Domain Adapted Text Summarization with Self-Generated Guidelines

Andrianos Michail[1,2*] **Bartosz Rudnikowicz**[1†] **Pavlos Fragkogiannis**[1†] **Cristina Kadar**[1†]

[1]Thomson Reuters Labs [2]University of Zurich
{FirstName.LastName}@thomsonreuters.com

Abstract

Text summarization systems face significant adaptation costs when deployed across diverse domains, requiring expensive few-shot learning or manual prompt engineering. We propose a cost-effective domain adaptation framework that generates reusable summarization guidelines using only two reference summaries and three LLM inferences. Our approach works by having the model compare its own generated summaries against domain specific reference summaries in a one time preparation step that derives concise natural language guidelines that capture the summarization patterns of the target domain. These guidelines are then appended to the summarization prompt to adapt the LLM to the target domain at a minimal cost. We evaluate our method across diverse model sizes on three distinct summarization domains: Lawsuits, ArXiv papers, and Patents. Automatic metrics show that guideline-based adaptation achieves comparable or superior performance compared to in-context learning and zero-shot baselines. An LLM preference evaluation using the latest models shows that summaries generated using such guidelines are superior to the zero-shot or in-context learning summarization prompts. Our method enables efficient domain adaptation of text summarizer LLMs with a minimal resource overhead, making specialized summarization particularly accessible for agentic systems that require to process heterogeneous texts in enterprise environments.

1 Introduction

Text summarization modules are integral to many modern agentic systems (Feng et al., 2023; Chen et al., 2025) as they enable agents to process larger volumes of information efficiently by condensing content within the limited input context. This allows agents to iteratively refine their understanding without repeatedly processing the full source material. Often, these summarization systems must process content from diverse domains and formats, where the optimal summary depends on the specific domain and intended use case. After deployment, processing long texts from previously unseen domains or formats often requires substantial manual adaptation of the summarization module. Common industry strategies for this adaptation include providing in-context demonstrations or creating labor-intensive, domain-specific prompts (Fonseca and Cohen, 2024). In contrast to approaches that incur high input token costs from in-context demonstrations or require extensive manual crafting of domain-specific prompts, we propose a generic framework for adapting summarization prompts to new domains or formats using only two reference examples and a total of three LLM inferences. In this framework, the Large Language Model (LLM) summarization system first reviews two (K=2) pairs of its generated outputs and the corresponding reference summaries from the target domain. From these exemplars, it derives a concise set of summarization guidelines, which are appended to the summarization prompt during inference. Fonseca and Cohen (2024) have shown that LLMs can successfully follow specific instructions in their prompts to adjust their writing style and content based on the communication goals of the summary. This demonstrates that providing LLMs with clear language

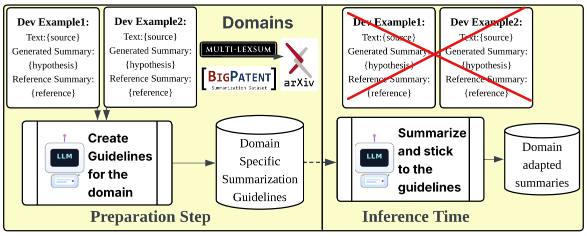

Figure 1: Our domain adaptation framework for text summarization with self-generated guidelines.

[1] *Work done during internship at Thomson Reuters Labs.
[2] †Corresponding authors.

Proceedings of the Natural Legal Language Processing Workshop 2025, pages 226–239
November 8, 2025 ©2025 Association for Computational Linguistics

guidelines can effectively help them adapt their summarization to different purposes. We evaluate the effectiveness of this approach by treating each text summarization dataset as a distinct domain and comparing its performance for each domain against generic prompt baselines, namely zero-shot and in-context learning.

Contributions

- We introduce an efficient domain adaptation framework for text summarization that generates reusable natural language guidelines using only two reference examples and three LLM inferences, significantly reducing inference costs compared to traditional in-context learning approaches.

- We demonstrate the effectiveness of our self-generated guidelines across three distinct domains (legal, scientific, and patent documents) and multiple model sizes (4B, 12B, 27B parameters), showing consistent improvements both in terms of metrics and semi automatic qualitative checks.

- We provide a comprehensive evaluation methodology combining ROUGE metrics with LLM-based preference evaluation, revealing complementary insights about summarization quality and demonstrating low agreement between token-overlap and human-like LLM preference evaluations (Cohen's kappa: -0.05 to 0.20).

2 Related Work

Within the past couple of years, it has become increasingly apparent that LLMs outperform specialized seq2seq models at summarization tasks (Pu et al., 2023; Zhang et al., 2025). A natural benefit of using LLMs for text generation tasks such as summarization is their long-context support and the ability to perform tasks in a manner that is more adapted to the specific use case through instruction prompt specifications.

```
Your job is to analyze example pairs of
Source Texts with Generated Summaries
and Target Summaries. You will write a
newline separated short list of
sentences (up to 8) as GUIDELINES on
generating better summaries that match
the DESIRED text length, formatting,
grammatical person, level of abstraction
 and sentence complexity shown in the
target examples, while avoiding the
mistakes present in the generated
summaries. These GUIDELINES will help a
summarizer produce better summaries
without seeing any of the example
summaries. Focus on general principles,
not specific example details or narrow
sub-domain instructions.

EXAMPLE {i}:

Source Text: {source_text}

Generated Summary: {zero_shot_summary}

Target Summary: {reference_summary}

Write a newline separated short list of
sentences (up to 8) as GUIDELINES on
generating summaries that better match
the DESIRED text length, formatting,
grammatical person, level of abstraction
 and sentence complexity shown in the
target examples, while avoiding the
mistakes in the generated summaries.
KEEP THE LIST SHORT. ONLY produce the
GUIDELINES and no additional text. Never
 mention specific examples, target
summaries, or generated summaries in the
 guidelines. The guidelines should be
general and applicable to the dataset as
 a whole, providing clear direction that
 can be followed independently.

GUIDELINES:
```

Prompt 1: Prompt used to generate guidelines. We always use two examples.

The search for effective prompt instructions to guide LLMs to improve task performance and to better fit the user's needs has been widely explored by practitioners and documented through resources such as prompting handbooks. The research community has also developed interest in automatically crafting prompts through various studies that in general terms combine exploration and evaluation (on a training set) of new variants given initial prompts.

We now introduce some relevant works in auto-prompting: Prasad et al. (2023) explore alternative phrasings of the initial prompt through word- and phrase-level edits. Pryzant et al. (2023) iteratively refine the initial prompt using natural language "batch gradients" that critique the current prompt whilst this prompt is being adapted to the opposite semantic direction of the gradient.

