LCFO:
Long Context and Long Form Output Dataset and Benchmarking

Marta R. Costa-jussà, Pierre Andrews, Mariano Coria Meglioli, Joy Chen, Joe Chuang, David Dale, Christophe Ropers, Alexandre Mourachko, Eduardo Sánchez, Holger Schwenk, Tuan Tran, Arina Turkatenko, Carleigh Wood
FAIR, Meta
{costajussa,mortimer,mfcoria,joyqchen
joe.chuang,daviddale,chrisropers,alexmourachko
eduardosanchez,schwenk,tuantran,arinatur,carleighwood}@meta.com

Abstract

This paper presents the Long Context and Form Output (LCFO) benchmark, a novel evaluation framework for assessing gradual summarization and summary expansion capabilities across diverse domains. LCFO consists of long input documents (5k words average length), each of which comes with three summaries of different lengths (20%, 10%, and 5% of the input text), as well as approximately 15 questions and answers (QA) related to the input content. Notably, LCFO also provides alignments between specific QA pairs and corresponding summaries in 7 domains.

The primary motivation behind providing summaries of different lengths is to establish a controllable framework for generating long texts from shorter inputs, i.e. summary expansion. To establish an evaluation metric framework for summarization and summary expansion, we provide human evaluation scores for human-generated outputs, as well as results from various state-of-the-art large language models (LLMs).

GPT-4o-mini achieves best human scores among automatic systems in both summarization and summary expansion tasks (≈ +10% and +20%, respectively). It even surpasses human output quality in the case of short summaries (≈ +7%). Overall automatic metrics achieve low correlations with human evaluation scores (≈ 0.4) but moderate correlation on specific evaluation aspects such as fluency and attribution (≈ 0.6).

1 Introduction

Robust long text generation capabilities are required to meet user demand for extensive content creation, including story writing and essay composition (Xie and Riedl, 2024), which is why recent models such as GPT-4 (et al., 2024b) are expanding the output lengths from 4k tokens in GPT-4o to 64k in the latest versions.

However, evaluating the performance of Large Language Models (LLMs) on summarization and summary expansion tasks (see definitions in Section 2.1) is particularly challenging, especially when it comes to summarizing very long input documents and generating either long summaries or long summary expansions. Although there is a lot of work and interest in studying summarization evaluation (e.g. Zhang et al., 2024a), evaluation of long text outputs is an emerging area (Que et al., 2024). Indeed, long-context input processing tasks (such as summarization or comprehension question answering applied to long documents) and long-form output production both involve high cognitive loads for humans. This may be why evaluation work in these areas is not as mature as in others.

To complement the summarization task, we include an "inverted" task of *summary expansion*: generating a longform text based on its shorter sketch. It requires more creativity than summarization (which can be approached extractively) but imposes more constraints than open-ended language modeling. A practical application for this task could be with a writing assistant expanding the text sketch created by the human writer into a longer output that the human subsequently revises. Another potential use is "back-translation" for summarization, with a model training to summarize the synthetic expanded documents into their original short forms.

The LCFO dataset that we are presenting in this paper is an exclusively human annotated benchmark and challenge dataset involving natural language understanding and generation across multiple domains in the aforementioned tasks. Our dataset is carefully manually crafted with human revision from the selection of the documents to the final annotation, without relying on LLMs at any point. We provide detailed linguistic guidelines and abstractive QA. Furthermore, we provide another set of linguistic guidelines to evaluate the tasks of

Findings of the Association for Computational Linguistics: ACL 2025, pages 10672–10700
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

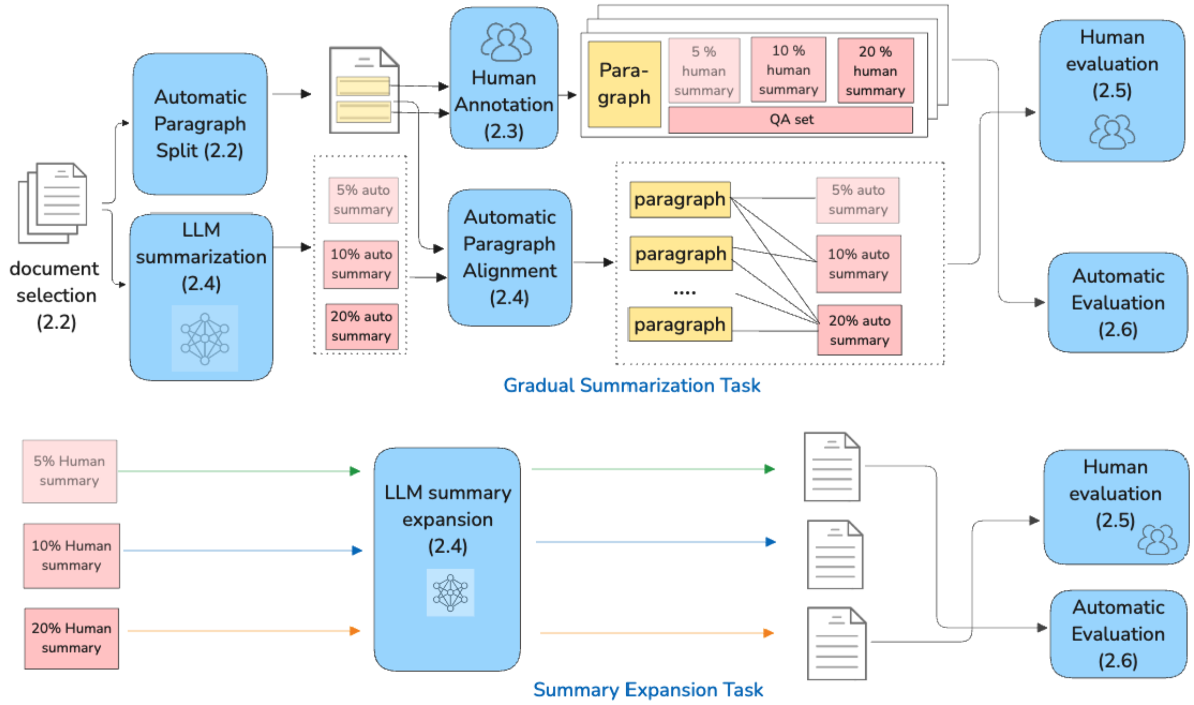

Figure 1: LCFO annotation and evaluation framework for the gradual summarization (top) and summary expansion (bottom) tasks. The resulting benchmark can also be used for other tasks such as reading comprehension and automatic evaluation metric development. In gradual summarization, long documents are split into paragraphs and annotated by humans to generate 3 level of summaries, plus a questions-answers set. Additionally, different LLMs were used to construct automated summaries at corresponding levels. In summary expansion, different human summaries are prompted to LLMs to generate the output targeting the same length as the source document. The components are described in different sections (marked in parentheses)

summarization and summary expansion.

Together, we present the LCFO benchmark. Given a source of long structured documents, we generate multiple long outputs and associated QA pairs, and we evaluate human and model outputs with human annotations. The main contributions of this work are the following (see Figure 1 for a schematic representation of the dataset and tasks):

- Dataset creation with structured inputs and alternative references; each input document is associated with 3 summaries of different lengths (20%, 10%, 5%). Gradual summarization is useful in that it provides both long and short summaries references. Moreover, the availability of summaries of various length is expected to improve the development of summary expansion approaches, which allows us to provide a more controllable summary expansion framework.
- A set of QA pairs for each input document aligned with each of the different length summaries. Our QA pairs are in free-form short-answer format (i.e. not multiple choice) and are of the abstractive type (i.e. they are not copies of parts of the source document). This QA can potentially be used to evaluate the model outputs, based on the appropriateness of the responses, similarly to previous proposals (Wang et al., 2022).
- Selection of automatic metrics. Most of these metrics can be used to evaluate at the paragraph level, and more widely can be used to evaluate summary expansion or long-form generation in multiple tasks and languages.
- Evaluation of several LLMs on our dataset both automatically and manually; and evaluation of automatic metrics on summarization and summary expansion.

2 The LCFO Creation Framework: Principles and Methods

This section describes the detailed principles and methods that underlie LCFO (Figure 1 (right)).

2.1 Definitions

Long context/form. We define long text as text that exceeds 5k words. For reference, the attention span for reading and taking notes has been considered 10 to 15 minutes since a 1978 seminal paper (Hartley and Davies, 1978). More recent research cited in a survey paper (Bradbury, 2016) shows that attention paid to lectures declines significantly after 20 minutes. The tasks we describe here are closer to note taking than lecture attendance; therefore, we should keep the 10–15 minute reference. If we base ourselves on the reported reading average time for first-language, secondary-education level readers (125 words per minute on average), we can consider that we will reach high cognitive load at around 1.5k words, and peak cognitive load at around 2.5k words. The quality of cognitive processing then starts to decrease significantly after.

Structured/hierarchical. The input and output documents are partitioned into sections and, if necessary, nested sub-sections. Their structure is determined by task- and domain-specific guidelines.

Gradual summarization (GS). The input is a long document (defined here as 5k words or more). The summarization task can be described as the act of generating a much shorter corresponding document that includes the essential information contained in the long document and the same logical structure linking the various pieces of essential information. The summarization task presented here consists in taking long documents as inputs and generating three corresponding summaries of length that represent 20%, 10%, or 5% of the input document.[1]

Summary expansion (SE). The input is a short and concise document that has similar properties to those of a summary (that is, it is mainly a standalone document that abstracts from details). The summary expansion task can be described as the act of generating a much longer document that preserves the essential elements found in the corresponding short document as well as the logical structure that connects such elements. More specifically, the task presented here consists in taking summaries as inputs and generating 3 long documents of different lengths. Each of the 3 lengths is set such that an input summary represents either 5%, 10%, or 20% of its respective expanded documents. As this is a more freely generative task, an additional requirement to be taken into consideration is that of coherence (for example, the detailed information included in one generated sentence should not contradict that included in another sentence). We prompt the model only with one single summary and specify the length of the output we want (5x, 10x...). We do not include in the prompt longer summaries or the whole document, since they may be a possible output.

2.2 Data selection and preprocessing

Selection. We select input documents that cover different domains, which meet the desired average length and the structure requirement.

- We cover 7 domains including politics, news, Wikipedia, scientific, literature, conversational, and legal documents, with the format of documents and conversations. We source from 10 datasets: LexGLUE (Chalkidis et al., 2022), BookSum (Kryscinski et al., 2022), SQuality (Wang et al., 2022), FacetSum (Meng et al., 2021) JRC-Acquis (Steinberger et al., 2006), MultiUN, Wikipedia, GovReport (Huang et al., 2021), Summscreen (Chen et al., 2022), and Seahorse (Clark et al., 2023). The correspondence between the source data sets and the domains is shown in Table 1.
- The source documents are selected to be on average 5k words / document. We prefer documents with a hierarchical structure and containing relatively few numbers[2], which are better suited for summarization and summary expansion tasks.
- We prioritize recent documents when the domains allow (e.g., Wikipedia, where articles have a significant amount of new information since 2024). We preprocess documents in structured domains to provide a flattened structure while keeping hierarchical markers that are readable to annotators and models. It also ensures a consistent format across datasets.
- We filter out documents that contain toxicity using the ETOX package (Costa-jussà et al., 2023) and add manual verification to ensure the high integrity of the selected documents.

[1] We did not explicitly instruct the human annotators to generate the shorter summaries hierarchically, by compressing the longer ones, but this is what they apparently did most of the time. Hence the term “gradual”: our data allows gradually interpolating between various levels of details in a text.

[2] We de-prioritize documents with many numbers (such as Wikipedia pages with large tables of various statistics), because our focus is less on structured data and more on natural language, and also because comparing large sets of numbers in the source and in the summary would be a difficult task for human annotators.

Preprocessing. To reduce the cognitive load for human annotation, we split paragraphs automatically (APS). Details on this paragraph splitting differ from corpus to corpus and are reported in the Appendix A.

2.3 Human summary and QA-pair generation

To obtain human-written summaries of long-form texts, detailed guidelines are developed (Appendix E). All summary writers must be native English speakers and have writing or editing experience.

These writers receive 252 long form documents (each around 5k words), and they are asked to read each document in its entirety and write three summaries for each document: the first summary representing around 20% of the length of the source document; the second and the third summaries representing further summarization — around 10% and 5% of the source document, respectively. When writing these summaries, the writers are tasked with compressing and retaining all the core ideas of the source. Giving a definition of a "core" or "main" idea of a text presented one of the challenges of our work with the writers. Each summary is supposed to be a cohesive standalone text that could be read and understood on its own.

The fact that the source documents represent different domains poses another challenge for the writers: they need to possess some knowledge and expertise in each of these. The documents are split into sections and paragraphs, and the writers are asked to keep the flow of the section/paragraph structure of the source text, trying to summarize it from top to bottom. However, we emphasize that the sentence by sentence summarization is exactly what we do not need, and the summaries need to be abstractive rather than extractive, which means copying the source text is strongly discouraged.

