Decoupling Generation and Selection for Budget-Constrained Faithful Summarization

Anonymous ACL submission

Abstract

Recent abstractive summarization systems based on pretrained language models achieve strong fluency but often suffer from factual inconsistency, redundancy, and limited controllability over summary length. We propose a budget-constrained combinatorial framework that decouples summary generation from content selection for faithful summarization. A language model first generates multiple candidate summaries, which are decomposed into a sentence-level candidate pool. We then formulate summary construction as a combinatorial optimization problem that jointly models semantic coverage, factual consistency, and redundancy under an explicit budget constraint. To perform diversity-aware subset selection, we adopt a DPP-inspired optimization strategy that favors informative and non-redundant candidate subsets. Experiments on CNN/DailyMail and Multi-News demonstrate improved factual consistency and redundancy control while maintaining competitive summarization quality across multiple pretrained generation models. Our code is publicly available at: `https://anonymous.4open.science/r/bcfs-D05E/`.

1 Introduction

Recent abstractive summarization systems based on pretrained language models and large language models (LLMs) achieve strong fluency and semantic quality (Li et al., 2024; Kim et al., 2024). However, despite these advances, modern summarization systems still suffer from three persistent limitations: *factual inconsistency*, *redundancy*, and *limited controllability*. Because most approaches formulate summarization as autoregressive sequence generation, the final summary is produced token-by-token without explicit global content planning. As a result, generated summaries may contain hallucinated information, repeated content, or poorly controlled length allocation (Kim et al., 2024; Tang et al., 2024b).

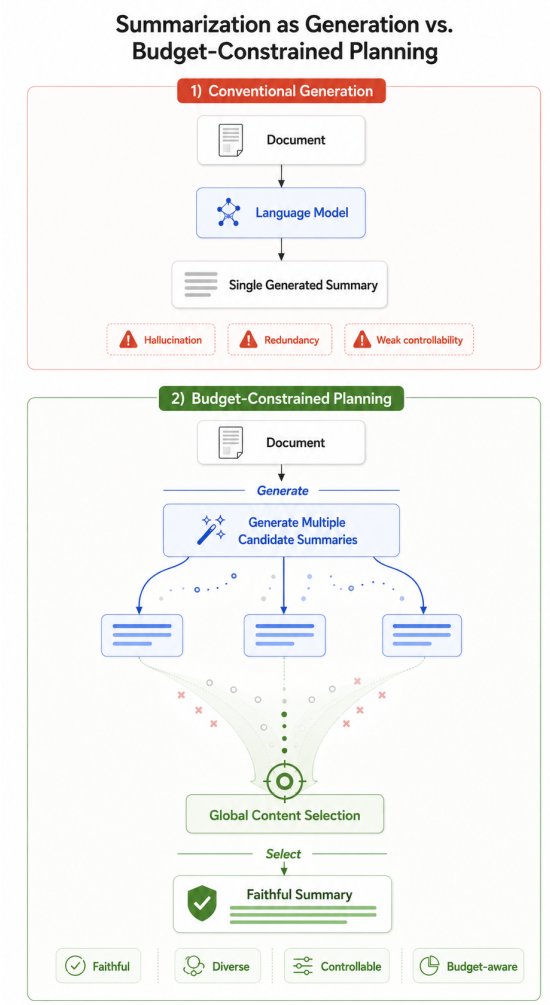

Figure 1: From generation to planning for abstractive summarization. Conventional autoregressive summarization relies on a single decoding trajectory, often leading to hallucination, redundancy, and weak controllability. Our framework decouples generation and selection by exploring multiple candidate summaries and globally selecting content under explicit budget constraints to produce faithful and diverse summaries.

Figure 1 illustrates this contrast. Conventional generation-based summarization directly maps a document into a single generated summary through

conditional language modeling. While effective for fluency, this formulation provides limited mechanisms for globally optimizing semantic coverage, factual consistency, and diversity.

In this work, we instead formulate abstractive summarization as a *budget-constrained combinatorial content planning problem*. Rather than relying on a single decoding trajectory, a pretrained language model first generates multiple candidate summaries, which are decomposed into a sentence-level candidate pool. A combinatorial optimization stage then selects a subset of candidate sentences under an explicit budget constraint while jointly considering semantic coverage, factual consistency, and redundancy reduction. Finally, the selected candidates are ordered to produce the final summary.

Our key observation is that salient and factually grounded information tends to recur across multiple decoding trajectories with higher probability, whereas hallucinated or weakly supported content appears less consistently. By explicitly decoupling *generation* from *selection*, the proposed framework enables global subset-level reasoning over candidate content instead of relying solely on local autoregressive decoding decisions.

To operationalize this framework, we introduce a unified optimization objective that combines coverage, factuality, and redundancy within a diversity-aware combinatorial selection framework. We further adopt a determinantal point process (DPP)-inspired optimization strategy to favor informative and non-redundant candidate subsets under explicit budget constraints.

Compared with conventional summarization systems, the proposed framework offers several advantages: (i) explicit and controllable summary budgeting independent of the generation model, (ii) improved factual consistency through selection-level grounding, and (iii) modular integration with pretrained summarization models and instruction-tuned LLMs without retraining.

We evaluate the proposed framework on CNN/DailyMail (Nallapati et al., 2016) and Multi-News (Fabbri et al., 2019). Experimental results demonstrate improved factual consistency and redundancy control while maintaining competitive summarization quality across multiple pretrained generation backbones.

Our contributions are summarized as follows:

- We propose a budget-constrained combinatorial framework that decouples summary generation from global content selection for faithful abstractive summarization.
- We introduce a unified optimization objective that jointly models semantic coverage, factual consistency, and redundancy for sentence-level candidate selection.
- We develop a DPP-inspired diversity-aware subset selection strategy over language-model-generated candidate pools.
- We demonstrate the effectiveness and modularity of the proposed framework across both single-document and multi-document summarization benchmarks.