In parallel work, literature of In-Context Reinforcement Learning has demonstrated that LLMs can improve and even acquire new abilities

Model	Dataset	Generated Guidelines
Gemma-3-27B	***ArXiv***	Focus on conveying the core research question and primary findings. Prioritize summarizing the overall approach and key results over detailed methods. Maintain a concise and direct writing style, avoiding unnecessary elaboration. Use declarative sentences and active voice to clearly state information. Emphasize the significance and potential implications of the work. Adopt a level of abstraction that highlights the main contributions, omitting granular details. Keep summaries relatively short, typically within a defined word or sentence limit. Frame the summary as a cohesive overview of the study's purpose and conclusions.
Gemma-3-27B	***BIGPATENT***	Focus on capturing the core invention and its key features. Maintain a formal and technical tone, mirroring patent-like language. Prioritize describing *what* the invention does over *how* it works in detail. Use complex sentence structures and precise terminology. Summaries should be concise, typically within a single paragraph. Employ the active voice and avoid excessive pronouns. Retain the original document's grammatical person (often third person). Emphasize the problem the invention solves and its advantages.

Table 1: Self-generated (for the 27B Model) summarization guidelines for the two distinct domains. The complete set of self-generated guidelines is presented in Table 4.

by receiving numerical or verbal feedback on their past generations often through multiple self-iterations(Lee et al., 2023; Monea et al., 2025; Song et al., 2025; Madaan et al., 2023).

Our approach builds on the auto-prompt tuning paradigm by incorporating insights from In-Context Reinforcement Learning. However, unlike iterative approaches, our method adapts the summarization prompt to a new domain through a single preparatory step that requires only two development samples and two summaries generated through a generic prompt for a total of just three LLM inferences.

3 Methodology

3.1 Prompts

Minimal Prompt We use a *Minimal* prompt that generically requests for a summary, similar to what the model has seen during its instruction tuning. The *Minimal* (zero-shot) prompt template is illustrated in the Prompt 2.

In-Context Learning (ICL) Prompt We create a *Minimal* prompt variant that also receives texts and their reference summaries in text. The *ICL prompt* template is available in the Prompt 4.

Summarization with Self-Generated Guidelines As denoted in Figure 1, we propose a two-step summarization pipeline. In the preparation step, the model contrastively analyzes example sets containing 1. Source texts 2. Summaries it has generated through the *Minimal* prompt 3. The Reference (called Target) Summaries. As depicted in Prompt 1, the model receives these three components and then it is requested to identify its previous mistakes and generate a short set of summarization guidelines for future inferences to produce more suitable summaries that better fit the style of the reference summaries. Example guidelines for two domains are showcased in Table 1. These self-generated guidelines are produced once for each combination of domain and dataset and then stored for use during inference time.

Guidelines Prompt Our proposed solution is a *Minimal* prompt extension that also instructs the model to follow the aforementioned self-generated guidelines. The *Guideline* prompt template is available in Prompt 3.

3.2 Summarization Systems

To compare scaling effects across different model sizes within the same architecture, we experiment with Google's Gemma-3 (**I**nstruction-**T**uned) at **4B**, **12B**, and **27B** parameters (Team et al., 2025). All models provide a **128k-token context window**, which is critical for handling the input size required for long-document summarization.

3.3 Datasets

We evaluate all models using only the test sets of each respective dataset. For the necessary examples of both in-context learning and guideline generation process, we use the same two examples arbitrarily selected from the development set of the corresponding dataset.

MultiLexSum (MLS): Multi-document Civil

Rights Lawsuits (908 sources drawn from the Civil Rights Litigation Clearinghouse) paired with expert-authored reference summaries (Shen et al., 2022). Reference summaries are provided at three different granularities: ***long*** (typically multiple paragraphs, 630 words), ***short*** (only one paragraph, 130 words) and ***tiny*** (one sentence, 25 words).

ArXiv: Full length scientific papers (total of 6440) taken from arXiv.org and PubMed.com scientific repositories paired with their abstracts as reference summaries (Cohan et al., 2018).

BIGPATENT: U.S. patent documents paired with human-written abstractive summaries as their reference summaries (Sharma et al., 2019). The patents come from nine different technological areas, however, in our research we limit ourselves to the 6911 patents of the area *"y: General tagging of new or cross-sectional technology"*.

Datasets Lengths The datasets vary in source text length. ***BIGPATENT*** average document length is 6585 tokens and median of 5290 tokens. For ***ArXiv*** documents average 8713 tokens and median of 7161 tokens, with documents extending beyond 100K tokens for both datasets. ***MultiLexSum*** contains substantially longer texts, averaging 95,998 tokens, with median on 41926 tokens and source texts reaching up to 4M tokens. These lengthy documents constrain in-context learning methods, as individual examples can completely fill or exceed conventional LLM input context length.

3.4 Evaluation Metrics

ROUGE: Automatic summarization evaluation metric that assesses summary quality by quantifying token overlap between a system-generated summary and one or more human reference summaries. Specifically, ROUGE-1 (**R-1**) reports recall/precision/F1 based on overlapping unigrams, ROUGE-2 (**R-2**) based on overlapping bigrams, and ROUGE-L (**R-L**) computes recall/precision/F1 based on the length of the Longest Common Subsequence. All ROUGE scores reported in this work are **F1** scores (x100 for readability).

LLM Preference: Complementary to ROUGE, we use Claude 4 Sonnet to measure pairwise preference between summaries generated through different prompting approaches relative to the reference summary. While traditional metrics like ROUGE focus on token overlap, LLM-based evaluation can capture more nuanced aspects of text quality and has been widely adopted in recent literature (Bavaresco et al., 2025; Liu et al., 2024a). Research has demonstrated that LLM evaluators often show higher agreement with human evaluation than conventional automatic metrics (Nguyen et al., 2024; Tan et al., 2024; Shen et al., 2023). To minimize positional bias (Wang et al., 2024), we randomly shuffle the order of the presented generated summaries during inference and perform each evaluation (total of 500 samples) three times, selecting the most frequent prediction. Our complete evaluation criteria and precise prompt is available in Prompt 5.

3.5 Compute Costs

For the summarization systems, all models were hosted at full precision on an x4 L40(48 GB) GPU cluster for a total of 430 hours[1]. For the LLM preference evaluation, we ran 44,808 Anthropic API requests averaging approximately 1100 input tokens per call with a total estimated cost of $150.

4 Results

This section presents our results in five parts. We begin with the ROUGE evaluation of our main results, followed by an LLM preference evaluation. Subsequently, we measure the agreement between ROUGE and the LLM preference within the subset. Afterwards, we analyze a pitfall and an opportunity of our *Guidelines* approach and conclude with remarks on the characteristics/irregularities of our generated text.

