Document Summarization with Conformal Importance Guarantees

Bruce Kuwahara[*]
Signal 1 AI
Toronto, Canada

Chen-Yuan Lin[*]
Signal 1 AI
Toronto, Canada

Xiao Shi Huang[*]
Signal 1 AI
Toronto, Canada

Kin Kwan Leung
Layer 6 AI
Toronto, Canada

Jullian Arta Yapeter
Signal 1 AI
Toronto, Canada

Ilya Stanevich
Signal 1 AI
Toronto, Canada

Felipe Perez
Signal 1 AI
Toronto, Canada

Jesse C. Cresswell
Layer 6 AI
Toronto, Canada

Abstract

Automatic summarization systems have advanced rapidly with large language models (LLMs), yet they still lack reliable guarantees on inclusion of critical content in high-stakes domains like healthcare, law, and finance. In this work, we introduce Conformal Importance Summarization, the first framework for importance-preserving summary generation which uses conformal prediction to provide rigorous, distribution-free coverage guarantees. By calibrating thresholds on sentence-level importance scores, we enable extractive document summarization with user-specified coverage and recall rates over critical content. Our method is model-agnostic, requires only a small calibration set, and seamlessly integrates with existing black-box LLMs. Experiments on established summarization benchmarks demonstrate that Conformal Importance Summarization achieves the theoretically assured information coverage rate. Our work suggests that Conformal Importance Summarization can be combined with existing techniques to achieve reliable, controllable automatic summarization, paving the way for safer deployment of AI summarization tools in critical applications. Code is available at github.com/layer6ai-labs/conformal-importance-summarization.

1 Introduction

Summarization is a widely performed task in many domains, from media [68] and legal documents [51, 34] to scientific articles [11] and clinical reports [19]. Recent advances in large language models (LLMs) have significantly improved the quality of summary generation [15, 1, 29, 64], with methods such as prompt-based generation [12, 66] and fine-tuned transformer models [61] exhibiting superior generalization and adaptability over classical natural language processing (NLP) methods [42, 67]. However, in critical domains any error in an AI-generated summary can have serious consequences [5]. For example, even with evidence that AI summarizers can reduce physician workloads and alleviate burnout [25], lack of consistency and the need to verify the AI's work remains a concern for physicians in practice [60]. Despite the improvements mentioned above, no existing method *guarantees* retention of important content which could, for example, ensure safety in a high-stakes application like healthcare [18].

Conformal prediction [62, 59] has recently risen in popularity as it provides distribution-free, finite-sample coverage guarantees [2], and has shown promise in classification [56, 3, 33], regression [10, 43, 55], and language tasks such as factual question answering [37, 52, 45, 14, 23]. In this work, we introduce **Conformal Importance Summarization**, the first application of conformal

[*]Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

prediction to document summarization which provides statistical guarantees on the inclusion of important content.

Our contributions are as follows:

- We formalize the problem of importance-preserving document summarization with statistical guarantees through the conformal prediction framework.
- We introduce a method that calibrates sentence-level importance scores, allowing summary generation with user-specified error (α) and recall (β) rates.
- We evaluate our method across multiple summarization benchmarks, demonstrating empirical importance coverage as expected from our theory, and quantifying the utility of our approach.

2 Background and Terminology

2.1 Document Summarization

Summarization is the task of producing a concise version of a source document that preserves its most important content. There are two major categories of approaches [35]: **extractive summarization**, which selects spans of text (typically sentences) taken verbatim from the source text, and **abstractive summarization** which generates new sentences that paraphrase or synthesize information from the source text. While abstractive summarization is extremely accessible due to the advent of instruction-tuned LLMs [63, 58], extractive summarization can be more suitable to high-stakes domains because it limits the possibility of hallucinations and remains more faithful to the source's meaning. While our main focus is on extractive summarization, we also show how the two approaches can be combined to benefit from the improved fluency and conciseness that abstractive summarization offers.

Classical extractive methods such as TextRank [44] rely on heuristics or graph-based techniques. Modern extractive models, such as BERTSum [41], leverage pretrained language models like BERT [21] to encode sentence-level representations and classify sentence importance. Recent trends in summarization also include reinforcement learning for optimizing summary-level objectives directly (e.g., ROUGE [50, 39]), fact-consistency tuning using entailment models [36], and hybrid extractive-abstractive pipelines [13].

In this work, we show that extractive summarization, which scores and ranks text spans, is naturally compatible with the calibration step of conformal prediction and can achieve statistical guarantees on the retention of important sentences.

2.2 Conformal Prediction

Consider a classification or regression problem with inputs $x \in \mathcal{X}$ associated with ground-truth values $y^* \in \mathcal{Y}$ drawn jointly from a distribution $(x, y^*) \sim \mathbb{P}$. Conformal prediction (CP) [62, 59] first calibrates a threshold $\hat{q}$ based on labeled data, then predicts a set of output values $C_{\hat{q}}(x_{\text{test}}) \subseteq \mathcal{Y}$ for any new datapoint x_{test}, while guaranteeing *coverage* with a user-defined error rate α,

$$\mathbb{P}[y^*_{\text{test}} \in C_{\hat{q}}(x_{\text{test}})] \geq 1 - \alpha. \tag{1}$$

Remarkably, the coverage guarantee is distribution-free and valid in finite samples, making CP a versatile tool to provide robust guarantees on correctness in a wide variety of scenarios [2].

Given a model $f_\theta : \mathcal{X} \to \mathcal{Y}$, CP defines a *conformal score* function $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where larger values indicate worse agreement between $f_\theta(x)$ and y^*. Upon computing S over n calibration datapoints, the threshold $\hat{q}$ is set as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the conformal scores. For any new input, prediction sets are generated by including all output values for which the conformal score is below the threshold $\hat{q}$,

$$C_{\hat{q}}(x_{\text{test}}) = \{y \in \mathcal{Y} \mid s(x_{\text{test}}, y) < \hat{q}\}. \tag{2}$$

As long as x_{test} is exchangeable with the calibration data drawn from $\mathbb{P}$, Eq. (1) will hold. For equal coverage levels $1 - \alpha$, the usefulness of prediction sets can be judged by their size [56, 3, 33], with smaller average set sizes $\mathbb{E}|C_{\hat{q}}|$ indicating prediction sets that are more useful in downstream tasks [16, 17]. The quality of $C_{\hat{q}}$ is largely driven by the accuracy and calibration of f_θ, and the design of S for expressing the model's confidence.

2.3 Conformal Factuality for Question-Answering

Extending CP to language tasks requires rethinking the problem setup and design of the conformal score S. Whereas classification tasks have a finite label set $\mathcal{Y}$, open-ended question-answering has an effectively infinite output space with many semantically equivalent responses, making it incompatible with the standard CP framework described in Section 2.2.

Conformal factuality [45] overcomes these challenges by replacing prediction sets with entailment sets, aiming to give responses that are entailed by the ground-truth y^* with high probability [8]. Given a natural language question x and response generated by an LLM, conformal factuality first decomposes the generated text into subclaims $\hat{y} = \{c_1, \dots, c_p\}$. We use $T(x, y^*) \subseteq \hat{y}$ to denote the subset of generated claims which are entailed by the ground-truth y^* for question x. Then, individual claims are filtered out based on a calibrated threshold $\hat{q}$. The remaining claims constitute a response $y \subseteq \hat{y}$ which is factual with high probability,

$$\mathbb{P}[y \subseteq T(x, y^*)] \geq 1 - \alpha. \tag{3}$$

To set $\hat{q}$, each claim's factuality is first evaluated based on heuristics such as model confidence or self-evaluation. Then, the conformal score is assigned as the greatest factuality level out of all claims not entailed by y^*. After computing the conformal score for each (x, y^*) in the calibration set, the overall conformal threshold $\hat{q}$ is set as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of scores.

On test data, claims with assessed factuality level less than $\hat{q}$ are filtered out. The sets of retained claims satisfy Eq. (3) by Theorem 3.1 of [45]. Since longer, more detailed answers are more helpful, the quality of the final response is judged by how many of the generated claims can be retained, subject to meeting the coverage guarantee in Eq. (3). Quality is dictated by how well the factuality of individual claims is assessed, and the prevalence of non-factual claims in the generated response, which can be improved with a stronger LLM, adjustments to the score function, better grounding through retrieved context, or reasoning [14, 23, 57].

3 Extractive Summarization via Conformal Importance

We take inspiration from conformal factuality to go beyond question-answering and provide statistical guarantees on extractive summarization. By calibrating a threshold on an importance score, we ensure that the final summary preserves the most salient information while maintaining a bounded risk of omitting critical content.

Formally, our goal is to produce a shorter version y of the long-text x which, with high probability, retains all important information. We take the long-text $x \in \mathcal{X}$ to consist of multiple sentences, $x = \{c_1, \dots, c_p\}$, and let $y^* \subseteq x$ be the ground-truth set of important sentences from x.[2] We then filter out sentences from x based on a calibrated threshold $\hat{q}$ leaving a subset $y \subseteq x$ which retains all important information with high probability,

$$\mathbb{P}[y^* \subseteq y] \geq 1 - \alpha. \tag{4}$$

We note a key difference between Eq. (3) and Eq. (4): conformal factuality aims for high precision that retained claims are factual, while conformal importance aims for high recall that important sentences are retained. As long as high recall is ensured, shorter summaries are preferable which allows us to measure the quality of y as the proportion of sentences removed. In the remainder of this section we develop a more general framework for error control than used in conformal factuality, describe our method to extract summaries, and theoretically prove that it obeys a coverage guarantee.

Generalizing the Coverage Guarantee. One limitation of the conformal factuality framework is that Eq. (3) rigidly considers a response with any number of non-factual claims to be a failure case. A more general framework would allow the user to set their own tolerance on how many non-factual claims could appear in an acceptable response [4]. Hence, for conformal importance we relax the coverage guarantee Eq. (4) so that summaries are acceptable if they retain at least a fraction β of the important sentences. Let

$$B(y; y^*) = \frac{|y \cap y^*|}{|y^*|} \tag{5}$$

[2]For simplicity we break x by sentences, but any span could be used. We discuss cases where $y^* \not\subseteq x$ below.

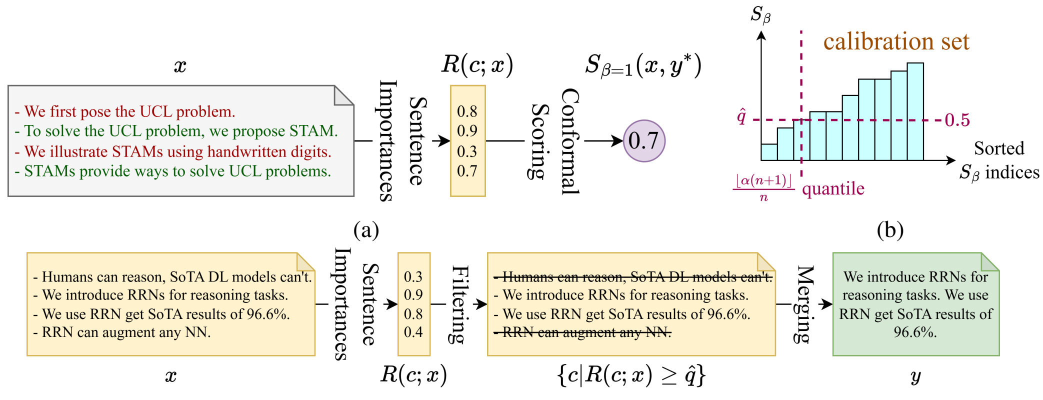


Figure 1: (a) Steps to compute the conformal score for a labeled datapoint. Ground-truth important sentences are coloured in green. $S_{\beta=1}$ is the smallest value of $R(c;x)$ for any important sentence. (b) The conformal threshold $\hat{q}$ is computed as a quantile of the sorted conformal scores over the calibration set. (c) At inference, only sentences with importance score $R(c;x) \geq \hat{q}$ are retained in the summary.

be the *recall*, i.e. the fraction of important sentences retained by y. Then the relaxed coverage guarantee is

$$\mathbb{P}[B(y;y^*) \geq \beta] \geq 1-\alpha. \tag{6}$$

Of course, this recovers Eq. (4) when $\beta = 1$.

Conformal Importance Summarization. Given the long-text x we assign an importance score to each sentence c as $R(c;x)$ such that $R(c;x) \geq 0$ with larger scores indicating higher estimated importance. In practice, the design of the importance score is key to the performance of our method, and we discuss various options below. Similar to [45], we define a filtering function based on importance scores

$$F_q(x) = \{c \in x \mid R(c;x) \geq q\}. \tag{7}$$

This function satisfies both $F_0(x) = x$ and $F_\infty(x) = \varnothing$ so that F_q filters out more sentences as q increases, and satisfies a nesting property: $F_q(x) \subseteq F_{q'}(x)$ for $q \geq q'$ [31].

To determine the appropriate threshold $\hat{q}$ we use CP calibration with conformal scores over a labeled calibration dataset. For each pair of long-text x_i and ground-truth summary y_i^* we compute

$$S_\beta(x_i, y_i^*) := \max\{q \in \mathbb{R}^+ \mid B(F_q(x_i); y_i^*) \geq \beta\}. \tag{8}$$

We note that the maximum always exists by the definition of $F_q(x)$. This score computes the largest threshold q such that at least a fraction β of the important sentences are retained after filtration. A larger q (and hence larger S_β) is preferable as it will produce a more concise summary $y_i = F_q(x_i)$.

Noting that $B(F_q(x);y^*)$ is a non-increasing function of q, we have the following observation.

Lemma 1. *For a fixed $\hat{q}$, we have*

$$S_\beta(x, y^*) \geq \hat{q} \iff B(F_{\hat{q}}(x); y^*) \geq \beta. \tag{9}$$

Proof. First, assume that $S_\beta(x,y^*) \geq \hat{q}$. By the definition in Eq. (8), there exists $q' \geq \hat{q}$ such that $B(F_{q'}(x);y^*) \geq \beta$. Since $B(F_q(x);y^*)$ is non-increasing in q, and $q' \geq \hat{q}$, it follows that $B(F_{\hat{q}}(x);y^*) \geq \beta$. On the other hand, now assume $B(F_{\hat{q}}(x);y^*) \geq \beta$. By the definition of maximum, we have $S_\beta(x,y^*) \geq \hat{q}$ directly from Eq. (8). □

With all conformal scores $S_\beta(x_i, y_i^*)$ computed on the calibration set, we choose the overall conformal threshold $\hat{q}$ as the $\frac{\lfloor \alpha(n+1) \rfloor}{n}$ quantile.[3] Given a new long-text x_{test} we score each sentence's importance as $R(c; x_{\text{test}})$, and filter out any sentence with importance less than $\hat{q}$, returning $y_{\text{test}} = F_{\hat{q}}(x_{\text{test}})$, as shown in Figure 1. This procedure satisfies our generalized coverage guarantee in Eq. (6).

[3]The different quantile compared to conformal factuality is necessary, as we have shifted to focus on recall instead of precision in the coverage guarantee and conformal score function.

