# COSMIC: Mutual Information for Task-Agnostic Summarization Evaluation

**Maxime DARRIN**[1,2,3,4] **Philippe FORMONT**[1,2,4,5]
**Jackie Chi Kit CHEUNG**[2,3,7] **Pablo PIANTANIDA**[1,2,4,6]
[1]International Laboratory on Learning Systems, [2]Mila - Quebec AI Institute
[3]McGill University [4]Université Paris-Saclay, [5]École de technologie supérieure (ETS)
[6]CNRS, CentraleSupélec [7]Canada CIFAR AI Chair
maxime.darrin@mila.quebec philippe.formont@mila.quebec
jackie.cheung@mcgill.ca pablo.piantanida@mila.quebec

## Abstract

Assessing the quality of summarizers poses significant challenges—gold summaries are hard to obtain and their suitability depends on the use context of the summarization system. Who is the user of the system, and what do they intend to do with the summary? In response, we propose a novel task-oriented evaluation approach that assesses summarizers based on their capacity to produce summaries while preserving task outcomes. We theoretically establish both a lower and upper bound on the expected error rate of these tasks, which depends on the mutual information between source texts and generated summaries. We introduce COSMIC, a practical implementation of this metric, and demonstrate its strong correlation with human judgment-based metrics, as well as its effectiveness in predicting downstream task performance. Comparative analyses against established metrics like BERTScore and ROUGE highlight the competitive performance of COSMIC.

## 1 Introduction

Assessing the quality of summarizers in different settings, tasks, and datasets is critical for better understanding these models and for studying their strengths and weaknesses. In many text generation scenarios, assessing model quality is arduous and resource-intensive, often necessitating human annotations and evaluations. Consequently, developing automatic metrics that align closely with human judgments is paramount (Graham and Baldwin, 2014; Tratz and Hovy, 2008; Giannakopoulos et al., 2008; Deutsch et al., 2021).

Standard automatic evaluation methods for summarization rely on the idea that a good summary should have some (semantic) overlap with either a gold standard or the source text (El-Kassas et al., 2021; Allahyari et al., 2017). They leverage similarity metrics such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004) to evaluate the quality of the generated summaries. However, these methods often do not correlate well with human judgements (Kryscinski et al., 2020; Kocmi et al., 2021). To enhance alignment with human judgment and capitalize on recent advancements in large language models, recent efforts have concentrated on learned metrics for scoring summaries (Zhang et al., 2020b; Rei et al., 2020; Liu et al., 2023). For instance, Clark et al. (2023) introduced six new learned metrics by finetuning a pretrained MT5 model (Xue et al., 2021) to predict human judgment along different axes. Another line of research focuses on reconstruction-based metrics, such as BLANC (Vasilyev et al., 2020) and Information Difference (Egan et al., 2021), the former evaluating the actual reconstruction error of the source text with and without a summary, and the latter evaluating the information gain obtained when conditioning the generative model on the summary.

Standard evaluation methods, therefore, suffer from two major shortcomings. They rarely formally define a clear notion of quality and the methods used to evaluate said notion lack theoretical foundations. This leads to potential discrepancies between the intended evaluation and what is actually measured, leading to a lack of validity of the evaluation methods.

We introduce a task-oriented evaluation setup where summaries are meant to allow an agent to perform downstream tasks without reading the longer source text. For example, if a political advisor drafts a briefing on a subject to enable a politician to make informed decisions, our evaluation considers the advisor successful if the decisions made using the summary align with those made using the initial source text (Pu et al., 2023; Vanderwende et al., 2007). Providing such a well-defined notion of success enables careful analysis of the validity of the evaluation for different possible use cases. In addition, it enables us to perform mechanical evaluations of our method, reducing the noise



*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12696–12717

from the evaluation process.

Furthermore, many summarization techniques typically operate under the assumption that the downstream task is unknown during the summarization process, yet an implicit notion of summary quality exists. However, this assumption is seldom explicitly addressed or substantiated. In our task-oriented approach, we provide an information-theoretic rationale for the existence of such a metric. We demonstrate that it essentially involves assessing the mutual information (MI) between the source texts' distribution and the summaries generated by a given summarizer.

This paper does not aim to study the syntactic, semantic, and pragmatic aspects of features of summary information, which are essential for capturing the rich notion of information in human communication. For instance, we do not account for the summaries' fluency, grammatical correctness, or coherence. Instead, our approach emphasizes how effectively, in the best-case scenario, the output of a summarizer can be used to perform downstream tasks. The results in this paper suggest that the formal definition of mutual information successfully achieves this goal. Our approach leverages embeddings to abstract the surface form of text, following recent studies by Pillutla et al. (2021) and Pimentel et al. (2023).

We evaluate the quality of our approach in two ways. First, we show that summarizers that induce a summary distribution with higher MI with the source texts' distribution are higher quality in the following sense — *they tend to produce summaries that preserve outcomes on downstream tasks* as compared to using the source texts. Second, we compare the MI to metrics trained on human judgments and show that it displays consistent correlations. Our results are consistent with and extend previous work that leverages MI to understand relationships in datasets (Ethayarajh et al., 2022; Bugliarello et al., 2020), predictive performance of representations (Sui et al., 2023) and tool to construct or evaluate representations (Kim et al., 2022), models and generative processes.

**Contributions.** Our contributions are threefold:

1. **A theoretical setting for summarizer evaluation.** We frame the summarizer evaluation problem as a statistical inference problem and we derive a task-agnostic reference-free quality metric: the MI between the distribution of source texts and the distribution of the summarizer's outputs.

2. **A practical implementation of this metric: `COSMIC`.** We propose a practical implementation of `COSMIC` using an MI estimator and sentence embeddings [1].

3. **An experimental evaluation.** We examine how well MI predicts the performance of downstream tasks in comparison to conventional metrics like `BERTScore` and `BARTScore`. Our findings demonstrate that MI is competitive with these metrics. Additionally, we illustrate its strong correlation with metrics trained to emulate human judgment.

## 2 Related Work

Assessing the quality of summaries poses a unique challenge due to the contextual nature of 'quality', influenced by factors such as the audience, topic, and intended purpose of the summary. Even expertly crafted human summaries considered "gold standard" in one context may be perceived as subpar in a different setting (Saziyabegum and Sajja, 2017; Indu and Kavitha, 2016).

**Reference-free summary evaluation.** Reference-free evaluation methods mostly rely on comparing the content of the summary with the content of the source text (Louis and Nenkova, 2013; El-Kassas et al., 2021) and they rely on common overlap metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) or BERTScore (Zhang et al., 2020b). However, most reference-free metrics show some important limitations (Deutsch et al., 2022): they lack theoretical grounding. Work such as BLANC (Vasilyev et al., 2020) and Information Difference (Egan et al., 2021) are closely related to our work. The former evaluates the actual reconstruction error of the source text with and without a summary, while the latter evaluates the information difference evaluated by the generative model. However, both lack the theoretical justification we introduce in Section 3 and only evaluate the information based on the generative model. While these methods appear intuitive, they lack of theoretical support for their approaches, relying solely on empirical results and correlations with human evaluations. Thus, the results are often tied to a

[1] A plug&play python library built on top of HF transformers is available as supplementary material and will be released upon publication of this work.

dataset and are affected by variance from the human evaluation. Conversely, we offer a theoretical framework applicable to any dataset and measure a tangible, well-defined success criterion: do the tasks yield the same output when performed on summaries as they do on the source texts?

**Embedding-based evaluation.** MAUVE (Pillutla et al., 2021) first proposed a new information-theoretic metric to compare two text distributions based on embedding clustering; more recent work showed that the crucial element was the clustering step (Pimentel et al., 2023). They show that while embeddings (and the clusters they form) do not capture fluency or grammatical correction, they do grasp meaning and coherence, making them excellent tools for evaluation.

**Dataset difficulty and MI**. Measuring MI between the concepts and input in a dataset is not a novel idea. In fact, Ethayarajh et al. (2022) leverages the Arimoto information (Arimoto, 1971), rediscovered and dubbed $\mathcal{V}$-usable information by Xu et al. (2020) to assess the difficulty of a dataset. Similarly, Bugliarello et al. (2020) evaluate the difficulty of translating from one language to another. Following this trend, Kim et al. (2022) reuses the point-wise MI between Gaussian distributions to evaluate text-to-images and image-to-text generative models.

**Mutual information for summarization.** The MI is a natural metric to optimize in summarization. It has been used as a score to select the most informative or surprising sentences in extractive summarization (Padmakumar and He, 2021) or as an alternative objective for text decoding (van der Poel et al., 2022). In this paper, we revisit the use of MI but between the distribution of source texts and the distribution of the summarizer.

## 3 A Task-Driven Evaluation Framework

### 3.1 Background: Probabilistic models for text summarization

We consider models for language summarization tasks that define a probability distribution over strings. More formally, these models are probability distributions $s$ over an output space $\mathcal{S}$ conditioned on an input text $\mathbf{t}$, where $\mathcal{S}$ is the set consisting of all possible strings that can be constructed from the vocabulary $\Omega$: $\mathcal{S} \triangleq \left\{\texttt{BOS} \circ \mathbf{s} \circ \texttt{EOS} \,|\, \mathbf{s} \in \Omega^*\right\}$, BOS and EOS stand for special reserved beginning-of-sequence and end-of-sequence tokens, respectively, and $\Omega^*$ denotes the Kleene closure of $\mathcal{S}$.

Today's models for language summarization are typically parameterized by encoder-decoder or decoder-only architectures with attention mechanisms with trainable weights $\theta$. These models follow a local-normalization scheme, meaning that $\forall\, i > 0, p_\theta(\cdot|\mathbf{s}_{<i}, \mathbf{t})$ defines a probability distribution over $\bar{\mathcal{S}} = \mathcal{S} \cup \texttt{EOS}$. The probability of a sequence $\mathbf{s} = \langle s_0, s_1, \ldots\rangle$ can then be decomposed as:

$$p_\theta(\mathbf{s}|\mathbf{t}) = \prod_{i=1}^{|\mathcal{S}|} p_\theta(\mathbf{s}_i|\mathbf{s}_{<i}, \mathbf{t}), \tag{1}$$

and $\mathbf{s}_{<i} = \langle s_0, \ldots, s_{i-1}\rangle$, $s < 1 = s_0 \triangleq \texttt{BOS}$.

### 3.2 Background: Information Theory

Information theory (Cover and Thomas, 2006) provides several tools for analyzing data and their associated probability distributions, including entropy and MI. These metrics are typically defined based on a "true" probability distribution, denoted as $p(c)$, or the joint probability density function $p(\mathbf{t}, \mathbf{s})$ which may not be known but governs the behavior of random variables $\mathbf{T}$ and $\mathbf{S}$. The fundamental concept in information theory is **surprisal**, defined as $H(C = c) = -\log p(c)$, and its expected value is termed **entropy:**

$$H(C) = \sum_{c \in \mathcal{C}} p(c) H(C = c). \tag{2}$$

Finally, another important concept is the **mutual information** (MI) between two random variables:

$$I(\mathbf{T}; \mathbf{S}) = H(\mathbf{T}) - H(\mathbf{T}|\mathbf{S}). \tag{3}$$

It captures the amount of information we get about one random variable when observing a realization of the other. **The data-processing inequality** (Cover and Thomas, 2006) states that processing a random variable with a (possibly random) function $f(\cdot)$ can never increase its informativeness but only reduce its information content, expressed as:

$$I(C; f(\mathbf{T})) \leqslant I(C; \mathbf{T}). \tag{4}$$

The **rate-distortion** (RD) function of a discrete random variable $C$ for a given distortion function $\ell(c, \widehat{c})$ is defined as (Csiszár, 1974, eq. (1.4)):

$$R_{C,\ell}(D) \triangleq \min_{p(\widehat{c}|c)\,:\,\mathbb{E}[\ell(C,\widehat{C})] \leqslant D} I(C; \widehat{C}). \tag{5}$$

For further details, the reader is referred to Appendix A.2 and (Cover and Thomas, 2006).

