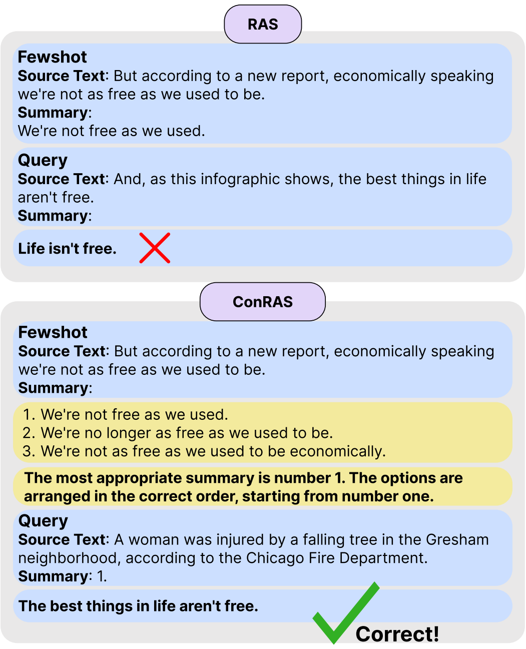
ConRAS: Contrastive In-context Learning Framework for Retrieval-Augmented Summarization

Juseon-Do[1†], **Sungwoo Han**[1†], [*]**Jingun Kwon**[1],
Hidetaka Kamigaito[2], **and Manabu Okumura**[3]
[1]Chungnam National University, [2]Nara Institute of Science and Technology (NAIST)
[3]Institute of Science Tokyo
{doju00, 77sungwhan}@o.cnu.ac.kr
jingun.kwon@cnu.ac.kr
kamigaito.h@is.naist.jp
oku@pi.titech.ac.jp

Abstract

Contrastive learning (CL) has achieved remarkable progress in natural language processing (NLP), primarily as a paradigm for pre-training and fine-tuning. However, its potential during the generation phase, particularly in in-context learning (ICL)-based retrieval-augmented summarization, remains largely unexplored. While previous studies have attempted to incorporate negative samples into ICL prompts, these methods do not enforce a true contrastive objective that encourages separation of positive and negative samples in the representation space. In this paper, we first demonstrate through preliminary experiments that small language models (SLMs) can interpret contrastive prompts and effectively distinguish between positive and negative samples during inference, without any parameter updates. Building on these findings, we propose ConRAS, a novel framework that injects contrastive objectives into ICL-based retrieval-augmented summarization. Extensive experiments and in-depth analysis on three summarization benchmarks using four SLMs show that ConRAS consistently outperforms state-of-the-art retrieval-augmented methods, achieving significant improvements in summary quality.



Fewshot
Source Text: But according to a new report, economically speaking we're not as free as we used to be.
Summary:
We're not free as we used.

Query
Source Text: And, as this infographic shows, the best things in life aren't free.
Summary:

Life isn't free.

Fewshot
Source Text: But according to a new report, economically speaking we're not as free as we used to be.
Summary:

1. We're not free as we used.
2. We're no longer as free as we used to be.
3. We're not as free as we used to be economically.

The most appropriate summary is number 1. The options are arranged in the correct order, starting from number one.

Query
Source Text: A woman was injured by a falling tree in the Gresham neighborhood, according to the Chicago Fire Department.
Summary: 1.

The best things in life aren't free.

Figure 1: Unlike the standard retrieval-augmented method, our ConRAS leverages contrastive prompts and a positive sample with ordered negative samples, enabling the model to correctly identify and generate concise summaries.

1 Introduction

Retrieval-augmented summarization (RAS) is a technique that enhances large language models (LLMs) and enables them to generate more informative and concise summaries by incorporating relevant information from external resources at inference time, often via in-context learning (ICL) (Qiu et al., 2022; Su et al., 2022; Wang et al., 2023; Shao et al., 2023). This paradigm has recently achieved strong performance across diverse summarization benchmarks, as it allows models to ground their outputs in a broader set of evidence than what is present in the original input (Juseon-Do et al., 2025; Edge et al., 2025).

Meanwhile, contrastive learning (CL) has emerged as a powerful approach for improving representation learning in natural language processing (Radford et al., 2021; Liu and Liu, 2021; Yan et al., 2021; Ray et al., 2024). By explicitly encouraging models to pull positive pairs together and push positive-negative pairs apart in the embedding space, CL consistently yields improvements in both pre-training and fine-tuning scenarios (Chen et al., 2020; Radford et al., 2021). In particular, CL-based approaches have proven effective for sentence embedding and summarization tasks (Zhong et al., 2020; Liu and Liu, 2021).

[*] corresponding author
[†] Equal Contribution

Findings of the Association for Computational Linguistics: EACL 2026, pages 4885–4900
March 24-29, 2026 ©2026 Association for Computational Linguistics

Despite these advances, the integration of contrastive learning into the inference stage of retrieval-augmented summarization remains largely unexplored. Previous work on CL in summarization focuses only on either pre-training or fine-tuning; it is still unclear whether contrastive signals can be leveraged at inference time, within ICL-based approaches, especially to benefit models that cannot be updated in post-deployment. The only attempts so far have incorporated negative samples into ICL prompts as a form of error correction (Mo et al., 2024), simply presenting both the correct answer and challenging distractors in the prompt. These approaches, however, do not explicitly encourage models to separate positive and negative samples during generation.

This question is especially important for small language models (SLMs), which are generally of under 8 billion parameters and are designed for efficient, real-world, or on-device deployment (Aminabadi et al., 2022; Pope et al., 2022; Sheng et al., 2023). While SLMs offer faster inference and lower resource costs (Sanh et al., 2020; Liu et al., 2024), they cannot benefit from further fine-tuning after deployment. Thus, it is crucial to know whether SLMs can exploit contrastive objectives purely through ICL prompts at inference time, without any parameter updates. In this paper, we address this gap through the following research questions: (1) Can SLMs leverage contrastive prompts in ICL to effectively distinguish between positive and negative samples at inference time? (2) Does injecting contrastive signals into ICL prompts during retrieval-augmented summarization improve summary quality without additional parameter updates?

To answer these questions, we first conduct preliminary experiments demonstrating that SLMs can indeed recognize and utilize contrastive prompts. By analyzing the embedding space induced by contrastive prompts, we find that SLMs systematically assign higher similarity to positive samples and can effectively separate them from negatives, even without parameter updates.

Building on these insights, we propose ConRAS, a principled yet simple contrastive in-context learning framework for retrieval-augmented summarization. ConRAS injects contrastive signals directly into ICL prompts, explicitly guiding SLMs to maximize the representational distance between positive and negative samples. This method improves summary quality with no additional training and minimal inference overhead. Figure 1 shows the conceptual differences between standard summarization and our contrastive approach.

Affix	Content
Prefix	Please summarize the following sentence. Sentence: {Random Source} Summary: 1. {Random Positive Sample}
NS	2. {Random Negative Sample$_1$} 3. {Random Negative Sample$_2$} ⋮ 9. {Random Negative Sample$_8$}
CE	The options are arranged in the correct order, starting from number one.
Postfix	This sentence "{Random Negative Sample$_i$}" means in one word:

Table 1: Instruction format for extracting embeddings for positive and negative samples. We repeatedly prompted the SLMs for each embedding by varying the postfix in the instruction.

We conduct extensive experiments and in-depth analysis across three standard summarization benchmarks using four different SLMs. The results demonstrate that ConRAS consistently outperforms state-of-the-art retrieval-augmented ICL methods, achieving significant improvements. Furthermore, we demonstrate that ConRAS can improve performance as model size increases and is also effective in reasoning settings. The code will be available at https://github.com/JuseonDo/ConRAS.

