

Decoupling Generation and Selection for Budget-Constrained Faithful Summarization

Zeyu Wang, Guanghua Wang*, Meng Xu
Kean University, USA

{wangzeyu, guanghua.wang, meng.xu}@kean.edu

Abstract

Abstractive summarization models remain vulnerable to factual inconsistency, redundancy, and weak length control. We propose a modular generation-and-selection framework for sentence-budget-constrained summarization. A pretrained generator produces multiple candidate summaries, which are decomposed into sentence-level candidates. A combinatorial selector then constructs the final summary by balancing relevance, factuality, and redundancy under an explicit budget. The framework supports MMR, ILP, and a DPP-inspired log-determinant objective without retraining the generator. Experiments on CNN/DailyMail, Multi-News, FaithBench, and TofuEval show consistent improvements in factuality and source-grounding metrics, especially for multi-document summarization, at the cost of lower reference-overlap scores. Human evaluation further indicates higher perceived consistency, relevance, clarity, and conciseness, with a small reduction in coherence. These results show that decoupling generation from selection provides a model-agnostic mechanism for improving factual grounding. Code is available at <https://anonymous.4open.science/r/bcfs-D05E/>.

1 Introduction

Pretrained language models and large language models (LLMs) have substantially improved the fluency and semantic quality of abstractive summarization systems (Li et al., 2024; Kim et al., 2024). However, autoregressive generation remains vulnerable to factual inconsistency, redundant content, and weak control over summary length. Because content decisions are made along a single token-level decoding trajectory, generated summaries may contain unsupported claims, omit salient information, or fail to satisfy a specified output budget (Kim et al., 2024; Tang et al., 2024b).

*Corresponding author.

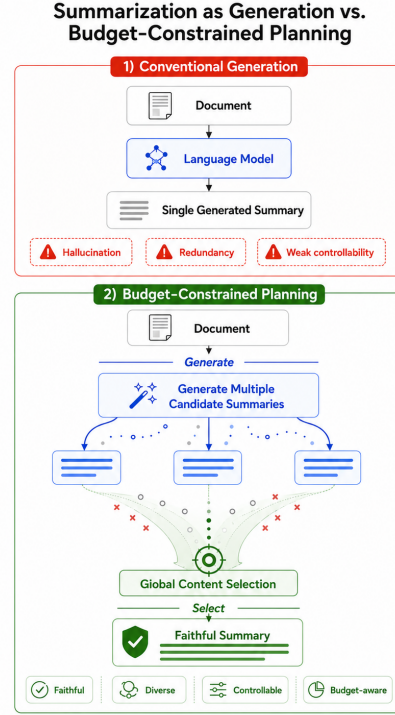


Figure 1: From generation to planning for abstractive summarization. Conventional autoregressive summarization produces a summary from a single decoding trajectory. Our framework decouples generation from selection by generating multiple candidate summaries, decomposing them into sentence-level candidates, and selecting a budget-constrained subset that balances coverage, factuality, and redundancy.

We formulate abstractive summarization as a *generation-and-selection* problem. Rather than treating a single generated sequence as the final output, a pretrained generator produces multiple candidate summaries, which are decomposed into sentence-level candidates. A combinatorial selector then constructs the final summary under an explicit sentence budget by balancing source relevance, factual consistency, and redundancy. The selected sentences are ordered according to their aligned source positions, requiring neither retraining of the

generator nor an additional rewriting stage.

The framework is motivated by the observation that salient and source-supported information often recurs across multiple decoding trajectories, whereas unsupported or weakly grounded content is typically less stable. Sentence-level selection can therefore combine complementary information from different generations while filtering repetitive or low-quality candidates. This design makes factuality, diversity, and budget constraints explicit at selection time rather than leaving them implicit in autoregressive decoding.

We instantiate the framework with a unified candidate-utility function and a DPP-inspired log-determinant selector that promotes high-quality and diverse subsets. The framework is modular with respect to both components: the generator may be an encoder-decoder summarization model or an instruction-tuned LLM, while the selector may be implemented using MMR, integer programming, or the proposed determinant-based objective. Because the similarity matrix is derived from heuristic redundancy scores, we use the log-determinant criterion as a deterministic diversity-aware objective rather than as a probabilistic DPP.

We evaluate the framework on the single-document CNN/DailyMail benchmark (Nallapati et al., 2016), the multi-document Multi-News benchmark (Fabbri et al., 2019), and the more recent FaithBench and TofuEval factuality-oriented benchmarks (Bao et al., 2025; Tang et al., 2024b). Across generators and datasets, global sentence selection consistently improves automatic factuality and source-grounding metrics, with particularly large gains in multi-document summarization. These improvements are accompanied by lower ROUGE and BERTScore, revealing a clear trade-off between source grounding and reference overlap. A blind human evaluation further shows that selected summaries are preferred more frequently and receive higher ratings for factual consistency, relevance, clarity, and conciseness, although direct generation retains a small advantage in coherence.