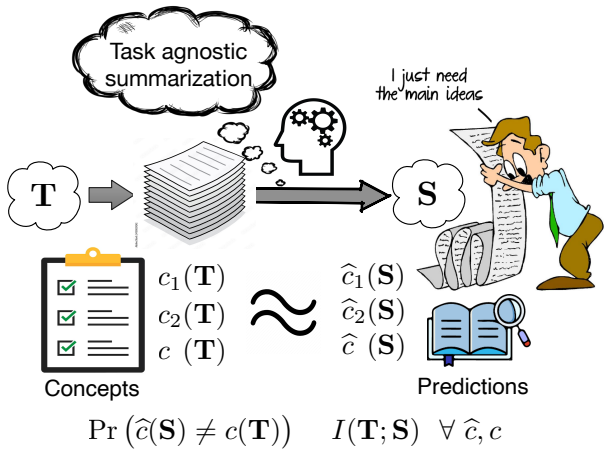


Figure 1: The summarizer is expected to generate summaries $\mathbf{S}$ for a given distribution of source texts $\mathbf{T}$, without prior knowledge of the specific application, e.g. predicting the concepts: $c_1, c_2, \ldots$. The objective is to assess the discrepancy incurred when predicting from the generated summaries instead of the original source texts. We demonstrate that, for any undisclosed task $c$, the likelihood of error is constrained within bounds determined by monotonous functions of the MI.

### 3.3 A task-oriented evaluation setting

Most summarization methods operate under the assumption that the downstream task of interest is unknown during the generation process, relying instead on a generic notion of summary quality. We assume in this work that the goal is to be able to perform similarly in terms of classification error on both the original texts $\mathbf{T}$ and the resulting summaries $\mathbf{S}$[2]. Next, we formalize this evaluation metric based on the assumption of an unknown downstream task.

Let $c : \Omega^* \to \{1, \ldots, m\}$ denote the target concept of interest which can be extracted from the initial texts $\mathbf{T}$ by applying $C \triangleq c(\mathbf{T})$; and similarly, let $\widehat{c}(\mathbf{S})$ denote the predicted concept from the summaries $\mathbf{S}$ according to the underlying model $p_\theta(\mathbf{s}|\mathbf{t})$. The evaluation of the summarizer's quality with respect to the downstream task (unknown from the summarization model) can be assessed using the expected error rate, as illustrated in Figure 1. In other words, it involves the classifier, determining the average probability of the extracted concept $\widehat{c}(\mathbf{S})$ differing from the original concept $c(\mathbf{T})$ in the source text:

$$P_e(c, \widehat{c}, \theta) \triangleq \mathbb{E}^{(\theta)}\left[\mathbb{1}[c(\mathbf{T}) \neq \widehat{c}(\mathbf{S})]\right], \qquad (6)$$

where the expectation is taken with respect to the joint distribution of text and summary $(\mathbf{T}, \mathbf{S})$ based on the source texts distribution and the distribution over the summaries induced by the summarizer. Interestingly, it is not difficult to check by data-processing inequality that (4): $I(C;\mathbf{S}) \leqslant I(C;\mathbf{T})$ and thus, the output summaries $\mathbf{S}$ by the summarization model may not preserve all relevant information necessary to predict $C$ from $\mathbf{S}$, unless the summarization model is aware of the downstream task. However, identifying the set of relevant downstream tasks for various texts is challenging in practice. Consequently, (6) does not provide a satisfactory evaluation metric for summarization systems. This observation raises the central question studied in this work: **How much information about any downstream task is preserved in the summaries?**

## 4 A Task-Agnostic Quality Measure for Summarization

### 4.1 Theoretical Results

In this section, we rigorously motivate the evaluation of summarization systems by measuring the MI between texts and the resulting summaries. Let's assume the existence of a random variable $\mathbf{T}$ representing source texts, which follows an unknown distribution inherent to the source text-domain. Given a stochastic summarization system $p_\theta(\mathbf{s}|\mathbf{t})$ as defined in expression (1), we denote $\mathbf{S} \sim p_\theta(\mathbf{s}|\mathbf{T})$ the random variable representing summaries generated by the summarization system for these sources texts $\mathbf{T}$.

Consider the task $c(\mathbf{T})$ that we intend to execute on the source texts, where $C = c(\mathbf{T})$ represents the random variable denoting the outcomes of this task on source texts. We assume that the same task can be performed on the summaries and let $\widehat{c}(\mathbf{S})$ denote the corresponding prediction. Intuitively, if we can accurately predict the outcome of $C$ from the summaries, then we will say that the summarization system is high-quality since it preserves the necessary information for performing the task. The next proposition frames our information-theoretic bounds on the performance of any arbitrary downstream task. In particular, it shows that the expected error rate $P_e(c, \widehat{c}, \theta)$ as defined in (6) can be upper and lower bounded by the MI between the texts and the summaries.

[2]Other goals are possible such as reducing the source text complexity, removing specific information etc... Our choice stems from the practical usability of the task-oriented definition for evaluation purposes

**Proposition 1** (Information-theoretic bounds)**.** *Let $C = c(\mathbf{T})$ denote the underlying concept variable and let $\widehat{c}(\mathbf{S})$ be the Bayes predictor of $C$ observing the output summaries $\mathbf{S}$, based on the underlying summarization model $p_\theta(\mathbf{s}|\mathbf{t})$. The expected error rate satisfies:*

$$P_e(c, \widehat{c}, \theta) \leqslant 1 - \kappa \exp\big(I(\mathbf{T}; \mathbf{S})\big), \quad (7)$$

$$P_e(c, \widehat{c}, \theta) \geqslant R^{-1}_{C,\ell_{01}}\big(I(\mathbf{T}; \mathbf{S})\big), \quad (8)$$

*where $\kappa \in (0, 1)$ is a constant which does not depend on summaries; and $R^{-1}_{C,\ell_{01}}(\cdot)$ is the inverse of the rate-distortion function using $\ell_{01}(c, \widehat{c}) = \mathbb{1}[c \neq \hat{c}]$. Furthermore, the lower bound holds for an arbitrary loss $\ell(\cdot, \cdot)$ measuring the disagreement between the concept and its predictive value:*

$$\inf_{\widehat{c}(\cdot)} \mathbb{E}[\ell(C, \widehat{c}(\mathbf{S}))] \geqslant R^{-1}_{C,\ell}\big(I(\mathbf{T}; \mathbf{S})\big).$$

*Proof.* The upper bound (7) relies on the fact that the predictor $\widehat{c}(\mathbf{S})$ is the optimal (Bayes) classifier for which the expected error rate admits a well-known expression. The lower bound (8) uses data-processing inequality which implies that $I(\mathbf{T}; \mathbf{S}) \geqslant I\big(c(\mathbf{T}); \widehat{c}(\mathbf{S})\big)$ and the definition of the rate-distortion function evaluated in the loss $\ell_{01}(c, \widehat{c})$ noticing that its expectation yields the expected error rate $P_e(c, \widehat{c}, \theta)$ as defined in (6). For further details, see Appendix A. □

*Remark* 1. The bounds in Proposition 1 imply that the expected error rate of any task predicted by observing the summaries is lower- and upper-bounded by functions of the MI between the text and the summaries $I(\mathbf{T}; \mathbf{S})$. More precisely, the upper bound (7) is a monotonically decreasing function of the MI while the lower bound (8) is a non-increasing function in the MI. The lower bound can be further simplified by evaluating the rate-distortion function, as shown in Appendix B. In other words, **greater MI corresponds to improved the expected prediction performance on the summaries. Conversely, the expected error rate is bounded from above when MI is limited.** Interestingly, the arguments regarding the bounds do not depend on the considered task, suggesting that the MI can be used as a task-agnostic metric to evaluate the quality of summaries.

In the next section, we propose a practical method to estimate MI. In Section 6, we empirically show that it correlates well with the performance of the downstream tasks with other human judgment-based metrics and compare it to standard metrics.

### 4.2 Estimating MI from samples

While the MI captures an intuitive notion of information and is theoretically grounded, estimating it accurately is notoriously challenging. Therefore, following prior research (Ethayarajh et al., 2022; Xu et al., 2020) we estimate Arimoto information (Arimoto, 1971) based on the KNIFE estimator (Pichler et al., 2022). Our method comprises three steps: (1) we project the source texts and the summaries into a continuous embedding space; (2) we fit a mutual information estimator onto these embeddings; and (3) we estimate the mutual information between the source texts and the summaries using the fitted estimator. We report the details of our method in Algorithm 1.

**Mutual information estimator.** We rely on the KNIFE estimator (Pichler et al., 2022) to estimate the differential entropy of the embeddings and then mutual information. This estimator effectively relies on Gaussian Mixtures with $K$ modes to fit the density function and induces a soft-clustering for text generation evaluation (Pillutla et al., 2021; Pimentel et al., 2023). We found experimentally that the number of modes $K$ did not impact the performance significantly. We report the results for $K = 4$ (see Appendix D.1 for further details).

**Embeddings.** In order to obtain continuous representations of the source texts and the summaries, we mainly rely on the AnglE-embedders (Li and Li, 2023) as they are at the top of the MTEB leaderboard (Muennighoff et al., 2023) with a rather small model. In addition, we experimented with different embedders from the sentence transformers library (Reimers and Gurevych, 2019), mainly the paraphrase and sentence similarity embeddings. We find that our method was robust in the choice of the embedder and thus reported the results for the AnglE-embedders.

**Algorithm 1** Evaluating the performance of a summarizer using KNIFE and sentence transformers

**Input:** A dataset $\mathcal{D}_N = \{(\mathbf{T}_i, \mathbf{S}_i)\}_{i=1}^N$
**Input:** A pre-trained embedder Emb
**Output:** An estimation of the mutual information between $\mathbf{T}$ and $\mathbf{S}$

$\mathbf{E_T} \leftarrow \{\text{Emb}(\mathbf{T}_i)\}_{i=1}^N$
$\mathbf{E_S} \leftarrow \{\text{Emb}(\mathbf{S}_i)\}_{i=1}^N$
$\widehat{I}_N(\mathbf{T}; \mathbf{S}) \leftarrow \text{KNIFE}(\mathbf{E_T}, \mathbf{E_S})$
**return** $\widehat{I}_N(\mathbf{T}; \mathbf{S})$

## 5 Experimental Settings

While the bounds derived in Section 4.1 provide certain theoretical guarantees regarding the mutual information, their practical consequences must still be evaluated empirically. We present here two evaluation methods. First, we show that MI effectively predicts whether performing a downstream on the summaries would lead to the same outcome as if it were performed on the source text. This validates our task-oriented evaluation approach for summaries. Furthermore, we contrast our metric with metrics trained to emulate human judgments across various dimensions. Remarkably, we observe that even without any training, our metric closely aligns with human preferences.

**Datasets.** We select three summarization datasets for the English language: CNN/DailyMail (See et al., 2017; Hermann et al., 2015), XSum (Narayan et al., 2018) and MultiNews (Fabbri et al., 2019) and perform all evaluations on all datasets. We provide additional experiments in French and Spanish using the MLSUM (Scialom et al., 2020) and XLSUM (Hasan et al., 2021) datasets. We report mixed results due to the lack of efficient multilingual embedders in Section E.2.

**Models.** We evaluate numerous summarizers from the HuggingFace hub, relying on different backbones, pretraining methods and finetuned on different datasets. We conduct our experiments on the PEGASUS suite of models (Zhang et al., 2020a), on BERT large models (Lewis et al., 2019) and on DistilBERT models (Shleifer and Rush, 2020). We also evaluated the models presented in the SEAHORSE benchmark (Clark et al., 2023) through the generated summaries they proposed, which include MT5 models (Xue et al., 2021), different variants of T5 models (Raffel et al., 2023), and of the PaLM models (Chowdhery et al., 2022). The reader is referred to Appendix C for models' details.

**Baseline metrics for quality assessment.** For all summaries, we evaluate the quality estimation metrics obtained by computing reference-free versions of the `ROUGE-L`, `BERTScore` (Zhang et al., 2020b) and `BARTScore` (Yuan et al., 2021) between the source texts and the summaries. We average over the dataset to get a summarizer-level score that we can compare with the MI.

**Human evaluation with SummEval.** While the SummEval (Fabbri et al., 2021) dataset provides only a very limited number of human judgments, it contains enough aligned source, and AI-generated summaries to evaluate the MI metric. We use these unannotated data to evaluate the MI and rank the model accordingly; then we compare this ranking with the human evaluation provided by the SummEval benchmark. We provide a comparison with previous metrics in Table 2. It is worth noting that the evaluation is not fair for our metric as other are evaluated only on the annotated data, while ours is evaluated on the unannotated data.