2 Preliminary Experiments

While contrastive learning has been widely adopted during model training, it remains unclear whether SLMs can benefit from contrastive signals introduced solely at inference time. To address this, we conducted preliminary experiments to investigate whether SLMs can distinguish between positive and negative samples and modulate the degree of separation between embeddings in response to explicit contrastive instructions, without any parameter updates.

Synthetic Dataset Construction. To isolate the model's response to contrastive signals without interference from prior knowledge, we constructed synthetic data by generating random strings for the source, positive, and negative samples. Specifically, we created 200 sets, each containing 10 random strings (1–50 characters each), serving as the source, positive (reference), and negative samples. For each instance, we prompted the SLMs with a specially designed contrastive prompt to obtain distinct representations for each sample. The embedding vectors were extracted from the final hidden state of the decoder for subsequent analysis.

Prompt Design. Table 1 shows the instruction format used for extracting embeddings of positive and negative samples. The prompt consists of a prefix that asks for a summary, a list of candidate summaries including one positive and multiple negatives, an explicit contrastive explanation (CE), and a postfix to focus the model's representation (Cheng et al., 2025). The CE component is structured natural language guidance that explicitly instructs the model to order negative samples by increasing distance from the positive sample. We evaluated the following three prompt settings:

(1) **w/o NS & CE:** A minimal prompt with only the prefix and postfix, excluding both negative samples (NS) and the contrastive explanation.

(2) **w/o CE:** A prompt with the prefix, NS, and postfix, but omitting the CE.

(3) **Contrastive Prompt (CP):** The full prompt including the prefix, NS, CE, and postfix.

Model and Evaluation Setup. We experimented with Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct (Grattafiori et al., 2024), and Qwen3-1.7B (Zhang et al., 2025). For each negative sample (Number 2–9), we calculated the average Euclidean distance from the positive sample across all sets. A higher average distance indicates greater separation between positive and negative samples in the embedding space. Appendix A includes used examples for prompts.

To evaluate the presence and strength of monotonic trends, we employed the Mann–Kendall test (Mann, 1945; Kendall and Gibbons, 1990), which is a widely used non-parametric method for detecting monotonic trends without assuming normality or linearity. We report Kendall's τ coefficient to quantify the strength of the trend, along with the corresponding p-value to assess statistical significance. To estimate the magnitude of the trend, we used the Theil–Sen slope estimator (Theil, 1950), which computes the median slope across all data point pairs. This approach provides robust and outlier-resistant estimates of trend slopes.

Results. Figure 2 shows the average Euclidean distance between the positive sample and each negative sample with its position in the instruction. Baseline correction was applied such that the first distance was set to zero. In the **w/o NS & CE** setting, we observe irregular patterns, and in some cases, distances even decrease with positions. In the **w/o CE** setting, the distances show an increasing trend, but the effect is not significant. By contrast, the **CP** setting shows a clear monotonic increase. These results indicate that providing negative samples can help the model better distinguish between positive and negative samples. Furthermore, explicitly including a contrastive explanation enables the model to accurately interpret the instruction and modulate embedding distances according to the specified order. Table 2 shows the results of assessing statistical significance for increasing trends. In the **CP** setting, the values of Kendall's τ were consistently higher than the **w/o CE** and **w/o NS & CE** settings, indicating a more robust monotonic trend when both **NS** and **CE** are applied. Furthermore, a statistically significant increasing trend ($p < 0.05$) was observed across all models. These findings directly answer our first research question: SLMs can indeed leverage contrastive prompts in ICL to produce systematically separated representations for positive and negative samples during inference.

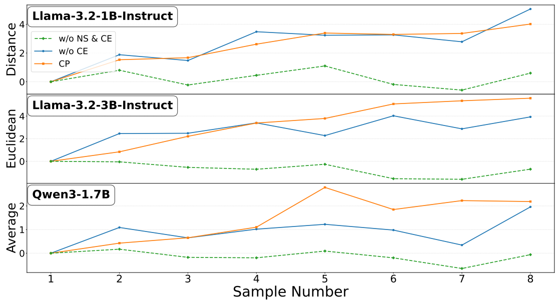
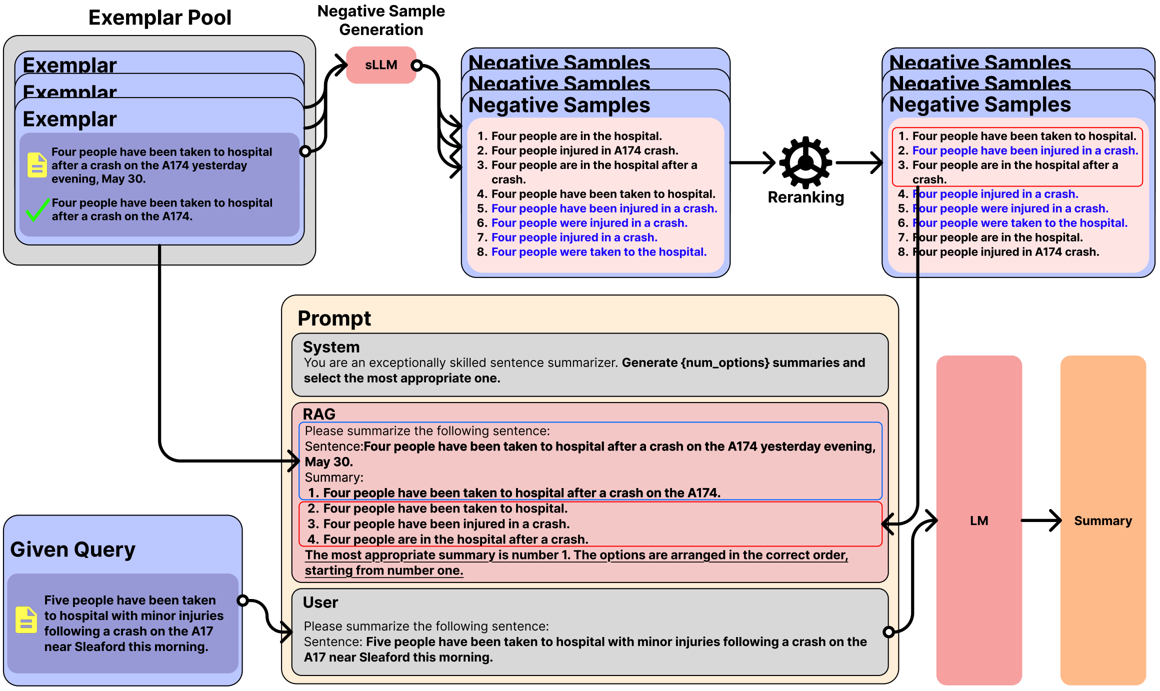


Figure 2: Results of preliminary experiments.

Model	Setting	Trend	p-value	τ	slope
Llama-3.2-1B-Instruct	NS ✗ & CE ✗	no	0.9015	-0.0714	-0.0353
	NS ✓ & CE ✗	no	0.0635	0.5714	**0.5626**
	CP	increasing	**0.0044**	**0.8571**	0.5040
Llama-3.2-3B-Instruct	NS ✗ & CE ✗	decreasing	0.0354	-0.6429	-0.2014
	NS ✓ & CE ✗	no	0.0635	0.5714	0.3529
	CP	increasing	**0.0008**	**1.0000**	**0.8415**
Qwen3-1.7B	NS ✗ & CE ✗	no	0.1735	-0.4286	-0.0390
	NS ✓ & CE ✗	no	0.3865	0.2857	0.1704
	CP	increasing	**0.0187**	**0.7143**	**0.3637**

Table 2: Results of the Mann–Kendall trend test and the Theil–Sen slope estimator. **Trend** indicates statistically significant trend.



Four people have been taken to hospital after a crash on the A174 yesterday evening, May 30.

✓ Four people have been taken to hospital after a crash on the A174.