4.1 ROUGE

We illustrate the ROUGE evaluation results in Table 2. Based on the token overlap evaluations, the following patterns emerge:

***ICL*(K = 2) decreases ROUGE**: The introduction of two full example demonstrations within the text can harm performance, with this effect being more prominent on the 4B and 12B models. We hypothesize that the lengthy examples in the *ICL* prompt are possibly acting as a haystack, a form of information clutter for the LLMs (Liu et al., 2024b; Hengle et al., 2025).

***Guidelines* bring improvements**: The introduction of *Guidelines* consistently improves summarization performance across most dataset and model combinations. This benefit is particularly

[1] Note that this infrastructure was needed for the most compute intense experiments (long input, 27B model) whilst a smaller cluster would suffice for most of the others

	Lawsuits (***MLS_long***)			Science (***ArXiv***)			Patents (***BIGPATENT***)			Average		
Model & Method	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
Gemma 4B												
Minimal	**29.6**	10.7	17.2	**40.9**	**12.1**	20.2	40.4	12.2	23.2	**37.0**	11.7	20.2
ICL (K=2)	28.6	10.1	16.9	34.7	8.0	18.2	38.2	11.5	23.0	33.4	9.9	15.5
Guidelines (K=2)	28.5	**11.1**	**17.5**	40.6	11.5	**20.8**	**40.6**	**12.7**	**24.1**	36.6	**11.8**	**20.8**
Gemma 12B												
Minimal	30.2	11.3	18.0	41.2	12.0	**21.2**	40.9	12.7	24.1	37.4	12.0	21.1
ICL (K=2)	28.5	10.5	17.3	31.6	7.5	17.6	43.7	**17.2**	**28.4**	34.6	11.8	17.4
Guidelines (K=2)	**31.3**	**12.7**	**19.3**	**41.5**	**12.3**	21.0	**45.1**	16.6	27.6	**39.3**	**13.9**	**22.6**
Gemma 27B												
Minimal	33.0	11.8	18.6	**43.2**	**13.3**	22.5	40.3	12.3	23.6	38.3	12.5	21.6
ICL (K=2)	28.2	10.1	16.9	39.9	11.6	21.6	**42.4**	**14.7**	**25.9**	36.8	12.1	21.5
Guidelines (K=2)	**38.1**	**15.8**	**21.3**	42.5	12.9	**22.6**	42.3	14.3	25.1	**41.0**	**14.3**	**23.0**

Table 2: Main evaluation table in ROUGE. Averages denote the arithmetic mean across the three datasets.

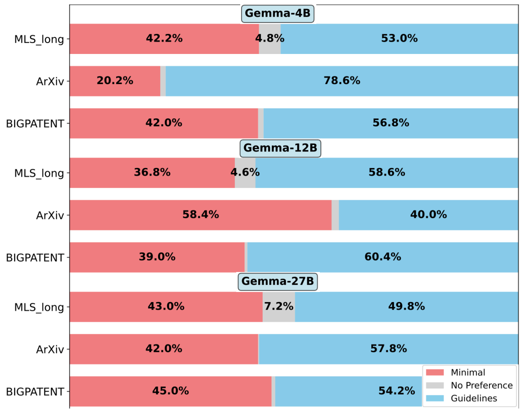

Figure 2: *Minimal* v *Guidelines* Claude's preference.

evident with Gemma 12B on the BIGPATENT dataset, which exhibits a ROUGE increase of **4.3** on average, and with Gemma 27B on the MLS_long dataset, which demonstrates an average improvement of **4.6**.

BIGPATENT 12B v 27B: Contrary to expectations, the 12B model achieved higher ROUGE scores than the 27B model on the BIGPATENT dataset when using the Guidelines prompt. To further examine this unexpected finding, we conducted an LLM preference evaluation between the two models and found that 51.4% of the summaries generated by the 27B model are preferred over those produced by the 12B model, contradicting the superior performance in terms of ROUGE scores.

4.2 LLM Preference

We illustrate Claude 4's preference between summaries generated by the *Minimal* and *Guidelines* prompts in Figure 2. In eight out of nine evaluations, the *Guidelines* summaries are preferred, with strong preferences ranging from 50% to 60%. Notably, an extreme preference of 78.6% was observed for the *Guidelines* approach on the ArXiv dataset using the 4B model, while ROUGE metrics showed similar performance across both methods. Upon examining these specific guidelines (second row, Table 4), we find that they constitute excellent advice for writing well-structured abstracts, emphasizing concise phrasing, logical sequencing, and structured organization around key categories such as methods, results, and implications. Our qualitative analysis on a sample confirms that this improved structure and formatting is indeed present, an enhancement that goes completely undetected by ROUGE evaluations but is strongly preferred by Claude.

Illustrated in Figure 6, when comparing *Guidelines* versus *ICL* summaries, Claude demonstrates a preference by a large margin for the *Guidelines* in seven evaluations, with *ICL* preferred only in the 12B and 27B BIGPATENT cases by smaller margins.

On the summaries produced by the 12B model on the ArXiv dataset, we observe the single evaluation where the *Minimal* prompt wins, being strongly preferred in 58.4% of the samples. This result suggests that low-quality or unfortunate generations of self-generated guidelines can degrade performance, even when applied to datasets where the same samples and prompts previously produced effective guidelines with a smaller model.

4.3 ROUGE v LLM Preference

To enable comparison with the Claude preference evaluations, we converted ROUGE into a preference-based metric by designating the summary with the higher ROUGE F1 score as preferred in each pair. We then measured the agreement be-