### 5.1 Downstream tasks

To demonstrate the efficacy of our metric in predicting the downstream task performance on the summaries, the ideal scenario would involve generating summaries and tasking various individuals with different tasks across diverse contexts. However, executing this at a larger scale is impractical. As an alternative, we suggest comparing the results of different algorithms (classifiers and embedders) applied to both the source texts and the generated summaries. A summarizer can be deemed effective if these algorithms yield consistent results when applied to both the summaries and the source texts.

**Classification tasks.** We selected four different tasks (sentiment analysis, policy classification, Emotion classification and ChatGPT detector) and corresponding classifiers from the Huggingface Hub to run on the source texts and summaries and compared their outputs. We report the expected error rate(6), *i.e.*, the classifier outputting a different label on the source text and the summary.

**Embedders.** We compare the embeddings obtained from the source texts and the summaries from different models, the output of the embedding by the classifiers models, and paraphrase and sentence similarity embeddings (Reimers and Gurevych, 2019). We show the correlation of our metrics with the cosine similarity between the embeddings of the source texts and the corresponding summaries. Since embedders are supposed to abstract the information in the texts, good summarization models should produce embeddings close to the source texts' embeddings.

### 5.2 Correlation With Learnt Metrics

There are not many available metrics trained on human judgment for summarization. We chose to evaluate our metric on the SEAHORSE metrics (Clark et al., 2023) as they are the most recent and provide interpretable metrics.

Table 1: Common quality estimation metrics correlation with the performance on the downstream classification tasks. Where Sent. analysis stands for sentiment analysis, GPT det. for GPT detector, Topic. for topic classification, Policy for policy classification, Emotion for emotion classification and Emb. for paraphrase embedding.

| | Metric | Sent. analysis | GPT det. | Policy | Emotion | Emb. |
|---|---|---|---|---|---|---|
| | $I(S;T)$ | 0.63 | 0.59 | 0.58 | 0.56 | 0.81 |
| Bas. | BERTScore Precision | 0.53 | 0.68 | 0.47 | 0.46 | 0.70 |
| | BERTScore Recall | 0.59 | 0.66 | 0.74 | 0.54 | 0.42 |
| | BLANC | 0.59 | 0.59 | 0.66 | 0.60 | 0.38 |
| | SMART1 | 0.55 | 0.63 | 0.47 | 0.47 | 0.28 |
| | SMART2 | 0.55 | 0.63 | 0.47 | 0.47 | 0.28 |
| | SMARTL | 0.55 | 0.63 | 0.47 | 0.47 | 0.28 |
| | BARTScore | 0.51 | 0.73 | 0.42 | 0.48 | 0.62 |
| | BERTScore | 0.64 | 0.75 | 0.64 | 0.58 | 0.61 |
| | ROUGE-L | 0.56 | 0.55 | 0.54 | 0.47 | 0.29 |
| SH. | Attribution | 0.49 | 0.71 | 0.37 | 0.49 | 0.62 |
| | Main ideas | 0.33 | 0.37 | 0.60 | 0.38 | 0.47 |

**Seahorse metrics.** These metrics — learnt from human judgement — assess summaries along 6 axes: **Main ideas; Attribution; Comprehensible; Grammar; Repetition and Conciseness**. `Main ideas` and `Attribution`, `respectively`, measure if the main ideas of the source text are present in the summary and if the information in the summary does indeed come from the source text. `Comprehensible` and `Grammar` measure if the summary is fluent and grammatically correct, while `Repetition` and `conciseness` measure if there are NO repetitions and the summary is concise. The $\Pr(\text{Yes})$ metric corresponds to the average probability over the dataset that the SEAHORSE model predicts the answer *Yes* to the corresponding question. While we would not expect our MI metric to correlate with the Grammar or Comprehensible scores, it should strongly correlate with the Main Ideas and Attribution scores as these are proxies to the information contained in the summary.

## 6 Experimental Results

**Correlation with downstream tasks performance.** In Table 1, we report the correlation between the different metrics and the expected error rate for different classification tasks and with the dot product for the Paraphrase embedding task. The MI is competitive with both the common quality estimation metrics such as `BERTScore` and `BARTScore`. In addition, we report the correlation of the metrics trained on human judgement from the SEAHORSE benchmark to predict whether the main idea of the text is present in the summary and if all elements of the summary are attributable to the source text.

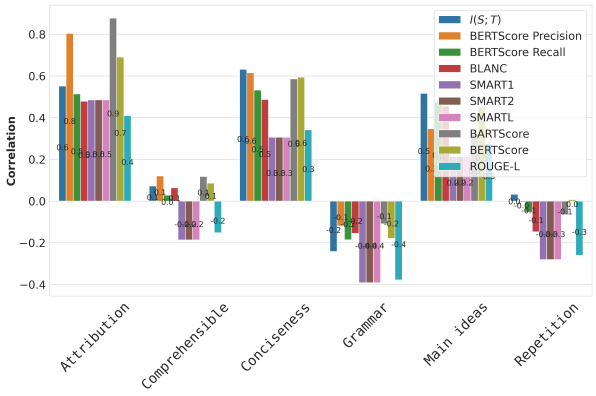


Figure 2: Spearman correlation with human judgment estimated by the SEAHORSE metrics. As one would expect the MI does correlate with `Attribution` and `Main ideas` but not with `comprehensible`, `grammar` or `repetition`.

**Correlation with human-judgement-based metrics.** As the SEAHORSE metrics are trained to mimic human judgment, we can use them to assess the behavior of the MI. We found consistent and expected correlations with the different SEAHORSE metrics (Figure 2). **The mutual information correlates well with `Attribution` and `Main Ideas` but not with `Comprehensible`, `Grammar` and `Repetition`.** This is not surprising as one would expect the mutual information to capture the amount of information in the summary that is attributable to the source text and not the grammatical correctness or fluency of the summary. The high correlation with `Conciseness` is a rather surprising result. We believe this is because the conciseness correlates with the strength of the summarizer, which correlates with the MI between the source texts and the summaries. The stronger the language model, the better the source texts' encoding as a summary. It is also plausible that the learned metrics are flawed in some ways, hindering the results.

We observe that the MI correlates more consistently with these `Main idea`, `Attribution` and `Conciseness` scores than the common quality assessment metrics, e.g. `BERTScore` and `BARTScore` (see Figure 2). It suggests that in addition to being a good predictor of the performance of the downstream tasks, MI is also a better predictor of the human judgment of the quality of the summaries. Notably, the MI, a theoretical quantity derived from Shannon's MI, reproduces human judgment expectations without training on human judgment data.

**Comparing summarizers with `COSMIC`.** In Figure 3, we compare the MI of different models and report their size. We observe that OOD models

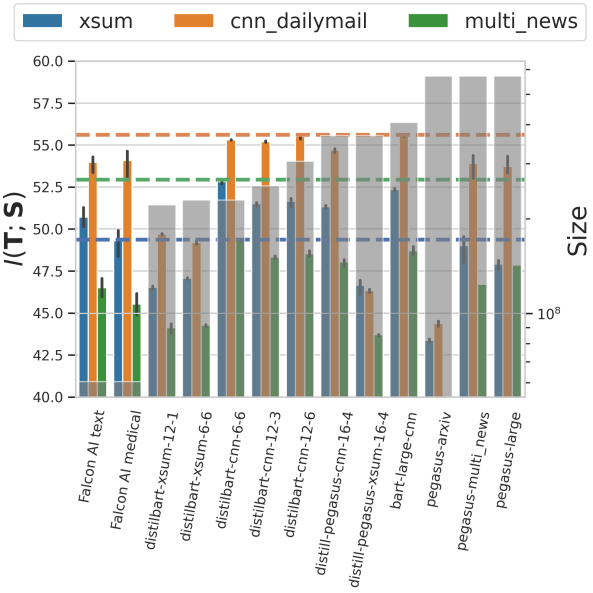


Figure 3: Comparison of models of different sizes and strengths, finetuned on different datasets regarding the MI between their summaries and the source text on different evaluation datasets. The dotted lines highlight the highest MI for each dataset reached by a model, and the grey area represents the number of parameters of the model

| Metric | Coher. | Cons. | Flu. | Rel. |
|---|---|---|---|---|
| | Reference Dependent | | | |
| ROUGE-1 | .35 | .55 | .53 | .58 |
| ROUGE-2 | .23 | .60 | 49 | .43 |
| ROUGE-L | .12 | .12 | .26 | .35 |
| BLEU | .22 | .05 | .33 | .38 |
| CHRF | .35 | .63 | .56 | .55 |
| BERTScore | .33 | -.03 | .14 | .20 |
| MoverScore | .23 | -.05 | .26 | .35 |
| BLEURT | .53 | .20 | .41 | 47 |
| SMS | .27 | 60 | .36 | .40 |
| SMART-1 | .43 | 67 | .64 | .67 |
| SMART-2 | .42 | .75 | .63 | .58 |
| SMART-L | .57 | .57 | .61 | .73 |
| | Reference Free | | | |
| PRISM | .23 | .60 | .36 | .37 |
| T5-ANLI | .25 | .58 | .54 | .52 |
| BARTScore | .35 | .62 | .49 | 45 |
| BARTScore+CNN | .55 | .32 | .59 | .58 |
| $Q^2$ | .25 | .75 | .58 | .45 |
| RISE $_{\text{extMulti-News}}$ | .53 | .73 | .71 | .70 |
| RISE $_{\text{SamSUM}}$ | .53 | .70 | .68 | .70 |
| $\text{RISE}_{CNN}$ | .53 | .73 | .75 | .70 |
| | Ours | | | |
| $I(\mathbf{T};\mathbf{S})$ | .23 | .53 | .47 | .54 |

Table 2: Comparison of our method against many baselines on the SummEval Human evaluation dataset. We report the system-level Kendall's Tau correlation with human judgments.

— trained on Arxiv or medical data — perform very poorly, whereas IN distribution models such as BART display significantly higher MI. Interestingly enough, the size of the model does not seem to be a good predictor of MI. In Appendix F we explore further use of the mutual information to compare the informativeness of different summarizers between themselves to construct a hierarchy based on their mutual informativeness.

## 7 Discussion of Quantitative Results

Our findings reveal that various conventional metrics do not consistently align with the effectiveness on downstream tasks. As illustrated in Figure 4, different metrics exhibit distinct behaviors and correlations with other metrics and the performance of downstream tasks.

**SEAHORSE Metrics.** Notably, most SEAHORSE metrics demonstrate limited correlations with the effectiveness of downstream tasks. Unexpectedly, the `Main idea` metric performs less effectively compared to the `Attribution` metric.

**`BERTScore`.** `BERTScore` displays good correlation with the task-preserving capabilities of the summaries but have poor correlations with human judgements (real or estimated); whereas the mutual information is theoretically grounded and an all-around more consistent metric.

**Behavior of MI.** The MI displays very similar behavior to the metric trained to detect whether the ideas presented in the summary come from the source text (Attribution) (Figure 4), but it surprisingly does not follow the same behavior as the metric trained to detect if the main idea is present (Main idea). Coincidentally, the main idea metric does not correlate with the expected error rate of the classification tasks. This may be an artifact of the training of SEAHORSE benchmark, a limitation of our metric, or could indicate that answering the question "Is the content of the summary fully attributable to the content of the source text" is more relevant for downstream tasks than "Is the main idea of the source text present in the summary".

**Independence of the metrics.** While it seems that `Attribution` and the MI follow a similar trajectory in Figure 4, we found that the MI is not correlated with the `Attribution` metric (Figure 5). This suggests that these two metrics are independent and could be used in conjunction to evaluate summarizers. When it comes to more standard metrics, we find two clusters of metrics that seem to be relatively independent of each other: one represented by `BERTScore` and comprising the MI and the other represented by `BARTScore`, which includes `Attribution`.

