpt	4.3032	4.3809	4.2619	4.2895	4.1930	4.3133	4.1581
reference	4.1040	4.0276	4.0752	4.1241	3.9348	4.1103	3.9578
POLITICS	overall	grammar	non-red.	clarity	focus	structure	inform.
mlsum	2.9449	3.3621	3.9474	3.3559	3.6291	3.3170	2.7998
ilpost	2.8822	3.2306	3.6804	3.3308	3.3496	3.1366	2.7415
bart	2.9210	3.0301	3.8258	3.1153	3.3609	2.9975	2.6729
gpt	4.1767	4.2644	3.9586	4.1428	4.0050	4.1366	4.1842
reference	3.7958	3.9261	3.9248	3.8947	3.7619	3.8346	3.9408
SCIENCE	overall	grammar	non-red.	clarity	focus	structure	inform.
mlsum	3.0165	3.4044	3.8251	3.1491	3.6898	3.2493	2.6434
ilpost	2.5602	3.1951	3.4408	2.9461	3.2827	2.8022	2.5160
bart	3.1766	3.1395	3.8019	3.3763	3.5992	3.2681	2.7943
gpt	4.1981	4.2556	4.0502	4.1388	4.0063	4.1121	4.1777
reference	3.7364	3.8796	3.6468	3.9044	3.6753	3.8220	3.8858
SPORT	overall	grammar	non-red.	clarity	focus	structure	inform.
mlsum	3.1136	3.4674	4.0435	3.3488	3.7337	3.4298	2.8251
ilpost	2.3754	2.7895v	3.5022	2.7433	3.0857	2.5849	2.5347
bart	3.0456	3.1310	3.7289	3.3155	3.3738	3.0917	2.9696
gpt	4.0334	4.1610	3.9430	4.0577	3.9927	4.0222	4.0240
reference	3.6591	3.6386	3.7877	3.6312	3.6008	3.6053	4.0984

Table 5.1: Average qualities' scores obtained from the human evaluation. The averages have been calculated for each quality (overall quality, grammaticality, non-redundancy, referential clarity, focus, structure&coherence, and informativeness). The results are shown for each model (mlsum, ilpost, bart, and gpt) and for the reference summary, over each set of articles (Culture, Politics, Science, Sport).

QUESTIONS	CULTURE	POLITICS	SCIENCE	SPORT
mlsum	47/135	53/129	44/133	53/146
ilpost	25/135	49/129	47/133	38/146
bart	33/135	46/129	30/133	75/146
gpt	129/135	113/129	131/133	139/146
reference	125/135	124/129	129/133	141/146
original	130/135	124/129	132/133	143/146

Table 5.2: Number of the correct answers to the questions within the human evaluation. The number of correct answers is shown for each model (mlsum, ilpost, bart, gpt), for the reference summary and the original article. The results are shown separately for each set of articles (Culture, Politics, Science, Sport).

6 | Correlation Analysis

After having collected all the data, this chapter aims to examine the results obtained. To accomplish this we have applied correlation to the automatic and human scores.

6.1. Spearman Correlation

To examine the gathered data, we have employed the Spearman correlation [33], a statistical method used to uncover potential relationships. The Spearman rank-order correlation coefficient is a nonparametric measure of the monotonicity of the relationship between two datasets. Like other correlation coefficients, this one varies between -1 and +1:

- *zero* → implies no correlation, or a curvilinear but not monotonic relationship.
- *positive values* → positive correlations signify that high ranks of one variable tend to coincide with high ranks of the other variable.
- *negative values* → negative correlations signify that high ranks of one variable tend to coincide with low ranks of the other variable.
- correlations of *-1* or *+1* imply an exact monotonic relationship.

Besides the Spearman correlation, it is usually calculated also the p-value. The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Spearman correlation high. Although the calculation of the p-value does not make strong assumptions about the distributions underlying the samples, it is only accurate for very large samples (>500 observations). Thus, it has not been taken into consideration during this analysis.

The outcomes of the Spearman correlation analysis will shed light on the effectiveness of the chosen metrics in capturing the nuances of summary quality as perceived by human evaluators. A high correlation would indicate that the metrics align well with human judgments, reinforcing their utility in automated model evaluation. On the other hand, a weaker correlation would prompt a reconsideration of the chosen metrics or indicate potential limitations in their ability to capture certain aspects of the summary's quality. This approach allows us to move beyond traditional metrics and gain valuable insights into the concordance between automated assessments and human perceptions of the summary's quality.

6.2. Correlation with automatic metrics

To understand if the metrics had some sort of connection with the real judgments of the summaries, obtained by human evaluation (considered to be the golden standard), we decided to compute the correlation to correlate the quality scores obtained from the human evaluation with the automatic metrics applied to the summaries generated by each model. The goal is to discern whether there exists a systematic relationship between the subjective evaluations by human raters and the objective metrics assessing summaries' quality.

The following figures (Figures 6.1, 6.2, 6.3, 6.4, 6.5, 6.6, and 6.7) show the correlation of each quality (intrinsic and extrinsic) with each metric applied. The correlation has been calculated taking into account the metrics calculated both with the old references and the new references.

As we can see from the figures, the correlations do not surpass the 0.25 correlation score, and some of them are even negative. Taking into account previous research [45], we would have expected a human correlation of at least 0.30 for good summaries. But, those summaries that have been considered way better from the human evaluation (which are the ones generated by gpt), have not received significantly higher scores in the automatic metrics. Not only that, the low correlation shows how those results do not even reflect properly the human judgements.

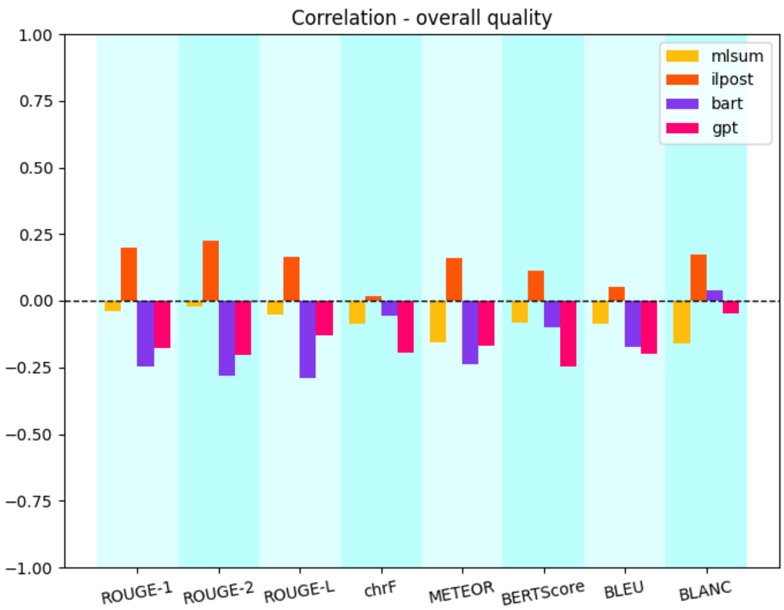
The use of the new reference set, which better represents a good example of abstractive generation, has indeed brought a better performance of the automatic metrics. But, being human evaluation considered a golden standard, and given the low scores of the correlation, it is clear that automatic metrics do not reflect properly the goodness of the summaries generated by the LLMs. We cannot say that the automatic metrics taken into account reflect properly the human judgements for the summarization of Italian articles.

Briefly, what we have just stated is that the correlation scores between automatic metrics and human evaluation are low. Given the fact that human evaluation is a golden standard, if the correlation between human judgment and automatic evaluations is so low it means that the latter does not accurately reflect the quality of a summary. This outcome let us understand that there is no link between human judgment and the scores of the automatic metrics obtained. Moreover, it strengthens our thesis for which traditional automatic metrics result to be unsuitable for summarization evaluation. For further comprehension, let's analyze the correlation plots.

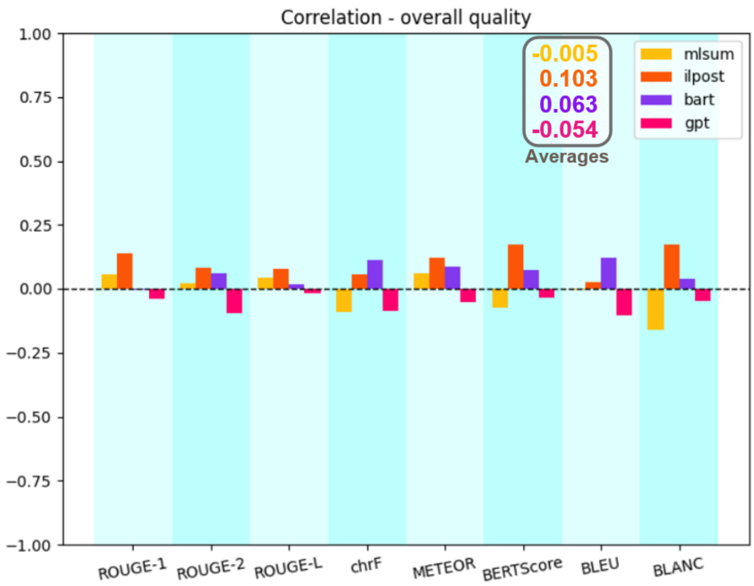
From the plots, it is easy to see that the correlations that we obtain between each quality and the various metrics are not strong. They appear to be very irregular, but more importantly with very low values, sometimes even negative. This happens both with the metrics over the old references and the metrics over the new references.

As for the overall quality, which is the quality that should enclose all the other ones, within the **old references** we can observe that the higher correlation is reached by the summaries generated by ilpost. More abstractive models like bart and gpt, show a quite strong negative behavior. The extractiveness of the old reference summaries tends to clash with human judgments on this end. In general, the highest correlation results with the old references are reached by ilpost model over the informativeness quality. Surprisingly, the informativeness quality reaches a positive correlation over all the metrics with gpt as well, while bart and mlsum fall on negative values. The quality that negatively correlated the most over the old references is the non-redundancy.

Taking into account the **new reference** sets of summaries within the overall quality correlation, on the top right corner of the plot in Figure 6.1 are reported the mean scores of correlation reached by the models over the various metrics. We can see from the average values that the models that reach a higher correlation are bart and ilpost. While the model that reaches the lower correlation is gpt. In general, compared to the situation obtained by the old references, the values of bart and gpt, gained points, making bart even not negative. The other correlations give a similar situation. Apart from the informativeness, the behavior of the correlations over the ilpost and bart summaries are almost always positive. The informativeness situation changes completely when passing from the old references to the new references. In fact, in the latter most correlation values are negative, over all the models. In general the correlation depicted within the new references show a low span of values. Correlations that were high before have a tendency to reduce their value, while correlations that were negative tend to increase their value.

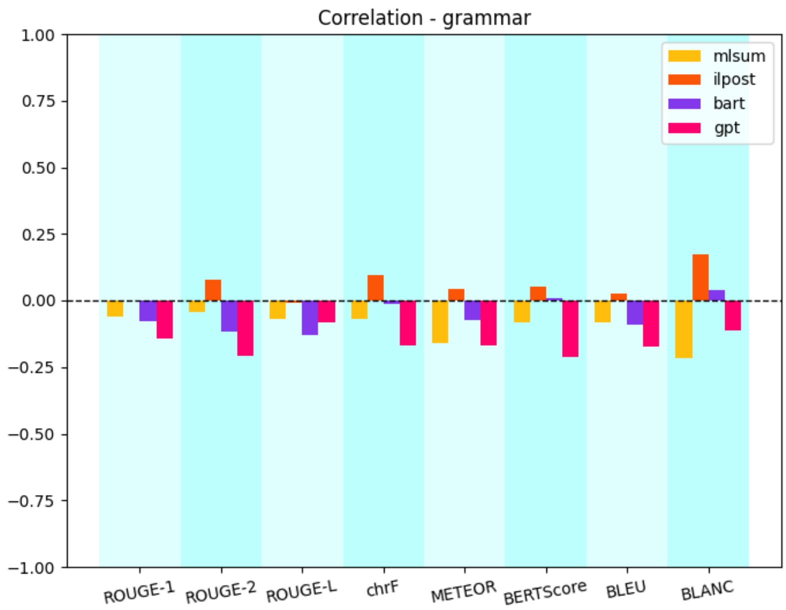


(a) Taking into account the **old reference**.

