Enhancing Long Document Long Form Summarisation with Self-Planning

Anonymous ACL submission

Abstract

We introduce a novel approach for long context summarisation, *highlight-guided generation*, that leverages sentence-level information as a content plan to improve the traceability and faithfulness of generated summaries. Our framework applies self-planning methods to identify important content and then generates a summary conditioned on the plan. We explore both an end-to-end and two-stage variants of the approach, finding that the two-stage pipeline performs better on long and information-dense documents. Experiments on long-form summarisation datasets demonstrate that our method consistently improves factual consistency while preserving relevance and overall quality. On GovReport, our best approach achieves up to 4.1 improvement in ROUGE-L and about 35% gains in SummaC scores. Qualitative analysis shows that highlight-guided summarisation helps preserve important details, leading to more accurate and insightful summaries across domains.

1 Introduction

Despite the strong text generation capabilities of current large language models (LLMs), generated long-form summaries often diverge significantly from human references in both content and style (Saxena et al., 2025). When prompted for conciseness and relevance, LLMs frequently fail to operationalise these, i.e., they struggle to identify key information and remove unnecessary details. Moreover, their outputs are prone to hallucinations (Askari et al., 2025; Belém et al., 2025; Chrysostomou et al., 2024).

Planning based approaches have been proposed to improve content selection (both in terms of saliency and coverage) as well as faithfulness in summarisation. Most of these approaches rely on complex intermediate plans of different granularity such as entity chains (Narayan et al., 2021), keyphrases (Xu et al., 2024), question-answer pairs (Narayan et al., 2023), events (Grenander et al., 2025), discourse relations (Liu et al., 2025), and topic templates (Perez-Beltrachini et al., 2019). In long-document summarisation, content selection is often implemented through an extract-then-generate pipeline, where sentences are selected using a trained classifier (Liu and Lapata, 2019; Ou and Lapata, 2025) or similarity heuristics (Erkan and Radev, 2004).

In this work, we argue that LLMs possess enough knowledge to identify summary worth content in input documents to make their own plans. We propose a simple and effective approach without training based on self-planning, *highlight-guided generation* (HIGEN). We instruct LLMs to generate a summary together with its plan, i.e., a set of sentences highlighting summary worth content from the input document to support the generation of a summary. We study two self-planning approaches. One where the sentence highlights are generated along with the summary and a revision-based one where the sentence highlights are fed back to the model together with the input document to generate a refined summary based on the highlights.

An alternative self-planning approach can be implemented with attribution methods, which identify parts of the input that the LLM relies on when generating summaries. We compare planning based on generative highlights versus planning based on extractive attribution methods. Concretely, we compare with a perturbation-based attribution method that extracts those input document sentences that yield a decrease in summary quality when they are removed from the input. Generated highlights offer key advantages over attribution-based methods: they preserve contextual coherence (e.g., maintaining speaker-utterance relationships in dialogues), are computationally more efficient than perturbation-based approaches, and can synthesise information rather than just extracting sentences.

We evaluate our approach on two long-context summarisation datasets, including GovReport and QMSum, and measure summary quality in terms of relevance and faithfulness. Our experiments and analysis show that self-planning can effectively improve the overall quality of the generated summaries by enumerating summary worth points. In query-based summarisation, LLMs are more likely to provide relevant and targeted information with the help of highlights.

2 Method

We propose a novel self-planning summarisation framework for long-document summarisation that leverages sentence plans derived from the input document to guide the summary generation. Our approach is motivated by the observation that while LLMs possess sufficient knowledge to identify relevant content in input documents, they struggle with maintaining focus and avoiding hallucination in long-context scenarios. By explicitly extracting important sentences as an intermediate content selection step, we aim to improve both the factual consistency and relevance of generated summaries.

Given an input document $D = \{s_1, s_2, \ldots, s_n\}$ consisting of n sentences, our goal is to generate a summary S that is both faithful to the source content and covers the most important information. Traditional approaches directly map $D \rightarrow S$, while we introduce an intermediate content selection step by first identifying a subset of important sentences $H = \{h_1, h_2, \ldots, h_k\} \subseteq D$ where $k \ll n$, and then generating the summary conditioned on these highlights: $D \rightarrow H \rightarrow S$.

Our framework consists of two main components: (i) *important sentence extraction*, which identifies the most important sentences from the input document, and (ii) *highlight-guided summarisation*, which generates the final summary based solely on the extracted highlights . We explore two architectural variants that differ in how these components are integrated: an end-to-end approach that performs both steps in a single generation pass, and a two-stage pipeline that separates the highlighting and summarisation processes.

2.1 End-to-end Approach

In the end-to-end variant, we prompt the LLM to sequentially perform highlight extraction and summary generation within a single inference call. The model is instructed to first identify and extract important sentences H from the input document, then immediately generate a summary S based only on the information contained in these highlights.