In addition, the writers are asked to provide a set of questions and answers for each long document. They need to compose at least 13–15 questions per 5,000 words. The answers are supposed to cover the points reflected in the summaries. We instruct the writers to produce open-ended, complex questions, which provide a good baseline for testing reasoning. For tracking and alignment purposes, each paragraph in each source document is given a number. The writers are asked to specify in which paragraph each answer can be found. They are also asked to indicate which of the summaries provides the answer to each question.

Besides the general guidance, we discovered that working with conversational content needed additional clarifications, so we prepared an additional document for working specifically with long-form text that contains conversations (such as plays or screenplays).

2.4 Automatic output and postprocessing

We want to understand how current state-of-the-art models perform on our new benchmark, both on the capability of comprehension a very long context and on the generation of long outputs. We conduct the automatic abstractive summarization for the former and summary expansion for the latter. We give details on the tasks below.

Gradual Summarization. We prompt the models with the human guidelines with a slight adaptation to be LLM friendly. We input the entire document without paragraph splitting. To give a fair evaluation, we prompt all LLMs in the zero-shot setting. To control the length of the LLM output, we have added additional instructions with the upper and lower bounds of the permissible words. For example, to ask the model to generate a summary of the R% length of the source text, the prompt contains `"Make sure the summary has {y} words or less.....Please write at least {x} words""""` , where `x` and `y` are determined per document with respect to the length and ratio `R`. In practice, we see that enforcing the length of the document right before and after the content block in the prompt gives consistent results. We give details of our summary prompts in the Appendix I.

Summary expansion. We customize prompts for each domain, plus the model-specific prompt templates. Similarly to the summarization task, the prompt contains instructions on the desired range of the generated text length. In addition, each prompt has instructions to guide the model in generating content of a certain quality (consistency, coherence, and keeping the main ideas in the summary). We prompt the model with specific formats for different domains as reported in the appendix I.

Automatic Paragraph Alignment (APA). We add this step to help human annotators evaluate the outputs that we are creating (as detailed in the next section 2.5). The task of comparing long inputs and outputs creates a high cognitive load on human evaluators. To reduce it, we provide an approximate alignment between the input paragraphs and the segments of the output, taking advantage of the assumption that a summary usually follows the structure of the source document. First, we use dynamic programming to find a monotonic alignment path between input and output sentences that would maximize the sum of cosine similarities of the SONAR embeddings (Duquenne et al., 2023) of the two sentences. An output sentence could be aligned with multiple consecutive input sentences, potentially from different paragraphs, but we assign it to a single input paragraph with which it is aligned the most frequently. Thus, each input paragraph gets aligned to a contiguous output segment (potentially empty) in a monotonic way. This alignment helps the annotators navigate the input and output documents jointly.

2.5 Human Evaluation

To perform human evaluation on previously generated output, we design human evaluation guidelines inspired by previous works (Clark et al., 2023; Krishna et al., 2023; Que et al., 2024) and fully reported in Appendices G and H.

Human evaluation on gradual summaries. Before starting the evaluation, annotators are allowed to reject a task if the output text is gibberish or obviously of low quality.

The generated summaries are evaluated in two tasks. In Task 1, the annotators first read the source document and the three summaries and then rate the generated text in four aspects, including attribution, coverage of the main ideas, conciseness and readability (similar to the 'checklist' in HelloBench by Que et al. (2024)). The annotators rate the summary on a 0-4 Likert scale and finally give an overall rating on a 0-10 Likert scale. Each summary receives its own separate set of scores.

In Task 2, the annotators validate the QA sets that were previously created by human writers. For each question in the QA sets (13–15 questions and answers), the annotators are required to determine whether the content of the summary contains enough information to answer the question (i.e., the answer is directly stated, heavily implied or logically entailed in the summary). The annotators give a YES or NO to each QA pair. For each summary, the annotators validate the whole set of QA once.

The whole evaluation is referenceless, which means that the human written summaries are not shown to annotators, and that they only see a single set of summaries from one anonymous model output each time.

Both tasks 1 and 2 involve human judgment, and to reduce the bias, 3 sets of rating from random annotators are required for each generated output. The same guidance should be used for all different domains. Detailed evaluation guidelines are included in the Appendix G.

Human evaluation on summary expansion. We use the same format as the previous summarization evaluation tasks and integrate some of the questions from Story Plot Generator (Zhu et al., 2023) and HelloBench (Que et al., 2024). For task 1, the annotators read the source summary and the generated long-form output, rate the output on 6 aspects, including the coverage of main core ideas, cohesion, richness in details, creativity, non-repetitiveness, and interest, and give an overall rating at the end. In task 2, they validate the QA set with the generated long-form text. Each output is evaluated separately without reference, and three sets of random annotation ratings are required. Detailed evaluation guidelines are included in the Appendix H.

Evaluation statistics. Summaries and summary expansions are each evaluated separately. For the evaluation of generated summaries, 252 documents from all domains are used as the source to generate the summaries (with 2 documents being excluded during the process). The summaries are generated using three different models (as reported in Section 3 and chosen to represent close and open models of different sizes): GPT-4o-mini-64k (et al., 2024b), LLAMA 3.1-70B, and LLAMA 3.1-8B (et al., 2024a). This results in 756 outputs and, along with 252 sets of human-written summaries, creates a dataset of 1,000 document-summary pairs for evaluation. A vendor sources 287 annotators, who are required (1) to be native speakers of English and (2) to hold a language-related degree. These annotators are selected from a pool that is different from that of the summary writers, ensuring that they have no prior knowledge of the source documents or the written summaries. Tasks are randomly assigned to annotators until every set of generated output receives three complete annotations. A limit of 10 evaluations for each model is set per annotator to mitigate biases in the results.

For evaluation of generated summary expansions, only a subset of data is selected,[3] including

[3] We exclude domains with high density of factual information, because we believe they are less appropriate for the summary expansion task. The problem here is potential hallucination of factual information (e.g. *Berlin is the capital of*

SummScreen, BookSum, SQuality and FacetSum (102 source documents in total). The expansions are generated with the same models as previously (GPT-4o-mini, LLaMA 3.1-70B, and LLaMA 3.1-8B), resulting in 306 long-form outputs. Ten experienced data analysts are selected to conduct the evaluation. Similarly to the evaluation of summaries, the tasks are assigned randomly until every long form output receives 3 complete annotations.

2.6 Automatic evaluation

The summarization outputs are typically evaluated by computing ROUGE scores (Lin, 2004) with respect to a reference. However, this approach is not sufficient for at least three reasons (Schluter, 2017): it depends too much on the reference, it offers only a comparison at the surface level, and it does not explain why a summary is good or bad. Thus, we compute several other reference-free metrics, each targeting a specific aspect of summarization quality (over the 6 aspects introduced in the SEAHORSE dataset by Clark et al. (2023)). For each aspect, we tried to pick a metric that is focused on this specific aspect, preferring ones that are computationally transparent and that have already been implemented elsewhere:

1. **Repetitiveness**: how much the summary repeats the same phrases. We report the count of all the word n-grams ($n \in \{1, 2, 3\}$) in the summary, divided by the count of such unique n-grams (REP-3) (Welleck et al., 2019).
2. **Fluency**: how grammatical the text is. We report the average probability of a summary sentence being grammatical (or linguistically acceptable in the Chomskyan sense of the term) computed with a CoLA classifier (Krishna et al., 2020).
3. **Coherence**: how similar are the sentences in the generated texts to each other. COH-2 averages similarity of the neighboring-over-one sentences in the embedding space (Parola et al., 2023).
4. **Attribution**: how much of the summary is directly attributable to the source (something like "precision" of the ideas in the summary). The average score of SEAHORSE Q4 (SH-4) model evaluates attribution (Clark et al., 2023).
5. **Coverage of the source**: how much of the source is reflected in the summary. The average SEAHORSE Q5 (SH-5) score reports this aspect (Clark et al., 2023).
6. **Overall** In order to evaluate the overall quality of the text, we use two metrics:
(1) For summarization, we aggregate a score (AVG) from the above metrics, namely, averaging $-REP\text{-}3, CoLA, COH\text{-}2, SH\text{-}4, SH\text{-}5$.
For summary expansion, the aggregated score (AVG) is the average score of $REP\text{-}3, CoLA, COH\text{-}2$. Note that REP-3 is negated to make the score monotonic. Also, for the summary expansion, REP-3 increases the value over the length of output, so the factor 0.2 is empirically set to normalize the value on the summary expansion task 20%.
(2) We use HelloEval (HE) score (Que et al., 2024): an LLM-as-judge model with various checklists trained w.r.t. human evaluation.

Selection of these metrics from a larger candidate set was partially motivated by their correlations with human annotations from Clark et al. (2023), described in Appendix D. Table 5 in Appendix C summarizes the list of metrics.

For the SEAHORSE scores, we had to feed the whole source text to a transformer model, which was neither feasible computationally with long context inputs nor made sense given the relatively short-form training data of those models. To bypass this problem, we segment sources and summaries into aligned fragments (using a modification of the alignment algorithm in Section 2.4) with at most 50 sentences on the source side and compute model-based metrics for the fragment pairs.

2.7 Data Statistics

LCFO covers 7 domains sourced from 10 datasets with an average document length of 5k words. Table 1 contains the distribution of the LCFO dataset in subsets and domains, as well as the average word length of the documents. More details are reported in the Appendix B.

3 Experiments

Settings. We experimented with closed and open LLMs. We chose GPT-4o-mini-64k[4] for the closed model and LLaMA 3.1-70B (Dubey et al., 2024) for the open-source one. For summarization, we

France), which can be detrimental for real-life use cases but is out of scope of our evaluation that does not check for extrinsic factuality but only for faithfulness to the source text.

[4] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

DATASET	N	DOMAIN	LEN
LexGLUE	25	Legal: supreme court opinions	4953
BookSum	27	Litertyre: books, novel, act	4114
SQuality	25	Literature: stories	4856
FacetSum	25	Scientific: journal articles on various domains	4904
JRC-Acquis	25	Legal: legislative text of the European Union	4825
MultiUN	25	Political: UN docs	4539
Wikipedia	25	Wikipedia: 22 docs on biomedicine	5266
GovReport	25	Political: Congressional Research Service and US Government Accountability Office	5078
Summscreen	25	Conversational: TV series transcript	5030
Seahorse	25	News: English BBC news	4576
Total	252	Average word count	4814

Table 1: LCFO Summary: domains and statistics (number of documents N and average length in words LEN)

ran the model with all length ratios (5%, 10%, 20%), while for summary expansion, we only expanded the summaries 20% to the full document. We also performed a postprocessing step to filter the templated response such as "`**Summary**`", "`Here is the summary:`", etc.

Summarization results. Table 2 shows the general results of the selected models at different levels of gradual summarization. Results broken down by domains are reported in the Appendix J. Note that LLMs tend to perform similarly regardless of the length of the output in terms of human scores. This is not the case for humans that show to lag behind when performing short summaries. The best results are consistently achieved with GPT-4o-mini and are consistent with previous research findings (Que et al., 2024). This model even surpasses human-level quality in short summaries. This may be explained by humans tending to perform worse when summarizing short documents and better when summarizing long ones.

Summary expansion results. Table 3 shows the overall results on the summary expansion task by a factor of 5, giving the 20% summary input. The performance of models is not coherent across metrics that look only at the output (i.e.%WC, REP-3, CoLA, COH-2, and AVG). In terms of HE and coherently with the human evaluation results, GPT-4o-mini is the best performing model. Additionally, we report the results of other combinations, for example, expanding by larger factors (10 and 20) giving 10% summary input and giving 5% summary input, respectively, in Appendix J.

When comparing across tasks (i.e. summarizing to 20% or doing summary expansion from 20% by a factor of 5), the results show better performance in the former. It is expected that summary expansion is a more challenging task across domains and all models. Current models struggle with this task. If we compare HE, the deltas in the same model vary from 6% for GPT-4o-mini to ≈ 30% for LLaMA 3.1-70B. When comparing output-based metrics, there are discrepancies in conclusions (i.e., LLaMA 3.1-8B better than 70B model). However, HE is still worse for the 8B model. This may indicate that selected output-based quality metrics are less reliable than the HE score (see the analysis below for metrics evaluation).

Metrics evaluation. In our study, we consider human evaluation, conducted according to the guidelines outlined in Appendices G and H, as the definitive measure of the overall quality score, as well as the scores for individual quality aspects such as coverage and attribution. To mitigate potential biases among the annotators, we calculate the average of three annotations for each task. Table 4 presents the Spearman correlation coefficients for various aspects and overall scores, comparing automatic metrics and human evaluations for summarization and summary expansion, respectively. A higher Spearman correlation coefficient signifies a stronger correlation between the automatic metrics and human annotation. The metric that shows the highest correlation with human annotation corresponds to SH-4, which measures attribution. When comparing metrics that measure overall performance, we observe that R-L is not very good at correlation, but it may also be due to the fact that our task is not the best suited to use human references. HE is the one with the highest correlation. AVG low-correlation (both in S and SE) may be explained by the fact that individual averaged metrics are not very good or they cover more specific aspects which may not end capturing the overall performance. This low correlation for R-L and AVG can explain the discrepancy observed in the model ranking (specially between LLaMA 3.1-70B and 8B in Tables 2 and 3. Correlations are low in all cases, which shows the difficulty of the evaluation. Beyond the challenge of automatizing it, we should add the fact that humans struggle in generating short summaries, which may imply that humans also struggle in evaluating them.