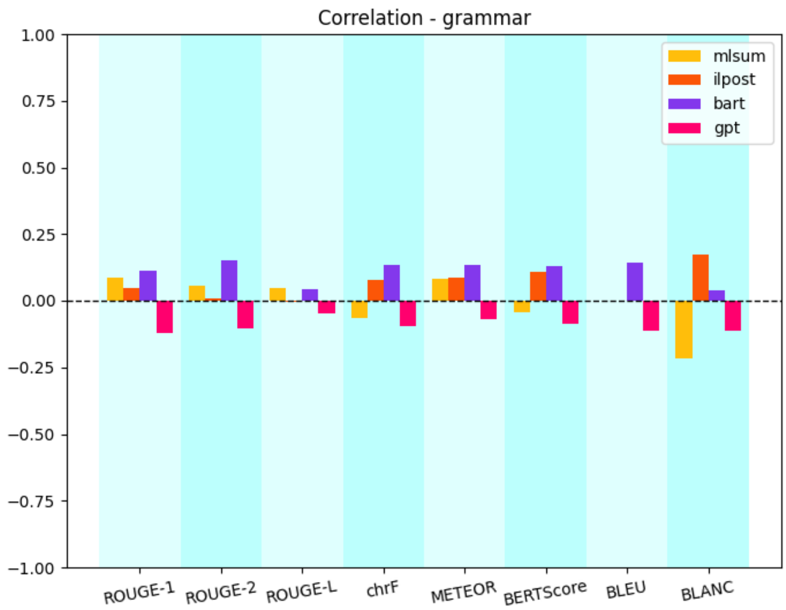


(b) Taking into account the **new reference**.

Figure 6.1: Spearman Correlation over all the dataset, between the **overall quality** scores (obtained by crowdsourcing) and each automatic metric score (within each model summary, calculated over both the old references and the new references). In the upper right corner for the correlations with the new references are present the averages of the correlations over the various models.

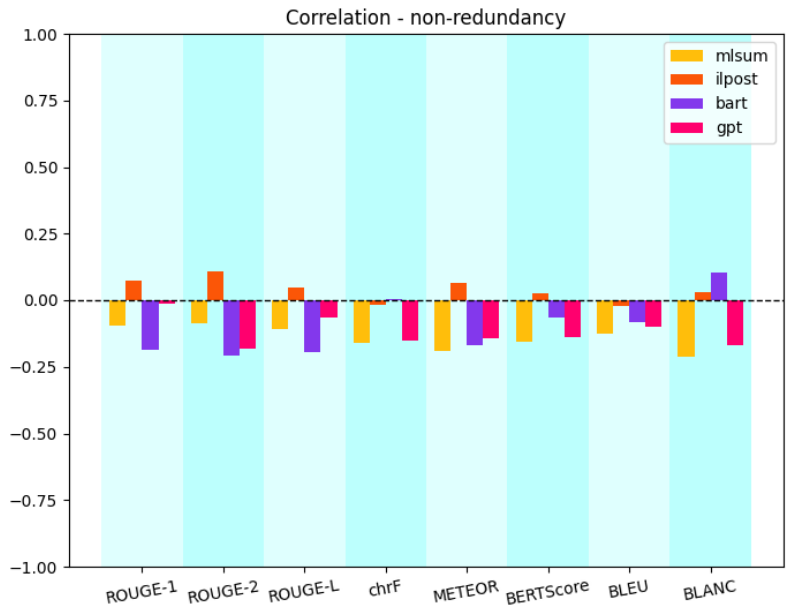


(a) Taking into account the **old reference**.

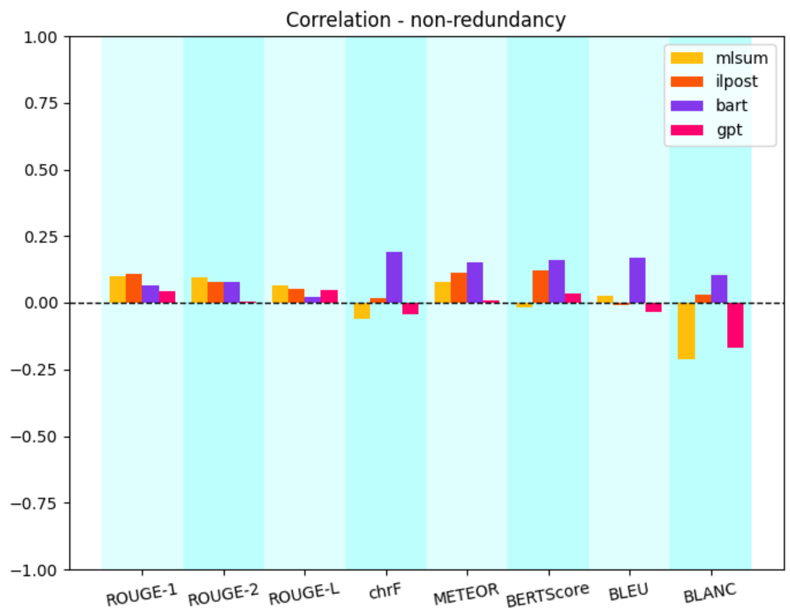


(b) Taking into account the **new reference**.

Figure 6.2: Spearman Correlation over all the dataset, between the **grammaticality** scores (obtained by crowdsourcing) and each automatic metric score (within each model summary, calculated over both the old references and the new references).

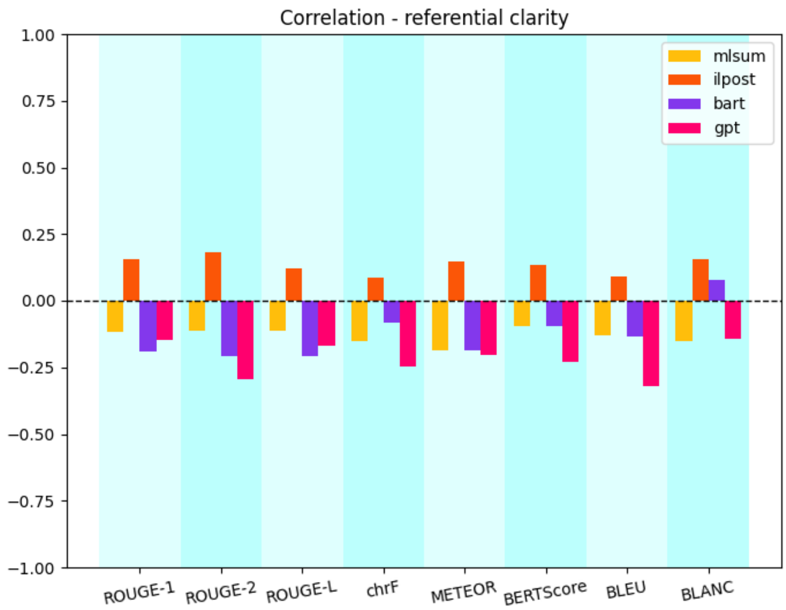


(a) Taking into account the **old reference**.

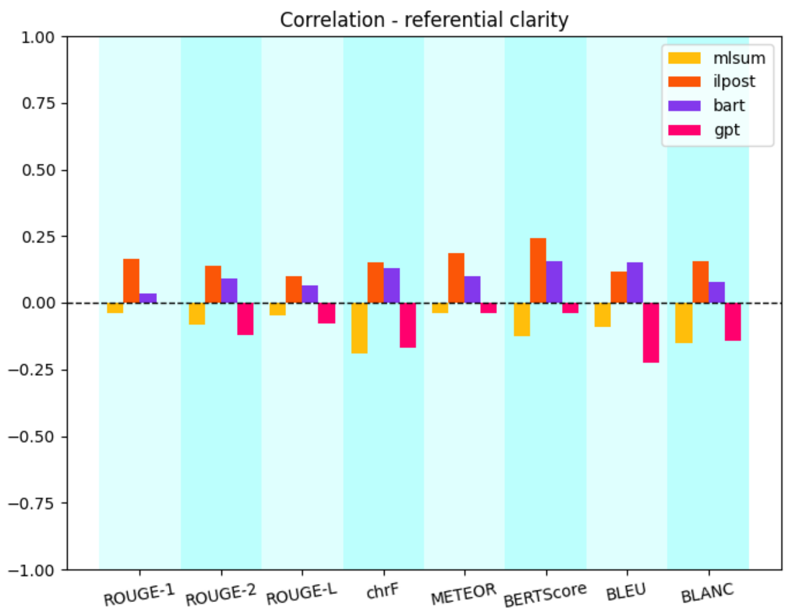


(b) Taking into account the **new reference**.

Figure 6.3: Spearman Correlation over all the dataset, between the **non-redundancy** scores (obtained by crowdsourcing) and each automatic metric score (within each model summary, calculated over both the old references and the new references).

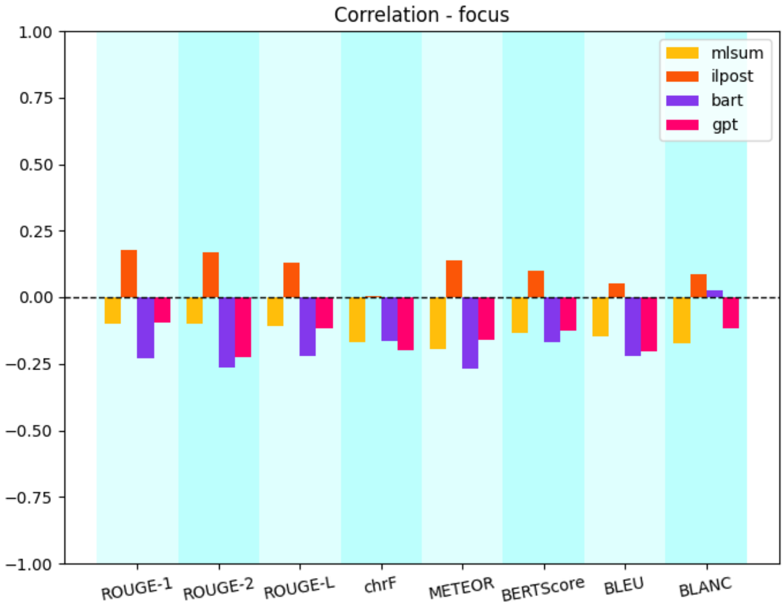


(a) Taking into account the **old reference**.

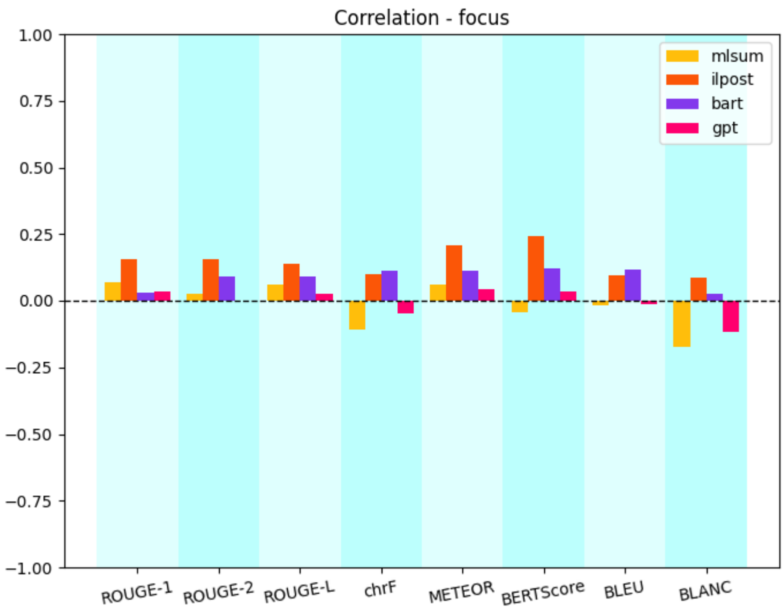


(b) Taking into account the **new reference**.

Figure 6.4: Spearman Correlation over all the dataset, between the **referential clarity** scores (obtained by crowdsourcing) and each automatic metric score (within each model summary, calculated over both the old references and the new references).