1. Four people are in the hospital.
2. Four people injured in A174 crash.
3. Four people are in the hospital after a crash.
4. Four people have been taken to hospital.
5. Four people have been injured in a crash.
6. Four people were injured in a crash.
7. Four people injured in a crash.
8. Four people were taken to the hospital.

1. Four people have been taken to hospital.
2. Four people have been injured in a crash.
3. Four people are in the hospital after a crash.
4. Four people injured in a crash.
5. Four people were injured in a crash.
6. Four people were taken to the hospital.
7. Four people are in the hospital.
8. Four people injured in A174 crash.

System
You are an exceptionally skilled sentence summarizer. **Generate {num_options} summaries and select the most appropriate one.**

RAG
Please summarize the following sentence:
Sentence:**Four people have been taken to hospital after a crash on the A174 yesterday evening, May 30.**
Summary:
1. **Four people have been taken to hospital after a crash on the A174.**
2. **Four people have been taken to hospital.**
3. **Four people have been injured in a crash.**
4. **Four people are in the hospital after a crash.**

The most appropriate summary is number 1. The options are arranged in the correct order, starting from number one.

User
Please summarize the following sentence:
Sentence: **Five people have been taken to hospital with minor injuries following a crash on the A17 near Sleaford this morning.**

Five people have been taken to hospital with minor injuries following a crash on the A17 near Sleaford this morning.

Figure 3: Overview of the ConRAS framework. Given a query and an exemplar pool, negative samples are generated using an SLM and arranged in a specific order. The instruction-based summarization prompt, which includes the query, exemplars, and reranked negative samples, is then provided to the SLM within a contrastive in-context learning framework for retrieval-augmented summarization.

3 Our ConRAS

Preliminary experiments lead us to focus on developing a contrastive in-context learning framework specifically tailored for retrieval-augmented summarization. Building on these insights, we propose **ConRAS**, a principled and conceptually simple method that leverages contrastive prompts to explicitly guide SLMs in distinguishing between positive and negative samples. Figure 3 shows an overview of the ConRAS framework. By incorporating both negative samples and a contrastive explanation into the prompt, ConRAS enables the model to maximize the representational distance between positive and negative samples, thereby enhancing summarization quality without additional parameter updates.

Retrieval-Augmented Summarization. Given an input source q and an exemplar pool of source-positive pairs $\mathcal{D} = \{(q'_i, p_i) \mid 0 \leq i \leq n\}$, the task of retrieval-augmented summarization aims to generate a concise summary s that is comprehensive with respect to the source q. Specifically, the model retrieves k source-positive pairs $\mathcal{R} = \{(q'_j, p_j) \mid 0 \leq j \leq k\}$ from $\mathcal{D}$ based on the relevance between q and q'_j. The summary s is then generated conditioned on the source q and the retrieved exemplar pool $\mathcal{R}$.

Negative Sample Generation and Ordering. To generate negative samples, we prompt the SLM with all source texts q' from the exemplar pool, generating summaries for each. To obtain multiple model-generated summaries per source, we employ beam search during generation. This approach allows us to construct multiple negative summaries for use in contrastive learning for summarization (Liu and Liu, 2021). We assign ROUGE scores between the gold and negative samples in order to obtain their ranking.

Instruction-based Summarization. We adopt an instruction-based prompt format to improve the discriminative ability of SLMs in retrieval-augmented summarization. Given a query and a set of retrieved exemplars, we construct a prompt in which the model is presented with the query and a list of candidate summaries. Among these candidates, the first one corresponds to the reference summary, which is a positive sample, while the others are model-generated negative samples. This instruction-based prompting encourages the model to explicitly compare candidate summaries, thereby improving its ability to distinguish high-quality summaries from distractors.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluated our method on three summarization benchmarks: Google (**Google**), Broadcast (**Broad**), and BNC (**BNC**) (Filippova and Altun, 2013; Clarke and Lapata, 2008). Table 3 shows the statistics of each dataset. The **Google** dataset comprises automatically generated sentence summaries using syntactic dependency trees extracted from news headlines and lead sentences of each article, with a gold compression ratio of 0.45 in the test set. In contrast, the **Broad** and **BNC** datasets consist of human-written summaries, with gold compression ratios of 0.71 and 0.72, respectively.

Evaluation Metrics. We evaluated summary quality using the F_1 scores of ROUGE-1 (**R-1**), ROUGE-2 (**R-2**), and ROUGE-L (**R-L**) (Lin, 2004), as well as BERTScore (**BS**) (Zhang et al., 2020). In addition, we assessed *faithfulness* of generated summaries. For this, we employed AlignScore (**AS**) (Zha et al., 2023) and MiniCheck (**MC**) (Tang et al., 2024).

Implementation Details. We employed the Llama and Qwen model families in our experiments. Specifically, we used Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, and Llama-3.1-8B-Instruct models (Grattafiori et al., 2024), as well as Qwen3-1.7B and Qwen3-4B models (Zhang et al., 2025). We used all-MiniLM-L6-v2 (Wang et al., 2020) to obtain embeddings for constructing the exemplar pool. To measure similarities and distances between texts, we employed FAISS (Douze et al., 2024). For generation, we set the beam search width to 8. In the retrieval-augmented summarization setting, we used four exemplars (shots), and for each shot, we included four candidates: one correct answer (positive) and three negative samples. For the Broad and BNC datasets, which do not have their own training set, we followed previous work (Juseon-Do et al., 2025) by using the BNC dataset as the exemplar pool for Broad, and the Broad dataset as the exemplar pool for BNC, respectively.

Negative Sample Generation and Ordering. For the **Google** dataset, we employed a few-shot prompt to generate hard negative samples during negative sample generation. In contrastive learning, hard negatives, which are negative samples that are similar to positive ones and therefore more difficult for the model to distinguish, are known to be more effective (Gao et al., 2021). Since the **Google** dataset has a validation set, we used the first six samples from the validation set as few-shot examples. In contrast, for the **Broad** and **BNC** datasets, which do not have a validation set, we generated negative samples in a zero-shot manner. For the Llama family, we generated negative samples using the smallest SLM (1B), and for Qwen, we used the 1.7B model. For reranking negative samples, we used R-2 scores to arrange them in descending order relative to the positive sample for each retrieved example.

Dataset	Training	Valid	Test	*Avg Src Len*	*Avg Tgt Len*
Google	200,000	1,000	1,000	24.4 (±9.2)	9.8 (±3.1)
Broad	-	-	1,370	19.8 (±12.8)	13.2 (±8.2)
BNC	-	-	1,629	27.9 (±15.3)	19.3 (±10.7)

Table 3: Statistics of datasets. The values in parentheses indicate the standard deviation of both the source and target lengths, respectively.

Compared Methods. We evaluated several retrieval-augmented state-of-the-art (SOTA) strategies: **NN**, where exemplars are chosen based on semantic similarity to the query using a nearest-neighbor approach (Liu et al., 2022); **MMR**, which employs Maximal Marginal Relevance to balance query relevance and inter-exemplar diversity to encourage LLMs to demonstrate the required reasoning process (Ye et al., 2023); **DL-MMR**, which extends MMR by incorporating target-length diversity among exemplars during retrieval, thus providing LLMs with length-unbiased exemplars (Juseon-Do et al., 2025); and **C-ICL**, which enriches the in-context learning prompt with retrieved exemplars. It also includes examples that the model suffers difficulty in predicting in the training set, regardless of the retrieved exemplars, enabling the model to learn from its own typical mistakes as well as correct reasoning (Mo et al., 2024). Because C-ICL essentially targets the information extraction task, we used R-2 to obtain such examples. We also evaluate our **ConRAS**, which augments the prompt with both positive and negative examples, as well as a contrastive explanation that instructs the model to separate positive and negative candidates in the embedding space. Hyperparameter settings for baselines are in Appendix B.