2.2 Two-stage Pipeline

To address the limitations of the end-to-end approach, we propose a two-stage pipeline that separates highlight extraction and summary generation into distinct inference calls. In the first stage, the model extracts important sentences from the input document. In the second stage, a fresh model context is used to generate the summary, with both the original document and the extracted highlights provided as input, but with explicit instructions to base the summary only on the highlighted content. The two-stage process can be formalised as: $H = \text{LLM}(\text{prompt}_h, D)$ and $S = \text{LLM}(\text{prompt}_s, D, H)$, where $\text{LLM}(\cdot)$ denotes the language model inference function, prompt_h and prompt_s are the task-specific prompts for highlight extraction and summary generation respectively (see Appendix B for more details), D is the input document, H represents the extracted highlights, and S is the final summary.

This separation offers several advantages: (i) it provides more reliable instruction following by focusing each generation step on a single task, (ii) it enables the use of different highlighting methods beyond generative extraction, and (iii) it allows for better control over the summary generation process by providing clear conditioning information .

2.3 Attribution Methods

Our two-stage framework supports multiple methods for extracting important sentences. A self-planning alternative to highlight generation is context attribution. Context attribution involves tracing and quantifying the influence of specific input segments on the generated output. In this work, we investigate attribution methods that rely solely on the model's internal mechanisms. We aim to investigate whether model attribution can effectively support content selection and guidance in long-context summarisation.

Perturbation-based methods quantify the importance of input sources by systematically perturbing the input and measuring the resulting changes in the model outputs, such as through occlusion (Zeiler and Fergus, 2014; Ribeiro et al., 2016; Mohebbi et al., 2023; Zhao and Shan, 2024; Cohen-Wang et al., 2024). In this work, we employ ContextCite (Cohen-Wang et al., 2024), a recently pro-

posed context attribution method that identifies which parts of the input context most causally influence a model's generation by systematically ablating context elements and measuring the changes in output probabilities of the original response.

3 Experiment Setting

Datasets. We report the results on two long-form summarisation datasets from SCROLLS benchmark (Shaham et al., 2022), including GovReport and QMSum.

Evaluation metrics. We report several automatic metrics to assess various aspects of the generated summaries. We use ROUGE-L (Lin, 2004) and BERTSCORE-F1 (Zhang et al., 2020) to measure the *relevance* of the summaries against human references. We employ SUMMAC (Laban et al., 2022) and FACTSCORE (Min et al., 2023) to assess the *factual consistency* between the generated summary and input document. SUMMAC measures the overall consistency based on sentence-level entailment. Scores reported in the paper are computed using the SUMMAC_{Conv} model. In this work, we adapt it to assess factual consistency by computing the percentage of atomic facts in the generated summary that are supported by the input document. We use `gpt-4o-mini` model to compute FACTSCORE values. Additionally, we report the average length of the generated summary, measured in terms of the number of tokens.

Models and Baselines. We evaluate the performance of `Llama3.1-8B`, `Qwen3-8B` and `Qwen3-32B` on the long-form summarisation datasets, with and without attribution-guided summarisation. The experiments are conducted in a zero-shot setting. As baselines for comparison, we consider direct prompting and Summary Chain-of-Thought (SumCoT, Wang et al., 2023). In the direct prompting setting, the model directly generates the summary given the input document, without content selection steps. SumCoT is a two-stage pipeline that leverages a QA-based plan to guide the summarisation process. LLMs are instructed to extract important information about entities and events by answering a list of guiding questions and then produce a summary with more fine-grained detail by integrating extracted information.

Hyperparameters. When generating the summaries, we apply greedy decoding with a temperature of 0 to produce deterministic outputs. For the results reported in Table Table 1, we extract 30 sentences as the attribution-based highlights.

4 Results

Results in Table 1 highlight consistent trends that demonstrate the effectiveness of our attribution-guided summarisation approach in long-context summarisation. Compared with the baselines, the proposed two-step pipeline significantly enhances both relevance and factual consistency for all models considered in the experiments on GovReport. For example, with `Qwen3-8B` model, the attribution-guided approach helps improve ROUGE-L from 43.08 to 47.20, indicating that the generated summaries more closely align with the content covered in the human references. Meanwhile, SUMMAC increases from 47.97 to 65.73 and FACTSCORE improves from 0.8999 to 0.9107, which suggests the generated summaries are more faithfully supported by the input document. Our two-stage pipeline consistently outperforms the end-to-end approach on the GovReport dataset, showing significantly better ROUGE-L and SUMMAC scores, while the two model variants show comparable performance on the QMSum dataset. This result implies that a separate content selection step benefits more in complex documents with dense information.