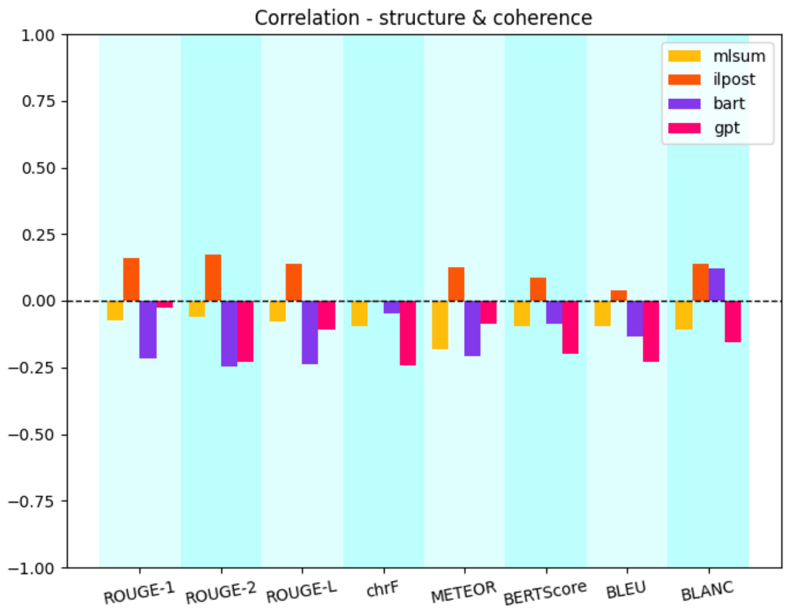


(a) Taking into account the **old reference**.

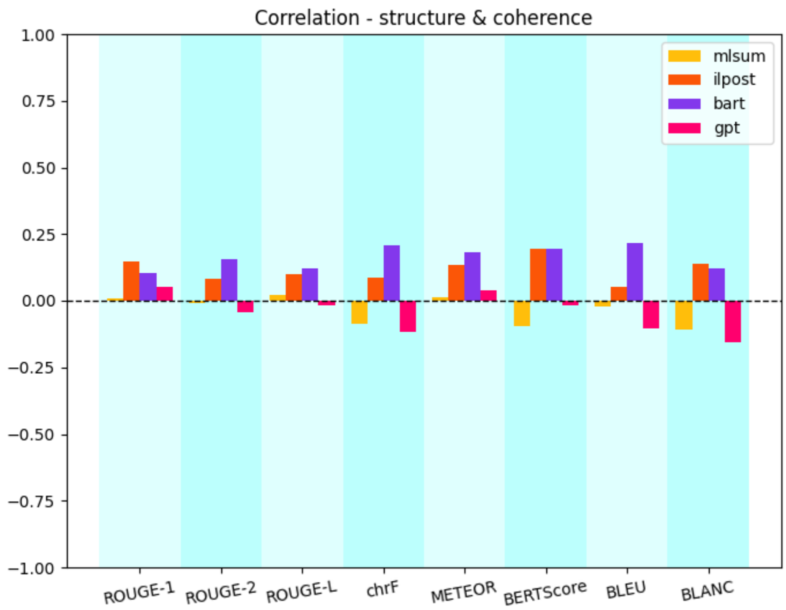


(b) Taking into account the **new reference**.

Figure 6.5: Spearman Correlation over all the dataset, between the **focus** scores (obtained by crowdsourcing) and each automatic metric score (within each model summary, calculated over both the old references and the new references).

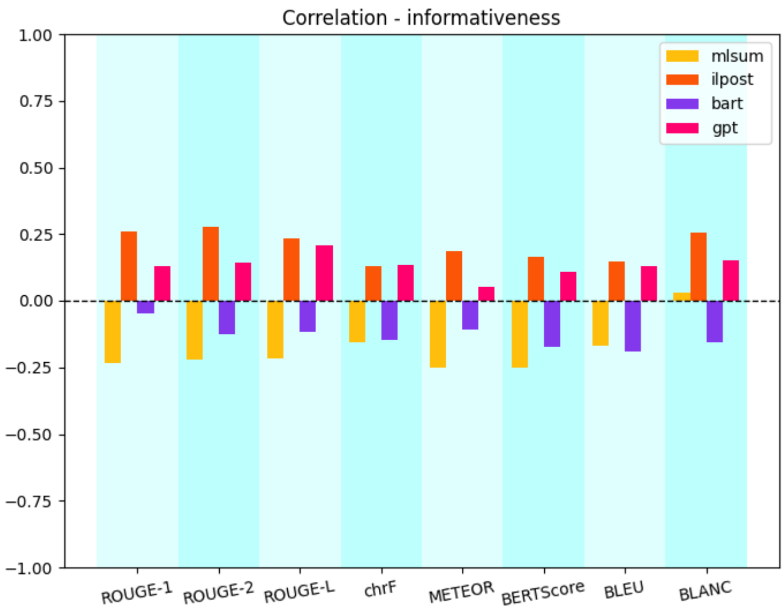


(a) Taking into account the **old reference**.

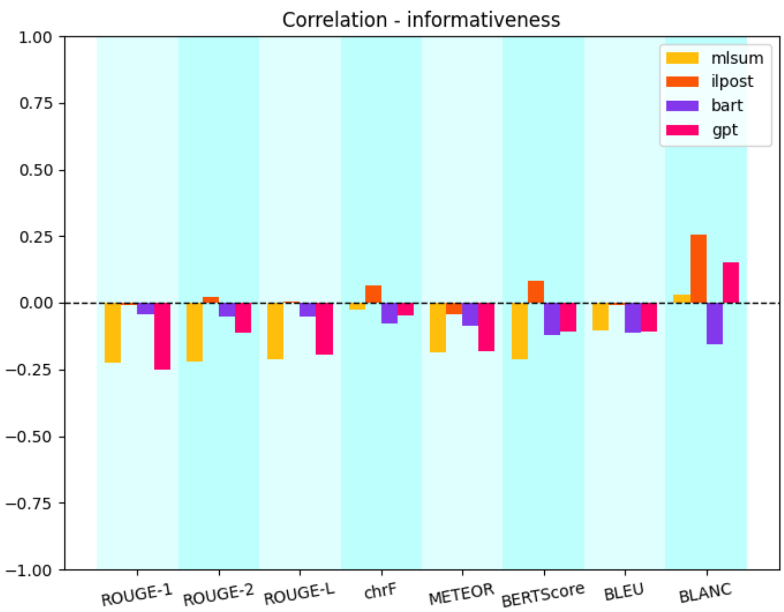


(b) Taking into account the **new reference**.

Figure 6.6: Spearman Correlation over all the dataset, between the **structure & coherence** scores (obtained by crowdsourcing) and each automatic metric score (within each model summary, calculated over both the old references and the new references).



(a) Taking into account the **old reference**.



(b) Taking into account the **new reference**.

Figure 6.7: Spearman Correlation over all the dataset, between the **informativeness** scores (obtained by crowdsourcing) and each automatic metric score (within each model summary, calculated over both the old references and the new references).

6.3. Correlation with the other qualities

In this section are shown the correlations between each quality. This calculation has been conducted to study if some qualities were particularly correlated to other qualities within the human judgment. The results are shown in Figures 6.8, 6.9, 6.10, 6.11, 6.12, 6.13, and 6.14.

Human judgments show that the overall quality is well described from the other requested qualities, and follows their trend. The only quality that has less correlation with the overall quality (very low one when it comes to gpt), as well as the other qualities, is informativeness. Structure & coherence is the quality that mostly correlates with the overall quality. Following we find a high correlation with the referential clarity, and then with the focus and grammar. The structure & coherence shows also an overall stronger correlation with the referential quality.

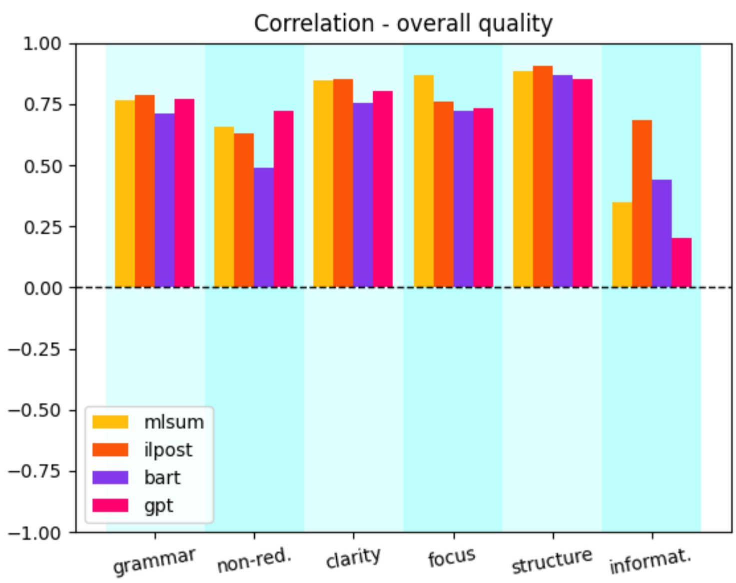


Figure 6.8: Spearman Correlation over all the dataset, between the **overall quality** scores and each other quality score (obtained by crowdsourcing).

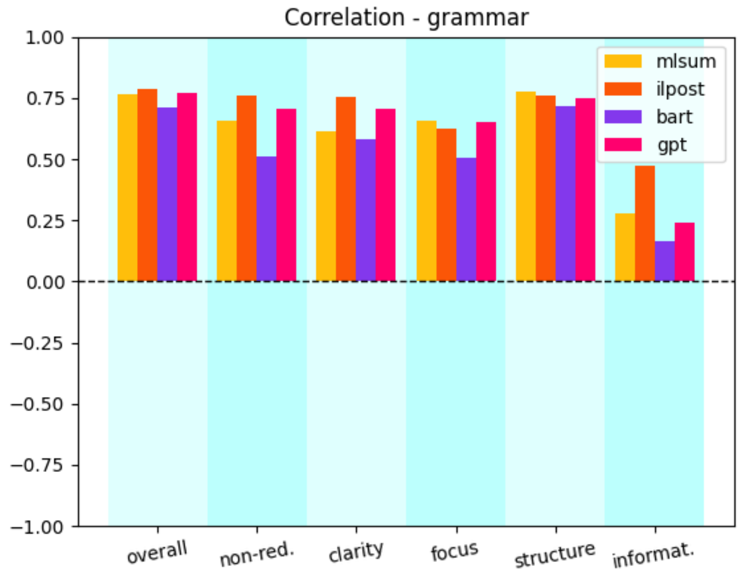


Figure 6.9: Spearman Correlation over all the dataset, between the **grammaticality** scores and each other quality score (obtained by crowdsourcing).

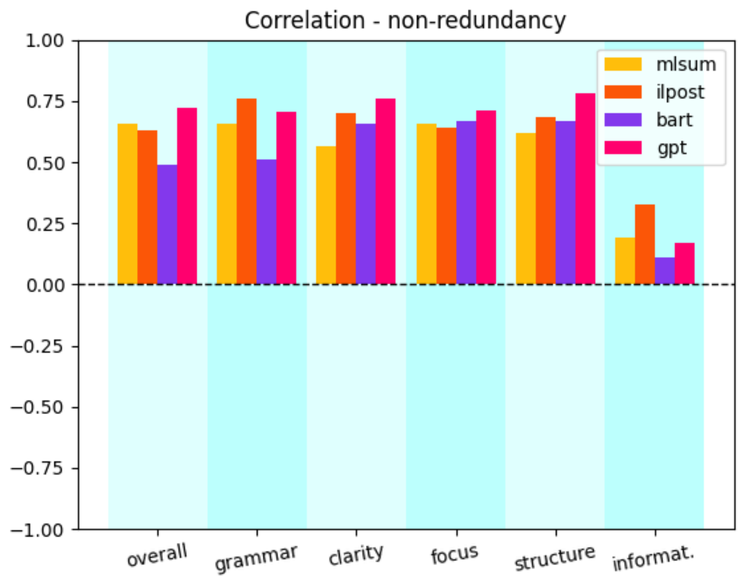


Figure 6.10: Spearman Correlation over all the dataset, between the **non-redundancy** scores and each other quality score (obtained by crowdsourcing).