Output	R-L(↑)	REP-3(↓)	CoLA↑	COH-2↑	SH-4↑	SH-5↑	AVG↑	HE↑	Hum↑
LCFO.5%									
Human	n/a	**0.308**	0.941	0.809	0.644	0.387	0.494	52.195	6.61
GPT-4o-mini	0.331	0.328	0.968	0.719	**0.635**	**0.487**	**0.496**	**76.917**	**7.25**
LLAMA 3.1-70B	0.384	0.383	0.965	0.861	0.622	0.377	0.488	72.468	6.27
LLAMA 3.1-8B	0.377	0.411	**0.969**	**0.865**	0.618	0.372	0.482	63.894	6.32
LCFO.10%									
Human	n/a	**0.395**	0.945	0.816	**0.661**	0.416	**0.489**	64.688	7.44
GPT-4o-mini	0.385	0.404	**0.964**	0.695	0.621	**0.471**	0.469	**77.863**	**7.50**
LLAMA 3.1-70B	0.434	0.515	0.944	**0.860**	0.614	0.369	0.454	72.497	6.42
LLAMA 3.1-8B	0.429	0.534	0.963	0.858	0.612	0.366	0.453	59.385	6.63
LCFO.20%									
Human	n/a	**0.244**	0.938	0.805	0.615	0.357	**0.494**	69.745	**7.78**
GPT-4o-mini	0.445	0.497	**0.961**	0.673	**0.616**	**0.464**	0.443	**76.706**	7.52
LLAMA 3.1-70B	0.467	0.631	0.928	0.860	0.596	0.357	0.422	71.603	6.32
LLAMA 3.1-8B	0.469	0.647	0.956	**0.861**	0.594	0.370	0.427	51.015	6.60

Table 2: Performance on the summarization task

Output	%WC	REP-3(↓)	CoLA↑	COH-2↑	AVG↑	HE↑	Hum↑
GPT-4o-mini	**1.931**	0.707	**0.913**	0.609	0.460	**70.896**	**6.431**
LLAMA 3.1-70B	1.058	**0.680**	0.877	0.750	0.497	39.199	4.469
LLAMA 3.1-8B	1.187	0.809	0.903	**0.779**	**0.507**	38.416	4.801

Table 3: Performance on the summary expansion task by a factor of 5, giving the 20% summary input.

	R-L	CoLA	SH-4	SH-5	AVG	HE
S	0.196 (0.065)	0.595 (6.337e-10)	0.616 (1.005e-10)	0.445 (1.105e-5)	0.159 (0.135)	0.428 (2.591e-05)
SE	n/a	n/a	n/a	n/a	0.285 (3.646e-05)	0.405 (1.957e-09)

Table 4: Spearman correlation coefficients (and p-value) for various aspects and overall scores between automatic metrics and human evaluation for summarization (S) and summary expansion (SE). For the former, we show correlations between CoLA and Human evaluation Q2d; SH-4 and Human evaluation Q2a; SH-5 and Human evaluation Q2b and R-L/AVG/HE and Human evaluation Q3 (appendix G). For the latter, we show correlations between AVG/HE and Human evaluation Q3 (appendix H).

4 Related Work

Related work on long context and long form output comes in many flavors. We cover a summary on long context and long-form output datasets.

Long-context datasets. Infinite length datasets such as NIAH, RULER (Hsieh et al., 2024) work with distracting information. Finite-length nondistractive-based datasets include: Multi-LexSum (Shen et al., 2022) is a collection of 9,280 expert-authored summaries drawn from single domain (Civil Rights Litigation Clearinghouse) writing the length of the source documents often exceeds two hundred pages per case and summaries are presented in two-sizes: a short and longer version; Longbench (Bai et al., 2024) and Marathon (Zhang et al., 2024b) that includes tasks with 5–25k context and, more recently, (Kwan et al., 2024) build a dataset up to 8k tokens context length to evaluate LLMs' long-context understanding across five key abilities: understanding of single or multiple relevant spans in long contexts based on explicit or semantic hints, and global context understanding. Loong is a multi-document QA dataset up to 200k context to assess RAG abilities. HelloBench (Que et al., 2024) includes summarization of a selection of long-input documents (3k to 6k word length).

Long-form output datasets. There is a lack of reference-based datasets on long form output. However, there are datasets that study prompting of different long-form generation; e.g., StoryGen (Zhu et al., 2023) includes prompts to generate stories, and HelloBench is one of the most diverse long form generation benchmarks including stories, screenplays, keyword writing.

Our contribution on datasets involve the manual collection of 3-length summaries from long input documents. This collection also includes abstractive QA (non-multiple choice) to test comprehension. Our contribution on metrics involves new human evaluation protocols on summarization and summary expansion, as well as annotations to develop supervised metrics on long-form outputs.

5 Conclusions

LCFO provides gradual summaries references from 5k input documents with QA pairs for each of the documents and summaries. Additionally, we provide human evaluation of human and model-generated summaries and model-generated summary expansions. Overall, LCFO enables the evaluation of several tasks and metrics in the setting of long-context input documents and long-form output. While the main contribution of this paper is to present the freely available LCFO dataset[5], we also evaluate model and human outputs, showing that LLMs are capable of surpassing human results when producing short summaries. Current evaluation results question the usefulness of manually generating human references for short summarization of long documents. To confirm this, as further work, we plan to exploit the capabilities of LCFO by using QA as part of automatic evaluation (i.e. scoring how many questions are correctly answered in model-generated summaries).

Limitations and Ethical considerations

Data contamination. Source documents may exist in the training data of the models, therefore, generation may be at risk. To mitigate this, we prioritize recent documents, since this is not enough, we annotate the correspondence of sections in summarization versions, so that we can generate only portions of the document. Therefore, if the model uses internal knowledge, we can quantify by spotting details from other sections.

Experiments. The experimental options that LCFO offers are much larger than the ones we explore in this paper. Also, the dataset can be easily expanded to have more summary references by matching with existing summaries in some of the domains. QA pairs have not been used in the paper but this is designed (but not limited) to serve for doing reading comprehension and/or for creating an evaluation metric.

Metrics. Summarization is a generative task with very diverse aspects of quality, and no single automatic evaluation metric captures them all adequately. To compensate for this, we report multiple evaluation metrics, but still, some of them are not well established; for example, there is no single metric of longform text coherence that the summarization community agrees upon. By providing the results of the human evaluation, we hope to help the community develop and validate better automatic evaluation metrics in the future.

Computing In terms of computing, evaluating LLMs on LCFO benchmark require larger memory due to both its big context size and the long-form output (should the models be capable to it). In case of LLAMA, we used 1 NVIDIA GPU A100 80 GB for the 8B model, and 8 GPUs for the 70B model. The resource was shared with the loading of scoring models (SH, CoLA) as well. Each evaluation run over 10 domains takes 90 minutes, including the computation of all the scores except HelloEval (where the computing time depends on the external availability of GPT-4 endpoint deployment).

Annotations Annotators were paid a fair rate. Each of the annotators signed a consent form agreeing on the dataset and its usage that they were participating in.

Acknowledgements

This paper is part of the LCM project[6] and authors would like to thank the entire LCM team for the fruitful discussions.

References

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

[5] Available at https://huggingface.co/datasets/facebook/LCFO

[6] https://github.com/facebookresearch/large_concept_models

N. Bradbury. 2016. Attention span during lectures: 8 seconds, 10 minutes, or more? *Advances in Physiology Education*, 40(4):509–513.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. Association for Computational Linguistics.

Marta Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023. Toxicity in multilingual machine translation at scale. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9570–9586, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations. *arXiv preprint*.

Abhimanyu Dubey et al. 2024a. The LLaMA 3 herd of models. *Preprint*, arXiv:2407.21783.

OpenAI et al. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

James Hartley and Ivor K. Davies. 1978. Note-taking: A critical review. *Programmed learning and educational technology*, 15(3):207–224.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *Preprint*, arXiv:2404.06654.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Empirical Methods in Natural Language Processing*.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wai-Chung Kwan, Xingshan Zeng, Yufei Wang, Yusen Sun, Liangyou Li, Yuxin Jiang, Lifeng Shang, Qun Liu, and Kam-Fai Wong. 2024. M4LE: A multi-ability multi-range multi-task multi-domain long-context evaluation benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15568–15592, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1080–1089, Online. Association for Computational Linguistics.

Alberto Parola, Jessica Mary Lin, Arndis Simonsen, Vibeke Bliksted, Yuan Zhou, Huiling Wang, Lana Inoue, Katja Koelkebeck, and Riccardo Fusaroli. 2023. Speech disturbances in schizophrenia: Assessing cross-linguistic generalizability of nlp automated measures of coherence. *Schizophrenia Research*, 259:59–70.

Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, Junran Peng, Zhaoxiang Zhang, Songyang Zhang, and Kai Chen. 2024. Hellobench: Evaluating long text generation capabilities of large language models. *Preprint*, arXiv:2409.16191.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multilexsum: real-world summaries of civil rights lawsuits at multiple granularities. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. 2022. SQuALITY: Building a long-document summarization dataset the hard way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1156, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

Kaige Xie and Mark Riedl. 2024. Creating suspenseful stories: Iterative planning with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2391–2407, St. Julian's, Malta. Association for Computational Linguistics.

Haopeng Zhang, Philip S. Yu, and Jiawei Zhang. 2024a. A systematic survey of text summarization: From statistical methods to large language models. *Preprint*, arXiv:2406.11289.

Lei Zhang, Yunshui Li, Ziqiang Liu, Jiaxi Yang, Junhao Liu, Longze Chen, Run Luo, and Min Yang. 2024b. Marathon: A race through the realm of long context with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5201–5217, Bangkok, Thailand. Association for Computational Linguistics.

Hanlin Zhu, Andrew Cohen, Danqing Wang, Kevin Yang, Xiaomeng Yang, Jiantao Jiao, and Yuandong Tian. 2023. End-to-end story plot generator. *Preprint*, arXiv:2310.08796.

A Automatic Paragraph Splitting Details

As part of the guidelines for summarization, annotators are instructed to read long documents from different domains and mentally distill key points into paragraphs and then into a cohesive document summary. This process requires the source text to be logically segmented into well-structured paragraphs that facilitate comprehension and synthesis.

During pilot studies, it became evident that the quality of the initial paragraph segmentation significantly impacted annotation outcomes. Poorly segmented paragraphs increased cognitive load and risked misinterpretation, while cohesive and logically structured paragraphs improved annotation consistency and efficiency.

Given the variability in formats and structures across datasets, a uniform approach to paragraph splitting was not feasible. Some datasets provided explicit structural markers (e.g., new lines, section headers), while others required more algorithmic intervention, such as employing the Segment Anything Text (SaT-l3) model. Furthermore, the SaT-l3 model's performance varied across text types, necessitating dataset-specific thresholds and post-processing techniques to optimize paragraph segmentation.

This section outlines the tailored methodologies applied to each dataset in the LCFO corpus, highlighting how their unique characteristics were addressed to produce high-quality, preprocessed documents for annotation.

LexGLUE Paragraphs were split using double newlines as separators, preserving the inherent paragraph structure in the dataset.

BookSum and SQualITY Lines were joined with a blank space to create continuous text blocks. Sentences and paragraphs were split using the SaT-l3 model with a threshold of 0.8, producing lists of sentences grouped into paragraphs. Paragraphs exceeding 3,000 characters were further split using the SaT-l3 model with a stricter threshold of 0.4. Consecutive short paragraphs (fewer than 2 sentences or under 400 characters) were merged to ensure coherence, especially for dialogue-heavy sections.

JRC-Acquis Lines were joined with a blank space to preserve the flow of text. Sentences and paragraphs were split using the SaT-l3 model with a standard threshold of 0.5. Consecutive paragraphs containing fewer than 2 sentences were merged. Sections and subsections were extracted from paragraph beginnings using the dataset's consistent numbered format (e.g., 1.1.2), serving as structural indicators.

MultiUN Lines were joined using blank spaces to form initial text blocks. Sentences and paragraphs were split using the SaT-l3 model with a threshold of 0.5. Short consecutive paragraphs (fewer than 2 sentences each, and up to 20 sentences total) were merged to improve readability and flow.

Wikipedia Original paragraphs were identified using empty lines (meaning double newline in the original text), which appeared as blank lines or in the CSV format. Long paragraphs (over 500 tokens) were split further using the SaT-l3 model with a threshold of 0.5 to improve segmentation accuracy for longer text units.

GovReport Same as LexGLUE, paragraphs were split using double newlines as separators.

Summscreen Initial paragraph segmentation was based on scene indicators ([SCENE-BREAK]) in the transcripts. However, this often resulted in excessively long paragraphs, with some documents containing only one or two paragraphs. Text formatting issues, such as double spaces in punctuation (e.g., " . "), were corrected to align with the SaT-l3 model's sensitivity. Long paragraphs exceeding 3,000 characters were re-segmented using the SaT-l3 model with a threshold of 0.9. Short consecutive paragraphs containing only one sentence were merged to form cohesive segments.