4.2 Main Results

Table 4 shows the performance of the Llama and Qwen model families on the Google, BNC, and Broad datasets. Except for the performance of

Model	Strategy	Google						BNC						Broadcast					
		R-1	**R-2**	**R-L**	**BS**	**AS**	**MC**	**R-1**	**R-2**	**R-L**	**BS**	**AS**	**MC**	**R-1**	**R-2**	**R-L**	**BS**	**AS**	**MC**
Llama-3.2 -1B-Instruct	NN	62.79	44.29	60.79	0.61	0.95	0.88	62.76	46.21	60.08	0.54	0.89	0.80	70.23	53.18	68.48	0.59	0.90	0.78
	MMR	64.69	47.29	62.85	0.64	0.96	0.88	63.04	46.83	60.42	0.54	0.90	0.80	70.42	53.39	68.35	0.58	0.90	0.78
	DL-MMR	63.39	45.26	61.54	0.62	0.95	0.88	65.47	47.70	62.41	0.56	0.90	0.82	70.41	54.84	70.05	0.61	0.93	0.81
	C-ICL	65.10	49.90	63.53	0.67	0.95	0.88	69.08	54.48	67.25	0.59	0.93	0.84	73.66	58.50	72.55	0.61	0.93	0.81
	ConRAS	**72.27***	**59.39***	**71.71***	**0.72***	**0.98***	**0.91***	**72.83***	**59.51***	**71.66***	**0.61***	**0.95***	**0.86**†	**78.42***	**64.14***	**77.73***	**0.64***	**0.95***	**0.83***
Llama-3.2 -3B-Instruct	NN	72.23	58.67	70.37	0.72	0.97	0.91	78.96	65.10	77.41	0.66	0.89	0.86	80.62	66.60	79.69	0.66	0.90	0.82
	MMR	73.42	61.28	71.64	0.73	0.98	0.91	78.91	65.51	77.55	0.66	0.90	0.86	80.34	66.05	79.36	0.65	0.90	0.82
	DL-MMR	72.74	58.88	70.64	0.73	0.97	0.90	77.65	62.07	74.93	0.65	0.90	0.85	79.75	64.94	78.22	0.64	0.93	0.83
	C-ICL	71.17	57.59	69.28	0.71	0.97	0.90	79.74	66.11	78.50	0.66	0.93	0.87	80.78	66.27	79.74	**0.66**	0.92	0.83
	ConRAS	**77.91***	**67.59***	**77.43***	**0.77***	**0.99***	**0.92***	**80.58***	**67.67***	**80.07***	**0.67**†	**0.95***	**0.89***	**81.38**†	**67.53***	**80.95***	**0.66**†	**0.96***	**0.85***
Llama-3.1 -8B-Instruct	NN	71.21	55.99	68.99	0.70	0.97	0.90	76.10	61.21	73.92	0.64	0.91	0.85	79.39	64.16	77.91	0.65	0.92	0.81
	MMR	73.66	60.78	71.72	0.73	0.98	0.90	76.29	61.33	74.00	0.64	0.91	0.84	79.14	63.67	77.41	0.65	0.91	0.81
	DL-MMR	71.71	56.46	69.40	0.70	0.97	0.90	72.72	54.62	68.74	0.62	0.91	0.83	76.97	60.23	74.54	0.64	0.93	0.80
	C-ICL	72.35	59.02	70.47	0.72	0.97	0.90	76.57	61.50	74.39	0.65	0.93	0.85	78.47	62.79	76.57	0.65	0.94	0.81
	ConRAS	**77.37***	**66.28***	**76.63***	**0.78***	**0.98**	**0.92***	**78.08***	**64.77***	**76.90***	**0.66***	**0.95***	**0.87***	**81.18***	**66.89***	**80.04***	**0.67***	**0.95***	**0.82**
Qwen3-1.7B	NN	72.11	59.84	70.79	0.71	**0.97**	0.90	76.25	61.10	74.17	0.69	0.98	0.86	79.20	63.54	77.77	**0.73**	0.98	0.82
	MMR	71.90	59.36	70.80	0.71	**0.97**	0.90	73.19	56.06	70.22	0.65	0.97	0.84	76.71	59.48	74.44	0.70	0.98	0.80
	DL-MMR	69.93	56.14	68.52	0.69	**0.97**	0.90	75.45	58.40	72.12	0.67	0.97	0.84	77.54	60.67	75.09	0.69	0.97	0.81
	C-ICL	70.09	59.22	69.12	0.70	**0.97**	**0.91**	78.58	64.09	76.68	0.70	0.98	0.86	79.86	64.18	78.06	0.70	0.97	0.82
	ConRAS	**72.73**	**62.66***	**71.67**	**0.72**	**0.97**	**0.91**	**79.93***	**66.55***	**78.98***	**0.72***	**0.99***	**0.88**	**80.50***	**66.24***	**79.88***	**0.73**	**0.99***	**0.84***
Qwen3-4B	NN	72.33	62.55	71.46	0.71	0.97	0.91	80.16	66.29	79.31	**0.73**	**0.98**	0.87	81.00	66.25	80.37	**0.74**	**0.98**	0.83
	MMR	**74.38**	63.09	**73.77**	**0.73**	0.97	0.90	78.41	62.86	76.77	0.71	0.97	0.87	79.83	63.79	78.53	0.72	0.97	0.82
	DL-MMR	71.14	59.17	69.79	0.70	0.97	0.90	78.53	62.45	76.43	0.71	0.97	0.87	78.60	62.20	77.24	0.70	0.97	0.83
	C-ICL	67.99	58.46	67.49	0.68	**0.98**	**0.92**	**81.15**	**67.87**	**80.54**	**0.73**	0.96	**0.89**	80.24	64.63	79.27	0.70	0.96	**0.84**
	ConRAS	73.67	**63.79**	72.79	**0.73**	**0.98**	0.90	80.16	66.43	79.44	0.71	**0.98**	0.87	**81.21**	**67.13**	**80.68**	0.73	0.97	**0.84**

Table 4: Experimental results using NN, MMR, DL-MMR, C-ICL, and ConRAS based on Llama and Qwen families. * and † denote statistically significant (*: p<0.01, †: p<0.05) improvements, compared to the underlined scores (typically the best baseline) on each dataset. We used paired bootstrap resampling with 100,000 random samples (Koehn, 2004) for the significance test.

Qwen3-4B on the BNC dataset, our ConRAS outperforms all baselines. In particular, ConRAS achieves significant improvements over the best baselines. These results directly answer our second research question that injecting contrastive signals into ICL prompts during retrieval-augmented summarization improves summary quality.

Table 5 shows the construction time and resource usage for each method when building the exemplar pool for retrieval-augmented summarization on the Google dataset. It also includes the inference time using Llama-3.2-1B-Instruct. Compared to NN, all other methods require additional memory. We observe that MMR, C-ICL, and ConRAS incur substantially higher pool-construction time; however, only NN and MMR pay an additional retrieval cost at inference whereas C-ICL and ConRAS do not. ConRAS exhibits a higher inference time than the baselines because its prompt includes more text, but this cost is moderate and is outweighed by its quality gains.

Table 6 shows the performance for human evaluation using Llama-3.2-3B-Instruct. We sampled 100 instances from the **Google** dataset. Using MTurk, 120 evaluators with U.S. high school and bachelor's degrees rated the outputs from 1 to 3 (3 is the best) for informativeness (Info.), conciseness (Conc.), and factual consistency (Faith.). Our ConRAS can generate informative, concise, and faithful summaries.

Strategy	Construction	Memory	Retrieve	Inference
NN	-	45.70M	1m19s	14m10s
MMR	9h24m	+361.62G	3h36m	16m40s
DL-MMR	0m0.1s	+8K	2m12s	15m50s
C-ICL	26h33m	+4K	1m19s	24m10s
ConRAS	30h22m	+32.23M	1m19s	35m50s

Table 5: Computation cost comparison for each method.