2 Related Work

2.1 Abstractive Summarization with Language Models

Recent abstractive summarization research has increasingly shifted from task-specific encoder–decoder models to large language models (LLMs) and instruction-tuned generation systems(Wang and Wu, 2024). Prior work has benchmarked LLMs for news summarization, analyzed their zero-shot summarization behavior, and studied how they use long input contexts (Zhang et al., 2024; Ravaut et al., 2024; Chhabra et al., 2024). Other studies examine LLMs as teachers or references for training smaller summarization models, as well as their ability to follow user-specified constraints such as length, topic, or style (Liu et al., 2024b,a; Ryu et al., 2026). Multi-document and long-document summarization remain especially challenging because models may fail to synthesize information across sources or may be sensitive to input ordering (DeYoung et al., 2024; Wang et al., 2024; Huang et al., 2024; Kirstein et al., 2024; Sonowal and Sadhu, 2025).

Our work differs from these generation-centered approaches by treating the language model as a candidate generator rather than the final decision maker. Instead of relying on a single decoded summary, we generate multiple candidate summaries, decompose them into sentence-level candidates, and apply a structured optimization layer to select a budget-constrained and faithful subset. This design is related to recent controllable and modular summarization studies, but our focus is on explicit

post-generation content selection under global constraints (Jiang and Dreyer, 2024; Le and Luu, 2024; Guan and Padmakumar, 2023).

2.2 Faithfulness and Hallucination Reduction

Faithfulness remains a central challenge in abstractive summarization. Earlier work showed that neural abstractive systems frequently generate unsupported content, motivating later studies on factual consistency evaluation and hallucination reduction (Maynez et al., 2020; Wang et al., 2020; Fabbri et al., 2022). Recent work has proposed stronger automatic evaluators and benchmarks, including alignment-based metrics, efficient grounded fact-checkers, and challenging hallucination benchmarks for summarization (Zha et al., 2023; Tang et al., 2024a; Tam et al., 2023; Bao et al., 2025). Other studies analyze positional bias and context-use failures in long-form summarization, showing that LLMs may summarize the beginning and end of documents more faithfully than middle content (Ravaut et al., 2024; Wan et al., 2025b).

Beyond evaluation, recent methods attempt to reduce hallucination through decoding, self-consistency, active learning, refinement, or factuality-aware reranking (Li et al., 2024; Huang et al., 2025; Chae et al., 2024; Xia et al., 2024; Wan et al., 2025a). These methods typically modify generation, decoding, training, or post-hoc verification. Our method is complementary: we do not retrain the generator or ask the language model to rewrite the selected content. Instead, we use factuality as an explicit sentence-level scoring signal inside a budget-constrained selection problem, allowing hallucination reduction to be handled at the decision stage.

2.3 Combinatorial Optimization for Summarization

Combinatorial optimization has a long history in summarization, particularly for extractive and multi-document settings. Classical approaches formulate summarization as sentence selection under coverage, diversity, and budget constraints, using methods such as maximal marginal relevance, integer linear programming, submodular maximization, and determinantal point processes (Carbonell and Goldstein, 1998; Gillick and Favre, 2009; Lin and Bilmes, 2011; Kulesza and Taskar, 2012). These methods provide interpretable mechanisms for controlling redundancy and budget, but they are often applied to source sentences rather than language-model-generated abstractive candidates.

Recent work has renewed interest in modular generation, reranking, and candidate selection in the LLM era. Reranking methods select among generated candidates, while modular approaches decompose summarization into extraction, selection, and rewriting or refinement stages (Jain et al., 2024; Ri et al., 2025; Guan and Padmakumar, 2023; Le and Luu, 2024). Submodular and optimization-based selection has also appeared in related NLP settings such as in-context example selection, instruction tuning data mixture, and data-efficient learning (Ji et al., 2024; Kumari et al., 2024; Renduchintala et al., 2024). Our work builds on this line but differs by applying a diversity-aware combinatorial selection layer directly over LLM-generated sentence candidates, jointly incorporating coverage, factuality, redundancy, and an explicit budget constraint.

3 Method

3.1 Problem Definition and Framework Overview

Given an input document or document set $\mathcal{D}$, abstractive summarization aims to produce a concise and faithful textual summary that captures the salient information of $\mathcal{D}$. Most existing approaches formulate this task as conditional sequence generation, which achieves strong fluency but provides limited control over output length and factual consistency.

We instead cast abstractive summarization as a *budget-constrained combinatorial selection problem* over a discrete candidate space. Specifically, we construct a candidate pool of sentences $\mathcal{C} = \{c_1, \ldots, c_n\}$ using a pretrained language model. The objective is to select a subset $A \subseteq \mathcal{C}$ that maximizes summary quality under an explicit budget constraint:

$$A^* = \arg\max_{A \subseteq \mathcal{C}} f(A) \quad \text{s.t.} \quad |A| \leq B, \qquad (1)$$

where l_i denotes the cost (e.g., length) of candidate c_i, and B is the summary budget.

We design $f(A)$ to capture three key properties of high-quality summaries:

- **Coverage:** promotes relevance and semantic coverage of the source document;
- **Factuality:** encourages consistency between candidate sentences and $\mathcal{D}$;