B Data Details

Our data collection

- 100% human annotated (no LLM preselection)
- 7 domains (political, wikipedia, scientific, literature, conversational, legal
- 252 Source Documents (5k)
- 4 lengths of the same Source Document (≈5k, ≈1k, ≈500, ≈250 words)
- 13-15 QA on each Long Context Source Document
- Annotation on the presence of these QA on each of the summaries
- Human evaluation of automatic and manual summaries
- Human evaluation of summary expansion

C Summary evaluation

Table 5 summarises the metrics used to evaluate.

D Selection of automatic metrics

At an early stage of our work (before collecting human annotations), we came up with a list of candidate metrics of summarization quality. We provide it in Table 6.

To select a single metric per aspect, we analyzed their Spearman correlations with human annotations of the English subset of SEAHORSE (Clark et al., 2023), focusing on Questions 2 to 6: absence of repetitions, grammatical correctness, attribution to the source, coverage of the source, and conciseness (which also reflects overall quality). We did not work with Question 1 (comprehensibility), because our metrics focus on inter-sentential comprehensibility, whereas 66% of SEAHORSE summaries consist of a single sentence.

We report the resulting correlations, grouped by the three English SEAHORSE subsets, in Table 7. One can see that the metrics we selected for summarization correlate reasonably with their corresponding aspects across all 3 subsets most of the time[7], which justifies our choice of them.

E Summarization Guidelines

Annotator proficiency requirements

- Native speaker of English
- Editor / writer / domain expert

Task You will receive document(s) that are approximately 5,000 words or longer from the following domains:

- Political (GovReports, MultiUN)
- News (Seahorse)

[7] An exception is the grammaticality/fluency aspect (Question 2) on the Wikihow subset, where no metric demonstrated adequate correlation with human annotations.

Task	Area	Metric	Description	Reference
Sum	Target similarity	R-L	ROUGE-L (longest common subsequence)	Lin (2004)
Sum/SumExp	Grammaticality	REP-3	Portion of duplicated N-grams (N=4)	Welleck et al. (2019)
Sum/SumExp	Fluency	CoLA	Sentence fluency classifier score	Krishna et al. (2020)
Sum/SumExp	Coherence	COH-2	2nd-order word-level coherence score	Parola et al. (2023)
Sum	Attribution	SH-4	Seahorse-Large-Q4 score	Clark et al. (2023)
Sum	Semantic coverage	SH-5	Seahorse-Large-Q5 coverage score	Clark et al. (2023)
SumExp	Word Count	WC		
Sum/SumExp	Overall	AVG	Empirical average of metrics	
Sum/SumExp	Overall	HE	HelloEval score	Que et al. (2024)

Table 5: Summary of automatic metrics used in different tasks.

Metric	Short name	Implementation
ref_rouge_fmeasure	R-L	ROUGE-L F1 score, computed with the `rouge-score` Python package (in the `rougeLsum` mode)
ref_rouge_recall		ROUGE-L recall (see above)
ref_rouge_precision		ROUGE-L precision (see above)
char_len_ratio		Summary-to-source lengths ratio, in characters
sent_len_ratio		Summary-to-source lengths ratio, in sentences
word_len_ratio		Summary-to-source lengths ratio, in words
td_coherence_1st		1st order coherence, computed with the `textdescriptives` Python package
td_coherence_2nd	COH-2	2nd order coherence, computed with the `textdescriptives` Python package
td_flesch_kincaid_grade		Flesch-Kincaid grade readability score (`textdescriptives`)
td_gunning_fog		Gunning-Fog readability score (`textdescriptives`)
td_dependency_distance_mean		Mean distance of syntactic dependencies (`textdescriptives`)
word_ngram_src_overlap		Fraction of source word n-grams (1 to 3) that appear in the summary
word_ngram_repetition_rate	REP-3	Fraction of summary word n-grams (1 to 3) that appear more than once
src_rouge_recall		ROUGE-2 recall of the summary w.r.t. the source
src_rouge_f1		ROUGE-2 F1 score of the summary w.r.t. the source
src_rouge_precision		ROUGE-2 precision of the summary w.r.t. the source
tgt_fluency	CoLA	Mean predicted probability of the sentence being linguistically acceptable, as per the model from Krishna et al. (2020)
src_fluency_diff		Difference of mean predicted probabilities of being acceptable in the summary and in the source
mean_self_sonar_sim		Mean cosine similarity SONAR embeddings of each summary sentence to its most similar summary sentence
mean_src_sonar_sim		Mean cosine similarity SONAR embeddings of each summary sentence to its most similar source sentence
mean_src_coverage_sonar_sim		Mean cosine similarity SONAR embeddings of each source sentence to its most similar summary sentence
mean_monotonic_src_sonar_sim		Mean cosine similarity SONAR embeddings of each summary sentence to its monotonically aligned source sentence
sonar_sim_monotonicity		Ratio of `mean_monotonic_src_sonar_sim` to `mean_src_sonar_sim`
p_entail_full		Probability of the summary being entailed by the source, predicted with `tasksource/deberta-small-long-nli` model
p_noncontradict_full		Probability of the summary not contradicting the source, predicted with `tasksource/deberta-small-long-nli` model
smollm_reconstruct_loss		Cross-entropy loss of the source given the summary, computed with the `HuggingFaceTB/SmolLM-360M-Instruct` model
seahorse_q4	SH-4	Probability of the “Yes” response extracted from the SEAHORSE Q4 model (`google/seahorse-large-q4`)
seahorse_q5	SH-5	Probability of the “Yes” response extracted from the SEAHORSE Q5 model
seahorse_q6		Probability of the “Yes” response extracted from the SEAHORSE Q6 model

Table 6: The automatic metrics we considered as candidates, along with their implementation details.

- Wikipedia (Wikipedia)
- Scientific/Technical (FacetSum)
- Literature (BookSum, SQuality)
- Conversational (Summscreen)
- Legal (LexGlue, JRCAcquis)

The documents will contain sections/chapters. You will need to summarize them retaining the section alignment. For certain domains, there will be additional guidance in the form of special guidelines (legal, medical etc.)

You will need to create 3 summaries:

- Summary 1: around 20% of the source text (1,000 words if total length is 5000)
- Summary 2: around 10% of the source text (500 words)
- Summary 3: around 5% of the source text (250 words)

After finishing summarizing, you will need to write a minimum of 15 questions with corresponding answers (QA) per each 5000 words.

Requirements for Summarization Here’s more information on what that means and how to summarize:

Please read the provided document in its entirety. Consider making notes of the main core ideas while you read. After you have finished reading, please write a short summary of the source text. The summary should:

SEAHORSE subset	wiki					xlsum					xsum				
Quality aspect	q2	q3	q4	q5	q6	q2	q3	q4	q5	q6	q2	q3	q4	q5	q6
ref_rouge_fmeasure (R-L)	0.33	0.03	0.11	0.25	0.27	0.17	0.12	0.12	0.23	0.18	0.11	0.11	0.13	0.16	0.15
ref_rouge_recall	0.31	0.02	0.15	0.26	0.25	0.16	0.11	0.10	0.23	0.18	0.11	0.09	0.12	0.17	0.14
ref_rouge_precision	0.29	0.02	0.03	0.16	0.20	0.17	0.11	0.14	0.20	0.18	0.12	0.10	0.14	0.14	0.15
char_len_ratio	0.03	0.08	-0.04	-0.04	-0.06	0.00	0.03	0.09	0.20	0.13	0.05	-0.01	-0.06	0.12	0.03
sent_len_ratio	-0.02	0.14	-0.06	-0.05	-0.07	-0.04	-0.02	0.08	0.13	0.08	0.08	-0.02	-0.05	0.05	0.01
word_len_ratio	0.03	0.08	-0.05	-0.05	-0.07	-0.02	0.03	0.07	0.18	0.11	0.06	-0.02	-0.07	0.11	0.02
td_flesch_kincaid_grade	0.08	-0.01	0.10	0.11	0.11	0.04	0.07	0.01	0.11	0.09	0.01	0.00	0.10	0.12	0.12
td_gunning_fog	0.14	-0.03	0.12	0.15	0.15	0.02	0.08	-0.00	0.07	0.05	0.05	-0.03	0.10	0.13	0.13
td_dependency_distance_mean	0.07	-0.05	-0.01	-0.01	-0.04	0.09	0.03	-0.02	0.09	0.04	0.06	-0.01	-0.03	0.07	0.05
word_ngram_src_overlap	0.18	0.09	0.08	0.06	0.05	0.06	0.09	0.14	0.18	0.15	0.03	-0.01	0.05	0.15	0.08
word_ngram_repetition_rate (REP-3)	-0.59	0.05	-0.13	-0.19	-0.23	-0.31	-0.08	-0.13	-0.05	-0.11	-0.27	-0.05	-0.02	0.04	-0.01
src_rouge_recall	0.06	0.09	0.08	0.06	0.03	0.04	0.05	0.11	0.09	0.11	-0.05	-0.01	0.07	0.13	0.08
src_rouge_f1	0.07	0.10	0.09	0.07	0.04	0.04	0.05	0.11	0.09	0.11	-0.05	-0.01	0.07	0.13	0.08
src_rouge_precision	0.10	0.07	0.20	0.12	0.09	0.05	0.03	0.09	-0.03	0.05	-0.09	-0.00	0.09	0.07	0.06
tgt_fluency (CoLA)	0.07	0.09	0.06	0.08	0.09	0.35	0.25	0.10	0.09	0.16	0.23	0.20	0.18	0.12	0.14
src_fluency_diff	0.06	0.08	0.00	-0.01	0.01	0.29	0.18	0.05	0.06	0.14	0.19	0.21	0.10	0.06	0.08
mean_self_sonar_sim	-0.57	0.09	-0.13	-0.14	-0.16	0.02	0.03	-0.02	-0.03	-0.02	0.05	0.02	0.00	0.06	-0.01
mean_src_sonar_sim	-0.10	0.06	0.16	0.03	0.02	0.01	0.07	0.08	0.00	0.01	-0.05	0.01	0.09	0.04	0.07
mean_src_coverage_sonar_sim	0.09	0.08	0.10	0.10	0.07	-0.01	0.09	0.11	0.09	0.06	-0.04	0.01	0.10	0.06	0.09
mean_monotonic_src_sonar_sim	0.10	0.03	0.21	0.13	0.11	0.01	0.07	0.08	0.00	0.01	-0.05	0.01	0.09	0.04	0.07
sonar_sim_monotonicity	0.37	-0.07	0.16	0.13	0.10	-0.04	-0.00	0.08	0.05	0.06	-0.02	0.05	-0.04	-0.07	0.02
p_entail_full	-0.02	0.02	0.38	0.17	0.20	-0.04	0.00	0.21	0.08	0.14	-0.02	-0.00	0.39	0.26	0.31
p_noncontradict_full	-0.02	-0.03	0.20	0.04	0.06	-0.08	-0.05	0.07	-0.03	-0.00	0.01	-0.04	0.25	0.18	0.20
smollm_reconstruct_loss	-0.09	-0.05	-0.11	-0.03	-0.03	-0.01	-0.01	-0.06	-0.11	-0.07	0.05	0.02	-0.17	-0.11	-0.13
seahorse_q4 (SH-4)	0.24	0.03	0.52	0.37	0.41	0.19	0.08	0.31	0.17	0.27	0.14	0.06	0.54	0.35	0.43
seahorse_q5 (SH-5)	0.43	0.03	0.33	0.51	0.47	0.20	0.12	0.25	0.27	0.29	0.16	0.14	0.36	0.30	0.31
seahorse_q6	0.45	0.01	0.42	0.51	0.52	0.26	0.10	0.31	0.22	0.31	0.20	0.10	0.47	0.35	0.41

Table 7: Spearman correlations of candidate automatic metrics with human judgments on the English subset of SEAHORSE (Clark et al., 2023), grouped by subsets of the data.

- Be much shorter than the source text. Please see the information about the length above.
- Convey ALL the main core ideas and information of the source document.
- Have a structure of a standalone cohesive text.
- Follow the flow of the section/paragraph structure of the source text, try to summarize it from top to bottom.
- Each paragraph in each summary should be marked with a number of the source text section/chapter showing where this information is from.

Here's a checklist which can help you with the task:

- **Understand the Main Idea**: Read the entire text to grasp the overall theme and the author's intent. Identify the main idea of the text.
- **Highlight Key Points**: Mark or note down the essential points and arguments.
- **Eliminate Redundancies**: Remove any repetitive information or examples that do not add value to the understanding of the main idea.
- **Use Your Own Words**: Paraphrase the key points in your own words instead of copying verbatim. This helps ensure the summary is concise.
- **Keep It Objective**: Focus on the information presented in the text without inserting personal opinions or interpretations.
- **Structure the Summary**: Organize the ideas logically, maintaining the flow of the original text.
- **Be Concise**: Aim for clarity and brevity; Use simple and direct language to convey the points.
- **Review and Revise**: Compare the summary to the original text to ensure accuracy and completeness. Edit for coherence, transitions, and readability.