Our contributions are threefold:

- We propose a modular generation-and-selection framework for sentence-budget-constrained abstractive summarization that operates without retraining the underlying generator.
- We formulate sentence-level content planning as a combinatorial selection problem

that combines relevance, factuality, and redundancy, and instantiate it with a DPP-inspired determinant-based quality-diversity objective.

- We provide extensive automatic, statistical, ablation, and human evaluation across single-document, multi-document, and factuality-oriented benchmarks, demonstrating improved source grounding together with an explicit reference-overlap trade-off.

2 Related Work

Our work connects three lines of research: neural abstractive summarization, faithfulness-oriented generation, and combinatorial sentence selection. Pretrained encoder-decoder models and instruction-tuned LLMs have substantially improved summarization fluency, but remain sensitive to long contexts, input ordering, and user-specified constraints such as length and topic (Zhang et al., 2024; Ravaut et al., 2024; Chhabra et al., 2024; Liu et al., 2024a). Recent modular approaches decompose summarization into extraction, generation, selection, or rewriting stages (Guan and Padmakumar, 2023; Le and Luu, 2024). In contrast, we treat the generator as a source of multiple abstractive hypotheses and perform global sentence-level selection without retraining or an additional rewriting stage.

Faithfulness has been addressed through factuality-aware decoding, self-consistency, post-editing, refinement, and reranking (Li et al., 2024; Chae et al., 2024; Wan et al., 2025a). Related reranking methods such as SimCLS, SummaReranker, and lightweight generation rerankers select one complete generated hypothesis from a candidate set (Liu and Liu, 2021; Ravaut et al., 2022; Jain et al., 2024). Recent factuality evaluators and benchmarks, including AlignScore, MiniCheck, FaithBench, and FaithLens, provide complementary mechanisms for assessing source grounding and identifying unsupported content (Zha et al., 2023; Tang et al., 2024a; Bao et al., 2025; Si et al., 2026). Our method instead decomposes multiple generated summaries into reusable sentence candidates and recombines complementary content under explicit factuality, redundancy, and budget constraints. This enables factuality to influence the final content plan directly rather than serving only as a summary-level reranking or post-hoc correction signal.

Combinatorial optimization has long been used in extractive and multi-document summarization through MMR, integer 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 typically select sentences from the source document. Our framework transfers this global selection perspective to language-model-generated abstractive candidates, jointly incorporating source relevance, consistency, redundancy, and an explicit sentence budget. A broader discussion of generation-based summarization, hallucination reduction, and optimization-based selection is provided in Appendix A.

3 Method

3.1 Problem Definition and Framework Overview

Given an input document or document set $\mathcal{D}$, the goal of abstractive summarization is to produce a concise summary $\hat{y}$ that captures salient information while remaining consistent with the source. Conventional systems typically generate $\hat{y}$ through a single autoregressive decoding trajectory. Although this formulation produces fluent text, it provides only indirect control over which content is selected, how much information is included, and whether different parts of the summary are redundant.

We instead formulate summarization as a generation-and-selection problem over a finite candidate space. As illustrated in Figure 2, the framework consists of three stages. First, a pretrained generator G produces multiple candidate summaries for the same input. Second, the generated summaries are decomposed into sentence-level candidates, which are scored for coverage, factuality, and pairwise redundancy. Third, a combinatorial selector chooses a subset of candidates under an explicit sentence budget, after which the selected sentences are ordered and concatenated to form the final summary.

Formally, the generator produces $Y = \{y_1, y_2, \dots, y_p\}$, where each y_j corresponds to a distinct decoding trajectory. The generated summaries are segmented into sentences and aggregated into a candidate pool $\mathcal{C} = \{c_1, c_2, \dots, c_n\}$. Exact duplicates are removed before optimization.

The final summary is constructed by selecting a

subset $A \subseteq \mathcal{C}$ under a sentence-level budget:

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

where B is the maximum number of selected sentences and $F(A)$ measures the quality of a candidate subset. The objective favors subsets that cover salient source information, contain source-supported claims, and avoid semantic repetition.

This formulation separates fluent text generation from global content selection. The generator explores multiple possible summaries, while the selector compares candidate content across decoding trajectories and imposes explicit constraints at the subset level. Because the selector operates over generated sentences rather than source sentences, the framework retains the abstractive capability of pretrained models while allowing global content planning without retraining the generator.

3.2 Candidate Generation

The candidate-generation stage converts the open-ended output space of a language model into a finite set of reusable sentence candidates. Rather than committing to a single decoded summary, we generate multiple hypotheses using beam search or stochastic decoding. Different trajectories may emphasize different facts, paraphrases, or content combinations, thereby exposing a broader set of possible summary units to the downstream selector.

