PELMS: Pre-training for Effective Low-Shot Multi-Document Summarization

Joseph J. Peper, Wenzhao Qiu, and **Lu Wang**
Computer Science and Engineering
University of Michigan
Ann Arbor, MI
{jpeper, qwzhao, wangluxy}@umich.edu

Abstract

We investigate pre-training techniques for abstractive multi-document summarization (MDS), which is much less studied than summarizing single documents. Though recent work has demonstrated the effectiveness of highlighting information salience for pre-training strategy design, they struggle to generate abstractive and reflective summaries, which are critical properties for MDS. To this end, we present **PELMS**, a pre-trained model that uses pre-training objectives based on semantic coherence heuristics and faithfulness constraints together with unlabeled multi-document inputs, to promote the generation of concise, fluent, and faithful summaries. To support the training of PELMS, we compile **MultiPT**, a multi-document pre-training corpus containing over 93 million documents to form more than 3 million unlabeled topic-centric document clusters, covering diverse genres such as product reviews, news, and general knowledge. We perform extensive evaluation of PELMS in low-shot settings on a wide range of MDS datasets. Our approach consistently outperforms competitive comparisons with respect to overall informativeness, abstractiveness, coherence, and faithfulness, and with minimal fine-tuning can match performance of language models at a much larger scale (e.g., GPT-4).

1 Introduction

Abstractive multi-document summarization (MDS) aims to generate coherent and concise summaries from sets of related documents. While transformer-based models have excelled in single-document summarization (Zhang et al., 2019a), their application to MDS often results in suboptimal cross-document salience detection and information aggregation. MDS-specific pre-trained models like Primera (Xiao et al., 2022) and Centrum (Puduppully et al., 2023) use information frequency to improve salience estimation, however, the quality

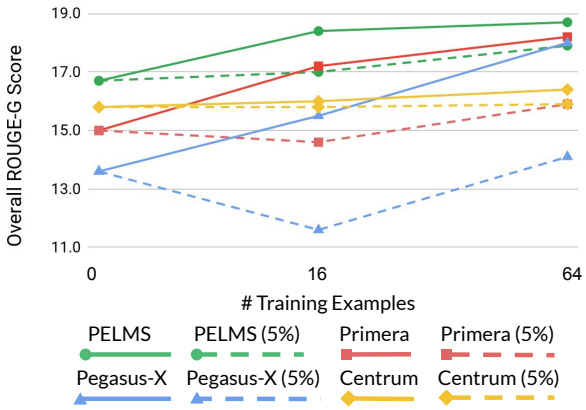

Figure 1: Comparison of long-input pre-trained model performance in the low-shot setting. We report ROUGE-G scores (geometric mean of ROUGE-1/2/L) aggregated over 6 MDS datasets. We evaluate fully-supervised training (solid lines) and adapter training (dashed lines) where only 5% of parameters are tuned. PELMS achieves stronger overall performance at all data quantities and training methods.

of such pre-training objectives is often reliant on brittle topic alignment methods (such as n-gram overlap, or cross-document entity linking) which struggle to generalize to an open set of domains, yielding outputs with with subpar summary *informativeness*. Existing MDS methods also typically overlook multi-document *faithfulness*, generating outputs that poorly represent the full input, particularly in domains requiring cross-document synthesis. Similarly, summary *coherence* is inadequately addressed, with most gap-sentence generation (GSG) style objectives simply denoising masked phrases in their original ordering to form the training target; these approaches may work well in conventional single-document settings, however they can introduce positional biases (DeYoung et al., 2023) in the multi-document setting due to the arbitrary ordering of documents within the input, impacting both *coherence* and *informativeness*.

While recent powerful large language models (LLMs) have demonstrated strong performance on

tasks including summarizaton, they do not fully address the need for efficient, pre-trained models for MDS, particularly for common real-world scenarios with domain-specific requirements, regulatory and data privacy concerns, or tight computational constraints. With these considerations in mind, we seek to *rapidly enable proficient MDS performance, generating high-quality summaries with minimal labeled data and/or compute requirements.*

We propose **PELMS**, a method of Pre-training for Effective Low-Shot Multi-document Summarization that promotes informativeness over a diverse range of text genres, ranging from consumer and editorial opinion, to news articles, and to scientific papers, while also improving abstractiveness, faithfulness, and coherence of summaries.[1] Our method leverages semantic clustering for sentence-level salience detection, overcoming the limitations of previous approaches reliant on lexical similarity. During pre-training, we generate summaries that balance topical *salience* with *source consistency*, while also maintaining strong abstractiveness and coherence by removing and intentionally ordering target sentences, a departure from traditional GSG methods. To support MDS pre-training, we introduce **MultiPT**, a comprehensive dataset with over 3 million topic-aligned document clusters from more than 93 million documents. We evaluate PELMS across six tasks, including the newly introduced MetaTomatoes dataset, which presents a complex task of meta-summarizing diverse movie reviews. Our results demonstrate the strong MDS performance of PELMS, particularly in low-shot settings, highlighting the alignment between our pre-training strategy and multi-document summarzation. Concretely,

- In zero-shot setups, PELMS *improves summary informativeness* (ROUGE and BertScore) over previous state-of-the-art MDS models, *while also enhancing faithfulness, abstractiveness, and coherence*. We see particular improvements in review domains that require extensive cross-document synthesis. Human judges further validate that our methods yield summaries with improved grammaticality, referential clarity, coherence, and faithfulness. We also incorporate length control in pre-training, yielding gains in zero-shot settings.

[1]Code and model are available at https://github.com/jpeper/pelms.

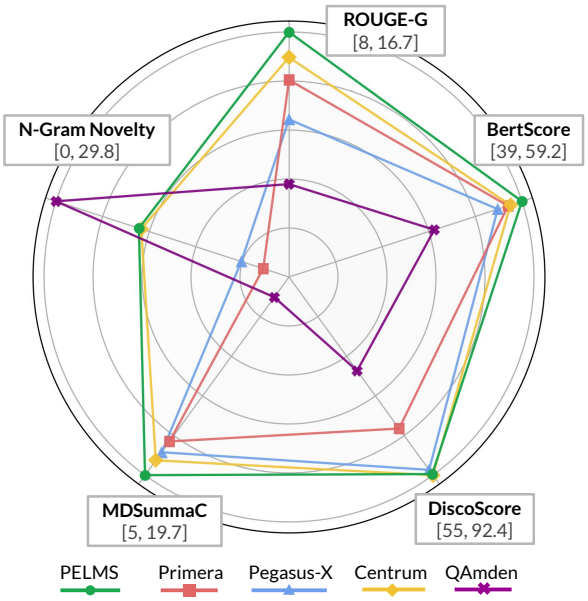

Figure 2: Zero-shot performance across **relevance** (ROUGE-G, BertScore), **abstractiveness** (N-gram Novelty), **coherence** (DiscoScore), and **faithfulness** (MDSummaC) metrics. Each metric is displayed on a unique scale, with $[a, b]$ respectively indicating the minimum and maximum values for each axis. PELMS achieves the best combination of performance over all key summary characteristics.

- In supervised scenarios, PELMS outperforms existing models in ROUGE and BertScore, while striking the best balance of faithfulness, abstractiveness, and coherence. We find full-parameter training improves informativeness and abstractiveness, whereas adapter-based training best maintains input faithfulness. Remarkably, we find *PELMS can match GPT-3.5 and GPT-4.0* (Ouyang et al., 2022; OpenAI, 2023) *performance with as few as 16 and 64 training examples*, respectively.

- Ablation studies reveal the significance of both our PELMS technique and the MultiPT dataset in enhancing MDS performance. PELMS proves effective across various model architectures, and existing models benefit from training on our dataset.

2 Related Work

2.1 Multi-Document Summarization

Multi-document summarization (MDS) involves summarizing extensive, varied content, often including a large number of articles. This task requires systems capable of efficiently handling inputs with complementary, redundant, or contradictory information (Otmakhova et al., 2022; Pasunuru

et al., 2021; Brazinskas et al., 2022; Hendrickx et al., 2009; Radev, 2000; Wolhandler et al., 2022). A key hurdle in MDS is the scarcity of extensive multi-document pre-training corpora, leading to reliance on small or synthesized corpora (Caciularu et al., 2021, 2023).

2.2 Summarization Pre-training

Pre-trained language models (PLMs) have become the dominant paradigm in natural language processing for tasks including abstractive text summarization. The popular Pegasus (Zhang et al., 2019a) uses gap sentence generation (GSG) in pre-training, masking key sentences in single documents based on lexical similarity. However, this can result in repetitive outputs when applied to MDS inputs.

In multi-document contexts, Primera (Xiao et al., 2022) builds on Pegasus using entity-based sentence grouping to reduce redundancy. However, it struggles with effectively linking semantically similar terms. Centrum (Puduppully et al., 2023) takes a different approach by selecting centroid documents as summary targets but may not fully represent diverse topics or opinions. Cross-document alignment via question answering has also been explored (Caciularu et al., 2023), showing promise in supervised settings.

Other works focus on improving targeted modeling improvements. For example, Wan and Bansal (2022) enhance Pegasus pre-training by incorporating factuality constraints, although these are shown to generalize poorly across different domains (Kryściński et al., 2019). Other works have explored improvements for specific summarization tasks such as opinion and dialog summarization (Brazinskas et al., 2022; Li et al., 2022).

Our method seeks to achieve effective MDS pre-training over a wide variety of domains, maintaining robust performance on not only conventional summary metrics such as ROUGE, but also on other key attributes such as abstractiveness, coherence, and factuality.

3 PELMS Pre-training Technique

The PELMS pre-training strategy for multi-document summarization is outlined in Figure 3. Briefly, we rely on a cluster-then-select pre-training objective to generate data for training transformer models. We then follow previous work by using a Gap Sentence Generation-style objective to form pre-training targets, but instead of masking, we remove the sentences to improve abstractiveness. Our key contributions are: 1) improved ranking and selection of target sentence candidates to encourage summary informativeness and faithfulness, and 2) injecting coherence constraints during formulation of the GSG target. Our technique consists of the following three steps:

Dataset	Domain	#Clusters	#Docs	Doc Len	Input Len
NewSHead	News	370,000	3.5	495.2	1,732
BigNews	News	1,060,000	4.3	656.1	2,820
WikiSum-40	Wiki	1,000,000	40	70.1	2,804
AmazonPT	Product Reviews	1,000,000	40.3	72.8	2,901
YelpPT	Business Reviews	142,000	61.3	60.6	3,714

Table 1: Overview of the MultiPT pre-training corpus. MultiPT is comprised of 5 sources of unlabeled topic-centric document clusters. Combined, they sum to over 3.5M clusters. We release MultiPT along with pre-computed sentence-level embeddings and entities.

1. *Clustering sentences into topics (§3.1).* We encode and cluster the input sentences to identify prevalent topics within the input, considering the top k topics for inclusion in the summary.

2. *Ranking cluster sentences by summary-worthiness (§3.2).* We score each cluster element based on a) distance to the cluster centroid and b) entailment-based intra-cluster consistency. These rankings determine which sentences are used in the pre-training target.

3. *Selecting and ordering target sentences to maintain summary coherence (§3.3).* Considering the c highest-scoring examples from each topic cluster, we select target sentences (one per topic), sourcing from the fewest number of documents possible. We use this, and a topic-position ordering heuristic, to specify the output ordering.

3.1 Topic Detection via Sentence Clustering

Information redundancy is a common phenomenon within multi-document inputs (Ma et al., 2022) and has been utilized to identify input salience, with the intuition being that topic frequency correlates with topic significance (Xiao et al., 2022; Nenkova et al., 2007). Methods like Primera use ROUGE similarity or align sentences using entity mentions. However, these are brittle and generalize poorly, motivating the need for a more refined selection mechanism. We propose to use **continuous semantic representations** when performing the sentence similarity comparison. Concretely,

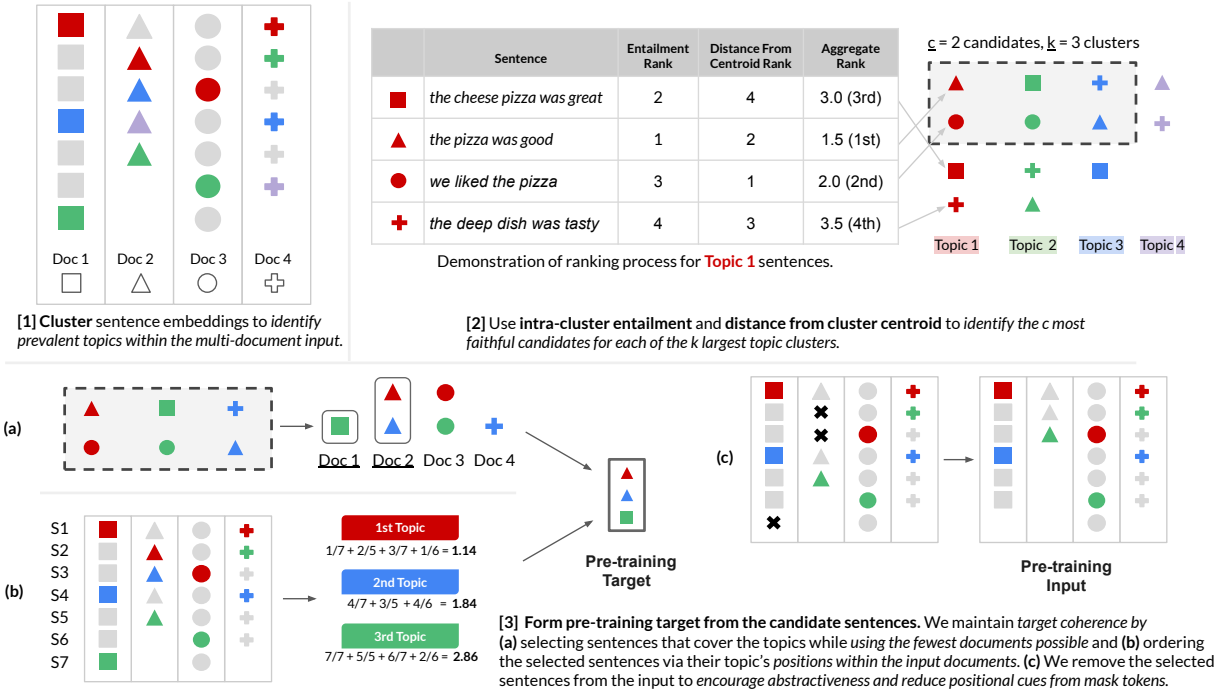

Figure 3: Overview of the PELMS pre-training approach. The pre-training target is formed by selecting a set of representative sentences covering frequently occurring topic clusters within the input. Target sentences are intentionally selected to enhance **faithfulness** and arranged to improve target **coherence**.