Theorem 1. *Let $\{x_i, y_i^*\}_{i=1}^{n+1}$ be exchangeable and $0 < \beta \leq 1$. Let $\hat{q}$ be the $\frac{\lfloor \alpha(n+1) \rfloor}{n}$-th quantile of the scores $\{S_\beta(x_i, y_i^*)\}_{i=1}^{n}$, which we assume to be distinct without loss of generality. Then for $\alpha \in [\frac{1}{n+1}, 1]$, we have*

$$1 - \alpha \leq \mathbb{P}[B(F_{\hat{q}}(x_{n+1}); y_{n+1}^*) \geq \beta] < 1 - \alpha + \frac{1}{n+1}. \tag{10}$$

Proof. Let $s_i = S_\beta(x_i, y_i^*)$ for $i = 1, \ldots, n$, and $s_{\text{test}} = S_\beta(x_{n+1}, y_{n+1}^*)$. Without loss of generality, we assume the scores are sorted $s_1 < s_2 \ldots < s_n$. By exchangeability of the x_i's, and for any $k = 1, \ldots, n$, we have

$$\mathbb{P}[s_{\text{test}} \geq s_k] = 1 - \frac{k}{n+1}. \tag{11}$$

Noting that $\hat{q} = s_{\lfloor \alpha(n+1) \rfloor}$ by definition of the quantile, and that $\alpha \geq \frac{1}{n+1}$, we obtain

$$\mathbb{P}[s_{\text{test}} \geq \hat{q}] = 1 - \frac{\lfloor \alpha(n+1) \rfloor}{n+1} \geq 1 - \alpha. \tag{12}$$

By Lemma 1 we then have

$$\mathbb{P}[B(F_{\hat{q}}(x_{n+1}); y_{n+1}^*) \geq \beta] \geq 1 - \alpha. \tag{13}$$

On the other hand, from Eq. (12) we see that

$$\mathbb{P}[s_{\text{test}} \geq \hat{q}] = 1 - \frac{\lfloor \alpha(n+1) \rfloor}{n+1} < 1 - \frac{\alpha(n+1) - 1}{n+1} = 1 - \alpha + \frac{1}{n+1}, \tag{14}$$

which shows that

$$1 - \alpha \leq \mathbb{P}[B(F_{\hat{q}}(x_{n+1}); y_{n+1}^*) \geq \beta] < 1 - \alpha + \frac{1}{n+1}. \tag{15}$$

□

Design Choices. Beyond the free parameters α and β which can be set according to the user's error tolerances, our method involves design choices around the ground-truth y^* and the importance score function $R(c; x)$. We defined important sentences as those sentences appearing in y^*, but the source of ground truth remains flexible. On benchmark datasets for extractive summarization, y^* may be provided at the sentence level as a subset of x. For other datasets, a summary may be given, but not as verbatim sentences selected from x, in which case we assume sentences in y^* can be matched to a unique source from x. When y^* is unavailable, and cannot be collected from domain experts, automated generation techniques using prompted LLMs can select important sentences. However, we note the strong possibility of bias if the same LLM is used for $R(c; x)$ [49, 20]. We describe concrete methods for sentence matching and ground truth generation in Section 4.1.

Even for the same long-texts x, different users may consider different parts important. For example, when summarizing clinical notes produced over the course of a patient's hospital stay, a doctor and an administrator may have different views on which details are important. Conformal Importance Summarization can accommodate these different viewpoints; for the same data x, each user can indicate their preferences y^* on the calibration set, and receive a threshold $\hat{q}$ tailored to their needs. It may also be beneficial to guide the meaning of importance via $R(c; x)$, for instance by describing within an LLM prompt the type of information which should be scored highly. Indeed, designing $R(c; x)$ is the most direct way to influence the ultimate performance of our method, so we offer and experimentally compare several possibilities that can be grouped into two classes.

LLM Scoring - An LLM is prompted to return a score between 0 and 1 of how important c is, given x as context. We test five different LLMs, **GPT-4o mini** [47], **Gemini 2.0 Flash-Lite** [6], **Gemini 2.5 Flash** [28], **Llama3-8B** [29], and **Qwen3-8B** [65], all with the same prompt given in Appendix B. The first three used public APIs, while the latter two were hosted locally on a 48 GB A6000 GPU.

Embedding Similarity Scoring - Sentence-level embeddings are commonly used in extractive summarization, and we demonstrate how they can be leveraged to give coverage guarantees. Using a sentence-level embedding model (e.g. SBERT [54]), distances between all embeddings are computed to form a graph. Then, one of four aggregation methods is used to produce importance scores: **Cosine Similarity Centrality** [53] builds a fully connected similarity graph and assigns importance as the weighted centrality of a node; **Sentence Centrality** [27] creates a directed graph where each sentence's importance is computed based on similarity to later sentences only; **GUSUM** [26] creates

Table 1: Dataset Details

Dataset	Subset Filtering	Labeling Method	$\lvert\mathcal{D}_{\text{cal}}\rvert$	$\lvert\mathcal{D}_{\text{test}}\rvert$	Average Length p	Label Positive Rate
ECT	All	Provided	100	2322	45.6	0.10
CSDS	Valid + Test	Provided	100	1500	25.5	0.27
CNN/DM	1000 samples	SBERT Opt.	100	900	31.0	0.10
TLDR-AIC	>1 summary	SBERT Opt.	100	1043	40.7	0.06
TLDR-Full	>1 summary	SBERT Opt.	100	1043	216	0.0014
MTS	>1 input sentence	GPT-4o mini	100	1029	6.4	0.81

an undirected graph where edges are cosine similarities, and the importance score is augmented with sentence-level features like position and length; **LexRank** [22] uses a similar undirected graph construction, but filters out weak connections, and then applies a PageRank-like algorithm [48] over node centrality to rank the importance of sentences.

Abstractive Post-processing. Extractive summarization limits the possibility of hallucination, but may produce text which does not flow smoothly. On the other hand, using LLMs for direct abstractive summarization does not give precise control over coverage and recall. Abstractive summarization essentially contains two subtasks: identifying important information, and synthesizing it into a concise and grammatically correct summary [13, 24]. Direct prompting forces the LLM to perform both tasks at once. We argue that disentangling these subtasks can lead to better information retention, similar to how retrieval-augmented generation [38] splits question-answering into the two subtasks of information retrieval and answer generation.

We propose a hybrid extractive-abstractive pipeline which first identifies important information, then synthesizes it. First, Conformal Importance Summarization extracts important information from x as $y = F_{\hat{q}}(x)$ with guarantees on coverage and recall. Then, extracted sentences y are passed to an LLM which is prompted to preserve *all* information, and only improve conciseness and flow. While there is no guarantee the LLM will maintain coverage, below we test how successful it can be in practice, and compare it to pure abstractive summarization. The prompts used are given in Appendix B.

4 Experimental Setup

Our experiments are designed to validate the conformal guarantee given by Theorem 1, and compare Conformal Importance Summarization to existing summarization methods that do not provide guarantees. We run ablations to understand (*i*) the influence of both α and β; (*ii*) the design of the importance score function $R(c; x)$; and, (*iii*) for datasets without explicit ground-truth labels, the label creation method. Finally, we compare pure abstractive summarization with an LLM to our hybrid extractive-abstractive approach.

4.1 Datasets

We use 5 datasets to evaluate the performance of our framework: **ECTSum** [46] contains complete transcripts from corporate earnings calls, as well as expert-curated extractive summaries at the sentence level; **CSDS** [40] is a dataset of Chinese language customer-client conversations. Although the summaries are abstractive, each conversation has sentence-level labels for use as an extractive benchmark; **CNN/DM** [32] covers news sourced from CNN and The Daily Mail with human-written summary sentences; SciTLDR [11] consists of summaries of scientific papers sourced from both authors and peer-reviewers, and we use two versions where the input is either the full text (**TLDR-Full**), or just the abstract, introduction, and conclusion (**TLDR-AIC**); **MTS-Dialog** [7] is a collection of doctor-patient conversations and corresponding summaries intended to cover dialogue material.

For each dataset, a random subset ($n = 100$) of all datapoints is sampled to form the calibration dataset. All remaining samples form the test dataset, except for CSDS where we use only the original validation and test sets, and CNN/DailyMail where we use 900 samples to reduce resource requirements, as shown in Table 1. Other details on dataset preparation are given in Appendix A.

While ECTSum and CSDS contain sentence-level labels, the other datasets require them to be generated by comparing the summary text to the original sentences. For CNN/DM and SciTLDR, a greedy labeling process based on SBERT [54] embedding cosine similarity is used, which we describe fully in Appendix A with Algorithm 1. Due to the heavy context-dependency of MTS, we instead queried GPT-4o mini to generate the labels, which tended to classify sentences as important at

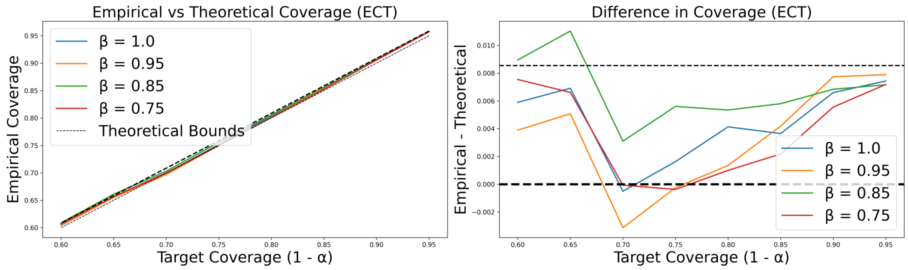


Figure 2: User-specified target coverage 1-α versus average empirical coverage, on ECTSum using Gemini 2.5 Flash scores. Dashed lines show theoretical bounds given in Theorem 1. Results are averaged over 400 random splits of calibration and test data.

a higher rate. We share the prompt used in Appendix B. While neither GPT-based nor SBERT-based labels would completely match a domain expert's preferences, it is sufficient to evaluate the theoretical guarantees and performance of Conformal Importance Summarization.

4.2 Metrics

To evaluate the quality of Conformal Importance Summarization, we use several complementary metrics. First, independent of the conformal framework, we assess the ranking quality of importance scores $R(c; x)$ using the area under the precision-recall curve (AUPRC), which evaluates how well each method distinguishes between important and unimportant sentences. AUPRC captures the trade-off between conciseness (precision) and completeness (recall).

Second, we evaluate the conformal framework by fixing values for the target error rate α and recall β, and measuring the proportion of sentences removed, akin to set size in traditional conformal prediction, which reflects how effectively content is compressed while preserving key information.

Finally, the empirical values of coverage and recall actually acheived are relevant. The recall of a given summary is $B(y; y^*)$ defined in Eq. (5), while the empirical coverage is binary, computed as $B(y; y^*) \geq \beta$. Both these measures are then averaged over the test set.

4.3 Baselines

No existing methods provide coverage guarantees for extractive summarization. In lieu of such existing baselines, we implement the several importance scoring methods described in Section 3, and also compare to the empirical coverage attained by LLMs without our conformal framework. In particular, we use GPT-4o mini to directly label each sentence as important or not, given the long-text as context but without access to the ground-truth summary. The prompt we use is provided in Appendix B. Whereas Conformal Importance Summarization can provide summaries for any choice of α, this baseline only provides a single value empirically, with no user control.

For the targeted evaluations of importance functions via AUPRC, we also include the ground-truth importance rate as a reference point, as it represents the performance of a random scoring function.

5 Results

5.1 Testing the Coverage Guarantee

First, we empirically verify the theoretical guarantees proved in Section 3. We run Conformal Importance Summarization on the calibration set for a specified α and β, and measure the empirical coverage on the test set. The coverage rate should lie between $1 - \alpha$ and $1 - \alpha + \frac{1}{n+1}$ in expectation, per Theorem 1. Since the coverage is a random variable depending on the calibration data, we repeat the process a total of 400 times with randomized data splits, and average the coverage. Figure 2 shows results for the ECTSum dataset with Gemini 2.5 Flash scores under several values of α and β. We see that the average coverage consistently lies between the theoretical bounds, as expected. Deviations from the bounds are minor, and result from the variance of the coverage random variable. The same plots for other datasets and scoring functions are shown in Appendix C.1.

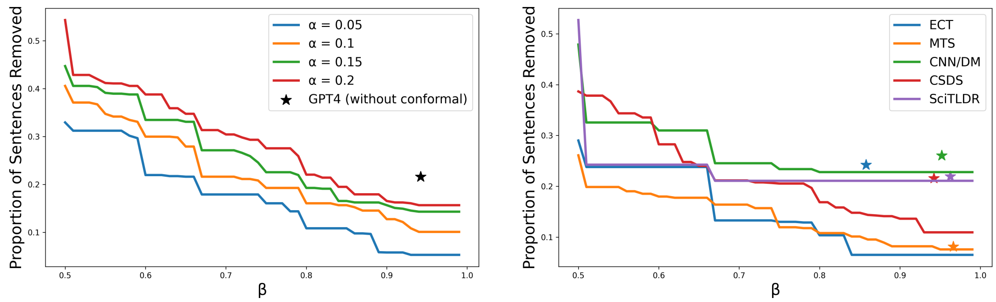


Figure 4: Target recall β vs. proportion of sentences removed (conciseness). **Left:** Lines indicate different values for the target error rate α on CSDS. **Right:** Lines indicate different datasets ($\alpha = 0.1$). Stars indicate the empirical recall and conciseness achieved by GPT-4o mini without conformal prediction.

5.2 Effect of α and β

Next, we demonstrate that by tuning α and β, one can control the tradeoff between conciseness and completeness of the resultant summary. Figure 3 using GPT-4o mini on CSDS shows that a higher allowable error rate α results in lower empirical recall, as does a lower target recall β. Note that β is the *minimum* target recall for a $1 - \alpha$ portion of the generated summaries, therefore the empirical recall is usually larger than β in practice for small α.

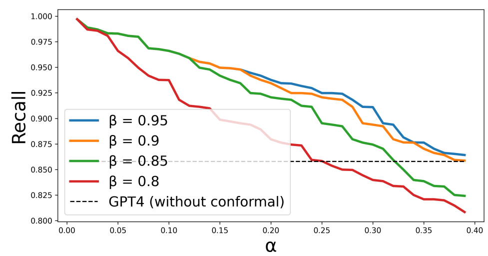


Figure 3: Target error rate α versus empirical recall $B(y; y^*)$ of important sentences in summaries, averaged over the CSDS test set. The dashed line shows GPT-4o mini performance without using conformal prediction.

Figure 3 (dashed line) also shows the baseline empirical recall when using GPT-4o mini as a standalone extractive summarization tool, as described in Section 4.3. This demonstrates how our conformal method allows more fine-grained control over recall and coverage than relying on LLMs directly. The same plots for other datasets and scoring functions are shown in Appendix C.2.

High recall alone is not the objective, since summaries also must be concise. Figure 4 shows the proportion of sentences removed as the target recall β is varied. Higher β values are more conservative and retain more sentences, while lower values lead to shorter summaries but may miss more important information. We see that for a given value of β, higher values of α also result in shorter summaries, since greater risk tolerance enables more sentences to be removed. These trends are consistent across datasets and scoring functions, as shown in Appendix C.3.

Again, we contrast this with the non-conformal baseline shown as stars in Figure 4. Direct extraction with GPT-4o mini produces more concise summaries than our method for a given recall level, but offers no control over what that recall level is. In cases where concise summaries are the top priority, our method allows the user to tune β down (or α up) to produce shorter summaries than the baseline.

5.3 Importance Score Function Performance

The efficacy of Conformal Importance Summarization is highly dependent on the design of the importance score $R(c; x)$. Here we compare the performance of the alternatives described in Section 3 across datasets. Table 2 (left columns) shows the sample-averaged AUPRC between the scores and ground-truth importance labels, as well as the dataset positive rate. Gemini 2 and 2.5's scores perform the best across all datasets, but GPT-4o mini also demonstrates superior performance compared to classical NLP methods. Smaller, open-source models still perform admirably compared to NLP approaches. Constraining GPT-4o mini to output a binary score rather than a float negatively affects its performance.