**Correlations with Human Judgement.** While

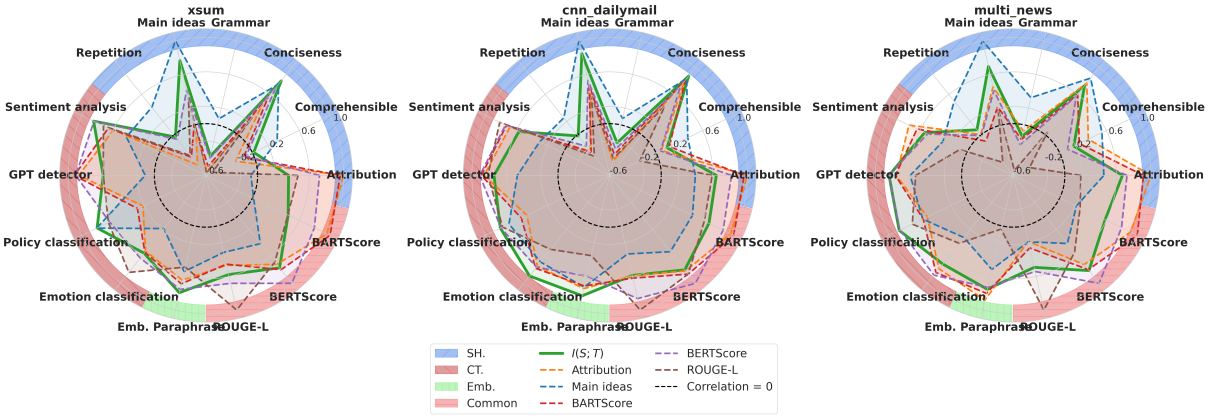


Figure 4: Correlation between the main metrics and different performance metrics for the different datasets. The MI (green) closely follows the behavior of the Attribution metric but is a better predictor of the performance of the downstream tasks (Sentiment analysis, GPT Detector, Topic classification, etc.). By contrast, ROUGE scores do not display consistent correlations. For metrics to be considered effective, they should consistently exhibit positive correlations with each other. In Section E.3 we report the aggregated numerical results.

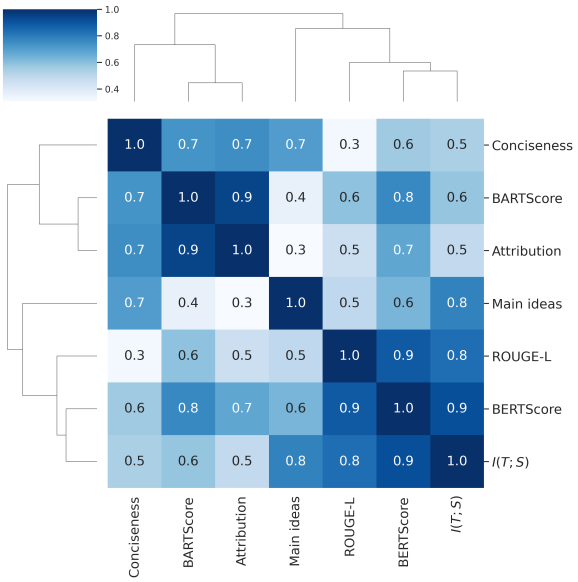


Figure 5: The correlation matrix illustrates the relationships among various metrics, with clustering based on their correlation similarity. This clustering indicates the degree of similarity between metrics in terms of their correlation with each other. The goal is to have a diverse set of grounded yet independent metrics that assess different aspects of text summarization quality.

the MI correlates less on the SummEval judgment dataset than competitors, such as SMART or RISE, it correlates more with the downstream tasks performance (Table 1) while not relying on golden summaries for comparison (SMART) or on very large pre-trained models (RISE). As mentioned previously, the comparison with other metrics is not completely fair as the MI is evaluated on unannotated data while the other metrics are evaluated on the (limited) annotated data. Nonetheless, the MI displays on-par performance with the other metrics on the SummEval dataset and is a promising metric for summarizer evaluation overall.

## 8 Summary and Discussion

We introduced a task-oriented setting for summary evaluation in which we can derive a principled and clear notion of the quality of a summarizer: the expected risk of performing a task on a summary instead of the source text. We connected this risk theoretically to the mutual information of the source texts and generated summaries and we bounded the risk on both sides using the mutual information. Even if these bounds are task-agnostic and thus potentially loose, we demonstrated experimentally that the mutual information indeed correlated well with the risk. High mutual information indicates that a task performed in the summaries is likely to produce the same outcome as in the source texts. COSMIC is therefore theoretically grounded in a reasonable task-oriented scenario. Our ability to estimate this mutual information practically and its correlation with downstream task performance further underscores its significance. Our proposed method extends beyond summarization systems and could also contribute to the broader field of multi-modal generation evaluation (Kim et al., 2022).

## Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011013290R2 made by GENCI.

## 9 Limitations & Ethical considerations

We have introduced a novel evaluation setting and a theoretically grounded metric for assessing summarizers, yet both have their limitations. Firstly, our setting assumes that the sole objective of a summary is to facilitate downstream task performance, and we define a good summary as one that preserves task outcomes. However, summaries can serve multiple purposes, such as aiding comprehension, acting as educational aids, or promoting the source text, which we do not account for in our approach.

While mutual information is theoretically grounded, it is not without flaws and fails to capture all nuances of the summarization task. It serves as a tool to evaluate a summarizer's informativeness compared to other metrics lacking theoretical grounding.

It is imperative to use mutual information in conjunction with other metrics to evaluate summaries comprehensively, as it solely addresses the informativeness of summaries about their source text. This metric does not assess grammaticality. Consequently, high mutual information values may arise from imperceptible artefacts that render the summary highly informative about the source text yet unintelligible to human readers.

Moreover, our method indirectly evaluates mutual information in the continuous domain by assessing the mutual information between embeddings generated by a fixed language model. The choice of this model significantly impacts mutual information estimation and the parameters of the estimation tool used.

## References


Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Suguru Arimoto. 1971. Information-theoretical considerations on estimation problems. *Information and control*, 19(3):181–194.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of english than into it: Measuring neural translation difficulty by cross-mutual information.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur P. Parikh. 2023. Seahorse: A multilingual, multifaceted dataset for summarization evaluation.

T. M. Cover and J. A. Thomas. 2006. *Elements of Information Theory*, 2nd edition. Wiley, New York, NY.

I Csiszár. 1974. On an extremum problem of information theory. *Studia Scientiarum Mathematicarum Hungarica*, 9(1):57–71.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nicholas Egan, Oleg Vasilyev, and John Bohannon. 2021. Play the shannon game with language models: A human-free approach to summary evaluation.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th*

*Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation.

George Giannakopoulos, Vangelis Karkaletsis, George Vouros, and Panagiotis Stamatopoulos. 2008. Summarization system evaluation revisited: N-gram graphs. *ACM Trans. Speech Lang. Process.*, 5(3).

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

M Indu and KV Kavitha. 2016. Review on text summarization evaluation methods. In *2016 international conference on research advances in integrated navigation systems (RAINS)*, pages 1–4. IEEE.

Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. 2022. Mutual information divergence: A unified metric for multimodal generative models.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

Vishakh Padmakumar and He He. 2021. Unsupervised extractive summarization using pointwise mutual information.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Georg Pichler, Pierre Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. 2022. A differential entropy estimator for training neural networks.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers.

Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2023. On the usefulness of embeddings, clusters and strings for text generator evaluation.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Is summary useful or not? an extrinsic human evaluation of text summaries on downstream tasks.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alfréd Rényi. 1961. On measures of entropy and information.

Saiyed Saziyabegum and PS Sajja. 2017. Review on text summarization evaluation methods. *Indian J. Comput. Sci. Eng*, 8(4):497500.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. Mlsum: The multilingual summarization corpus. *arXiv preprint arXiv:2004.14900*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Sam Shleifer and Alexander M. Rush. 2020. Pre-trained summarization distillation.

Ce Sui, Xiaosheng Zhao, Tao Jing, and Yi Mao. 2023. Evaluating summary statistics with mutual information for cosmological inference.

Stephen Tratz and Eduard H Hovy. 2008. Summarization evaluation using transformed basic elements. In *TAC*. Citeseer.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618. Text Summarization.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert.

## A Proof of the Upper and Lower Bounds

### A.1 Proof of the upper bound on the expected error rate

We begin by noticing that

$$\begin{aligned} 1 - \sum_{c \in \mathcal{C}} p^2(c|\mathbf{s}) &= 1 - \mathbb{E}\left[p(C|\mathbf{S})|\mathbf{S} = \mathbf{s}\right] \\ &\geqslant 1 - \mathbb{E}\left[\max_{c \in \mathcal{C}} p(c|\mathbf{s}) \middle| \mathbf{S} = \mathbf{s}\right] \\ &= 1 - \max_{c \in \mathcal{C}} p(c|\mathbf{s}). \end{aligned} \tag{9}$$

By taking the expectation over $\mathbf{S}$ at both sides and using the well-known relationship with the Bayes error, we obtain the following inequality:

$$\begin{aligned} P_e(c, \widehat{c}, \theta) &= 1 - \mathbb{E}\left[\max_{c \in \mathcal{C}} p(c|\mathbf{S})\right] \\ &\leqslant 1 - \mathbb{E}\left[\sum_{c \in \mathcal{C}} p^2(c|\mathbf{S})\right]. \end{aligned} \tag{10}$$

Similarly, it is possible to derive a lower bound:

$$\sum_{c \in \mathcal{C}} p^2(c|\mathbf{s}) = p^2(c^\star|\mathbf{s}) + \sum_{c \neq c^\star} p^2(c|\mathbf{s}) \geqslant \left(\max_{c \in \mathcal{C}} p(c|\mathbf{s})\right)^2, \tag{11}$$

where $c^\star(\mathbf{s}) = \arg\max_{c \in \mathcal{C}} p(c|\mathbf{s})$. By taking the expectation over $\mathbf{S}$ at both sides, we obtain:

$$\begin{aligned} \sqrt{\mathbb{E}\left[\sum_{c \in \mathcal{C}} p^2(c|\mathbf{S})\right]} &\geqslant \mathbb{E}\left[\sqrt{\sum_{c \in \mathcal{C}} p^2(c|\mathbf{S})}\right] \\ &\geqslant \mathbb{E}\left[\max_{c \in \mathcal{C}} p(c|\mathbf{S})\right] = 1 - P_e(c, \widehat{c}, \theta). \end{aligned} \tag{12}$$

Let us denote the second order Rényi's entropy (Rényi, 1961) conditioned on $\mathbf{s}$ as follow:

$$H_2(C|\mathbf{s}) \triangleq -\frac{1}{2} \log\left(\sum_{c \in \mathcal{C}} p^2(c|\mathbf{s})\right), \tag{13}$$

and thus,

$$\sum_{c \in \mathcal{C}} p^2(c|\mathbf{s}) = \exp\left(-2H_2(C|\mathbf{s})\right). \tag{14}$$

By replacing (14) in (10) and in (11), we obtain

$$1 - \sqrt{\mathbb{E}_{\mathbf{s} \sim p_S}\left[\exp\left(-2H_2(C|\mathbf{s})\right)\right]} \leqslant P_e(c, \widehat{c}, \theta) \leqslant 1 - \mathbb{E}_{\mathbf{s} \sim p_S}\left[\exp\left(-2H_2(C|\mathbf{s})\right)\right]. \tag{15}$$

Since $\log$ is a concave function:

$$\log\left(\sum_{c \in \mathcal{C}} p^2(c|\mathbf{s})\right) \geqslant \sum_{c \in \mathcal{C}} p(c|\mathbf{s}) \log p(c|\mathbf{s}), \tag{16}$$

we have that

$$H_2(C|\mathbf{s}) \leqslant -\sum_{c \in \mathcal{C}} p(c|\mathbf{s}) \log p(c|\mathbf{s}) \triangleq H(C|\mathbf{s}), \tag{17}$$

where $H(C|\mathbf{s})$ indicates the Shannon entropy conditioned to the given observation $\mathbf{s}$. Replacing (17) in the upper bound of (15) yields

$$\begin{aligned} P_e(c, \widehat{c}, \theta) &\leqslant 1 - \mathbb{E}_{\mathbf{s}\sim\mathbf{S}}\left[\exp\left(-H(C|\mathbf{s})\right)\right] \\ &\leqslant 1 - \exp\left(-H(C|\mathbf{S})\right), \qquad (18) \\ &\leqslant 1 - \exp\left(-H(\mathbf{T}|\mathbf{S})\right), \qquad (19) \\ &\equiv 1 - \kappa \exp\big(I(\mathbf{T};\mathbf{S})\big), \qquad (20) \end{aligned}$$

where (18) follows from the fact that the negative exponential function is convex ; (19) follows by Data-Processing Inequality (Cover and Thomas, 2006) since $C \triangleq c(\mathbf{T})$ and thus $C \leftrightarrow \mathbf{T} \leftrightarrow \mathbf{S}$ form a Markov Chain and $H(\mathbf{T}|\mathbf{S})$ denotes the differential entropy of the text $\mathbf{T}$ given the summary $\mathbf{S}$ ; and (20) follows by an appropriate definition of the constant $0 < \kappa < 1$ which does not depend on the summary random variable $\mathbf{S}$. This concludes the proof of the desired upper bound.