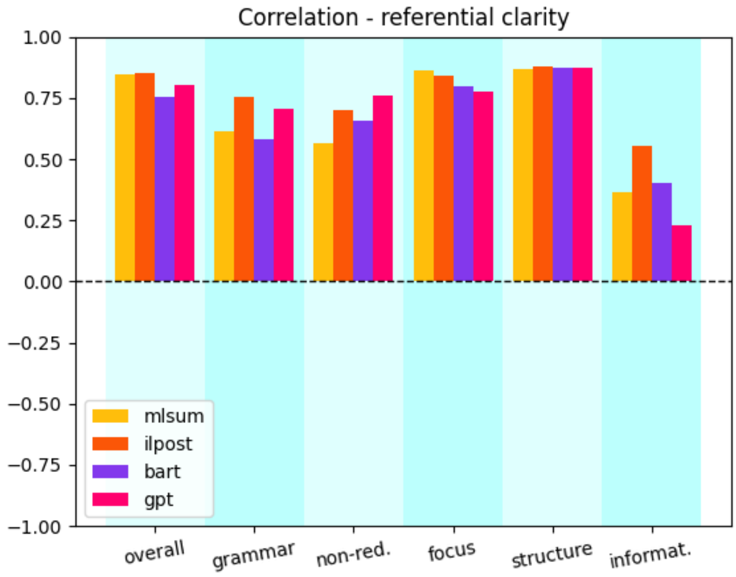


Figure 6.11: Spearman Correlation over all the dataset, between the **referential clarity** scores and each other quality score (obtained by crowdsourcing).

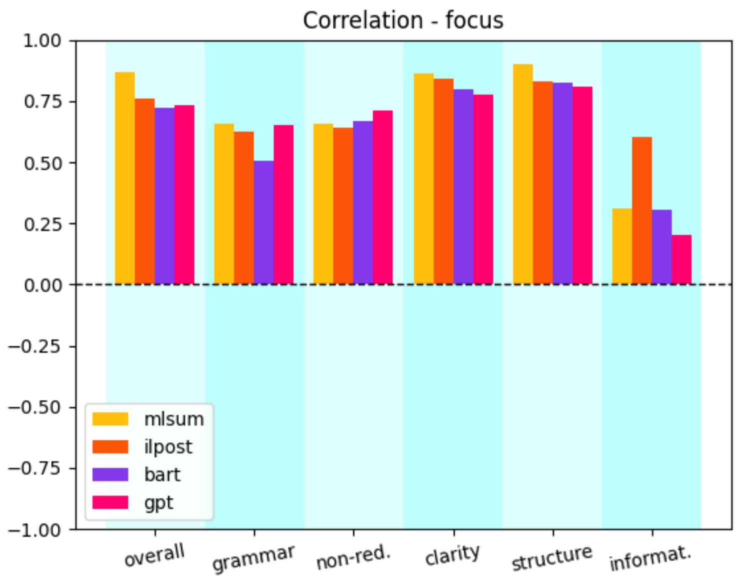


Figure 6.12: Spearman Correlation over all the dataset, between the **focus** scores and each other quality score (obtained by crowdsourcing).

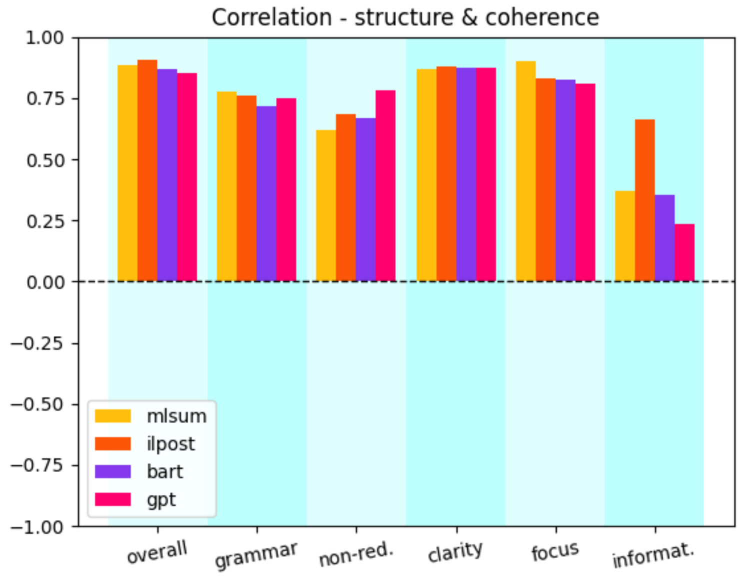


Figure 6.13: Spearman Correlation over all the dataset, between the **structure & coherence** scores and each other quality score (obtained by crowdsourcing).

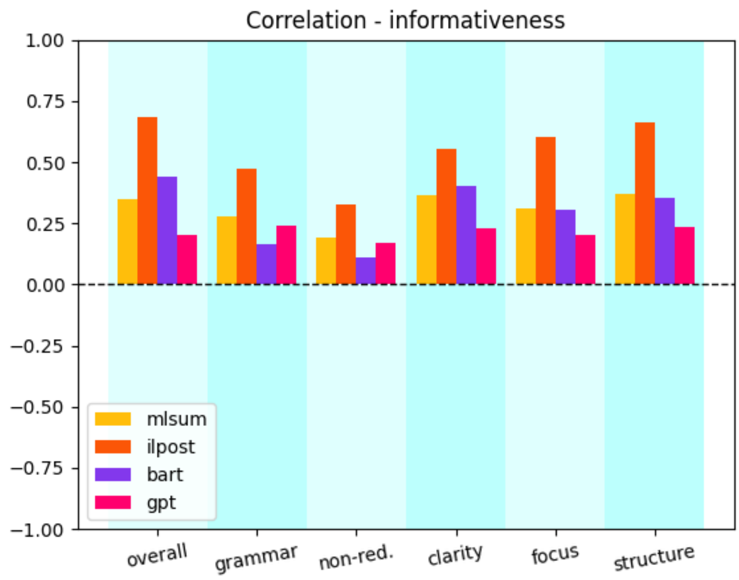


Figure 6.14: Spearman Correlation over all the dataset, between the **informativeness** scores and each other quality score (obtained by crowdsourcing).

6.4. Enhancement attempt

This section shows an attempt to reach better results from the automatic metrics application, trying to take into consideration that the texts that we are analyzing are based on the Italian language.

6.4.1. Preprocessing structure

Considering the low results obtained from using automatic metrics to evaluate model summaries, we decided to try to see if applying an Italian language tailored preprocessing would improve the results. This decision comes from the understanding that automatic metrics present in literature have been constructed primarily for the English language. With the application of preprocessing, we wanted to check if a significant improvement could be seen.

The preprocessing follows the following steps:

1. Lowercasing of all the words.
2. Removal of articles, punctuation, and apostrophes.
3. Removal of stopwords (taken from the NLTK library and a collection found in a GitHub repository [7]).
4. Stemming application using *SnowballStemmer('italian')* from the NLTK library [2].

6.4.2. Results with preprocessing

The results obtained by the preprocessing application and the calculation of automatic metrics with the new reference set can be seen in Tables 6.1, 6.2, 6.3, 6.4, 6.5, 6.6, 6.7, and 6.8. The results show an improvement of a few tenths compared to the results obtained previously (available in chapter 4), which have been considered not significantly high to conduct more research on this end.

ROUGE-1	Culture	Politics	Science	Sport
mlsum	0.2615	0.2858	0.3398	0.3024
ilpost	0.3025	0.3721	0.3364	0.3558
bart	0.4705	0.4465	0.4870	0.4545
gpt	0.4742	0.5056	0.5716	0.4446

Table 6.1: ROUGE-1 scores' average results with preprocessing, using the new references. The averages have been calculated for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

ROUGE-2	Culture	Politics	Science	Sport
mlsum	0.1575	0.1677	0.2091	0.1968
ilpost	0.1370	0.1943	0.1722	0.1870
bart	0.3151	0.2706	0.3291	0.3083
gpt	0.2233	0.2621	0.2956	0.1556

Table 6.2: ROUGE-2 scores' average results with preprocessing, using the new references. The averages have been calculated for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

ROUGE-L	Culture	Politics	Science	Sport
mlsum	0.2391	0.2528	0.3155	0.2843
ilpost	0.2522	0.2802	0.2716	0.3073
bart	0.4155	0.3715	0.4349	0.4144
gpt	0.3806	0.3752	0.4760	0.3445

Table 6.3: ROUGE-L scores' average results with preprocessing, using the new references. The averages have been calculated for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

BERTScore	Culture	Politics	Science	Sport
mlsum	0.7036	0.6999	0.7256	0.7023
ilpost	0.6995	0.7248	0.7202	0.7191
bart	0.7723	0.7530	0.7699	0.7566
gpt	0.7619	0.7705	0.7868	0.7428

Table 6.4: BERTScore scores' average results with preprocessing, using the new references. The averages have been calculated for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

BLEU	Culture	Politics	Science	Sport
mlsum	0.1275	0.1327	0.1675	0.1567
ilpost	0.1284	0.1846	0.1641	0.1696
bart	0.2886	0.2668	0.3194	0.2999
gpt	0.2907	0.3286	0.3620	0.2301

Table 6.5: BLEU scores' average results with preprocessing, using the new references. The averages have been calculated for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

BLANC	Culture	Politics	Science	Sport
mlsum	0.0785	0.0720	0.0852	0.0783
ilpost	0.0785	0.1014	0.0894	0.0935
bart	0.1741	0.1698	0.1931	0.1617
gpt	0.1859	0.1939	0.2430	0.1237

Table 6.6: BLANC scores' average results with preprocessing, using the new references. The averages have been calculated for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

chrF	Culture	Politics	Science	Sport
mlsum	0.2234	0.1983	0.2395	0.2375
ilpost	0.2274	0.2687	0.2510	0.2756
bart	0.4107	0.3818	0.4129	0.4004
gpt	0.4836	0.5279	0.5416	0.4388

Table 6.7: chrF scores' average results with preprocessing, using the new references. The averages have been calculated for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

METEOR	Culture	Politics	Science	Sport
mlsum	0.3716	0.4416	0.5187	0.4447
ilpost	0.3755	0.4734	0.4408	0.4417
bart	0.5348	0.4925	0.5603	0.5144
gpt	0.3958	0.4109	0.4661	0.3497

Table 6.8: METEOR scores' average results with preprocessing, using the new references. The averages have been calculated for each model (mlsum, ilpost, bart, gpt), divided by set of articles (Culture, Politics, Science, Sport).

7 | Conclusions and future developments

In the following sections, we will delve into the primary outcomes of our work thus far. We will also give a little prompt for possible future studies based on newly studied approaches that are coming in the literature.

7.1. Summary and Conclusions

7.1.1. Final Considerations

The analysis carried out showed clear results. Our findings can be summarized as follows:

- The evaluation with traditional metrics depicted an unsure picture of which model generates better summaries for Italian Web articles. The gpt and bart model are preferred, but their scores are characterized by a lot of fluctuations between the two over the different sets of articles, with a difference in scores not high.
- The preprocessing for the Italian language didn't help to reach a significant improvement in the automatic metrics application.
- The human evaluation states that gpt gains unequivocally the higher scores compared to all the other models.
- The correlation analysis revealed a weak correlation with the automatic metrics, particularly regarding the gpt summaries, which, according to the human evaluation results, best reflect user preferences. This outcome strengthens our thesis: a strong discrepancy exists between human and automatic evaluations. In fact, as human evaluation serves as the golden standard, the highlighted divergence underscores the inadequacy of traditional automatic metrics in summarization evaluation.

In the evaluations with traditional metrics, even when gpt had the highest overall score, the gap with the scores obtained by the other models was not high. Also, even if sometimes gpt obtained the highest score, most of the time it did not obtain it over all the sets considering the same metric. Gpt obtained the highest scores over all the sets only with the chrF metric, which, though, has been shown to have a negative correlation over all the qualities on gpt. Indeed the correlation analysis carried out showed a poor correlation between the evaluation with automatic metrics and the human evaluation, especially for high abstractive summaries. This means that traditional metrics cannot be trusted on our specific task.