Generative highlights achieve a better balance between relevance and faithfulness. We compare the performance of generative highlights against attribution-based highlights in our summarisation framework. For the experiments with ContextCite attribution, we only take into account important sentences with non-zero attribution scores when generating the summary. Table 1 shows that the relevance of the summaries guided by ContextCite attribution is comparable to the summaries guided by generative highlights. Generative highlights consistently outperform ContextCite attribution in terms of factual consistency, showing better SUMMAC and FACTSCORE scores across different models and datasets. Example summaries in Table 2 demonstrate that ContextCite attribution encourages the model to produce a comprehensive summary that is rich in detail, including both the final decisions and specific action items. The summary can be lengthy and verbose compared with the summaries guided by generative highlights.

Table 6 shows that in the meeting summarisation task, many highlights extracted by ContextCite attribution are not informative, while the generative highlights are able to extract the key facts by syn-

Model	Method	GovReport					QMSum				
		R-L	BS-F1	SumC	Fact	#Tokens	R-L	BS-F1	SumC	Fact	#Tokens
Llama3.1-8B	Direct	45.49	62.08	53.48	0.7880	532.08	21.99	57.22	36.61	0.7162	139.58
	SumCoT	44.58	61.83	50.53	0.8627	545.61	22.49	56.14	38.05	**0.8179**	131.71
	HiGen-CC	44.12	61.40	59.77	0.7687	479.18	22.28	57.30	36.45	0.6975	134.15
	HiGen (end-to-end)	39.62	62.14	59.23	**0.8845**	382.49	**23.78**	**58.05**	**38.39**	0.7720	118.53
	HiGen (two-step)	**47.18**	**63.08**	**65.68**	0.8338	566.41	22.76	57.78	36.51	0.7688	142.38
Qwen3-8B	Direct	43.08	63.35	47.97	0.8999	491.42	22.56	58.43	37.88	0.8157	121.53
	SumCoT	34.19	60.66	43.92	0.9014	307.04	22.02	57.67	**39.78**	0.8479	108.44
	HiGen-CC	46.60	64.46	56.34	0.8834	639.04	**23.38**	**58.74**	37.49	0.7942	127.24
	HiGen (end-to-end)	38.75	58.67	46.54	0.8570	407.37	22.35	58.39	39.39	0.8501	105.14
	HiGen (two-step)	**47.20**	**64.71**	**65.73**	**0.9107**	709.32	22.94	58.71	38.23	**0.8628**	131.39
Qwen3-32B	Direct	43.19	63.59	48.14	**0.9033**	453.16	21.76	**58.25**	37.01	0.8002	132.22
	SumCoT	36.40	61.90	43.48	0.8996	326.73	22.27	57.84	37.44	0.8039	125.50
	HiGen (end-to-end)	45.80	63.78	50.23	0.8641	570.81	**22.37**	58.24	**38.25**	0.7621	119.96
	HiGen (two-step)	**46.39**	**64.14**	**60.82**	0.8998	619.77	21.62	58.01	36.88	**0.8336**	144.40

Table 1: Model performance on GovReport and QMSum validation sets measured in terms of ROUGE-L, BERTScore-F1, SummaC, FactScore, and summary length in tokens. Bold indicates best results per approach.

thesising information in the local context. Generative highlights can preserve the speaker-utterance correspondence in the dialogue by converting the utterance into a concise statement.

Qualitative analysis. We conduct a qualitative analysis on 20 examples drawn from each of the GovReport and QMSum datasets. As shown in the examples in the Appendix, the baseline summaries often capture a broad and high-level overview of the meeting, with a focus on the key decisions made during the meeting, rather than addressing the specific query. While SumCoT summaries provide a consistent and clear structure that involves meeting attendees, discussion topics and main decisions, speaker attribution, concrete technical detail and the rationale behind decisions are often omitted in the summaries. Summaries guided by the highlights include not only the final decisions, but also the core rationale and trade-offs behind them. Our analysis reveals that highlight-guided summarisation can help preserve important details, such as entities, terminology, and quantitative data, which are often omitted in summaries generated by direct prompting. Our technique also proves beneficial in query-based summarisation, where the model leverages the extracted highlights to identify relevant information and generate more targeted and query-aligned summaries.

5 Conclusions

We introduced an highlight-guided summarisation framework for long-document summarisation that leverages important sentence-level information to improve both factual consistency and relevance of generated summaries. Our approach addresses key challenges in long-context summarisation by explicitly identifying important content before generation, mimicking human summarisation processes. Our experiments on GovReport and QMSum demonstrate consistent improvements across multiple models. The two-stage pipeline achieves substantial gains in ROUGE-L scores (up to 4.1 points on GovReport) and factual consistency metrics, with SummaC scores improving from 48.0 to 65.7. Our qualitative analysis reveals that the proposed framework can help preserve important details such as entities, terminology, and quantitative data that are often omitted in direct prompting approaches.