Model	Strategy	Infor.	Conc.	Faith.
Llama-3.1 -3B-Instruct	NN	2.48	2.73	2.80
	MMR	2.23	2.57	2.75
	DL-MMR	2.27	2.60	2.75
	C-ICL	2.50	2.59	2.82
	ConRAS	**2.80***	**2.91***	**2.91**

Table 6: Human evaluation results.

5 Analysis

For all analysis experiments, we test 300 randomly selected samples by following previous work (Mo et al., 2024).

Impact of the Number of Exemplars (Shots). We investigate the effect of the number of shots, which is the number of retrieved paired exemplars, on the performance using the Google dataset. For ConRAS, each exemplar contains one positive sample

	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS
Strategy	Llama-3.2-1B-Instruct				Llama-3.1-8B-Instruct			
ConRAS (0-shot)	52.07	30.42	49.40	0.56	63.92	49.25	61.07	0.65
ConRAS (1-shot)	59.45	43.08	58.47	0.62	73.56	60.59	72.34	0.73
ConRAS (2-shot)	64.77	49.08	64.02	0.66	74.09	60.95	73.39	0.74
ConRAS (4-shot)	66.67	51.97	65.67	0.68	75.53	63.22	74.57	0.75
ConRAS (8-shot)	71.70	59.10	71.02	0.72	76.03	64.10	75.11	0.76
ConRAS (16-shot)	**73.64***	**61.77***	**72.57***	**0.73***	**77.25***	**65.86***	**76.59***	**0.77***
NN (16-shot)	63.56	46.70	61.04	0.66	70.77	56.79	68.93	0.72
NN (32-shot)	67.73	53.33	65.61	0.69	71.77	57.53	69.36	0.72
NN (64-shot)	69.83	55.50	67.79	0.70	73.69	61.56	71.86	0.73

Table 7: Performance of ConRAS with varying numbers of shots on the **Google** dataset. 0 shot indicates a zero-shot summarization setting without retrieval.

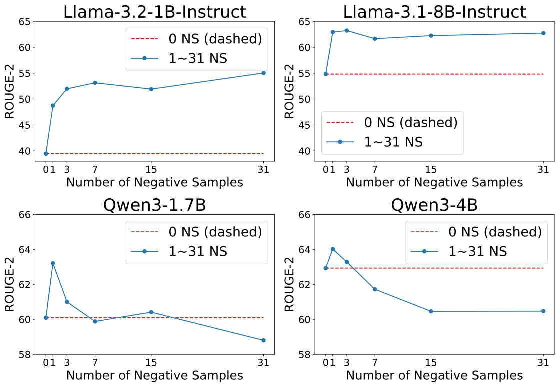


Figure 4: Effect of the number of negative samples on model performance. The dashed red line represents the performance with zero negative samples, while the blue line represents performance when varying the number of negative samples from 1 to 31.

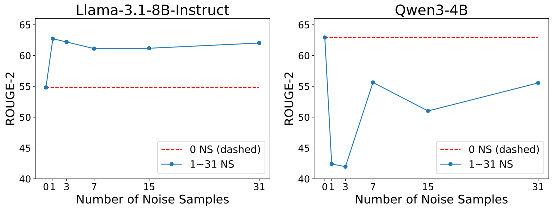


Figure 5: Robustness to noise samples on the Google dataset. The x- and y-axis are the same as Figure 4.

and three negative samples. Maintaining this ratio, in a 4-shot setting, for example, four positive summaries and twelve negative summaries are included. The results are shown in Table 7. As the number of shots increases, ConRAS demonstrates consistent improvements in performance across both model families. Notably, for the Llama family, 16-shot ConRAS surpasses 64-shot NN. For the Qwen family, the results are in Appendix C. These results suggest that ConRAS is particularly effective in low-resource scenarios, where only a limited number of exemplars are available.

Impact of the Number of Negative Samples. We investigate the impact of the number of negative samples on the performance of ConRAS using the Google dataset. The number of paired exemplars (shots) is fixed at four, but we vary the number of negative samples for each retrieved exemplar from 1 to 31. Figure 4 shows the results. For the Llama family, increasing the number of negative samples up to three consistently improves ConRAS performance. In contrast, for the Qwen family, the best performance is achieved with a single negative sample; adding more negative samples leads to a decrease in performance, although using three negative samples still yields better results than using none. To further understand the smaller gains in Qwen3-4B's performance, observed in Table 4, we evaluate the models' robustness to noise by replacing negative samples with random string noise used in the preliminary experiments. The results are shown in Figure 5. Llama-3.1-8B exhibits greater robustness to noise samples than Qwen3-4B. These findings suggest that a model's robustness to noise is closely associated with the effectiveness of ConRAS. Appendix D includes examples for the robustness evaluation and case studies.

Effect of the Order of Negative Samples and the Contrastive Explanation. The ablation results on the Google dataset for the negative sample order and the contrastive explanation are presented in Table 8. Removing either the ordering or the contrastive explanation leads to a significant decline in performance. Furthermore, the Reverse setting results in the lowest scores. This highlights the importance of using negative samples in an order that reflects their similarity to the positive sample, as intended by the ConRAS design. These results confirm that both the proper ordering of negative samples and the explicit contrastive explanation are critical for the effectiveness of ConRAS.

Impact of the Quality of Negative Samples. We first define negative sample quality using R-2 scores computed based on the gold summary on the Google dataset. We then sort the negatives in descending order of R-2, so that negatives with higher R-2 scores are considered higher quality or harder, as they are more similar to the gold summary. The left plot in Figure 6 shows the quality of negative samples. To analyze the impact of negative sample difficulty, we conduct experiments with ConRAS using three negative samples at different quality levels: hardest, intermediate, and easiest. As in contrastive learning, we observe that the use of higher-quality (harder) negative samples

Strategy	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS
	Llama-3.2-3B-Instruct				Llama-3.1-8B-Instruct			
ConRAS	**76.92***	**67.20***	**76.39***	**0.76***	**77.74***	**66.51***	**76.93***	**0.77***
w/o CE	73.22	60.11	72.41	0.74	74.50	60.92	73.30	0.75
w/o Ordering	74.36	62.86	73.73	0.75	74.91	61.91	74.17	0.75
w/o Ordering & CE	73.46	60.28	72.65	0.74	74.35	60.19	73.32	0.75
Reverse	69.14	57.16	67.99	0.68	73.80	60.63	72.68	0.74

Table 8: Results of the ablation study for ConRAS. In the w/o Ordering setting, negative samples are attached in the order of beam output ranking without reranking. The **Reverse** setting indicates that the reranked order of negative samples is reversed.

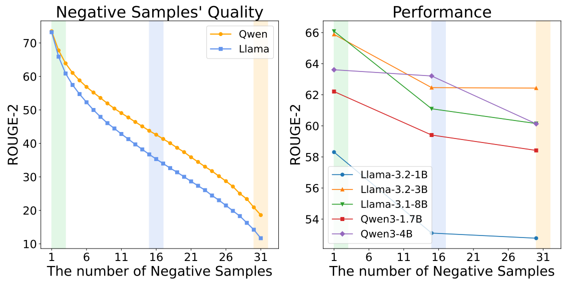


Figure 6: (Left) R-2 scores of generated negative samples, which is sorted from hardest to easiest. Green, Blue, and Orange indicate samples at different quality levels (hard, intermediate, easy). (Right) Performance of ConRAS when using three negative samples of varying quality.

leads to better performance even in ICL retrieval-augmented settings.