This design is motivated by the observation that salient and source-supported information often appears across multiple decoding trajectories, whereas weakly supported or hallucinated content is less stable. The framework does not assume that repeated content is necessarily correct; instead, it treats the candidate pool as a search space whose elements are subsequently evaluated against the source.

Each generated summary y_j is segmented into sentences. The resulting sentences are pooled and deduplicated to construct $\mathcal{C}$. We retain the originating summary and within-summary position of each candidate for analysis and realization, although subset selection is performed over the deduplicated pool.

Unlike summary-level reranking, which selects one complete hypothesis from Y , our framework treats generated sentences as reusable content units. This allows the final summary to combine complementary information from multiple candidate summaries under a shared budget. The generation

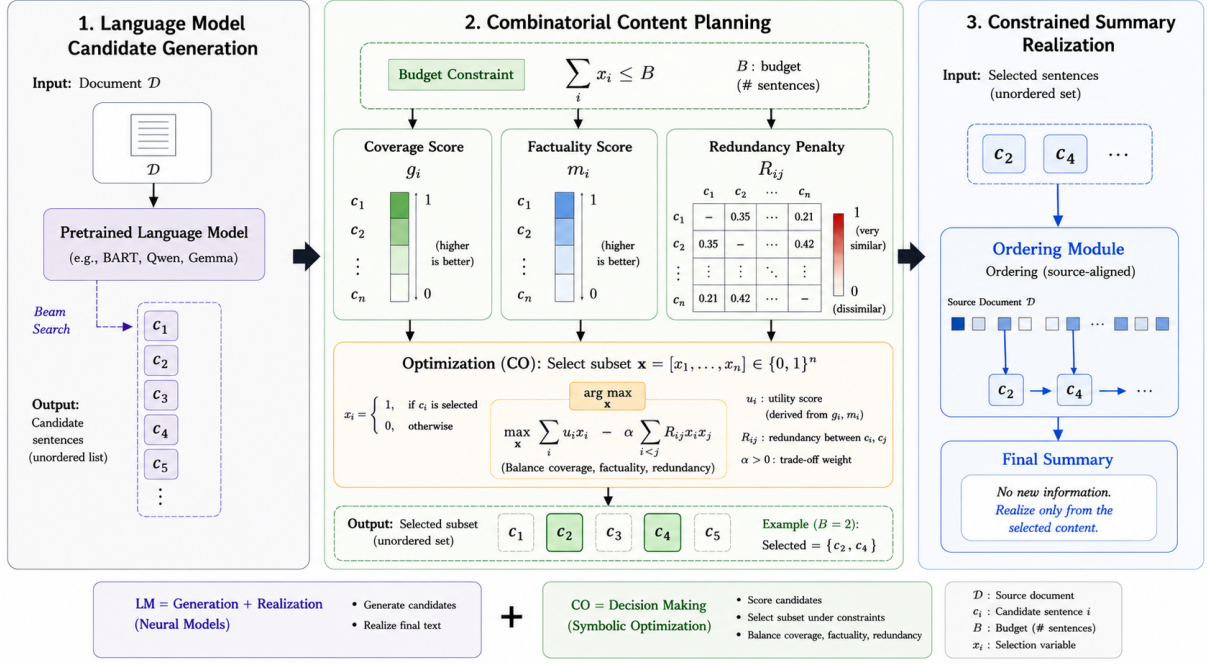


Figure 2: Overview of the proposed framework. A pretrained generator first produces multiple candidate summaries, which are decomposed into sentence-level candidates. A combinatorial selector then chooses a subset under a sentence budget by balancing coverage, factuality, and redundancy. Finally, the selected sentences are ordered according to the source document and concatenated to form the final summary.

stage is independent of the selector and can therefore be used with encoder–decoder summarization models or instruction-tuned LLMs without architectural modification.

3.3 Candidate Scoring

For each candidate sentence, we compute two sentence-level scores and one pairwise score. The sentence-level scores measure coverage and factuality, whereas the pairwise score estimates redundancy between candidates.

Coverage The coverage score measures how well a candidate captures salient information from the source:

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

where ϕ is a relevance function comparing candidate c_i with $\mathcal{D}$. Depending on the implementation, ϕ may use lexical overlap, semantic similarity, contextual embeddings, or a learned relevance estimator. Higher values indicate that the candidate is more representative of the source content.

Factuality To encourage source grounding, we compute a sentence-level factuality score:

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

where P_{fact} estimates whether the claim expressed by c_i is supported by the source. Candidates judged to be unsupported or weakly grounded therefore receive lower utility during selection.