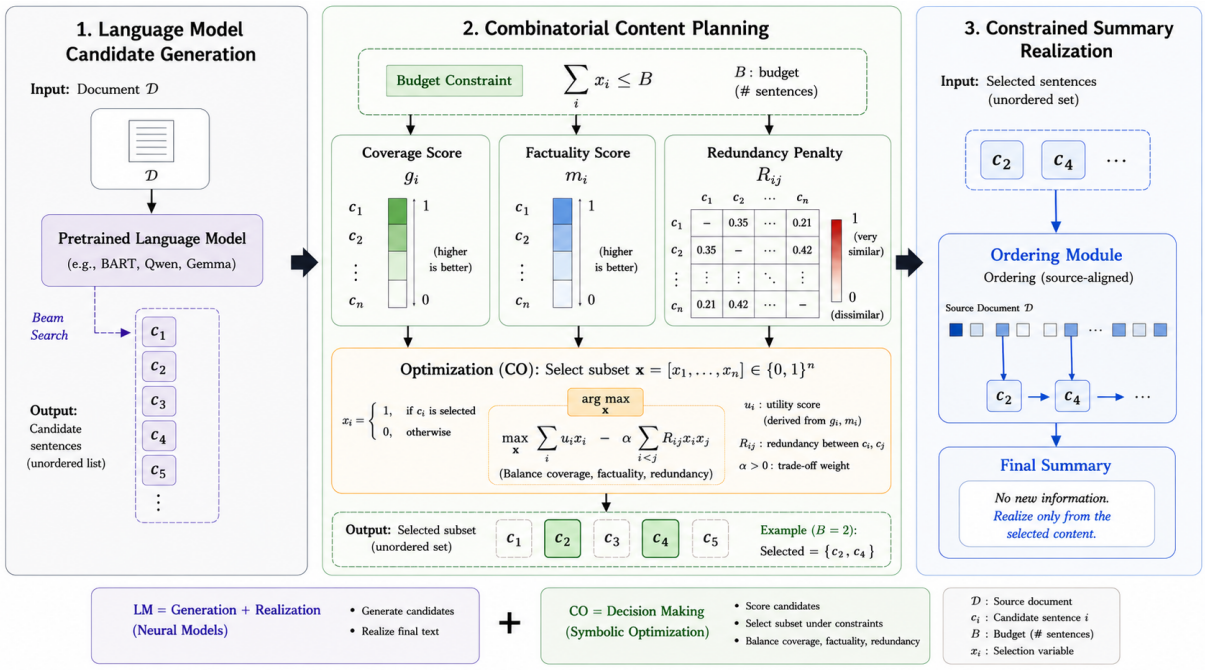

Figure 2: Overview of the proposed framework. A language model first generates a pool of candidate sentences. A combinatorial optimizer then selects a subset under a budget constraint by jointly considering coverage, factuality, and redundancy. Finally, the selected content is ordered to produce the final summary.

- **Redundancy:** penalizes semantic overlap among selected candidates to encourage diversity.

These components are integrated within a unified combinatorial objective to guide subset selection under the budget constraint.

Crucially, we design $f(A)$ using coverage and diversity principles commonly used in submodular optimization, which naturally capture diminishing returns and provide a principled foundation for modeling semantic coverage and redundancy reduction. In particular, we combine (i) a coverage-based component defined over semantic units or document representations, (ii) a modular factuality score, and (iii) a diversity-promoting similarity objective. This formulation enables diversity-aware combinatorial optimization methods such as DPP while retaining the desirable structural properties of submodular-style objectives.

Figure 2 illustrates the overall framework, which consists of three stages:

(1) *candidate generation*, where a language model produces a diverse set of candidate sentences;

(2) *combinatorial content planning*, where an optimizer selects a subset of candidates by jointly considering utility, factuality, and redundancy under a budget constraint; and

(3) *summary realization*, where the selected candidates are ordered to produce the final summary.

This formulation offers three key advantages. First, it provides *explicit and guaranteed control over summary length* through the budget constraint, independent of the generation model. Second, it improves *faithfulness* by restricting the final output to content that is explicitly evaluated against the source document. Third, it enables a *modular integration* of neural generation and combinatorial optimization, allowing pretrained language models to be used without retraining while enforcing global structural constraints explicitly at the selection stage.

3.2 Candidate Generation

The first stage of our framework converts the continuous generation space of a pretrained language model into a finite set of candidate sentences. Rather than relying on a single generated summary, we exploit the stochastic nature of neural text generation to construct a diverse candidate pool. Intuitively, when a language model generates multiple summaries for the same input, salient and well-supported information tends to appear repeatedly across different decoding trajectories with higher probability, while hallucinated or weakly grounded content is less likely to remain consistent.

Formally, given an input document or document set $\mathcal{D}$, a pretrained language model G generates a set of candidate summaries: $Y = \{y_1, y_2, \ldots, y_p\}$, where each y_p corresponds to a distinct decoding trajectory produced via beam search or stochastic decoding.

Each generated summary is segmented into individual sentences, and the resulting sentences are aggregated into a candidate pool: $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$.

Exact duplicate sentences are removed during preprocessing to reduce trivial redundancy. We additionally preserve the provenance of each candidate sentence, including its originating decoding path and sentence position within the generated summary, although optimization is performed over the deduplicated candidate pool.

Unlike conventional reranking approaches that select a single summary hypothesis from beam outputs, our framework decomposes generated summaries into reusable sentence-level content units. This enables the downstream optimization stage to recombine complementary information across multiple decoding trajectories under explicit budget constraints.

Importantly, the candidate generation stage is fully modular and independent of the downstream optimization procedure. Consequently, the proposed framework can be integrated with different pretrained generation models, including encoder-decoder summarization models and instruction-tuned large language models, without requiring retraining or architectural modifications.

3.3 Scoring Functions

After constructing the candidate pool $\mathcal{C} = \{c_1, \ldots, c_n\}$, we compute three complementary scoring components corresponding to utility, factuality, and redundancy. These scoring functions provide both sentence-level quality estimates and pairwise diversity information for the downstream optimization stage.

For each candidate sentence c_i, we compute: (i) a coverage score measuring relevance to the source document $\mathcal{D}$, and (ii) a factuality score estimating consistency with the source input. In addition, for each candidate pair (c_i, c_j), we compute a redundancy score estimating semantic overlap between the two candidates.

Coverage The coverage component measures the relevance and semantic coverage of a candidate sentence with respect to the source document. Formally, we define a coverage function:

$$g_i = \phi(\mathcal{D}, c_i), \tag{2}$$

where $\phi(\cdot)$ estimates how well candidate sentence c_i captures salient information from the source document $\mathcal{D}$.

The framework is agnostic to the specific implementation of $\phi(\cdot)$ and can incorporate lexical overlap measures, semantic similarity functions, embedding-based relevance scores, or learned relevance estimators.

Factuality To encourage factual consistency, we compute a sentence-level factuality score:

$$m_i = P_{\text{fact}}(\text{consistent} \mid \mathcal{D}, c_i), \tag{3}$$

where P_{fact} denotes a factuality estimation model that evaluates whether candidate sentence c_i is supported by the source document $\mathcal{D}$.

Candidates that are weakly grounded or hallucinated are therefore assigned lower factuality scores during optimization.

Redundancy To discourage repetitive information in the final summary, we compute a pairwise redundancy matrix: $R \in \mathbb{R}^{n \times n}$, where each entry estimates semantic overlap between two candidate sentences:

$$R_{ij} = \psi(c_i, c_j). \tag{4}$$

Here, $\psi(\cdot, \cdot)$ denotes a similarity function measuring lexical or semantic overlap between candidate sentences.