Scalability of ConRAS. To evaluate the scalability of ConRAS, we conduct additional experiments in three settings: document-level summarization, increasing LLM size, and LLMs with a reasoning setting. For document-level evaluation, we test ConRAS on the CNN/DM dataset, which is a standard benchmark for single-document summarization. The dataset consists of 287k, 13.4k, and 11.5k for training, development, and testing (Hermann et al., 2015). In this setting, we use only one retrieved exemplar from the training pool due to input length constraints when injecting long documents into LLMs. Because MMR and DL-MMR require multiple exemplars, using only one exemplar reduces them to the same setting as NN. Similarly, C-ICL also requires paired datasets to include examples that the model frequently mispredicts. Thus, these methods cannot be applied in this setting. As shown in Table 9, ConRAS outperforms the NN baseline. ConRAS uses only a single input exemplar and augments the prompt by adding negative samples, without requiring additional paired data for instruction. This result highlights the strong performance of ConRAS even when only one retrieved exemplar is available. To evaluate the performance of ConRAS on larger models and in reasoning settings, we conduct experiments with Llama-3.1-70B-Instruct and Qwen3-1.7B using a reasoning setting. Table 10 shows the results. We observe that ConRAS provides substantial benefits both as model size increases and in settings that provide effective reasoning.

Strategy	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS
	Llama-3.2-1B-Instruct				Llama-3.1-8B-Instruct			
NN	32.49	11.65	29.43	0.21	36.46	13.61	33.01	0.23
ConRAS	**34.58***	**13.73***	**31.58***	**0.22†**	**38.53***	**15.63***	**34.97***	**0.26†**
	Qwen3-1.7B				Qwen3-4B			
NN	35.24	11.39	31.70	0.23	38.84	14.82	35.56	0.26
ConRAS	**36.03†**	**12.76***	**32.59**	**0.24**	**40.23***	**16.42***	**36.45**	**0.27**

Table 9: Performance on the CNN/DM dataset.

Strategy	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS
	Llama-3.1-70B-Instruct				Qwen3-1.7B + Reasoning			
NN	71.33	57.36	69.56	0.71	62.36	49.69	61.28	0.65
MMR	72.59	60.54	71.41	0.71	63.37	50.72	62.04	0.65
DL-MMR	72.44	58.96	70.92	0.72	61.40	47.19	60.02	0.64
C-ICL	74.00	62.16	72.65	0.73	61.59	50.23	60.50	0.63
ConRAS	**77.21***	**65.73***	**76.44***	**0.77***	**64.65**	**54.11***	**63.70†**	**0.66**

Table 10: Performance for larger LLMs and reasoning settings.

Different Instruction To evaluate the generalizability of ConRAS's instruction format used in ConRAS, we conduct additional experiments on the **Google** dataset using Llama-3.2-1B. Instead of using our original summary prompt, we replace it with the following instruction: "Please provide a concise version of the sentence." Similarly, the contrastive prompt is replaced with the following instruction: "Option 1 provides the most accurate summary. The choices are listed in the correct order beginning with number one." Table 11 shows the results. Despite these substantial changes in wording, ConRAS significantly outperforms the other methods.

6 Related Work

Retrieval-Augmented Generation (RAG). Retrieval-Augmented Generation (RAG) has emerged as a promising approach, that has been extensively studied across various NLP tasks (Izacard and Grave, 2021; Guo et al., 2023; Jeong et al., 2024; In et al., 2025). By conditioning LLMs on retrieved exemplars, RAG has shown notable improvements in the quality of generated text. Recent research has explored methods for retrieving more diverse exemplars, such as MMR and DL-MMR. MMR explicitly promotes diversity

Strategy	R-1	R-2	R-L	BS
NN	62.29	45.62	60.88	0.65
MMR	65.25	49.95	63.70	0.67
DL-MMR	64.00	46.93	62.65	0.66
C-ICL	65.40	51.55	64.49	0.67
ConRAS	**74.60***	**63.47***	**74.01***	**0.74***

Table 11: Performance of different instructions for Llama-3.2-1B on the **Google** dataset.

during retrieval, thereby providing LLMs with exemplars that better demonstrate the required reasoning process (Ye et al., 2023). DL-MMR further extends MMR by incorporating target-length diversity, providing LLMs with exemplars that are unbiased with respect to target length for retrieval-augmented summarization (Juseon-Do et al., 2025).

Contrastive Learning (CL). CL has emerged as a powerful paradigm for representation learning in NLP, achieving remarkable progress across a range of tasks (Radford et al., 2021; Gao et al., 2021; Liu and Liu, 2021; Zhang et al., 2022; Bae et al., 2025). CL-based approaches have demonstrated significant effectiveness in sentence embeddings and summarization (Gao et al., 2021; Liu and Liu, 2021; Wu et al., 2022; Zhang et al., 2023; Zhuang et al., 2024). More recently, attempts have been made to incorporate the concept of contrastive learning into in-context learning (ICL) by adding more examples to prompts (Mo et al., 2024). However, these methods primarily serve as error correction mechanisms rather than explicitly enforcing contrastive objectives during the generation process. Furthermore, these examples are selected independently of the retrieved exemplars.

7 Conclusion

In this paper, we introduced ConRAS, a contrastive in-context learning framework for retrieval augmented summarization. Through extensive experiments on three summarization benchmarks, we demonstrated that ConRAS consistently outperforms strong retrieval-augmented baselines. Our analyses revealed that both the quality and ordering of negative samples with including the contrastive explanation are critical to improve performance. Also, ConRAS maintains its effectiveness across various model sizes and reasoning settings, which demonstrates its scalability and generalization.

8 Limitations

The preliminary study is a mechanism probe under synthetic control and is not intended as task-level evidence; it motivates the prompt design, while task relevance is assessed within ConRAS using realistic candidates at both sentence- and document-levels.

Our negatives are beam-search candidates. They are not injected errors but plausible yet less-preferred alternatives. In summarization, hard negatives are commonly defined as candidates close to the reference but weaker in salience (Zhong et al., 2020; Liu and Liu, 2021). We rank candidates by ROUGE-2 and expose this ordering with a contrastive explanation (CE) (Zhong et al., 2020; Liu and Liu, 2021). Removing ordering, reversing it, or dropping CE hurts performance (Table 8); gains also scale with hardness and drop when replacing candidates with noise (Figures 6, 5). Thus, these are effective contrastors, not mere rephrasings. Explicitly erroneous and/or adversarial negatives are left for future work.

9 Ethics Statement

This work relies on publicly available summarization datasets (Google, Broadcast, BNC, and CNN/DM). We did not collect new personal data, and no personal information is introduced by our pipeline.

References

Reza Yazdani Aminabadi, Samyam Rajbhandari, Minjia Zhang, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Jeff Rasley, Shaden Smith, Olatunji Ruwase, and Yuxiong He. 2022. Deepspeed inference: Enabling efficient inference of transformer models at unprecedented scale. *Preprint*, arXiv:2207.00032.

Suyoung Bae, YunSeok Choi, Hyojun Kim, and Jee-Hyong Lee. 2025. SALAD: Improving robustness and generalization through contrastive learning with structure-aware and LLM-driven augmented data. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12724–12738, Albuquerque, New Mexico. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *Preprint*, arXiv:2002.05709.

Zifeng Cheng, Zhonghui Wang, Yuchen Fu, Zhiwei Jiang, Yafeng Yin, Cong Wang, and Qing Gu. 2025. Contrastive prompting enhances sentence embeddings in llms through inference-time steering. *Preprint*, arXiv:2505.12831.

James Clarke and Mirella Lapata. 2008. Global inference for sentence compression an integer linear programming approach. *J. Artif. Int. Res.*, 31(1):399–429.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491, Seattle, Washington, USA. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhicheng Guo, Sijie Cheng, Yile Wang, Peng Li, and Yang Liu. 2023. Prompt-guided retrieval augmentation for non-knowledge-intensive tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10896–10912, Toronto, Canada. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15. MIT Press.