Coverage and factuality capture complementary properties. A sentence may be well supported but peripheral to the main topic, or highly relevant but contain an unsupported detail. We combine both signals into a unified sentence-level utility:

$$u_i = \lambda_{\text{cov}} g_i + \lambda_{\text{fact}} m_i, \quad (4)$$

where λ_{cov} and λ_{fact} control the relative importance of relevance and source grounding.

Redundancy Sentence-level utility alone may favor several candidates expressing the same fact. We therefore compute a pairwise redundancy matrix $R \in \mathbb{R}^{n \times n}$:

$$R_{ij} = \psi(c_i, c_j), \quad (5)$$

where ψ measures lexical or semantic similarity between candidates c_i and c_j . Large values indicate that the two candidates convey substantially overlapping content. Before optimization, the scoring components are normalized within each candidate pool so that coverage, factuality, and redundancy contribute on comparable scales.

3.4 Budget-Constrained Selection

Let

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

denote the binary selection vector, where $z_i = 1$ indicates that candidate c_i is selected. We first define a general quality–redundancy objective:

$$\begin{aligned} \max_{\mathbf{z} \in \{0,1\}^n} \quad & \sum_{i=1}^n u_i z_i - \alpha \sum_{i < j} R_{ij} z_i z_j \\ \text{s.t.} \quad & \sum_{i=1}^n z_i \leq B, \end{aligned} \quad (6)$$

where $\alpha \geq 0$ controls the strength of redundancy suppression. The first term rewards individually informative and source-supported candidates, whereas the second penalizes pairs that convey overlapping information.

Equation 6 defines the general selection problem, but the framework is not tied to a particular optimization algorithm. We compare three selectors using identical candidate pools, utilities, pairwise similarities, and sentence budgets. MMR performs greedy local selection by balancing candidate utility against similarity to the current subset. The integer-programming selector uses either hard pairwise redundancy constraints or a soft linearized redundancy penalty, depending on the scoring configuration. The DPP-inspired selector instead uses a log-determinant criterion to evaluate candidate quality and subset diversity jointly.

DPP-Inspired Selection The DPP-inspired selector uses the same candidate utilities and pairwise redundancy scores as the other selectors, but replaces the additive pairwise penalty in Eq. 6 with a determinant-based subset criterion. It is therefore an alternative instantiation of the general selection framework rather than an exact reformulation of Eq. 6.

For each candidate c_i , we derive a positive quality score from its coverage–factuality utility:

$$q_i = \max(u_i, \epsilon_q), \quad Q = \text{diag}(q_1, \dots, q_n), \quad (7)$$

where ϵ_q prevents zero-quality entries. Let $K \in \mathbb{R}^{n \times n}$ denote the similarity matrix obtained by scaling the pairwise redundancy scores in Eq. 5 according to the redundancy weight and clipping the resulting values to $[0, 1]$. We construct

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

where ϵ provides numerical regularization. The diagonal quality terms favor candidates with high coverage–factuality utility, while the log-determinant criterion is used to discourage subsets containing highly similar candidates.

For a target sentence budget B , the selector seeks

$$\begin{aligned} A^* &= \arg \max_{A \subseteq \mathcal{C}} \log \det(L_A) \\ \text{s.t.} \quad & |A| = \min(B, |\mathcal{C}|), \end{aligned} \quad (9)$$

where L_A denotes the principal submatrix indexed by A . Unlike the general constraint in Eq. 6, the implemented selector fills the available sentence budget whenever the candidate pool contains at least B candidates.

Exact optimization is combinatorial, so we use greedy log-determinant maximization. Starting from the empty set, the selector evaluates every remaining candidate and adds the one producing the largest valid value of $\log \det(L_{A \cup \{i\}})$. A candidate addition is valid only when the corresponding determinant has positive sign. The process continues until the target number of candidates has been selected.

If no remaining candidate yields a valid determinant, the selector adds the candidate with the largest remaining diagonal entry of L . If the greedy procedure encounters a numerical exception, it instead selects the candidates with the highest quality scores. When $|\mathcal{C}| \leq B$, all candidates are retained.

Because preprocessing does not guarantee that the heuristic similarity matrix K is positive semi-definite, the resulting L need not define a valid DPP kernel. We therefore use the log-determinant criterion as a deterministic diversity heuristic and refer to the method as *DPP-inspired*, rather than as a probabilistic DPP.

3.5 Summary Realization

The selected subset A^* is unordered, but a readable summary requires a coherent sentence sequence. For each selected candidate c_i , we identify the most semantically similar sentence in the source and assign the corresponding source position $p(c_i)$. The selected candidates are then sorted by $p(c_i)$ and concatenated to form the final summary $\hat{y}$.