We can conclude that the automatic metrics used are not suitable for evaluating summaries of Italian articles. Human judgment allowed us to establish that gpt output summaries are more comprehensible, easy to follow, and correct. The gpt summaries have been clearly preferred over all the other summaries presented, which makes GPT-3.5 the most suitable model for ConWeb between the models taken into account.

7.1.2. Lesson Learned

Based on the experiments conducted and the resulting outcomes, it's evident that human evaluation doesn't confirm the findings of traditional automatic metrics, making them unsuitable for evaluating summaries. From these findings, certain criteria emerge which a metric should adhere to in order to be deemed suitable for summary evaluation:

- Semantics play a crucial role. Many conventional automatic metrics primarily focus on aligning text spans with a reference summary. However, language encompasses the ability to express the same idea using various words and perspectives. Therefore, it's imperative for a metric to consider semantics rather than mere overlap in its assessment. This involves acknowledging concepts such as synonyms, antonyms, inflectional variants, and paraphrases.
- Continual updates to metrics are essential to keep pace with evolving idioms and language variations. For instance, within our dataset, a football team was identified by the color of its players' uniforms, highlighting the need for metrics to adapt to such nuances.
- Evaluation accuracy could significantly improve by incorporating multiple reference summaries for the same generated summary. These reference summaries may encompass diverse styles or simply vary through the use of synonyms, thereby accounting for some of the variations during evaluation.
- Alternatively, instead of utilizing multiple reference summaries, directly referring to the original article could be beneficial.

In light of these considerations, advancements in generative Large Language Models (LLMs) present a promising avenue. Leveraging the capabilities of LLMs for evaluating other LLMs eliminates the need for reference summaries and offers a viable alternative. Transitioning towards LLM-based evaluations offers numerous advantages. This approach not only streamlines the evaluation process but also mitigates issues such as crowd worker inattention (that we have experienced prominently during human evaluations), thereby reducing assessment costs and time expenditure.

7.2. New developments and future studies

7.2.1. Current trends on Summarization Evaluation

At the time of ending of this research, new approaches to summary evaluation are being developed. As previously stated, evaluating abstractive summaries is challenging. Within an evaluation is important to take into consideration different aspects like fluency, coherence, relevance, and consistency [26]. But issues like hallucination must be taken into consideration as well. The metrics explored in this research have been mostly reference-based, which means that the generation of the model is being evaluated by taking into consideration a reference output (or more than one), usually human-written. In the task of summarization, references are well-written summaries that would be considered a good generation output for a model. Collecting human-written references is an extremely time-consuming and expensive task. Furthermore, that kinds of approach have many limitations on understanding properly the goodness of the models' outputs, which are reflected in the poor correlation with human judgments.

To tackle these issues other approaches are being explored, which are **context-based** metrics. Context-based metrics do away with references completely and evaluate based on the context (i.e., source document) instead. The newly studied approach is to evaluate LLMs with LLMs. One such approach is **GPTScore** [21] which utilizes generative pre-trained models to score generated texts. The GPTScore evaluation framework is customizable, multifaceted (one evaluator performs multifaceted evaluations), and training-free. The idea behind the framework is that better quality text (in terms of meeting some desired criteria) is more likely to be generated when the context is clear. It measures this likelihood using conditional generation probability. To figure out what users really want, it first establishes a set of rules for evaluation, based on how the text is supposed to be generated and what aspects they care about, such as clarity or coherence. Then, examples are provided to help the model learn what to look for. Finally, the estimation of how likely it is for the text to meet the defined evaluation rules is calculated. GPTScore is a way to measure how well the text aligns with what the user wants, based on the rules they've set up.

Another similar approach is seen in **G-Eval** [32], which utilizes Large Language Models (LLMs) along with Chain-of-Thought (CoT) and a form-filling method to assess LLM outputs First, they provide evaluation criteria to the LLM, prompting it to generate a CoT for evaluation steps. Then, to evaluate coherence in news summarization, they concatenate the prompt, CoT, news article, and summary and ask the LLM to output a score between 1 to 5. The scores are normalized using token probabilities and weighted to produce a final result. G-Eval found that using GPT-4 as an evaluator yielded a high Spearman correlation of 0.514 with human judgments, outperforming prior methods.

7.2.2. Future Works

Future studies could try to use these newly explored approaches over the Italian Articles collected and previously evaluated. It would be interesting to check if the high correlation achieved in other languages by using those new evaluation approaches can be achieved for our task (based on the Italian language) as well. Also being those approaches based on contexts and not references, they allow to fine tune large language models in a faster way. Web scraping of Italian web pages, without the time-consuming task of human reference generation, would allow for a quicker gathering of Italian articles to use to fine-tune Large Language Models on our specific task. The generated summaries could be used to carry out a new human evaluation to correlate with the scores obtained with the context-based metrics.

Bibliography

[1] Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs — databricks.com. https://www.databricks.com/blog/mpt-7b. [Accessed 24-02-2024].

[2] NLTK :: nltk.stem.SnowballStemmer — nltk.org. https://www.nltk.org/api/nltk.stem.SnowballStemmer.html?highlight=stopwords. [Accessed 25-02-2024].

[3] bert-score — pypi.org. https://pypi.org/project/bert-score/. [Accessed 24-02-2024].

[4] GitHub - PrimerAI/blanc: Human-free quality estimation of document summaries — github.com. https://github.com/PrimerAI/blanc. [Accessed 24-02-2024].

[5] Fast Facts of Common Eye Disorders | CDC — cdc.gov. https://www.cdc.gov/visionhealth/basics/ced/fastfacts.htm. [Accessed 24-02-2024].

[6] Doshisha university. https://www.doshisha.ac.jp/en/. [Accessed 13-03-2024].

[7] GitHub - stopwords-iso/stopwords-it: Italian stopwords collection — github.com. https://github.com/stopwords-iso/stopwords-it. [Accessed 25-02-2024].

[8] ARTeLab/mbart-summarization-ilpost · Hugging Face — huggingface.co. https://huggingface.co/ARTeLab/mbart-summarization-ilpost, . [Accessed 25-02-2024].

[9] ARTeLab/mbart-summarization-mlsum · Hugging Face — huggingface.co. https://huggingface.co/ARTeLab/mbart-summarization-mlsum, . [Accessed 25-02-2024].

[10] NLTK :: nltk.translate.meteorscore module — nltk.org. https://www.nltk.org/api/nltk.translate.meteor_score.html. [Accessed 24-02-2024].

[11] rouge-score — pypi.org. https://pypi.org/project/rouge-score/. [Accessed 24-02-2024].

[12] sacrebleu — pypi.org. https://pypi.org/project/sacrebleu/. [Accessed 24-02-2024].

[13] Vision impairment and blindness — who.int. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment. [Accessed 24-02-2024].

[14] L. Caponi. Web augumentation for coordinating conversational and visual experiences on the web, 2022.

[15] A. Celikyilmaz, E. Clark, and J. Gao. Evaluation of text generation: A survey, 2021.

[16] N. Conte. Ranked: The Most Popular AI Tools — visualcapitalist.com. https://www.visualcapitalist.com/ranked-the-most-popular-ai-tools/. [Accessed 24-02-2024].

[17] H. T. Dang. Overview of duc 2005. 2005. URL https://api.semanticscholar.org/CorpusID:61825275.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL https://arxiv.org/abs/1810.04805.

[19] P. Eyal, R. David, G. Andrew, E. Zak, and E. Damer. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54, 09 2021. doi: 10.3758/s13428-021-01694-3.

[20] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev. Summeval: Re-evaluating summarization evaluation, 2021.

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

[22] K. Ganesan, C. Zhai, and J. Han. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. volume 2, 12 2010.

[23] T. Goyal, J. J. Li, and G. Durrett. News summarization and evaluation in the era of gpt-3, 2023.

[24] N. Iskender, T. Polzehl, and S. Möller. Best practices for crowd-based evaluation of German summarization: Comparing crowd, expert and automatic evaluation. In S. Eger, Y. Gao, M. Peyrard, W. Zhao, and E. Hovy, editors, *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 164–175, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.eval4nlp-1.16. URL https://aclanthology.org/2020.eval4nlp-1.16.

[25] N. Iskender, T. Polzehl, and S. Möller. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In A. Belz, S. Agarwal, Y. Graham, E. Reiter, and A. Shimorina, editors, *Proceedings of the Workshop on*

Human Evaluation of NLP Systems (HumEval), pages 86–96, Online, Apr. 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.humeval-1.10`.

[26] W. Kryściński, N. S. Keskar, B. McCann, C. Xiong, and R. Socher. Neural text summarization: A critical evaluation, 2019.

[27] N. Landro, I. Gallo, R. La Grassa, and E. Federici. Two new datasets for italian-language abstractive text summarization. *Information*, 13(5), 2022. ISSN 2078-2489. doi: 10.3390/info13050228. URL `https://www.mdpi.com/2078-2489/13/5/228`.

[28] A. Lavie and A. Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In C. Callison-Burch, P. Koehn, C. S. Fordyce, and C. Monz, editors, *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `https://aclanthology.org/W07-0734`.

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

[30] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013`.

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

[32] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023.

[33] L. Myers and M. Sirois. Differences between spearman correlation coefficients. *Encyclopedia of Statistical Evidence*, 07 2004. doi: 10.1002/0471667196.ess5050.

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://doi.org/10.3115/1073083.1073135`.

[35] A. Pasquale. Automatic generation of a navigation tree for conversational web browsing, 2022.

[36] M. Popović. chrF: character n-gram F-score for automatic MT evaluation. In O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, and P. Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL `https://aclanthology.org/W15-3049`.

[37] O. Press, N. A. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables input length extrapolation, 2022.

[38] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2018. URL `https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf`.

[39] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano. Mlsum: The multilingual summarization corpus, 2020.

[40] R. C. Streijl, S. Winkler, and D. S. Hands. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22: 213–227, 2016. URL `https://api.semanticscholar.org/CorpusID:15510814`.

[41] G. Team. Gemini: A family of highly capable multimodal models, 2023.

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

[43] F. Tresoldi. An improved natural language understanding pipeline for conversational web browsing, 2023.

[44] C. van der Lee, A. Gatt, E. van Miltenburg, and E. Krahmer. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech Language*, 67:101151, 2021. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2020.101151. URL `https://www.sciencedirect.com/science/article/pii/S088523082030084X`.

[45] O. Vasilyev, V. Dharnidharka, and J. Bohannon. Fill in the BLANC: Human-free quality estimation of document summaries. In S. Eger, Y. Gao, M. Peyrard, W. Zhao, and E. Hovy, editors, *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.eval4nlp-1.2. URL `https://aclanthology.org/2020.eval4nlp-1.2`.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017.

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

[48] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto. Benchmarking large language models for news summarization, 2023.

A | Dataset

The following reports some tuples of the dataset used for our experiments. It has been reported a tuple for each set of articles (Culture, Politics, Science, Sport). Each tuple is characterized by:

- **original**: it's the original text of the article.
- **old reference**: it's the reference summary previously put on the disposal by Polytechnic of Milan Univerisity's researchers.
- **new reference**: it's the newly used reference summary realized for our experiments.
- **mlsum**: is the summary realized by the model *mbart-summarization-mlsum*.
- **ilpost**: is the summary realized by the model *mbart-summarization-ilpost*.
- **bart**: is the summary realized by the model *BART*.
- **gpt**: is the summary realized by the model *GPT-3.5*.

The content of the dataset is originally in Italian, but for comprehension reasons, it has been reported the English-translated version as well. Furthermore, to ease the comprehension, some parts of the text have been highlighted with the following meaning:

- light colored parts (light orange, green, violet, pink) - parts of the original text that are written exactly (or almost exactly) in the summaries. Even if some parts are not highlighted, it's worth noticing that sometimes what changes is the order of a few words or the verbal tense. Therefore please take a look at the tuples of the dataset below to understand more what it contains.
- red - indicates a grammatical or content mistake. Content mistake means that the information provided by the summary is wrong with respect to the original text.
- yellow - the title of the model related to the summary has been yellow highlighted when the content of the summary has been considered wrong, in content or meaning. That means that the information given by the summary is wrong compared to the the

original text, or they have been given in a way that, even if apparently grammatically correct, is hard to understand by the reader.