### A.2 Review of the Distortion-Rate Function

The rate-distortion (RD) function of a random variable $C$ for a given distortion function $\ell(\cdot,\cdot)$ is defined as (Csiszár, 1974, eq. (1.4))

$$R_{C,\ell}(D) \triangleq \inf_{\substack{p(\widehat{c}|c):\\ \mathbb{E}[\ell(C,\widehat{C})]\leqslant D}} I(C;\widehat{C}). \qquad (21)$$

For convenience, we assume that

$$\inf_{\widehat{c}} \ell(c,\widehat{c}) = 0, \quad \forall c.$$

Furthermore, we suppose that there exists $D > 0$ such that $R_{C,\ell}(D)$ is finite. We denote the infimum of those $D$ by $D_{\min}$ and $R_{\max} \triangleq R_{C,\ell}(D_{\min})$ (or, more precisely, $R_{\max} \triangleq \lim_{D\to D_{\min}+} R(D)$).

The following properties (see (Csiszár, 1974, Lem. 1.1)) of the RD function will be used.

**Theorem 1.** *The RD function $R_{C,\ell}(D)$ is a non-increasing convex function of $D$ on the interval $(D_{\min}, \infty)$. It is monotonically decreasing on the interval $(D_{\min}, D_{\max})$ and constant with $R_{C,\ell}(D) = R_{\min}$ on $[D_{\max}, \infty)$ (here $D_{\max} = \infty$ and $D_{\min} = 0$ are possible).* ***The inverse function $R^{-1}_{C,\ell}(r)$ is well defined on $(R_{\min}, R_{\max})$ and is monotonically decreasing. It is known as the distortion rate (DR) function of the random variable $C$ for the given distortion function $\ell(\cdot,\cdot)$.***

### A.3 Proof of the lower bound on the average of a general loss

For any suitably loss or evaluation metric denoted by $\ell(c,\widehat{c})$, the quality of the predicted concept $\widehat{c}(\mathbf{S})$, which is based on the random summary $\mathbf{S}$, compared to a desired target concept $C \triangleq c(\mathbf{T})$ from the original text $\mathbf{T}$, can be expressed by the average loss $\mathbb{E}[\ell(c(\mathbf{T}), \widehat{c}(\mathbf{S}))]$ with respect to the joint distribution of the source text and its summary $(\mathbf{T}, \mathbf{S})$.

From Data-Processing Inequality and the definition of the RD function (21), the following proposition provides a lower bound on the performance of any arbitrary predictor $\widehat{c}(\mathbf{S})$ of the target concept $C$:

$$\begin{aligned} I(\mathbf{T};\mathbf{S}) &\geqslant I(c(\mathbf{T});\widehat{c}(\mathbf{S})) \qquad (22) \\ &\geqslant \inf_{\substack{p(\widehat{c}|c):\\ \mathbb{E}[\ell(C,\widehat{C})]\leqslant\mathbb{E}[\ell(c(\mathbf{T}),\widehat{c}(\mathbf{S}))]}} I(C;\widehat{C}) \qquad (23) \\ &= R_{C,\ell}\left(\mathbb{E}[\ell(c(\mathbf{T}),\widehat{c}(\mathbf{S}))]\right), \qquad (24) \end{aligned}$$

where the inequality in (22) follow from Data-Processing since $c(\mathbf{X}) \leftrightarrow \mathbf{X} \leftrightarrow \mathbf{S} \leftrightarrow \widehat{c}(\mathbf{S})$ form a Markov Chain ; and (23) follows by the definition of the RD function (21).

- For $\mathbb{E}[\ell(c(\mathbf{T}), \widehat{c}(\mathbf{S}))] \in (D_{\min}, D_{\max})$, we can invert the RD function (21), and thus we obtain from (22) the fundamental bound $R^{-1}_{C,\ell}\left(I(\mathbf{T};\mathbf{S})\right) \leqslant \mathbb{E}[\ell(c(\mathbf{T}), \widehat{c}(\mathbf{S}))]$ or, equivalently,

$$\mathbb{E}[\ell(c(\mathbf{T}), \widehat{c}(\mathbf{S}))] \;\geq R^{-1}_{C,\ell}\big(I(\mathbf{T};\mathbf{S})\big), \tag{25}$$

  which holds for any predictor $\widehat{c}(\mathbf{S})$ and thus, for the one minimizing the left-hand size of (25).

- For $\mathbb{E}[\ell(c(\mathbf{T}), \widehat{c}(\mathbf{S}))] < D_{\min}$ equation (22) reduces to $I(\mathbf{T};\mathbf{S}) \geqslant +\infty$ which shows that to achieve an expected distortion below $D_{\min}$ the random variables $(\mathbf{T}, \mathbf{S})$ must have a joint distribution that is not absolutely continuous with respect to the product of their marginal distributions.

- For $\mathbb{E}[\ell(c(\mathbf{T}), \widehat{c}(\mathbf{S}))] \geqslant D_{\max}$ we obtain the trivial bound $I(\mathbf{T};\mathbf{S}) \geqslant 0$.

*Remark* 2. Inequality (25) shows that for arbitrary random concept $c(\mathbf{T})$ about the text to be inferred with $\widehat{c}(\mathbf{S})$ using the random summary $\mathbf{S}$ generated from $\mathbf{T}$, the expected loss of any predictor $\widehat{c}(\cdot)$ is lower bounded by a monotonically decreasing function of the mutual information between $\mathbf{T}$ and $\mathbf{S}$. **Thus, irrespective of the precise formulation of the loss function or the task defined by $c(\cdot)$ for execution on the summary, maximizing the mutual information $I(\mathbf{T};\mathbf{S})$ stands as a requisite condition for achieving commendable inference performance.** Our result suggests that estimating mutual information of a given summarizer can be a good proxy to assess its quality in the sense of preserving relevant information about concepts.

## B Rate-Distortion Bound for Classification Tasks

We assume that $C$ is uniformly distributed on the finite set $\mathcal{C} \triangleq \{1, 2, \ldots, m\}$, and we use the Hamming distortion function defined by

$$\ell(c, \widehat{c}) \triangleq \begin{cases} 0, & \text{if } c = \widehat{c} \\ 1, & \text{else}\,. \end{cases}$$

Note that the expected distortion equals the expected error rate, i.e.,

$$P_e(c, \widehat{c}, \theta) \triangleq \inf_{\widehat{c}\colon \Omega^* \to \mathcal{C}} \Pr\big(C \neq \widehat{c}(\mathbf{S})\big) = \inf_{\widehat{c}\colon \Omega^* \to \mathcal{C}} \mathbb{E}[\ell\big(C, \widehat{c}(\mathbf{S})\big)]\,. \tag{26}$$

The RD function is given by (Cover and Thomas, 2006, Problem 10.5) (not solved there)

$$R(D) = \begin{cases} \log m - H_b(D) - D\log(m-1), & \text{if } 0 \leqslant D \leqslant 1 - \frac{1}{m} \\ 0, & \text{if } 1 - \frac{1}{m} < D\,. \end{cases} \tag{27}$$

Here, $H_b(D)$ is the binary entropy function of $D$, i.e., $H_b(D) \triangleq -D\log(D) - (1-D)\log(1-D)$. Inserting (26) and (27) into (22), we obtain

$$I(\mathbf{T};\mathbf{S}) \;\geqslant\; \log m - H_b\big(P_e(c, \widehat{c}, \theta)\big) - P_e(c, \widehat{c}, \theta)\log(m-1)\,,$$

which is the well known Fano's inequality (Cover and Thomas, 2006, Th. 2.10.1). This can be put into the form of the general lower bound (25):

$$P_e(c, \widehat{c}, \theta) \;\geq\; R^{-1}_{C,\ell_{01}}\big(I(\mathbf{T};\mathbf{S})\big)\,.$$

However, a closed-form expression of $R^{-1}_{C,\ell_{01}}(I)$ is not available. The function is plotted for different $m$ in Figure 6.

## C Model Specifications

All the evaluated models are listed with their characteristics in Table 3.

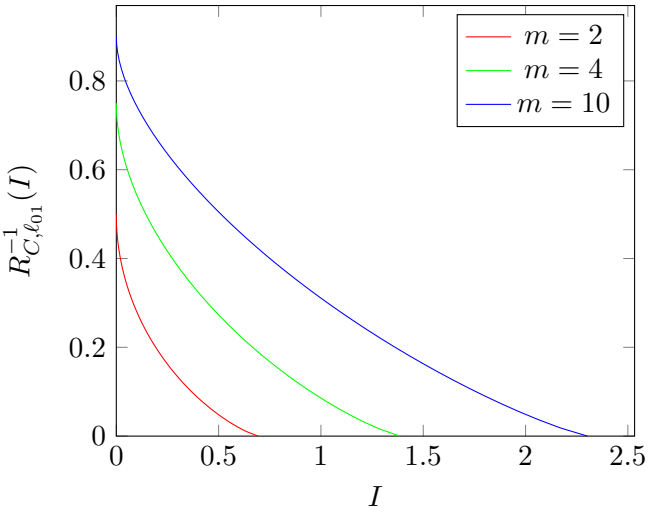


Figure 6: $R(I)$ for different values of $m$.

Table 3: Summary of the models we benchmarked with their name on the Huggingface hub, size and performance metrics.