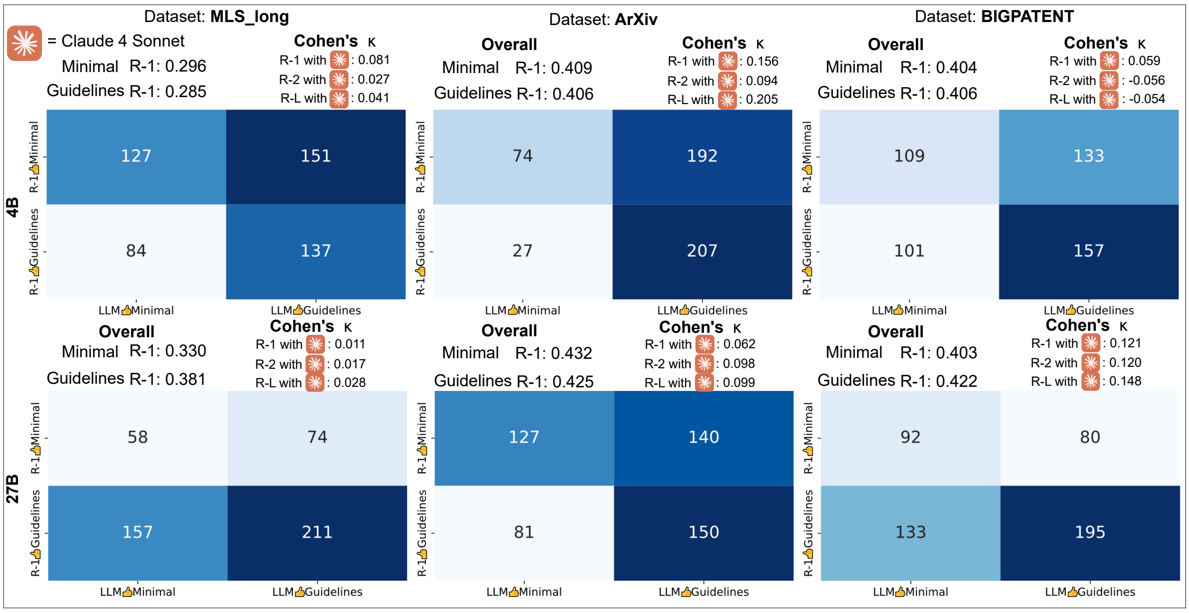

Figure 3: Agreement between Claude and ROUGE preferences for the 4B and 27B models. The notation {evaluation metric}👍{summarization method} indicates cases where the evaluation metric prefers summaries generated by the corresponding method. For the 12B model, see Figure 5.

tween these ROUGE-based preferences and the LLM preference evaluations for both the 4B and 27B models, as shown in Figure 3.

We observe low agreement between the two evaluation approaches, with Cohen's kappa scores ranging from -0.05 to 0.20. In datasets where the two approaches produce summaries with similar overall average R-1 scores with the most frequent source of disagreement is that Claude prefers summaries produced through the *Guidelines* prompt while R-1 favors those generated through the *Minimal* prompt. Conversely, when there are larger overall R-1 differences (>0.02), the most frequent disagreement occurs when Claude prefers summaries from the *Minimal* prompt while R-1 the *Guidelines* prompt.

These interactions between the metrics shows the importance of both token overlap ROUGE and human-like LLM preference evaluations to obtain a broader picture of summarization quality.

4.4 Pitfall: Bad-Batch Guidelines

During the preparatory step, the auto-regressive generation of summarization guidelines is susceptible to sequence generation pitfalls previously identified in the literature (Arora et al., 2022). Exposure bias can lead to error accumulation, where initial minor mistakes or improbable continuations compound over time, resulting in counterproductive generations. We call these problematic guidelines generations as "Bad-Batch Guidelines" as they may mislead the summarization module and cause undesired behavior in the generated summaries.

Our evaluation thus far has relied on the first set of guidelines generated by the system. However, how robust is our approach when the initial guideline generation produces suboptimal results? To address this, we generate multiple guideline sets for the 12B model on ArXiv and BIGPATENT datasets, selecting a set of guidelines that we identify as potentially misleading to the summarization process. These selected "Bad-Batch Guidelines" are presented in Table 5.

We then perform summarization with these "Bad-Batch Guidelines" and evaluate the results using both ROUGE and LLM preference evaluation against both the *Guidelines* and *Minimal* prompts. The results are illustrated in Figure 4.

We can see that the "Bad-Batch Guidelines" perform similarly to the *Minimal* prompt in terms of ROUGE on the ArXiv dataset or even outperform it on the BIGPATENT dataset. However, the LLM preference evaluation reveals the subpar quality of summaries generated using faulty guidelines, particularly prominently in the BIGPATENT dataset.

To further investigate the issues introduced by the "Bad-Batch Guidelines", we performed a semi-automatic qualitative analysis of the LLM preference evaluation's "reasoning" output using Claude Code. We found that the problematic guidelines promoted two key issues: (1) in ArXiv datasets,

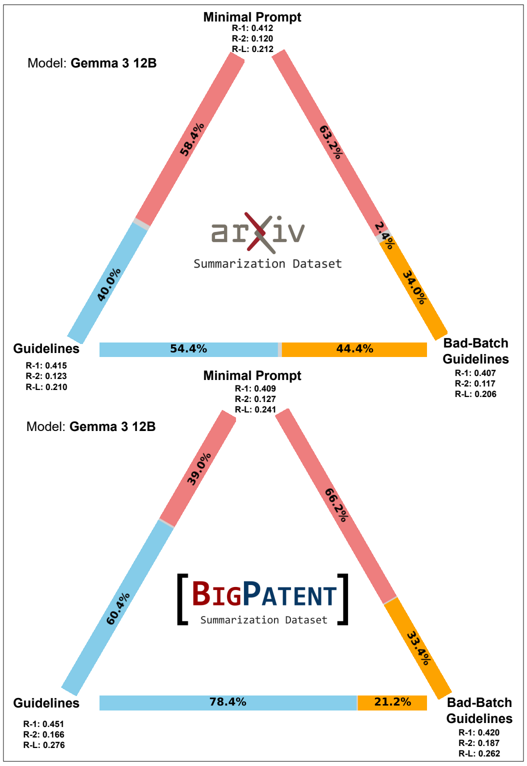

Figure 4: *Minimal* v *Guidelines* v Bad-Batch Guidelines. Bar segments indicate Claude's preference percentages whilst corner values show average ROUGE scores.

the introduction of background context and content not present in the reference material, and (2) in BIGPATENT datasets, excessive verbosity that led to repetition. These insights were subsequently confirmed through additional qualitative checks.

This highlights the risks of providing unchecked summarization guidelines to the model, which can be minimized by keeping a human in the loop to review the guidelines and automatically running initial experiments to validate performance.

4.5 Opportunity: LenSpecific Guidelines

The guidelines produced by the model are written in natural language, allowing human post-editing. We conducted an experiment on the MultiLexSum dataset, a dataset with three reference summaries of different lengths. We generate independently guidelines for each length variant, then applied minimal human edits (detailed in Table 6) to introduce/alter the guidelines so that it specifies the target summary length based on the reported average.

Observing the raw length of the produced summaries, we compare the word counts between *Guidelines* and *LenSpecificGuidelines* across three target lengths. For tiny summaries (25-word target), word lengths decreased from 90–113 to 60–65 words. Short summaries (130-word target) shifted from 83–120 to 116–125 words. For long summaries (630-word target), we observe a monotonic increase in average length based on model size: 4B model (108 → 151 words), 12B model (121 → 239 words), and 27B model (193 → 323 words). This pattern in the long target summaries (in a dataset with 96,000 tokens long texts) might suggest that the amount of information a model can represent and summarize is proportional to its parameter size, even when given explicit instructions to produce longer (630 words) summaries (issue is also documented in Fonseca and Cohen (2024)).

In ROUGE scores, the *LenSpecificGuidelines* outperform the original *Guidelines* in most experiments as anticipated since they better match the target summary length.