PELMS embeds and clusters the input sentences, with each cluster representing a set of topic-aligned sentences. We leverage the Sentence Transformers library (Reimers and Gurevych, 2019) which offers *lightweight* text embeddings and a fast local-community clustering method, which are required for processing large-scale pre-training data in a tractable fashion. Once we have identified semantic clusters, we use this structured cluster representation to identify salient topics. Similar to the Entity Pyramid method (Xiao et al., 2022), we use frequency as a proxy for salience. Concretely, larger clusters represent topics that are more prominent within the input, and are therefore more summary-worthy. We select from the k largest clusters. See Appendix A for details on the clustering method and parameters.

3.2 Entailment-aware Target Sentence Selection

As each top-k cluster represents a unique summary-worthy topic, we must choose a representative sentence from each cluster for inclusion within the pre-training target. **To improve the consistency of our selected sentence with the rest of the cluster**, we use a combination of cluster centrality and intra-cluster entailment to score the candidate sentences (Fig. 3-[2]). Pegasus and Primera leverage simple lexical overlap to identify the most significant sentence. Similarly, we could consider simply selecting the medoid element within each topic cluster. However, methods such as Wan and Bansal (2022) find it possible to improve the faithfulness of summarization systems by imposing additional selection constraints. We rank the candidate sentences in two ways: 1) by distance to the cluster centroid, to capture the average semantic meaning of the cluster, and 2) by average NLI entailment with the rest of the cluster elements, as a means of maintaining input-consistent summary examples. We aggregate these two rankings using a simple unweighted Borda count (Emerson, 2013) to generate a ranking of all cluster elements. We cap the NLI calculation to the 5 most central cluster elements due to the quadratic computational complexity of the intra-cluster entailment score.

3.3 Pre-training Formulation for Improved Coherence

Coherence is a key characteristic for summaries and has been extensively studied for summarization tasks including MDS (Christensen et al., 2013; Wang et al., 2016). However, GSG simply masks the highest-scoring sentences and de-noises in order of appearance within the input, which can result in incoherent outputs for arbitrarily-ordered

multi-document inputs. With our method, after identifying summary-worthy sentences, we then **select and order a set of sentences to comprise the pre-training target**.

Considering only the c highest-ranked sentences from each topic cluster, we select target sentences, one per topic, using minimum set cover to source from as few documents as possible to improve target coherence (Fig. 3-[3a]).

After selecting target sentences, we order them subject to the following constraints in this order of precedence: 1) sentences selected from the same document should maintain their original relative ordering, and 2) sentences should be ordered by average topic position – e.g., 'lead' topics should appear early within the target (Fig. 3-[3b]). Finally we remove the selected sentences from the pre-training input (Fig. 3-[3c]).

4 Pre-training Details

We pre-train a new MDS model with the PELMS technique, forming pre-training examples from the unlabeled document clusters in our MultiPT corpora. Below we overview MultiPT and the pre-training architecture used for training PELMS. See Appendix A.2 for full pre-training details.

MultiPT Pre-training Dataset MDS is designed to support lengthy multi-document inputs, yet there are few available data sources for large-scale unlabeled multi-documents for pre-training. Notably, NewsHead (Gu et al., 2020) has been used, containing clusters of news articles. However, it is limited in magnitude for a pre-training dataset, containing only 370k document clusters.

Extending on this, we compile MultiPT, a new multi-document dataset comprised of over 3 million document clusters from public data sources. It contains a wide diversity of genres (including news, general knowledge, and opinionated content), and covers a broad range of inputs with respect to document lengths and cluster sizes. Table 1 outlines the pre-training dataset. During MultiPT pre-processing, we use simple text overlap heuristics to filter out any documents that exist within the evaluation datasets.

Base Model As done in Primera, we use the popular sparse-attention long-input Longformer Encoder Decoder (Beltagy et al., 2020) as our base architecture, initializing from the 464M LED-large base model. We follow Primera in inserting a global `<doc-sep>` token after each input document for improved cross-document communication (Caciularu et al., 2021).

5 Evaluation Setup

5.1 Evaluation Datasets

We use five existing MDS datasets spanning news, opinion, and scientific domains in our evaluation, and additionally curate **MetaTomatoes**, a meta-summary generation dataset for critics' film reviews. Table 6 overviews the evaluation dataset statistics. Appendix C.2 provides further dataset details.

5.2 Evaluation Metrics

Previous analysis of MDS techniques has largely focused on ROUGE evaluation, supplemented by expensive human evaluation. In this work, we are the first to systematically explore the behavior of pre-trained MDS models across a wide variety of summary evaluation metrics. We evaluate summary *informativeness* with **ROUGE** and **BertS**core, *coherence* with **DiscoS**core, *faithfulness* with our new **MDSummaC** metric, and *abstractiveness* with **N-gram Novelty**. Full evaluation metrics details are covered in Appendix E.

5.3 Baseline Models

We compare our PELMS with four leading long-input pre-trained summarization models in zero-shot and supervised scenarios:

Pegasus-X (Phang et al., 2022) extends the 568M Pegasus model (Zhang et al., 2019a) to long-input tasks by continued long-input pre-training. It incorporates block-staggered local attention for efficient processing of long inputs and has demonstrated superior performance on long-input tasks compared to models like LongT5 (Guo et al., 2022).

QAMDen (Caciularu et al., 2023), also a 464M model initialized from LED-large, leverages the Primera-style sparse attention. Its pre-training consits of multi-document question answering, using silver-labeled question-context-answer tuples automatically derived from NewSHead document clusters.

Primera (Xiao et al., 2022) and **Centrum** (Puduppully et al., 2023) are both 464M models based on the LED-large architecture. Primera is trained on NewSHead (Gu et al., 2020) with an entity salience-based pre-training objective. Namely,

Dataset	Model	R1	R2	RL	RG	BertS	DiscoS	MDSummaC	N-gram Novelty
MultiNews	Pegasus-X	39.1	11.2	17.8	19.8	59.2	91.5	37.8	4.6
	Primera	41.9	12.7	19.5	21.8	60.8	93.1	**41.3**	3.6
	QAmden	31.1	7.9	15.7	15.6	49.9	63.5	5.3	**40.2**
	Centrum	**44.2**	**15.1**	**21.6**	**24.3**	**62.2**	**95.7**	39.2	11.5
	PELMS	41.7	13.5	20.0	22.4	61.0	95.3	38.6	7.8
Multi-XScience	Pegasus-X	28.8	4.5	14.9	12.5	55.7	89.8	30.5	3.3
	Primera	29.6	4.6	15.2	12.8	55.5	81.7	27.0	1.6
	QAmden	25.4	3.5	14.3	10.9	51.9	71.9	9.6	**22.8**
	Centrum	29.2	4.6	15.3	12.7	55.2	90.8	**31.5**	7.2
	PELMS	**30.0**	**4.8**	**15.4**	**13.0**	**56.0**	**90.9**	31.1	3.0
Amazon	Pegasus-X	27.2	4.2	15.1	12.0	58.0	**90.8**	**15.3**	5.1
	Primera	28.1	5.1	16.4	13.3	59.0	84.1	8.1	2.5
	QAmden	24.1	3.3	14.5	10.5	54.9	77.2	6.2	16.2
	Centrum	28.9	4.8	16.8	13.3	58.5	90.4	12.3	25.3
	PELMS	**31.3**	**7.1**	**18.7**	**16.1**	**62.0**	90.6	12.8	**30.8**
Yelp	Pegasus-X	25.0	3.9	14.3	11.2	59.0	89.9	**14.5**	6.9
	Primera	27.4	4.9	15.7	12.8	59.8	85.2	10.7	2.7
	QAmden	21.2	3.3	13.2	9.7	54.8	75.7	6.9	17.5
	Centrum	27.4	5.4	16.7	13.5	58.8	89.9	9.8	**39.1**
	PELMS	**31.7**	**8.0**	**19.0**	**16.9**	**62.4**	**91.6**	14.2	24.9
DUC2004	Pegasus-X	31.5	5.0	15.5	13.5	56.4	92.8	4.6	4.1
	Primera	32.9	7.0	16.9	15.8	58.1	77.1	12.6	2.3
	QAmden	28.2	4.3	15.3	12.3	53.8	75.7	8.9	**12.1**
	Centrum	**34.7**	**7.9**	**18.2**	**17.1**	**59.3**	**94.0**	13.8	3.4
	PELMS	34.0	6.6	16.7	15.6	58.8	93.0	**15.4**	7.0
MetaTomatoes	Pegasus-X	31.1	4.6	13.8	12.5	54.4	93.4	5.4	12.5
	Primera	32.1	5.0	14.4	13.2	54.9	80.5	3.5	7.1
	QAmden	21.6	2.6	11.6	8.6	44.5	72.3	1.9	**70.2**
	Centrum	32.5	5.6	15.0	13.9	54.9	**93.4**	4.7	26.9
	PELMS	**34.1**	**7.4**	**16.5**	**16.1**	**55.1**	92.0	**6.2**	41.4
Overall	Pegasus-X	30.4	5.6	15.2	13.6	57.1	91.4	18.0	6.1
	Primera	32.0	6.6	16.3	15.0	58.0	83.6	17.2	3.3
	QAmden	25.3	4.1	14.1	11.3	51.6	72.7	6.5	**29.8**
	Centrum	32.8	7.2	17.3	15.8	58.2	**92.4**	18.6	18.9
	PELMS	**33.8**	**7.9**	**17.7**	**16.7**	**59.2**	92.2	**19.7**	19.2

Table 2: Zero-shot MDS results. **Bold** and underline respectively indicate the best and second-best model. Green indicates where PELMS is the best model. PELMS achieves strong informativeness (ROUGE, BertScore) and coherence (DiscoScore) while maintaining the best combination of faithfulness (MDSummaC) and abstractiveness (N-gram Novelty).

it performs entity extraction to identify common topics, and uses a ROUGE consistency metric to select one representative sentence for each frequently-occuring entity. Centrum, using the same base model and pre-training data, instead employs a "leave-one-out" training approach, selecting a ROUGE-calculated centroid document as the target summary, treating the task as document-granularity denoising.

5.4 Supervised Fine-tuning Methods

In our supervised experiments, we train the models using labeled MDS data. We perform standard training, updating all model parameters during training. Results are averaged over 5 runs, each with unique random seeds. We also perform parameter-efficient fine-tuning (PEFT) to understand whether the evaluated models can converge when updating fewer model parameters; we use the popular Adapter (Houlsby et al., 2019) method which freezes the original model weights and trains only a small percentage of new parameters.

6 Results

We perform a rigorous comparison of PELMS and competitive baseline models in both **zero-shot** and **supervised** settings (using both full supervision and PEFT adapter training). Appendix B outlines the full details of our zero-shot and supervised evaluation configuration.

6.1 Zero-Shot Evaluation

In the zero-shot evaluation (Table 2), PELMS consistently excels, surpassing baselines on key metrics. Table 8 displays example zero-shot outputs from each model. The PELMS performance underscores the alignment of our pre-training objective with the MDS (Multi-Document Summariza-

# Shots	Model	R1	R2	RL	RG	BertS	DiscoS	MDSummaC	N-gram Novelty
0	Pegasus-X	30.4	5.6	15.2	13.6	57.1	91.4	18.0	6.1
	Primera	32.0	6.6	16.3	15.0	58.0	83.6	17.2	3.3
	QAmden	25.3	4.1	14.1	11.3	51.6	72.7	6.5	29.8
	Centrum	32.8	7.2	17.3	15.8	58.2	**92.4**	18.6	18.9
	PELMS	**33.8**	**7.9**	**17.7**	**16.7**	**59.2**	92.2	**19.7**	19.2
16	Pegasus-X (5%)	26.0	4.5	14.0	11.6	56.8	83.9	18.0	5.1
	Primera (5%)	31.1	6.4	16.2	14.6	58.2	86.7	20.2	2.3
	QAmden (5%)	19.0	3.1	12.1	8.8	53.1	68.2	17.4	17.4
	Centrum (5%)	32.6	7.3	17.2	15.8	58.1	**92.3**	18.7	**18.7**
	PELMS (5%)	**34.0**	**8.1**	**18.1**	**17.0**	**60.0**	**92.3**	**20.4**	18.1
	Pegasus-X	31.9	6.9	17.4	15.5	60.5	89.1	13.0	**34.1**
	Primera	34.5	8.1	18.7	17.2	61.1	92.4	18.5	24.0
	QAmden	19.8	3.3	12.6	9.3	53.3	68.4	17.4	17.0
	Centrum	33.0	7.4	17.4	16.0	58.4	**92.7**	**19.2**	16.2
	PELMS	**36.0**	**9.0**	**19.5**	**18.4**	**61.6**	92.6	16.5	30.1
64	Pegasus-X (5%)	29.3	6.0	16.1	14.1	58.5	86.5	17.6	11.4
	Primera (5%)	32.7	7.4	17.3	15.9	59.5	90.2	**21.0**	7.6
	QAmden (5%)	19.5	3.2	12.4	9.2	53.3	68.4	17.3	17.0
	Centrum (5%)	32.7	7.4	17.3	15.9	58.4	**92.6**	19.1	16.7
	PELMS (5%)	**35.0**	**8.8**	**19.0**	**17.9**	**60.9**	92.5	20.9	**18.4**
	Pegasus-X	35.4	8.6	19.4	18.0	62.2	92.0	10.2	**46.1**
	Primera	35.8	8.8	19.3	18.2	61.8	**93.0**	16.2	31.0
	QAmden	29.6	7.1	17.3	15.3	59.0	83.2	17.6	20.4
	Centrum	33.5	7.7	17.8	16.4	59.3	92.7	**20.0**	12.2
	PELMS	**36.2**	**9.3**	**19.8**	**18.7**	**62.5**	92.3	15.5	34.4
Full	Pegasus-X (5%)	31.4	7.3	16.7	15.4	59.2	89	15.4	18.2
	Primera (5%)	32.2	7.5	17.2	15.9	59.5	88.2	18.9	12.5
	QAmden (5%)	27.6	6.2	15.6	13.6	57.4	80.8	16.9	17.3
	Centrum (5%)	33	7.8	17.7	16.4	59	92	18.4	**21.5**
	PELMS (5%)	**34.5**	**8.4**	**18.4**	**17.4**	**60.3**	**92.1**	**19.8**	19.4
	Pegasus-X	36.6	9.9	20.1	19.2	62.4	92.1	13.0	**31.7**
	Primera	37.1	10	20.4	19.4	62.6	**93.3**	16.6	30.7
	QAmden	36.6	9.9	20.2	19.2	62.2	91.9	16.1	28.4
	Centrum	35	9.0	19.0	18.0	60.7	92.8	16.7	23.4
	PELMS	**37.6**	**10.5**	**20.8**	**20.0**	**63.3**	93.0	**17.0**	31.0
LLM	GPT-3.5-Long	34.6	7.6	18.1	16.8	61.2	91.5	13.6	50.9
	GPT-4	35.2	7.9	18.4	17.1	61.8	92.0	11.6	52.1

Table 3: Overall results on zero-shot and supervised splits, averaged over all 6 datasets. We explore both adapter fine-tuning (updating only 5% of parameters) and full-parameter tuning. Blue indicates our model is best for a given data and tuning method split. Values in green indicate our model achieves top performance *overall* for its data quantity split. We observe most metrics improve with more data, although often at the expense of faithfulness (MDSummaC). Other than N-Gram novelty, we see 16-shot PELMS and 64-shot PELMS respectively outperform GPT-3.5-Long and GPT-4 over all metrics.

tion) task. Notably, we achieve superior results in ROUGE-G (geometric mean of ROUGE-1/2/L) and BertScore compared to Primera and Pegasus-X, indicating better content representation and semantic understanding.