Table 2 (right columns) shows the performance of different score functions as measured by the proportion of sentences removed with $\alpha = 0.2$ and $\beta = 0.8$. Once again, the Gemini models and GPT-4o mini generally perform the best by having the greatest reduction in length for this level of coverage. One example summary with filtered and retained sentences is shown in Figure 5.

Table 2: Performance comparison of importance scoring methods. **Left:** AUPRC of importance rankings compared to ground truth. AUPRC of the original article indicates the proportion of all sentences labeled as important. **Right:** Conciseness of summaries (proportion of sentences removed) under Conformal Importance Summarization with $\alpha = 0.2$ and $\beta = 0.8$. Higher is better.

Importance Score	AUPRC ↑						Proportion of Sentences Removed ↑					
	ECT	CSDS	CNN/DM	TLDR-AIC	TLDR-Full	MTS	ECT	CSDS	CNN/DM	TLDR-AIC	TLDR-Full	MTS
Original Article	0.10	0.27	0.10	0.06	0.014	0.81	0.00	0.00	0.00	0.00	0.00	0.00
Cos. Sim. Centrality	0.22	0.34	0.34	0.35	0.14	0.86	0.22	0.11	0.18	0.29	0.5	0.18
Sentence Centrality	0.14	0.34	0.29	0.28	0.09	0.86	0.17	0.08	0.22	0.30	0.50	0.10
GUSUM	0.21	0.44	0.33	0.21	0.09	0.90	0.11	0.24	0.27	0.15	0.20	0.13
LexRank	0.22	0.43	0.32	0.32	0.14	*[4]	0.16	0.12	0.20	0.37	0.35	*[4]
GPT-4o mini (binary)	0.12	0.34	0.13	0.08	0.02	0.83	0.24	0.22	0.26	0.22	0.28	0.08
GPT-4o mini	0.30	0.49	0.34	0.33	0.20	0.93	0.24	0.25	**0.30**	0.40	0.26	0.16
Llama3-8B	0.18	0.39	0.22	0.15	0.05	0.92	0.13	0.11	0.14	0.11	0.17	0.14
Qwen3-8B	0.17	0.38	0.22	0.16	0.04	0.91	0.13	0.11	0.09	0.14	0.12	**0.22**
Gemini 2.0 Flash-Lite	0.35	0.68	**0.42**	**0.39**	**0.25**	**0.95**	0.28	0.46	0.25	0.40	0.45	0.13
Gemini 2.5 Flash	**0.43**	**0.69**	0.36	0.34	0.24	0.94	**0.37**	**0.49**	0.26	**0.41**	**0.47**	0.14

We shifted our financial forecast for prioritizing cash and liquidity given the uncertainties and we delivered another year of outstanding cash flow, our fourth consecutive year of cash flow greater than $1 billion. In addition, in our 2020 Sustainability Report, we committed to the ambitious goals of reducing our Scope 1 and 2 greenhouse gas emissions by one-third by 2030 and achieving carbon neutrality by 2050. ~~As a result, EBIT declined by about $100 million just related to this additional inventory actions we took. If volume is flat in '21 compared with '20, we would have about a $100 million tailwind from this improved utilization as we go into this year or about $0.60 a share. Looking at our cost structure, you'll recall that we reduced costs by approximately $150 million in 2020 versus '19, and we estimate about $100 million of this was temporary.~~ When we put this together, we expect our '21 adjusted earnings per share will increase between 20% and 30% compared to 2020. ~~You recall in the first quarter of 2020, our earnings per share was up 15% year-over-year, a very strong performance for our industry at that time.~~ Finally, on cash, a high priority for Eastman, we expect '21 to be our fifth consecutive year of free cash flow above $1 billion. Over the next two years, Eastman will invest approximately $250 million in the facility, which will support Eastman's commitment to addressing the global waste crisis and mitigating challenges created by climate change, while also creating value for shareholders. Using the company's polyester renewal technology, this new facility will use 110 kmt [Phonetic] of plastic waste to produce premium high-quality specialty plastics made with recycled content. This will not only reduce the company's use of fossil fuels feedstocks, but it will also reduce our greenhouse gas emissions by 20% to 30%.

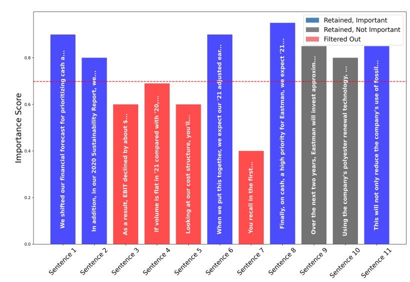


Figure 5: Example of Conformal Importance Summarization using Gemini 2.5 Flash scores. **Left:** Ground-truth summary sentences in blue; text with a strikethrough indicates sentences filtered out by Conformal Importance Summarization ($\alpha = 0.2$, $\beta = 0.8$). On this example, all important sentences were retained ($B(y; y^*) = 1$) while 36% of all sentences were filtered out. **Right:** Sentence-wise importance scores $R(c; x)$ compared to the conformal threshold $\hat{q}$ (dashed line).

5.4 Ablations

First, we perform ablations over the labeling method for datasets lacking explicit ground truth. Specifically, in Appendix C.4 we test ROUGE-1, -2, and -L scores [39] instead of SBERT cosine similarity, in conjunction with Algorithm 1 to create importance labels, then re-evaluate AUPRC and conciseness performance for all importance scoring methods. Aside from small changes in the scores, the performance trend is similar, and no fundamental difference in our conclusions would have been made with another labeling method.

Our method relies on a labeled calibration set which could be expensive to curate. In Appendix C.5 we perform an ablation over the calibration set size n. Compared to $n = 100$ used throughout this work, having as few as 50 labeled examples can still produce comparable results.

5.5 Comparison to Direct Abstractive and Hybrid Extractive-Abstractive Summarization

Next, we evaluate existing LLMs prompted to directly summarize text while retaining at least a fraction β of important information. To give these models guidance on *what* information is considered important, we add 10 examples from the calibration set to the prompt, enabling in-context learning [9]. We also test our proposed hybrid pipeline from Section 3 where an LLM starts with our extractive summary and is prompted to retain *all* information. If successful, this would preserve the empirical coverage level, while improving paragraph flow and potentially shortening the text further. We use Gemini 2.5 Flash for the extractive component, as it showed the strongest results in Table 2, and test GPT-4o mini as the abstractive model (similar plots for Gemini 2.5 Flash are given in Appendix C.6).

Unlike extractive summarization, in abstractive summarization determining whether the important aspects of a ground-truth sentence have been preserved is subjective. Hence, we use a proxy for recall $B(y; y^*)$ based on semantic entailment: each ground-truth important sentence in y^* is compared to

[4] LexRank's bag of words embedding step fails on this dataset, hence no scores are available.

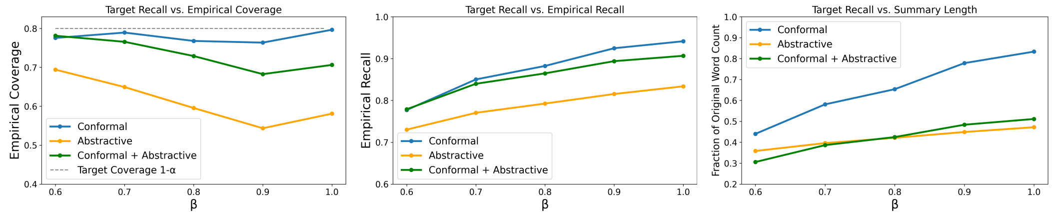


Figure 6: Comparison between extractive summarization with our method, abstractive summarization with an LLM, and our hybrid proposal on ECTSum. Here the target coverage is $1 - \alpha = 0.8$, the conformal approach uses Gemini 2.5 Flash scoring, and the abstractive model is GPT-4o mini.

the generated summary y using an LLM-based evaluator to check if the important aspects of the sentence have been retained. In practice we use GPT-4o mini with prompts given in Appendix B.

In Figure 6 using ECTSum we first observe that direct prompting to retain a fraction β of important information is not effective, achieving coverage on only 55-70% of summaries. While the hybrid approach has lower coverage than conformal alone, it greatly outperforms direct prompting. Notably, its improved recall and coverage are achieved with the same level of conciseness as direct prompting. We attribute this to how the hybrid approach separates out the two subtasks of identifying important information, and synthesizing it into a concise summary. Perhaps surprisingly, we note that direct prompting for different levels of β does give some level of control over recall.

6 Conclusions and Limitations

Our results show that Conformal Importance Summarization provides distribution-free coverage guarantees for extractive summarization. It allows integration with both LLMs and classical NLP techniques to achieve flexible and reliable summary generation with control over the tradeoff between completeness and conciseness. Furthermore, we show that these results hold with a variety of importance scores and labeling methods, which could reflect different user preferences for importance.

Our experiments suggest that recent LLMs are well-suited to judge sentence-level importance. LLMs prompted to directly judge importance can achieve higher sentence removal rates than our conformal method for a single recall level, but do not provide any control over the desired recall or conciseness. An LLM prompted to directly summarize inputs while retaining all important information tends to produce very concise summaries, but with much lower coverage and recall than requested. Separating the subtasks of judging importance and synthesizing information greatly improves recall without sacrificing conciseness. Since we did not perform extensive prompt tuning, we suspect that greater performance could be achieved with our method through tuning, or through reasoning models such as Deepseek-R1 [30].

While our benchmark suite spanned the domains of finance, customer service, news, science, and medicine with both English and Chinese examples, we note that it only contained two datasets with explicit ground-truth labels for extractive summarization, while the other three datasets required label refinement. Our datasets are somewhat limited in maximum length - with the exception being SciTLDR-Full at an average length of 216 sentences. Meanwhile, electronic health records may exceed 200,000 words [19]. Implementing Conformal Importance Summarization for such a problem would be a valuable step towards validating its practical utility and establishing standards for α and β.

Towards the goal of summarizing longer documents, it is possible to extend our framework to spans of text other than sentences, which we focused on for convenience. Simply break the long-text into spans, score their importance, and filter out spans with scores lower than the calibrated threshold. For example, paragraph-sized spans would reduce the number of spans that need to be scored for long documents. Due to the lack of existing datasets with long documents and labeled extractive summaries at the paragraph level, we have not performed these experiments.

Future work could extend our framework to more general forms of labels. One possibility is to perform extractive summarization with non-binary labels, such as an annotator's Likert scale rating of importance, and provide guarantees of including sentences with a specified fraction of the total importance weight. A more ambitious goal is to extend the framework fully to abstractive summarization. This would enable us to better leverage LLMs' natural abstractive capabilities, allow better evaluations on a wide range of datasets, and provide more natural sounding summaries.

References

[1] Josh Achiam et al. GPT-4 Technical Report. *arXiv:2303.08774*, 2023.

[2] Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv:2107.07511*, 2021.

[3] Anastasios N. Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

[4] Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024.

[5] Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, and Dominic Pimenta. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine*, 8(1), 2024. doi: 10.1038/s41746-025-01670-7.

[6] Shrestha Basu Mallick and Logan Kilpatrick. Gemini 2.0: Flash, Flash-Lite and Pro. `https://developers.googleblog.com/en/gemini-2-family-expands/`, February 2025. Accessed May 15, 2025.

[7] Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.168.

[8] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics, 2015. doi: 10.18653/v1/D15-1075.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[10] Evgeny Burnaev and Vladimir Vovk. Efficiency of conformalized ridge regression. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 605–622, 2014.

[11] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.428.

[12] Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*, 2024.

[13] Yen-Chun Chen and Mohit Bansal. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, 2018. doi: 10.18653/v1/P18-1063.

[14] John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. Large language model validity via enhanced conformal prediction methods. In *Advances in Neural Information Processing Systems*, volume 37, pages 114812–114842, 2024.

[15] 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. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[16] Jesse C. Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. Conformal Prediction Sets Improve Human Decision Making. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 9439–9457, 2024.

[17] Jesse C. Cresswell, Bhargava Kumar, Yi Sui, and Mouloud Belbahri. Conformal Prediction Sets Can Cause Disparate Impact. In *The Thirteenth International Conference on Learning Representations*, 2025.

[18] Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen Wong, Graham Wills, Elliot First, Frank Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. Current and future state of evaluation of large language models for medical summarization tasks. *npj Health Systems*, 2 (1):6, 2025.

[19] Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen K. Wong, Graham Wills, Elliot First, Miranda Schnier, Kyle Burton, Cris G. Ebby, Jillian Gorskic, Matthew Kalscheur, Samy Khalil, Marie Pisani, Tyler Rubeor, Peter Stetson, Frank Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. Development and validation of the provider documentation summarization quality instrument for large language models. *Journal of the American Medical Informatics Association*, pages 1050–1060, 2025.

[20] Daniel Deutsch, Rotem Dror, and Dan Roth. 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. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.753.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019. doi: 10.18653/v1/N19-1423.

[22] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

[23] Naihe Feng, Yi Sui, Shiyi Hou, Jesse C. Cresswell, and Ga Wu. Response quality assessment for retrieval-augmented generation via conditional conformal factuality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2832–2836, 2025. ISBN 9798400715921. doi: 10.1145/3726302.3730244. URL `https://doi.org/10.1145/3726302.3730244`.

[24] Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. Bottom-Up Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. doi: 10.18653/v1/D18-1443.

[25] Shilpa Ghatnekar, Adam Faletsky, and Vinod E Nambudiri. Digital scribe utility and barriers to implementation in clinical practice: A scoping review. *Health and Technology*, 11(4):803–809, 2021.

[26] Tuba Gokhan, Phillip Smith, and Mark Lee. GUSUM: Graph-based Unsupervised Summarization Using Sentence Features Scoring and Sentence-BERT. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, pages 44–53. Association for Computational Linguistics, 2022.

[27] Shuai Gong, Zhenfang Zhu, Jiangtao Qi, Chunling Tong, Qiang Lu, and Wenqing Wu. Improving extractive document summarization with sentence centrality. *PLOS ONE*, 17(7):1–16, 07 2022. doi: 10.1371/journal.pone.0268278.

[28] Google Developers. Developers can now start building with Gemini 2.5 Flash. `https://blog.google/products/gemini/gemini-2-5-flash-preview/`, April 2025. Accessed May 15, 2025.

[29] Aaron Grattafiori et al. The Llama 3 Herd of Models. *arXiv:2407.21783*, 2024.

[30] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv:2501.12948*, 2025.

[31] Chirag Gupta, Arun K. Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022. ISSN 0031-3203. doi: 10.1016/j.patcog.2021.108496.

[32] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[33] Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[34] Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. Text summarization from legal documents: A survey. *Artificial Intelligence Review*, 51:371–402, 2019.

[35] Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Comput. Surv.*, 55(8), December 2022. ISSN 0360-0300. doi: 10.1145/3545176.

[36] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 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, 2020. doi: 10.18653/v1/2020.emnlp-main.750.

[37] Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv:2305.18404*, 2023.

[38] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

[39] Chin-Yew 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.

[40] Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. CSDS: A Fine-Grained Chinese Dataset for Customer Service Dialogue Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4436–4451, 2021. doi: 10.18653/v1/2021.emnlp-main.365.

[41] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, 2019. doi: 10.18653/v1/D19-1387.

[42] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, 2019. doi: 10.18653/v1/D19-1387.

[43] Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Copula-based conformal prediction for multi-target regression. *Pattern Recognition*, 120:108101, 2021. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2021.108101.

[44] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411. Association for Computational Linguistics, July 2004.