Limitations

Our approach has several limitations that we aim to address in future work. The computational overhead of the two-stage pipeline increases inference time and resource requirements compared to direct prompting, particularly when using perturbation-based attribution methods. The generative highlighting approach, while effective, may introduce hallucinations and other problems inherent to the underlying LLM. Finally, our experiments are limited to two datasets and three model architectures, which may limit the generalisability of our findings across different domains and model scales.

References

Hadi Askari, Anshuman Chhabra, Muhao Chen, and Prasant Mohapatra. 2025. Assessing LLMs for zero-

shot abstractive summarization through the lens of relevance paraphrasing. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2187–2201, Albuquerque, New Mexico. Association for Computational Linguistics.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. In *ACL (1)*, pages 3119–3137. Association for Computational Linguistics.

Catarina G Belém, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. 2025. From single to multi: How LLMs hallucinate in multi-document summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5276–5309, Albuquerque, New Mexico. Association for Computational Linguistics.

George Chrysostomou, Zhixue Zhao, Miles Williams, and Nikolaos Aletras. 2024. Investigating hallucinations in pruned large language models for abstractive summarization. *Trans. Assoc. Comput. Linguistics*, 12:1163–1181.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: Attributing model generation to context. In *NeurIPS*.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Matt Grenander, Siddharth Varia, Paula Czarnowska, Yogarshi Vyas, Kishaloy Halder, and Bonan Min. 2025. Exploration of plan-guided summarization for narrative texts: the case of small language models. *CoRR*, abs/2504.09071.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Trans. Assoc. Comput. Linguistics*, 10:163–177.

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.

Dongqi Liu, Xi Yu, Vera Demberg, and Mirella Lapata. 2025. Explanatory summarization with discourse-driven planning. *CoRR*, abs/2504.19339.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP/IJCNLP (1)*, pages 3728–3738. Association for Computational Linguistics.

Louis Mahon and Mirella Lapata. 2024. A modular approach for multimodal summarization of TV shows. In *ACL (1)*, pages 8272–8291. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*, pages 12076–12100. Association for Computational Linguistics.

Hosein Mohebbi, Willem H. Zuidema, Grzegorz Chrupala, and Afra Alishahi. 2023. Quantifying context mixing in transformers. In *EACL*, pages 3370–3392. Association for Computational Linguistics.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Trans. Assoc. Comput. Linguistics*, 11:974–996.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan T. McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Trans. Assoc. Comput. Linguistics*, 9:1475–1492.

Litu Ou and Mirella Lapata. 2025. Context-aware hierarchical merging for long document summarization. In *ACL (Findings)*, pages 5534–5561. Association for Computational Linguistics.

Laura Perez-Beltrachini, Yang Liu, and Mirella Lapata. 2019. Generating summaries with topic templates and structured convolutional decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5107–5116, Florence, Italy. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *HLT-NAACL Demos*, pages 97–101. The Association for Computational Linguistics.

Rohit Saxena, Hao Tang, and Frank Keller. 2025. End-to-end long document summarization using gradient caching. In *ACL 2025*.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding. In *EMNLP*

(Findings), pages 7977–7989. Association for Computational Linguistics.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: standardized comparison over long language sequences. In *EMNLP*, pages 12007–12021. Association for Computational Linguistics.

David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2025. On positional bias of faithfulness for long-form summarization. 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 8791–8810, Albuquerque, New Mexico. Association for Computational Linguistics.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Lei Xu, Mohammed Asad Karim, Saket Dingliwal, and Aparna Elangovan. 2024. Salient information prompting to steer content in prompt-based abstractive summarization. In *EMNLP (Industry Track)*, pages 35–49. Association for Computational Linguistics.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *ECCV (1)*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*. OpenReview.net.

Zhixue Zhao and Boxuan Shan. 2024. Reagent: A model-agnostic feature attribution method for generative language models. *CoRR*, abs/2402.00794.

A Experimental Setup Details

A.1 Dataset Details

The licenses for the datasets used in our experiments are as follows. QMSum is available under MIT License, and the original GovReport dataset is available under CC-BY-4.0 License. For both datasets, we use the version from SCROLLS benchmark, which is under MIT License. We run experiments on 300 samples from the GovReport validation set. Experiments on QMSum are run on the whole validation set (272 samples).

A.2 Implementation Details

We implement ROUGE-L and BERTSCORE using `evaluate` library. BERTSCORE is computed by `DeBERTa-xlarge-mnli model` (He et al., 2021), We adopt the implementation of FACTSCORE from PRISMA code repository (Mahon and Lapata, 2024) and use `GPT-4o-mini` for both atomic fact extraction and claim verification. We adapt the implementation of ContextCite from (Cohen-Wang et al., 2024) to extract ContextCite attributions. Input sentences are ranked by their attribution scores, and the top-k sentences are selected as highlights. We extract 30 attributed sentences for each instance. We only take into account the important sentences with non-zero attribution scores when producing the summaries.