Unified Sentence Utility Coverage and factuality are combined into a unified sentence-level utility score:

$$u_i = \lambda_{\text{cov}} g_i + \lambda_{\text{fact}} m_i, \tag{5}$$

where λ_{cov} and λ_{fact} control the relative importance of semantic coverage and factual consistency.

The resulting utility score reflects both the informativeness and faithfulness of a candidate sentence.

3.4 Budget-Constrained Optimization

Given the candidate pool $\mathcal{C} = \{c_1, \ldots, c_n\}$ and the scoring functions defined in Section 3.3, the goal of the optimization stage is to select a subset of candidate sentences that maximizes informativeness and factual consistency while minimizing redundancy under an explicit budget constraint.

Let

$$\mathbf{z} = [z_1, \ldots, z_n] \in \{0,1\}^n$$

denote the binary selection vector, where

$$z_i = \begin{cases} 1, & \text{if candidate } c_i \text{ is selected,} \\ 0, & \text{otherwise.} \end{cases}$$

We define the budget-constrained subset selection objective as:

$$f(A) = \max_{\mathbf{z}} \sum_{i=1}^{n} u_i z_i - \alpha \sum_{1 \le i < j \le n} R_{ij} z_i z_j, \quad (6)$$

subject to

$$|A| \le B. \quad (7)$$

The first term rewards candidates that are informative and factually consistent with the source document, while the second term penalizes semantic overlap among selected candidates. The trade-off parameter $\alpha > 0$ controls the strength of redundancy suppression during subset selection.

The proposed objective is motivated by coverage and diversity principles commonly used in submodular optimization. However, directly optimizing pairwise redundancy interactions can become computationally challenging for large candidate pools. Instead of directly optimizing Eq. 6, we use a DPP-style diversity objective as a tractable approximation for subset selection.

Determinantal Point Processes We construct a DPP-inspired quality-diversity kernel:

$$L = QKQ + \epsilon I, \quad (8)$$

where $Q = \mathrm{diag}(q_1, \ldots, q_n)$ contains sentence-level quality scores, $K \in \mathbb{R}^{n \times n}$ is a pairwise similarity matrix derived from redundancy estimates, I is the identity matrix, and ϵ is a small positive constant for numerical stability.

The quality score for each candidate is derived from its unified sentence utility:

$$q_i = \max(u_i, \epsilon_q), \quad (9)$$

where $\epsilon_q > 0$ prevents zero-quality candidates from producing degenerate kernel entries. In our implementation, ϵ_q is set to a small constant.

The similarity matrix K is obtained by scaling the pairwise redundancy matrix and clipping its entries to the interval $[0, 1]$. Thus, K_{ij} reflects the semantic similarity between candidate sentences c_i and c_j. The quality matrix Q encourages the selection of informative and factually grounded candidates, while the determinant-based objective discourages selecting near-duplicate candidates by favoring subsets with diverse similarity structure.

The selected subset is obtained by greedy MAP-style inference:

$$A^* = \arg\max_{A \subseteq \mathcal{C}} \log\det(L_A) \quad \text{s.t.} \quad |A| \le B, \quad (10)$$

where L_A denotes the principal submatrix indexed by candidates in A.

Since the similarity matrix is constructed from heuristic lexical or semantic overlap estimates, it does not necessarily guarantee a strictly valid positive semi-definite DPP kernel. We therefore use the selector as a deterministic DPP-inspired diversity objective rather than as an exact probabilistic DPP model. In practice, candidates are added greedily by selecting the sentence that maximizes the log-determinant score of the current subset. If determinant computation fails due to numerical issues, the selector falls back to the highest-quality remaining candidates.

The optimization stage outputs a selected subset $A^* \subseteq \mathcal{C}$, which is subsequently passed to the summary realization stage.

3.5 Summary Realization

After optimization, the selected subset $A^* = \{c_{i_1}, \ldots, c_{i_k}\}$ is converted into the final summary through an ordering-based realization step.

Since optimization operates on unordered sentence-level candidates, we apply a simple *source-similarity ordering* strategy. For each selected sentence c_i, we identify the most semantically related source sentence and use its source position $p(c_i)$ as an ordering signal.

The selected candidates are then sorted as:

$$A^*_{\text{ordered}} = \mathrm{sort}\big(A^*; p(c_i)\big). \quad (11)$$

Finally, the ordered sentences are concatenated to produce the summary:

$$\hat{y} = [c_{i_1}; c_{i_2}; \ldots; c_{i_k}]. \quad (12)$$

This realization strategy preserves the optimized subset directly without requiring additional rewriting, thereby maintaining both the budget constraint and the factual content selected during optimization. Although summaries assembled from multiple decoding trajectories may still exhibit minor discourse inconsistencies, the approach provides a

Dataset	Method	ROUGE (↑)				BERTScore (↑)	Faithfulness (↑)			
		R-1	R-2	R-L	R-Lsum	F1	FactCC	MiniCheck	AlignScore	FactKB
CNN/DM	BART	44.05	21.07	30.62	40.99	88.21	75.83	94.90	91.56	98.49
	FactEdit	42.53	20.48	39.74	40.76	88.17	76.03	94.82	91.46	98.68
	SimCLS	46.67	22.15	43.54	43.09	66.14	59.49	83.34	82.37	80.75
	BRIO-Mul	47.78	23.55	44.57	44.12	87.84	54.06	84.67	74.48	85.20
	BRIO-Ctr	47.28	22.93	44.15	44.10	89.11	62.62	88.42	84.98	78.89
	EFactSum	44.37	21.28	40.92	41.07	88.36	60.74	88.33	82.02	91.38
	SummaReranker	46.62	22.39	43.59	43.80	88.47	68.17	93.81	88.27	98.10
	Llama-3-8B	38.98	14.82	24.20	35.28	87.72	46.78	76.00	75.08	98.51
	Qwen3.5-9B	34.88	10.91	20.33	30.77	86.96	40.16	56.82	65.17	98.12
	Gemma-4-E4B-it	30.51	6.41	18.07	27.40	86.19	53.99	69.49	71.32	89.59
	BART+DPP	41.67	19.36	27.96	38.85	87.72	**76.92**	**96.93**	**93.14**	**98.83**
	Llama-3-8B+DPP	26.89	5.27	16.54	23.48	85.21	52.27	85.78	78.61	98.20
Multi-News	PRIMERA	41.10	15.88	22.17	37.30	86.89	37.40	63.06	53.30	94.36
	SimCLS	45.78	17.31	23.14	41.59	85.68	32.47	54.35	48.19	57.35
	BRIO-Ctr	46.04	16.90	22.73	41.78	85.61	30.39	51.50	46.01	53.81
	SummaReranker	44.56	16.72	22.61	40.45	87.14	35.18	61.49	51.88	94.72
	Llama-3-8B	38.84	8.68	18.16	34.81	85.42	38.35	60.46	59.04	90.99
	Qwen3.5-9B	38.31	6.74	15.96	34.05	84.82	42.01	48.40	53.08	86.98
	Gemma-4-E4B-it	35.44	6.53	16.27	31.86	84.89	48.49	66.79	66.35	87.68
	PRIMERA+DPP	34.31	10.99	18.90	31.05	84.02	40.10	77.88	65.98	89.20
	Llama-3-8B+DPP	34.18	6.11	16.27	30.11	84.05	**52.69**	**79.22**	**70.07**	**98.02**