[45] Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[46] Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. ECTSum: A New Benchmark Dataset For Bullet Point Summarization of Long Earnings Call Transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, 2022. doi: 10.18653/v1/2022.emnlp-main.748.

[47] OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`, 2025. Accessed May 15, 2025.

[48] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Infolab, 1999.

[49] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM Evaluators Recognize and Favor Their Own Generations. In *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802, 2024.

[50] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*, 2018.

[51] David Preti, Cristina Giannone, Andrea Favalli, and Raniero Romagnoli. Automatic Summarization of Legal Texts, Extractive Summarization using LLMs. *Ital-IA 2024: 4th National Conference on Artificial Intelligence*, 2024.

[52] Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*, 2024.

[53] Juan Ramirez-Orta and Evangelos Milios. Unsupervised document summarization using pre-trained sentence embeddings and graph centrality. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 110–115. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.sdp-1.14.

[54] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410.

[55] Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[56] Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[57] Maxon Rubin-Toles, Maya Gambhir, Keshav Ramji, Aaron Roth, and Surbhi Goel. Conformal language model reasoning with coherent factuality. In *The Thirteenth International Conference on Learning Representations*, 2025.

[58] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022.

[59] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[60] Shreya J. Shah, Trevor Crowell, Yejin Jeong, Anna Devon-Sand, Margaret Smith, Betsy Yang, Stephen P. Ma, April S. Liang, Clarissa Delahaie, Caroline Hsia, Tait Shanafelt, Michael A. Pfeffer, Christopher Sharp, Steven Lin, and Patricia Garcia. Physician Perspectives on Ambient AI Scribes. *JAMA Network Open*, 8(3):e251904–e251904, 2025.

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[62] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

[63] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

[64] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report. *arXiv:2412.15115*, 2024.

[65] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

[66] Haopeng Zhang, Philip S Yu, and Jiawei Zhang. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 2024.

[67] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11328–11339, 2020.

[68] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide evidence supporting each claim made in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 outlines the limitations that affect our proposed method. Other limitations and the choices we made to surmount them are outlined throughout the text.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Full proofs are provided in the main text.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We disclose full details on the selection and processing of datasets, algorithm used to producing labels, scores, thresholds, and selections of sentences for the final evaluation. Datasets are publicly available.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our codebase is available at github.com/layer6ai-labs/conformal-importance-summarization. Datasets are publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental details are disclosed in Section 4 and Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: While error bars and statistical analysis are preferred, we were not able to provide them on all experiments due to the cost associated with rerunning experiments that utilize commercial LLMs. Select experiments do come with error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Computational costs for conformal prediction are negligible and not a barrier to reproduction. The other computational costs come from LLM inference for which we used commercial APIs, and locally hosted open-source models run on an A6000 GPU with 48 GB of memory.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms with the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Summarization is a standard NLP task performed by AI and non-AI means - our method does not introduce new negative societal impacts. Benefits are discussed in the motivation.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release models or data with a high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the creators of all models and datasets we use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code has been released at github.com/layer6ai-labs/conformal-importance-summarization, including documentation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not use human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not use human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Our uses of LLMs are standard.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Additional Experimental Details

A.1 Dataset Processing Details

Some modifications were required to make datasets amenable to our extractive summarization setting.

- **TLDR-AIC**: We selected the Abstract-Introduction-Conclusion subset to provide an appropriate amount of input context for summarization. A large portion of this dataset consists of single-sentence summaries. We focus on samples where multiple versions of single-sentence summaries have been written from different perspectives (e.g. author, reviewer), a total of 1143 samples, and pool those sentences into a single summary as the ground truth.
- **TLDR-Full**: This version of SciTLDR uses the full text of a scientific paper as the input, making inputs much larger and summaries a much smaller fraction of the long-text. Otherwise, processing is the same as TLDR-AIC.
- **MTS-Dialog**: This dataset contains question and answer-style conversations between doctors and patients. Hence, certain sentences require context for correct interpretation. To accommodate this, each question from the doctor and all subsequent patient responses were merged into a single sentence unit, following the pattern "Doctor: *questions*. Patient: *answers*". If a sample input contained two or fewer sentences after this merge, the entire sample is removed from the dataset; 1129 samples remained after this filtering.

Dataset Licenses ECT is released under a GNU GPL license. CSDS is provided for use without a specific license. CNN/DM and SciTLDR are both released under an Apache-2.0 License. MTS-Dialog is released under a Creative Commons Attribution 4.0 International license. Hence, our usages of these five datasets are permissible under their respective licenses.

A.2 Greedy Optimization for Extractive Summarization Labeling

The following greedy oracle labeling method is used to provide sentence-level importance labels for datasets that lack sentence-level ground-truth annotations. All we require is that a reference summary r be available, which could be abstractive rather than extractive. The output of the algorithm is an extractive summary that can be used as ground truth y^* in Conformal Importance Summarization. First, for a long-text x we compute a similarity score V between each sentence c_i and the reference summary, then sort sentences by score in descending order. We then iterate through the ranked sentences in a greedy manner and add them to the extractive summary if the overall similarity of the combined extractive summary to r increases by at least δ. Mathematically, if the current extractive summary is y^*_{curr} with similarity to r of $V(y^*_{\text{curr}}; r)$, the sentence c_i is added to y^*_{curr} if

$$V(y^*_{\text{curr}} \cup \{c_i\}; r) - V(y^*_{\text{curr}}; r) > \delta. \tag{16}$$

After one iteration through the sentences in x, we return y^*_{curr} to be used as the ground truth extractive summary y^*. This process is depicted in Algorithm 1.

Algorithm 1: Greedy Optimization for Extractive-Summarization Labeling

```
Input: Sentences x = [c_1, ..., c_p], reference summary r, scoring function V(·;·), threshold δ
Output: Extractive summary y*
Compute v_i = V(c_i; r) for all i;
Sort indices by descending v giving the permutation π = [π_1, ..., π_n];
y*_curr ← ∅
V_curr ← 0
for j = 1 to p do
    i ← π_j;
    Δ ← V(y*_curr ∪ {c_i}; r) − V_curr;
    if Δ > δ then
        y*_curr ← y*_curr ∪ {c_i};
        V_curr ← V_curr + Δ;
return y*_curr
```

B LLM Prompts

Ground Truth Labelling. To generate LLM-based ground-truth labels for important sentences in extractive summarization where sentence-level labels were not available, we provided input sentences as separate strings to GPT-4o mini, along with the existing summary text from the dataset. The prompt in Listing 1 was used to produce a list of ground-truth labels corresponding to each input sentence.

```
"""
Evaluate whether each input claim is included in the summary text. The output labels
    , corresponding to each input claim, should be either 0 or 1, indicating
    whether the corresponding claim, or the information it carries, is indeed
    included in the actual summary. For example, if claim_1's information is
    contained in the summary, then label_1 should be 1; if information carried in
    claim_3 cannot be found in the summary text, then label_3 should be 0.

Summary text:
{summary_text}

List of claims:
{[claim_text]}
"""
```

Listing 1: System Prompt for Creating Sentence-Level Ground-Truth Labels

Importance Scoring. To generate LLM-based scores $R(c_i; y^*)$ for Conformal Importance Summarization, the source text was provided to an LLM along with the individual sentences from that text as separate strings. The prompt in Listing 2 was used to produce a list of importance scores between 0 and 1 corresponding to each input sentence.

```
"""
Please evaluate the importance of each input claim in the original text, based on
    how the information carried in the claim is aligned with the overall message.
    Please provide a importance score for EACH input claim. Each output score
    should be a two decimal float number ranged between 0 and 1, indicating how
    important the corresponding input claim is in the context of the text document.
     For example, if claim 1's information is highly aligned with that of the input
     text, and very likely to be included in the summary, then score 1 should be
    close to 1, say greater than 0.8; if information carried in claim 3 is trivial
    or only remotely related to the central message of the text, and is not worthy
    of inclusion in the summary, then score 3 should be close to 0, say less than
    0.2.

Original text:
{original_text}

List of claims:
{[claim_text]}
"""
```

Listing 2: System Prompt for Creating Importance Scores

We also experimented with prompting the LLM to give binary scores $R(c_i; y^*)$, rather than floats. The prompt in Listing 3 was used to produce a list of importance scores corresponding to each input sentence as above.

```
"""
Evaluate the importance of each input claim in the original text, based on how the
    information carried in the claim is aligned with the overall message. Please
    provide a binary importance score for EACH input claim. Each output score
    should be either 0 or 1, indicating whether the corresponding input claim is
    important enough in the context of the text document to be included in the
    summary. For example, if claim 1's information is highly aligned with that of
    the input text, and very likely to be included in the summary, then score 1
```

```
    should be 1; if information carried in claim 3 is trivial or only remotely
    related to the central message of the text, and is not worthy of inclusion in
    the summary, then score 3 should be 0.

Original text:
{original_text}

List of claims:
{[claim_text]}
"""
```

Listing 3: System Prompt for Creating Importance Scores with Binary Restriction

Direct Abstractive Summarization. We tested the ability of instruction-tuned LLMs to directly create abstractive summaries that retain a specified fraction β of important information. To guide the LLM as to what type of information was important for a given dataset, ten examples from the calibration dataset were provided to enable in-context learning [9].

```
"""
Here are examples of what constitutes important information to include in the
    summary:

{examples_text}

Create an abstractive summary of the following text.

Requirements:
- Aim to retain at least {beta}% of the important information
- Use your own words and phrasing (abstractive, not extractive)

Input text to summarize:
{input_text}
"""
```

Listing 4: System Prompt for Direct Abstractive Summarization

Hybrid Extractive-Abstractive Summarization. To disentangle the subtasks of importance assessment and summary synthesis, we applied abstractive summarization with an LLM as a post-processing step after Conformal Importance Summarization. The abstractive step used the following prompt that specifies *all* information should be retained from the extractive summary input.

```
"""
Requirements:
- Use more concise language to make the text shorter
- Retain all of the information from the input text
- Improve flow, coherence, and readability
"""
```

Listing 5: System Prompt for Abstractive Summarization as a Post-Processing Step

Sentence-level Recall Estimation. Unlike for extractive summaries, determining if an abstractive summary has retained a specific piece of important information requires a judgement call. Hence, we estimated recall of sentence-level information using GPT4o-mini prompted to determine if the content of a sentence is retained in an abstractive summary.

```
"""
You will be given an important sentence from the original text and a generated
    summary. Your goal is to determine whether the important sentence given is
    retained in the generated summary.

Important sentence:
{important_sentence}

Generated summary:
```

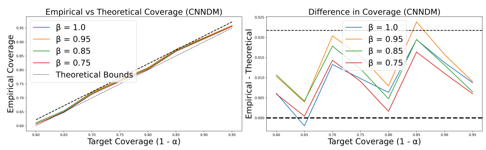
```
{summary}

Output True if the important sentence is retained in the generated summary. Output
    False otherwise.
"""
```

Listing 6: System Prompt for Determining if a Ground-truth Important Sentence is Retained in an Abstractive Summary

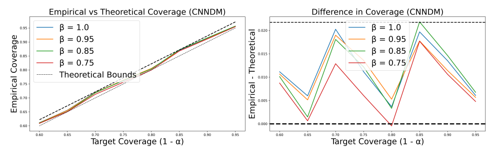
C Additional Experimental Results

C.1 Coverage Plots for all Datasets and Methods

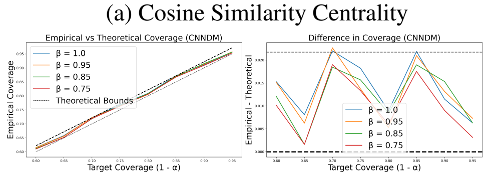
To supplement the results in Section 5.1, here we show the empirical coverage obtained by all methods across all datasets. This confirms that all numerical comparisons in our work are fair in that they all achieve the expected coverage level. In Figures 7 - 11 we shows plots analogous to Figure 2 in the main text. All parameters are identical to that in the main text. We see a similar trend where nearly all datapoints fall within the bounds for all plots, with occasional small deviations due to the inherent randomness involved and finite sample sizes.



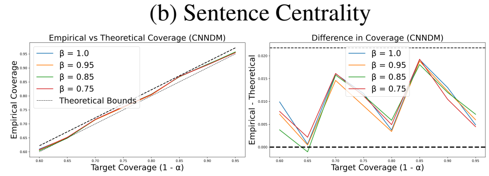
(a) Cosine Similarity Centrality



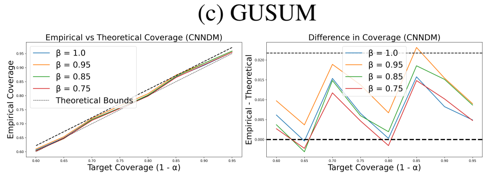
(b) Sentence Centrality



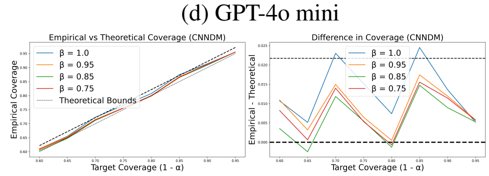
(c) GUSUM



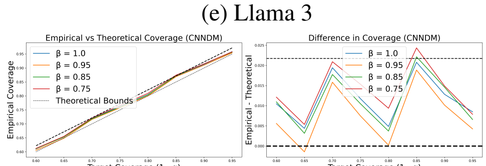
(d) GPT-4o mini



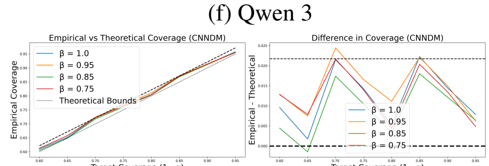
(e) Llama 3



(f) Qwen 3

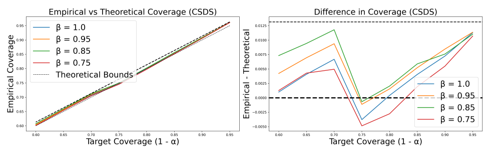


(g) Gemini 2.0 Flash-Lite

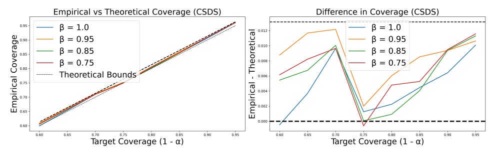


(h) Gemini 2.5 Flash

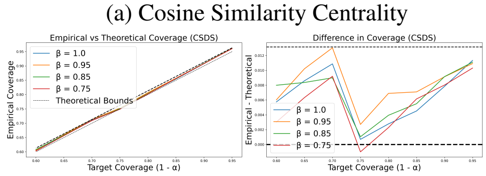
Figure 7: User-specified probability of achieving coverage (1-α) vs. empirical probability of achieving coverage, on the CNN/DM dataset. Dashed lines show theoretical bounds given in Theorem 1. Results are averaged over 400 random splits of calibration and test data.