B Prompt Templates

We present the prompt templates used for highlight extraction and summary generation in this section. Fig. 1 and Fig. 2 show the prompt templates used for the experiments on GovReport dataset. We adapted the standard prompt used in LongBench (Bai et al., 2024) and ZeroSCROLLS Benchmark (Shaham et al., 2023) and added instructions to enforce structured output.

Fig. 3 and Fig. 4 show the prompt templates used for the experiments on QMSum dataset. The prompt format is adapted from (Wan et al., 2025).

C Computation Details

Experiments with ContextCite attribution as highlights were run on four NVIDIA A100 GPUs with 80GB of GPU memory. Other experiments were run on two NVIDIA A100 GPUs. The GPU hours vary depending on the model size and average context length in the dataset. Extracting ContextCite attribution using `Qwen3-8B` model on GovReport validation set takes about 10 hours. Generating the highlights and summaries on QMSum or GovReport takes about 30 minutes to 1 hour with `vllm`. (Kwon et al., 2023)

Hightlight Extraction + Summary Generation

You are given a report by a government agency. Extract a list of `{Number of sentences}` key sentences from the input document and then write a one-page summary of the report only focusing on the extracted sentences. You must give your answer in a structured format: "Key Sentences:
1. `{Sentence Text}`
2. `{Sentence Text}`
...
Summary: [your summary]", where [your summary] is your generated summary.

Report:
`{Document Text}`

Figure 1: Prompt used for end-to-end highlight extraction and summary generation on GovReport

D Qualitative Examples

This section provides qualitative examples of summaries generated by different methods and different types of highlights.

Table 6 demonstrates the difference between the salient sentences extracted by ContextCtie attribution and highlight sentences generated by LLM on a random instance from QMSum validation set. Both ContextCite attribution and generated highlights are computed using `Qwen3-8B` model.

Summary Generation

You are given a report by a government agency. Write a one-page summary of the report focusing on the main points. You must give your answer in a structured format: "Summary: [your summary]", where [your summary] is your generated summary.

Report:
`{Document Text}`
You should only focus on the following key points:
1. `{Sentence Text}`
2. `{Sentence Text}`
...

Figure 2: Prompt used for generating the summary with the two-step pipeline on GovReport.

Hightlight Extraction + Summary Generation

Read the following meeting transcript. Extract a list of `{Number of sentences}` key sentences from the input document and then produce a summary in 4 sentences only focusing on the extracted sentences. You must give your answer in a structured format: "Key Sentences:
1. `{Sentence Text}`
2. `{Sentence Text}`
...
Summary: [your summary], where [your summary] is your generated summary.
==========
`{Meeting Transcript}`
==========

Figure 3: Prompt used for end-to-end highlight extraction and summary generation on QMSum

Summary Generation

Read the following meeting transcript. Produce a summary in 4 sentences focusing on key decisions, action items, and important discussion points. You must give your answer in a structured format: "Summary: [your summary]", where [your summary] is your generated summary.
==========
`{Meeting Transcript}`
==========
You should only focus on the following key points:
1. `{Sentence Text}`
2. `{Sentence Text}`
...

Figure 4: Prompt used for generating the summary with the two-step pipeline on QMSum.