Yeonjun In, Sungchul Kim, Ryan A. Rossi, Mehrab Tanjim, Tong Yu, Ritwik Sinha, and Chanyoung Park. 2025. Diversify-verify-adapt: Efficient and robust retrieval-augmented ambiguous question answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1212–1233, Albuquerque, New Mexico. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.

Juseon-Do, Jaesung Hwang, Jingun Kwon, Hidetaka Kamigaito, and Manabu Okumura. 2025. Considering length diversity in retrieval-augmented summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2489–2500, Albuquerque, New Mexico. Association for Computational Linguistics.

Maurice G. Kendall and Jean D. Gibbons. 1990. *Rank Correlation Methods*, 5 edition. A Charles Griffin Title.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

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

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. 2024. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. *Preprint*, arXiv:2402.14905.

Henry Mann. 1945. Nonparametric tests against trend. *Econometrica*, 13(3):245–259.

Ying Mo, Jiahao Liu, Jian Yang, Qifan Wang, Shun Zhang, Jingang Wang, and Zhoujun Li. 2024. C-ICL: Contrastive in-context learning for information extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10099–10114, Miami, Florida, USA. Association for Computational Linguistics.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Anselm Levskaya, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. 2022. Efficiently scaling transformer inference. *Preprint*, arXiv:2211.05102.

Linlu Qiu, Peter Shaw, Panupong Pasupat, Tianze Shi, Jonathan Herzig, Emily Pitler, Fei Sha, and Kristina Toutanova. 2022. Evaluating the impact of model scale for compositional generalization in semantic parsing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9157–9179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Pretam Ray, Jivnesh Sandhan, Amrith Krishna, and Pawan Goyal. 2024. CSSL: Contrastive self-supervised learning for dependency parsing on relatively free word ordered and morphologically rich low resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8458–8466, Miami, Florida, USA. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Daniel Y. Fu, Zhiqiang Xie, Beidi Chen, Clark Barrett, Joseph E. Gonzalez, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. 2023. Flexgen: High-throughput generative inference of large language models with a single gpu. *Preprint*, arXiv:2303.06865.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2022. Selective annotation makes language models better few-shot learners. *Preprint*, arXiv:2209.01975.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

H. Theil. 1950. *A Rank-invariant Method of Linear and Polynomial Regression Analysis, 3; Confidence Regions for the Parameters of Polynomial Regression Equations: (proceedings Knaw, _5_3(1950), Nr 9, Indagationes Mathematicae, _1_2(1950), P 467-482)*. Stichting Mathematisch Centrum. Statistische Afdeling.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Xing Wu, Chaochen Gao, Zijia Lin, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. InfoCSE: Information-aggregated contrastive learning of sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3060–3070, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023. Complementary explanations for effective in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4469–4484, Toronto, Canada. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Junlei Zhang, Zhenzhong Lan, and Junxian He. 2023. Contrastive learning of sentence embeddings from scratch. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3916–3932, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Zhenyu Zhang, Yuming Zhao, Meng Chen, and Xiaodong He. 2022. Label anchored contrastive learning for language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1437–1449, Seattle, United States. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Haojie Zhuang, Wei Emma Zhang, Chang Dong, Jian Yang, and Quan Sheng. 2024. Trainable hard negative examples in contrastive learning for unsupervised abstractive summarization. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1589–1600, St. Julian's, Malta. Association for Computational Linguistics.

A Appendix A

Table 12 shows the example prompt used in preliminary experiments.

B Appendix B

Table 13 shows the hyperparameters used for baselines.

C Appendix C

Table 15 reports the performance of the Qwen family under different numbers of shots. Qwen3-1.7B significantly outperforms the 32-shot NN and achieves performance comparable to the 64-shot NN. Similarly, Qwen3-4B shows comparable to the 32-shot NN.

w/o NS & CE
Please summarize the following sentence. Sentence: cDaCErYcEe7UwAiwAZqkswC1oBxKom4HVoPBD Summary: K0jnzcEYOH1jJxLhWfkY6AzprMlhCkOYWDvAaSuteB This sentence "d5BDCBLvAZ7FygNc1NlA43I" means in one word:
w/o CE
Please summarize the following sentence. Sentence: cDaCErYcEe7UwAiwAZqkswC1oBxKom4HVoPBD Summary: 1. K0jnzcEYOH1jJxLhWfkY6AzprMlhCkOYWDvAaSuteB 2. j5E3T9qDEkP3iHvyRIg8Z2BevjLZXi1W5vqWCu 3. 4joMQNKZDY7A1S3c4V78rPiK5lvemmf93C 4. UZSBUmLhWaAzxYmVXZvwAmUCsE4zNvr40KfHwS4kLzIKTsKO 5. TLrlk 6. IYdJeAJ 7. 4XKyv3l5IVZ1OJcw5uNVhWvkSHXfB 8. d5BDCBLvAZ7FygNc1NlA43I 9. Lk7n This sentence "d5BDCBLvAZ7FygNc1NlA43I" means in one word:
Contrastive Prompt (CP)
Please summarize the following sentence. Sentence: cDaCErYcEe7UwAiwAZqkswC1oBxKom4HVoPBD Summary: 1. K0jnzcEYOH1jJxLhWfkY6AzprMlhCkOYWDvAaSuteB 2. j5E3T9qDEkP3iHvyRIg8Z2BevjLZXi1W5vqWCu 3. 4joMQNKZDY7A1S3c4V78rPiK5lvemmf93C 4. UZSBUmLhWaAzxYmVXZvwAmUCsE4zNvr40KfHwS4kLzIKTsKO 5. TLrlk 6. IYdJeAJ 7. 4XKyv3l5IVZ1OJcw5uNVhWvkSHXfB 8. d5BDCBLvAZ7FygNc1NlA43I 9. Lk7n The options are arranged in the correct order, starting from number one. This sentence "d5BDCBLvAZ7FygNc1NlA43I" means in one word:

Table 12: Used examples for prompt to extract the embedding vectors.

D Appendix D

Robustness. Table 14 shows examples we used to evaluate the robustness of the models.

Case Study. Figure 7 presents an example query, its gold answer, and the exemplars retrieved by each strategy. The first block shows the example query and its corresponding gold answer for summarization, while other blocks contain the retrieved exemplars from each method with 4-shot settings.

C-ICL	
Sorting score	ROUGE-2
Top_p	0.7
Temperature	0.6
Sampling	3
MMR	
Lambda	0.5
DL-MMR	
Lambda	0.1

Table 13: Hyperparameters for MMR, DL-MMR, and C-ICL methods.

Sentence: The Israeli navy has detained four Palestinian fishermen off the Gaza coast, capturing their boat and taking them to an undisclosed location. Summary: 1. The Israeli navy has detained four Palestinian fishermen. 2. PuNePCbR1pOlGlVoUJ9t0i3CAZ 3. pB2EbSeT 4. Sa7yOaZnP18pJLYPLmH6NzjbmCJGI5rogWTQKpKUEFz9y The most appropriate summary is number 1. The options are arranged in the correct order, starting from number one.
Sentence:Egyptian authorities released 20 Yemeni fishermen on Friday that were arrested in June when they were found sailing in Egyptian territorial waters. Summary: 1. Egyptian authorities released 20 Yemeni fishermen. 2. 9eAOOgXIFquQzZSe8nfHqCmy631JUHeEIGaQqc4EVqi6I4cMY 3. mDGbcVhW 4. 2ERKnaCHKJleeQ0nQV0myvFzqPBO The most appropriate summary is number 1. The options are arranged in the correct order, starting from number one.
Sentence:The Spanish navy arrested early Sunday two Somali pirates who took part in the hijacking of a tuna trawler which remained in the hands of bandits near the Somali coast, the defence ministry said. Summary: 1. The Spanish navy arrested two Somali pirates. 2. 6jkowWaJFgL9MgAtLY79nn8FhXuDHvNaaQwdCGl78jmqsS 3. wNxgnQMbLg7PTYNxya8Cq8AKnOt3bG7BY12v 4. sjUuWsLjIAYtL The most appropriate summary is number 1. The options are arranged in the correct order, starting from number one.
Sentence:Israeli forces Wednesday arrested 14 Palestinians, including three minors, from across the West Bank and Gaza Strip, according to local and security sources. Summary: 1. Israeli forces arrested 14 Palestinians, including three minors, from across the West Bank and Gaza Strip. 2. 8TKGiNt8GU 3. 2MGFToW 4. pHUye6KMgfvyL8jvkFoGYuRggCsICtGVqALaP The most appropriate summary is number 1. The options are arranged in the correct order, starting from number one.