Table 1: Main results on CNN/DailyMail and Multi-News. The proposed DPP-based selection framework consistently improves factuality and source grounding across multiple evaluation metrics while maintaining competitive summarization quality. Best results within each dataset block are shown in bold.

lightweight and controllable mechanism for converting optimized sentence subsets into coherent summaries.

4 Results

Detailed experimental settings, datasets, implementation details, and evaluation configurations are provided in Appendix A.

4.1 Main Results

Table 1 presents the main experimental results on the CNN/DailyMail and Multi-News benchmarks. We compare the proposed DPP-based combinatorial selection framework against pretrained summarization systems and recent faithfulness-oriented baselines.

Overall, the proposed framework consistently improves factual grounding across multiple evaluation metrics while maintaining competitive summarization quality. Unlike conventional reranking approaches that select a single beam hypothesis, our framework decomposes generated summaries into reusable sentence-level units and recombines them through budget-constrained subset selection.

CNN/DailyMail On CNN/DailyMail, `BART+DPP` achieves the strongest overall factuality performance among all compared systems, obtaining the best FactCC (76.92), MiniCheck (96.93), AlignScore (93.14), and FactKB (98.83) scores. Compared with direct `BART` generation, DPP-based selection improves MiniCheck from 94.90 to 96.93 and AlignScore from 91.56 to 93.14, indicating improved source grounding and entailment consistency.

Although DPP-based selection slightly reduces ROUGE scores relative to generation-oriented systems such as BRIO and SimCLS, the resulting summaries are substantially more reliable and better supported by the source document. Similar trends are observed for instruction-tuned language models. Compared with direct `Llama-3-8B` generation, `Llama-3-8B+DPP` improves MiniCheck from 76.00 to 85.78 and AlignScore from 75.08 to 78.61 without requiring additional training.

Multi-News On the more challenging Multi-News benchmark, the proposed framework produces larger factuality improvements. In particular, `Llama-3-8B+DPP` achieves the best FactCC (52.69), MiniCheck (79.22), AlignScore (70.07), and FactKB (98.02) scores among all evaluated systems.

Compared with direct `Llama-3-8B` generation,

Dataset	Method	ROUGE (↑)				BERTScore (↑)	Faithfulness (↑)			
		R-1	R-2	R-L	R-Lsum	F1	FactCC	MiniCheck	AlignScore	FactKB
CNN/DM	BART+MMR	41.85	19.44	28.03	38.99	87.75	76.55	96.22	92.42	98.67
	BART+ILP	42.84	19.91	28.46	39.87	87.91	76.71	96.76	92.56	98.65
	BART+DPP	41.67	19.36	27.96	38.85	87.72	76.92	96.93	93.14	98.83
	Llama-3-8B+MMR	29.18	6.14	17.33	25.53	85.68	49.94	83.69	77.52	97.98
	Llama-3-8B+ILP	27.96	5.75	16.95	24.35	85.39	50.06	84.61	77.98	98.30
	Llama-3-8B+DPP	26.89	5.27	16.54	23.48	85.21	52.27	85.78	78.61	98.20
Multi-News	PRIMERA+MMR	41.18	14.86	20.98	37.28	85.61	33.82	62.74	54.12	87.72
	PRIMERA+ILP	39.58	14.04	20.55	35.63	85.39	36.77	70.95	59.45	88.94
	PRIMERA+DPP	34.31	10.99	18.90	31.05	84.02	40.10	77.88	65.98	89.20
	Llama-3-8B+MMR	39.36	7.97	17.15	35.13	85.17	48.40	65.56	63.83	95.40
	Llama-3-8B+ILP	38.28	7.63	17.00	33.97	85.03	49.43	75.58	68.62	96.21
	Llama-3-8B+DPP	34.18	6.11	16.27	30.11	84.05	52.69	79.22	70.07	98.02

Table 2: Ablation study comparing different combinatorial optimization strategies under the same candidate generation framework. DPP consistently achieves the strongest factuality performance across both datasets and model families, while MMR and ILP generally preserve higher ROUGE scores.

DPP-based selection improves MiniCheck from 60.46 to 79.22 and FactKB from 90.99 to 98.02. These gains suggest that explicit combinatorial content planning becomes increasingly beneficial in long-context and multi-document summarization settings where redundancy and hallucination are more difficult to control.

For encoder-decoder models, `PRIMERA+DPP` similarly improves factual consistency relative to direct `PRIMERA` generation, although with a larger reduction in ROUGE scores. Overall, the results demonstrate that decoupling generation and selection provides a robust mechanism for improving factual grounding across both encoder-decoder models and instruction-tuned language models.

To further illustrate how combinatorial sentence selection improves factual grounding and content planning, we provide a qualitative case study in Appendix B.

4.2 Ablation Study

Table 2 compares different combinatorial optimization strategies under the same candidate generation framework, including Maximal Marginal Relevance (MMR), Integer Linear Programming (ILP), and the proposed DPP-based selector.

Overall, MMR tends to preserve higher ROUGE scores by favoring locally relevant candidates, while ILP provides a stronger global balance between utility and redundancy. In contrast, the proposed DPP-based selector consistently achieves the strongest factuality performance across both datasets and model families.