Gold Summary
The industrial designer recommended batteries instead of solar cells because it would be problematic if solar cells didn't have enough energy. Titanium was proposed rather than rubber as the latter would make it difficult to do the moulding of buttons. Besides, the industrial designer wanted a simple push button because it only needed a simple chip and was less expensive than LCD.
Baseline Summary
The industrial designer recommended using simple batteries instead of solar cells and titanium for the remote control's material due to manufacturing and cost concerns. They also suggested push buttons over LCD displays to reduce cost and complexity, while acknowledging the possibility of an LCD for information output if budget allows. The team discussed incorporating trendy, spongy, and fruit-like designs, considering both titanium and rubber materials, and explored the idea of making the remote control customizable with additional features like color options or speech recognition. The final decision was to focus on creating a fancy and innovative product that aligns with current design trends, even if it means sacrificing some functionality for aesthetic appeal.
SumCoT Summary
The document describes a conceptual design meeting involving the Project Manager, Marketing, User Interface, and Industrial Designer, focusing on the design of a remote control. Key discussions revolve around materials, interface options, and incorporating trend elements like a spongy, fruit-shaped aesthetic. The team decides to use simple batteries and push buttons instead of solar cells or LCD displays due to cost and manufacturing constraints. They plan to explore two design options—titanium with a fruit scent or a spongy rubber body—while keeping the product aligned with current trends and within budget.
Highlight-driven planning (End-to-end)
The industrial designer recommended using simple batteries and push buttons for the remote control due to cost and manufacturing constraints, while also considering the use of titanium for design. They discussed the feasibility of using a speech recognition chip to enhance functionality without increasing costs too much. The team also explored the idea of incorporating trendy, natural elements like fruit and vegetable shapes and colors, and considered optional features such as color-changing displays or fingerprint recognition. Ultimately, they decided to focus on creating a fancy, user-friendly remote control that aligns with current design trends.
Highlight-driven planning (Two-step, ContextCite)
The team discussed the design of industrial components for the remote control, deciding to use simple batteries instead of solar cells due to energy concerns, and opted for titanium over rubber or plastic for its aesthetic appeal, despite manufacturing challenges. They agreed to use push buttons rather than LCD displays to reduce costs and simplify the technology, while also considering the inclusion of speech recognition for added functionality. Marketing emphasized the importance of making the remote control f̈ancyänd highlighted the trend of using fruit and vegetable-inspired colors and materials to differentiate the product. The project manager assigned individual tasks for Rama, Mark, and Sammy, with the group planning to create a prototype using modeling clay and explore both titanium and spongy design options.
Highlight-driven planning (Two-step, Generative)
The industrial designer recommended using simple batteries for energy and titanium for the material due to manufacturing feasibility and cost-effectiveness, while avoiding double-curved shapes. They also suggested using push buttons instead of an LCD display to reduce costs and simplify technology, as the LCD requires more advanced and expensive components. The team discussed the possibility of incorporating speech recognition technology, which would allow for more features but needed to be balanced with budget constraints. The designer proposed exploring a spongy, fruit-shaped design with titanium panels and a rubber body, aiming to align with current trends and enhance the product's appeal.

Table 2: Comparison between different summaries generated by `Qwen3-8B` model for a random sample from QMSum validation set. Summaries produced using highlight-driven planning are guided by 30 sentences.

Gold Summary
Marketing suggested putting a small LCD screen on top of the remote but Industrial Designer against it because a remote should be started with the power button and User Interface thought it would be unnatural to put the LCD on the top. Then, Marketing mentioned that would be where users' handball might be. So Industrial Designer thought they could put the whole interface a bit down so there would be room for the interface.
Baseline Summary
The group discussed the pros and cons of including an LCD screen on the remote, emphasizing its importance for the younger target audience despite cost concerns. They agreed to prioritize the most frequently used functions like power, channel, and volume, placing them in a central and accessible location to reduce RSI issues. The design should be familiar yet modern, incorporating the company's corporate image and ensuring usability. Action items include evaluating the cost of LCD screens and finalizing the placement of the screen and buttons before the next design round.
SumCoT Summary
The meeting involves the Marketing team, Project Manager, Industrial Designer, and User Interface Designer discussing the design and functionality of a remote control. They are considering the inclusion of an LCD screen and debating its placement, as well as the layout and usability of buttons. The group agrees to include an LCD screen but needs to finalize its position and cost. They also decide to prioritize the most commonly used functions and focus on a younger audience, aiming for a user-friendly and brand-recognizable design.
Highlight-driven planning (End-to-end)
The group discussed the design and functionality of a remote control, emphasizing the need for a familiar yet modern look, with a focus on user-friendly features like an LCD screen. They considered the placement of buttons, aiming to reduce RSI by spacing out frequently used ones. The team also debated the importance of a recognizable corporate image and the inclusion of features like a beeping button to locate the remote. Despite the cost concerns, they agreed that an LCD screen is essential to meet the preferences of their target younger audience.
Highlight-driven planning (Two-step, ContextCite)
The group discussed the pros and cons of including an LCD screen on the remote, noting that while it is important for the younger audience and offers useful features like program recognition, it may be costly and not essential for all users. They agreed that the most important functions—power, channel, and volume—should be placed near the thumb to reduce the risk of RSI, but also considered spacing them apart to avoid repetitive strain. The Industrial Designer mentioned the need to consider the cost and design implications of the LCD screen, while the Marketing team emphasized the importance of a user-friendly design and the need to prioritize key features due to time constraints. The group ultimately decided to include an LCD screen but acknowledged the need to explore its feasibility and cost before finalizing the design.
Highlight-driven planning (Two-step, Generative)
The group discussed the pros and cons of including an LCD screen on the remote, noting its importance for the younger audience but acknowledging the cost concerns. They agreed that the remote should be designed to fit the hand, be original yet familiar, and incorporate a recognizable corporate image, possibly through a slogan or color. The decision was made to include an LCD screen, though its placement and functionality will be further discussed, with the focus on making it user-friendly and intuitive. The team also emphasized the need for a short learning curve, with the most frequently used buttons (power, channel, volume) placed for easy access and to reduce RSI issues.