This source-aligned ordering preserves the approximate information flow of the input document and avoids introducing an additional generation step. We do not apply neural rewriting after selection because rewriting could modify the selected factual content or violate the sentence budget. The

final output therefore directly realizes the optimized subset, although combining sentences from different decoding trajectories may occasionally introduce minor discourse or stylistic inconsistencies.

4 Results

Detailed experimental settings, evaluation protocols, complete statistical results, and additional analyses are provided in Appendix B.

4.1 Main Results

Table 1 compares the proposed DPP-based selection framework with direct-generation baselines, prior summarization systems, faithfulness-oriented methods, and instruction-tuned LLMs on CNN/DailyMail and Multi-News.

Overall, DPP-based selection consistently improves factuality-oriented metrics while reducing reference-overlap scores. On CNN/DailyMail, BART+DPP achieves the strongest factuality performance across all five reported metrics, obtaining FactCC 79.93, MiniCheck 97.73, AlignScore 94.63, FactKB 99.46, and FaithLens 99.63. Relative to direct BART, the largest gains occur on FactCC and AlignScore, indicating that sentence-level selection improves both entailment-based and semantic-alignment measures of source grounding.

The gains are more pronounced on Multi-News. Compared with direct PRIMERA, PRIMERA+DPP improves MiniCheck from 63.06 to 85.38, AlignScore from 53.30 to 71.27, and FaithLens from 64.46 to 79.70. Similarly, Llama-3-8B+DPP improves FactCC from 38.35 to 53.67, MiniCheck from 60.46 to 83.15, AlignScore from 59.04 to 72.56, FactKB from 90.99 to 98.29, and FaithLens from 68.52 to 95.14. These larger improvements suggest that global sentence selection is particularly effective when salient information and redundant content are distributed across multiple source documents.

Nearly all paired differences are statistically significant after Holm correction. The only unmarked comparisons are FactKB and FaithLens for Llama-3-8B+DPP on CNN/DailyMail. Complete paired differences, confidence intervals, and adjusted p values are reported in the appendix.

The factuality gains are accompanied by lower ROUGE and BERTScore. This trade-off reflects the selection objective, which prioritizes source-supported and non-redundant content rather than

lexical agreement with a single reference summary. Reference overlap and factual grounding should therefore be interpreted as complementary dimensions of summary quality rather than interchangeable measures.

4.2 Additional Faithfulness Benchmarks

Table 2 evaluates the framework on FaithBench and TofuEval, which provide complementary tests of factual consistency beyond the standard CNN/DailyMail and Multi-News settings.

On FaithBench, BART+DPP improves over direct BART on every metric, increasing FactCC from 80.13 to 89.25, MiniCheck from 86.62 to 92.89, AlignScore from 88.92 to 93.94, FactKB from 97.25 to 99.90, and FaithLens from 86.67 to 94.67. The Llama-based selector also produces substantial gains, particularly on MiniCheck, AlignScore, FactKB, and FaithLens.

A similar pattern appears on TofuEval. BART+DPP achieves the strongest overall factuality performance, including MiniCheck 96.47 and AlignScore 90.17. Llama-3-8B+DPP likewise improves all five metrics over its direct-generation baseline. FaithLens reaches 100.00 for both DPP systems on this benchmark, making the remaining factuality metrics important for distinguishing their performance.

These results show that the source-grounding improvements extend beyond the two standard news benchmarks and are observed across both task-specific summarization models and instruction-tuned LLMs.

4.3 Comparison of Selection Strategies

Table 3 compares MMR, ILP, and DPP using identical candidate pools, scoring functions, and sentence budgets. This comparison isolates the effect of the selector from differences in candidate generation.

MMR generally preserves higher ROUGE and BERTScore, whereas ILP and DPP tend to obtain stronger factuality-oriented results. For BART, DPP achieves the highest MiniCheck, AlignScore, and FactKB, while ILP obtains the highest FactCC. For the Llama-based systems, ILP performs slightly better on several CNN/DailyMail factuality metrics, whereas DPP obtains the strongest FactKB. On Multi-News, ILP and DPP perform similarly for PRIMERA, while DPP achieves the highest FactCC, AlignScore, and FactKB for Llama-3-8B.