Additional Guidelines for Conversational Text You may be assigned to work on conversational type of text, such as meeting transcripts, screenplays, and novels. Since the text structure is quite different from documents, this is an additional guideline to help you working on conversational text:

- Skim through the whole document: Try to get a rough idea of the whole plot

- Identify the characters and main core ideas: Identify the main characters and focus on their interaction

- Omit the trivial details: there maybe side plot or supporting characters in the source text, carefully decide if it is related to the main core ideas (plot)

- Group and summarize with respect to main core ideas: There could be plot twists or related hints in the source documents. Remember the summary should be clear and straightforward, the plot outline

- Should be clear in the summary without referencing to the source document

Requirements for Question and Answer sets

- The questions need to be abstractive not extractive: this means they need to be directed at the ideas in the text, not words and sentences as such.

- The questions should be open-ended, not Yes/No questions

- The correct answers should cover the main points in the source text: the questions should roughly correspond to paragraphs/sections in the text

- Thus, the correct answer should cover points reflected in the summary (for your convenience, you can refer to your longest summary, but please mind your shorter summary also need to be able to answer at least some of the questions)

- The answers should not be short (30 words or more)

- The correct answer should be found namely IN THE SUMMARY and not able to be just pulled from general knowledge

- If possible, refrain from factual questions, but try composing questions for reasoning, such as WHY- questions

- You are encouraged to combine information from different sections together

- Avoid only asking questions about the beginning and the end of the section, use the information in the middle as well

- If possible, the questions should not have several answer possibilities.

F Human Summary Sample

This is one sample of 3 human written summaries and QA set of the document 4586 from GovReport.

F.1 Source Document Excerpt

[This is the first and last paragraphs of the source document. The whole document is 5411 words long.]

This report examines technological innovation in payment systems generally and particular policy issues as a result of retail (i.e., point of sale) payment innovation. The report also discusses wholesale payment, clearing, and settlement systems that send payment messages between banks and transfer funds, including the "real-time payments" service being introduced by the Federal Reserve. This report includes an Appendix that describes interbank payment, clearing, and settlement systems related to U.S. payments.

...

To address systemic risk concerns, a private RTP system could be designated as a systemically important Financial Market Utility (FMU) under Title VIII of the Dodd-Frank Act (P.L. 111-203). The Dodd-Frank Act allows the Financial Stability Oversight Council , a council of financial regulators led by the Treasury Secretary, to designate a payment, clearing, or settlement system as systemically important on the grounds that "the failure of or a disruption to the functioning of the FMU could create or increase the risk of significant liquidity or credit problems spreading among financial institutions or markets and thereby threaten the stability of the U.S. financial system." FMUs, currently including the Clearing House Interbank Payments System, are subject to heightened regulation, and the Fed has supervisory and enforcement powers to ensure those standards are met. Policymakers could consider whether systemic risk concerns are better addressed through Fed operation of payment and settlement systems or Fed regulation of private systems.

F.2 Long, Medium and Short Summaries

Each paragraph of the summaries is paired with the source paragraph id (e.g. p1, p2, etc) to indicate the information source.

Long Summary (20%) To the average consumer, swiping their credit card seems simple, because the complex the infrastructure involved in is 'hidden'. These deceptively "simple" electronic payments are comprised of three main steps. First, the sender makes the payment through an online payment service or an app, which instructs the sender's bank to make the payment to the recipient. Second, the bank sends a payment message to the recipient's bank through a payment system or clearing service. Finally, the payment is completed (settled) when the funds are received by the recipient. (p2)

Some of the bank-to-bank (ACH) payment, clearing, and settlement (PCS) systems are operated by the Federal Reserve, and others by private-sector organizations. Recently, the use of electronic payment methods (credit card, debit card, and ACH) has grown, while the use cash and check payments has declined. Electronic payments have been made easier and more convenient with digital wallets and payment apps like Venmo, Cash App, and Zelle - all of which require users to link a bank account, credit card, or debit card. (p4, p5, p6, p7)

There are concerns about whether current regulations are equipped to handle electronic payments. If not, this poses potential risks to cybersecurity, data privacy, industry competition, and consumer access and protection. Current payment regulations depend, in part, on if the service is provided by a bank, who have many strict regulatory requirements. As such, Nonbank payment systems are not subject to existing regulatory enforcement and can only be supervised - as money transmitters at state level and money service businesses at federal level. (p8, p9)

Electronic payment regulatory concerns could be addressed by including nonbank payment companies into the bank regulatory regime. One way could be via the Office of the Comptroller of the Currency (OCC) special purpose national bank charter. And another, through a state-level industrial loan company (ILC) charter with the Federal Deposit Insurance Corporation (FDIC). Both methods could provide nonbank firms access to the Fed wholesale payment systems, which could be advantageous. However, some state regulators have filed lawsuits to block nonbank companies access to these charters, arguing that it allows companies to circumvent state consumer protections. So far, no companies have applied for an OCC charter, likely due to the legal uncertainty surrounding it. (p11, p12)

The main argument against nonbank payment companies filing ILC charters is that it would allow them to own banks - and the FDIC has not approved deposit insurance for a new ILC since 2006. Opponents argue that allowing a company to own a bank could expose the US economy to risks like imprudent underwriting. Proponents assert that these concerns are exaggerated, noting that several other countries allow similar arrangement with no ill effects. So far, Square is the only company with a pending application and two other companies have withdrawn their applications. (p13)

New technology reduces some risks related to payments but creates new ones. The risk of having one's wallet stolen is reduced, but payment information is subject to more sophisticated risks such as malware attacks. Furthermore, storing payment information on a variety of websites, apps, and devices creates more opportunities for hackers. After recent security breaches which allowed user information to be stolen, several solutions have been proposed. For example, a federal breach notification law could be enacted, to create federal cybersecurity standards or to increase penalties for companies with inadequate security measures. (p15, p16, p17)

Payment systems need to collect detailed information about customer transactions in order to function properly. This data can be used by companies to target ads. Scammers can also use this information for fraudulent purposes. The constantly increasing use of Electronic payments has led to questions about how user data is used and whether consumers are sufficiently informed and given enough control about how their data is used. (p18)

There are some consumer benefits to storing consumer data. It can help them track payments and budget more easily by importing to budgeting apps. They can also share financial information with banks more easily when applying for loans. But, given the benefits and the risks, the question remains: how much access should companies have to individuals' information? (p19)

Privacy policies are another area of concern with respect to consumer protection and electronic payments. According to the Bureau of Consumer Financial Protection (CFPB), it is difficult to provide disclosures that are clear and easy to understand, partly due to the small screens on phones. Clearer privacy policies and allowing consumers more control over how their data is used could help. (p20)

The Electronic Fund Transfer Act, Regulation

E implemented by the CFPB, is the most relevant law aimed at protecting consumers who are making electronic payments. It mandates consumer disclosures, limits consumer liability for unauthorized payments, and maintains procedures for resolving errors. Further regulations are being considered. (p22)

Consumers could also be protected through financial education, especially for more at-risk older and lower-income groups. This could include learning how to use new payment systems safely and how to protect against financial harm, as well as knowing how to get help if something goes wrong. (p24)

Payment system innovations may affect consumers differently based on income. Consumers who mainly pay with cash, don't have bank accounts, or don't have internet or mobile access won't be able to benefit. Neither will those who are not comfortable using new technology. (p25)

However, surveys reveal that 83% of underbanked, and 50% of unbanked, consumers have smart phone access. So, as costs of these payment services decline, some marginalized groups could experience better access to the the financial system through access to digital currency channels via cash equivalents like pre-paid cards. But, the cost of internet and mobile data plans may limit access to faster payment systems, so this also needs to be considered. (p26, p27, p28)

Faster payment systems may also benefit low-income consumers by allowing them faster access to their paychecks and other fund transfers. But a potential drawback is that withdrawals from their accounts would occur more quickly as well. (p28)

In 2019, the Fed announced that it plans to create a wholesale real-time payment (RTP) system. (p32)

Originally, the Fed's primary function was to provide bank-to-bank check-clearing services. Private clearing houses were experiencing issues that led to the creation of the Fed. As payment methods have evolved, the Fed has begun providing other types of payment systems. It does this by linking the accounts that all banks keep at the Fed so that it can complete the transfers. The new system that the Fed is developing, called FedNow, would allow payments to occur in real time, rather than later in the day - or even the next day, as is the case currently. (p33, p35)

However, there are some concerns regarding implementation of FedNow. Many worry that it will undermine private sector development of similar systems. Others fear that failing to implement FedNow will lead to a monopoly of a private-sector company, to the detriment of consumers and smaller banks. (p42, p44)

Medium Summary (10%) Electronic payments have three stages. First, the sender makes the payment through an online payment service or an app, which instructs the sender's bank to make the payment to the recipient. Second, the bank sends a payment message to the recipient's bank through a payment system or clearing service. Finally, the payment is completed (settled) when the funds are received by the recipient. (p2)

Some of the bank-to-bank (ACH) payment, clearing, and settlement (PCS) systems are operated by the Federal Reserve, and others by private-sector organizations. Recently, the use of electronic payment methods (credit card, debit card, and ACH) has grown, while the use cash and check payments has declined. Electronic payments have been made easier and more convenient with digital wallets and payment apps like Venmo, Cash App, and Zelle - all of which require users to link a bank account, credit card, or debit card. (p4, p5, p6, p7)

There is concern about whether current regulations are equipped to handle these technological advances. If not, they could pose risks to cybersecurity, data privacy, industry competition, and consumer access and protection. (p8)

One way to address these concerns is to add nonbank companies to the bank regulatory regime. Another is via the Office of the Comptroller of the Currency (OCC) special purpose national bank charter. And another, through a state-level industrial loan company (ILC) charter with the Federal Deposit Insurance Corporation (FDIC). Both could provide nonbank firms access to the Fed wholesale payment systems, which could be advantageous. However, some state regulators have tried to block nonbank access to these charters, arguing that it allows companies to circumvent state consumer protections. So far, no companies have applied for an OCC charter, likely due to the legal uncertainty surrounding it. (p11, p12)

The main argument against ILC charters is that it would allow companies to own banks. The FDIC has not approved deposit insurance for a new ILC since 2006. Opponents argue that allowing a retailer to own a bank could expose the US economy to risks such as imprudent underwriting. Proponents assert that these concerns are exaggerated and that several other countries allow similar arrangement with no ill effects. Currently, Square is the only company with a pending application. (p13)

Privacy policies are another area of concern. According to the Bureau of Consumer Financial Protection (CFPB), it is difficult to provide disclosures that are clear and easy to understand, partly due to the small screens on phones. Clear privacy policies and allowing consumers more control over how their data is used could help. (p20)

The Electronic Fund Transfer Act, Regulation E implemented by the CFPB, is the most relevant law aimed at protecting consumers who are making electronic payments. It mandates consumer disclosures, limits consumer liability for unauthorized payments, and maintains procedures for resolving errors. (p22)

In 2019, the Fed announced that it plans to create a wholesale real-time payment (RTP) system.The proposed system, called FedNow, would allow payments to occur in real time, rather than later in the day - or even the next day, as is the case currently. (p32, p33, p35)

There are some concerns regarding FedNow. Many worry that it will undermine private sector development of similar systems. Others fear that failing to implement FedNow will lead to a monopoly of a private-sector company, to the detriment of consumers and smaller banks. (p42, p44)

Short Summaries (5%) Electronic payments have three stages. First, the sender makes the payment through an online payment service or an app, which instructs the sender's bank to make the payment to the recipient. Second, the bank sends a payment message to the recipient's bank through a payment system or clearing service. Finally, the payment is completed (settled) when the funds are received by the recipient. (p2)

Some of the bank-to-bank payment, clearing, and settlement (PCS) systems are operated by the Federal Reserve, and others by private-sector organizations. Recently, the use of electronic payment methods has grown, while the use cash and check payments has declined. (p4)

There is concern about whether current regulations are equipped to handle these technological advances. If not, they could pose risks to cybersecurity, data privacy, industry competition, and consumer access and protection. (p8, p9)

A potential way to resolve concerns regarding regulation is to add certain nonbank payment companies into the bank regulatory regime. One way to accomplish this is through the OCC or Office of the Comptroller of the Currency special purpose national bank charter. The second is through a state-level industrial loan company (ILC) charter with the Federal Deposit Insurance Corporation (FDIC). (p11, p12)

Privacy policies are another area of concern. According to the Bureau of Consumer Financial Protection (CFPB), it is difficult to provide disclosures that are clear and easy to understand, partly due to the small screens on phones. Clear privacy policies and allowing consumers more control over how their data is used could help. (p20)

In 2019, the Fed announced that it plans to create a wholesale real-time payment (RTP) system. The proposed system, called FedNow, would allow payments to occur in real time, rather than later in the day - or even the next day as is the case currently. (p32, p33, p35)

F.3 Question and Answer Set

Question 1: What are the three parts of a payment system?

- **Answer**: First, there is the sender or the person making the payment through an online payment service or an app, which instructs the sender's bank to make the payment to the recipient. Second, the bank sends a payment message to the recipient's bank through a payment system or clearing service. Finally, the payment is completed when the funds are transferred, or settled.
- (Information contained in 20% Summary, 10% Summary, 5% Summary.)
- (Source paragraph number: p2)

Question 2: Who operates bank-to-bank payment, clearing, and settlement systems?