Table 3: Comparison between different summaries generated by `Qwen3-8B` model for a random sample from QMSum validation set. Summaries produced using highlight-driven planning are guided by 30 sentences.

Gold Summary
When the group found that it was hard to balance between the cost and functions, the Project Manager proposed to give up the voice recognition and keep the touch screen. But according to Marketing's research, even if the interviewees were not informed of the touch screen, they showed a great preference for voice recognition. So just for catering to popular tastes, the Marketing insisted that they should keep the voice recognition in any case.
Baseline Summary
The team discussed the feasibility of incorporating a touch screen into the remote control design but ultimately decided to prioritize voice recognition due to its lower cost and existing in-house technology. They agreed to use rubber buttons and a non-traditional, organic design for the casing to align with current market trends and appeal to the youth market. A backup battery was suggested to complement solar power, ensuring reliability in low-light conditions. The next steps involve developing a prototype with a simple, non-remote-looking design, focusing on voice recognition and a sleek, versatile interface, with Nathan and Ron collaborating on the design and layout.
SumCoT Summary
The document describes a meeting between the Marketing team, Project Manager, Industrial Designer, and User Interface team to discuss the design of a new remote control. The main topic of debate is whether to use a touch screen or voice recognition technology, with the latter being chosen due to budget constraints. They also consider materials for the case, such as rubber or natural options, and aim for a simple, non-traditional design. The result is a decision to prioritize voice recognition and focus on a sleek, user-friendly remote control that aligns with current market trends.
Highlight-driven planning (End-to-end)
The meeting discusses the conceptual design of a remote control, focusing on the touch screen and voice recognition features. Marketing is concerned about the cost of the touch screen, while the Industrial Designer suggests it could be a selling point. They debate between touch screen and voice recognition, ultimately deciding to prioritize voice recognition due to its lower cost and existing in-house technology. The team also considers alternative designs and materials for the remote, aiming to create a unique and user-friendly product.
Highlight-driven planning (Two-step, ContextCite)
The team discussed the limitations of the manufacturer in component design, particularly regarding the touch screen and energy sources, and considered alternatives like solar or kinetic power. They decided to provisionally go with the touch screen as a main selling point but acknowledged the high costs involved. After evaluating the costs and market appeal, they agreed to prioritize voice recognition over the touch screen due to its fiscal viability and existing in-house technology. The Project Manager assigned Nathan to focus on the basic look and feel of the remote, while Ron was tasked with designing the layout and prototype, aiming to target the youth market with a sleek, non-traditional remote design.
Highlight-driven planning (Two-step, Generative)
The Marketing team disagreed with the Project Manager regarding the touch screen due to budget constraints, as it would significantly increase costs. The meeting focused on conceptual design, aiming to finalize decisions for the prototype and clarify roles. The Industrial Designer discussed component options, including energy sources and case materials, while highlighting the cost implications of touch screen technology. The team ultimately decided to prioritize voice recognition over touch screens to balance innovation with budget feasibility, and agreed to proceed with a simple, non-traditional remote design incorporating rubber and multiple case colors.

Table 4: Comparison between different summaries generated by `Qwen3-8B` model for a random sample from QMSum validation set. Summaries produced using highlight-driven planning are guided by 30 sentences.

Gold Summary
Mental spaces can be tackled with mechanisms that can also deal with context issues (time, space etc.): creating a base space and rules of interaction with other interconnected spaces. However, the complexity of these mechanisms has to be bound as well: it is necessary to define the range of constructions to be studied.
Baseline Summary
The group discussed the updated semantic specification and construction formalism, noting that while the notation is still evolving, it aims to capture both form and meaning types. They agreed to defer detailed work on mental spaces until Robert returns, focusing instead on refining the semantic constraints and constructions. The team also considered the distinction between focus and topic in information structure, acknowledging the need for further exploration. Finally, they planned to meet again on Thursday at 1 PM to review and refine the thesis proposal and related work.
SumCoT Summary
The discussion revolves around the development of a semantic formalism and the integration of mental spaces in linguistics, with researchers like Grad B, Grad F, Grad A, Grad E, Grad D, and Professor C collaborating on the topic. Grad A is set to leave on Thursday, and the group plans to meet again at a time to be determined, possibly at 1 PM or 4 PM. The group is working on refining their notation and semantic constraints, acknowledging the need for further discussion and adjustments. The result is an evolving framework that aims to capture both form and meaning, with a focus on constructions, discourse context, and the potential use of mental spaces for modeling linguistic phenomena.
Highlight-driven planning (End-to-end)
The meeting discusses the development of a semantic specification and construction formalism, highlighting the need for a notation that can represent both form and meaning. There is a focus on the distinction between semantic and formal constraints, as well as the use of terms like "designate" and "profile" in the context of mental spaces. The group acknowledges the complexity of handling information structure, including focus and topic, and agrees to defer detailed work on these aspects. They also plan to proceed with defining the scope of their work, focusing on specific domains and using feedback to refine their approach.
Highlight-driven planning (Two-step, ContextCite)
The group discussed the revised semantic specification and construction formalism for mental spaces, emphasizing the need to distinguish between form and meaning types. They acknowledged the complexity of mental spaces and decided to defer detailed work until Robert returns, focusing instead on defining the semantic constraints and constructions. The team agreed to prioritize getting the current framework right before considering more abstract or inheritance-based structures. They also noted the importance of handling discourse elements like focus and topic, and mentioned that the notation should allow for flexibility in capturing these aspects.
Highlight-driven planning (Two-step, Generative)
The meeting discussed the revised semantic specification and construction formalism, noting that the current notation is similar to previous versions with minimal changes. The group acknowledged the need to clarify terminology, such as "semantic constraints" and "designates," and agreed to defer further decisions on these terms. They also addressed the integration of mental spaces and discourse context, emphasizing the importance of focusing on the core aspects of the formalism before tackling more complex issues. Action items included refining the notation, discussing the semantic side of constructions, and planning a follow-up meeting to review progress and gather feedback.