| Model | Dataset | Size | ROUGE-L | BERTScore | BARTScore | M. I. | Attr. | Rep. | Compr. | Conc. | Gram. | $I(T,S)$ | H(T\|S) | H(S\|T) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Falconsai/medical_summarization | cnn_dailymail | 60 M | 0.17 | 0.16 | -1.70 | 0.33 | 0.79 | 0.46 | 0.76 | 0.42 | 0.30 | 54.09 | -14.60 | -7.24 |
| | multi_news | 60 M | 0.09 | 0.02 | -2.06 | 0.22 | 0.74 | 0.55 | 0.70 | 0.36 | 0.34 | 45.53 | -11.49 | -2.29 |
| | xsum | 60 M | 0.31 | 0.24 | -1.69 | 0.33 | 0.77 | 0.36 | 0.74 | 0.33 | 0.34 | 49.30 | -15.25 | -6.26 |
| Falconsai/text_summarization | cnn_dailymail | 60 M | 0.15 | 0.17 | -1.48 | 0.36 | 0.82 | 0.64 | 0.81 | 0.44 | 0.38 | 53.97 | -14.49 | -8.22 |
| | multi_news | 60 M | 0.10 | 0.06 | -1.67 | 0.23 | 0.80 | 0.60 | 0.75 | 0.38 | 0.38 | 46.50 | -12.46 | -4.00 |
| | xsum | 60 M | 0.31 | 0.29 | -1.53 | 0.33 | 0.81 | 0.58 | 0.77 | 0.35 | 0.37 | 50.71 | -16.67 | -8.84 |
| sshleifer/distilbart-xsum-12-1 | xsum | 221 M | 0.08 | 0.03 | -3.27 | 0.29 | 0.35 | 0.86 | 0.90 | 0.22 | 0.80 | 46.52 | -12.49 | 17.01 |
| | multi_news | 221 M | 0.03 | -0.09 | -3.08 | 0.17 | 0.42 | 0.82 | 0.77 | 0.23 | 0.68 | 44.13 | -10.10 | 16.22 |
| | cnn_dailymail | 221 M | 0.04 | -0.04 | -3.16 | 0.21 | 0.41 | 0.87 | 0.83 | 0.22 | 0.67 | 49.70 | -10.20 | 15.21 |
| sshleifer/distilbart-cnn-6-6 | cnn_dailymail | 229 M | 0.16 | 0.22 | -1.35 | 0.47 | 0.85 | 0.92 | 0.85 | 0.55 | 0.50 | 55.31 | -15.80 | -11.84 |
| | multi_news | 229 M | 0.10 | 0.15 | -1.37 | 0.34 | 0.86 | 0.90 | 0.84 | 0.52 | 0.57 | 49.34 | -15.31 | -8.99 |
| | xsum | 229 M | 0.32 | 0.35 | -1.30 | 0.47 | 0.84 | 0.91 | 0.90 | 0.49 | 0.60 | 52.81 | -18.77 | -13.27 |
| sshleifer/distilbart-xsum-6-6 | cnn_dailymail | 229 M | 0.04 | -0.02 | -2.88 | 0.38 | 0.55 | 0.96 | 0.96 | 0.41 | 0.87 | 49.19 | -9.68 | 12.37 |
| | multi_news | 229 M | 0.03 | -0.05 | -2.75 | 0.33 | 0.57 | 0.95 | 0.94 | 0.44 | 0.88 | 44.26 | -10.23 | 10.02 |
| | xsum | 229 M | 0.09 | 0.06 | -2.95 | 0.46 | 0.46 | 0.97 | 0.97 | 0.38 | 0.91 | 47.07 | -13.02 | 13.67 |
| sshleifer/distilbart-cnn-12-3 | xsum | 255 M | 0.31 | 0.32 | -1.58 | 0.49 | 0.74 | 0.82 | 0.94 | 0.47 | 0.64 | 51.50 | -17.45 | -11.65 |
| | cnn_dailymail | 255 M | 0.18 | 0.24 | -1.36 | 0.48 | 0.82 | 0.87 | 0.93 | 0.54 | 0.60 | 55.21 | -15.71 | -13.15 |
| | multi_news | 255 M | 0.10 | 0.13 | -1.49 | 0.36 | 0.82 | 0.86 | 0.92 | 0.53 | 0.66 | 48.32 | -14.28 | -9.03 |
| sshleifer/distilbart-cnn-12-6 | cnn_dailymail | 305 M | 0.18 | 0.25 | -1.31 | 0.52 | 0.84 | 0.91 | 0.93 | 0.58 | 0.62 | 55.51 | -16.00 | -13.79 |
| | multi_news | 305 M | 0.11 | 0.15 | -1.40 | 0.39 | 0.84 | 0.89 | 0.93 | 0.57 | 0.69 | 48.54 | -14.51 | -9.60 |
| | xsum | 305 M | 0.31 | 0.33 | -1.50 | 0.54 | 0.76 | 0.91 | 0.96 | 0.51 | 0.71 | 51.65 | -17.58 | -12.11 |
| sshleifer/distill-pegasus-xsum-16-4 | xsum | 369 M | 0.08 | 0.05 | -2.88 | 0.44 | 0.45 | 0.97 | 0.97 | 0.36 | 0.91 | 46.65 | -12.57 | 16.31 |
| | multi_news | 369 M | 0.03 | -0.06 | -2.60 | 0.28 | 0.56 | 0.95 | 0.93 | 0.39 | 0.86 | 43.70 | -9.68 | 16.49 |
| | cnn_dailymail | 369 M | 0.04 | -0.04 | -2.92 | 0.27 | 0.50 | 0.96 | 0.95 | 0.32 | 0.85 | 46.35 | -6.84 | 13.27 |
| sshleifer/distill-pegasus-cnn-16-4 | xsum | 369 M | 0.23 | 0.28 | -1.49 | 0.47 | 0.81 | 0.84 | 0.94 | 0.49 | 0.70 | 51.32 | -17.29 | -6.70 |
| | multi_news | 369 M | 0.08 | 0.11 | -1.44 | 0.33 | 0.84 | 0.82 | 0.90 | 0.51 | 0.69 | 48.03 | -13.99 | -4.63 |
| | cnn_dailymail | 369 M | 0.13 | 0.19 | -1.40 | 0.48 | 0.83 | 0.81 | 0.92 | 0.54 | 0.64 | 54.69 | -15.19 | -8.70 |
| facebook/bart-large-cnn | cnn_dailymail | 406 M | 0.18 | 0.25 | -1.19 | 0.49 | 0.86 | 0.95 | 0.94 | 0.59 | 0.67 | 55.54 | -16.05 | -13.55 |
| | multi_news | 406 M | 0.10 | 0.15 | -1.30 | 0.37 | 0.86 | 0.95 | 0.94 | 0.58 | 0.73 | 48.71 | -14.69 | -9.45 |
| | xsum | 406 M | 0.32 | 0.34 | -1.31 | 0.52 | 0.81 | 0.96 | 0.97 | 0.53 | 0.74 | 52.36 | -18.33 | -13.06 |
| google/pegasus-multi_news | multi_news | 570 M | 0.06 | 0.02 | -2.51 | 0.48 | 0.69 | 0.93 | 0.79 | 0.52 | 0.58 | 46.71 | -12.68 | -4.24 |
| | xsum | 570 M | 0.27 | 0.20 | -2.58 | 0.53 | 0.49 | 0.94 | 0.85 | 0.39 | 0.72 | 49.01 | -15.00 | -10.98 |
| google/pegasus-arxiv | xsum | 570 M | 0.14 | -0.22 | -3.52 | 0.11 | 0.32 | 0.25 | 0.43 | 0.13 | 0.28 | 43.38 | -9.33 | 0.79 |
| google/pegasus-large | cnn_dailymail | 570 M | 0.20 | 0.25 | -1.21 | 0.26 | 0.87 | 0.66 | 0.84 | 0.43 | 0.51 | 53.73 | -14.21 | -9.28 |
| | multi_news | 570 M | 0.07 | 0.12 | -1.52 | 0.18 | 0.82 | 0.82 | 0.72 | 0.37 | 0.42 | 47.86 | -13.81 | -5.25 |
| | xsum | 570 M | 0.27 | 0.31 | -1.19 | 0.24 | 0.88 | 0.86 | 0.89 | 0.37 | 0.70 | 47.92 | -13.89 | -4.18 |
| google/pegasus-multi_news | cnn_dailymail | 570 M | 0.16 | 0.16 | -2.26 | 0.47 | 0.63 | 0.94 | 0.83 | 0.43 | 0.65 | 53.89 | -14.39 | -13.12 |
| google/pegasus-arxiv | cnn_dailymail | 570 M | 0.10 | -0.26 | -3.30 | 0.08 | 0.35 | 0.26 | 0.45 | 0.12 | 0.30 | 44.37 | -4.86 | -0.28 |
| | multi_news | 570 M | 0.05 | -0.27 | -3.59 | 0.06 | 0.32 | 0.36 | 0.46 | 0.12 | 0.39 | 39.87 | -5.82 | 1.21 |

# D Mutual Information Estimation with KNIFE

## D.1 Predictive mutual information

The estimation of mutual information is widely acknowledged to be challenging, and in practical scenarios, we often resort to approximating it with a proxy measure known as Arimoto information (Arimoto, 1971) or recently rediscovered as predictive mutual information (Xu et al., 2020). Instead of computing the mutual information in the general case, it is computed under computational constraints enforced by a class

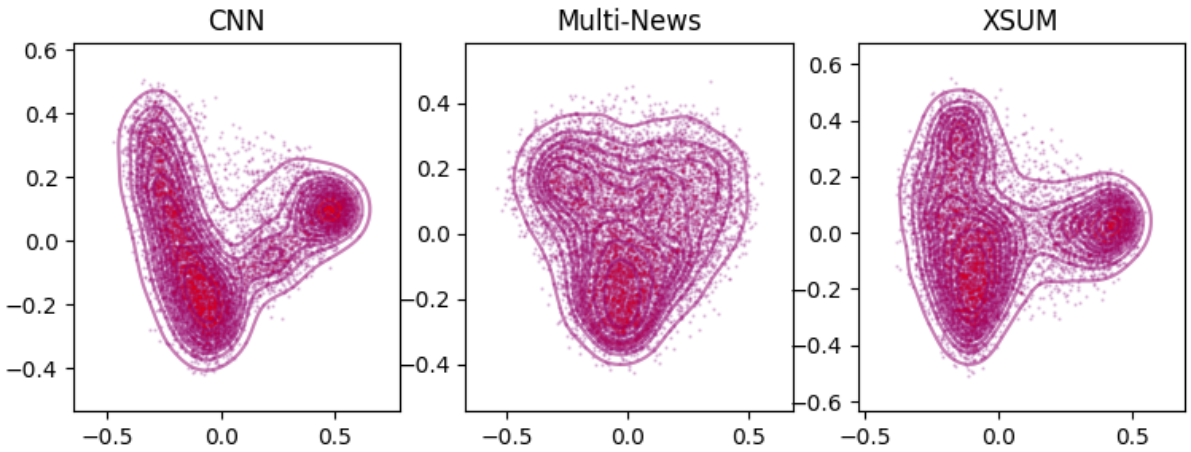


Figure 7: PCA was performed on the embeddings of the source texts and summaries for three datasets under consideration. It is evident from the plots that the embeddings do not exhibit a Gaussian distribution but rather resemble a mixture of Gaussians. This characteristic makes the Gaussian estimator of MI unsuitable for our purposes.

of predictive functions.

**Definition 1** (Predictive conditional entropies)**.** Let $\mathbf{T}$ and $\mathbf{S}$ be two random variables respectively over $\Omega^*$ and a class of functions $\mathcal{F}$:

$$h_{\mathcal{F}}(\mathbf{T} \mid \mathbf{S}) = \inf_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{TS}}[-\log f_{[\mathbf{S}]}(\mathbf{T})],$$
$$h_{\mathcal{F}}(\mathbf{T} \mid \varnothing) = \inf_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{T}}[-\log f_{[\varnothing]}(\mathbf{T})].$$

For the sake of brevity, we denote $h_{\mathcal{F}}(\mathbf{T} \mid \varnothing)$ by $h_{\mathcal{F}}(\mathbf{T})$.

**Definition 2** (Predictive $\mathcal{F}$-information)**.**

$$I_{\mathcal{F}}(\mathbf{S} \to \mathbf{T}) \triangleq h_{\mathcal{F}}(\mathbf{T}) - h_{\mathcal{F}}(\mathbf{T} \mid \mathbf{S}), \tag{28}$$
$$I_{\mathcal{F}}(\mathbf{T} \to \mathbf{S}) \triangleq h_{\mathcal{F}}(\mathbf{S}) - h_{\mathcal{F}}(\mathbf{S} \mid \mathbf{T}). \tag{29}$$

If $\mathcal{F}$ represents the set of all possible functions, then we expect $I_{\mathcal{F}}(\mathbf{T} \to \mathbf{S}) = I_{\mathcal{F}}(\mathbf{S} \to \mathbf{T}) = I_{\mathcal{F}}(\mathbf{S}; \mathbf{T})$. However, due to computational limitations imposed by $\mathcal{F}$, these estimators are not symmetrical. Therefore, we have two options for estimating the mutual information.

*Remark* 3. The predictive mutual information is asymmetric with respect to $\mathbf{S}$ and $\mathbf{T}$. In our context, we opt for using the predictive mutual information $I_{\mathcal{F}}(\mathbf{S} \to \mathbf{T})$ to gauge the degree of information preservation about the source texts through the summarization process. Thus, we define $\widehat{I}(\mathbf{T}; \mathbf{S}) \equiv I_{\mathcal{F}}(\mathbf{S} \to \mathbf{T})$. Experimentally, we observed that the predictive mutual information $I_{\mathcal{F}}(\mathbf{T} \to \mathbf{S})$ did not yield consistent outcomes. Further details are provided in Appendix A. We leverage this asymmetry in Appendix F.

**Mutual information estimator.** We utilize the KNIFE estimator (Pichler et al., 2022) to estimate the predictive mutual information between continuous random variables. This estimator defines $\mathcal{F}$ as the class of Gaussian Mixtures with $K$ modes, introducing a soft-clustering approach for text generation evaluation (Pillutla et al., 2021; Pimentel et al., 2023). Through experimentation, we observed that varying the number of modes $K$ did not notably affect the results. Therefore, we present our findings with $K = 4$.

Initially, we examine the applicability of the basic Gaussian estimator of mutual information proposed in (Kim et al., 2022). However, we find it unsuitable for our scenario since the embeddings of both the source texts and summaries exhibit multimodal distributions, as illustrated in Figure 7. Instead, we find the KNIFE estimator (Pichler et al., 2022) to be better suited for our context as it is designed to estimate mixtures of Gaussians.