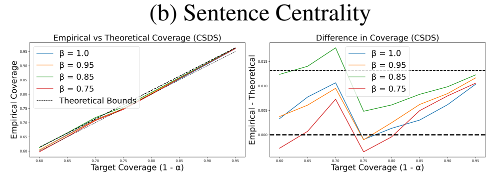
(a) Cosine Similarity Centrality



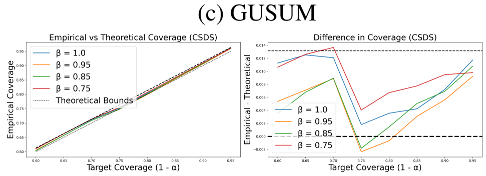
(b) Sentence Centrality



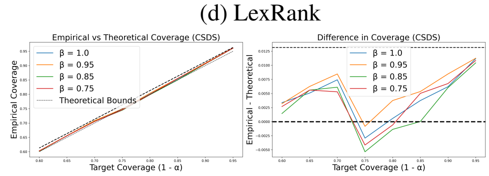
(c) GUSUM



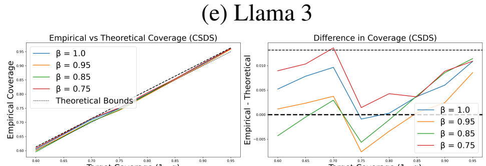
(d) LexRank



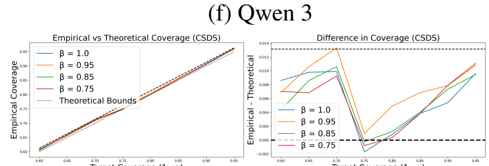
(e) Llama 3



(f) Qwen 3

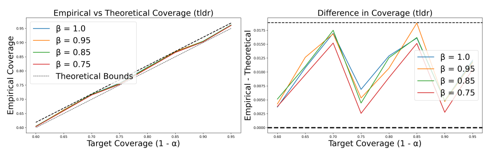


(g) Gemini 2.0 Flash-Lite

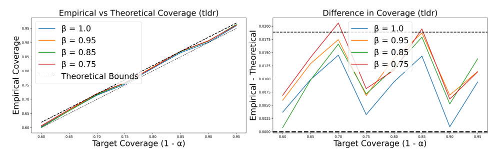


(h) Gemini 2.5 Flash

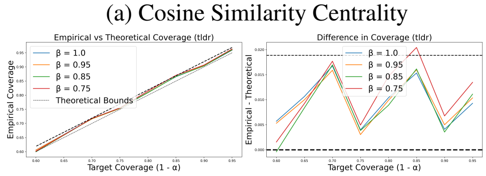
Figure 8: User-specified probability of achieving coverage (1-α) vs. empirical probability of achieving coverage, on the CSDS dataset. Dashed lines show theoretical bounds given in Theorem 1. Results are averaged over 400 random splits of calibration and test data.



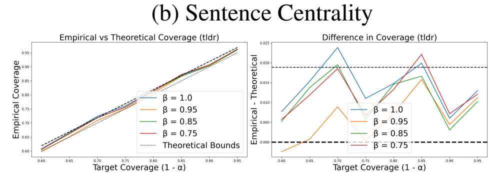
(a) Cosine Similarity Centrality



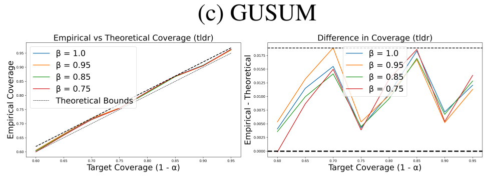
(b) Sentence Centrality



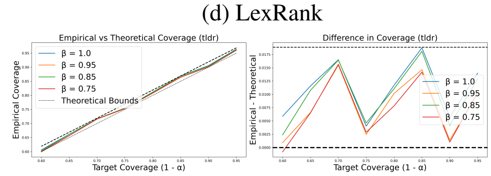
(c) GUSUM



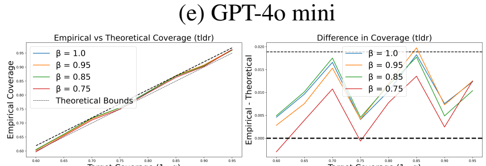
(d) LexRank



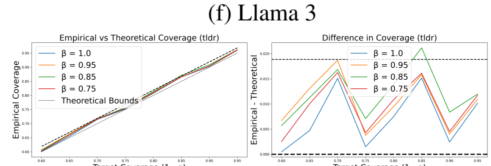
(e) GPT-4o mini



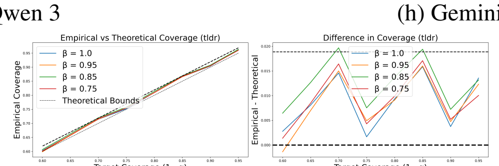
(f) Llama 3



(g) Qwen 3

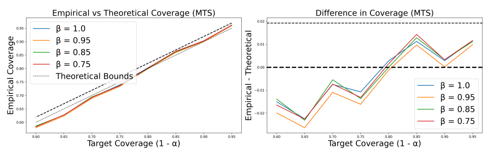


(h) Gemini 2.0 Flash-Lite

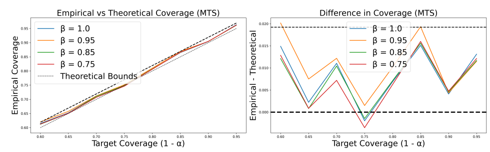


(i) Gemini 2.5 Flash

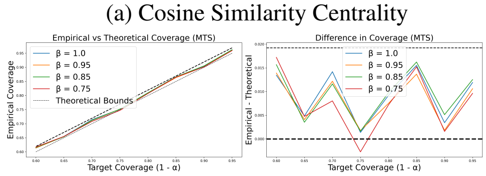
Figure 9: User-specified probability of achieving coverage (1-α) vs. empirical probability of achieving coverage, on the TLDR-AIC dataset. Dashed lines show theoretical bounds given in Theorem 1. Results are averaged over 400 random splits of calibration and test data.



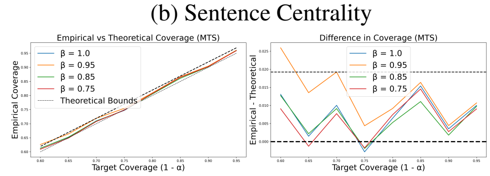
(a) Cosine Similarity Centrality



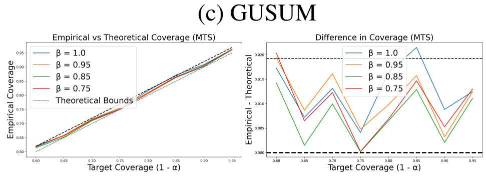
(b) Sentence Centrality



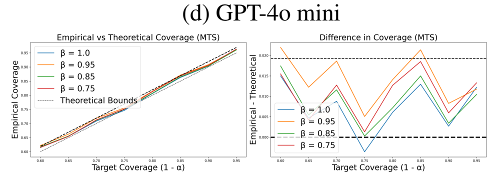
(c) GUSUM



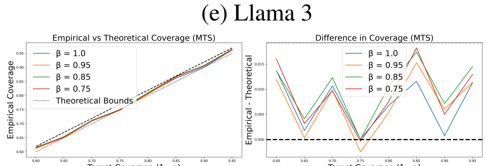
(d) GPT-4o mini



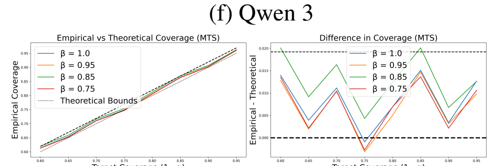
(e) Llama 3



(f) Qwen 3

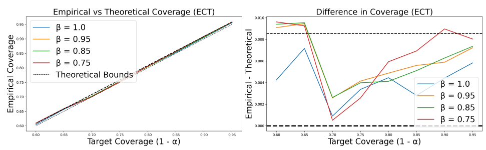


(g) Gemini 2.0 Flash-Lite

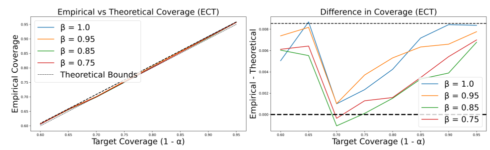


(h) Gemini 2.5 Flash

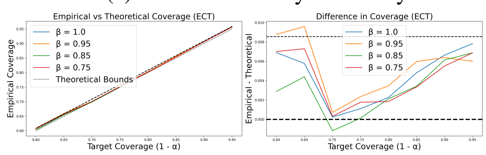
Figure 10: User-specified probability of achieving coverage (1-α) versus empirical probability of achieving coverage, on MTS dataset. Dashed lines show theoretical bounds given in Theorem 1. Results are averaged over 400 random splits of calibration and test data.



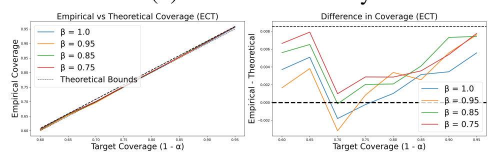
(a) Cosine Similarity Centrality



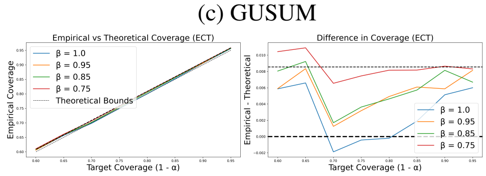
(b) Sentence Centrality



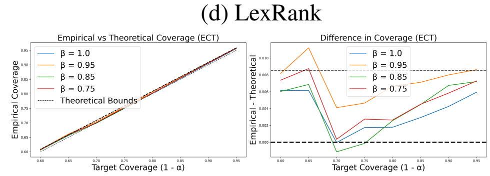
(c) GUSUM



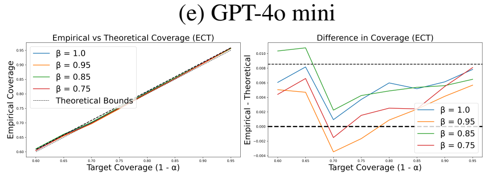
(d) LexRank



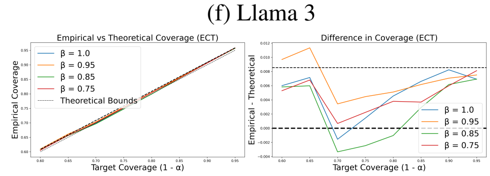
(e) GPT-4o mini



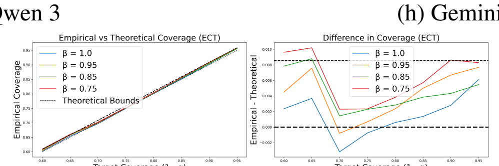
(f) Llama 3



(g) Qwen 3



(h) Gemini 2.0 Flash-Lite



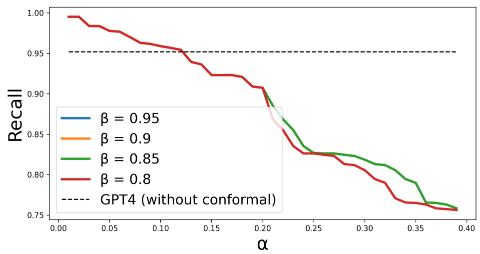
(i) Gemini 2.5 Flash

Figure 11: User-specified probability of achieving coverage (1-α) vs. empirical probability of achieving coverage, on the ECT dataset. Dashed lines show theoretical bounds given in Theorem 1. Results are averaged over 400 random splits of calibration and test data.

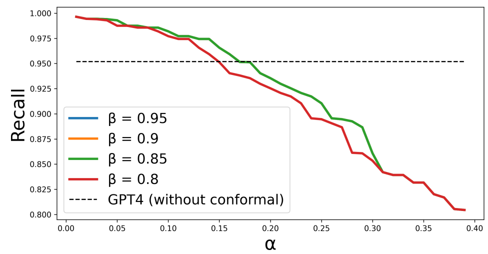
C.2 Error Rate vs. Recall Plots for all Datasets and Methods

Figures 12 - 16 show the empirical recall $B(y; y^*)$ based on the choice of error rate α for all datasets and methods[5], analogous to Figure 3 in Section 5.1. The trend is similar, with higher α leading to lower recall of important content. We notice that for TLDR-AIC and CNN/DM, many of the lines for different β values overlap one another. This may be due to the low number of sentences in the long-texts for these datasets, making some β values effectively equivalent for those x.

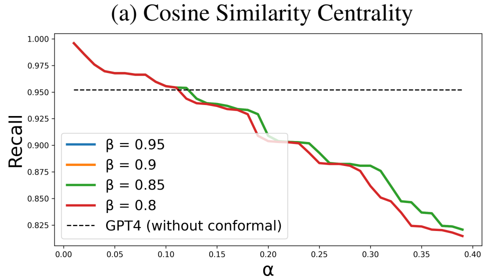
[5]Due to computational constraints, we only compute this plot for LexRank on the CNN/DM dataset



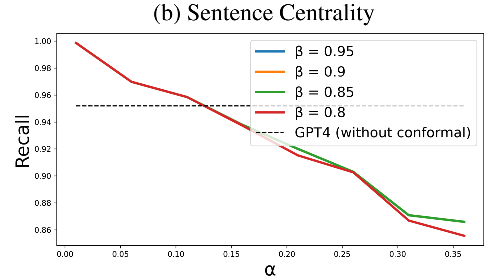
(a) Cosine Similarity Centrality



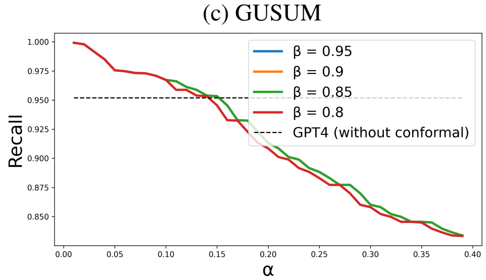
(b) Sentence Centrality



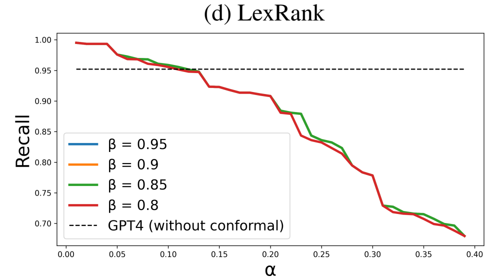
(c) GUSUM



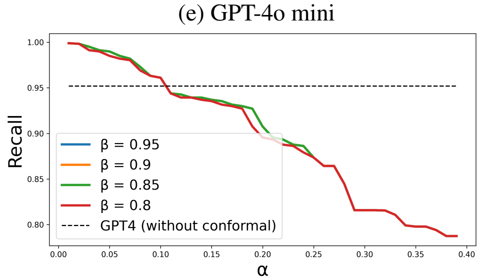
(d) LexRank



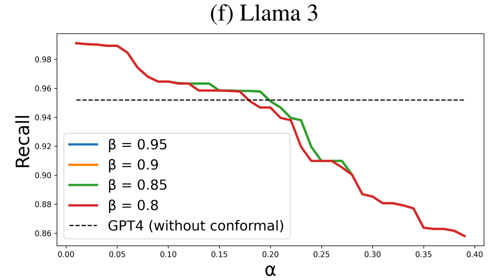
(e) GPT-4o mini



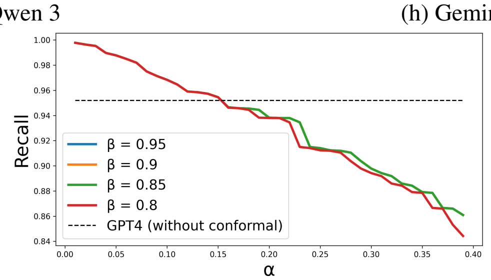
(f) Llama 3



(g) Qwen 3

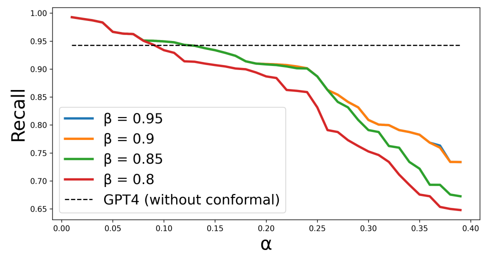


(h) Gemini 2.0 FlashLite

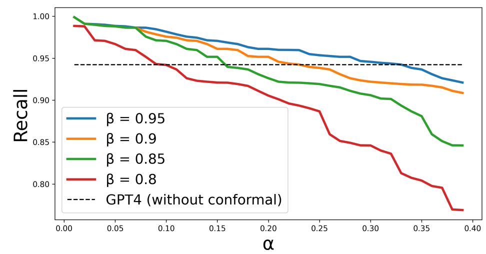


(i) Gemini 2.5 Flash

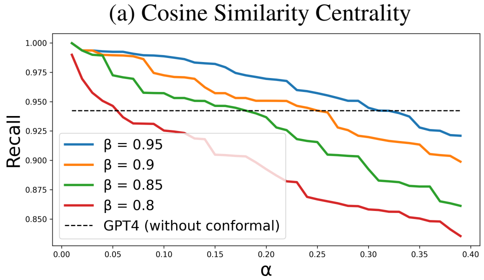
Figure 12: Target error rate α versus empirical recall $B(y; y^*)$ of important sentences in summaries, averaged over the CNN/DM test set. The dashed line shows GPT-4o mini performance without using conformal prediction. Several curves may overlap when there are only a few discrete levels of empirical recall possible, making some values of β equivalent.