Table 14: Examples used for evaluating robustness

Strategy	**Qwen3-1.7B**				**Qwen3-4B**			
	R-1	**R-2**	**R-L**	**BS**	**R-1**	**R-2**	**R-L**	**BS**
ConRAS (0-shot)	59.84	47.95	58.76	0.62	59.64	47.45	58.50	0.61
ConRAS (1-shot)	69.07	57.74	67.65	0.69	70.17	58.69	69.07	0.70
ConRAS (2-shot)	72.46	61.73	71.46	0.72	70.86	60.30	69.89	0.70
ConRAS (4-shot)	72.83	62.06	71.22	0.72	72.73	62.98	71.52	0.72
ConRAS (8-shot)	72.20	62.06	70.64	0.71	74.71	65.78	74.01	0.73
ConRAS (16-shot)	74.47	**64.90***	73.27	0.73	77.64	68.35	76.81	0.76
NN (16-shot)	73.17	60.99	71.89	0.73	76.77	65.27	75.95	0.75
NN (32-shot)	73.85	61.65	72.67	**0.74**	78.36	67.59	77.60	0.77
NN (64-shot)	**75.26**	63.63	**73.93**	**0.74**	**78.74**	**68.50**	**77.90**	**0.78**

Table 15: Performance of ConRAS with varying numbers of shots on the **Google** dataset based on both Qwen3-1.7B and Qwen3-4B. 0 shot indicates a zero-shot summarization setting without retrieval.

Given query and gold summary.

Given query
Luis Suarez was spotted in London this afternoon and this has led the Daily Star to link the Liverpool striker to a potential move to Chelsea or Arsenal.
Gold
Luis Suarez was spotted in London.

Retrieved exemplars using NN.

NN

Source. Luis Suarez has revealed he wants to stay at Liverpool after speculation Paris Saint-Germain were planning an offer for the Uruguayan, the Liverpool Echo has reported.
Target. Luis Suarez wants to stay at Liverpool.

Source. Liverpool have once again been linked with a shock move for Barcelona striker David Villa who is no longer guaranteed of first team place at the Nou Camp according to the Daily Mirror.
Target. Liverpool have been linked with a shock move.

Source. Luis Suárez has confirmed that he wants to leave Liverpool this summer – but the Anfield club insist the striker is “not for sale”.
Target. Luis Suárez has confirmed he wants to leave Liverpool this summer.

Source. Juventus are ’not interested’ in Liverpool striker Luis Suarez as he looks to extend his stay at Anfield, ending speculation surrounding the Uruguayan’s future.
Target. Juventus are not interested in Liverpool striker Luis Suarez as he looks.

Retrieved exemplars using MMR.

MMR

Source. Luis Suarez has revealed he wants to stay at Liverpool after speculation Paris Saint-Germain were planning an offer for the Uruguayan, the Liverpool Echo has reported.
Target. Luis Suarez wants to stay at Liverpool.

Source. SWANSEA City goalkeeper Michel Vorm was spotted in Barcelona this weekend, according to reports in Spain.
Target. SWANSEA City goalkeeper Michel Vorm was spotted in Barcelona.

Source. David Beckham has been linked with sensational return to Real Madrid, according to the Daily Mail.
Target. David Beckham has been linked with sensational return to Real Madrid.

Source. Luis Suarez bit an opponent’s arm, then scored on a header in the seventh minute of second-half stoppage time Sunday to give Liverpool a 2-2 tie against Chelsea in the Premier League.
Target. Luis Suarez bit an opponent’s arm, then scored.

Retrieved exemplars using DL-MMR. **NS** indicates a negative sample.

DL-MMR

Source. Luis Suarez has revealed he wants to stay at Liverpool after speculation Paris Saint-Germain were planning an offer for the Uruguayan, the Liverpool Echo has reported.
Target. Luis Suarez wants to stay at Liverpool.

Source. Luis Suárez has confirmed that he wants to leave Liverpool this summer – but the Anfield club insist the striker is “not for sale”.
Target. Luis Suárez has confirmed he wants to leave Liverpool this summer.

Source. Liverpool have once again been linked with a shock move for Barcelona striker David Villa who is no longer guaranteed of first team place at the Nou Camp according to the Daily Mirror.
Target. Liverpool have been linked with a shock move.

Source. Liverpool captain Steven Gerrard says striker Luis Suarez has the potential to become one of the world’s best players.
Target. Steven Gerrard says Luis Suarez has the potential to become one of the world’s best players.

Retrieved exemplars using C-ICL.

C-ICL

Source. Luis Suarez has revealed he wants to stay at Liverpool after speculation Paris Saint-Germain were planning an offer for the Uruguayan, the Liverpool Echo has reported.
Target. Luis Suarez wants to stay at Liverpool.

Source. Liverpool have once again been linked with a shock move for Barcelona striker David Villa who is no longer guaranteed of first team place at the Nou Camp according to the Daily Mirror.
Target. Liverpool have been linked with a shock move.

Source. Luis Suárez has confirmed that he wants to leave Liverpool this summer – but the Anfield club insist the striker is “not for sale”.
Target. Luis Suárez has confirmed he wants to leave Liverpool this summer.

Source. Colombian act Chocquibtown has been nominated for a 2011 Grammy, the organization of the award show announced.
Wrong Summary. Colombian act Chocquibtown has been nominated for a 2011 Grammy.
Target. Colombian act Chocquibtown has been nominated for a 2011 Grammy, the organization of the award show announced.

Retrieved exemplars using NN ConRAS. **NS** indicates a negative sample.

ConRAS

Source. Luis Suarez has revealed he wants to stay at Liverpool after speculation Paris Saint-Germain were planning an offer for the Uruguayan, the Liverpool Echo has reported.
Target. Luis Suarez wants to stay at Liverpool.
NS. Suarez wants to stay at Liverpool.
NS. He wants to stay at Liverpool.
NS. Suarez wants to stay at Liverpool after PSG offer.

Source. Liverpool have once again been linked with a shock move for Barcelona striker David Villa who is no longer guaranteed of first team place at the Nou Camp according to the Daily Mirror.
Target. Liverpool have been linked with a shock move.
NS. David Villa is no longer guaranteed to play for Barcelona.
NS. Barcelona striker David Villa is no longer guaranteed to play for Liverpool.
NS. Barcelona striker David Villa is no longer guaranteed of first team place.

Source. Luis Suárez has confirmed that he wants to leave Liverpool this summer – but the Anfield club insist the striker is "not for sale".
Target. Luis Suárez has confirmed he wants to leave Liverpool this summer.
NS. Suárez wants to leave Liverpool this summer.
NS. Suarez wants to leave Liverpool this summer.
NS. Suárez wants to leave Liverpool.

Source. Juventus are 'not interested' in Liverpool striker Luis Suarez as he looks to extend his stay at Anfield, ending speculation surrounding the Uruguayan's future.
Target. Juventus are not interested in Liverpool striker Luis Suarez as he looks.
NS. Juventus are not interested in Liverpool striker Luis Suarez.
NS. Juventus are not interested in Liverpool striker Suarez.
NS. Juventus are not interested in Luis Suarez.

Figure 7: Prompt examples used for the compared methods and ConRAS.