On CNN/DailyMail, `BART+DPP` achieves the best FactCC (76.92), MiniCheck (96.93), AlignScore (93.14), and FactKB (98.83) scores among all selectors. Similarly, `Llama-3-8B+DPP` achieves the highest factuality scores among the LLM-based selectors, indicating that determinant-based diversity selection more effectively suppresses unsupported or weakly grounded content.

The differences between optimization strategies become more pronounced on Multi-News. For both `PRIMERA` and `Llama-3-8B`, DPP substantially improves factual consistency compared with MMR and ILP. In particular, `Llama-3-8B+DPP` achieves the strongest FactCC (52.69), MiniCheck (79.22), AlignScore (70.07), and FactKB (98.02) scores overall.

These results indicate that diversity-aware subset selection is particularly important in long-context and multi-document summarization settings, where candidate redundancy and hallucination become more difficult to control.

5 Conclusion

We presented a combinatorial content planning framework for budget-constrained and faithful abstractive summarization. By decoupling generation and selection, the proposed approach jointly optimizes semantic coverage, factual consistency, and redundancy reduction under an explicit budget constraint. Experiments on CNN/DailyMail and Multi-News demonstrate improved factual consistency while maintaining competitive summarization quality across multiple pretrained summarization models and instruction-tuned language models.

6 Limitations

Despite achieving improvements in factual consistency and controllability, the proposed framework has several limitations.

First, the quality of the final summary remains fundamentally dependent on the quality of the generated candidate pool. Since the optimization stage only selects from generated candidates, factual information omitted by the underlying language model cannot be recovered during selection. Consequently, severe generation failures or systematic hallucinations in the candidate pool may limit downstream optimization effectiveness.

Second, the current framework operates at the sentence level and does not explicitly model discourse coherence or cross-sentence dependencies. Although the proposed realization strategy preserves a stable document-level ordering, summaries assembled from multiple decoding trajectories may still contain stylistic inconsistencies or minor discourse discontinuities. Moreover, although the proposed framework improves factual consistency, automatic factuality metrics remain imperfect and do not guarantee fully error-free summaries in high-stakes domains.

Third, the proposed DPP-based optimization uses heuristic similarity functions and a greedy MAP approximation rather than an exact probabilistic DPP formulation. As a result, the optimization objective provides diversity-oriented subset selection but does not guarantee globally optimal combinatorial solutions. In addition, the current implementation enforces a sentence-level budget constraint rather than a finer-grained token-level constraint.

Fourth, the framework introduces additional inference cost compared with direct generation baselines because multiple candidate summaries must be generated before optimization. The computational overhead increases with larger candidate pools and stronger factuality evaluation models.

Finally, the current study focuses primarily on news summarization benchmarks. Although the framework is model-agnostic, its effectiveness on other summarization settings, such as long-form scientific summarization, dialogue summarization, or highly abstractive generation tasks, remains to be explored in future work.

References

AI@Meta. 2024. Llama 3 model card.

Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. 2022. Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9818–9830, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Forrest Sheng Bao, Miaoran Li, Renyi Qu, Ge Luo, Erana Wan, Yujia Tang, Weisi Fan, Manveer Singh Tamber, Suleman Kazi, Vivek Sourabh, Mike Qi, Ruixuan Tu, Chenyu Xu, Matthew Gonzales, Ofer Mendelevitch, and Amin Ahmad. 2025. FaithBench: A diverse hallucination benchmark for summarization by Modern LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 448–461, Albuquerque, New Mexico. Association for Computational Linguistics.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

Kyubyung Chae, Jaepill Choi, Yohan Jo, and Taesup Kim. 2024. Mitigating hallucination in abstractive summarization with domain-conditional mutual information. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1809–1820, Mexico City, Mexico. Association for Computational Linguistics.

Anshuman Chhabra, Hadi Askari, and Prasant Mohapatra. 2024. Revisiting zero-shot abstractive summarization in the era of large language models from the perspective of position bias. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 1–11, Mexico City, Mexico. Association for Computational Linguistics.

Jay DeYoung, Stephanie C. Martinez, Iain J. Marshall, and Byron C. Wallace. 2024. Do multi-document summarization models synthesize? *Transactions of the Association for Computational Linguistics*, 12:1043–1062.

Tanay Dixit, Fei Wang, and Muhao Chen. 2023. Improving factuality of abstractive summarization without sacrificing summary quality. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 902–913, Toronto, Canada. Association for Computational Linguistics.

Gemma Team et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. FactKB: Generalizable factuality evaluation using language models enhanced with factual knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 933–952, Singapore. Association for Computational Linguistics.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 10–18, Boulder, Colorado. Association for Computational Linguistics.

Google DeepMind. 2026. Gemma 4 e4b it. `https://huggingface.co/google/gemma-4-E4B-it`. Accessed: 2026-05-01.

Shuo Guan and Vishakh Padmakumar. 2023. Extract, select and rewrite: A modular sentence summarization method. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 41–48, Singapore. Association for Computational Linguistics.

Kung-Hsiang Huang, Philippe Laban, Alexander Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2024. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 570–593, Mexico City, Mexico. Association for Computational Linguistics.

Yanwen Huang, Yong Zhang, Ning Cheng, Zhitao Li, Shaojun Wang, and Jing Xiao. 2025. Dynamic attention-guided context decoding for mitigating context faithfulness hallucinations in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5174–5193, Vienna, Austria. Association for Computational Linguistics.

Siddhartha Jain, Xiaofei Ma, Anoop Deoras, and Bing Xiang. 2024. Lightweight reranking for language model generations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6960–6984, Bangkok, Thailand. Association for Computational Linguistics.

Baijun Ji, Xiangyu Duan, Zhenyu Qiu, Tong Zhang, Junhui Li, Hao Yang, and Min Zhang. 2024. Submodular-based in-context example selection for LLMs-based machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15398–15409, Torino, Italia. ELRA and ICCL.

Xiang Jiang and Markus Dreyer. 2024. CCSum: A large-scale and high-quality dataset for abstractive news summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7306–7336, Mexico City, Mexico. Association for Computational Linguistics.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Fables: Evaluating faithfulness and content selection in book-length summarization. *arXiv preprint arXiv:2404.01261*.

Frederic Kirstein, Terry Ruas, Robert Kratel, and Bela Gipp. 2024. Tell me what I need to know: Exploring LLM-based (personalized) abstractive multi-source meeting summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 920–939, Miami, Florida, US. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2-3):123–286.