However, the LLM Preference evaluation finds that the summaries produced by *Guidelines* are preferred over the *LenSpecificGuidelines* on the short and long target summaries with large margins across all models (3-21%). Instead, the single sentence tiny target summaries exhibit the opposite behavior with *LenSpecificGuidelines* being preferred over the *Guidelines* with even broader margins across all models with (2-34%). We conducted a semi-automatic qualitative analysis of the LLM preference evaluation's "reasoning" output using Claude Code. For "tiny" summary targets, we observed that the LenSpecificGuidelines promoted higher levels of conciseness without sacrificing completeness. However, for lengthier "long" and "short" summary targets, the LenSpecificGuidelines performed worse, as they tended to introduce redundant information and excessive legal specifics.

4.6 Remarks on Generations

We examine the stability and quality of our text generation results. While it would be optimal to compute each experiment multiple times to account for stochastic variation, this is prohibitively expensive at scale. We therefore conduct a stability check and analyze additional generation quality issues, including instruction-following artifacts and in-context learning errors.

Stability of Results In terms of ROUGE, we run three rounds of generations using the 4B model on the *Minimal*, *Guidelines*, and *ICL* prompts across the ArXiv and BIGPATENT datasets. All

	Lawsuits (***MLS_long***)			Lawsuits (***MLS_short***)			Lawsuits (***MLS_tiny***)			Average		
Model & Method	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
Gemma 4B												
Minimal	29.6	10.7	17.2	25.9	9.8	15.0	**24.2**	**9.4**	13.9	26.6	10.0	15.4
Guidelines	28.5	11.1	17.5	**37.2**	11.9	22.3	18.4	5.1	12.8	28.0	**11.9**	17.5
LenSpecificGuidelines	**33.4**	**12.6**	**19.0**	38.8	**12.3**	**22.4**	20.0	5.3	**14.0**	**34.9**	9.9	**18.5**
Gemma 12B												
Minimal	30.2	11.3	18.0	27.1	10.5	15.7	**25.8**	**10.4**	**14.9**	27.7	10.7	16.2
Guidelines	31.3	12.7	19.3	**39.9**	**13.9**	**23.9**	16.0	4.4	11.2	29.1	10.3	18.1
LenSpecificGuidelines	**40.4**	**15.8**	**21.9**	39.7	13.2	23.4	20.9	5.5	14.8	**33.7**	**11.5**	**20.0**
Gemma 27B												
Minimal	33.0	11.8	18.6	30.0	10.8	16.3	**28.5**	**10.5**	**15.2**	30.5	**12.5**	11.0
Guidelines	38.1	15.8	**21.3**	37.8	**11.5**	**21.8**	15.8	4.3	10.9	30.6	12.1	14.4
LenSpecificGuidelines	**40.6**	**16.1**	20.9	**37.9**	11.3	21.2	20.9	5.5	14.7	**33.1**	11.0	**18.9**

Table 3: Summarization performance on the MultiLexSum dataset across varying target summary lengths with length-specific instruction guidelines.

three runs for each method show similar average ROUGE scores, indicating that our reported results are stable across the other experiments as well.

Beyond ROUGE stability, we also validated the consistency of our LLM preference evaluations. We conducted additional LLM preference evaluations on the summaries produced by the 27B model using two alternative evaluation setups: (1) role-specific system prompts and (2) GPT-4.1 as the underlying evaluator. Both alternative variants agree with the primary LLM preference variant, with the role-specific system prompt showing particularly strong preferences for guideline-generated summaries (80% for MultiLexSum and 66% for BIGPATENT datasets).

Generation Artifacts We performed string matching to identify cases where models fail to follow generation instructions and produce artifacts such as "here is the summary:". On the smaller ArXiv and BIGPATENT datasets, the 12B and 27B models produce no artifacts, while the 4B model contains artifacts in approximately 0.7% of its generations. However, on **MultiLexSum**, the dataset with the largest source texts exhibited a 7-13% of generations containing artifacts across all three model sizes, with particularly high artifact rates in the *ICL* prompt. This suggests that models struggle to comprehend instructions when processing longer inputs, consistent with findings in the literature of information representation in Long-Context LLMs (Liu et al., 2024b; Hsieh et al., 2024a,b; Hengle et al., 2025).

In-Context Learning Repetitions We observe an infrequent but notable error in the *ICL* prompt on BIGPATENT: the 4B and 12B models occasionally (<0.02%) reproduce the summary from the last in-context example rather than generating a summary for the target text. Surprisingly, this error does not occur with the 27B model. This suggests that smaller models are more prone to confusion when processing in-context demonstrations, while larger models better handle this input format. Such errors may partially explain why the *ICL* prompt underperforms in smaller models but achieves competitive results within the bigger models.

5 Discussion

Computational Efficiency Our approach to generating summarization guidelines has similar attributes to in-context learning prompting but with a computational cost that is close to a *Minimal* prompt. We process the two demonstration examples once to produce reusable guidelines, so we do not have to include the K-shot examples at inference time. For K=2, that means we can omit prepending the two source texts and their reference summaries every time we perform summarization.

This efficiency gain is especially prevalent in long-form summarization. For instance, in MultiLexSum the two source texts plus their reference summaries can add up to about 190,000 input tokens, whereas the resulting guidelines are roughly 300 tokens. When performing summarization at scale, this minimization of input tokens would significantly reduce costs.

Flexible Initialization In this work, we focus on deriving summarization guidelines with an LLM by comparing model-generated summaries to reference summaries. That said, the guidelines can also be initialized in other ways, for example, from a combination of editorial standards and generated summaries, or from generated summaries paired with human post-edits. In practice, the guidelines could even be written entirely by humans; at that point, the process is more accurately described as prompt engineering.

Human Post-Editable Because the LLM-produced summarization guidelines are written in natural language, the patterns and guidelines inferred from the provided samples can be reviewed and customized by a subject matter expert. We therefore propose a human-LLM collaboration framework comprising three steps: (1) a human author provides a reference summary; (2) the LLM identifies where it deviates from the desired summarization and formulates corresponding guidelines; and (3) the subject matter expert post-edits these guidelines to refine and operationalize the desired summarization behavior.

6 Conclusions

We present a framework that adapts summarization models to new domains using only two reference summaries and three LLM inferences. The three evaluation domains: *lawsuits, arXiv papers, and patents* are deliberately distinct, motivating the need for domain-sensitive adaptation. Across these domains, the method yields mostly consistent gains in ROUGE and LLM preference. Guideline-based summaries are preferred in the majority of the evaluations across 4B, 12B, and 27B models. The self-reflection step enables models to compare their outputs with references, derive actionable, domain-specific guidelines, and use those guidelines in prompts to consistently outperform zero-shot and standard in-context learning baselines. Our joint evaluation shows low agreement between ROUGE and LLM preference (Cohen's kappa -0.05 to 0.20), suggesting that structural and coherence improvements favored by LLM judges are not captured by token-overlap metrics. Minimal examples, when converted into self-generated guidelines, enable effective domain adaptation without costly few-shot setups while maintaining competitive quality. The approach also yields interpretable guidelines and may inform future work on adaptive deployment and the role of self-reflection in domain-specific summarization.