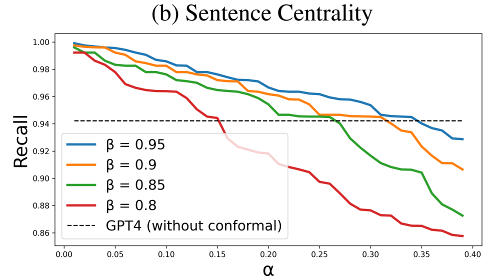
(a) Cosine Similarity Centrality



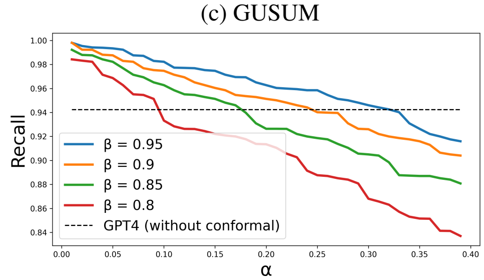
(b) Sentence Centrality



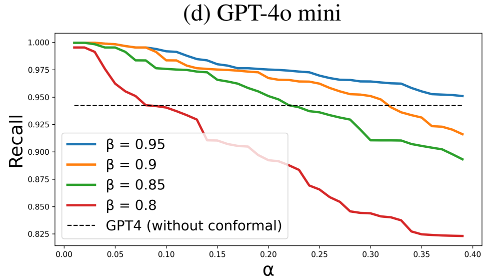
(c) GUSUM



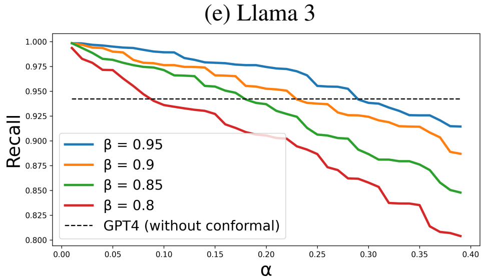
(d) GPT-4o mini



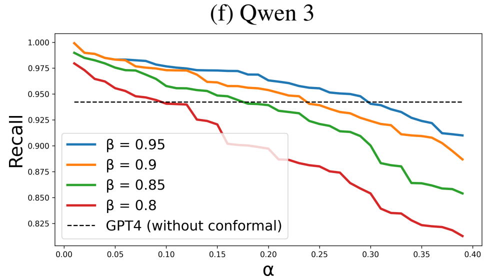
(e) Llama 3



(f) Qwen 3

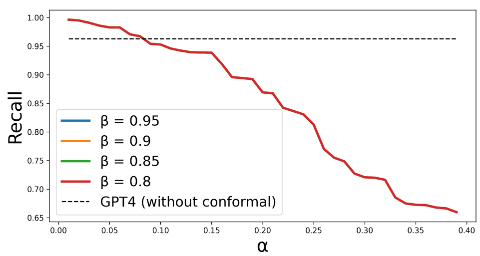


(g) Gemini 2.0 Flash-Lite

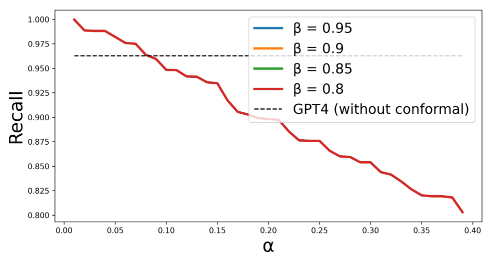


(h) Gemini 2.5 Flash

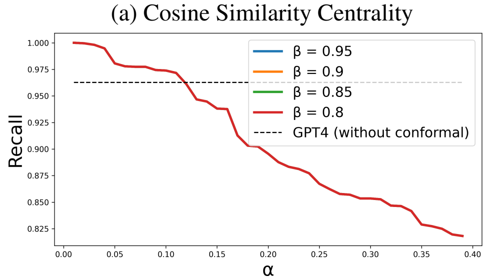
Figure 13: Target error rate α versus empirical recall $B(y; y^*)$ of important sentences in summaries, averaged over the CSDS test set. The dashed line shows GPT-4o mini performance without using conformal prediction.



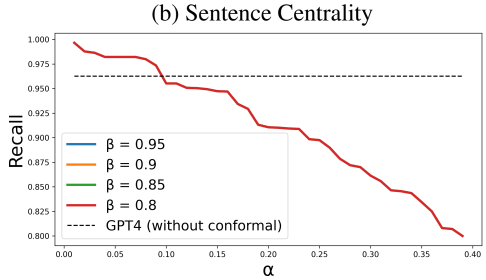
(a) Cosine Similarity Centrality



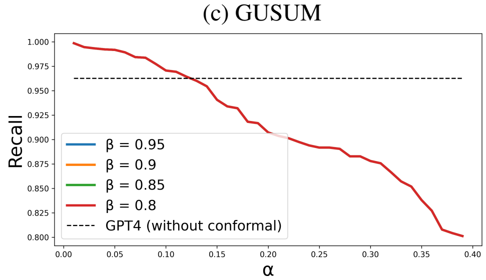
(b) Sentence Centrality



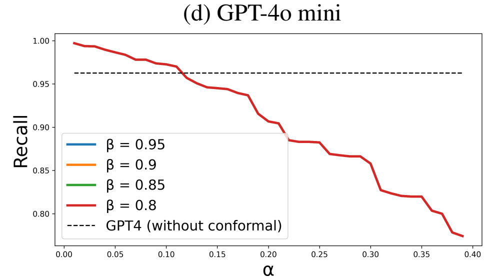
(c) GUSUM



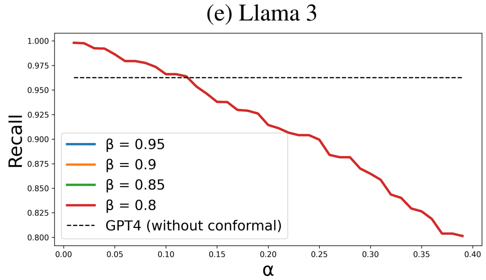
(d) GPT-4o mini



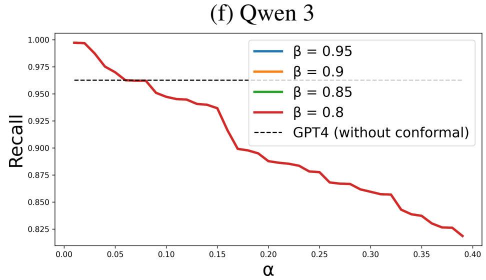
(e) Llama 3



(f) Qwen 3

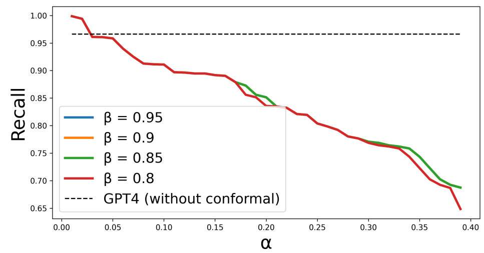


(g) Gemini 2.0 Flash-Lite

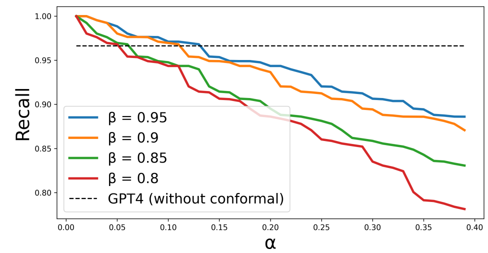


(h) Gemini 2.5 Flash

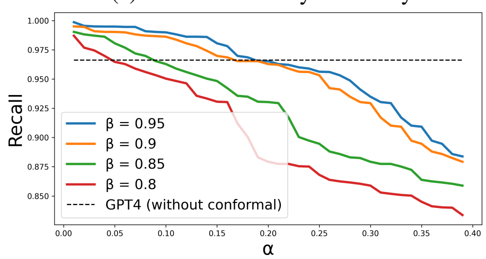
Figure 14: Target error rate α versus empirical recall $B(y; y^*)$ of important sentences in summaries, averaged over the TLDR-AIC test set. The dashed line shows GPT-4o mini performance without using conformal prediction. Several curves overlap because all datapoints in TLDR-AIC contain a small number of ground-truth sentences, meaning there are only a few discrete levels of empirical recall possible, making some values of β equivalent.



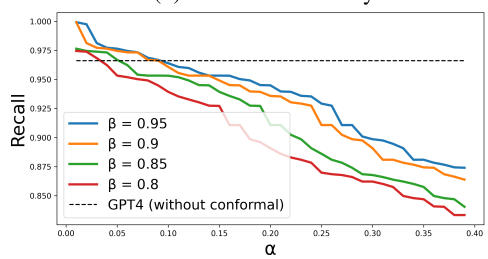
(a) Cosine Similarity Centrality



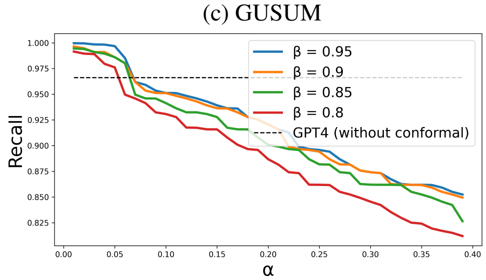
(b) Sentence Centrality



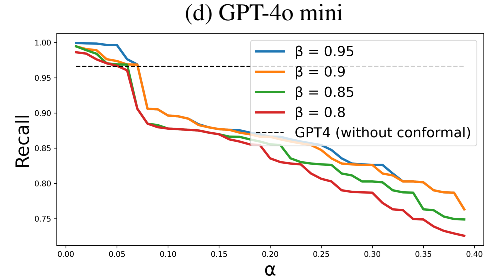
(c) GUSUM



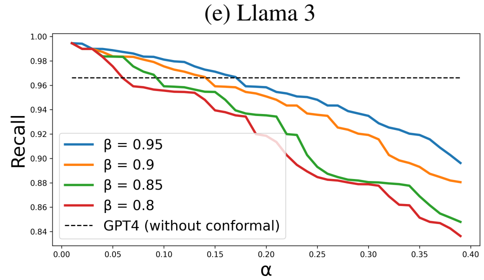
(d) GPT-4o mini



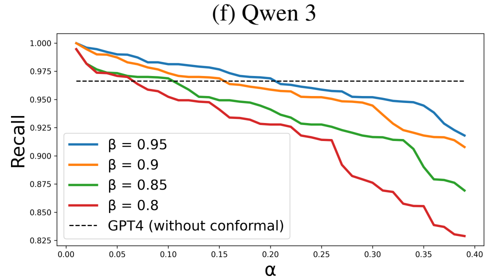
(e) Llama 3



(f) Qwen 3

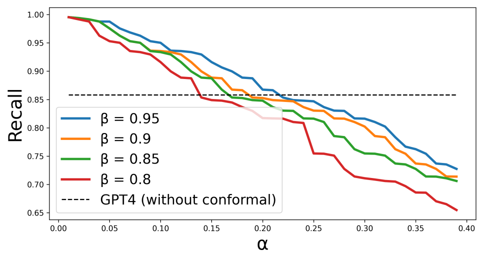


(g) Gemini 2.0 Flash-Lite

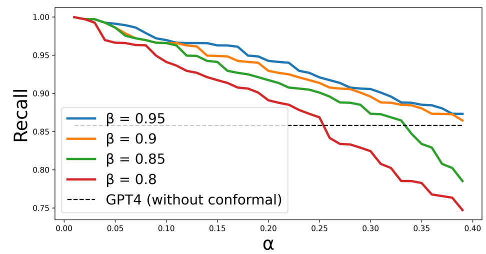


(h) Gemini 2.5 Flash

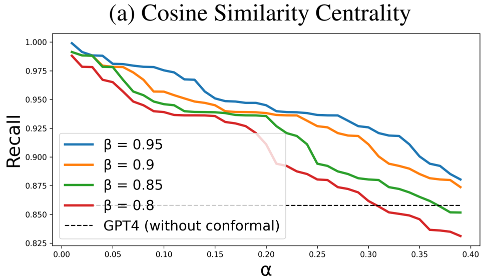
Figure 15: Target error rate α versus empirical recall $B(y; y^*)$ of important sentences in summaries, averaged over the MTS test set. The dashed line shows GPT-4o mini performance without using conformal prediction. Several curves may overlap when there are only a few discrete levels of empirical recall possible, making some values of β equivalent.



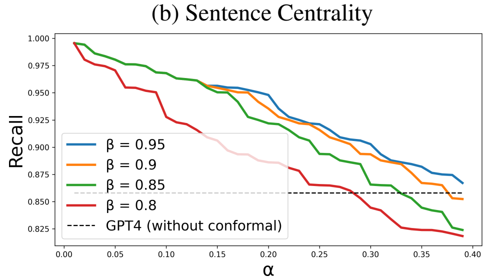
(a) Cosine Similarity Centrality



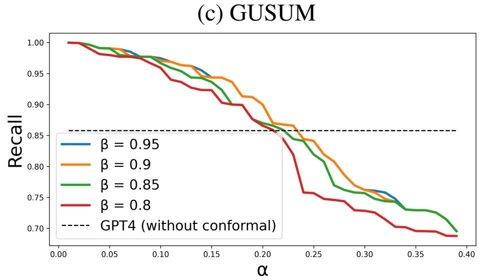
(b) Sentence Centrality



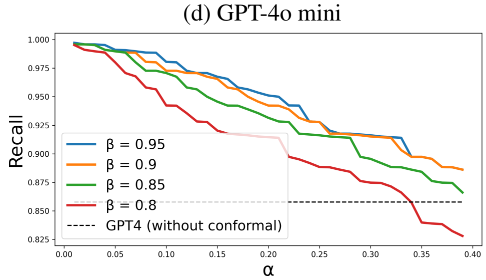
(c) GUSUM



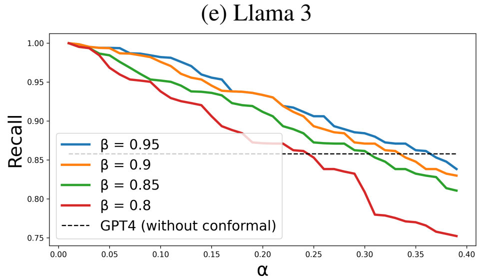
(d) GPT-4o mini



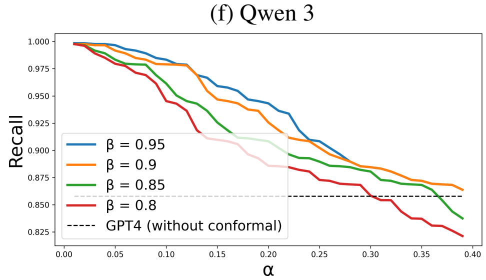
(e) Llama 3



(f) Qwen 3



(g) Gemini 2.0 Flash-Lite



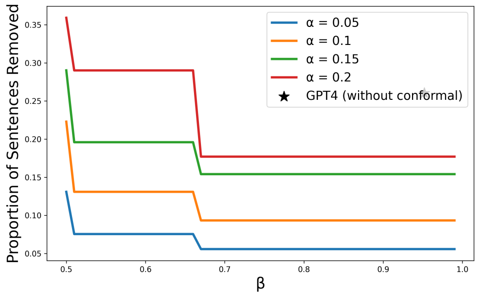
(h) Gemini 2.5 Flash

Figure 16: Target error rate α versus empirical recall $B(y; y^*)$ of important sentences in summaries, averaged over the ECT test set. The dashed line shows GPT-4o mini performance without using conformal prediction.