About the yellow highlights, if a summary model title has not been yellow highlighted it doesn't mean that it is a summary that can be considered correct. Sometimes there are summaries where the sentence has correct grammar, and it stands well on its own, but it doesn't represent properly the content of the article (for example in the Politics tuple, *mlsum* summary).

After the example tuples of the dataset, there will be an example of how the forms presented on Prolific for the crowdsourcing task were structured.

A.1. Dataset

ITALIAN	ENGLISH
CULTURE	
original:	
In occasione del "Natale di Roma", che ricorre oggi 21 aprile per celebrare la fondazione della città nel 753 a. C., riapre al pubblico oggi la "Schola del Traiano" nel Parco archeologico di Ostia Antica, dopo diversi interventi di messa in sicurezza delle murature e degli apparati decorativi e dopo un intervento di bonifica dell'amianto. L'edificio, che prende il nome dalla grande statua dell'imperatore Traiano rinvenuta al suo interno durante gli scavi del 1938-39 ha una lunghissima vita: la prima fase dell'edificio vede la costruzione di un'elegante domus, chiamata Domus dei Bucrani, tra il 60 e il 50 a.C. Già alla fine del I secolo a.C. però la domus è distrutta e interrata, per via dell'innalzamento dell'acqua di falda. Al di sopra viene costruita una nuova domus, chiamata Domus del Peristilio, che ricalca all'incirca la struttura della residenza precedente. La nuova domus appartiene all'influente famiglia dei Fabii e così rimane fino all'inizio del III secolo a.C. quando l'ultimo esponente della gens Fabia, che aveva congiurato contro l'imperatore Elagabalo, fu fatto assassinare e i suoi beni confiscati. La domus cambia così destinazione d'uso, diventando forse una schola, ovvero una sede di corporazione".	On the occasion of the "Christmas of Rome", which occurs today 21 April to celebrate the foundation of the city in 753 BC. C., the "Schola del Traiano" reopens to the public today in the archaeological park of Ostia Antica, after several interventions to make the walls and decorative elements safe and after an asbestos reclamation operation."The building, which takes its name from the large statue of Emperor Trajan found inside it during the excavations of 1938-39, has a very long life: the first phase of the building saw the construction of an elegant domus, called Domus dei Bucrani, between 60 and 50 BC. Already at the end of the 1st century BC, however, the domus was destroyed and buried, due to the rise of the groundwater. A new domus was built above, called the Domus of the Peristyle, which roughly follows the structure of the previous residence. The new domus belongs to the influential Fabii family and remained so until the beginning of the 3rd century BC when the last exponent of the gens Fabia, who had conspired against the emperor Elagabalus, was murdered and his assets confiscated. The domus thus changes its intended use, perhaps becoming a schola, or a corporate headquarters".

old reference:

In occasione del "Natale di Roma" riapre al pubblico oggi la "Schola del Traiano" nel Parco archeologico di Ostia Antica.L'edificio, che prende il nome dalla grande statua dell'imperatore Traiano rinvenuta al suo interno durante gli scavi del 1938-39 ha una lunghissima vita.

On the occasion of the "Christmas of Rome", the "Schola del Traiano" reopens to the public today in the archaeological park of Ostia Antica. The building, which takes its name from the large statue of Emperor Trajan found inside during the excavations of 1938- 39 has a very long life.

new reference:

Dopo gli interventi di sicurezza, in occasione del "Natale di Roma", riapre il 21 aprile la "Schola

del Traiano" nel Parco archeologico di Ostia Antica. L'edificio prende il nome dalla grande statua dell'imperatore Traiano rinvenuta al suo interno. Nasce come domus tra il 60 e il 50 a.C.. La famiglia dei Fabii ne divenne proprietaria dopo una ricostruzione, ma quando i loro beni vennero confiscati divenne una schola.

After the security interventions, on the occasion of the "Christmas of Rome", the "Schola del Traiano"

reopens on April 21st in the archaeological park of Ostia Antica. The building takes its name from the large statue of Emperor Trajan found inside. It was born as a domus between 60 and 50 BC. The Fabii family became owners after a reconstruction, but when their assets were confiscated it became a schola.

mlsum:

La città riapre al pubblico la "schola del Traiano" nel parco archeologico di Ostia Antica

The city reopens to the public the "schola del Traiano" in the archaeological park of Ostia Antica

ilpost:

La "scheola del Traiano" di Ostia Antica. A 70 anni dal natale di Roma, in occasione del "Natale di Roma" riapre al pubblico oggi l\'edificio che ha reso famosa la grande statua dell\'imperatore Traiano.

The "Scheola del Traiano" of Ostia Antica. 70 years after the birth of Rome, on the occasion of the "Christmas of Rome", the\building that made the great statue of\Emperor Trajan famous reopens to the public today.

bart:	
"Schola del Traiano" riapre al pubblico oggi nel Parco archeologico di Ostia Antica. "L\'edificio" prende il nome dalla grande statua dell\'imperatore Traiano rinvenuta al suo interno.	"Schola del Traiano" reopens to the public today in the archaeological park of Ostia Antica. "The\building" takes its name from the large statue of\Emperor Trajan found inside.
gpt:	
In occasione del "Natale di Roma", la "Schola del Traiano" nel Parco archeologico di Ostia Antica riapre al pubblico dopo interventi di messa in sicurezza. L'edificio ha una storia lunga e complessa, passando da una domus dell'antica Roma a una possibile sede di corporazione. Questa riapertura offre ai visitatori l'opportunità di esplorare un pezzo di storia romana e apprezzare la conservazione del patrimonio culturale.	On the occasion of the "Christmas of Rome", the "Schola del Traiano" in the archaeological park of Ostia Antica reopens to the public after safety measures. The building has a long and complex history, going from a domus of ancient Rome to a possible headquarters for a corporation. This reopening offers visitors the opportunity to explore a piece of Roman history and appreciate the preservation of cultural heritage.

POLITICS

original:

La tutela dell'ambiente, della biodiversità e degli ecosistemi entra in Costituzione. L'Aula della Camera ha definitivamente approvato la proposta di legge costituzionale che modifica in tal senso due articoli della Carta, il 9 ed il 41. Il testo, alla seconda lettura alla Camera, è passato a Montecitorio con 468 voti a favore, un contrario e sei astenuti. Il Senato lo aveva approvato con la maggioranza dei due terzi lo scorso 3 novembre. Di conseguenza, entra subito in vigore e non è sottoponibile a referendum. Gli astenuti erano tutti di Fdi. In dichiarazione di voto, Emanuele Prisco aveva annunciato che il suo partito aveva lasciato libertà di coscienza ai suoi deputati. Il voto finale è stato salutato da un lungo applauso dell'Assemblea di Montecitorio. "Questo voto del Parlamento segna una giornata epocale: testimonio qui la presenza del governo che crede in questo cambiamento, grazie al quale la nostra Repubblica introduce nei suoi principi fondanti la tutela dell'ambiente". Così il ministro per la transizione ecologica, Roberto Cingolani, commenta a Montecitorio l'imminente voto della Camera che introduce la tutela dell'ambiente in Costituzione. "Grande soddisfazione per l'ok del Parlamento alla modifica della Costituzione con inserimento della tutela dell'ambiente e del principio di giustizia intergenerazionale".

The protection of the environment, biodiversity and ecosystems becomes part of the Constitution. The Chamber of the Chamber has definitively approved the constitutional bill which modifies two articles of the Charter in this sense, 9 and 41. The text, at the second reading in the Chamber, passed to Montecitorio with 468 votes in favour, one against and six abstained. The Senate approved it with a two-thirds majority on November 3. Consequently, it comes into force immediately and cannot be put to a referendum. The abstentions were all from Fdi. In his explanation of the vote, Emanuele Prisco announced that his party had granted freedom of conscience to its deputies. The final vote was greeted by long applause from the Montecitorio Assembly. "This vote by Parliament marks an epochal day: I testify here to the presence of the government that believes in this change, thanks to which our Republic introduces environmental protection into its founding principles". Thus the minister for the ecological transition, Roberto Cingolani, comments in Montecitorio on the imminent vote of the Chamber which introduces environmental protection into the Constitution. "Great satisfaction with Parliament's approval for the amendment of the Constitution with the inclusion of environmental protection and the principle of intergenerational justice".

old reference:

La tutela dell'ambiente, della biodiversità e degli ecosistemi entra in Costituzione. L'Aula della Camera ha definitivamente approvato la proposta di legge costituzionale. Il testo, alla seconda lettura alla Camera, è passato a Montecitorio con 468 voti a favore, un contrario e sei astenuti. Il Senato lo aveva già approvato con la maggioranza a novembre, di conseguenza, entra subito in vigore e non è sottoponibile a referendum.	The protection of the environment, biodiversity and ecosystems becomes part of the Constitution. The Chamber has definitively approved the constitutional bill. The text, at its second reading in the Chamber, passed Montecitorio with 468 votes in favour, one against and six abstentions. The Senate had already approved it with the majority in November, consequently it comes into force immediately and cannot be subjected to a referendum.

new reference:

La tutela dell'ambiente, della biodiversità e degli ecosistemi entra in Costituzione. L'Aula della Camera ha definitivamente approvato con 468 voti a favore la proposta di legge costituzionale che modifica due articoli della Carta. La modifica entra subito in vigore, gli astenuti alla votazione sono stati tutti di Fdi. A Montecitorio dopo l'applauso al voto, il ministro per la transizione ecologica, Roberto Cingolani, riconosce l'importanza di questo giorno.	The protection of the environment, biodiversity and ecosystems becomes part of the Constitution. The Chamber definitively approved with 468 votes in favor the constitutional bill which modifies two articles of the Charter. The change comes into force immediately, the abstentions in the vote were all from Fdi. In Montecitorio after the applause for the vote, the minister for the ecological transition, Roberto Cingolani, recognizes the importance of this day.

mlsum:

L'Aula della Camera approva la proposta di legge che modifica in senso due articoli della Carta, il 9 e il 41	The Chamber approves the bill which modifies two articles of the Charter, 9 and 41

ilpost:

la tutela dell'ambiente entra in Costituzione. La Camera ha definitivamente approvato la proposta di legge che modifica in tal senso due articoli della Carta: il voto finale è stato salutato da un lungo applauso dell'Assemblea di Montecitorio.

environmental protection becomes part of the Constitution. The Chamber has definitively approved the bill which modifies two articles of the Charter in this sense: the final vote was greeted by long applause from the Montecitorio Assembly.

bart:

L'Aula della Camera ha definitivamente approvato la proposta di legge costituzionale. La tutela dell'ambiente, della biodiversità e degli ecosistemi entra in Costituzione. Il testo, alla seconda lettura alla Camera, è passato a Montecitorio with 468 voti a favore, un contrario and sei astenuti.

The Chamber has definitively approved the constitutional bill. The protection of the environment, biodiversity and ecosystems becomes part of the Constitution. The text, at its second reading in the Chamber, passed Montecitorio con 468 votes in favour, one against e six abstentions.

gpt:

La Camera ha approvato una modifica costituzionale per inserire la tutela dell'ambiente, della biodiversità e degli ecosistemi nella Costituzione italiana. La proposta di legge costituzionale ha ottenuto un ampio sostegno con 468 voti a favore. Questa modifica entra subito in vigore senza bisogno di un referendum. Il ministro per la transizione ecologica, Roberto Cingolani, ha elogiato il voto come un momento epocale e ha sottolineato l'impegno del governo nel proteggere l'ambiente e garantire la giustizia intergenerazionale.

The Chamber has approved a constitutional amendment to include the protection of the environment, biodiversity and ecosystems in the Italian Constitution. The constitutional bill received broad support with 468 votes in favour. This change comes into force immediately without the need for a referendum. The Minister for Ecological Transition, Roberto Cingolani, praised the vote as an epochal moment and underlined the government's commitment to protecting the environment and ensuring intergenerational justice.

SCIENCE

original:

Venere è per dimensioni molto simile alla Terra e ha un mantello composto da silicati attorno a un nucleo ferroso, possiede un'atmosfera e l'attività sulla sua superficie rende evidente la presenza di attività geologica interna. Tuttavia è molto più asciutto della Terra e la sua atmosfera è novanta volte più densa. Esso è il pianeta più caldo del sistema solare, con temperature superficiali superiori ai 450 °C, molto probabilmente a causa della quantità di gas che provoca effetto serra nell'atmosfera. Non sono state individuate prove definitive delle attuali attività geologiche su Venere, ma si potrebbe pensare che la sua densa atmosfera sia regolarmente alimentata da eruzioni vulcaniche.