Limitations

Our approach and examination have several limitations. Firstly, we used only two reference summaries for guideline generation. This small sample may not capture the full range of summarization styles within each domain. Our comparison with in-context learning is subject to the same two-sample constraint, which may not reflect the optimal performance of either approach, as both could benefit from larger example sets. Secondly, our experimental evaluation focused on the Gemma-3 family (4B, 12B, 27B parameters) across three domains (legal, scientific, patents). While these results demonstrate the approach's effectiveness within this scope, generalization to other model architectures or domains is not guaranteed.

Third, our evaluation methods have weaknesses. ROUGE captures only surface-level token overlap and misses qualitative aspects such as coherence and structure. Our LLM preference evaluation using Claude 4 Sonnet might also be introducing model- and prompt-specific biases. We attempted to minimize the bias introduction risks by re-running evaluations with alternative prompt variants (including system prompt roleplaying) and by using GPT 4.1 as a secondary evaluator in which we observed similar conclusions.

Fourth, our approach does not consider prompt caching techniques, which could provide an alternative pathway for domain adaptation. Prompt caching allows LLMs to store and reuse precomputed attention states for frequently used prompt prefixes, potentially enabling domain adaptation by caching domain-specific context without requiring guideline generation. While our guideline-based approach provides a lightweight solution, prompt caching might offer complementary benefits.

These limitations highlight several promising directions for future research. The evaluation challenges we identified point to the need for more comprehensive assessment frameworks that better capture the multifaceted nature of summarization quality. Finally, the generality of our guideline-generation approach suggests extensions to other domain-specific text generation tasks beyond text summarization.

References

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Cheung. 2022. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 700–710, Dublin, Ireland. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Nuo Chen, Hongguang Li, Jianhui Chang, Juhua Huang, Baoyuan Wang, and Jia Li. 2025. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 755–773, Abu Dhabi, UAE. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2023. The role of summarization in generative agents: A preliminary perspective. *Preprint*, arXiv:2305.01253.

Marcio Fonseca and Shay Cohen. 2024. Can large language model summarizers adapt to diverse scientific communication goals? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8599–8618, Bangkok, Thailand. Association for Computational Linguistics.

Amey Hengle, Prasoon Bajpai, Soham Dan, and Tanmoy Chakraborty. 2025. Multilingual needle in a haystack: Investigating long-context behavior of multilingual large language models. 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 5165–5180, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024b. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14982–14995, Bangkok, Thailand. Association for Computational Linguistics.

Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. 2023. Supervised pretraining can learn in-context reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 43057–43083. Curran Associates, Inc.

Junyuan Liu, Zhengyan Shi, and Aldo Lipani. 2024a. SummEQuAL: Summarization evaluation via question answering using large language models. In *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024)*, pages 46–55, Bangkok, Thailand. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Giovanni Monea, Antoine Bosselut, Kianté Brantley, and Yoav Artzi. 2025. LLMs are in-context bandit reinforcement learners. *Preprint*, arXiv:2410.05362.

Huyen Nguyen, Haihua Chen, Lavanya Pobbathi, and Junhua Ding. 2024. A comparative study of quality evaluation methods for text summarization. *Preprint*, arXiv:2407.00747.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2023. GrIPS: Gradient-free, edit-based instruction search for prompting large language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3845–3864, Dubrovnik, Croatia. Association for Computational Linguistics.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *Preprint*, arXiv:2309.09558.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multilexsum: Real-world summaries of civil rights lawsuits at multiple granularities. In *Advances in Neural Information Processing Systems*, volume 35, pages 13158–13173. Curran Associates, Inc.

Kefan Song, Amir Moeini, Peng Wang, Lei Gong, Rohan Chandra, Yanjun Qi, and Shangtong Zhang. 2025. Reward is enough: LLMs are in-context reinforcement learners. *Preprint*, arXiv:2506.06303.

Shao Min Tan, Quentin Grail, and Lee Quartey. 2024. Towards an automated pointwise evaluation metric for generated long-form legal summaries. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 129–142, Miami, FL, USA. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Haopeng Zhang, Philip S. Yu, and Jiawei Zhang. 2025. A systematic survey of text summarization: From statistical methods to large language models. *ACM Comput. Surv.*, 57(11).

A Appendix

```
Summarize the following text. ONLY
produce the summary and no additional
text:

Text: {source_text}

Summary:
```

Prompt 2: Zero-shot summarization prompt.

```
Your task is to summarize a text. When
summarizing, adhere to the GUIDELINES
when possible and relevant.

GUIDELINES:

{guidelines}

Text to summarize:

{source_text}

ONLY produce the summary whilst adhering
 to the GUIDELINES and DON'T PRODUCE
additional text.

Summary:
```

Prompt 3: Summarization prompt with guidelines.

```
Your task is to summarize a text. Here
are a few examples of Source Texts and
Target Summaries:

Source Text: {source_text_example}
Target Summary: {
reference_summary_example}

Summarize the following text.
Source Text: {source_text}
ONLY produce the summary and no
additional text.

Summary:
```

Prompt 4: Summarization prompt with few-shot examples. Examples section (source text and target summary pairs) is repeated for each example.