PELMS's enhanced coherence (DiscoScore) and faithfulness (MDSummaC) are particularly evident, along with its ability to generate more abstract summaries. This indicates a well-rounded proficiency in summarization, balancing content fidelity and novel synthesis effectively.

In comparison, QAmden, with its QA-centric pre-training, struggles in zero-shot settings, producing incoherent outputs. Its design is less suited for summarization without specific training, highlighting the importance of task-aligned pre-training objectives.

Centrum, although a strong contender, especially in news domains, shows mixed results in other areas like Multi-XScience and opinion summarization datasets. While it matches PELMS in terms of coherence due to its approach of selecting entire documents as summaries, this hampers faithfulness, unlike our sentence-level selection which offers greater flexibility across diverse domains.

Overall, PELMS's robust performance across various metrics and domains underscores the effectiveness of its pre-training objective, particularly in settings that deviate from traditional news-focused summarization tasks – a domain requiring little cross-document synthesis (Wolhandler et al., 2022). This adaptability is crucial for comprehensive and

reliable multi-document summarization.

6.2 Supervised Learning Evaluation

In the supervised analysis, we explore both full parameter tuning and parameter-sparse updates using adapters (training only 5% of model parameters). We evaluate on 16-shot, 64-shot, and full-shot data quantities. All experiments are averaged over 5 runs with unique seeds. Table 3 shows overall results averaged over the six datasets. Per-dataset results w/ statistical significance tests are provided in Appendix H. We see PELMS demonstrates notable superiority across all different training data scales and tuning configurations.

Full-parameter Tuning With increasing supervision, PELMS consistently leads in ROUGE-G, showing notable gains even in small data scenarios like the 16-shot split (1.2+ point improvement). This trend continues with larger data quantities, maintaining a steady edge in both ROUGE and BertScore metrics over competitors like Primera.

While Pegasus-X shows high abstractiveness, it lags in faithfulness, highlighting a trade-off in summary characteristics. QAmden, although initially underperforming in few-shot settings, gradually matches other methods with full supervision, suggesting its potential in data-rich environments though indicating a misalignment between pre-training and the MDS setting. Centrum, despite its strengths, shows only modest improvements with supervision, lagging behind PELMS, particularly in ROUGE and BertScore.

Adapter Fine-tuning PELMS's performance is especially pronounced in adapter fine-tuning scenarios. It significantly surpasses others in ROUGE-G across all data scenarios, from 16-shot to full-shot, demonstrating its capacity to adapt with limited trainable parameters. This is a key strength for resource-constrained environments.

While all models generally converge to lower ROUGE and BertScore values with adapter training, PELMS maintains a balanced profile, preserving both abstractiveness and faithfulness. This balance is critical for practical summarization applications where both abstractiveness and accuracy are important.

In summary, PELMS excels in both fully-supervised and PEFT contexts, achieving top summary quality and effectively adapting to varying training conditions. This performance demonstrates PELMS as a versatile method for multi-document summarization tasks, capable of handling both data and compute constraints effectively.

7 Further Analyses

Comparison with LLMs We include comparison of MDS models versus OpenAI GPT-3.5 and GPT-4 (Ouyang et al., 2022; OpenAI, 2023) in Table 3. We see that PELMS *achieves comparable or better overall performance in as few as 16 shots versus GPT-3.5-Long and in 64 shots versus GPT-4 for the ROUGE, BertScore and DiscoScore metrics.* We observe abstractiveness (N-gram Novelty) is very high for the GPT, although this is contrasted with relatively poor MDSummaC faithfulness. Appendix G contains the full experiment details and per-dataset results.

Pre-training Ablation We investigate the individual benefits of both our MultiPT data and PELMS pre-training objective on other models (Table 4). We find Primera benefits from pre-training on our diverse large-scale dataset (versus on NewSHead). We also see our method extends well to an alternate architecture (Pegasus-X), outperforming the LED architecture model on most metrics, although faithfulness is decreased.

Human Evaluation We validate our results through human evaluation, following Xiao et al. (2022) in evaluating both summary fluency and and faithfulness. We analyze a random 25-example subset of MetaTomatoes due to the time-consuming nature of the scoring process, particularly for faithfulness evaluation which requires extensive sentence-level input vs. output comparison. PELMS demonstrates improved grammaticality, referential clarity, and coherence compared to the top competitive MDS baselines. We also find PELMS has highest input vs output faithfulness, generating summaries that were the most reflective of their inputs. Appendix F contains the full human evaluation details.

Length Control Experiment We briefly explore length control during pre-training, varying the k value which sets the number of target sentences and training with a corresponding a length-prefix. We achieve an average ROUGE-G improvement of 0.6 points. Further details and results of this experiment can be found in Appendix D.

Architecture	Technique	w/ MultiPT	RG	BertScore	DiscoScore	MDSummaC	N-gram Novelty
LED	Primera	No	15.0	58.0	91.4	18.0	3.3
LED	Primera	Yes	15.4	58.4	91.5	**20.7**	4.6
LED	PELMS	Yes	**16.7**	**59.2**	**92.2**	19.7	**19.2**
Pegasus-X	Pegasus-X	No	13.6	57.1	91.4	**18.0**	6.1
Pegasus-X	PELMS	Yes	**17.3**	**59.8**	**92.8**	17.6	**22.6**

Table 4: We pre-train with the Primera technique using our MultiPT data, and pre-train our PELMS technique with the Pegasus-X architecture (initializing our model with Pegasus-X weights). We report the overall zero-shot results.

	MetaTomatoes (25 Examples)			
	Gram. ↓	Ref. ↓	Str. & Coher ↓	Faithf. ↑
Pegasus-X	1.84	1.88	1.86	65.9
Primera	1.76	1.72	1.94	58.4
Centrum	1.43	1.73	1.77	59.8
PELMS	**1.32**	**1.43***	**1.58***	**78.3***

Table 5: Human evaluation on MetaTomatoes dataset. For Grammaticality, Referential Clarity, and Structure & Coherence, we report the average ranking when comparing methods. Ties are allowed. For Faithfulness, we report the percentage of system-generated SCUs that are holistically entailed by human-verified input SCUs. *: Results are statistically significant for Referential Clarity, Structure & Coherence, and Faithfulness ($p < 0.05$).

8 Conclusion

We introduce PELMS, a novel new multi-document summarization pre-training method. PELMS leverages semantic topic clustering to improve summary informativeness and uses coherence and faithfulness constraints during pre-training target formulation to produce concise, fluent, and faithful summaries. Both automatic evaluation and human evaluation demonstrate that PELMS achieves consistent improvements over competitive baseline pre-trained models, yielding especially strong performance in low-shot and parameter-efficient training settings.

Limitations and Risks

As our emphasis was on building a proficient pre-trained model, our technique is complementary to fine-tuning-specific methods that attempt to further inject specific summarization constraints during fine-tuning. For example, inference-time or train-time techniques such as FactPEGASUS (Wan and Bansal, 2022) and CLIFF (Cao and Wang, 2021) can still be used during supervised fine-tuning with our base model. While we are not aware of any significant risks associated with PELMS, there is always a risk that biases in the pre-training data, such as MultiPT, could lead to the model inadvertently generating summaries with skewed perspectives or unintended biases. Additionally, the model's ability to create fluent summaries could potentially be misused for spreading misinformation if applied in manipulative contexts.

Acknowledgements

This work is supported in part by the National Science Foundation through grants CMMI-2050130 and IIS-2046016. We thank the reviewers for their valuable comments.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-shot learning for opinion summarization. *arXiv preprint arXiv:2004.14884*.

Arthur Brazinskas, Ramesh Nallapati, Mohit Bansal, and Markus Dreyer. 2022. Efficient few-shot fine-tuning for opinion summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1509–1523, Seattle, United States. Association for Computational Linguistics.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E. Peters, Arie Cattan, and Ido Dagan. 2021. Cross-document language modeling. *CoRR*, abs/2101.00406.

Avi Caciularu, Matthew Peters, Jacob Goldberger, Ido Dagan, and Arman Cohan. 2023. Peek across: Improving multi-document modeling via cross-document question-answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1989, Toronto, Canada. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards coherent multi-document summarization. In *Proceedings of the 2013 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, pages 1163–1173, Atlanta, Georgia. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12. Citeseer.

Jay DeYoung, Stephanie C Martinez, Iain J Marshall, and Byron C Wallace. 2023. Do multi-document summarization models synthesize? *arXiv preprint arXiv:2301.13844*.

Peter Emerson. 2013. The original borda count and partial voting. *Social Choice and Welfare*, 40(2):353–358.

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, Hongkun Yu, You Wu, Cong Yu, Daniel Finnie, Jiaqi Zhai, and Nicholas Zukoski. 2020. Generating representative headlines for news stories.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. Longt5: Efficient text-to-text transformer for long sequences.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Iris Hendrickx, Walter Daelemans, Erwin Marsi, and Emiel Krahmer. 2009. Reducing redundancy in multi-document summarization using lexical semantic similarity. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+ Sum 2009)*, pages 63–66.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Summac: Re-visiting nli-based models for inconsistency detection in summarization.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yu Li, Baolin Peng, Pengcheng He, Michel Galley, Zhou Yu, and Jianfeng Gao. 2022. Dionysus: A pre-trained model for low-resource dialogue summarization.

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

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States. Association for Computational Linguistics.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. *arXiv preprint arXiv:2010.14235*.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *ACM Computing Surveys*, 55(5):1–37.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4–es.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Yulia Otmakhova, Thinh Hung Truong, Timothy Baldwin, Trevor Cohn, Karin Verspoor, and Jey Han Lau. 2022. LED down the rabbit hole: exploring the potential of global attention for biomedical multi-document summarisation. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 181–187, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Ramakanth Pasunuru, Mengwen Liu, Mohit Bansal, Sujith Ravi, and Markus Dreyer. 2021. Efficiently summarizing text and graph encodings of multi-document clusters. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4768–4779, Online. Association for Computational Linguistics.

Jason Phang, Yao Zhao, and Peter J. Liu. 2022. Investigating efficiently extending transformers for long input summarization.

Ratish Surendran Puduppully, Parag Jain, Nancy Chen, and Mark Steedman. 2023. Multi-document summarization with centroid-based pretraining. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 128–138, Toronto, Canada. Association for Computational Linguistics.

Dragomir Radev. 2000. A common theory of information fusion from multiple text sources step one: Cross-document structure. In *1st SIGdial Workshop on Discourse and Dialogue*, pages 74–83, Hong Kong, China. Association for Computational Linguistics.

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

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding.

David Wan and Mohit Bansal. 2022. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, Seattle, United States. Association for Computational Linguistics.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics.

Xun Wang, Masaaki Nishino, Tsutomu Hirao, Katsuhito Sudoh, and Masaaki Nagata. 2016. Exploring text links for coherent multi-document summarization. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 213–223.

Ruben Wolhandler, Arie Cattan, Ori Ernst, and Ido Dagan. 2022. How "multi" is multi-document summarization?

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Michael Strube, and Steffen Eger. 2022. Discoscore: Evaluating text generation with bert and discourse coherence.

A PELMS Pre-training Details

A.1 PELMS Technique Configuration

Topic Clustering For semantic topic clustering, we use the Sentence Transformers *all-mpnet-base-v2* embedding model, which is derived from Song et al. (2020). We cluster using the Sentence Transformers community-based 'fast clustering' algorithm, using cosine similarity as the distance metric. We specify a distance threshold of 0.6, and only consider clusters with at least 2 elements.

As described in section 3, PELMS topic selection is parameterized by k, enabling us to vary the

length per example. In half of the training examples, we set k to 8 for enabling general purpose summarization. For the remaining examples, we sample k to enable length-controlled summarization, as detailed in Appx. D.

Sentence Ranking + Selection For the NLI entailment step, we use the *albert-base-vitaminc* (Schuster et al., 2021) model. We use only the positive entailment probability when calculating intra-cluster entailment. Additionally, the ranking and ordering steps are parametrized by c, which is the number of elements considered for selection from each topic cluster. We set this to 2 in our pre-training.

A.2 Additional Pre-training Details

We train PELMS from the 464M LED-large base model, using the default sliding-window local attention configuration supplemented by inserting `<doc-sep>` tokens with full attention added after each document. For all examples, we cap the input sizes at 4,096 tokens to enable tractable training. This cap affects approximately 10% of pre-training inputs. When truncating long inputs, we distribute the tokens equally among the documents.

We train on 16 GPUs with a total effective batch size of 1,024, learning rate of 5e-5, and 1,250 warmup steps. Leveraging 16 NVIDIA A40 GPUs, this takes approximately 4 days. We train over the full pre-training dataset for one epoch. Preprocessing of the 3-million+ pre-training examples with the PELMS technique takes approximately 14 hours using the same computing environment.

We perform basic dataset pre-processing of MultiPT prior to pre-training target creation, removing extremely short and/or noisy documents containing HTML content.

B Evaluation Details

Input Preparation For fair comparison of all models, we pre-process the input documents similarly, with input sizes capped at 4,096 tokens during training and evaluation. If truncation is necessary, we distribute the tokens equally among the documents, truncating the end of each document.

Generation Settings Largely following the existing methods, for all experiments, we use beam search as the generation decoding strategy, with the number of beams set to 5. We enable tri-gram blocking during decoding as this consistently yields improved results for all models. Following Xiao et al. (2022), we set a maximum generation length per dataset to ensure reasonable and comparable output lengths. For Amazon and Yelp we use 96, DUC2004 and Multi-XScience we use 128, MetaTomatoes we use 192, and MultiNews we use 256.

Model Training We use a learning rate of 3e-4 in our experiments, training for a maximum of 30 epochs. We base this off of hyperparameters from the the four baseline models which all leverage the same LED-large model. We apply early stopping based on validation ROUGE-G score. For few-shot experiments, we scale down the validation set size to match the few-shot size (for example, in 16-shot training, we use a validation set of 16 examples). Results are averaged over 5 runs. For datasets with multiple ground-truth references, we unpack them and treat each reference as a separate example or "shot".

Test Sets We cap test set size to 1,000 examples to ensure tractable evaluation as the neural DiscoScore and MDSummaC automated metrics used in our evaluation are GPU compute-intensive. This impacts only the Multi-XScience and Multi-News evaluation. MDSummaC takes approximately 45 minutes to evaluate 1,000 examples.

Model Selection Due to computational limitations, we evaluate only on recent strong baselines: Primera and Pegasus-X, Centrum, and QAMDen. These outperform other baselines like LED and BART (Lewis et al., 2019). We explored Flan-T5, a T5-based method pre-trained for multiple tasks (Chung et al., 2022), but it showed poor generalization to multi-document summarization in our pilot studies, likely due to lack of long-input pre-training.