Lilly Kumari, Shengjie Wang, Arnav Das, Tianyi Zhou, and Jeff Bilmes. 2024. An end-to-end submodular framework for data-efficient in-context learning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3293–3308, Mexico City, Mexico. Association for Computational Linguistics.

Thang Le and Anh Tuan Luu. 2024. Extractive summarization with text generator. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies (Volume 1: Long Papers), pages 157–174, Mexico City, Mexico. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Taiji Li, Zhi Li, and Yin Zhang. 2024. Improving faithfulness of large language models in summarization via sliding generation and self-consistency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8804–8817, Torino, Italia. ELRA and ICCL.

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.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520, Portland, Oregon, USA. Association for Computational Linguistics.

Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024a. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4481–4501, Mexico City, Mexico. Association for Computational Linguistics.

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

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024b. On learning to summarize with large language models as references. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8647–8664, Mexico City, Mexico. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022. SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524, Dublin, Ireland. Association for Computational Linguistics.

Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2764–2781, Bangkok, Thailand. Association for Computational Linguistics.

H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Ganesh Ramakrishnan. 2024. SMART: Submodular data mixture strategy for instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12916–12934, Bangkok, Thailand. Association for Computational Linguistics.

Narutatsu Ri, Nicholas Deas, and Kathleen McKeown. 2025. Reranking-based generation for unbiased perspective summarization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24701–24723, Vienna, Austria. Association for Computational Linguistics.

Sangwon Ryu, Heejin Do, Daehui Kim, Hwanjo Yu, Dongwoo Kim, Yunsu Kim, Gary Lee, and Jungseul Ok. 2026. Exploring iterative controllable summarization with large language models. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 512–528, Rabat, Morocco. Association for Computational Linguistics.

Himadri Sonowal and Saisab Sadhu. 2025. Structure-aware chunking for abstractive summarization of long legal documents. In *Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025)*, pages 171–178, Mumbai, India. Association for Computational Linguistics.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

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

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024b. TofuEval: Evaluating hallucinations of LLMs on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5-omni technical report. *Preprint*, arXiv:2604.15804.

David Wan, Justin Chen, Elias Stengel-Eskin, and Mohit Bansal. 2025a. MAMM-refine: A recipe for improving faithfulness in generation with multi-agent collaboration. 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 9882–9901, Albuquerque, New Mexico. Association for Computational Linguistics.

David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2025b. 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.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Guanghua Wang, Priyanshi Garg, and Weili Wu. 2024. Segmented summarization and refinement: A pipeline for long-document analysis on social media. *Journal of social computing*, 5(2):132–144.

Guanghua Wang and Weili Wu. 2024. Surveying the landscape of text summarization with deep learning: A comprehensive review. *Discrete Mathematics, Algorithms and Applications*, 16(03):2330004.

Yu Xia, Xu Liu, Tong Yu, Sungchul Kim, Ryan Rossi, Anup Rao, Tung Mai, and Shuai Li. 2024. Hallucination diversity-aware active learning for text summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8665–8677, Mexico City, Mexico. Association for Computational Linguistics.

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

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

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

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

A Experimental Setup

A.1 Datasets

We evaluate the proposed framework on both single-document and multi-document abstractive summarization benchmarks.

CNN/DailyMail We conduct experiments on the CNN/DailyMail dataset (Nallapati et al., 2016), a widely used benchmark for single-document news summarization. The dataset consists of news articles paired with multi-sentence abstractive summaries written by human annotators. Following standard preprocessing protocols, non-anonymized versions of the articles are used for training and evaluation.

Multi-News To evaluate the framework in long-context and multi-document settings, we additionally conduct experiments on the Multi-News dataset (Fabbri et al., 2019). Multi-News contains clusters of related news articles paired with professionally written summaries that synthesize information across multiple documents. Compared

with CNN/DailyMail, this dataset presents substantially greater challenges in terms of redundancy reduction, cross-document content aggregation, and long-context faithfulness.

These two datasets provide complementary evaluation settings for studying the effectiveness of budget-constrained combinatorial content planning under both standard and long-context summarization scenarios.

A.2 Baselines

We compare the proposed framework against representative pretrained summarization models, factuality-aware generation systems, reranking approaches, and recent instruction-tuned large language models.

Pretrained Summarization Models We include **BART** (Lewis et al., 2020), a widely adopted encoder–decoder summarization model, and **PRIMERA** (Xiao et al., 2022), a long-context summarization model designed for multi-document summarization.

Factuality-Aware Summarization Methods To evaluate faithfulness-oriented generation approaches, we compare against **FactEdit** (Balachandran et al., 2022), which performs post-editing for factual correction, **BRIO-Mul** and **BRIO-Ctr** (Liu et al., 2022), which optimize generation quality through contrastive ranking objectives, and **EFactSum** (Dixit et al., 2023), which incorporates factuality-aware optimization into abstractive summarization.

Reranking and Candidate Selection Methods We additionally compare against **SimCLS** (Liu and Liu, 2021), a contrastive learning-based reranking framework, and **SummaReranker** (Ravaut et al., 2022), which improves summarization quality through summary-level reranking.

Instruction-Tuned Large Language Models To evaluate the generality of the proposed framework across generation backbones, we further experiment with recent instruction-tuned large language models, including **Qwen3.5-9B** (Team, 2026), **Llama-3-8B** (AI@Meta, 2024), and **Gemma-4-E4B-it** (et al., 2024; Google DeepMind, 2026).

A.3 Implementation Details

All experiments are implemented using `PyTorch` and the `Hugging Face Transformers` library. For each input document, the generation model first produces multiple candidate summaries, which are subsequently decomposed into sentence-level candidates to construct the optimization pool $\mathcal{C}$. Unless otherwise specified, all optimization-based methods operate on the same cached candidate pool to ensure fair comparison across different subset selection strategies.

Generation Models For CNN/DailyMail experiments, we use `facebook/bart-large-cnn` (Lewis et al., 2020) with the checkpoint's official summarization configuration, including beam-search decoding, length normalization, no-repeat trigram blocking, and early stopping. The direct-generation baseline uses the model's standard beam setting, while optimization-based runs use a larger beam size to expose a more diverse candidate pool for downstream selection.