- **Answer**: Some of these systems are operated by the Federal Reserve and some are operated by private-sector organizations.
- (Information contained in 20% Summary, 10% Summary, 5% Summary.)
- (Source paragraph number: p4)

Question 3: What issues could there be if current regulations are not equipped to handle these payment system innovations?

- **Answer**: If regulations are inadequate, there could be issues related to cybersecurity, data privacy, industry competition, and consumer access and protection.
- (Information contained in 20% Summary, 10% Summary, 5% Summary.)
- (Source paragraph number: p8)

Question 4: What are two ways to bring nonbank companies into the bank regulatory regime?

- **Answer**: One way to accomplish this is through the OCC or Office of the Comptroller of the Currency special purpose national bank charter. The second is through a state-level industrial loan company (ILC) charter with the Federal Deposit Insurance Corporation (FDIC).
- (Information contained in 20% Summary, 10% Summary, 5% Summary.)
- (Source paragraph number: p11,12)

Question 5: According to the Bureau of Consumer Financial Protection, what are some of the difficulties with privacy policies?

- **Answer**: It is difficult to provide disclosures that are clear and easy to understand, partly due to the small screens on phones.
- (Information contained in 20% Summary, 10% Summary, 5% Summary.)
- (Source paragraph number: p20)

Question 6: What is FedNow?

- **Answer**: In 2019, the Fed announced that it plans to create a wholesale real-time payment (RTP) system. The proposed system, called FedNow, would allow payments to occur in real time, rather than later in the day or even the next day as is the case currently.
- (Information contained in 20% Summary, 10% Summary, 5% Summary.)
- (Source paragraph number: p29)

Question 7: What are some reasons for the increase in electronic payments?

- **Answer**: Electronic payments have increased because of payment apps such as Venmo, Cash App, and Zelle make it convenient and easy for consumers to send payments. Digital wallets stored on phones are another reason for increased electronic payments due their ease of use and convenience.
- (Information contained in 20% Summary, 10% Summary.)
- (Source paragraph number: p6)

Question 8: What is necessary in order for a consumer to be able to use electronic payment services?

- **Answer**: The consumer must have a debit card, credit card, or bank account linked to an electronic payment system.
- (Information contained in 20% Summary, 10% Summary.)
- (Source paragraph number: p7)

Question 9: Why have state regulators filed lawsuits to block the OCC?

- **Answer**: Regulators feel that the OCC charter would allow companies to avoid state regulations that protect consumers.
- (Information contained in 20% Summary, 10% Summary.)
- (Source paragraph number: p11, p12)

Question 10: What is the main argument against the ILC charter?

- **Answer**: The ILC would allow companies such as retailers to own banks. Opponents are concerned that this could lead to imprudent underwriting and could hurt the US economy by exposing it to risk.
- (Information contained in 20% Summary, 10% Summary.)
- (Source paragraph number: p13)

Question 11: What does the Electronic Funds Transfer Act Regulation E do?

- **Answer**: Regulation E mandates consumer disclosures, limits consumer liability for unauthorized payments, and maintains procedures for resolving errors.

- (Information contained in 20% Summary, 10% Summary.)
- (Source paragraph number: p22)

Question 12: How could financial education help consumers use electronic payment systems safely?

- **Answer**: Consumers could be taught how to use new payment systems safely and how to protect against financial harm, as well as how to get help if something goes wrong.
- (Information contained in 20% Summary.)
- (Source paragraph number: p24)

Question 13: What are is an argument against the FedNow?

- **Answer**: Many worry that it will undermine private sector development of similar systems.
- (Information contained in 20% Summary, 10% Summary.)
- (Source paragraph number: p42, p44)

Question 14: How can storing more consumer data benefit consumers?

- **Answer**: It can help consumers track payments and budget more easily using budgeting apps. They can also share financial information with banks more easily when applying for loans.
- (Information contained in 20% Summary.)
- (Source paragraph number: p19)

Question 15: How could faster payment systems affect low-income consumers?

- **Answer**: Faster payment systems may benefit low-income consumers by allowing them faster access to their paychecks and other fund transfers. But a potential drawback is that withdrawals from their accounts would occur more quickly as well.
- (Information contained in 20% Summary.)
- (Source paragraph number: p28)

G Summarization Human Evaluation Guidelines

Annotator proficiency requirements All annotators must meet ALL of the following requirements:

- Native speaker of English AND
- Language related degree holder or related professionals

Background Information What is a good summary? A good summary should meet the following criteria:

- **Conciseness**: The summary should only contain the most important information while maintaining readability. Trivial information should not be included, even in the longest summary. Additionally, the summary should be comprehensible on its own, without needing to refer to additional documentation.
- **Coverage of main core ideas**: The summary should preserve the most important ideas, regardless of its length. In our task, summaries are created by gradually omitting less important information. Therefore, we expect that the core ideas will be retained in all summaries, even the shortest ones. Main core ideas should be the key ideas that help the reader to understand the main topic. Depending on the type of documents, the definition of idea would be slightly different. For example, if the source document is a meeting note, the summary should include the main topic, the discussion, the result / final decision. The trivial details like greetings or small talks should not be included. If the source document is a novel, the summary should focus on main characters and important events rather than trivial description of the character or side events.
- **Attribution**: the information in the summary can be accurately referred back to the source documents. All the information in the summary should be an abstraction from the source documents. No additional information that can not be found in the source document should be included in the summary.
- **Cohesion as a document**: Each summary should be an abstraction of the entire source

document. All the information or ideas should be digested from different parts of the source document and combined into a new paragraph. Merely shortening a document paragraph by paragraph will not be considered as a good summary. Similarly, a bulletin-like document jumping from point to point also will not be considered as a good summary.

Annotation Our summarization structure is as follows: The source text which is approximately 5,000 words long gets summarized three times: The first summary is 20% of the original length of the doc. It should retain all the core ideas of the source document. The second summary is 10% of the length of the source document. It should also retain all the core ideas of the source. The third summary is short, it should be 5% of the source length. We understand there will be some information loss, but again, all the core ideas should be present in the summary.

There are two tasks related to evaluating the summary.

Task 1 In this task, you need to rate the overall quality of the summaries regarding several aspects. Here is the detailed workflow:

Step 1 Screening

Please spend no more than 5 minutes skimming through the longest (20%) summary and answer the question below.

- Q1: Is it a cohesive text? Can you fully understand it?
  - If NOT, reject the task completely.
  - If YES, continue with the following steps

Step 2 Read the texts and take notes

Please read the whole source document carefully and take notes in your own way. It could be highlighting the key points or jotting down the ideas in your own words or any means that can help you digest the document. While reading please do not skip any line. After reading and taking note, you should be able to identify several main core ideas or more (You can spot more main core ideas if the text is longer). Please continue to read the summary and identify if the main ideas also exist in the summaries. Now check how many ideas can be found in the summary.

(This procedure is to help rate the summaries more objectively, you are not required to submit the highlights or the notes.)

Step 3 Rate the summaries

Answer all the following questions with a 4-point scale:

- Q2a Check the attribution of the summary. Can all the information in the summary be attributed to the source text?
  - Give 4 points if yes, all the information can be directly attributed to the source text.
  - Give 3 points if mostly yes, only 1 idea seems to not be found in the source text.
  - Give 2 points if not really, more than 1 idea cannot be attributed to the source text.
  - Give 1 point if not, most ideas cannot be found in the source text and seem to be completely new.
  - Give 0 points if not, none of the ideas can be found in the source text.
- Q2b Check the coverage of main core ideas of the source text in the summary. Are all the main core ideas of the source document retained?
  - Give 4 points if yes, all the main core ideas of the source are retained.
  - Give 3 points if mostly yes, only 1 or 2 main core ideas are not found in the summary.
  - Give 2 points if not really, more than 2 main core ideas are not found in the summary.
  - Give 1 point if not, most main core ideas are not found in the summary.
  - Give 0 points if not, none of the ideas can be found in the summary
- Q2c Check the conciseness of the summary. Is the summary short and clear without repetition and redundancy?
  - Give 4 points if yes, the summary is not wordy but clear.
  - Give 3 points if mostly yes, but 1 part is unnecessary.
  - Give 2 points if not really, more than 1 part is unnecessary.
  - Give 1 point if not, the summary is lengthy and most passages can be omitted without losing the core ideas.

  - Give 0 points if not, the summary is lengthy and most passages can be omitted without losing the core ideas.

- Q2d Check the readability of the summary. Is the summary fluent and understandable?
  - Give 4 points if yes, the summary is fluent and understandable, and well written.
  - Give 3 points if mostly yes, but there is room for improvement.
  - Give 2 points if not really, not quite fluent and sometimes hard to understand.
  - Give 1 point if not, the summary is hard to read and understand.
  - Give 0 points if not, the summary is impossible to read and understand.

  After evaluating all the aspects of the summary, please give an overall score of 0-10 on the quality of the summary.

- Q3 Do you think it is a good summary? On a scale of 0-10, how would you rate the overall quality of the summary?
  - 10: The summary is perfect in every aspect
  - 8-9: The summary is considered good. It contains minor issues in certain aspects but it meets all requirements with room for improvement.
  - 6-7: The summary is moderate, it contains non-critical errors but to help the reader understand the source documents
  - 4-5: The summary is below acceptable level. It contains critical errors that could potentially mislead the reader.
  - 2-3: The summary contains very limited information that is relevant to the source document
  - 0-1: The summary is barely readable and comprehensible or it barely contains relevant information to the source document.

  Make sure you have answered all the questions for every summary.

Task 2 You will be provided with 15 questions depending on the length of the documents. Please identify if the answer is directly stated, heavily implied, or logically entailed in the summary. You need to answer YES or NO only.

H Summary Expansion Human Evaluation Guidelines

Annotator proficiency requirements All annotators must meet ALL of the following requirements:

- Native speaker of English AND
- Language related degree holder or related professionals

Background Information What is a good summary expansion? A good summary expansion should meet the following criteria:

- **Coverage of main core ideas**: Main core ideas should be the key ideas covered in the original summary. During each expansion more details will be added, however the main core ideas should remain the same. For example, if the setting of the original summary is an office comedy, the ultimate text should not be an unrelated superhero movie screenplay.

What is a good long form text? On top of coverage of main core ideas, a good summary expansion should also meet the following criteria: Cohesion as a document:

- The ultimate document should be a standalone document so that the reader doesn't need an additional document to understand the text. The ultimate document should be well-structured and formatted. For example, if the ultimate document is a screenplay, it should have clear scene sections, character direction and dialogs of characters.
- Non-repetitive and rich in details: Although the ultimate document is based on the summary, additional information / details is allowed. The ultimate document should not just repeat the core ideas.
- Being Interesting: A good screenplay and novel should be able to capture the reader's attention and keep them engaged throughout the story / plot. We are not searching for an Oscar winning novel / screenplay, as long as the plot makes sense and the added details serve the purpose of the story, it is considered as an interesting plot. For example, you can check the following questions depending on the story: If it's a comedy, does it sound funny

to you? If it's a romance story, does it evoke the proper sentiment? Are the added details aligned with the plot, or do they feel out of place? . . . etc

Task 1 In this task, you need to rate the overall quality of the summary expansions regarding several aspects. Here is the detailed workflow:

Step 1 Screening

Please spend no more than 5 minutes skimming through the long form text and answer the question below.

- Q1: Is it a cohesive text? Can you fully understand it?
  - If NOT, reject the task completely.
  - If YES, continue with the following steps

Step 2 Read the texts and highlight key points

Please read the original summary carefully and highlight the key points and make notes. Do not skip any line. Continue to read other summaries and long form text, highlighting the key points that are the same as the original summary

(This procedure is to help rating the summaries more objectively, you are not required to submit the highlights or the notes.)

Step 3 Rate the long form text

Answer all the following questions with a 4-point scale separately for the long form text:

- Q2a Check the coverage of main core ideas. Are all the core concepts of the original summary retained?
  - Give 4 points if yes, all the main core ideas are retained.
  - Give 3 points if mostly yes, only 1 or 2 core ideas are lost.
  - Give 2 points if not really, more than 2 ideas are lost.
  - Give 1 point if not, most ideas are lost.
  - Give 0 points if not, none of the ideas can be found in the source text.
- Q2b Check the cohesion of text. Is it well structured? Does it contain all the necessary components? (Scene description, dialog, main characters, etc) Does it flow logically and maintain consistency?
  - Give 4 points if yes, it is a well structure screenplay / novel
  - Give 3 points if mostly yes, but 1 part is missing.
  - Give 2 points if not really, more than 1 part is missing.
  - Give 1 point if text does not follow the structure of a screenplay / novel
  - Give 0 points if not, the text doesn't not read as a cohesive text at all
- Q2c Check the richness in details. Does it contain enough details?
  - Give 4 points if yes, the text contains a lot of details.
  - Give 3 points if yes, the text contains details but has room for improvement.
  - Give 2 points if not really, the text contains limited details.
  - Give 1 point if not, the text contains very few details
  - Give 0 points if not, the text does not provide any additional details at all.
- Q2d Check the creativity. Does the added details novel and original while being relevant to the core main ideas?
  - Give 4 points if yes, all of the added details are novel and original
  - Give 3 points if yes, most of the added details are novel and original.
  - Give 2 points if not really, only some of the added details are novel and original.
  - Give 1 point if no, very few added details are repetitive.
  - Give 0 points if no, no added details are novel and original .
- Q2e Check the non-repetitiveness. Does it repeat a lot?
  - Give 4 points if no, all of the details are unique and different from each other
  - Give 3 points if no, most of the details are unique, only one is repeated
  - Give 2 points if not really, some of the details are repetitive
  - Give 1 point if yes, most the added details are repetitive
  - Give 0 points if not, all the added details are repetitive

- Q2f Rate the story plot. How interesting is it to you? Is it engaging and compelling?
    - Give 4 points if the text is very interesting.
    - Give 3 points if the text is quite interesting.
    - Give 2 points if the text is somewhat interesting.
    - Give 1 point if the text is only slightly interesting.
    - Give 0 points if the text is dull and not interesting at all.