Venus is very similar in size to Earth and has a mantle composed of silicates around a ferrous core, it has an atmosphere and the activity on its surface makes the presence of internal geological activity evident. However, it is much drier than Earth and its atmosphere is ninety times denser. It is the hottest planet in the solar system, with surface temperatures above 450°C, most likely due to the amount of greenhouse gases in the atmosphere. No definitive evidence of current geological activity on Venus has been found, but one might think that its dense atmosphere is regularly fueled by volcanic eruptions.

old reference:

Venere è per dimensioni molto simile alla Terra e ha un mantello composto da silicati attorno a un nucleo ferroso. Esso è il pianeta più caldo del sistema solare, con temperature superficiali superiori ai 450 °C.

Venus is very similar in size to Earth and has a mantle composed of silicates around a ferrous core. It is the hottest planet in the solar system, with surface temperatures above 450°C.

new reference:

Venere è per dimensioni simile alla Terra. Ha un mantello di silicati attorno a un nucleo ferroso e attività geologia interna. Le eruzioni vulcaniche alimentano un'atmosfera densa. E' il pianeta più caldo del sistema solare, con temperature superficiali superiori ai 450 °C, infatti è molto asciutto.

Venus is similar in size to Earth. It has a silicate mantle around a ferrous core and active internal geology. Volcanic eruptions fuel a thick atmosphere. It is the hottest planet in the solar system, with surface temperatures above 450 °C, in fact it is very dry.

mlsum:

La temperatura dell'ambiente è inferiore ai 450 °C e la sua densa atmosfera è novanta volte più densa

The ambient temperature is less than 450°C and its dense atmosphere is ninety times denser

ilpost:

L'eruzione di Venere. Il pianeta più caldo del sistema solare continua a essere riempito di gas serra, ma non ci sono prove scientifiche per dire che la sua atmosfera sia regolarmente alimentata da eruzioni vulcaniche.

The eruption of Venus. The hottest planet in the solar system continues to be filled with greenhouse gases, but there is no scientific evidence to say its atmosphere is regularly fueled by volcanic eruptions.

bart:

Venere è per dimensioni molto simile alla Terra e ha un mantello composto da silicati attorno a un nucleo ferroso. La sua densa atmosfera sia regolarmente alimentata da eruzioni vulcaniche. Esso è il pianeta più caldo del sistema solare, con temperature superficiali superiori ai 450 °C.

Venus is very similar in size to Earth and has a mantle composed of silicates around a ferrous core. Its dense atmosphere being regularly fueled by volcanic eruptions. It is the hottest planet in the solar system, with surface temperatures above 450°C.

gpt:

Venere è simile in dimensioni alla Terra ma presenta notevoli differenze. Ha un mantello di silicati intorno a un nucleo ferroso ed è caratterizzato da attività geologica visibile sulla sua superficie. L'atmosfera di Venere è densa e composta principalmente di anidride carbonica, generando un effetto serra che porta a temperature superiori ai 450°C. La superficie è arida e non sono state osservate prove definitive di attività vulcanica attuale, ma potrebbe esservi stata in passato.	Venus is similar in size to Earth but has notable differences. It has a silicate mantle around a ferrous core and is characterized by visible geological activity on its surface. Venus' atmosphere is dense and composed mainly of carbon dioxide, generating a greenhouse effect that leads to temperatures above 450°C. The surface is arid and no definitive evidence of current volcanic activity has been observed, but there may have been in the past.

SPORT

original:

L'incubo di Alessia Orro, la palleggiatrice del Vero Volley di Monza e della Nazionale vittima di stalking, si è materializzato nuovamente, ma lei ha avuto ancora il coraggio di denunciare. É stato preso in flagranza di reato, infatti, il cinquantacinquenne originario del novarese, proprio mentre era intento a perseguitare la sua vittima. E il suo club le ha manifestato grande vicinanza. "Siamo orgogliosi di Alessia, che ancora una volta ha dimostrato la sua personalità e tutto il suo valore anche come persona denunciando la situazione e affidandosi tempestivamente ai Carabinieri per la sua tutela e la soluzione del caso". Queste il messaggio della società brianzola di serie A1 che dimostra tutta la sua vicinanza all'atleta. Oltre a sottolineare il ruolo pedagogico che il gesto di Orro può avere, come esempio per le numerose persone vittime di stalking come lei. La 23enne aveva già dimostrato la sua determinazione tre anni fa, quando segnalò la stessa persona che la seguiva e la minacciava alle forze dell'ordine. Il Gip di Varese, in quell'occasione, aveva condannato l'uomo agli arresti domiciliari, nonostante il parere contrario del sostituto procuratore, che voleva la conferma

The nightmare of Alessia Orro, the setter of Vero Volley di Monza and the national team who was the victim of stalking, materialized again, but she still dared to report. In fact, the fifty-five-year-old from the Novara area was caught in the act of committing a crime, just as he was intent on persecuting his victim. And her club showed great closeness to her. "We are proud of Alessia, who once again demonstrated her personality and all her value as a person by reporting the situation and promptly entrusting herself to the Carabinieri to protect her and resolve the case." This is the message from the A1 series Brianza club which demonstrates its closeness to the athlete. In addition to underlining the pedagogical role that Orro's gesture can have, as an example for the numerous people who are victims of stalking like her. The 23-year-old had already demonstrated her determination three years ago, when she reported the same person who followed her and threatened her to the police. The investigating judge of Varese, on that occasion, had sentenced the man to house arrest, despite the contrary opinion of the deputy prosecutor, who wanted the prison sentence to be confirmed. Too little because Angelo

del carcere. Troppo poco perché Angelo Persico, professionista di Novara, è tornato sulle tracce della pallavolista, con gli stessi metodi: appostamenti durante le gare e gli allenamenti e messaggi insistenti sui social, anche insulti e minacce.	Persico, a professional from Novara, went back on the volleyball player's trail, with the same methods: stakeouts during matches and training sessions and insistent messages on social media, including insults and threats.

old reference:

L'incubo di Alessia Orro, la palleggiatrice del Vero Volley di Monza e della Nazionale vittima di stalking, si è materializzato nuovamente, ma lei ha avuto ancora il coraggio di denunciare. É stato preso in flagranza di reato, infatti, il cinquantacinquenne originario del novarese, proprio mentre era intento a perseguitare la sua vittima. La 23enne aveva già dimostrato la sua determinazione tre anni fa, quando segnalò la stessa persona che la seguiva e la minacciava alle forze dell'ordine.	The nightmare of Alessia Orro, the setter of Vero Volley di Monza and the national team who was the victim of stalking, materialized again, but she still had the courage to report. In fact, the fifty-five-year-old from the Novara area was caught in the act of committing a crime, just as he was intent on persecuting his victim. The 23-year-old had already demonstrated her determination three years ago, when she reported the same person who followed her and threatened her to the police.

new reference:

Alessia Orro, la palleggiatrice del Vero Volley di Monza e della Nazionale, dopo 3 anni è nuovamente vittima di stalking da parte della stessa persona, ma lei ha avuto ancora il coraggio di denunciare. Il colpevole Angelo Persico, condannato precedentemente agli arresti domiciliari, è stato preso in flagrante mentre perseguitava la vittima. La società brianzola di serie A1 dimostra vicinanza all'atleta.	Alessia Orro, the setter of Vero Volley di Monza and the national team, is once again the victim of stalking by the same person after 3 years, but she still had the courage to report it. The guilty Angelo Persico, previously sentenced to house arrest, was caught red-handed while persecuting the victim. The Brianza A1 series club demonstrates closeness to the athlete.

mlsum:

L'atleta è stato preso in flagranza di reato, infatti, il cinquantacinquenne originario del novarese, che stava per perseguitare la sua vittima	The athlete was caught in flagrante delicto, in fact, the fifty-five year old originally from Novara, who was about to persecute his victim

ilpost:

Alessia Orro ha ancora il coraggio di denunciare. La palleggiatrice del Vero Volley di Monza e la Nazionale vittima di stalking si è materializzata nuovamente, ma lei ha avuto ancora il coraggio.	Alessia Orro still has the courage to report. The setter of Vero Volley di Monza and the national team who was the victim of stalking materialized again, but she still had the courage.

bart:

Alessia Orro, palleggiatrice del Vero Volley di Monza, vittima di stalking. Orro aveva dimostrato la sua determinazione tre anni fa. Angelo Persico, professionista di Novara, è tornato sulle tracce della pallavolista.	Alessia Orro, setter of the Vero Volley of Monza, victim of stalking. Orro had demonstrated her determination three years ago. Angelo Persico, a professional from Novara, is back on the trail of the volleyball player.

gpt:

La palleggiatrice Alessia Orro, membro del Vero Volley Monza e della Nazionale, è nuovamente vittima di stalking. Tuttavia, dimostrando coraggio, ha denunciato l'individuo coinvolto. L'uomo è stato arrestato in flagrante mentre perseguiva la sua vittima. Il club Monza ha espresso sostegno e orgoglio per la determinazione di Orro nel denunciare la situazione.	The setter Alessia Orro, a member of Vero Volley Monza and the national team, is once again the victim of stalking. However, showing courage, she reported the individual involved. The man was arrested red-handed while stalking his victim. The Monza club expressed support and pride in Orro's determination to report the situation.

B | Prolific Form

The following is an example of a form sent to be completed to the crowdworkers. The example is given completely in English but take into consideration that the form was originally written in Italian, as well as the text of summaries shown. For this example, it has been used the form made for the first tuple of the Dataset shown in Appendix A. Google Form was used to create each form.

Research on the quality of summaries generated by neural network models

Section 1

Introduction:
Hello and thank you for participating in our research study! My name is Awad Yasmin, and I am a student at the Doshisha University and Polytechnic of Milan University who is conducting a master's degree research. This study focuses on evaluating summaries generated by neural network models.

Purpose of the Research:
The main aim of this research is to demonstrate that current evaluation methods, such as ROUGE, which are based on direct comparison of textual features rather than semantic analysis, may not be suitable for accurately evaluating the output of language models of large size (LLM), such as chatGPT.

Summary Evaluation Process:
You will evaluate summaries generated by various neural networks. For this search, you will not be told which neural network produced each summary. We will ask you to evaluate different aspects of each summary. The text from which the summary was generated will not always be shown, as it is superfluous for the aspects for which you will be asked to judge. Your responses will remain confidential and any personal information will be treated with the utmost care. Your sincerity is invaluable to the success of this research. Please answer the survey questions as faithfully as possible. Thank you for your participation, and I wish you a pleasant experience during the survey!

Enter your Prolific ID here
At the end of the questionnaire you will find the URL to click to demonstrate completion of the questionnaire on Prolific. Once your response is validated, payment will be accepted.

[space for text answer]

Section 2

Intrinsic Qualities
The Intrinsic Qualities of the summary are aspects that are measured based on the summary itself without considering the reference document.

The description of each aspect is as follows:

- **Overall Quality**: refers to the overall impression of the summary, taking into account various factors such as clarity, coherence, and informativeness (i.e. the degree of information conveyed by the text). For example, given a news summary, a concise, well-written summary that effectively captures the main points of the story is rated high overall.
- **Grammar**: refers to the fact that the summary should not have capitalization errors, dictation, or ungrammatical sentences (e.g. fragments, missing components) that make the text difficult to read.
- **Non-Redundancy**: Assess the presence of unnecessary repetition in the summary, such as repeated phrases, repeated facts, or the excessive use of specific nouns or noun phrases. Repeating the same information in consecutive sentences or using a person's full name when a pronoun would suffice are examples of redundancy.
- **Referential Clarity**: Evaluate how easy it is to identify pronouns and noun phrases and connect them to what they refer to in the summary. The lack of clarity would occur if a person, thing, place, time, etc. is mentioned, but its role or connection to the story is unclear.
- **Focus**: Evaluate whether the summary maintains a clear focus, containing only information relevant to the rest of the summary. Including irrelevant details or information not directly related to the main topic would indicate a lack of focus.
- **Structure and Coherence**: evaluate the organization and flow of the summary, ensuring that it is well structured and develops from sentence to sentence coherently. An example of poor structure would be a summary that appears as a random collection of related information with no clear progression.