For Multi-News experiments, we use `allenai/PRIMERA-multinews` (Xiao et al., 2022). The direct-generation baseline follows the default decoding configuration provided by the checkpoint, whereas optimization runs generate a larger number of candidate summaries to improve candidate diversity.

For instruction-tuned large language models, we use concise summary-oriented prompts with dataset-specific instructions. The evaluated checkpoints include `Qwen/Qwen3.5-9B` (Team, 2026), `meta-llama/Llama-3.1-8B-Instruct` (AI@Meta, 2024), and `google/gemma-4-E4B-it` (et al., 2024; Google DeepMind, 2026). Direct-generation baselines use sampling-based decoding configurations, while optimization-based variants generate multiple candidate summaries before sentence-level selection.

For Multi-News inputs, source documents are concatenated in their original dataset order using document delimiters. Inputs are truncated only when necessary to satisfy model-specific maximum context lengths.

Candidate Pool Construction Generated summaries are segmented into sentence-level candidates using standard sentence tokenization. Exact duplicate sentences are removed during preprocessing to reduce trivial redundancy. The resulting deduplicated candidate pool is subsequently used by all downstream optimization methods.

Score Normalization Coverage, factuality, and redundancy scores are normalized independently within each candidate pool before optimization.

For sentence-level coverage and factuality scores, we apply min–max normalization.

For redundancy estimation, the same normalization procedure is applied to the off-diagonal entries of the pairwise redundancy matrix, while diagonal entries are retained as self-similarity values. This normalization places all scoring components on comparable scales while preserving relative ranking within each candidate pool.

Optimization The optimization stage uses the unified utility and redundancy objective described in Section 3.4. For DPP-based selection, we construct the quality-diversity kernel from normalized sentence utilities and pairwise redundancy scores, and perform greedy MAP inference under the budget constraint.

A.4 Evaluation Metrics

We evaluate summarization quality using both generation-quality metrics and factual consistency metrics.

Generation Quality Metrics We report **ROUGE** (Lin, 2004), including ROUGE-1, ROUGE-2, and ROUGE-L, to evaluate lexical overlap between generated summaries and reference summaries. We additionally report **BERTScore** (Zhang et al., 2019), which measures semantic similarity using contextualized embeddings.

Faithfulness Metrics To evaluate factual consistency, we report several complementary faithfulness metrics. **FactCC** (Kryscinski et al., 2020) evaluates whether generated summaries are entailed by the source document. **MiniCheck** (Tang et al., 2024a) provides efficient sentence-level factuality estimation for grounded generation. **AlignScore** (Zha et al., 2023) measures semantic alignment between generated summaries and source documents using alignment-based scoring. We additionally report **FactKB** (Feng et al., 2023), a knowledge-grounded factual consistency metric for abstractive generation.

Together, these metrics provide a comprehensive evaluation of summary informativeness, semantic quality, and factual faithfulness.

B Case Study

B.1 Case Studies

To better understand how combinatorial content planning affects the final summary, we present a representative example from the CNN/DailyMail test set using `meta-llama/Llama-3.1-8B-Instruct`. We compare the direct generation baseline with the proposed DPP-based sentence selection framework. Table 3 presents the source article excerpt, the reference summary, both system outputs, and the corresponding quality and factuality metrics.

The example illustrates how post-generation sentence selection can recover source-supported information that is underemphasized or omitted in a single decoding trajectory, while simultaneously improving factual grounding. The direct generation baseline produces a fluent summary and captures the overall legal dispute, but allocates substantial space to compensation details and privacy-rights paraphrases that are not emphasized in the reference summary. In contrast, the DPP-selected summary more closely follows the reference structure by preserving several key factual elements, including the illegal interception, the legal-privilege violation, the destruction order, and the kidnapping-and-torture allegation.

This improvement is reflected in both overlap-based and factuality-oriented evaluation metrics. Compared with the direct generation baseline, the DPP-selected summary improves ROUGE-Lsum from 25.68 to 41.86, indicating substantially stronger alignment with the reference summary. More importantly, factual consistency metrics also improve consistently: AlignScore increases from 77.23 to 86.39, FactKB improves from 89.51 to 99.32, and MiniCheck increases from 95.22 to 95.75.

The example highlights the central intuition behind the proposed framework. Rather than relying on a single autoregressive generation trajectory, the combinatorial selector aggregates complementary factual units across multiple generated candidates under an explicit budget constraint. As a result, the final summary achieves stronger factual grounding and improved content coverage while avoiding redundant or weakly supported information.

Component	Content
Source Article	British intelligence agency GCHQ was ordered to destroy documents obtained from the unlawful interception of communications between Libyan dissident Sami al-Saadi and his lawyers. The Investigatory Powers Tribunal found that the intercepted material involved **legally privileged communications** and ordered the agency to destroy copies of the eavesdropped material. Al-Saadi claimed that the British government was involved in his **kidnapping** and return to Libya, where he said he was **tortured** by the Gaddafi regime.
Reference Summary	Sami al-Saadi's communications with his lawyer were **illegally intercepted**. The Investigatory Powers Tribunal ordered intelligence agencies to **destroy copies** of the intercepted communications. Mr. al-Saadi claimed that the British government was complicit in **kidnapping** and returning him to Libya, where he was allegedly **tortured**.
Direct Baseline	A British watchdog ordered intelligence agencies to destroy illegally obtained information from communications between a Libyan dissident and his lawyers. GCHQ admitted intercepting confidential calls and emails involving Sami al-Saadi, violating legal professional privilege. The tribunal ruled that the dissident should not receive **compensation** for the violation of his privacy rights. GCHQ was instructed to destroy the intercepted documents to protect the dissident's privacy.
DPP Summary	A British watchdog ordered intelligence agencies to destroy illegally obtained information from a dissident's confidential communications with his lawyers. The intercepted material violated rules of **legal privilege**. The tribunal ordered the **destruction of the intercepted communications**. Sami al-Saadi also claimed that the British government was involved in his **kidnapping and torture** by the Libyan regime.

Table 3: Case study from CNN/DailyMail using `Llama-3.1-8B-Instruct`. Compared with direct generation, the proposed DPP-based selection framework better preserves reference-aligned factual structure, including the legal-privilege violation, destruction order, and kidnapping-and-torture allegation. The selected summary also achieves substantially higher factual consistency scores while improving reference alignment.