  After evaluating all the aspects of the expanded text, please give an overall score of 0-10 on the quality of the expanded text.

- Q3 Do you think expanded text is well written? Do you think it is a good read? On a scale of 0- 10, how would you rate the overall quality?
    - 10: The text is perfect in every aspects
    - 8-9: The text is considered good. It contains minor issues in certain aspects but it meets all requirements with room for improvement.
    - 6-7: The text is moderate, it contains non-critical errors but to help the reader understand the source documents
    - 4-5: The text is below acceptable level. It contains critical errors that cause trouble to read
    - 2-3: The text contains very limited information that is relevant to the source summary
    - 0-1: The text is barely readable and comprehensible or it barely contains relevant information to the source summary.

Make sure you have answered all the questions for every expanded summary and long form text.

Task 2 You will be provided with around 15 questions depending on the length of the documents. You need to answer YES or NO to each QA set. Answer YES only when the answer is directly stated, heavily implied, or logically entailed in the text.

I Prompting Details

The prompts contain three parts: General guideline, domain-specific prompts, and input context. The general guideline adapts the human guidelines (Appendix E) for the summarization and summary expansion, while the domain specific prompts give extra information about the domain as instructions of expected output. In the prompt template below, the general guideline is provided, `{{domain-X}}` denotes the domain-specific prompt. `{{input}}` is for the input document for the summarization task and human summaries for the summary expansion task.

Prompt for summarization

```
"""
You are a professional editor and reader.
You are reading a {{domain}}
{{domain-meta}}.
The {{domain}} starts with [START]
and ends with [END].
After you have finished reading, please
provide a summary of the {{domain}}.
{{domain-expect}}.
Make sure the summary has
{{len(input) * ratio + 200}} words or
less.
[START]
{{input}}
[END].
Write at least {{len(input) * ratio}}
words.
"""
```

Prompt for summary expansion

```
"""
You are a professional editor and reader.
You are reading a summary of {{domain}}
{{domain-meta}}.
The {{domain}} starts with [START]
and ends with [END].
After you have finished reading, write
a well-structured, consistent {{domain}}
that extends the summary.
{{domain-expect-expand}}.
[START]
{{input}}
[END].
Write at least {{len(source)}} words.
"""
```

The model-specific prompts for each domain are listed below. Note that not all domains have the prompt template for summary expansion.

- **BookSum:**

```
{{domain}}: "book chapter"
{{domain-meta}}: """about the book
  [BOOK-TITLE], chapter [CHAP-NO],
  title [[CHAP-TITLE]]."""
{{domain-expect}}: ""
{{domain-expect-expand}}: """Please
  keep the main plot and characters
  if found in the summary.
""".
```

- **LexGLUE:**

```
{{domain}}: "legal document"
{{domain-meta}}: ""
{{domain-expect}}: """Keep the main
  ideas and terms in the document.
"""
```

- **SQuALITY:**

```
{{domain}}: "short story"
{{domain-meta}}: ""
{{domain-expect}} : """Keep the main
  character names and narratives
  of the story.
"""
{{domain-expect-expand}}: (same)
```

- **Seahorse:**

```
{{domain}}: "news article"
{{domain-meta}}: ""
{{domain-expect}}: ""
```

- **FacetSum:**

```
{{domain}}: "academic article"
{{domain-meta}}: "about [TITLE]"
{{domain-expect}}: """Keep the
  structure of sections [SECTIONS]
"""
{{domain-expect-expand}}: (same)
```

- **JRC-Acquis:**

```
{{domain}}: "document"
{{domain-meta}}: "from  European Commision"
{{domain-expect}}: ""
```

- **MultiUN:**

```
{{domain}}: "document"
{{domain-meta}}: "from United Nation"
{{domain-expect}}: ""
```

- **GovReport:**

```
{{domain}}: "government report"
{{domain-meta}}: ""
{{domain-expect}}: ""
```

- **Wikipedia:**

```
{{domain}}: "Wikipedia article"
{{domain-meta}}: "about [TITLE]"
{{domain-expect}}: ""
```

- **Summscreen:**

```
{{domain}}: "screenplay"
{{domain-meta}}: "about [TITLE]"
{{domain-expect}}: """Keep the main
  plot and characters in the screenplay
"""
{{domain-expect-expand}}: """Keep the
  main plot and characters in the
  summary.
  Write in the dialogue form with
  multiple utterances
"""
```

J Detailed Results

Complementing Table 3 from Section 3, Tables 8 and 9 show results of different models in the summary expansion tasks on a different level of summary and expansion. Tables 10, 11 and 12 show the detailed results breaked down by domain on the summary expansion task given the summaries (5%, 10% and 20%) and expanding by a respective factor of (20, 10, 5). It is shown that, despite having repeated instructions on the length, all models behave greatly differently in different domains. In particular, GPT-4o-mini can generate longer scientific/technical texts, but struggle to generate longer texts in conversational texts without sacrifying the qualities. On the other hand, medium-sized models such as LLaMA 3.1-8B generate more texts consistently across domains, but at a higher repetition.

The results of the summarization task of different levels are detailed in tables 13, 14 and 15. According to human evaluation, the best performing result in Table 13 is with GPT-4o-mini in the wikipedia domain and in Table 14 is with the same model in the legal domain (LexGlue). For table 15, the best results are with human output in the conversational domain (Summscreen).

Output	%WC	REP-3(↓)	CoLA↑	COH-2↑	AVG↑	HE↑
GPT-4o-mini	**4.789**	0.623	**0.910**	0.564	0.450	67.617
LLaMA 3.1-70B	2.807	1.830	0.913	0.625	0.391	54.344
LLaMA 3.1-8B	1.788	0.781	0.906	0.752	0.501	42.572

Table 8: Performance on the summary expansion task by a factor of 10, given the 10% summary input.

Output	%WC	REP-3(↓)	CoLA↑	COH-2↑	AVG↑	HE↑
GPT-4o-mini	**7.960**	0.592	0.909	0.574	0.455	65.951
LLaMA 3.1-70B	2.296	0.631	0.871	0.720	0.488	51.539
LLaMA 3.1-8B	3.805	0.720	0.967	0.862	0.562	45.141

Table 9: Performance on the summary expansion task by a factor of 20, given the 5% summary input.

DATASET	Model	% WC	REP-3(↓)	CoLA↑	COH-2↑	AVG↑	HE↑
	GPT-4o-mini	7.116	0.420	0.953	0.701	0.523	82.421
BookSum	LLaMA 3.1-70B	3.307	0.631	0.931	0.863	0.556	66.252
	LLaMA 3.1-8B	5.070	2.170	0.975	0.787	0.442	60.393
	GPT-4o-mini	7.335	0.398	0.954	0.695	0.523	55.620
SQuALITY	LLaMA 3.1-70B	3.010	0.530	0.946	0.769	0.536	47.703
	LLaMA 3.1-8B	3.166	0.624	0.947	0.795	0.539	42.532
	GPT-4o-mini	9.348	0.635	0.936	0.593	0.467	59.238
FacetSum	LLaMA 3.1-70B	0.253	0.792	0.769	0.816	0.475	47.145
	LLaMA 3.1-8B	0.369	0.998	0.843	0.841	0.495	27.660
	GPT-4o-mini	8.042	0.914	0.794	0.309	0.307	74.779
Summscreen	LLaMA 3.1-70B	2.615	0.572	0.836	0.432	0.385	50.914
	LLaMA 3.1-8B	3.370	0.712	0.840	0.474	0.391	44.121

Table 10: Performance on the summary expansion task by a factor of 20, given the 5% summary input, per dataset.

DATASET	Model	% WC	REP-3(↓)	CoLA↑	COH-2↑	AVG↑	HE↑
	GPT-4o-mini	5.155	0.443	0.955	0.689	0.518	82.421
BookSum	LLaMA 3.1-70B	3.596	1.453	0.981	0.817	0.503	74.280
	LLaMA 3.1-8B	2.451	0.778	0.950	0.859	0.551	60.163
	GPT-4o-mini	3.531	0.418	0.952	0.665	0.511	56.867
SQuALITY	LLaMA 3.1-70B	3.104	1.098	0.957	0.718	0.485	50.855
	LLaMA 3.1-8B	2.056	0.659	0.951	0.813	0.544	38.008
	GPT-4o-mini	5.220	0.675	0.939	0.573	0.459	62.472
FacetSum	LLaMA 3.1-70B	0.504	3.005	0.949	0.707	0.352	33.248
	LLaMA 3.1-8B	0.356	0.952	0.882	0.860	0.517	29.713
	GPT-4o-mini	5.251	0.954	0.794	0.331	0.311	68.707
Summscreen	LLaMA 3.1-70B	4.023	1.762	0.766	0.259	0.224	58.992
	LLaMA 3.1-8B	2.290	0.736	0.843	0.475	0.390	42.404

Table 11: Performance on the 10% summary expansion task per dataset.

Dataset	Model	% WC	REP-3(↓)	CoLA↑	COH-2↑	AVG↑	HE↑	Hum↑
BookSum	GPT-4o-mini	0.641	0.459	0.960	0.739	0.536	87.161	6.691
	LLaMA 3.1-70B	1.539	0.650	0.867	0.842	0.526	50.969	5.123
	LLaMA 3.1-8B	1.695	0.736	0.936	0.861	0.550	53.692	5.160
SQuALITY	GPT-4o-mini	3.014	0.513	0.961	0.738	0.532	60.385	6.320
	LLaMA 3.1-70B	1.107	0.582	0.952	0.780	0.539	37.775	5.434
	LLaMA 3.1-8B	1.351	0.735	0.955	0.817	0.542	35.017	5.360
FacetSum	GPT-4o-mini	0.502	0.642	0.954	0.640	0.488	70.061	7.693
	LLaMA 3.1-70B	0.138	0.705	0.874	0.877	0.537	26.587	3.453
	LLaMA 3.1-8B	0.205	0.935	0.871	0.871	0.518	28.462	4.173
Summscreen	GPT-4o-mini	3.569	1.215	0.776	0.319	0.284	65.977	5.000
	LLaMA 3.1-70B	1.449	0.782	0.814	0.500	0.386	41.466	3.800
	LLaMA 3.1-8B	1.495	0.828	0.851	0.568	0.418	36.493	4.480

Table 12: Performance on the summary expansion task by a factor of 5, given the 20% summary input.