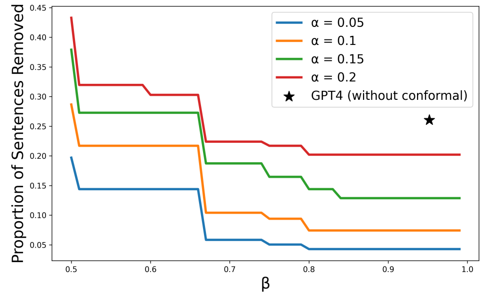
C.3 Target Recall vs. Conciseness Plots for all Datasets and Methods

Figures 17 - 20 show the conciseness, the proportion of sentences removed, based on the choice of β for all datasets and methods[6], analogous to Figure 4 in Section 5.1. Once again, the trend is highly similar across settings, with higher β leading to a smaller reduction in length.

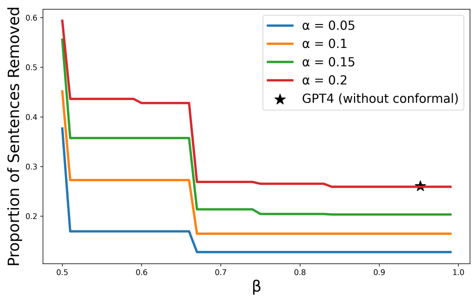
[6]Due to computational constraints, we only compute this plot for LexRank on the CNN/DM dataset



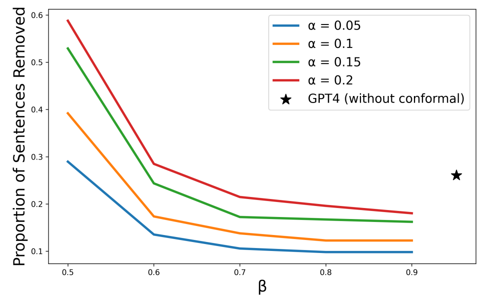
(a) Cosine Similarity Centrality



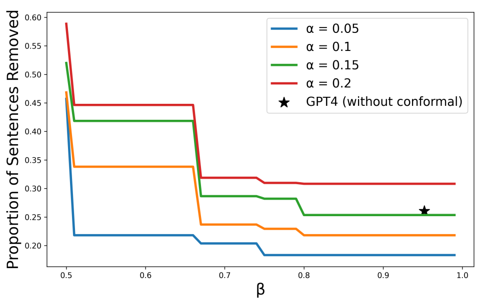
(b) Sentence Centrality



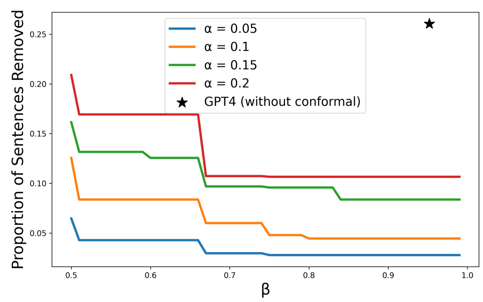
(c) GUSUM



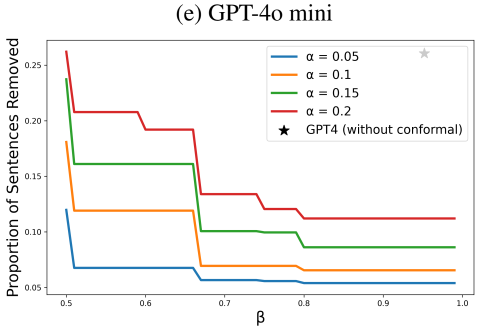
(d) LexRank



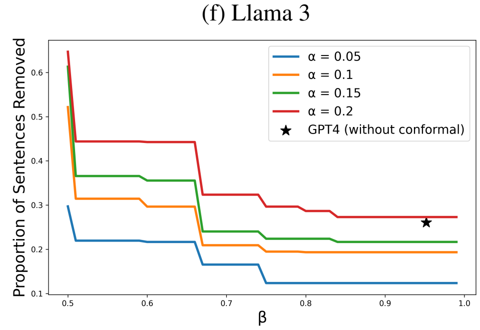
(e) GPT-4o mini



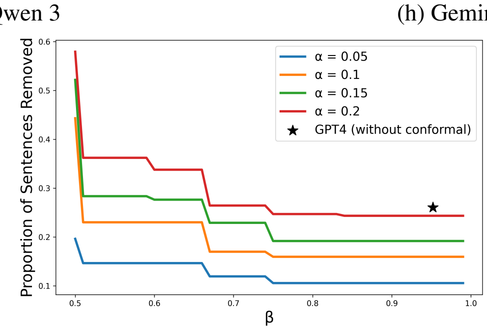
(f) Llama 3



(g) Qwen 3

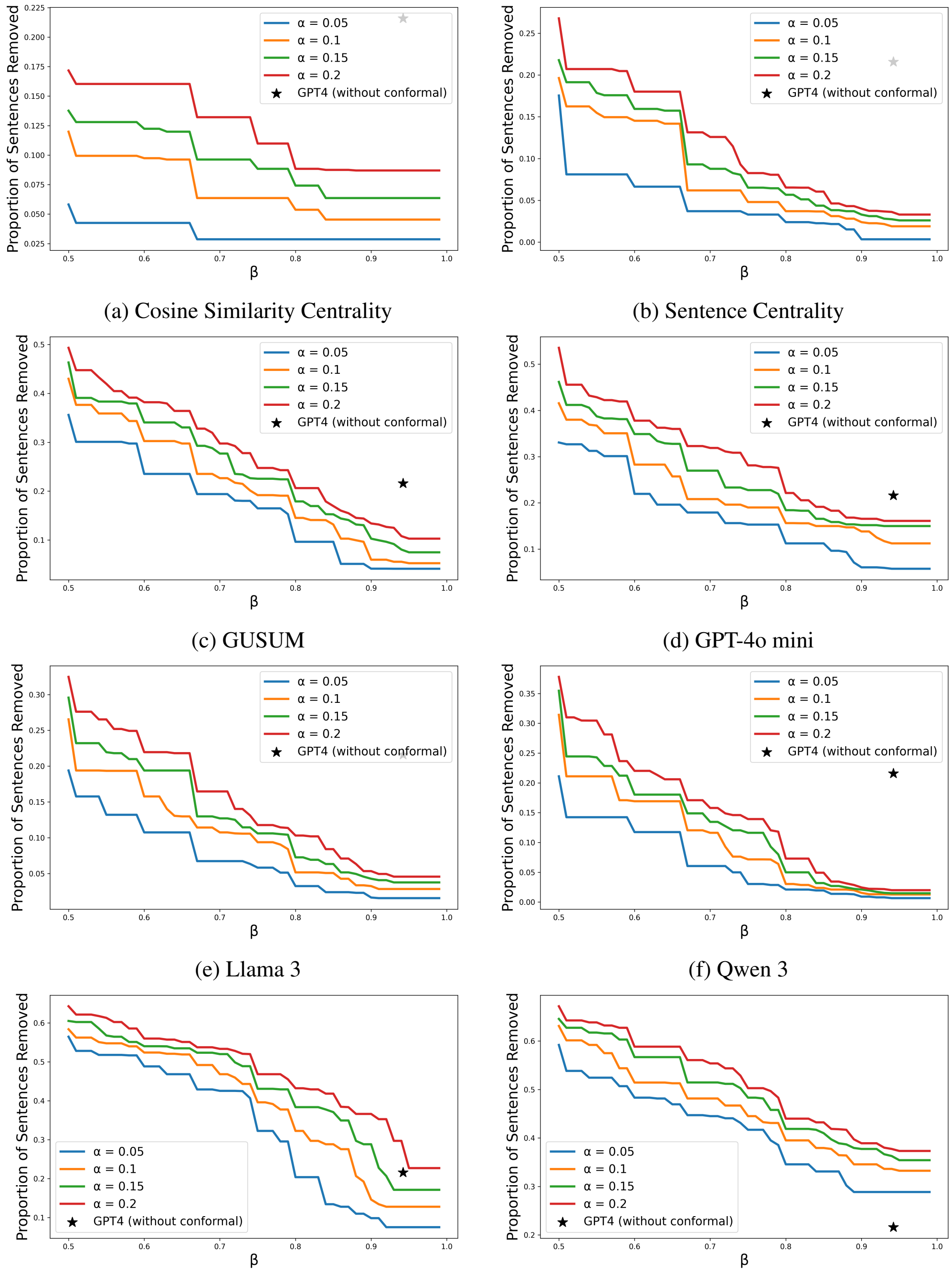


(h) Gemini 2.0 Flash-Lite



(i) Gemini 2.5 Flash

Figure 17: Target recall β vs. proportion of sentences removed (conciseness). Lines indicate different values for the target error rate α on CNN/DM.



(a) Cosine Similarity Centrality

(b) Sentence Centrality

(c) GUSUM

(d) GPT-4o mini

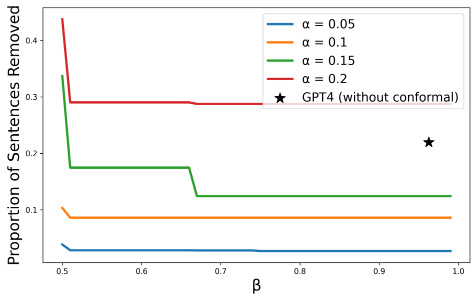
(e) Llama 3

(f) Qwen 3

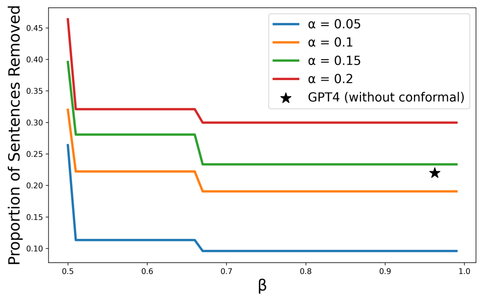
(g) Gemini 2.0 Flash-Lite

(h) Gemini 2.5 Flash

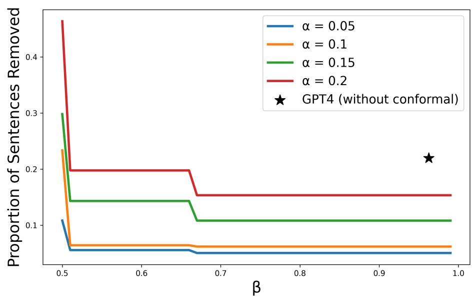
Figure 18: Target recall β vs. proportion of sentences removed (conciseness). Lines indicate different values for the target error rate α on CSDS.



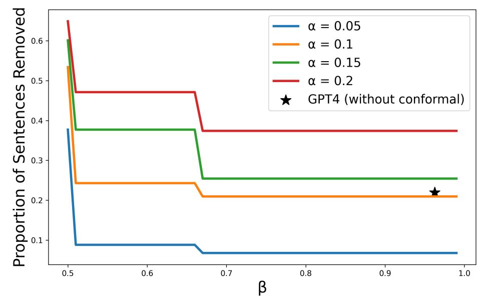
(a) Cosine Similarity Centrality



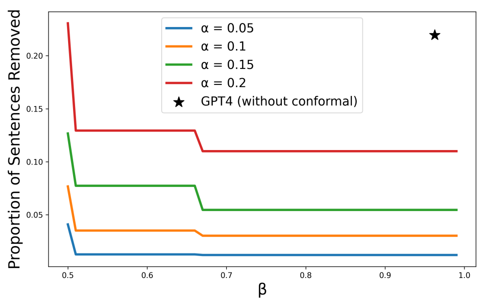
(b) Sentence Centrality



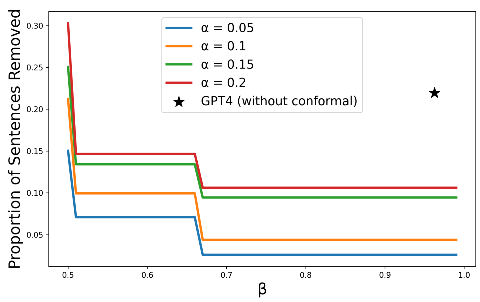
(c) GUSUM



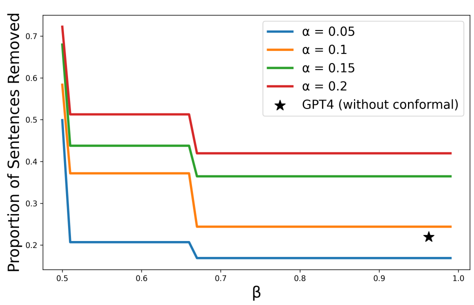
(d) GPT-4o mini



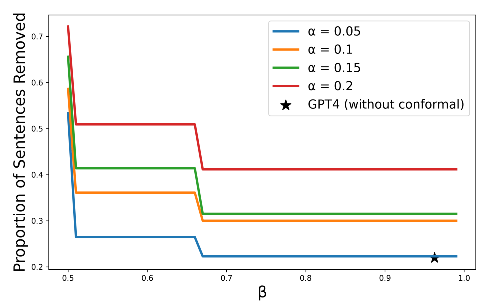
(e) Llama 3



(f) Qwen 3

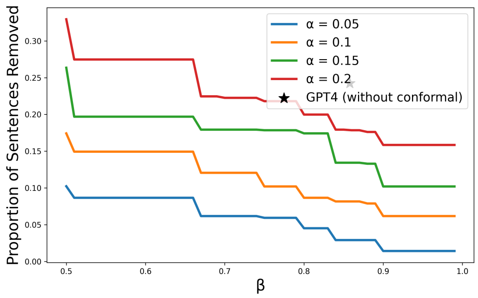


(g) Gemini 2.0 Flash-Lite

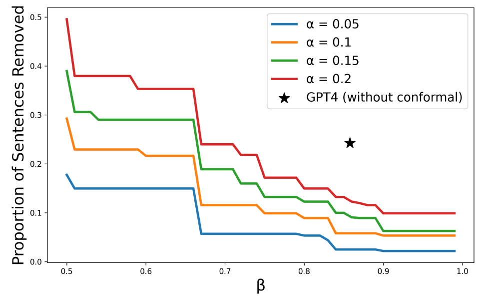


(h) Gemini 2.5 Flash

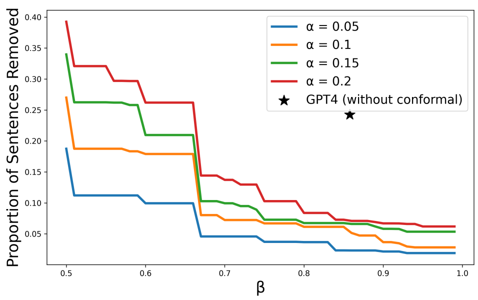
Figure 19: Target recall β vs. proportion of sentences removed (conciseness). Lines indicate different values for the target error rate α on TLDR-AIC.