C Dataset Information

C.1 Pre-training Datasets

The datasets used in our MultiPT corpus are described here.

- NewSHead (Gu et al., 2020)—A dataset of news stories published between between 2018-2019, grouped together to form topic-centric document clusters.
- BigNews-Aligned (Liu et al., 2022)—A large-scale news dataset containing over 3 million

english political news articles which are clustered by event. These articles are gathered from 11 large US news sites.

- WikiSum-40 (Liu and Lapata, 2019)—Derived from WikiSum (Liu et al., 2018) a Wikipedia corpus containing wikipedia articles and their source articles. WikiSum-40 is a filtered variant containing only the 40 most-relevant documents for every wikipedia article.

- AmazonPT (Ni et al., 2019)—A subset of the massive dataset of Amazon product reviews. We use a 1,000,000 cluster subset of this, sampling uniformly from the available categories.

- YelpPT (`https://www.yelp.com/dataset`)—a reviews dataset containing consumer reviews of businesses.

C.2 Evaluation Datasets

We provide further details on the datasets used during evaluation, and introduce our new MetaTomatoes dataset.

- Multi-News (Fabbri et al., 2019)—A large-scale MDS news summarization dataset containing 56,216 articles-summary pairs.

- DUC2004 (Dang, 2005)—A small carefully-curated news summarization dataset containing 50, 10-document news article clusters and corresponding summaries.

- Multi-XScience (Lu et al., 2020)—A related-work generation task, with the goal of consolidating information from scientific article abstracts cited by a given paper.

- Amazon, Yelp (Bražinskas et al., 2020)—Two small crowdworker-curated opinion summarization datasets with focus on aggregating opinions expressed in consumer reviews of consumer products (Amazon), and businesses (Yelp).

C.2.1 MetaTomatoes Meta-review Dataset

Wang and Ling (2016) previously released RottenTomatoes, an MDS meta-review dataset in which inputs consisted of many short editorial summaries produced by RottenTomatoes contributors. The objective was to generate a brief sentence that captures the overall opinion towards the new film. In contrast, we identify and scrape longer-form meta-reviews produced by the RottenTomatoes editorial team using similar inputs. Figure 4 displays an example meta-summary. These meta-reviews are significantly longer than any one input 'contributor review summary', with the input posing a unique challenge due to the large number of documents, necessitating cross-document information aggregation and understanding.

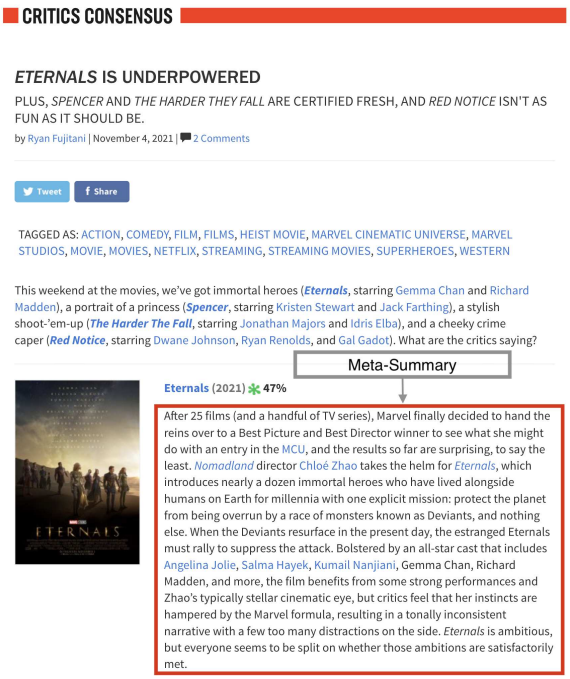

Figure 4: Example Rotten Tomatoes Critics Consensus Meta-Summary

D Length-Controlled Summarization

As mentioned in Section 7, we perform length-controlled summarization experimentation in the zero-shot setting. During pre-training we train with a fixed k of 8 for half of the examples. For the remainder, we randomly sample k from a normal distribution (mean=7, std_dev=5) to encourage flexibility within the model output; in these cases, we prepend the input with a corresponding length prefix, corresponding to one of five length bins. We bound k within the range [1,14]. Table 7 overviews the results and best prefixes for each dataset.

The five length prefixes and corresponding bins are as follows:

- 'short': [1,2]

- 'short-medium': [3,5]

- 'medium': [6,8]

Evaluation Datasets	Domain	#Clusters	#Docs / C	Doc_len	Input_len	Summ. Length
MultiNews	News	56,000	2.8	640.4	1793	217
Multi-XScience	Academic Literature	40,000	4.4	160	700	105
Amazon	Product Reviews	180	8	49.7	397	50.3
Yelp	Business Reviews	300	8	49.8	398.4	52.3
DUC2004	News	50	10	588.2	5882	115
MetaTomatoes	Movie Meta-Reviews	1,497	84.3	23.2	1956	142

Table 6: Overview of the six datasets used in our MDS evaluation.

- 'medium-long': [9,11]
- 'long': [12, 14]

E Additional Evaluation Metrics Details

We provide additional information on the 5 metrics used in our evaluation. All metrics report values in range [0,1] with displayed results multiplied by 100 for presentation purposes. For all, higher scores are better.

Summary Informativeness

- ROUGE (Lin, 2004), an n-gram overlap metric commonly used for summarization evaluation. We follow Pegasus-X in using ROUGE-G (geometric mean of R1/R2/RL) as their overall ROUGE metric, and we additionally use ROUGE-G as our stopping criteria during training. We note that we observed nearly identical behavior and trends when instead using the arithmetic mean of R1/R2/RL.
- BertScore (Zhang et al., 2019b), a BERT-based text similarity metric, also popular for text generation evaluation. We use *DeBERTa-XLarge-MNLI*[2] He et al. (2020) as the base model.

Coherence

- DiscoScore (Zhao et al., 2022), a recent BERT-based method for evaluating discourse coherence. In particular, we use the DS-SENT (NN) variant—which is shown to correlate well with human judgment, particularly in the news domain—to measure the discourse similarities between system and reference summaries.

[2] https://huggingface.co/microsoft/deberta-xlarge-mnli

Faithfulness

- MDSummaC - To better measure consistency between multiple-document inputs and a system summary, **we introduce MDSummaC**, an entailment-based consistency metric that we extend from the SummaC (Laban et al., 2021) consistency metric. SummaC uses entailment to identify whether a text is consistent with another. In particular, we repurpose the $SummaC_{ZS}$ variant, which uses sentence-wise comparison of the input text and generated summary, calculating consistency as the maximum entailment score between a given summary sentence and any input sentence, then reporting the average of this over all of the summary sentences. Unfortunately, this formulation struggles to fit the multi-document case, as SummaC will return a high score even if a summary is maximally consistent with sentences from just one document. To ensure the summary is independently consistent with each document, we instead report the average $SummaC_{ZS}$ score over each individual input document Doc_i in the input I as follows:

$$\text{MDSummaC}\,(I, S) = \frac{1}{n}\sum_{i=1}^{n}\text{SummaC}_{\text{ZS}}\,(Doc_i, S) \tag{1}$$

Abstractiveness

- N-gram novelty - Calculates the abstractiveness of a summary as the proportion of novel n-grams within the summary. For example, unigram novelty captures the proportion of summary unigrams not seen in the input. To simplify our reporting, we report the arithmetic mean of 1-gram, 2-gram and 3-gram novelty as our proxy for summary abstractiveness.

Dataset	Best Prefix	R1	R2	RL	RG	BertS	DiscoS	MDSumC	N-gram Novelty
MultiNews	long	43.4	14.1	20.6	23.3	61.4	95.2	38.1	9.7
Multi-X-Science	none	29.7	5.1	15.6	13.3	56.4	90.8	30.9	5.5
Amazon	medium-long	35.4	8.7	21.2	18.7	63.2	90.9	15.3	32.4
Yelp	short-medium	30.8	6.6	18.1	15.4	61.3	91.2	10.2	38.2
Duc2004	medium	36.1	8.4	18.0	17.6	58.3	93.7	14.8	7.3
MetaTomatoes	medium-long	33.8	7.1	16.3	15.7	55.0	92.0	6.8	43.6
Controlled Avg.		**34.9**	**8.3**	**18.3**	**17.3**	**59.3**	**92.3**	19.4	**22.8**
Uncontrolled Avg.		33.8	7.9	17.7	16.7	59.2	92.2	**19.7**	19.2

Table 7: Comparison of uncontrolled vs length-controlled zero-shot performance. The controlled setting varies only in the prefix supplied during inference. We report the best prefix per dataset, noting that the best prefixes aligned well with the expected best (i.e., longer prefixes for datasets with longer summaries). Our selection metric is RG score.

F Human Evaluation Details

We hired a group of three native English speakers to act as human evaluators of the generated summaries. We use Xiao et al. (2022)'s guidelines for fluency evaluation, measuring grammaticality, referential clarity, and structure & coherence. As their instructions confusingly mixed both absolute scoring with suggestions to perform comparative scoring, we simplified the task to a comparative ranking of the system outputs from our three evaluated models, with ties allowed.

Summaries were presented to annotators in random order. We performed the faithfulness evaluation in two phases. First, in the filtering phase, annotators were asked to identify all Summary Content Units (SCUs) within the inputs. We used majority vote to merge these annotations. Next, once SCUs had been selected, the annotators were provided the system summaries and asked to score each each summary sentence as either 1 or 0, with 1 meaning the sentence was holistically entailed by the input SCUs. To produce an overall score, we report the average percentage of summary sentences that were entailed by the input SCUs.

Figures 5 and 6 contain the guidelines provided during the human evaluation. Our inter-annotator agreement scores (Krippendorf's Alpha) were 0.41 for Grammaticality, 0.48 for Referential Clarity, 0.29 for Structure & Coherence, and 0.51 for Faithfulness.

G Full LLM Results

Table 10 displays the results of GPT-3.5-Long and GPT-4 on each of the six evaluation datasets. We use the "0613" versions of GPT-3.5-Long (16k tokens) and GPT-4 (8k tokens) in a zero-shot setting. For a fair comparison, inputs are all truncated to the same 4,096 tokens as supported by the baseline models. We follow the MDS prompts used by Caciularu et al. (2023), with a simple instruction to "generate a multi-document summary for the following input: <input documents>".

H Full Supervised Results

We provide the full results for all datasets with our supervised setups (all combinations of 16/64/full-shot splits and full-parameter/adapter training) in Tables 11, 12, 13, 14, 15, and 16.

Ground Truth	Pegasus-X	Primera	PELMS
What do you do with an ex-pope? If you happen to know, call Rome, because the Church isn't sure yet. Many fear Benedict will become an implicit rival to the new pope, despite his apparent desire to keep a low profile. He told Roman diocese priests yesterday that he would be "withdrawing into prayer," Reuters reports. "I will always be close to all of you," he said, "even if I remain hidden from the world." The "close to you" bit is somewhat literal; Benedict has announced he'll be living in the Vatican. That's caused some controversy, the Wall Street Journal reports, but one senior papal adviser says it's "better to have him here than somewhere else, where he could become another center of power." Benedict is likely to exert influence over the next pope by picking a new head of the embattled Vatican Bank, and he'll keep writing as a respected theologian. The Church also hasn't decided what vestments he'll wear, whether he'll give audiences, or even what to address him 'a decision that has "juridical" ramifications', a spokesman said. One thing's for sure: He won't be infallible anymore.	VATICAN CITY – The looming retirement of Pope Benedict XVI, who will soon become the first living ex-pontiff in some six centuries, has created a quandary for the Roman Catholic Church: What to make of a former pope who will continue to live under the roof of a new one? Speaking unscripted to thousands of priests from the diocese of Rome, in what turned out to be a farewell address in his capacity as bishop of the Italian capital, Benedict outlined a cloistered life ahead, once he steps down in two weeks time: "Even if I am withdrawing into prayer, I will always been close to all of you and I am sure that you will be close to me, even if I remain hidden to the world," he said. After February 28, when he becomes the first pontiff to resign instead of ruling for life, Benedict will first go to the papal summer retreat at Castel Gandolfo, south of Rome and then move permanently into the four-storey Mater Ecclesiae convent.	VATICAN CITY – The looming retirement of Pope Benedict XVI, who will soon become the first living ex-pontiff in some six centuries, has created a quandary for the Roman Catholic Church: What to make of a former pope who will continue to live under the roof of a new one? Wednesday provided two examples of how the former officeholder could cast a shadow over his successor's work. Pope Benedict was likely to take the sensitive step of appointing a new chief of the Vatican bank, Vatican spokesman Rev. Federico Lombardi said – a move that would leave the next pontiff with a just-picked custodian to a secretive bank that, under Pope Benedict, has denied allegations of money-laundering. The pope underscored that Wednesday. "I have done this in full freedom for the good of the church," he told a general audience, one of his last. Later at Mass, the pontiff received a long and emotional ovation from the hundreds of rank-and-file faithful who packed the pews, including dozens of cardinals who removed their miters in a sign of respect to the outgoing pontiff. One cleric wiped away tears.	Pope Benedict XVI's decision to resign is fueling an unusual discussion within the Vatican: What status he'll hold once he retires? The Vatican, which is navigating uncharted waters since his shock announcement, said experts have still not decided what his title will be or whether he will wear the white of a pope, the red of a cardinal or the black of an ordinary priest. Pope Benedict also made his first public appearances in the wake of his resignation announcement, issuing a strong warning to the cardinals who will elect his successor, decrying the internal rivalries among senior churchmen that have plagued his eight-year papacy. The rivalries burst into public view a year ago when the Vatican confirmed the authenticity of letters to the pope, leaked to the Italian media, detailing the ouster of a top Vatican official who tried to blow the whistle on wasteful spending in the world's smallest state. Father Lombardi said he didn't think the pope consulted the College of Cardinals on his decision to stay inside the Vatican. Added a senior papal adviser: "It's better to have him here than somewhere else, where he could become another center of power."