Dataset	Model	R-L(↑)	REP-3(↓)	CoLA↑	COH-2↑	SH-4↑	SH-5↑	AVG↑	HE↑	Hum↑
					LCFO.5%					
LexGLUE	Human	n/a	0.258	0.930	0.807	0.617	0.339	0.528	46.690	6.360
	GPT-4o-mini	0.342	0.407	0.956	0.688	0.657	0.500	0.479	78.916	7.747
	LLaMA 3.1-70B	0.386	0.415	0.954	0.875	0.625	0.369	0.482	63.280	6.987
	LLaMA 3.1-8B	0.378	0.471	0.972	0.879	0.617	0.383	0.476	59.455	6.907
BookSum	Human	n/a	0.226	0.913	0.762	0.572	0.315	0.503	71.006	6.691
	GPT-4o-mini	0.302	0.257	0.977	0.857	0.599	0.485	0.532	93.168	6.815
	LLaMA 3.1-70B	0.377	0.362	0.976	0.846	0.578	0.374	0.483	76.871	6.272
	LLaMA 3.1-8B	0.372	0.400	0.973	0.846	0.581	0.347	0.469	72.999	6.049
SQuALITY	Human	n/a	0.263	0.922	0.760	0.520	0.334	0.497	33.534	5.173
	GPT-4o-mini	0.285	0.284	0.980	0.841	0.548	0.375	0.492	74.618	6.600
	LLaMA 3.1-70B	0.340	0.472	0.961	0.802	0.463	0.201	0.391	64.237	5.227
	LLaMA 3.1-8B	0.339	0.535	0.968	0.819	0.488	0.233	0.395	57.288	5.827
FacetSum	Human	n/a	0.260	0.945	0.835	0.691	0.436	0.571	57.456	7.053
	GPT-4o-mini	0.404	0.354	0.921	0.568	0.682	0.524	0.468	73.968	7.434
	LLaMA 3.1-70B	0.412	0.387	0.962	0.884	0.696	0.508	0.533	67.585	6.213
	LLaMA 3.1-8B	0.419	0.425	0.967	0.888	0.704	0.518	0.530	69.176	6.733
JRC-Acquis	Human	n/a	0.247	0.949	0.849	0.672	0.464	0.577	52.092	7.180
	GPT-4o-mini	0.352	0.383	0.952	0.539	0.682	0.593	0.477	82.239	7.347
	LLaMA 3.1-70B	0.390	0.424	0.942	0.883	0.690	0.470	0.512	60.948	6.306
	LLaMA 3.1-8B	0.368	0.427	0.945	0.882	0.673	0.449	0.504	59.209	6.514
MultiUN	Human	n/a	0.255	0.927	0.862	0.592	0.276	0.521	44.466	6.861
	GPT-4o-mini	0.352	0.364	0.968	0.549	0.630	0.528	0.462	76.639	7.347
	LLaMA 3.1-70B	0.402	0.400	0.955	0.903	0.618	0.303	0.476	76.121	6.611
	LLaMA 3.1-8B	0.378	0.443	0.965	0.907	0.608	0.320	0.471	59.683	6.806
Wikipedia	Human	n/a	0.246	0.961	0.810	0.664	0.246	0.527	68.484	6.893
	GPT-4o-mini	0.341	0.332	0.974	0.756	0.693	0.423	0.503	80.633	7.754
	LLaMA 3.1-70B	0.382	0.405	0.968	0.821	0.660	0.299	0.469	59.334	6.551
	LLaMA 3.1-8B	0.379	0.439	0.963	0.839	0.672	0.282	0.463	59.259	5.841
GovReport	Human	n/a	0.226	0.958	0.803	0.639	0.336	0.538	35.157	6.720
	GPT-4o-mini	0.340	0.333	0.978	0.722	0.696	0.538	0.520	81.420	7.280
	LLaMA 3.1-70B	0.407	0.363	0.973	0.870	0.651	0.430	0.512	54.626	6.080
	LLaMA 3.1-8B	0.407	0.353	0.971	0.855	0.620	0.354	0.489	52.231	6.44
Summscreen	Human	n/a	0.243	0.927	0.739	0.532	0.384	0.507	62.003	7.040
	GPT-4o-mini	0.294	0.289	0.984	0.832	0.514	0.346	0.478	60.347	6.627
	LLaMA 3.1-70B	0.390	0.328	0.985	0.849	0.523	0.259	0.458	70.638	6.173
	LLaMA 3.1-8B	0.375	0.373	0.976	0.854	0.526	0.266	0.450	69.116	5.667
Seahorse	Human	n/a	0.213	0.950	0.819	0.651	0.440	0.563	51.057	6.200
	GPT-4o-mini	0.295	0.279	0.985	0.832	0.647	0.556	0.548	67.220	7.613
	LLaMA 3.1-70B	0.352	0.382	0.965	0.842	0.661	0.434	0.504	65.295	6.293
	LLaMA 3.1-8B	0.354	0.369	0.978	0.846	0.649	0.472	0.515	65.893	6.427

Table 13: Performance on the 5% summarization task per dataset.

Dataset	Model	R-L(↑)	REP-3(↓)	CoLA↑	COH-2↑	SH-4↑	SH-5↑	AVG↑	HE↑	Hum↑
LCFO.10%										
LexGLUE	Human	n/a	0.351	0.940	0.829	0.660	0.362	0.544	62.141	7.387
	GPT-4o-mini	0.419	0.494	0.947	0.599	0.633	0.500	0.437	80.094	8.120
	LLaMA 3.1-70B	0.452	0.566	0.942	0.882	0.625	0.397	0.456	59.666	7.080
	LLaMA 3.1-8B	0.439	0.641	0.967	0.876	0.621	0.376	0.440	59.349	7.200
BookSum	Human	n/a	0.278	0.907	0.757	0.610	0.342	0.512	83.776	7.601
	GPT-4o-mini	0.327	0.308	0.978	0.858	0.578	0.442	0.510	94.867	7.062
	LLaMA 3.1-70B	0.427	0.456	0.967	0.835	0.573	0.335	0.451	82.114	6.469
	LLaMA 3.1-8B	0.415	0.511	0.966	0.844	0.551	0.313	0.432	73.681	6.420
SQuALITY	Human	n/a	0.313	0.917	0.773	0.548	0.312	0.497	51.501	6.000
	GPT-4o-mini	0.327	0.329	0.975	0.819	0.525	0.341	0.466	76.441	6.467
	LLaMA 3.1-70B	0.382	0.367	0.974	0.836	0.518	0.320	0.456	46.731	4.613
	LLaMA 3.1-8B	0.373	0.406	0.979	0.856	0.526	0.324	0.456	50.129	5.280
FacetSum	Human	n/a	0.328	0.942	0.840	0.710	0.425	0.570	69.570	7.680
	GPT-4o-mini	0.461	0.409	0.945	0.658	0.666	0.506	0.473	78.381	7.882
	LLaMA 3.1-70B	0.455	0.538	0.934	0.890	0.696	0.527	0.502	63.663	6.501
	LLaMA 3.1-8B	0.449	0.547	0.954	0.892	0.698	0.496	0.499	63.768	6.987
JRC-Acquis	Human	n/a	0.339	0.959	0.845	0.702	0.512	0.590	61.618	7.680
	GPT-4o-mini	0.433	0.479	0.947	0.548	0.668	0.543	0.445	81.398	7.819
	LLaMA 3.1-70B	0.440	0.579	0.922	0.888	0.662	0.455	0.470	52.570	6.542
	LLaMA 3.1-8B	0.443	0.586	0.938	0.859	0.673	0.441	0.465	49.986	7.02
MultiUN	Human	n/a	0.312	0.942	0.871	0.612	0.334	0.539	57.357	7.902
	GPT-4o-mini	0.422	0.455	0.961	0.629	0.621	0.518	0.455	74.459	7.875
	LLaMA 3.1-70B	0.447	0.546	0.912	0.914	0.622	0.329	0.446	52.920	6.903
	LLaMA 3.1-8B	0.446	0.557	0.950	0.900	0.606	0.295	0.439	55.186	7.014
Wikipedia	Human	n/a	0.286	0.969	0.812	0.723	0.286	0.547	78.316	7.747
	GPT-4o-mini	0.388	0.428	0.963	0.640	0.690	0.446	0.462	78.149	7.696
	LLaMA 3.1-70B	0.445	0.557	0.931	0.830	0.670	0.278	0.430	57.310	6.681
	LLaMA 3.1-8B	0.444	0.489	0.963	0.822	0.691	0.322	0.462	58.038	6.246
GovReport	Human	n/a	0.296	0.956	0.815	0.670	0.361	0.548	52.627	6.720
	GPT-4o-mini	0.402	0.420	0.972	0.639	0.683	0.529	0.480	81.374	7.72
	LLaMA 3.1-70B	0.454	0.511	0.923	0.874	0.667	0.406	0.472	49.294	6.57
	LLaMA 3.1-8B	0.459	0.498	0.972	0.871	0.650	0.415	0.482	52.712	6.987
Summscreen	Human	n/a	0.308	0.930	0.732	0.557	0.414	0.514	68.971	7.733
	GPT-4o-mini	0.314	0.364	0.974	0.778	0.511	0.345	0.449	61.149	6.667
	LLaMA 3.1-70B	0.417	0.436	0.984	0.849	0.505	0.281	0.437	49.294	6.413
	LLaMA 3.1-8B	0.402	0.501	0.987	0.855	0.506	0.295	0.428	62.788	6.067
Seahorse	Human	n/a	0.271	0.952	0.811	0.651	0.518	0.576	60.999	7.067
	GPT-4o-mini	0.353	0.353	0.977	0.786	0.632	0.542	0.517	72.321	7.720
	LLaMA 3.1-70B	0.417	0.491	0.963	0.831	0.660	0.480	0.489	60.241	6.440
	LLaMA 3.1-8B	0.402	0.474	0.968	0.839	0.642	0.476	0.490	60.415	7.080

Table 14: Performance on the 10% summarization task per dataset

DATASET	Model	R-L(↑)	REP-3(↓)	COLA↑	COH-2↑	SH-4↑	SH-5↑	AVG↑	HE↑	HUM↑
				LCFO.20%						
LexGLUE	Human	n/a	0.455	0.940	0.842	0.688	0.417	0.559	65.036	7.800
	GPT-4o-mini	0.516		0.5822	0.9446	0.5874	0.6286	0.4820	0.4121	8.093
	LLAMA 3.1-70B	0.507	0.702	0.927	0.860	0.632	0.369	0.417	54.710	7.027
	LLAMA 3.1-8B	0.501	0.824	0.943	0.882	0.637	0.410	0.409	49.870	6.973
BookSum	Human	n/a	0.344	0.918	0.766	0.621	0.349	0.517	89.376	7.605
	GPT-4o-mini	0.355	0.385	0.975	0.831	0.573	0.441	0.487	95.169	7.432
	LLAMA 3.1-70B	0.455	0.544	0.956	0.842	0.511	0.314	0.416	69.068	6.296
	LLAMA 3.1-8B	0.453	0.634	0.971	0.842	0.550	0.394	0.425	76.472	6.457
SQuALITY	Human	n/a	0.395	0.919	0.782	0.565	0.339	0.505	61.257	5.800
	GPT-4o-mini	0.382	0.425	0.969	0.774	0.518	0.328	0.433	79.698	6.720
	LLAMA 3.1-70B	0.412	0.498	0.963	0.797	0.454	0.205	0.384	41.584	4.027
	LLAMA 3.1-8B	0.426	0.601	0.979	0.835	0.469	0.233	0.383	47.011	5.587
FacetSum	Human	n/a	0.415	0.940	0.829	0.745	0.490	0.584	72.317	8.147
	GPT-4o-mini	0.477	0.483	0.953	0.685	0.659	0.526	0.468	80.937	7.711
	LLAMA 3.1-70B	0.474	0.622	0.944	0.899	0.705	0.507	0.487	55.197	6.501
	LLAMA 3.1-8B	0.456	0.565	0.952	0.894	0.698	0.471	0.490	46.081	6.813
JRC-Acquis	Human	n/a	0.435	0.971	0.860	0.713	0.566	0.605	72.317	7.902
	GPT-4o-mini	0.513	0.566	0.952	0.551	0.685	0.578	0.440	75.351	7.681
	LLAMA 3.1-70B	0.493	0.792	0.854	0.884	0.648	0.422	0.403	38.876	6.653
	LLAMA 3.1-8B	0.490	0.788	0.929	0.882	0.633	0.446	0.420	49.870	6.833
MultiUN	Human	n/a	0.422	0.942	0.875	0.605	0.317	0.531	64.540	8.139
	GPT-4o-mini	0.484	0.604	0.954	0.623	0.615	0.482	0.414	72.007	7.917
	LLAMA 3.1-70B	0.483	0.654	0.907	0.918	0.625	0.289	0.417	37.311	6.694
	LLAMA 3.1-8B	0.512	0.665	0.925	0.908	0.603	0.326	0.419	45.155	7.028
Wikipedia	Human	n/a	0.355	0.967	0.817	0.738	0.333	0.557	81.923	7.653
	GPT-4o-mini	0.471	0.516	0.960	0.596	0.687	0.432	0.432	76.772	7.609
	LLAMA 3.1-70B	0.467	0.779	0.877	0.849	0.638	0.300	0.377	55.349	6.493
	LLAMA 3.1-8B	0.476	0.701	0.940	0.786	0.606	0.257	0.378	51.110	6.014
GovReport	Human	n/a	0.397	0.954	0.823	0.712	0.405	0.563	60.368	8.027
	GPT-4o-mini	0.488	0.521	0.968	0.605	0.684	0.521	0.451	76.887	7.680
	LLAMA 3.1-70B	0.489	0.638	0.916	0.872	0.634	0.425	0.442	43.421	6.347
	LLAMA 3.1-8B	0.479	0.531	0.971	0.881	0.623	0.384	0.466	40.544	7.093
Summscreen	Human	n/a	0.395	0.939	0.739	0.552	0.414	0.513	63.691	8.373
	GPT-4o-mini	0.347	0.443	0.969	0.756	0.503	0.346	0.426	60.306	6.200
	LLAMA 3.1-70B	0.432	0.487	0.979	0.845	0.502	0.294	0.426	60.564	6.627
	LLAMA 3.1-8B	0.432	0.519	0.987	0.851	0.522	0.322	0.433	49.467	6.413
Seahorse	Human	n/a	0.336	0.964	0.828	0.673	0.533	0.586	66.028	7.907
	GPT-4o-mini	0.415	0.439	0.970	0.722	0.607	0.507	0.474	71.568	8.173
	LLAMA 3.1-70B	0.454	0.589	0.957	0.834	0.615	0.441	0.452	54.075	6.600
	LLAMA 3.1-8B	0.469	0.642	0.960	0.847	0.601	0.456	0.444	54.142	6.760

Table 15: Performance on the 20% summarization task per dataset.