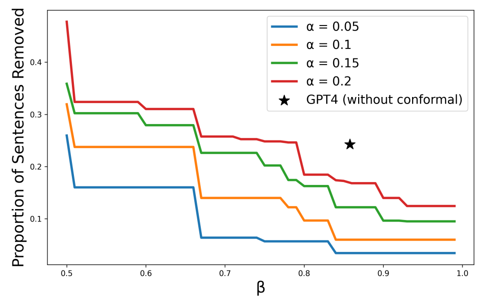
(a) Cosine Similarity Centrality



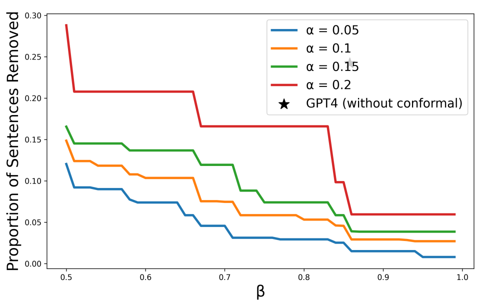
(b) Sentence Centrality



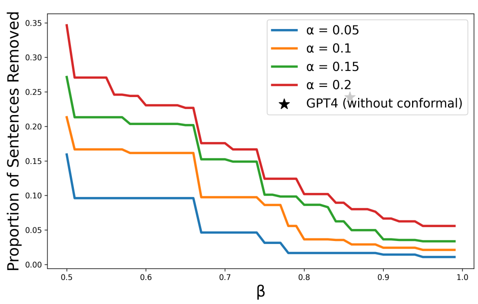
(c) GUSUM



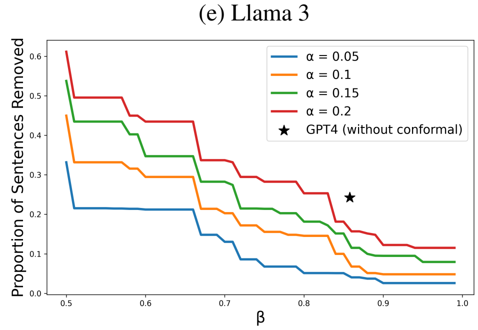
(d) GPT-4o mini



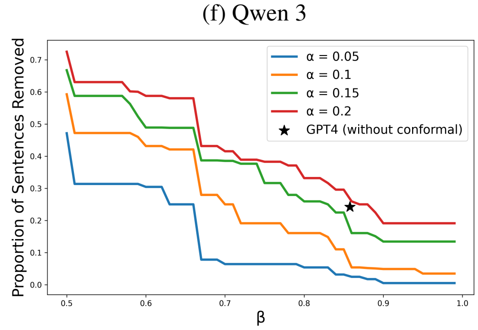
(e) Llama 3



(f) Qwen 3

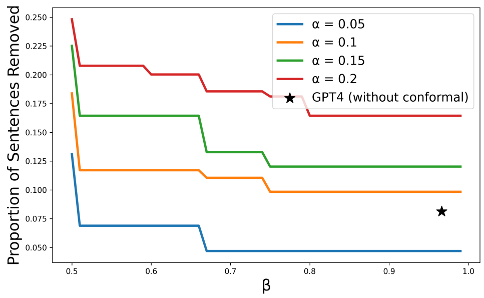


(g) Gemini 2.0 Flash-Lite

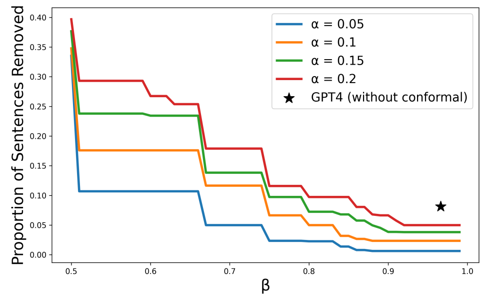


(h) Gemini 2.5 Flash

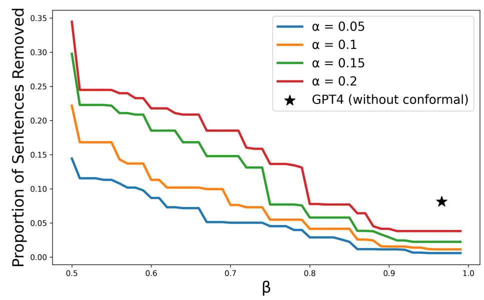
Figure 20: Target recall β vs. proportion of sentences removed (conciseness). Lines indicate different values for the target error rate α on ECT.



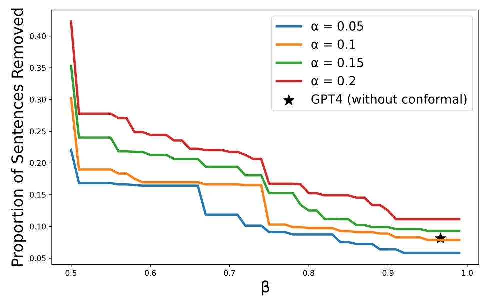
(a) Cosine Similarity Centrality



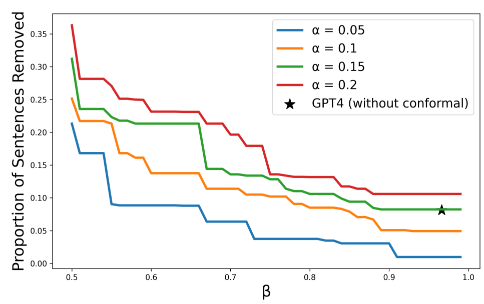
(b) Sentence Centrality



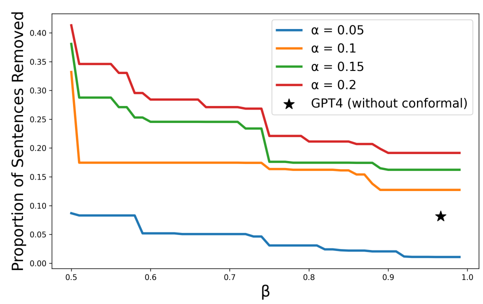
(c) GUSUM



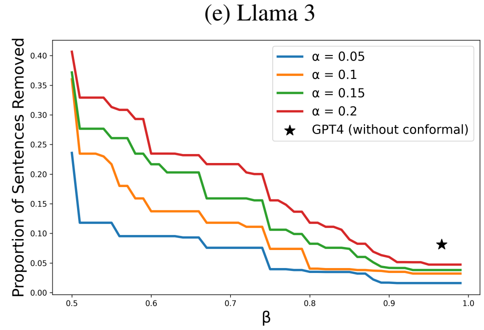
(d) GPT-4o mini



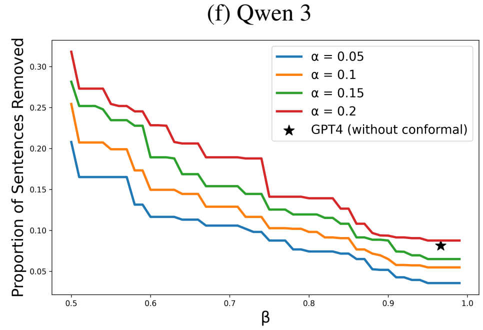
(e) Llama 3



(f) Qwen 3



(g) Gemini 2.0 Flash-Lite



(h) Gemini 2.5 Flash

Figure 21: Target recall β vs. proportion of sentences removed (conciseness). Lines indicate different values for the target error rate α on MTS.

Table 3: Performance comparison of importance scoring methods, measured in AUPRC of claim rankings compared to ROUGE-1/2/L based ground truth labels. Higher is better.

Importance Score	ROUGE-1		ROUGE-2		ROUGE-L	
	CNN/DM	TLDR-AIC	CNN/DM	TLDR-AIC	CNN/DM	TLDR-AIC
Dataset Positive Rate	0.10	0.06	0.10	0.06	0.10	0.06
Cos. Sim. Centrality	0.13	0.30	0.27	0.26	0.31	0.29
Sentence Centrality	0.28	0.27	0.24	0.24	0.27	0.27
GUSUM	0.31	0.19	0.30	0.18	0.30	0.19
LexRank	0.32	0.30	0.28	0.26	0.32	0.28
GPT-4o mini (binary)	0.13	0.08	0.13	0.08	0.13	0.08
GPT-4o mini	0.38	0.37	0.34	0.32	0.37	0.33
Llama3-8B	0.23	0.18	0.21	0.16	0.24	0.19
Qwen3-8B	0.23	0.20	0.22	0.17	0.23	0.20
Gemini 2.0 Flash-Lite	**0.42**	**0.40**	**0.38**	**0.39**	**0.40**	**0.38**
Gemini 2.5 Flash	0.35	0.36	0.32	0.34	0.34	0.36

Table 4: Performance comparison of importance scoring methods, measured in conciseness of summaries (proportion of sentences removed) under Conformal Importance Summarization compared to ROUGE-1/2/L based ground truth labels. Higher is better.

Importance Score	ROUGE-1		ROUGE-2		ROUGE-L	
	CNN/DM	TLDR-AIC	CNN/DM	TLDR-AIC	CNN/DM	TLDR-AIC
Original Article	0.00	0.00	0.00	0.00	0.00	0.00
Cos. Sim. Centrality	0.22	0.24	0.23	0.12	0.21	0.22
Sentence Centrality	0.22	0.20	0.25	0.22	0.19	0.22
GUSUM	0.21	0.07	0.32	0.10	0.18	0.07
LexRank	0.17	0.26	0.17	0.29	0.16	0.25
GPT-4o mini (binary)	0.26	0.22	0.26	0.22	0.26	0.22
GPT-4o mini	**0.32**	0.25	**0.33**	0.28	**0.29**	0.21
Llama3-8B	0.25	0.27	0.31	**0.37**	0.27	0.27
Qwen3-8B	0.07	0.06	0.15	0.08	0.11	0.08
Gemini 2.0 Flash-Lite	0.11	0.09	0.09	0.10	0.09	0.11
Gemini 2.5 Flash	0.22	**0.28**	0.30	0.31	0.22	**0.29**

C.4 ROUGE Score-based Ground Truth Performance

Table 3 and 4 respectively show the AUPRC and conciseness of summaries when we test our methods using labels generated from ROUGE scores, rather than cosine similarity, using Algorithm 1. Since we only use this algorithm for CNN/DM and SciTLDR, we only display results for CNN/DM and TLDR-AIC.

The results are similar to using cosine similarity: Gemini 2.0 Flash-Lite once again performs best in terms of AUPRC, and Gemini 2.5 Flash still performs very well in terms of sentence reduction length. The best numerical values for AUPRC and conciseness on each dataset are also comparable to the cosine similarity-based ground truth from Table 2.

C.5 Ablation over Calibration Set Size

Throughout the paper, we used a fixed calibration set size of $n = 100$ samples to demonstrate that the method can operate with very little labeled data. However, in some regimes the availability of labeled data can be extremely limited, so in this section we test our method with even fewer calibration datapoints.

The effect of calibration dataset size in conformal prediction is well understood theoretically; the coverage guarantee is famously "valid in finite samples", meaning that it holds statistically for any

Table 5: Ablation of empirical coverage over calibration dataset size n.

Target Coverage	Mean				Standard Deviation			
$1-\alpha$	$n=25$	$n=50$	$n=75$	$n=100$	$n=25$	$n=50$	$n=75$	$n=100$
0.60	0.61	0.61	0.60	0.61	0.09	0.07	0.05	0.05
0.65	0.65	0.67	0.66	0.65	0.09	0.07	0.06	0.05
0.70	0.73	0.71	0.71	0.70	0.09	0.06	0.05	0.05
0.75	0.77	0.77	0.75	0.75	0.09	0.06	0.05	0.05
0.80	0.81	0.81	0.80	0.80	0.08	0.05	0.05	0.04
0.85	0.89	0.86	0.86	0.85	0.06	0.05	0.04	0.04
0.90	0.92	0.90	0.91	0.90	0.05	0.04	0.03	0.03
0.95	0.96	0.96	0.96	0.95	0.04	0.03	0.02	0.02

Table 6: Ablation of summary conciseness (proportion of sentences removed) over calibration dataset size n. Results are taken over 20 random calibration/test splits.

n	Mean	Std
25	0.28	0.09
50	0.31	0.08
75	0.32	0.08
100	0.33	0.05

finite calibration dataset. In practice, n controls the variance of the coverage viewed as a random variable over the calibration data. For a textbook-style explanation of these details, see Section 3.2 of [2].

We match the experimental setting of Figure 2 which uses $\beta = 0.8$ and the Gemini 2.5 Flash scoring function to generate results on the ECTSum dataset, shown in Table 5. As guaranteed by Theorem 1, the empirical coverage is no less than $1-\alpha$ for all values of n. Lower values of n tend to overshoot the minimum coverage $1-\alpha$ by a bit more, because the upper bound of $1-\alpha+\frac{1}{n+1}$ becomes looser with n, but we still find all values within theoretical bounds (for example, $\frac{1}{25+1} \approx 0.04$, and $\frac{1}{50+1} \approx 0.02$). The primary reason to increase n is to reduce the variance of the empirical coverage so that it is less likely any given instantiation has lower than expected coverage.

Given that we find the coverage guarantee to be satisfied, we can also check the main metric of our method's performance: the conciseness of summaries it produces (as the proportion of sentences removed). In Table 6 we find little difference in the performance when using 50 to 100 samples, although variance is increased on smaller datasets. Overall, this ablation demonstrates that our method is applicable in the very low labeled data regime.

C.6 Direct Abstractive and Hybrid Extractive-Abstractive Comparison

Here we provide additional plots using the same settings as in Section 5.5. Figure 22 shows the comparison between extractive summarization with our conformal method, abstractive summarization with an LLM, and our hybrid proposal of applying abstractive summarization to our extractive summary, this time with Gemini 2.5 Flash used for both conformal scoring, and abstractive summarization. The results are highly comparable to Figure 6.

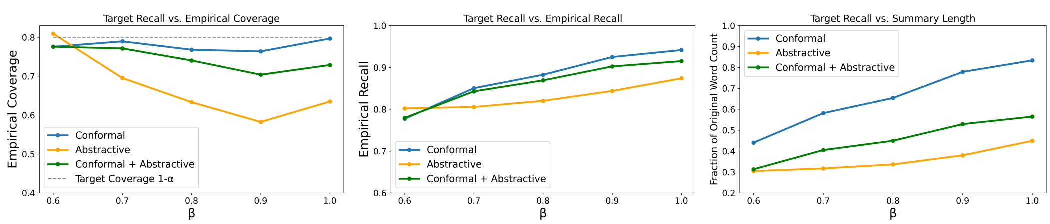


Figure 22: Comparison between extractive summarization with our method, abstractive summarization with an LLM, and our hybrid proposal on ECTSum. Here the target coverage is $1-\alpha=0.8$, the conformal approach uses Gemini 2.5 Flash scoring, and the abstractive model is also Gemini 2.5 Flash.