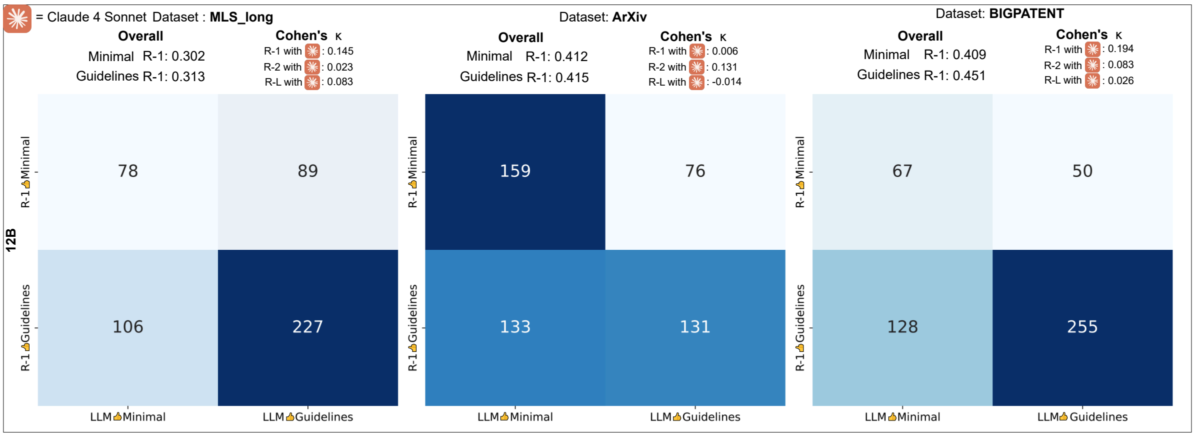

Figure 5: Agreement between Claude and ROUGE preferences for the 12B model. The notation {evaluation metric}👍{summarization method} indicates cases where the evaluation metric prefers summaries generated by the corresponding method.

```
You are evaluating two AI-generated
summaries. Compare them against the
reference summary and determine which
one is better.

**Reference Summary:** {
reference_summary}

**Summary A:** {summary_a}

**Summary B:** {summary_b}

Please evaluate which summary is better
based on:
1. **Accuracy**: How well does it
capture the key information?
2. **Clarity**: Is it well-structured
and easy to understand?
3. **Conciseness**: Does it avoid
redundancy while maintaining essential
information?
4. **Faithfulness**: How well does it
align with the reference summary?

**Instructions:**
- Choose "A" if Summary A is better
- Choose "B" if Summary B is better
- Choose "None" if both summaries are of
 equal quality

Please respond in the following JSON
format:
{{
    "preference": "A|B|None",
    "reasoning": "Brief explanation (2-3
 sentences) of why you made this choice,
 focusing on the key differentiating
factors."
}}
```

Prompt 5: LLM preference evaluation prompt. Temperature is set to 0.1

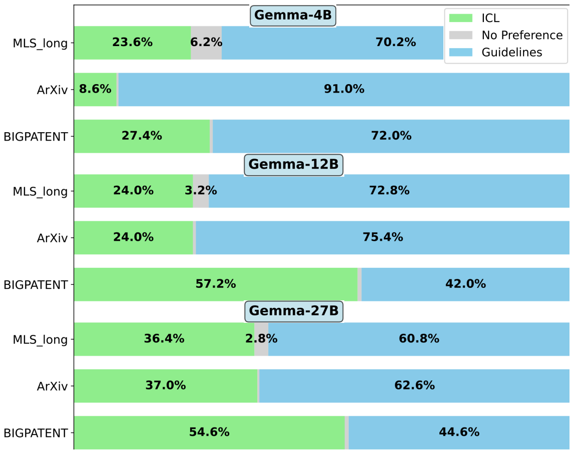

Figure 6: *ICL* v *Guidelines* LLM Preference results.

Model	Dataset	Generated Guidelines
Gemma-3-4B	***MultiLexSum***	* Focus on conveying the core legal issue and the outcome of the case concisely. * Use the passive voice to maintain a formal and objective tone. * Limit each sentence to approximately 15-20 words, prioritizing clarity and brevity. * Include key factual details (parties involved, location, legal basis) without unnecessary elaboration. * State the settlement or ruling directly and clearly. * Avoid legal jargon where possible, and define any necessary terms briefly. * Maintain a neutral and impartial perspective throughout the summary. * Structure the summary with a logical flow, starting with the initial action and ending with the resolution. * Keep the summary at approximately 100-150 words in length.
Gemma-3-4B	***ArXiv***	* Prioritize concise phrasing, avoiding overly detailed explanations. * Maintain a neutral, objective tone and use the passive voice where appropriate. * Focus on conveying the core findings and methods, omitting background context unless essential. * Structure summaries around key steps or categories (e.g., methods, results, implications). * Limit each sentence to a maximum of 20 words for clarity and brevity. * Present information in a logical sequence, reflecting the flow of the original text. * Use precise terminology and avoid jargon unless clearly defined. * Summarize the purpose and scope of the study upfront. * Conclude with a brief statement of the overall significance or impact.
Gemma-3-4B	***BIGPATENT***	* Prioritize concise phrasing, eliminating redundant words and phrases. * Maintain a formal, objective tone and use third-person perspective. * Focus on key concepts and avoid excessive detail. * Structure summaries with a clear beginning, middle, and end, reflecting the text's logical flow. * Employ declarative sentences and avoid overly complex sentence structures. * Capture the core function or purpose of the original text. * Summarize in complete sentences, ensuring grammatical correctness. * Aim for a length proportional to the original text's significance – generally, a fraction of the original word count.
Gemma-3-12B	***MultiLexSum***	Focus on the core allegations and resolution. Maintain a formal, objective tone. Use concise language and avoid unnecessary detail. Structure the summary chronologically. Clearly state the parties and their roles. Summarize legal proceedings and outcomes succinctly. Limit the summary to essential facts and rulings. Employ declarative sentences with moderate complexity.
Gemma-3-12B	***ArXiv***	Focus on the core purpose and key findings of the source text. Use concise language and avoid unnecessary detail. Employ a formal and objective tone. Structure the summary logically, mirroring the source's flow. Prioritize broader concepts over specific instances. Maintain a moderate level of sentence complexity. Avoid making interpretations or drawing conclusions beyond the source. Use passive voice where appropriate to maintain objectivity.
Gemma-3-12B	***BIGPATENT***	Focus on conveying the core functionality and key components of the described system. Maintain a formal and objective tone throughout the summary. Use concise sentences with a moderate level of complexity. Prioritize describing the system's purpose and features over detailed mechanisms. Structure the summary logically, typically starting with overall purpose and then detailing components. Avoid overly technical jargon or unnecessary detail. Aim for a summary length that provides a sufficient overview without being excessively long. Present information in a declarative style, focusing on what the system is and does.
Gemma-3-27B	***MultiLexSum***	Prioritize conveying key legal and procedural details, including case names, court locations, and specific actions taken by the court. Maintain a formal and objective tone, avoiding interpretive language or subjective assessments. Focus on summarizing the *sequence* of events, rather than simply listing facts. Use complete sentences and avoid overly concise or telegraphic phrasing. Include monetary amounts and specific dates when they are central to the case's outcome. Employ precise legal terminology where appropriate, but explain it if necessary for clarity. Summaries should generally be between 150-300 words to adequately cover the essential information. Retain passive voice where it reflects legal documentation style and avoids attributing agency unnecessarily.
Gemma-3-27B	***ArXiv***	Focus on conveying the core research question and primary findings. Prioritize summarizing the overall approach and key results over detailed methods. Maintain a concise and direct writing style, avoiding unnecessary elaboration. Use declarative sentences and active voice to clearly state information. Emphasize the significance and potential implications of the work. Adopt a level of abstraction that highlights the main contributions, omitting granular details. Keep summaries relatively short, typically within a defined word or sentence limit. Frame the summary as a cohesive overview of the study's purpose and conclusions.
Gemma-3-27B	***BIGPATENT***	Focus on capturing the core invention and its key features. Maintain a formal and technical tone, mirroring patent-like language. Prioritize describing *what* the invention does over *how* it works in detail. Use complex sentence structures and precise terminology. Summaries should be concise, typically within a single paragraph. Employ the active voice and avoid excessive pronouns. Retain the original document's grammatical person (often third person). Emphasize the problem the invention solves and its advantages.