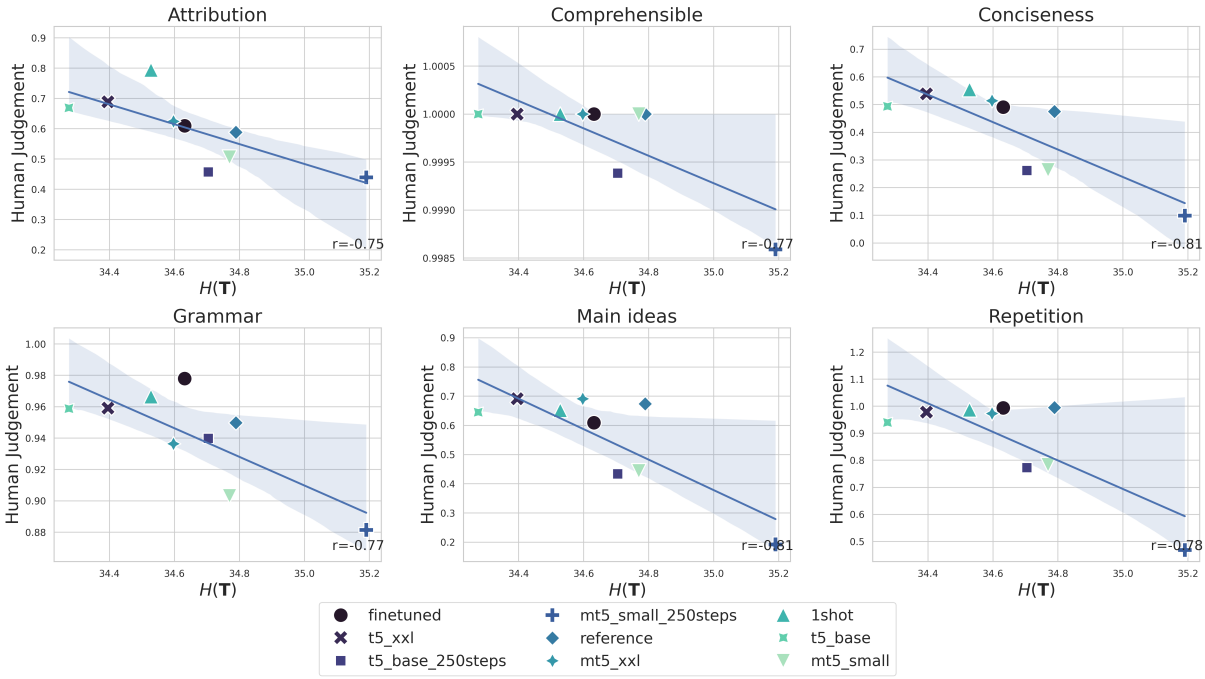


Figure 8: Correlation between the source texts entropies as estimated per KNIFE and the answers to the SEAHORSE benchmark. We observe that the entropy of the source texts correlates negatively with the answers.

## E Comparison Across Datasets

One might seek to compare models evaluated on different datasets. However, the mutual information, in its current form, is not suitable for such comparisons as it relies on the entropy of the dataset. In Figure 8, we demonstrate that for the SEAHORSE benchmark, there are significant variations in the entropies of the source texts, introducing bias into the comparison of mutual informations estimated as $I(\mathbf{T};\mathbf{S}) = H(\mathbf{T}) - H(\mathbf{T}|\mathbf{S})$. To address this issue, we propose normalizing the mutual information by the entropy of the dataset. We show that the normalized version of the mutual information correlates with the responses to the questions of the SEAHORSE benchmark (cf. Figure 9).

*Remark* 4. While the SEAHORSE benchmarks contains similar texts for all their models, the models have been each evaluated on different samples. For instant, in the english subset, only 91 samples out of the 10000 are common to all models. This leads to biaises in the evaluation of the mutual information.

We observed that variations in the source text datasets significantly affect the estimation results of mutual information. There's a tendency for smaller models to exhibit higher source text entropies, leading to misleading comparisons. To address this, we suggest computing the normalized MI between the source texts and summaries. This normalized MI is defined as follows:

$$\text{Normalized MI} \triangleq \frac{I(\mathbf{T};\mathbf{S})}{H(\mathbf{T})} = 1 - \frac{H(\mathbf{T}|\mathbf{S})}{H(\mathbf{T})}, \tag{30}$$

where $H(\mathbf{T}|\mathbf{S}) \leqslant H(\mathbf{T})$.

In Figure 9, we observed weak correlations between this normalized MI and the human judgments reported in the SEAHORSE benchmark. This discrepancy might arise from the evaluation of different datasets for each model, suggesting that the normalized MI might not be the most suitable normalization method. Hence, comparing models evaluated on different datasets should be avoided for now. However, when evaluated on the same datasets, MI correlates well with the metrics trained on the SEAHORSE benchmark. This indicates that MI is an promising tool for evaluating summarizers.

### E.1 SummEval dataset (Fabbri et al., 2021)

The SummEval dataset is well-suited for our task due to its inclusion of both summaries and corresponding source texts for identical documents. However, its limited size, comprising only 1700 samples, renders it

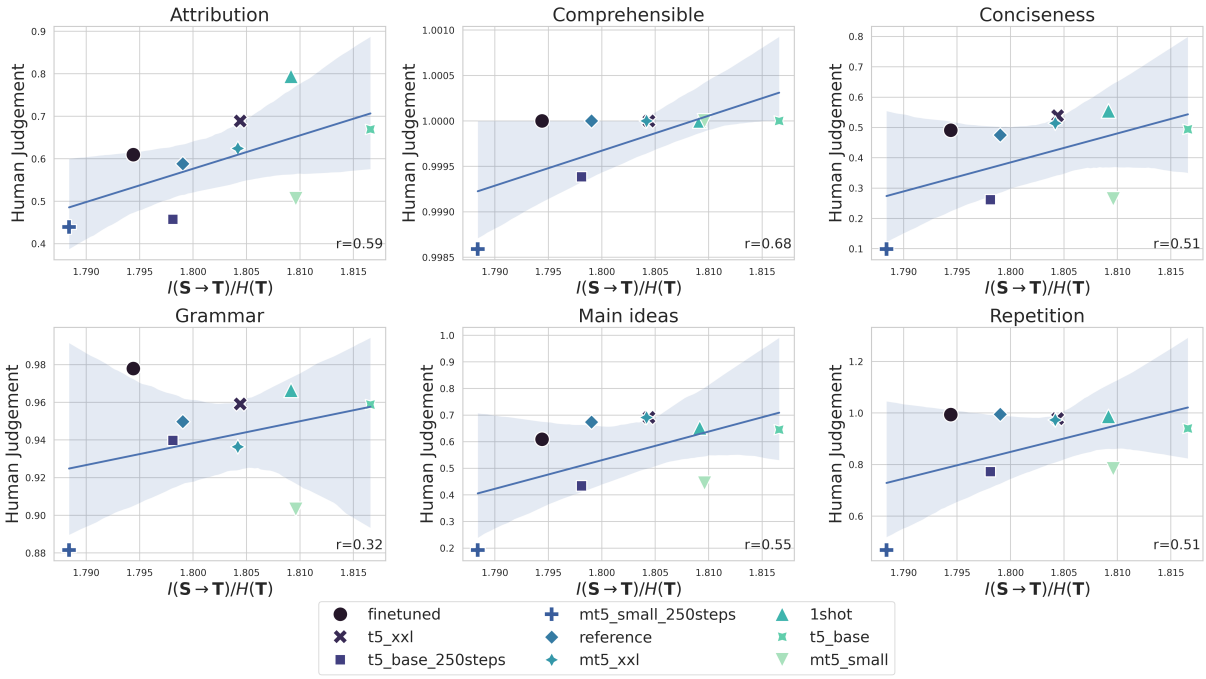


Figure 9: Correlation between the normalized mutual information and the answers to the SEAHORSE benchmark.

Table 4: Correlation between MI and ROUGE, and Seahorse metrics and probability of success of the classifcation task, grouped by datasets for non-trivial decoding strategies. SH. stands for Seahorse metrics and CT. for classification tasks.

| | Metric | $I(\mathbf{S};\mathbf{T})$ | Attribution | BERTScore | Main ideas | ROUGE-L |
|---|---|---|---|---|---|---|
| SH. | Attribution | 0.39 | 1.00 | 0.78 | 0.82 | 0.38 |
| | Comprehensible | 0.10 | 0.38 | 0.49 | 0.47 | 0.08 |
| | Conciseness | 0.50 | 0.81 | 0.83 | 0.90 | 0.46 |
| | Grammar | 0.14 | 0.28 | 0.39 | 0.43 | 0.04 |
| | Main ideas | 0.51 | 0.82 | 0.90 | 1.00 | 0.50 |
| | Repetition | -0.31 | -0.17 | -0.14 | -0.12 | -0.57 |
| CT. | Topic | 0.33 | 0.19 | 0.28 | 0.26 | 0.65 |
| | Emotions | 0.10 | 0.06 | 0.09 | 0.08 | 0.37 |
| | Sentiment Analysis | -0.04 | -0.09 | -0.10 | -0.15 | -0.00 |
| | GPT Detector | 0.51 | 0.32 | 0.49 | 0.42 | 0.72 |
| | Policy | 0.69 | 0.64 | 0.72 | 0.74 | 0.57 |
| Emb. | MPNET | 0.69 | 0.64 | 0.75 | 0.76 | 0.57 |
| | all-MiniLM | 0.65 | 0.65 | 0.74 | 0.75 | 0.53 |
| | Paraphrase | 0.22 | 0.28 | 0.22 | 0.21 | -0.07 |
| Common | ROUGE-L | 0.54 | 0.38 | 0.51 | 0.50 | 1.00 |
| | BERTScore | 0.50 | 0.78 | 1.00 | 0.90 | 0.51 |

insufficient for estimating mutual information.

### E.2 Additional languages

We performed additional experiments in French, German and Spanish using multlingual embedder to evaluate the mutual informatio. We obtained mixed-results. While the overall trends are similar, the lack of good multilingual embedders certainly hinders the results we can hope to obtain. It is a clear limit of our work since our method is highly dependent on the existence of a viable embedder for the text distribution at hand.

### E.3 Full results

## F Deciphering Summarizer Hierarchy

We proposed to evaluate the mutual information $I(\mathbf{T};\mathbf{S})$, where $\mathbf{S} \sim p_\theta(\mathbf{s}|\mathbf{t})$ being a summarizer – in our case, a finetuned language model– and $\mathbf{T}$ is the random variable of source texts. If we have two summarizers $p_\theta(\mathbf{s}|\mathbf{t})$ and $q_\phi(\mathbf{s}|\mathbf{t})$, we can evaluate the mutual information $I(\mathbf{S}_p \to \mathbf{S}_q)$, where $\mathbf{S}_p \sim p_\theta(\mathbf{s}|\mathbf{t})$ and $\mathbf{S}_q \sim q_\phi(\mathbf{s}|\mathbf{t})$. The mutual information here indicates how much information about $\mathbf{S}_q$ conveys about $\mathbf{S}_p$ and vice-versa. Interestingly, this observation enables us to build a hierarchy of

Table 5: Correlation between MI and ROUGE, and Seahorse metrics and probability of success of the classifcation task, grouped by datasets for non-trivial decoding strategies. SH. stands for Seahorse metrics and CT. for classification tasks.

| | Metric | $I(\mathbf{S};\mathbf{T})$ | Attribution | Main ideas | BARTScore | BERTScore | ROUGE-L |
|---|---|---|---|---|---|---|---|
| SH. | Attribution | 0.56 | 1.00 | 0.26 | 0.95 | 0.75 | 0.42 |
| | Comprehensible | 0.11 | 0.07 | 0.42 | 0.10 | 0.02 | -0.37 |
| | Conciseness | 0.80 | 0.67 | 0.81 | 0.66 | 0.67 | 0.24 |
| | Grammar | -0.24 | -0.35 | 0.16 | -0.34 | -0.30 | -0.50 |
| | Main ideas | 0.77 | 0.26 | 1.00 | 0.23 | 0.45 | 0.28 |
| | Repetition | 0.01 | -0.23 | 0.39 | -0.23 | -0.08 | -0.34 |
| CT. | Sentiment analysis | 0.65 | 0.68 | 0.34 | 0.68 | 0.70 | 0.54 |
| | GPT detector | 0.73 | 0.83 | 0.39 | 0.86 | 0.89 | 0.65 |
| | Policy classification | 0.83 | 0.40 | 0.69 | 0.46 | 0.77 | 0.71 |
| | Emotion classification | 0.72 | 0.68 | 0.40 | 0.72 | 0.75 | 0.58 |
| Emb. | Emb. Paraphrase | 0.79 | 0.76 | 0.60 | 0.74 | 0.69 | 0.29 |
| Common | ROUGE-L | 0.55 | 0.42 | 0.28 | 0.45 | 0.70 | 1.00 |
| | BERTScore | 0.79 | 0.75 | 0.45 | 0.82 | 1.00 | 0.70 |
| | BARTScore | 0.57 | 0.95 | 0.23 | 1.00 | 0.82 | 0.45 |

summarizers. Some summarizers produce very informative summaries that can be used to predict the ones from other models while being so informative that other summaries cannot provide enough information to build them. We build the directed graph of the predictive power of the summaries of each model on the summaries of other models. A model's average outgoing mutual information is the average mutual information between this model's summaries and other models' summaries. A model's average incoming mutual information is the average mutual information between its summaries and those of other models.