Table 5: Comparison between different summaries generated by `Qwen3-8B` model for a random sample from QMSum validation set. Summaries produced using highlight-driven planning are guided by 30 sentences.

Source Document
Industrial Designer: so these are the different options that we have . Okay . So the batteries , I'll start with the battery , right ? Project Manager: Mm-hmm . Industrial Designer: So they can be simple which is like uh the normal batteries in uh our disfmarker uh the cells , yeah ? Project Manager: Yeah . Industrial Designer: Uh thes these are the kind disfmarker different kind of batteries that the company makes , right ? So . And dynamos . Um vocalsound Marketing: Does that mean like a wind-up one ? Industrial Designer: yeah , yeah . Marketing: vocalsound A wind-up remote . [...] Industrial Designer: on pressing this button I disfmarker a circuit completes , the information goes to the chip , which is somewhere here and the chip that tra then translates the code into an infra infrared radiation , which goes goes out through there . vocalsound So uh the important point that I read over the website was uh that the configurations of these printed circuit circuit boards uh are quite cheap to make , you can ge get them printed as you want to , [...] Industrial Designer: Yeah . Marketing: It it depends on the whole ergonomics of it , you know , it's like how you put your hands so y it's the least movement basically . Industrial Designer: Yeah . Yeah , singe single side curved or double side curved does not say too much , does it ? [...] Industrial Designer: Or or curved at one end and flat on the top , because I I'm not sure if it is flat on both both the sides , then ho how much easy would it be to reach for buttons , etcetera . Um disfmarker Marketing: You have to have a certain element of flatness , I think . [...] User Interface: Okay . So um I thought um I would also include the definition of user interface um so it's the aspects of a of of a computer system or programme which can be seen uh by the user um and and which disfmarker uh the mechanisms that the user uses to control its operation and input data . So this would p includes things like shape and size and buttons and um voice recognition as well , and colour , and so on . [...] User Interface: Um um the method I vocalsound employed this time was vocalsound a again having a look to related products and mainly on the internet and then vocalsound um vocalsound analyse them uh from the point of view of user fen friendliness and vocalsound also um vocalsound whether their appearance was was pleasant . Um vocalsound and then um vocalsound this uh this um vocalsound this can help us to decide which features we want to incorporate in our product . Project Manager: Mm-hmm . [...] Marketing: Imagine it would be inside . [...] User Interface: Mm-hmm . Marketing: Cool , I'm gonna go and sit on my own . Project Manager: Y ah nobody wants to talk to you .
ContextCite attribution
1. Project Manager: Mm-hmm . 2. Yeah , singe single side curved or double side curved does not say too much , does it ? 3. Marketing: Cool , I'm gonna go and sit on my own . 4. Marketing: You have to have a certain element of flatness , I think . 5. Project Manager: Yeah . 6. Project Manager: Mm-hmm . 7. Marketing: Imagine it would be inside . [...]
Generated highlights
1. The chip translates the code into infrared radiation. 2. The Industrial Designer noted that printed circuit boards are cheap to make. 3. The remote can use different types of batteries, including simple, dynamo, and kinetic energy. 4. Marketing suggested a wind-up remote as an idea for kinetic energy. 5. The User Interface Designer mentioned the definition of user interface. 6. The User Interface Designer analyzed products based on user friendliness and appearance. [...]

Table 6: Comparison between highlight sentences based on ContextCite attribution and highlights generated by LLMs on QMSum validation set