For each summary, the evaluation of the 6 aspects is required on a scale from 1 to 5 (1 - Very low, 2 - Low, 3 - Moderate, 4 - High, 5 - Very High). If you are taking the questionnaire from your phone, be careful to scroll sideways to see all the answer options.

The city reopens to the public the "schola del Traiano" in the archaeological park of Ostia Antica

	1 - Very Low	2 - Low	3 - Moderate	4 - High	5 - Very High
Overall quality	○	○	○	○	○
Grammar	○	○	○	○	○
Non-Redundan...	○	○	○	○	○
Referential Clar...	○	○	○	○	○
Focus	○	○	○	○	○
Structure and c...	○	○	○	○	○

After the security interventions, on the occasion of the "Christmas of Rome", the "Schola del Traiano" reopens on April 21st in the archaeological park of Ostia Antica. The building takes its name from the large statue of Emperor Trajan found inside. It was born as a domus between 60 and 50 BC. The Fabii family became owners after a reconstruction, but when their assets were confiscated it became a schola.

	1 - Very Low	2 - Low	3 - Moderate	4 - High	5 - Very High
Overall quality	○	○	○	○	○
Grammar	○	○	○	○	○
Non-Redundan...	○	○	○	○	○
Referential Clar...	○	○	○	○	○
Focus	○	○	○	○	○
Structure and c...	○	○	○	○	○

The “Scheola del Traiano” of Ostia Antica. 70 years after the birth of Rome, on the occasion of the "Christmas of Rome", the\building that made the great statue of\Emperor Trajan famous reopens to the public today.

	1 - Very Low	2 - Low	3 - Moderate	4 - High	5 - Very High
Overall quality					
Grammar					
Non-Redundan...					
Referential Clar...					
Focus					
Structure and c...					

"Christmas of Rome", which occurs today April 21st "Schola del Traiano" in the archaeological park on April 212th of Ostia Antica. which takes its name between 60 and 50 BC. Already at the end of the 1st century BC. but the domus is destroyed. This question is used to check the level of attention of the person taking the questionnaire, please answer 2 to all the assessments. and buried, due to the rising groundwater.

	1 - Very Low	2 - Low	3 - Moderate	4 - High	5 - Very High
Overall quality					
Grammar					
Non-Redundan...					
Referential Clar...					
Focus					
Structure and c...					

"Schola del Traiano" reopens to the public today in the archaeological park of Ostia Antica. "The\building" takes its name from the large statue of\Emperor Trajan found inside.

	1 - Very Low	2 - Low	3 - Moderate	4 - High	5 - Very High
Overall quality	○	○	○	○	○
Grammar	○	○	○	○	○
Non-Redundan...	○	○	○	○	○
Referential Clar...	○	○	○	○	○
Focus	○	○	○	○	○
Structure and c...	○	○	○	○	○

On the occasion of the "Christmas of Rome", the "Schola del Traiano" in the archaeological park of Ostia Antica reopens to the public after safety measures. The building has a long and complex history, going from a domus of ancient Rome to a possible headquarters for a corporation. This reopening offers visitors the opportunity to explore a piece of Roman history and appreciate the preservation of cultural heritage.

	1 - Very Low	2 - Low	3 - Moderate	4 - High	5 - Very High
Overall quality	○	○	○	○	○
Grammar	○	○	○	○	○
Non-Redundan...	○	○	○	○	○
Referential Clar...	○	○	○	○	○
Focus	○	○	○	○	○
Structure and c...	○	○	○	○	○

Section 3

Informativeness
Summary informativeness measures how much of the information contained in the source document is preserved in the extracted summary.

The first question will show the source document of the generated summaries (it is not necessary to score the source document). You will then be shown a comprehension question to answer based on that document.

The various summaries will then be shown. Please rate how well each summary preserves the crucial information of the source document by giving a score from 1 to 5 (1 - Very little, 2 - A little, 3 - Moderately, 4 - A lot, 5 - Very much), and answer the comprehension question below based only on the summary in question.

On the occasion of the "Christmas of Rome", which occurs today 21 April to celebrate the foundation of the city in 753 BC. C., the "Schola del Traiano" reopens to the public today in the archaeological park of Ostia Antica, after several interventions to make the walls and decorative elements safe and after an asbestos reclamation operation. "The building, which takes its name from the large statue of Emperor Trajan found inside it during the excavations of 1938-39, has a very long life: the first phase of the building saw the construction of an elegant domus, called Domus dei Bucrani, between 60 and 50 BC. Already at the end of the 1st century BC, however, the domus was destroyed and buried, due to the rise of the groundwater. A new domus was built above, called the Domus of the Peristyle, which roughly follows the structure of the previous residence. The new domus belongs to the influential Fabii family and remained so until the beginning of the 3rd century BC when the last exponent of the gens Fabia, who had conspired against the emperor Elagabalus, was murdered and his assets confiscated. The domus thus changes its intended use, perhaps becoming a schola, or a corporate headquarters".

	1	2	3	4	5	
Very little	○	○	○	○	○	Very much

Based on the text above, were there any safety measures prior to reopening?

- yes
- no
- maybe
- There is no such information in the text

The city reopens to the public the "schola del Traiano" in the archaeological park of Ostia Antica

	1	2	3	4	5	
Very little	○	○	○	○	○	Very much

Based on the summary above, were there any safety measures prior to reopening?

- yes
- no
- maybe
- There is no such information in the text

After the security interventions, on the occasion of the "Christmas of Rome", the "Schola del Traiano" reopens on April 21st in the archaeological park of Ostia Antica. The building takes its name from the large statue of Emperor Trajan found inside. It was born as a domus between 60 and 50 BC. The Fabii family became owners after a reconstruction, but when their assets were confiscated it became a schola.

	1	2	3	4	5	
Very little	○	○	○	○	○	Very much

Based on the summary above, were there any safety measures prior to reopening?

- yes
- no
- maybe
- There is no such information in the text

The “Scheola del Traiano” of Ostia Antica. 70 years after the birth of Rome, on the occasion of the "Christmas of Rome", the\building that made the great statue of\Emperor Trajan famous reopens to the public today.

	1	2	3	4	5	
Very little	○	○	○	○	○	Very much

Based on the summary above, were there any safety measures prior to reopening?

- yes
- no
- maybe
- There is no such information in the text

"Christmas of Rome", which occurs today April 21st "Schola del Traiano" in the archaeological park on April 212th of Ostia Antica. which takes its name between 60 and 50 BC. Already at the end of the 1st century BC. but the domus is destroyed. This question is used to check the level of attention of those who are carrying out the questionnaire, please answer 2. and underground, due to the rising groundwater.

	1	2	3	4	5	
Very little	○	○	○	○	○	Very much

Based on the summary above, were there any safety measures prior to reopening?

- yes
- no
- maybe
- There is no such information in the text

"Schola del Traiano" reopens to the public today in the archaeological park of Ostia Antica. "The\building" takes its name from the large statue of\Emperor Trajan found inside.

	1	2	3	4	5	
Very little	○	○	○	○	○	Very much

Based on the summary above, were there any safety measures prior to reopening?

- yes
- no
- maybe
- There is no such information in the text

On the occasion of the "Christmas of Rome", the "Schola del Traiano" in the archaeological park of Ostia Antica reopens to the public after safety measures. The building has a long and complex history, going from a domus of ancient Rome to a possible headquarters for a corporation. This reopening offers visitors the opportunity to explore a piece of Roman history and appreciate the preservation of cultural heritage.

	1	2	3	4	5	
Very little	○	○	○	○	○	Very much

Based on the summary above, were there any safety measures prior to reopening?

- yes
- no
- maybe
- There is no such information in the text

Section 4

Final Considerations

If you have any comments, observations, or considerations you would like to leave us, please write them here. We will be happy to read them!

[space for text answer]

Confirmation Message

Final Considerations

Thank you so much for your participation! Here is the URL to click to demonstrate your completion of the questionnaire on Prolific!

https://app.prolific.com/submissions/complete?cc=examplecode

In case of problems, this is the code to copy and paste to the platform: *examplecode*

We wish you a good day :)

List of Figures

List of Tables

Acknowledgements

Tra non molto finiranno i miei giorni da universitaria. Anche se al momento della stesura di questi ringraziamenti non ne sono certa. Sono in attesa. Ma se queste parole verranno lette sarà perchè sarà successo. Sarà arrivata la parola fine. Sono sicura di volerla raggiungere quella parola, so di essere pronta per un nuovo capitolo, nuove energie, nuovi sorrisi. Avete presente il meme "queste rocce sono rocce"? Ecco, questo percorso è stato un percorso, durato anni e quindi ospitante momenti felici e momenti infelici. Come è normale che sia. Sono contenta di aver conosciuto tante nuove persone, e aver visto posti mai veduti. Tra poche settimane mi laureerò in Giappone. Kyoto è stata la mia residenza negli ultimi 2 anni, e spero continuerà ad esserlo per un altro po'. Da pochi giorni mi sono trasferita alla Comune, luogo fatato abitato da dolci streghette che mi hanno fatto sentire a casa dall'altra parte del mondo: Dafne, Laura e Giada. Mi fa sempre ridere come spesso le belle persone si afficcino nella mia vita nei modi più improbabili. E poi c'è Marta, la mia amica spagnola di fiducia, vicina di stanza e di vita durante gli anni da studente alla Doshisha. Marta mi ha insegnato che coche amariglio è più bello di dire macchina gialla, e che non devo iniziare questo gioco con lei perchè sono una persona troppo distratta. Dall'Italia con furore c'è stata Alice, il ghepardo con cui ho affrontato la giungla di Milano. Con cui ho studiato, giocato, e parlato, spesso da dietro uno schermo. Non c'era molta scelta. Ma a noi che ce ne frega? Quei pomeriggi di studio sono stati accompagnati anche da Lorenzo, e qualche volta il piccolo Spettro. E quanto mi sono divertita anche con i ROPgais. Supporti dalla famiglia non sono mai mancati. Mamma, Omar, Elia, Baba. Siete sempre dentro al mio cuore. E sono abbastanza fortunata da poter dire che se dovessi scrivere tutti i nomi dei componenti della mia famiglia, la lista sarebbe talmente lunga che non entrerebbe in una pagina. Ma sappiate che vi penso, vi ho pensato, e vi penserò sempre. Niki, la mia compagna di avventure per eccellenza, spero di rivederti presto. Strano ma vero, ho continuato a conoscere persone dall'Italia mentre stavo dall'altra parte del mondo. C'è stato un romano che mi ha ricordato che tutto può cambiare. E poi c'è Ele. Ele ha reso i colori di questo mondo più accesi. Ele fa nascere in me tanti sorrisi. E mi ricorda che il buono è una scelta. Ho ricevuto tanto buono da tutti voi. E spero di essere stata capace nel tempo di dare a voi il buono che voi avete dato a me. E' bello sapere che il mio cuore si stia allargando, diventando capace di ospitare sempre più persone. Grazie per aver deciso di affrontare insieme questa strana cosa chiamata vita, anche se solo in parte. Grazie per essere stati o essere diventati la mia famiglia in questi ultimi 3 anni. Vi voglio bene profondamente.

Questa ricerca mi ha dato l'occasione di lavorare con persone dal cuore gentile. Non era qualcosa che mi sarei aspettata, ma sono contenta che sia successo. E spero di non aver esageratamente deluso le loro aspettative. Ringrazio la Professoressa Maristella, Emanule e Marco. Mi avete regalato supporto e comprensione da migliaia di chilometri di distanza. Grazie. E grazie per avermi dato l'opportunità di lavorare su un progetto che può concretamente fare del bene. Spero che ConWeb continui a crescere, nelle mani di persone capaci e con felici obbiettivi, come voi e i vostri studenti e collaboratori. Siete un team fortissimo! E io tifo per voi.

A voi, citati in queste brevi parole, e a voi altri, non citati ma a cui dedico un pensiero. Ecco, a tutti voi, auguro una vita piena di avventure, sorrisi, e fiori.

Preferibilmente, fiori gialli.