No selector dominates every metric. Instead, MMR, ILP, and DPP provide different operating

Dataset	Method	ROUGE ($\uparrow$)				BERTScore ($\uparrow$)	Faithfulness ($\uparrow$)				
		R-1	R-2	R-L	R-Lsum	F1	FactCC	MiniCheck	AlignScore	FactKB	FaithLens
CNN/DM	BART	44.09	21.13	30.68	41.05	88.22	75.91	94.96	91.55	98.47	98.02
	FactEdit	42.53	20.48	39.74	40.76	88.17	76.03	94.82	91.46	98.68	97.85
	SimCLS	46.67	22.15	43.54	43.09	66.14	59.49	83.34	82.37	80.75	95.74
	BRIO-Mul	47.78	23.55	44.57	44.12	87.84	54.06	84.67	74.48	85.20	94.23
	BRIO-Ctr	47.28	22.93	44.15	44.10	89.11	62.62	88.42	84.98	78.89	96.63
	EFactSum	44.37	21.28	40.92	41.07	88.36	60.74	88.33	82.02	91.38	97.38
	SummaReranker	46.62	22.39	43.59	43.80	88.47	68.17	93.81	88.27	98.10	97.68
	Llama-3-8B	38.98	14.82	24.20	35.28	87.72	46.78	76.00	75.08	98.51	99.09
	Qwen3.5-9B	34.88	10.91	20.33	30.77	86.96	40.16	56.82	65.17	98.12	98.69
	Gemma-4-E4B-it	30.51	6.41	18.07	27.40	86.19	53.99	69.49	71.32	89.59	75.80
	BART+DPP	39.18 $\pm 0.17^\ddagger$	17.76 $\pm 0.16^\ddagger$	26.23 $\pm 0.17^\ddagger$	36.30 $\pm 0.17^\ddagger$	87.23 $\pm 0.03^\ddagger$	79.93 $\pm 0.43^\ddagger$	97.73 $\pm 0.16^\ddagger$	94.63 $\pm 0.21^\ddagger$	99.46 $\pm 0.21^\ddagger$	99.63 $\pm 0.25^\ddagger$
	Llama-3-8B+DPP	26.40 $\pm 0.18^\ddagger$	5.15 $\pm 0.14^\ddagger$	16.39 $\pm 0.13^\ddagger$	22.91 $\pm 0.16^\ddagger$	85.06 $\pm 0.04^\ddagger$	51.57 $\pm 0.64^\ddagger$	87.56 $\pm 0.39^\ddagger$	79.56 $\pm 0.40^\ddagger$	98.55 ± 0.21	98.89 ± 0.25
Multi-News	PRIMERA	41.10	15.88	22.17	37.30	86.89	37.40	63.06	53.30	94.36	64.46
	SimCLS	45.78	17.31	23.14	41.59	85.68	32.47	54.35	48.19	57.35	58.32
	BRIO-Ctr	46.04	16.90	22.73	41.78	85.61	30.39	51.50	46.01	53.81	53.18
	SummaReranker	44.56	16.72	22.61	40.45	87.14	35.18	61.49	51.88	94.72	60.26
	Llama-3-8B	38.84	8.68	18.16	34.81	85.42	38.35	60.46	59.04	90.99	68.52
	Qwen3.5-9B	38.31	6.74	15.96	34.05	84.82	42.01	48.40	53.08	86.98	49.20
	Gemma-4-E4B-it	35.44	6.53	16.27	31.86	84.89	48.49	66.79	66.35	87.68	57.93
	PRIMERA+DPP	36.69 $\pm 0.24^\ddagger$	11.86 $\pm 0.18^\ddagger$	19.56 $\pm 0.17^\ddagger$	32.79 $\pm 0.23^\ddagger$	84.49 $\pm 0.05^\ddagger$	45.68 $\pm 0.74^\ddagger$	85.38 $\pm 0.56^\ddagger$	71.27 $\pm 0.55^\ddagger$	97.54 $\pm 0.44^\ddagger$	79.70 $\pm 1.45^\ddagger$
	Llama-3-8B+DPP	32.24 $\pm 0.20^\ddagger$	5.42 $\pm 0.09^\ddagger$	15.93 $\pm 0.09^\ddagger$	28.18 $\pm 0.18^\ddagger$	83.59 $\pm 0.04^\ddagger$	53.67 $\pm 0.80^\ddagger$	83.15 $\pm 0.56^\ddagger$	72.56 $\pm 0.53^\ddagger$	98.29 $\pm 0.57^\ddagger$	95.14 $\pm 1.30^\ddagger$

Table 1: Main results on CNN/DailyMail and Multi-News. Values following $\pm$ denote half-widths of the 95% paired-bootstrap confidence intervals for the +DPP-minus-baseline differences. ‡ indicates $p_{\text{Holm}} < 0.01$ under a two-sided paired sign-flip permutation test with Holm correction over the ten reported metrics. Unmarked +DPP results are not statistically significant after correction. Bold values indicate the highest score for each metric within each dataset. Complete paired differences, confidence intervals, and adjusted p values are reported in the appendix.