Table 8: Zero-shot MultiNews Summary Output Example (Ground Truth, Pegasus-X, Primera, PELMS)

Centrum	QAMDen	GPT-3.5-Long	GPT-4
VATICAN CITY (Reuters) – Pope Benedict XVI's decision to resign is fueling an unusual discussion within the Vatican: What status he'll hold once he retires? FILE PHOTO: Pope Benedict Benedict XVI attends Ash Wednesday mass at the Vatican February 13, 2013. Thousands of people are expected to gather in the Vatican for Pope Benedict's Ash Wednesday Mass, which is expected to be his last before leaving office at the end of February. VATICAN CITY – Pope Benedict will see out his life in prayer, 'hidden from the world', he said on Thursday in his first personal comment on his plans since he stunned Roman Catholics by announcing his retirement. His remarks, in a voice that was hoarse at times, followed Monday's resignation notice which spoke of 'a life dedicated to prayer'. The Vatican has said the 85-year-old German will live within its walls. His seclusion may allay concern that the first living former pope in centuries might trouble Church unity. Speaking unscripted to thousands of priests from the diocese of Rome, in what turned out to be a farewell address in his capacity as bishop of the Italian capital, Benedict outlined a cloistered life ahead, once he steps down in two weeks time.	a cloistered life ahead, becauese speaking unscripted to thousands of priests from the diocese of Rome, in what turned out to be a farewell address in his capacity as bishop of the Italian capital, Benedict outlined a close relationship with his flock, once he steps down in two weeks time – – Advertisement – T̈he face of the church is at times disfigured by the sins against the unity of the Church, who will never stop guiding it and I am sure that you will be close to all of you, even if I remain hidden to the world,Ḧe said, Ëven if I am withdrawing into prayer, I will always be close, and I'm sure that I, myself, will always, I'm very sorry that I remain, I hope that I'll always be very close to you, I'll see you, but I'll be very very, very close, very very very close and I hope you'll be close. .	Pope Benedict XVI's decision to resign has sparked discussion about his status once he retires. There are concerns about whether he will become a rival to his successor, as he plans to live inside the Vatican. Pope Benedict has made it clear that he will not interfere with his successor's work and will give up all formal powers of his office. However, there are still questions about what title to use for him and whether he will hold audiences with potential admirers. After his resignation, Pope Benedict will live a secluded life of prayer and will move into the Mater Ecclesiae convent within the Vatican. The Vatican has not yet decided on his title or what he will wear. The election of the new pope will take place in a secret conclave between March 15 and 20.	Pope Benedict XVI's decision to resign has sparked discussions within the Vatican about his status post-retirement. This is the first time in six centuries that a living ex-pontiff will coexist with a new Pope, raising questions about potential conflicts. Pope Benedict has made plans to live inside the Vatican after his resignation, which some analysts believe could unwittingly make him a rival to his successor. However, he has stated that he has no intention of interfering with his successor's work and will give up all formal powers of his office. The Pope will continue to live inside Vatican City, where he will write and pray. His future home will be a renovated convent within Vatican walls. The Vatican has stated that he will not influence the election of his successor. However, his title post-resignation and his attire are still under consideration. The Pope has stated that he will spend his time in prayer, "hidden from the world".

Table 9: Zero-shot MultiNews Summary Output Example (Centrum, QAMDen, GPT-3.5-Long, GPT-4)

Dataset	Model	R1	R2	RL	RG	BertS	DiscoS	MDSummaC	N-gram Novelty
MultiNews	GPT-3.5-Long	40.6	12.2	20.2	21.6	63.0	94.0	27.4	38.9
	GPT-4	41.3	12.5	20.3	21.9	63.4	94.5	26.3	37.5
Multi-XScience	GPT-3.5-Long	30.7	5.2	15.8	13.7	56.5	91.1	20.6	40.6
	GPT-4	30.8	5.1	15.9	13.6	56.7	91.4	16.9	42.9
Amazon	GPT-3.5-Long	31.3	5.7	17.9	14.7	62.4	90.6	6.6	70.1
	GPT-4	33.2	6.9	19.1	16.3	64.6	91.9	5.2	71.3
Yelp	GPT-3.5-Long	33.6	6.6	19.3	16.2	64.0	91.9	6.9	65.2
	GPT-4	32.7	6.0	18.5	15.4	64.3	91.5	4.6	70.2
DUC2004	GPT-3.5-Long	40.0	10.4	20.2	20.3	63.3	95.4	14.3	34.4
	GPT-4	40.8	11.2	21.2	21.3	64.0	95.4	12.1	34.2
MetaTomatoes	GPT-3.5-Long	31.2	5.7	15.2	14.0	57.9	85.9	6.0	55.9
	GPT-4	32.2	5.7	15.4	14.1	57.8	87.5	4.5	56.2
Overall	GPT-3.5-Long	34.6	7.6	18.1	16.8	61.2	91.5	13.6	50.9
	GPT-4	35.2	7.9	18.4	17.1	61.8	92.0	11.6	52.1

Table 10: Results of LLMs on the multi-document summarization task. We see GPT-4 achieves slightly higher results compared to GPT-3.5-Long. In particular, GPT-4 is 0.6 points better than GPT-3.5 for BertScore informativeness. Coherence and abstractiveness are also moderately improved, although there is some decrease in faithfulness.

Overall objective

Your task is to assess the quality of AI-written summaries. The summaries will be of a set of input articles. For this particular task, you will be working with summaries of movie reviews.

High-level overview of the two phases of this task

You will be tasked with two types of summary evaluation.

- **Evaluation 1** - We will measure whether summaries are faithful to the provided reviews; for each important fact or opinion expressed in the summaries, we will determine whether it can be resonably attributed back to input as a whole.
  - Step 1
    - To understand how consistent a summary is, we first need to understand what information is expressed in the articles. The first step will consist of a filtering step to finding important or highly-relevant information within the input. Will will refer to these as **S**ummary **C**ontent **U**nits (SCU's).
  - Step 2
    - For each expression / SCU in the AI-generated summaries, we will determine whether they reflect the relevant SCUs in the input reviews.
- **Evaluation 2** - Summary comparison consists of comparing and ranking summaries over several types of summary characteristics, such as fluency and coherence.

Evaluation 1, Step 1: Identifying summary-worthy information

I. The Goal: Finding relevant information within the input articles

Before rating the summaries, we first want to understand the inputs. Specifically, we need to identify the sentences in each document that contain informative and well-written ideas/opinions, which we will refer to as Summary Content Units (SCUs).

For this task, you **will not** be asked to write your own summary -- you will simply **select well-formed sentences (i.e. SCUs) that you would want to see in a summary of the documents you are presented with**.

II. The Task: Document Sentence Selection

You will be given a spreadsheet containing all of the documents for each example:

See Fig. 1

Each row of the spreadsheet will contain one sentence from a document. For every document, you should enter "1" to select the SCUs / sentences that are well-formed and that you believe capture the most important and useful information. Enter "0" to exclude the sentences that do not (see above example). You should judge the contents of each document independently -- **don't worry if similar, identical or contradictory information** is selected across documents.

Even selecting *zero* or *all* sentences in a single document is acceptable, if you think it's the right choice. We will not enforce limits on the number of sentences you select, but we expect you to select, *on average*, **20%-40% of the available sentences**. To help you keep track of how many sentences you've selected, the percent selected column calculates the total fraction of sentences you selected so far (i.e. total sentences selected divided by total sentences labeled).

Judging whether a particular sentence is worth selecting is, to some extent, subjective and **you should select the segments that you would want to see in a summary** of the topic covered by all the documents within that example. However, some basic rules should be taken into account:

1. Do select sentences that express **key ideas/opinions** in the document *or* give **useful details which support these ideas/opinions**.
   Do not select sentences that contain **less important elaborations or further details** about the topic.

The location is actually excellent.	[useful opinion]
Within walking distance of the Magnificent Mile.	[useful details]
Good restaurants and high end shops all around.	[useful details]
I never got bored of walking in that neighbourhood.	[elaboration]
The rooms were beautiful and massive!	[useful opinion]
Also, the most comfy beds I've ever slept on.	[useful opinion]
It was so hard to get out of bed every morning.	[elaboration]
I slept like a baby for a whole week :)	[elaboration]

2. **General background information** about the topic **should be avoided**, as it provides no useful information about the **key ideas/opinions** the document is presenting. The same applies to sentences which are about tangential subtopics.

We were looking for a family-friendly hotel.	[not about the hotel]
It feels harder and harder to find one these days.	[not about the hotel]
This one was just perfect!!	[general comment]
Beautiful and spacious rooms with a view.	[useful opinion]
And an excellent pool for the kids!	[useful opinion]
I'd definitely recommend this place to anyone!	[general comment]

3. **Avoid selecting poorly-written sentences** (e.g., misspelled, or with wrong punctuation). Also avoid very long sentences that combine useful opinions with less interesting information.

The staff were very friendly and even though we booked through expedia to get a great deal on the hotel we were treated like a somebody and even received a room upgrade to a lake view.	[mix of useful with other]
First....location!close to Pike Market....a few blocks...!shopping ALLOVER.	[poorly written]

EXAMPLE	DOC	INPUTS	SCU LABEL	PERCENT SELECTED
0	0	Kissing may have evolved to help people pick a compatible mate -- and to keep the love going in long-term relationships.	1	1
0	0	You've got to kiss a lot of frogs to find your prince, as the saying goes.	0	0.5
0	0	New research suggests the cliché is true on an evolutionary level.	0	0.3333333333
0	0	Kissing might have evolved as a way to assess the quality of potential mates, according to two new studies.	0	0.25
0	0	Women, who tend to be pickier about romantic entanglements than men, also care more about kissing in the first phases of a relationship, suggesting that make-outs may weed out duds.	0	0.2
0	0	What's more, women are especially attuned to the importance of kissing during fertile phases of the menstrual cycle.	0	0.1666666667
0	0	Kissing exists in virtually every culture on Earth, said study researcher Rafael Wlodarski, a doctoral candidate at the University of Oxford.	1	0.2857142857
0	0	Some of the oldest records left by humanity, including the Hindu Veda and ancient Egyptian wall murals, depict kissing.	0	0.25
0	0	"Because it's so common," Wlodarski told LiveScience, "it might serve a purpose."	0	0.2222222222
0	0	The evolution of make-outs Theories about why kissing matters fall into three categories.	0	0.2
0	0	Some believe kissing evolved to help people assess potential mates, perhaps by transmitting pheromones, or chemical signals that could carry information about health or immune compatibility.	1	0.2727272727
0	0	[50 Sultry Facts About Sex] "It's just an excuse to get two people who are interested in each other close enough to have a sniff," Wlodarski said.	0	0.25
0	0	No particular compound has been proven to be a human pheromone, but there is evidence that scent carries information.	1	0.3076923077
0	0	One study published in April 2013 found that women prefer the scent of men who have high levels of the masculine hormone testosterone.	0	0.2857142857
0	0	Kissing also may have evolved to keep romantic pairs bonded, or to increase arousal prior to sex.	0	0.2666666667
0	0	To test these theories, Wlodarski and his colleagues recruited 902 American and British adults to answer questions about their attitudes toward kissing.	0	0.25
0	0	The participants rated how important they considered kissing at various stages in relationships.	0	0.2352941176
0	0	The approximately half of participants who were in relationships also reported how much they and their partners kissed, and how satisfied they were in the relationship.	0	0.2222222222
0	0	The results gave little support to the notion that kissing evolved to ease the way to sex (even if it may often be used that way).	0	0.2105263158

Fig. 1

Figure 5: Human Evaluation Guidelines Pt. 1

Evaluation 1 Step 2: Scoring the faithfulness of the generated summaries

The next step is to utilize the filtered input SCUs faithfulness analysis.
For this, we will measure (one phrase at a time) whether the content of each summary can be attributed back to input SCUs. An highlighted example is provided below the instructions.

For each phrase in a generated summary:

- Step 1: Compare the **summary phrase** ("e.g. Emily Blunt and her female co-stars are outstanding here") with the list of **input SCUs** on the left-most column.
- Step 2: Enter a **1** or **0** in an empty column below the summary you are rating. **1** if the input is reflective of the inputs as whole, and **0** if not.
  - Label 1 if the summary phrase is reflective of the inputs. **Be sure to consider all of the input SCUs when making this judgement.** For example, if the input SCUs contain "*This is the worst movie I've seen*" and "*This is a terrible film*" then the summary phrase "*Reviewers are mixed on Top Gun Maverick*" would be labeled as **1.**
    - Consider the **most apparent** takeaway when viewing the Input SCUs. For example, if 20 input SCUs speak positively and 1 speaks negatively towards a topic, then a generated summary should speak positively. However, if there are many varied or more nuanced perspectives, then the summary should reflect this diversity.
  - Label 0 otherwise.
    - Note that a summary containing information not found in the SCUs should be labeled 0.

Complete this process for each of the 3 summaries independently.

See Fig. 2

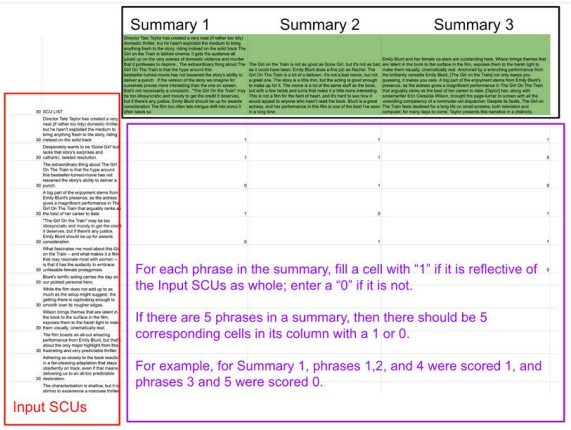

Fig. 2

Evaluation 2 - Measuring summary fluency

Example of spreadsheet you will complete. (Note, ignore the Pyramid Score column).
Your spreadsheet assignments along with example IDs to annotate will be e-mailed to you.

See Fig. 3

There are multiple candidate summaries (3) for each example and we would like to rank the summaries from most most fluent to least. There are **three rows per movie**, with the SUMMARY column being the AI-generated summaries that you will evaluate. The GROUND TRUTH column contains the summaries you will compare the generated summaries with. Note that for one example (3 rows), we are comparing to the same GROUND TRUTH summary for all.

Specifically, for each summary, score them according to the three following guidelines:

1. **Grammaticality -** The summary should have no datelines, system-internal formatting, random text or URLs, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.

2. **Referential Clarity -** It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.

3. **Structure and Coherence -** The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

Please note that we want to ***rank*** the models, where a score of '1' means the model was the best. "2" would indicate second best, and "3" would indicate worst. Ties are allowed. See the following examples to understand how to enter ties.