Table 4: Self-generated summarization guidelines for all models and domains.

Model	**Dataset**	**Bad-Batch Generated Guidelines**
Gemma-3-12B	***ArXiv***	Focus on the core purpose and key findings. Use a formal, objective tone. Maintain a high level of abstraction, avoiding excessive detail. Employ relatively complex sentence structures. Prioritize conveying scientific concepts and methodologies. Present information in a concise and structured manner. Limit the inclusion of background context or tangential details. Avoid overly enthusiastic or speculative language.
Gemma-3-12B	***BIGPATENT***	Focus on conveying the core purpose and key features of the described invention. Use formal and technical language appropriate for a patent-style description. Structure sentences to present information in a logical, sequential order. Maintain a third-person perspective and avoid personal opinions. Include specific details about components and their functions. Employ precise terminology and avoid vague or ambiguous phrasing. Summarize structural elements and their interactions within the system. Target a summary length that comprehensively covers the invention's scope.

Table 5: Bad-Batch self-generated guidelines used in Figure 4.

Model	**Dataset**	**LenSpecific Guidelines**
Gemma-3-4B	***MultiLexSum Long***	* Focus on conveying the core legal issue and the outcome of the case concisely. * Use the passive voice to maintain a formal and objective tone. * Limit each sentence to approximately 15-20 words, prioritizing clarity and brevity. * Include key factual details (parties involved, location, legal basis) without unnecessary elaboration. * State the settlement or ruling directly and clearly. * Avoid legal jargon where possible, and define any necessary terms briefly. * Maintain a neutral and impartial perspective throughout the summary. * Structure the summary with a logical flow, starting with the initial action and ending with the resolution. * ~~Keep the summary at approximately 100-150 words in length.~~ Target a summary consisting of multiple paragraphs with a total length of 600-650 words.
Gemma-3-4B	***MultiLexSum Short***	* Summarize the core legal issue (discrimination) concisely, focusing on the type of discrimination (gender) and the parties involved. * Maintain a formal, objective tone, mirroring the legal document's style. * Limit the summary to ~~approximately 60-80~~ a single paragraph consisting of approximately 130-150 words, prioritizing essential information. * Use the passive voice where appropriate to maintain a neutral perspective. * Avoid overly detailed descriptions of procedural steps (e.g., "filed a motion"). * Focus on the outcome of the case (consent decree, dismissal) rather than the extensive litigation details. * Include key terms like "Equal Employment Opportunity Commission," "Consent Decree," and "discrimination." * Structure the summary in a logical sequence: issue, parties, resolution. * Ensure the summary is self-contained and understandable without reference to the original document.
Gemma-3-4B	***MultiLexSum Tiny***	* Summarize the core legal issue (discrimination) concisely, focusing on the central claim. * Maintain a formal, objective tone appropriate for legal documents. * Use the third person to describe the events and parties involved. * Provide a high-level overview of the resolution (Consent Decree). * Limit the summary to ~~approximately 60-80 words~~ a single sentence of up to 25-30 words long. * Include key details like the parties involved (EEOC, defendants, plaintiff). * Mention the outcome of the case (dismissal with prejudice). * Focus on the essential facts and legal findings, omitting minor details. * Employ clear and concise language, avoiding jargon where possible.
Gemma-3-12B	***MultiLexSum Long***	Focus on the core allegations and resolution. Maintain a formal, objective tone. Use concise language and avoid unnecessary detail. Structure the summary chronologically. Clearly state the parties and their roles. Summarize legal proceedings and outcomes succinctly. Limit the summary to essential facts and rulings. Employ declarative sentences with moderate complexity. Target a summary consisting of multiple paragraphs with a total length of 600-650 words.
Gemma-3-12B	***MultiLexSum Short***	Focus on the core legal claims and outcomes. Maintain a formal and objective tone. Use concise sentences and avoid excessive detail. Present information in a chronological order. Highlight key parties and their roles. Adhere to a neutral grammatical person (third person). Summarize the final disposition of the case. Prioritize factual accuracy and legal relevance. Target a single paragraph of approximately 130-150 words length.
Gemma-3-12B	***MultiLexSum Tiny***	Focus on the core legal claims and resolution. Maintain a formal and objective tone. Summarize key actions and outcomes concisely. Use declarative sentences and avoid excessive detail. Structure the summary chronologically or thematically. Limit the summary to essential facts and legal terminology. Adhere to a professional and concise writing style. Avoid subjective interpretations or opinions. Limit the summary to a single sentence of up to 25-30 words long.
Gemma-3-27B	***MultiLexSum Long***	Prioritize conveying key legal and procedural details, including case names, court locations, and specific actions taken by the court. Maintain a formal and objective tone, avoiding interpretive language or subjective assessments. Focus on summarizing the *sequence* of events, rather than simply listing facts. Use complete sentences and avoid overly concise or telegraphic phrasing. Include monetary amounts and specific dates when they are central to the case's outcome. Employ precise legal terminology where appropriate, but explain it if necessary for clarity. ~~Summaries should generally be between 150-300 words to adequately cover the essential information.~~ Target a summary consisting of multiple paragraphs with a total length of 600-650 words. Retain passive voice where it reflects legal documentation style and avoids attributing agency unnecessarily.
Gemma-3-27B	***MultiLexSum Short***	Prioritize conveying key legal outcomes and agreements over detailed procedural history. Focus on the core claims, resolutions, and financial settlements. Maintain a formal and objective tone, avoiding subjective language. Use complete sentences and structured paragraphs for clarity. Summarize in the present tense whenever possible to convey lasting impact. Keep summaries concise, ~~generally within 3-5 sentences.~~ targeting a single paragraph of approximately 130-150 words length. Emphasize the actions of entities (e.g., the court, the parties) rather than individual actors. Avoid including extraneous details like docket numbers or filing dates.
Gemma-3-27B	***MultiLexSum Tiny***	Focus on core legal claims and resolutions, omitting detailed procedural history unless crucial to the outcome. Maintain a formal, objective tone and grammatical person (typically third person). Summarize in complete sentences, mirroring the complexity of legal writing. Prioritize conveying the key parties, allegations, and ultimate disposition of the case. Limit ~~length to approximately 3-5 sentences~~ the summary to a single sentence of up to 25-30 words long, emphasizing conciseness. Avoid excessive detail regarding motions or intermediate steps; focus on final rulings. Clearly state the legal basis for the claims and the outcome of the case. Use precise language reflecting legal terminology where appropriate.

Table 6: Edits to the self-generated guidelines of *MultiLexSum* to target specific lengths, used in Table 3.