Dataset	Method	FactCC	MiniCheck	AlignScore	FactKB	FaithLens
FaithBench	BART	80.13	86.62	88.92	97.25	86.67
	PRIMERA	34.11	31.22	33.44	81.85	28.00
	Llama-3-8B	37.25	47.54	55.76	60.98	66.67
	Qwen3.5-9B	38.11	29.81	41.48	66.64	40.00
	Gemma-4-E4B-it	49.13	54.02	59.74	60.71	53.33
	SimCLS	29.45	30.11	28.42	37.52	40.00
	BRIO-Ctr	30.23	28.57	27.94	36.40	44.00
	SummaReranker	29.27	34.06	34.66	74.76	45.33
	BART+DPP	89.25	92.89	93.94	99.90	94.67
	Llama-3-8B+DPP	42.43	73.34	70.86	90.83	84.00
TofuEval	BART	59.15	86.41	81.18	94.01	98.00
	PRIMERA	35.82	44.11	41.15	92.52	33.00
	Llama-3-8B	39.38	56.31	54.61	84.14	90.00
	Qwen3.5-9B	34.09	37.65	44.42	78.82	79.00
	Gemma-4-E4B-it	49.51	67.47	65.09	76.47	80.00
	SimCLS	32.50	46.30	41.10	41.80	40.00
	BRIO-Ctr	33.15	44.49	40.03	42.70	36.00
	SummaReranker	37.93	51.73	45.20	94.12	44.00
	BART+DPP	66.37	96.47	90.17	96.48	100.00
	Llama-3-8B+DPP	52.79	83.91	70.97	95.63	100.00

Table 2: Results on FaithBench and TofuEval. DPP-based selection improves factuality metrics relative to the corresponding direct-generation backbone. Best results within each dataset are shown in bold.

points within the same generation-and-selection framework: MMR retains more reference overlap, while ILP and DPP place greater emphasis on global factuality and diversity. This result supports the modularity of the framework and shows that the observed gains do not arise solely from generating a larger candidate pool.

4.4 Human Evaluation

To complement the automatic metrics, we conduct a blind pairwise evaluation on 100 CNN/DailyMail examples, including 50 comparisons for BART and 50 for Llama-3-8B. Twenty annotators compare a

direct-generation summary with its DPP-selected counterpart. System identities and output order are hidden. Annotators provide an overall preference and five-point ratings for consistency, relevance, coherence, clarity, and conciseness. Full annotation instructions are provided in Appendix E.

Table 4 shows that DPP-selected summaries are preferred in 61 comparisons, compared with 26 preferences for direct generation; 10 pairs are judged equivalent and 3 are marked uncertain. The preference pattern is consistent across both backbones, with DPP preferred in 32 of 50 BART comparisons and 29 of 50 Llama-3-8B comparisons.

DPP-selected summaries also receive higher mean ratings for consistency, relevance, clarity, and conciseness. Direct generation has a small numerical advantage in coherence, suggesting that recombining sentences from multiple candidate summaries may occasionally weaken discourse flow. Overall, the human evaluation provides supporting evidence that the selection framework improves perceived content quality and conciseness despite lower reference overlap.

4.5 Additional Analysis

Appendix D reports sensitivity analyses for optimization weights, sentence budget, and candidate-pool size. The sentence-budget analysis shows that the DPP-inspired selector reliably satisfies the requested number of sentences and retains its fac-

Dataset	Method	ROUGE ($\uparrow$)				BERTScore ($\uparrow$)	Faithfulness ($\uparrow$)				
		R-1	R-2	R-L	R-Lsum		FactCC	MiniCheck	AlignScore	FactKB	FaithLens
CNN/DM	BART+MMR	41.47	18.86	27.16	38.43	87.58	78.86	96.20	93.00	98.95	98.22
	BART+ILP	41.03	18.86	27.12	38.11	87.55	80.18	97.62	94.46	99.31	99.55
	BART+DPP	39.18	17.76	26.23	36.30	87.23	79.93	97.73	94.63	99.46	99.63
	Llama-3-8B+MMR	28.51	5.88	17.08	24.81	85.51	50.32	85.81	78.60	98.28	98.68
	Llama-3-8B+ILP	27.54	5.41	16.71	23.95	85.30	52.04	87.89	79.64	98.18	98.96
	Llama-3-8B+DPP	26.40	5.15	16.39	22.91	85.06	51.57	87.56	79.56	98.55	98.89
Multi-News	PRIMERA+MMR	38.87	13.27	20.30	35.04	85.06	45.23	83.85	69.89	96.49	74.14
	PRIMERA+ILP	37.26	12.18	19.73	33.36	84.62	45.77	85.39	71.27	97.40	78.87
	PRIMERA+DPP	36.69	11.86	19.56	32.79	84.49	45.68	85.38	71.27	97.54	79.70
	Llama-3-8B+MMR	32.77	5.60	16.02	28.70	83.72	53.46	83.03	72.39	98.14	94.88
	Llama-3-8B+ILP	32.99	5.64	16.02	28.91	83.77	53.43	83.30	72.46	98.10	95.11
	Llama-3-8B+DPP	32.24	5.42	15.93	28.18	83.59	53.67	83.15	72.56	98.29	95.14

Table 3: Comparison of MMR, ILP, and DPP under identical candidate pools, scoring functions, and sentence budgets. MMR generally preserves more reference overlap, whereas ILP and DPP provide stronger factuality-oriented performance.