**OOD models.** Underperforming models, which were trained on disparate data distributions such as Arxiv or medical summarization, generally display low mutual information with other models and prove challenging to predict from conventional specialized systems (see FFigure 10). This outcome is unsurprising, given that these models exhibit significantly divergent behavior compared to others. Consequently, their outputs offer minimal insight into the outputs of other models.

**Strong models.** Robust models like distilBart and Bart demonstrate high informativeness regarding other models, while also posing challenges for prediction (refer to Figure 10 and Figure 3). This outcome is anticipated, given that robust models can encapsulate significantly more information within their summaries compared to other models. As a result, their summaries prove valuable for predicting the outputs of weaker summarizers. However, these summaries are also considerably challenging to predict from the perspectives of those weaker models.

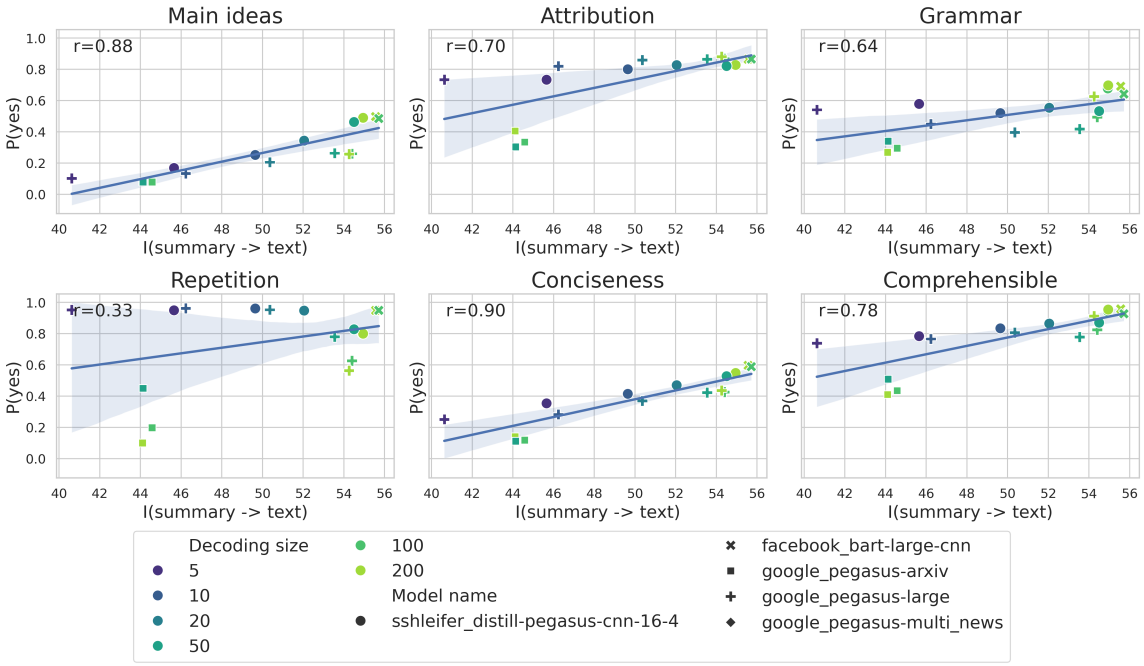


Figure 11: Correlation with the SEAHORSE metrics on the CNN DailyMail dataset is measured by $\Pr(\text{Yes})$, which represents the average probability across the dataset that the SEAHORSE model predicts the answer "Yes" to the corresponding question.

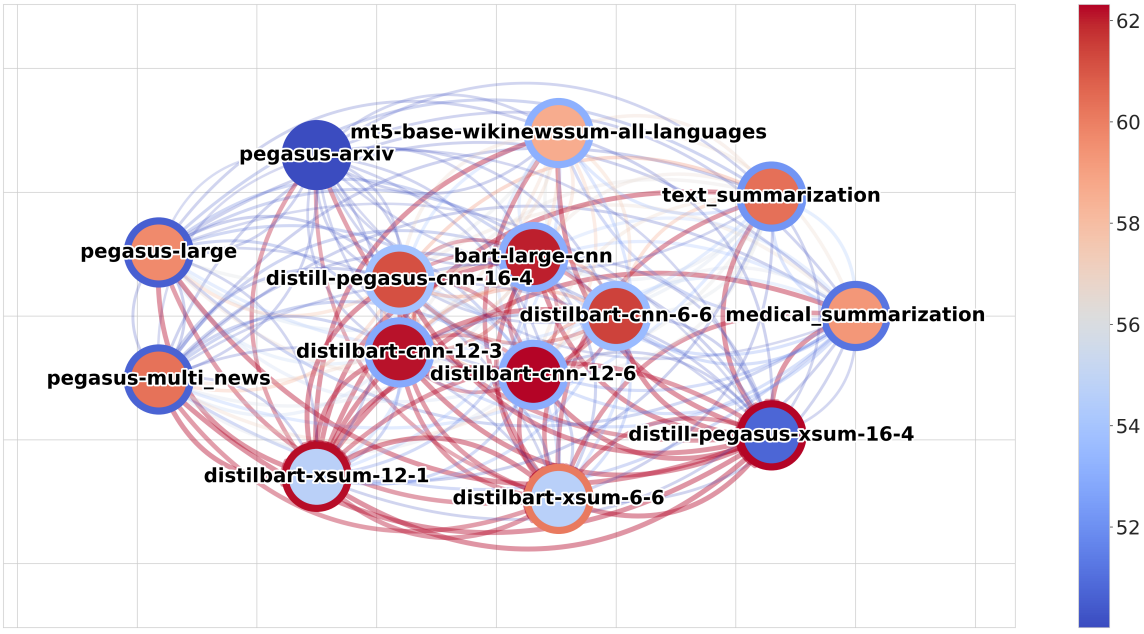


Figure 10: The predictive power of each model's summaries on the summaries of other models is depicted in the visualization. The central color denotes the average predictive power of that summarizer regarding the others, while the border color indicates the average predictive power of the other summarizers concerning that summarizer. A red center and blue border signify high informativeness, indicating a summarizer that is highly informative and difficult to predict. Conversely, a node with a blue center and red border implies low informativeness about the other summarizers but easy predictability by them.

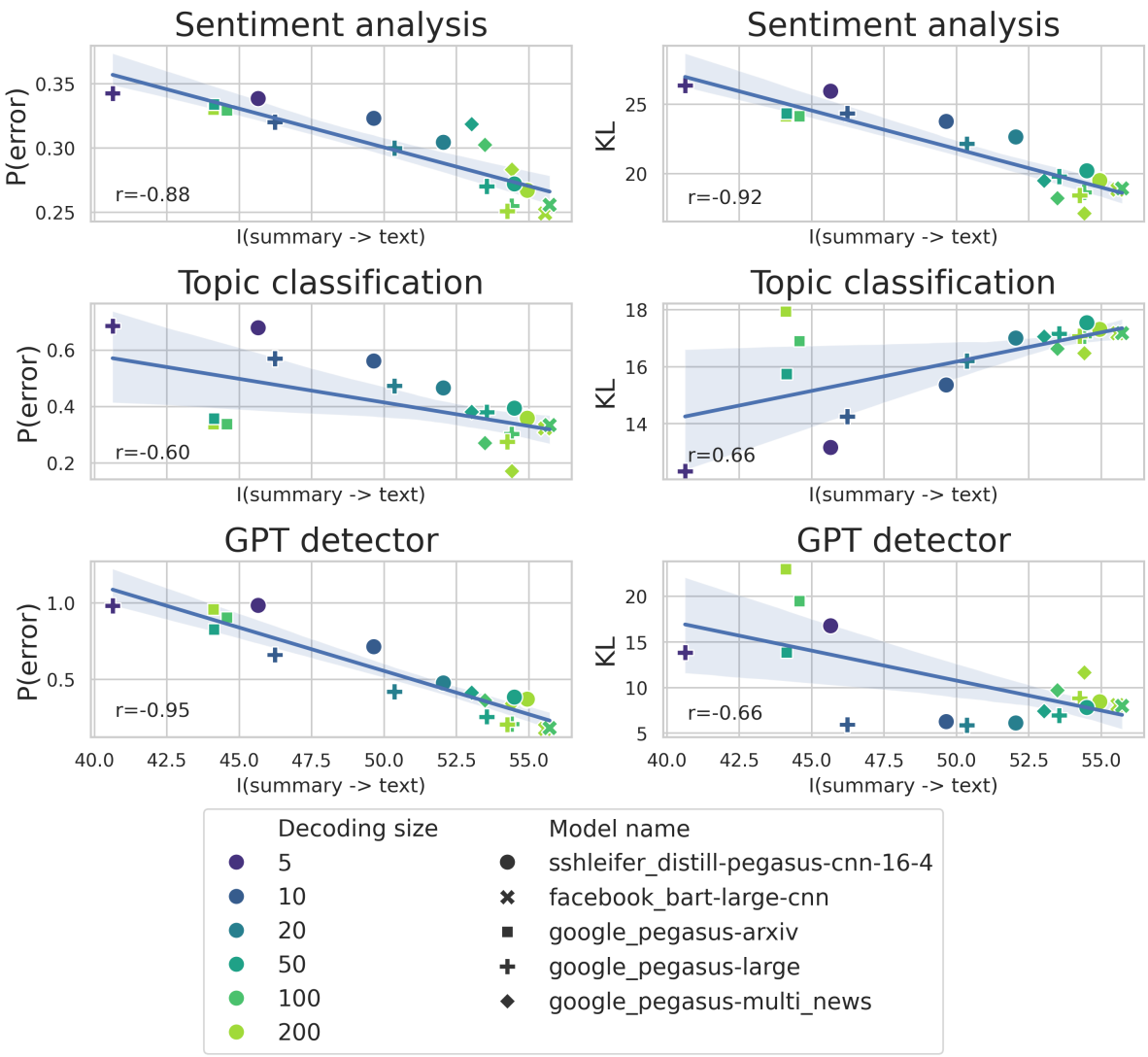


Figure 12: As one would expect, the performance of the classification tasks on the summaries increases with the decoding size as it allows the model to pack more information into the summary.

## G Ablations

### G.1 Decoding size

**Impact of the length of the summary.** The longer a summary, the more likely a downstream classifier is to produce the same output on the source text and on the summary. However, this trend is not always verified for weakers or OOD models. In Figure 12 and Figure 11, we can observe that the Pegasus model finetuned on arxiv papers tends to be less informative even when generating extended summaries when applied to the CNN-Dailymail dataset. This shows that the mutual information captures more than just the length of the summary but also its actual informativity.

## H Negative Results

### H.1 Pointwise mutual information

While the mutual information gives good insights about a summarizer, the point-wise mutual information, computed for each pair of source texts and summaries did not result in interesting correlations with the downstream tasks. Previous work have shown that it was a sound metric when the generative model is used to compute the mutual information (Bugliarello et al., 2020; Ethayarajh et al., 2022), however in our scenario we fit an ad-hoc mutual information estimator.