The following are all valid ways of ranking the three summaries:

[1,1,1] (All are equal)
[1,2,3] (There is a clear first, second and third best)
[1,2,2] (There is a clear first, and the other two are tied for second)
[1,1,2] (Two are equally good, and one is worse)

See Fig. 4

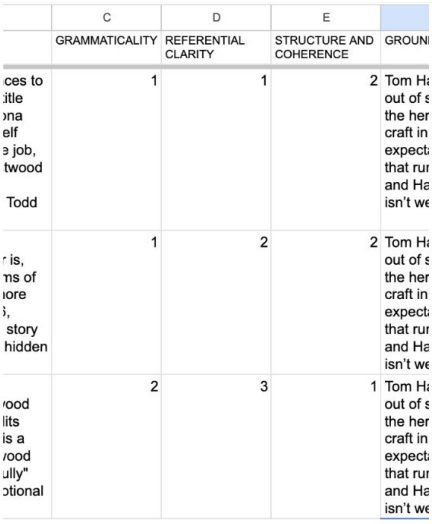

Fig. 4

A	B	C	D	E	
EXAMPLE	SUMMARY	GRAMMATICALITY	REFERENTIAL CLARITY	STRUCTURE AND COHERENCE	GROUND TRUTH
0	Sully could have very easily been a film about a tragic plane crash, and Eastwood invites audiences to simply stop and consider the miracle for a while longer.. If Tom Hanks hadn't existed to play the title role, Eastwood would have had to invent him: This is one of those cases where movie-star persona and real-life legend There's a good idea for a movie here, in fact, there are two.But the movie itself never quite takes off.There is something compelling about someone who simply shows up for the job, day after day, carving out the remarkable from the unremarkable.A movie that unlocks other Eastwood movies, showing them all to be nuanced studies of tormented masculinity - yes, even "American Sniper."The so-called 'untold story' is essentially nonsense...The struggle is real for screenwriter Todd Komarnicki...Tense and exciting - as Clint Eastwood, once again,	-	-	-	Tom Hanks excels at playing decent fellows, out of stories about simple guys rising to diffic the heroic efforts of real-life US Airways pilot craft in the Hudson River — looks like a comp expectations, although not quite to the extent that runs more toward the "solidly entertainin and Hanks have both made better movies, th isn't well worth a trip to your local cineplex.
0	But the movie itself never quite takes off. The closest [Sully] gets to 'emotive' is a brief welling-of-the-eyes at the hospital, after hearing that all of his passengers survived. Sullenberger is, fundamentally, not that interesting of a protagonist. He's so heroic, so evidently in the right in terms of his actions, that his struggles don't resonate. Sully wants to be life-affirming Oscar bait, but it's more like Clint Eastwood's Book Club For Dads Who Shop In Cargo Pants At Costco. Eastwood, at 86, continues to be a marvel, and this stands as one of the four-time Oscar winner's best films. True story of "The Miracle on the Hudson" Exhilarating, Moving and Well Done Weaving an enthralling tale hidden behind the one everyone read about in the press, "Sully" is a	-	-	-	Tom Hanks excels at playing decent fellows, out of stories about simple guys rising to diffic the heroic efforts of real-life US Airways pilot craft in the Hudson River — looks like a comp expectations, although not quite to the extent that runs more toward the "solidly entertainin and Hanks have both made better movies, th isn't well worth a trip to your local cineplex.
0	Sully is one of Clint Eastwood's best films. The closest [Sully] gets to 'emotive' is a brief welling-of-the-eyes at the hospital, after hearing that all of his passengers survived. It's an Eastwood film that very nearly strikes the same unsophisticated but delicate tone of his obligatory end-credits ditty. Eastwood is able to tell a story so big in a tight 96-minute package and do it so effortlessly is a true achievement. Sully could have very easily been a film about a tragic plane crash, and Eastwood invites audiences to simply stop and consider the miracle for a while longer. Inevitable is how "Sully" feels. The story of the Miracle on the Hudson is only one of the rewards of this earnest, slyly emotional biography.	-	-	-	Tom Hanks excels at playing decent fellows, out of stories about simple guys rising to diffic the heroic efforts of real-life US Airways pilot craft in the Hudson River — looks like a comp expectations, although not quite to the extent that runs more toward the "solidly entertainin and Hanks have both made better movies, th isn't well worth a trip to your local cineplex.

Fig. 3

Figure 6: Human Evaluation Guidelines Pt. 2

Dataset	Model	R1	R2	RL	RG	BertS	DiscoS	MDSummaC	N-gram Novelty
MultiNews	Pegasus-X	35.5 (0.2)	11.3 (0.5)	18.0 (0.4)	19.3 (0.5)	60.6 (0.5)	85.0 (0.8)	28.0 (2.8)	17.3 (4.1)
	Primera	44.2 (0.3)	15.1 (0.4)	21.8 (0.4)	24.4 (0.4)	63.0 (0.3)	95.6 (0.2)	36.1 (3.7)	11.4 (3.7)
	QAmden	18.4 (0.2)	5.2 (0.1)	11.7 (0.1)	10.4 (0.1)	53.1 (0.1)	57.8 (0.2)	37.5 (0.2)	10.7 (0.2)
	Centrum	44.4 (0.3)	15.6 (0.4)	22.1 (0.4)	24.8 (0.4)	62.5 (0.3)	95.9 (0.0)	39.8 (0.5)	10.8 (0.5)
	PELMS	43.8 (0.9)	14.6 (0.6)	21.3 (0.5)	23.9 (0.6)	62.7 (0.6)	95.9 (0.2)^	34.6 (3.3)	15.0 (4.1)
Multi-XScience	Pegasus-X	24.7 (1.3)	3.7 (0.2)	13.9 (0.4)	10.8 (0.4)	55.2 (0.2)	82.4 (1.9)	21.8 (6.3)	18.1 (12.8)
	Primera	29.1 (0.4)	4.6 (0.2)	15.3 (0.2)	12.6 (0.3)	55.7 (0.2)	89.8 (1.2)	30.0 (3.0)	6.9 (7.0)
	QAmden	20.6 (0.2)	2.9 (0.1)	12.8 (0.1)	9.2 (0.1)	52.7 (0.1)	70.1 (0.1)	27.3 (0.3)	10.9 (0.1)
	Centrum	29.3 (0.2)	4.7 (0.1)	15.3 (0.1)	12.8 (0.1)	55.3 (0.1)	91.1 (0.1)	31.4 (0.4)	6.9 (0.5)
	PELMS	**30.6 (0.8)^***	**5.1 (0.3)^***	**16.0 (0.3)^***	**13.6 (0.4)^***	**56.0 (0.3)^***	90.1 (1.0)	18.4 (2.6)	27.9 (6.6)^
Amazon	Pegasus-X	33.4 (1.5)	6.6 (0.9)	20.6 (1.0)	16.5 (1.3)	64.9 (0.9)	91.9 (0.4)	9.3 (1.3)	54.0 (5.9)
	Primera	34.2 (1.4)	7.3 (0.5)	21.4 (0.4)	17.5 (0.6)	65.3 (0.4)	91.8 (1.1)	11.7 (1.6)	39.3 (11.4)
	QAmden	21.7 (0.2)	2.7 (0.1)	14.0 (0.1)	9.4 (0.0)	56.0 (0.0)	75.3 (0.0)	12.2 (0.2)	8.8 (0.2)
	Centrum	28.7 (0.3)	4.8 (0.3)	16.6 (0.3)	13.1 (0.4)	58.6 (0.3)	90.7 (0.1)	12.3 (0.3)	22.9 (1.7)
	PELMS	**35.8 (0.8)^***	**8.8 (0.5)^***	**22.6 (0.3)^***	**19.2 (0.5)^***	**66.6 (0.2)^***	92.2 (0.5)^	11.0 (2.1)	46.5 (5.6)
Yelp	Pegasus-X	31.5 (1.8)	6.6 (0.7)	19.8 (1.3)	16.0 (1.2)	65.3 (0.9)	90.0 (1.1)	11.5 (1.9)	46.6 (8.6)
	Primera	32.8 (1.6)	7.4 (0.8)	20.8 (1.2)	17.1 (1.2)	64.9 (1.2)	91.8 (0.3)	12.6 (2.0)	38.0 (13.7)
	QAmden	17.5 (0.3)	2.9 (0.1)	12.3 (0.2)	8.5 (0.1)	55.3 (0.3)	70.4 (0.2)	11.3 (0.2)	10.5 (0.2)
	Centrum	28.4 (0.2)	5.6 (0.1)	17.0 (0.1)	13.9 (0.2)	59.9 (0.3)	90.5 (0.2)	12.2 (0.3)	29.3 (1.6)
	PELMS	**35.5 (1.1)^***	**9.3 (0.4)^***	**22.1 (0.5)^***	**19.4 (0.5)^***	**66.2 (0.4)^***	91.3 (1.8)	12.8 (1.8)^	32.7 (5.1)
DUC2004	Pegasus-X	34.8 (1.6)	7.2 (0.7)	17.4 (0.4)	16.3 (0.9)	59.4 (0.9)	92.3 (0.5)	4.5 (0.4)	13.5 (3.6)
	Primera	35.6 (0.8)	8.2 (0.6)	18.1 (0.7)	17.4 (0.7)	59.8 (0.9)	93.6 (1.0)	15.9 (0.8)	5.8 (2.6)
	QAmden	22.6 (0.3)	3.6 (0.1)	13.7 (0.2)	10.3 (0.2)	54.7 (0.2)	68.6 (0.6)	12.8 (0.5)	6.0 (0.2)
	Centrum	34.5 (0.1)	8.2 (0.0)	18.2 (0.1)	17.3 (0.1)	59.2 (0.0)	94.1 (0.0)	14.8 (0.3)	4.4 (0.2)
	PELMS	**37.7 (1.4)^***	**9.3 (1.1)^***	**19.2 (0.7)^***	**18.8 (1.2)^***	**61.8 (0.7)^***	93.9 (1.7)	15.7 (2.2)	**21.7 (4.8)^***
MetaTomatoes	Pegasus-X	31.7 (0.8)	6.0 (0.3)	15.0 (0.3)	14.1 (0.4)	57.5 (0.6)	93.0 (0.9)	2.6 (0.6)	55.3 (10.0)
	Primera	31.3 (1.4)	6.2 (0.5)	15.0 (0.4)	14.3 (0.7)	58.0 (0.7)	92.1 (1.1)	4.5 (0.7)	42.9 (8.9)
	QAmden	18.1 (0.5)	2.5 (0.2)	11.1 (0.2)	8.0 (0.3)	48.2 (0.1)	68.0 (0.2)	3.6 (0.0)	55.4 (0.7)
	Centrum	32.5 (0.1)	5.6 (0.1)	14.9 (0.1)	14.0 (0.1)	55.1 (0.1)	93.6 (0.3)	4.8 (0.1)	23.2 (2.0)
	PELMS	**32.8 (0.2)^***	**7.0 (0.1)^***	**15.8 (0.1)^***	**15.4 (0.1)^***	56.3 (1.5)	91.9 (0.7)	**6.4 (1.7)^***	36.6 (13.7)

Table 11: 16-shot results with full-parameter training. In our experiments, we average over 5 runs, each with unique random seeds. We report the mean and (std) values. Green and ^ indicate PELMS outperforms all baselines. **Bold** and * indicate improvement is statistically significant (one-tailed paired t-test with each baseline, $p < 0.05$).

Dataset	Model	R1	R2	RL	RG	BertS	DiscoS	MDSummaC	N-gram Novelty
MultiNews	Pegasus-X	25.1 (0.4)	7.2 (0.2)	13.6 (0.2)	13.5 (0.3)	56.9 (0.1)	71.0 (0.5)	40.0 (0.3)	1.7 (0.3)
	Primera	41.1 (0.6)	12.8 (0.4)	19.7 (0.4)	21.8 (0.5)	61.0 (0.3)	93.5 (1.1)	43.3 (0.4)	1.8 (0.2)
	QAmden	18.5 (0.2)	5.3 (0.1)	11.7 (0.1)	10.4 (0.1)	53.1 (0.1)	57.9 (0.2)	37.3 (0.2)	10.7 (0.1)
	Centrum	44.4 (0.2)	15.6 (0.3)	22.1 (0.3)	24.8 (0.3)	62.5 (0.3)	95.8 (0.1)	39.7 (0.4)	11.1 (0.3)
	PELMS	41.9 (0.5)	13.8 (0.3)	20.4 (0.3)	22.7 (0.3)	61.2 (0.2)	95.2 (0.3)	39.7 (0.3)	6.5 (0.3)
Multi-XScience	Pegasus-X	23.7 (0.3)	3.6 (0.1)	13.4 (0.2)	10.4 (0.2)	55.1 (0.1)	81.9 (0.4)	30.0 (0.4)	3.2 (0.3)
	Primera	28.7 (0.4)	4.4 (0.1)	15.1 (0.2)	12.4 (0.2)	55.5 (0.1)	87.7 (1.5)	31.8 (0.7)	1.2 (0.1)
	QAmden	20.6 (0.2)	2.9 (0.1)	12.8 (0.1)	9.2 (0.1)	52.6 (0.1)	70.1 (0.2)	27.4 (0.3)	10.9 (0.2)
	Centrum	29.1 (0.2)	4.6 (0.1)	15.2 (0.1)	12.7 (0.1)	55.2 (0.0)	91.0 (0.1)	31.3 (0.4)	7.4 (0.2)
	PELMS	**29.8 (0.2)^***	**4.8 (0.1)^***	15.3 (0.1)^	**13.0 (0.1)^***	**55.9 (0.1)^***	91.0 (0.2)^	31.5 (0.4)	3.1 (0.3)
Amazon	Pegasus-X	24.9 (0.3)	3.0 (0.1)	14.8 (0.2)	10.4 (0.2)	58.6 (0.1)	85.5 (0.2)	14.9 (0.1)	3.9 (0.1)
	Primera	27.1 (0.3)	4.3 (0.1)	15.9 (0.1)	12.3 (0.1)	59.2 (0.2)	86.8 (0.7)	11.3 (0.8)	2.0 (0.2)
	QAmden	21.7 (0.0)	2.8 (0.0)	14.0 (0.0)	9.5 (0.0)	56.0 (0.0)	75.3 (0.0)	12.0 (0.0)	8.6 (0.0)
	Centrum	28.7 (0.1)	4.7 (0.1)	16.7 (0.1)	13.1 (0.1)	58.6 (0.1)	90.4 (0.1)	12.6 (0.1)	23.9 (0.9)
	PELMS	**32.0 (0.5)^***	**6.9 (0.3)^***	**19.5 (0.5)^***	**16.2 (0.4)^***	**63.3 (0.5)^***	**91.2 (0.2)^***	13.8 (0.3)	**28.8 (2.4)^***
Yelp	Pegasus-X	21.1 (0.4)	3.2 (0.3)	13.0 (0.4)	9.6 (0.4)	58.3 (0.3)	82.0 (0.6)	13.9 (0.1)	5.6 (0.1)
	Primera	26.7 (0.3)	4.7 (0.2)	16.2 (0.3)	12.7 (0.3)	60.3 (0.3)	87.7 (3.3)	13.5 (2.2)	1.6 (0.4)
	QAmden	17.2 (0.0)	2.8 (0.0)	12.2 (0.0)	8.3 (0.0)	54.8 (0.1)	70.2 (0.0)	11.5 (0.0)	10.5 (0.0)
	Centrum	26.7 (0.1)	5.1 (0.0)	16.4 (0.1)	13.1 (0.0)	58.5 (0.1)	89.3 (0.0)	9.7 (0.2)	39.4 (0.4)
	PELMS	**32.5 (0.5)^***	**8.6 (0.4)^***	**20.2 (0.5)^***	**17.8 (0.5)^***	**64.6 (0.5)^***	**91.6 (0.2)^***	14.2 (0.4)^	27.2 (5.2)
DUC2004	Pegasus-X	31.5 (0.5)	5.6 (0.4)	16.0 (0.3)	14.1 (0.5)	57.6 (0.3)	90.6 (0.5)	4.5 (0.1)	3.4 (0.2)
	Primera	32.8 (0.1)	7.0 (0.1)	16.6 (0.1)	15.6 (0.1)	58.4 (0.0)	80.8 (0.4)	15.5 (0.5)	1.5 (0.2)
	QAmden	21.9 (0.0)	3.2 (0.0)	13.3 (0.0)	9.8 (0.0)	54.5 (0.0)	69.2 (0.1)	11.8 (0.0)	5.8 (0.0)
	Centrum	34.2 (0.0)	8.0 (0.0)	18.0 (0.0)	17.0 (0.0)	59.0 (0.0)	94.0 (0.0)	14.3 (0.3)	4.1 (0.1)
	PELMS	34.7 (1.0)^	7.3 (0.5)	17.1 (0.5)	16.3 (0.7)	**59.5 (0.5)^***	93.5 (0.5)	15.6 (0.4)^	**6.5 (0.8)^***
MetaTomatoes	Pegasus-X	29.4 (0.1)	4.2 (0.1)	13.4 (0.1)	11.9 (0.1)	54.2 (0.2)	92.5 (0.2)	5.0 (0.3)	12.5 (0.8)
	Primera	30.5 (0.3)	4.9 (0.1)	14.0 (0.1)	12.8 (0.1)	55.1 (0.1)	83.6 (1.8)	5.9 (0.2)	5.6 (0.6)
	QAmden	14.1 (0.1)	1.7 (0.0)	8.8 (0.1)	5.9 (0.1)	47.5 (0.1)	66.7 (0.0)	4.2 (0.1)	58.0 (0.2)
	Centrum	32.5 (0.1)	5.6 (0.0)	14.9 (0.0)	13.9 (0.0)	55.0 (0.1)	93.3 (0.0)	4.8 (0.0)	26.0 (0.5)
	PELMS	**32.9 (0.1)^***	**7.3 (0.1)^***	**16.2 (0.1)^***	**15.7 (0.1)^***	**55.3 (0.2)^***	91.4 (0.1)	**7.6 (0.1)^***	36.7 (1.6)

Table 12: 16-shot results with adapter (5%) training. In our experiments, we average over 5 runs, each with unique random seeds. We report the mean and (std) values. Green and ^ indicate PELMS outperforms all baselines. **Bold** and * indicate improvement is statistically significant (one-tailed paired t-test with each baseline, $p < 0.05$).

Dataset	Model	R1	R2	RL	RG	BertS	DiscoS	MDSummaC	N-gram Novelty
MultiNews	Pegasus-X	42.0 (0.5)	14.1 (0.2)	20.8 (0.3)	23.1 (0.3)	62.8 (0.3)	93.5 (0.5)	23.2 (3.5)	26.5 (4.6)
	Primera	45.0 (0.4)	15.9 (0.5)	22.3 (0.5)	25.2 (0.4)	63.7 (0.3)	96.0 (0.2)	33.2 (3.8)	17.0 (3.9)
	QAmden	29.2 (1.0)	9.2 (0.4)	16.0 (0.4)	16.3 (0.5)	57.2 (0.4)	76.1 (1.4)	36.3 (0.4)	8.6 (0.2)
	Centrum	44.4 (0.2)	15.5 (0.3)	22.1 (0.3)	24.8 (0.3)	62.6 (0.3)	96.0 (0.1)	40.6 (0.9)	9.4 (1.5)
	PELMS	45.0 (0.2)^	15.5 (0.1)	21.9 (0.1)	24.8 (0.1)	63.4 (0.3)	96.2 (0.2)^	30.6 (1.7)	20.5 (1.9)
Multi-XScience	Pegasus-X	27.9 (1.1)	4.6 (0.2)	15.4 (0.3)	12.6 (0.5)	56.0 (0.2)	86.1 (1.6)	11.9 (1.4)	41.4 (4.9)
	Primera	29.8 (0.3)	4.9 (0.1)	15.8 (0.1)	13.2 (0.1)	56.1 (0.1)	90.3 (0.4)	20.7 (4.5)	26.6 (10.1)
	QAmden	26.1 (1.0)	4.2 (0.3)	14.7 (0.3)	11.7 (0.5)	54.8 (0.3)	82.4 (2.2)	26.4 (0.7)	9.8 (0.9)
	Centrum	29.3 (0.2)	4.7 (0.1)	15.4 (0.2)	12.8 (0.2)	55.5 (0.1)	91.0 (0.1)	31.7 (0.6)	5.9 (0.7)
	PELMS	29.9 (0.7)^	5.0 (0.2)^	15.8 (0.2)^	13.3 (0.3)^	56.2 (0.1)^	89.1 (1.1)	20.0 (2.2)	25.1 (5.3)
Amazon	Pegasus-X	36.4 (1.5)	8.3 (1.1)	22.9 (1.2)	19.0 (1.4)	66.9 (0.6)	92.9 (0.2)	9.4 (1.1)	57.9 (3.5)
	Primera	35.3 (0.8)	7.9 (0.5)	21.4 (0.7)	18.1 (0.6)	65.2 (0.8)	92.3 (0.4)	10.8 (1.7)	38.8 (8.8)
	QAmden	32.9 (0.4)	7.4 (0.4)	21.4 (0.4)	17.3 (0.5)	63.3 (0.4)	90.4 (1.2)	11.7 (0.7)	21.4 (2.9)
	Centrum	29.9 (0.9)	5.8 (0.4)	18.0 (0.4)	14.6 (0.5)	61.0 (0.6)	90.4 (0.4)	13.9 (0.8)	11.7 (1.7)
	PELMS	36.8 (1.0)^	9.2 (0.5)^	23.4 (0.6)^	19.9 (0.6)^	67.1 (0.6)^	92.2 (0.7)	11.4 (1.7)	43.5 (9.0)
Yelp	Pegasus-X	35.4 (0.4)	8.6 (0.5)	22.4 (0.5)	19.0 (0.5)	66.9 (0.4)	92.3 (0.3)	11.5 (1.0)	54.8 (6.4)
	Primera	35.1 (0.8)	8.7 (0.4)	22.1 (0.7)	18.9 (0.5)	66.6 (0.4)	91.9 (0.6)	12.4 (2.2)	42.4 (4.4)
	QAmden	33.2 (0.8)	8.4 (0.3)	21.2 (0.7)	18.1 (0.5)	65.2 (0.7)	90.2 (0.5)	14.2 (0.3)	26.2 (4.5)
	Centrum	30.5 (0.5)	6.7 (0.1)	18.0 (0.4)	15.5 (0.3)	61.9 (0.3)	91.2 (0.3)	14.4 (0.2)	15.1 (1.4)
	PELMS	**36.0 (0.6)^***	**9.8 (0.7)^***	22.8 (0.7)^	**20.0 (0.8)^***	67.3 (0.6)^	91.0 (1.2)	13.5 (2.4)	40.3 (7.5)
DUC2004	Pegasus-X	37.5 (0.9)	8.7 (0.5)	19.0 (0.6)	18.4 (0.6)	61.4 (0.6)	94.1 (1.3)	3.6 (0.7)	32.9 (4.7)
	Primera	37.9 (0.3)	9.3 (0.3)	19.3 (0.3)	19.0 (0.3)	60.9 (0.5)	94.7 (0.1)	15.5 (1.6)	13.9 (7.1)
	QAmden	34.1 (0.5)	7.6 (0.3)	18.0 (0.1)	16.7 (0.3)	58.6 (0.2)	84.0 (1.5)	13.5 (0.5)	5.9 (0.8)
	Centrum	34.2 (0.5)	7.8 (0.2)	18.0 (0.2)	16.8 (0.2)	59.1 (0.3)	93.8 (0.4)	15.0 (0.4)	3.7 (0.5)
	PELMS	37.7 (0.5)	9.3 (0.8)^	19.3 (0.7)^	18.9 (0.8)	61.4 (0.7)^	93.3 (0.8)	14.6 (0.9)	21.1 (3.0)
MetaTomatoes	Pegasus-X	33.5 (0.6)	7.3 (0.5)	16.1 (0.3)	15.8 (0.5)	59.0 (0.4)	93.2 (0.5)	1.7 (0.3)	63.1 (4.2)
	Primera	31.6 (0.7)	6.4 (0.3)	15.2 (0.3)	14.5 (0.4)	58.6 (0.5)	92.5 (0.2)	4.4 (0.6)	47.0 (6.3)
	QAmden	22.0 (1.0)	5.6 (0.3)	12.6 (0.4)	11.6 (0.5)	54.8 (0.2)	75.8 (1.5)	3.4 (0.1)	50.8 (1.8)
	Centrum	32.5 (0.2)	5.9 (0.2)	15.0 (0.2)	14.2 (0.2)	55.6 (0.3)	93.6 (0.1)	4.2 (0.3)	27.3 (3.5)
	PELMS	31.6 (0.8)	7.0 (0.3)	15.5 (0.3)	15.1 (0.2)	59.4 (0.3)^	91.8 (1.0)	3.1 (0.3)	55.6 (3.0)

Table 13: 64-shot results with full-parameter training. In our experiments, we average over 5 runs, each with unique random seeds. We report the mean and (std) values. Green and ^ indicate PELMS outperforms all baselines. **Bold** and * indicate improvement is statistically significant (one-tailed paired t-test with each baseline, $p < 0.05$).

Dataset	Model	R1	R2	RL	RG	BertS	DiscoS	MDSummaC	N-gram Novelty
MultiNews	Pegasus-X	26.2 (0.6)	7.6 (0.4)	14.1 (0.4)	14.1 (0.5)	57.1 (0.2)	73.1 (0.8)	39.1 (0.5)	2.6 (0.3)
	Primera	42.2 (0.5)	13.8 (0.5)	20.6 (0.5)	22.9 (0.5)	61.8 (0.5)	94.8 (0.1)	43.9 (0.2)	2.2 (0.3)
	QAmden	18.4 (0.2)	5.3 (0.1)	11.7 (0.1)	10.4 (0.1)	53.1 (0.1)	57.8 (0.1)	37.5 (0.2)	10.7 (0.2)
	Centrum	44.4 (0.2)	15.6 (0.3)	22.1 (0.3)	24.8 (0.3)	62.5 (0.3)	95.9 (0.1)	39.8 (0.4)	10.8 (0.4)
	PELMS	43.4 (0.5)	14.8 (0.4)	21.2 (0.5)	23.9 (0.5)	62.1 (0.5)	95.8 (0.1)	40.3 (0.6)	6.8 (0.3)
Multi-XScience	Pegasus-X	24.5 (0.8)	3.8 (0.2)	13.8 (0.4)	10.9 (0.4)	55.2 (0.2)	82.6 (0.9)	29.5 (0.4)	3.4 (0.3)
	Primera	28.9 (0.3)	4.4 (0.0)	15.1 (0.1)	12.5 (0.1)	55.6 (0.1)	89.0 (1.4)	32.3 (0.7)	1.0 (0.2)
	QAmden	20.6 (0.2)	2.9 (0.1)	12.8 (0.1)	9.2 (0.1)	52.7 (0.1)	70.1 (0.1)	27.3 (0.3)	10.8 (0.1)
	Centrum	29.1 (0.2)	4.6 (0.1)	15.2 (0.1)	12.7 (0.1)	55.3 (0.0)	91.0 (0.1)	31.3 (0.4)	7.4 (0.3)
	PELMS	**29.8 (0.2)^***	**4.8 (0.1)^***	**15.4 (0.1)^***	**13.0 (0.1)^***	**55.9 (0.1)^***	91.0 (0.3)^	31.4 (0.4)	3.2 (0.3)
Amazon	Pegasus-X	33.0 (1.0)	7.4 (0.5)	19.9 (0.8)	17.0 (0.8)	63.6 (1.0)	90.8 (0.4)	13.2 (0.8)	20.9 (3.5)
	Primera	31.3 (0.4)	6.3 (0.5)	18.2 (0.5)	15.3 (0.6)	62.5 (0.5)	91.2 (0.2)	13.4 (0.4)	12.5 (2.0)
	QAmden	21.8 (0.0)	2.8 (0.0)	14.0 (0.0)	9.5 (0.0)	56.0 (0.0)	75.3 (0.0)	12.0 (0.0)	8.6 (0.0)
	Centrum	28.1 (0.4)	4.8 (0.1)	16.4 (0.1)	13.1 (0.2)	59.0 (0.1)	90.7 (0.1)	12.3 (0.2)	20.5 (1.1)
	PELMS	**34.9 (0.6)^***	**8.8 (0.3)^***	**22.4 (0.2)^***	**19.0 (0.4)^***	**65.9 (0.3)^***	91.2 (0.2)^	**14.7 (0.6)^***	**36.7 (2.0)^***
Yelp	Pegasus-X	31.3 (0.9)	7.3 (0.6)	19.8 (0.8)	16.5 (0.8)	63.7 (0.6)	89.3 (0.2)	14.1 (0.4)	24.3 (2.9)
	Primera	29.8 (0.9)	6.2 (0.5)	18.0 (0.5)	14.9 (0.7)	62.4 (0.5)	91.0 (0.2)	14.9 (0.5)	10.5 (1.9)
	QAmden	17.5 (0.4)	2.9 (0.1)	12.2 (0.3)	8.5 (0.2)	55.4 (0.3)	70.4 (0.2)	11.1 (0.5)	10.6 (0.2)
	Centrum	28.1 (0.8)	5.6 (0.3)	16.8 (0.2)	13.8 (0.5)	59.5 (0.5)	90.2 (0.6)	11.6 (1.2)	33.6 (4.2)
	PELMS	**34.2 (0.3)^***	**9.3 (0.2)^***	**21.5 (0.5)^***	**19.0 (0.3)^***	**65.8 (0.3)^***	**91.7 (0.3)^***	15.5 (0.6)^	24.2 (1.7)
DUC2004	Pegasus-X	31.7 (0.1)	5.6 (0.1)	15.9 (0.0)	14.1 (0.1)	57.6 (0.1)	91.1 (0.5)	4.7 (0.1)	3.4 (0.1)
	Primera	33.4 (1.1)	7.7 (0.6)	17.2 (0.8)	16.4 (0.9)	58.6 (0.4)	86.6 (5.8)	16.0 (1.1)	1.4 (0.8)
	QAmden	22.5 (0.0)	3.5 (0.0)	13.6 (0.0)	10.3 (0.0)	54.5 (0.0)	68.9 (0.0)	11.9 (0.0)	5.9 (0.0)
	Centrum	34.3 (0.1)	8.1 (0.1)	18.2 (0.1)	17.2 (0.1)	59.1 (0.1)	94.1 (0.0)	14.9 (0.1)	4.2 (0.2)
	PELMS	**34.8 (0.5)^***	7.7 (0.2)	17.6 (0.4)	16.8 (0.3)	**60.0 (0.2)^***	93.8 (0.2)	15.8 (0.6)	4.5 (0.6)
MetaTomatoes	Pegasus-X	29.1 (0.4)	4.2 (0.0)	13.3 (0.1)	11.8 (0.1)	54.0 (0.3)	92.1 (0.7)	4.8 (0.3)	13.8 (2.3)
	Primera	30.6 (0.4)	5.7 (0.4)	14.5 (0.2)	13.6 (0.4)	56.1 (0.5)	88.7 (1.8)	5.8 (0.3)	18.2 (7.5)
	QAmden	16.3 (0.2)	2.1 (0.1)	10.2 (0.2)	7.1 (0.1)	48.0 (0.1)	67.7 (0.2)	4.1 (0.1)	55.5 (0.4)
	Centrum	32.4 (0.0)	5.6 (0.1)	14.9 (0.0)	13.9 (0.1)	55.1 (0.1)	93.5 (0.1)	4.7 (0.1)	23.5 (1.6)
	PELMS	**32.7 (0.2)^***	**7.1 (0.1)^***	**16.0 (0.1)^***	**15.5 (0.1)^***	55.5 (0.1)	91.3 (0.1)	**7.5 (0.1)^***	34.8 (0.3)

Table 14: 64-shot results with adapter (5%) training. In our experiments, we average over 5 runs, each with unique random seeds. We report the mean and (std) values. Green and ^ indicate PELMS outperforms all baselines. **Bold** and * indicate improvement is statistically significant (one-tailed paired t-test with each baseline, $p < 0.05$).

Dataset	Model	R1	R2	RL	RG	BertS	DiscoS	MDSummaC	N-gram Novelty
MultiNews	Pegasus-X	47.4 (0.4)	18.8 (0.4)	24.3 (0.2)	27.9 (0.3)	65.4 (0.4)	95.9 (0.4)	26.8 (1.4)	18.4 (1.4)
	Primera	47.6 (0.3)	18.4 (0.6)	24.2 (0.2)	27.7 (0.4)	65.2 (0.4)	96.5 (0.7)	34.6 (1.1)	16.6 (0.8)
	QAmden	48.2 (0.1)	19.5 (0.2)	24.5 (0.0)	28.4 (0.2)	65.5 (0.2)	96.6 (0.0)	31.9 (1.4)	19.3 (1.7)
	Centrum	46.5 (0.7)	18.0 (0.6)	23.8 (0.4)	27.1 (0.5)	64.5 (0.5)	96.4 (0.2)	35.0 (2.0)	14.6 (1.6)
	PELMS	47.7 (0.3)	18.7 (0.3)	24.1 (0.4)	27.8 (0.3)	65.2 (0.2)	96.6 (0.4)^	33.5 (0.8)	17.2 (1.2)
Multi-XScience	Pegasus-X	32.5 (0.6)	7.1 (0.3)	18.0 (0.4)	16.1 (0.5)	57.9 (0.3)	89.2 (0.6)	19.5 (0.7)	30.2 (1.4)
	Primera	32.5 (0.3)	6.8 (0.3)	17.6 (0.3)	15.7 (0.3)	57.6 (0.1)	90.6 (0.2)	21.4 (1.0)	26.8 (1.3)
	QAmden	33.2 (0.4)	7.1 (0.2)	18.0 (0.1)	16.2 (0.1)	57.8 (0.2)	90.4 (0.3)	21.8 (0.7)	25.7 (0.9)
	Centrum	32.4 (0.3)	6.6 (0.1)	17.4 (0.1)	15.5 (0.1)	57.3 (0.1)	90.7 (0.2)	23.1 (0.9)	22.3 (1.4)
	PELMS	32.8 (0.2)	7.1 (0.3)^	18.0 (0.4)^	16.1 (0.3)	57.9 (0.3)^	90.1 (0.4)	23.2 (0.8)^	23.0 (1.2)
Amazon	Pegasus-X	34.8 (0.5)	8.2 (0.4)	22.0 (0.2)	18.4 (0.4)	65.1 (0.6)	91.0 (0.4)	11.9 (0.8)	32.0 (2.0)
	Primera	35.4 (0.8)	8.2 (0.4)	21.6 (0.6)	18.4 (0.6)	64.9 (0.6)	92.0 (0.4)	10.8 (1.1)	34.1 (7.3)
	QAmden	34.1 (0.5)	7.7 (0.2)	21.3 (0.2)	17.8 (0.1)	64.0 (0.2)	91.8 (0.3)	12.2 (0.1)	24.4 (2.6)
	Centrum	30.6 (0.6)	6.1 (0.3)	18.7 (0.3)	15.1 (0.4)	61.0 (0.3)	90.6 (0.3)	13.0 (0.8)	16.5 (4.0)
	PELMS	**36.9 (0.5)^***	**9.2 (0.5)^***	**23.8 (0.3)^***	**20.1 (0.5)^***	**67.4 (0.4)^***	**92.5 (0.3)^***	13.6 (1.0)^	38.6 (4.4)^
Yelp	Pegasus-X	34.9 (0.4)	9.5 (0.2)	22.1 (0.3)	19.4 (0.3)	66.2 (0.3)	91.3 (0.3)	13.7 (0.3)	33.4 (1.5)
	Primera	36.3 (0.6)	9.8 (0.5)	23.2 (0.7)	20.2 (0.7)	67.2 (0.4)	92.4 (0.2)	12.8 (0.5)	40.0 (2.8)
	QAmden	34.7 (0.4)	9.5 (0.3)	22.8 (0.3)	19.6 (0.3)	66.3 (0.2)	91.2 (0.3)	15.1 (0.6)	30.0 (0.9)
	Centrum	31.7 (0.6)	7.4 (0.3)	19.3 (0.4)	16.5 (0.4)	62.8 (0.4)	91.3 (0.1)	13.8 (0.3)	19.7 (1.2)
	PELMS	36.3 (0.4)^	**10.3 (0.3)^***	23.0 (0.3)	20.5 (0.3)^	67.2 (0.3)^	91.3 (0.3)	14.4 (0.4)	28.5 (4.6)
DUC2004	Pegasus-X	33.2 (0.5)	6.6 (0.5)	16.8 (0.3)	15.4 (0.6)	58.4 (0.4)	91.2 (0.2)	4.7 (0.2)	6.4 (2.0)
	Primera	36.0 (0.5)	8.1 (0.3)	18.7 (0.3)	17.6 (0.3)	60.1 (0.3)	94.3 (0.4)	17.5 (0.6)	3.1 (0.5)
	QAmden	35.1 (1.2)	7.8 (0.3)	18.2 (0.2)	17.1 (0.4)	59.4 (0.5)	88.3 (2.0)	13.9 (0.8)	5.9 (0.5)
	Centrum	34.5 (0.1)	7.8 (0.1)	18.2 (0.2)	17.0 (0.2)	59.4 (0.3)	93.8 (0.2)	14.3 (0.3)	3.6 (0.2)
	PELMS	**37.7 (1.0)^***	**9.5 (0.5)^***	**19.3 (0.5)^***	**19.1 (0.6)^***	**61.4 (0.5)^***	94.5 (0.6)^	15.7 (0.6)	**10.6 (1.2)^***
MetaTomatoes	Pegasus-X	36.5 (0.3)	9.3 (0.1)	17.6 (0.1)	18.2 (0.1)	61.1 (0.3)	94.2 (0.1)	1.1 (0.1)	69.6 (0.9)
	Primera	34.7 (0.4)	8.5 (0.4)	17.1 (0.2)	17.1 (0.3)	60.7 (0.2)	93.8 (0.4)	2.7 (0.8)	63.7 (2.9)
	QAmden	34.1 (0.3)	7.8 (0.3)	16.5 (0.2)	16.3 (0.3)	60.0 (0.2)	93.2 (0.4)	1.6 (0.2)	65.1 (1.9)
	Centrum	34.4 (0.3)	8.1 (0.3)	16.8 (0.2)	16.7 (0.3)	59.3 (0.2)	93.7 (0.4)	1.2 (0.1)	63.9 (0.8)
	PELMS	34.1 (0.5)	8.1 (0.5)	16.7 (0.1)	16.7 (0.3)	60.6 (0.1)	92.9 (0.5)	1.7 (0.4)	68.4 (4.8)

Table 15: Full-shot results with full-parameter training. In our experiments, we average over 5 runs, each with unique random seeds. We report the mean and (std) values. Green and ^ indicate PELMS outperforms all baselines. **Bold** and * indicate improvement is statistically significant (one-tailed paired t-test with each baseline, $p < 0.05$).

Dataset	Model	R1	R2	RL	RG	BertS	DiscoS	MDSummaC	N-gram Novelty
MultiNews	Pegasus-X	44.6 (0.4)	16.3 (0.4)	22.6 (0.3)	25.4 (0.3)	63.8 (0.5)	95.8 (0.3)	35.6 (0.4)	10.4 (2.1)
	Primera	44.4 (0.4)	15.4 (0.3)	22.0 (0.5)	24.7 (0.4)	63.1 (0.3)	96.1 (0.2)	42.2 (0.6)	6.0 (0.9)
	QAmden	44.1 (0.2)	15.7 (0.1)	22.1 (0.3)	24.9 (0.2)	63.3 (0.2)	95.6 (0.1)	38.8 (0.1)	9.4 (0.0)
	Centrum	44.2 (0.5)	15.7 (0.8)	22.2 (0.7)	24.9 (0.8)	62.9 (0.6)	95.9 (0.0)	41.0 (0.3)	7.9 (0.3)
	PELMS	44.6 (0.2)^	15.4 (0.5)	21.9 (0.3)	24.7 (0.2)	63.1 (0.2)	96.2 (0.4)^	41.4 (0.6)	8.5 (0.6)
Multi-XScience	Pegasus-X	30.4 (0.2)	6.1 (0.3)	16.8 (0.3)	14.6 (0.3)	57.1 (0.2)	87.6 (0.3)	21.9 (0.3)	20.8 (0.8)
	Primera	30.9 (0.1)	5.8 (0.1)	16.4 (0.3)	14.3 (0.2)	56.5 (0.4)	90.8 (0.4)	28.2 (0.1)	11.0 (0.4)
	QAmden	29.4 (0.4)	5.2 (0.1)	16.1 (0.1)	13.5 (0.2)	56.2 (0.2)	88.3 (0.1)	26.2 (0.1)	12.2 (0.3)
	Centrum	30.6 (0.2)	5.3 (0.2)	16.2 (0.1)	13.8 (0.2)	56.0 (0.1)	90.7 (0.3)	27.6 (0.3)	12.6 (0.3)
	PELMS	30.7 (0.2)	5.7 (0.4)	16.4 (0.4)	14.2 (0.2)	56.5 (0.4)	89.0 (0.2)	27.2 (0.2)	11.0 (0.9)
Amazon	Pegasus-X	25.4 (0.1)	3.3 (0.0)	14.9 (0.0)	10.8 (0.0)	58.6 (0.0)	85.5 (0.0)	14.8 (0.1)	3.9 (0.1)
	Primera	26.8 (0.7)	4.2 (0.0)	15.6 (0.5)	12.1 (0.2)	59.2 (0.2)	86.9 (0.1)	12.5 (0.1)	2.5 (1.7)
	QAmden	21.7 (0.0)	2.7 (0.1)	13.9 (0.1)	9.4 (0.2)	56.0 (0.2)	75.3 (0.1)	12.0 (0.0)	8.7 (0.1)
	Centrum	27.9 (0.1)	4.8 (0.0)	16.3 (0.1)	13.0 (0.0)	58.9 (0.0)	90.5 (0.1)	12.5 (0.4)	20.5 (1.0)
	PELMS	**32.0 (0.2)^***	**7.0 (0.3)^***	**19.4 (0.1)^***	**16.3 (0.2)^***	**63.1 (0.2)^***	**91.2 (0.1)^***	13.3 (0.5)	**27.6 (0.7)^***
Yelp	Pegasus-X	21.0 (0.2)	3.1 (0.1)	12.8 (0.2)	9.5 (0.2)	58.2 (0.1)	81.4 (0.3)	13.9 (0.1)	5.5 (0.3)
	Primera	27.1 (0.0)	4.9 (0.0)	16.6 (0.1)	13.0 (0.0)	60.5 (0.0)	84.3 (0.1)	11.3 (0.1)	2.0 (0.0)
	QAmden	17.4 (0.2)	2.5 (0.1)	12.0 (0.1)	8.1 (0.2)	55.2 (0.1)	70.5 (0.1)	11.2 (0.1)	10.4 (0.1)
	Centrum	28.5 (0.2)	5.6 (0.2)	16.9 (0.2)	13.9 (0.2)	59.9 (0.3)	90.8 (0.2)	12.9 (0.3)	24.6 (3.0)
	PELMS	**32.5 (0.2)^***	**8.4 (0.1)^***	**20.1 (0.1)^***	**17.7 (0.1)^***	**64.3 (0.1)^***	**91.6 (0.2)^***	**14.4 (0.2)^***	**27.9 (0.5)^***
DUC2004	Pegasus-X	31.7 (0.1)	5.8 (0.1)	15.9 (0.1)	14.3 (0.2)	57.7 (0.0)	91.1 (0.3)	4.6 (0.2)	3.3 (0.1)
	Primera	32.1 (0.3)	6.8 (0.2)	16.3 (0.4)	15.2 (0.3)	58.1 (0.2)	79.7 (0.2)	14.6 (1.5)	2.9 (2.6)
	QAmden	22.9 (0.0)	3.5 (0.0)	13.5 (0.0)	10.3 (0.0)	54.7 (0.0)	68.2 (0.0)	11.6 (0.0)	5.8 (0.0)
	Centrum	34.3 (0.3)	8.1 (0.2)	18.1 (0.2)	17.1 (0.2)	59.1 (0.2)	94.1 (0.0)	14.6 (0.3)	4.2 (0.1)
	PELMS	34.1 (0.1)	6.9 (0.1)	16.7 (0.1)	15.8 (0.1)	**59.3 (0.1)^***	93.1 (0.1)	15.1 (0.2)^	**7.3 (0.3)^***
MetaTomatoes	Pegasus-X	35.3 (0.2)	9.0 (0.2)	17.3 (0.2)	17.6 (0.2)	60.0 (0.2)	92.6 (0.2)	1.3 (0.1)	65.4 (0.6)
	Primera	32.0 (0.1)	7.9 (0.2)	16.2 (0.1)	16.0 (0.2)	59.7 (0.3)	91.4 (0.3)	4.4 (0.2)	50.8 (2.5)
	QAmden	30.0 (0.5)	7.8 (0.1)	15.7 (0.3)	15.4 (0.2)	58.7 (0.1)	87.0 (0.7)	1.8 (0.2)	57.6 (0.4)
	Centrum	32.4 (0.1)	7.5 (0.1)	16.2 (0.1)	15.8 (0.1)	57.1 (0.1)	90.1 (0.5)	1.6 (0.1)	59.4 (0.8)
	PELMS	32.9 (0.2)	7.0 (0.1)	16.0 (0.2)	15.5 (0.2)	55.5 (0.1)	91.6 (0.1)	**7.5 (0.1)^***	34.3 (2.2)

Table 16: Full-shot results with adapter (5%) training. In our experiments, we average over 5 runs, each with unique random seeds. We report the mean and (std) values. Green and ^ indicate PELMS outperforms all baselines. **Bold** and * indicate improvement is statistically significant (one-tailed paired t-test with each baseline, $p < 0.05$).