(a) Pairwise preference counts				
Backbone	DPP	Direct	Same	Unsure
BART	32	12	5	1
Llama-3-8B	29	14	5	2
Overall	61	26	10	3

(b) Mean quality ratings			
Dimension	DPP	Direct	Δ
Coherence	4.360	4.422	-0.062
Consistency	4.466	4.034	+0.432
Relevance	4.341	3.708	+0.633
Clarity	4.379	3.614	+0.765
Conciseness	4.218	3.326	+0.892

Table 4: Human evaluation results. Panel (a) reports pairwise preference counts over 100 comparisons. Panel (b) reports mean ratings on a 1–5 scale, with Δ computed as DPP-inspired selection minus direct generation. Bold values indicate the preferred system or the higher mean rating.

tuality advantage across budgets from two to five sentences. Faithfulness-oriented weighting generally improves FactCC, MiniCheck, and AlignScore, whereas diversity-oriented weighting preserves more ROUGE-Lsum. Increasing the candidate pool further improves factuality, particularly on Multi-News, but also increases generation and scoring cost and strengthens the trade-off with reference overlap. A qualitative case study is provided in Appendix C.

5 Conclusion

We presented a modular generation-and-selection framework for sentence-budget-constrained abstractive summarization. The framework generates

multiple candidate summaries and applies global sentence-level selection to balance relevance, factuality, and redundancy without retraining the underlying generator. Experiments across single-document, multi-document, and factuality-oriented benchmarks show consistent improvements in automatic source-grounding metrics, particularly for multi-document summarization, while revealing a trade-off with reference-overlap scores. Human evaluation further shows that selected summaries are preferred more frequently and receive higher ratings for consistency, relevance, clarity, and conciseness. These results demonstrate the potential of decoupling generation from global content selection for controllable, factuality-oriented summarization.

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 framework operates at the sentence level and does not explicitly model discourse relations or cross-sentence dependencies. Although source-aligned ordering preserves the approximate information flow of the input, summaries assem-

bled from different decoding trajectories may contain stylistic inconsistencies or abrupt transitions. This limitation is also reflected in the human evaluation, where direct generation obtains a small numerical advantage in coherence. Moreover, because each comparison is evaluated by a single annotator, inter-annotator agreement cannot be estimated.

Third, the framework uses automatic relevance and factuality scores during selection. Improvements on related automatic evaluation metrics may therefore partly reflect metric alignment and should not be interpreted as guarantees of error-free factuality. The DPP-inspired selector also relies on greedy log-determinant maximization, which does not guarantee a globally optimal subset. Finally, the current implementation enforces a sentence-count budget rather than a finer-grained token- or word-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, although our evaluation spans news summarization and topic-focused dialogue summarization, it does not establish generalization to scientific, legal, clinical, or other long-form and highly abstractive domains. Evaluating the framework in these settings remains an important direction for 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.
- 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.
- Gemma Team. 2024. [Gemma: Open models based on Gemini research and technology](#).
- 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 Gülçehre, 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.
- Shuzheng Si, Qingyi Wang, Haozhe Zhao, Yuzhuo Bai, Guanqiao Chen, Kangyang Luo, Gang Chen, Fan-chao Qi, Minjia Zhang, Baobao Chang, and Maosong Sun. 2026. [FaithLens: Detecting and explaining faithfulness hallucination](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 14068–14099, San Diego, California, United States. 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 Shalymov, 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 Related Work

A.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).

A.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.

A.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.

B Experimental Setup

B.1 Datasets

We evaluate the proposed framework on two standard news summarization datasets and two factuality-oriented benchmarks. Together, they cover single-document, multi-document, and topic-focused dialogue summarization, as well as challenging hallucination settings involving modern LLMs.

CNN/DailyMail CNN/DailyMail (Nallapati et al., 2016) is a widely used benchmark for single-document news summarization. It contains news articles paired with multi-sentence abstractive summaries written from article highlights. We use the standard non-anonymized test split to evaluate summary quality and factual consistency in a conventional single-document setting.

Multi-News Multi-News (Fabbri et al., 2019) is a multi-document summarization dataset containing clusters of related news articles paired with professionally written summaries. Compared with CNN/DailyMail, it requires models to aggregate information across multiple documents while controlling cross-document redundancy and maintaining consistency over substantially longer inputs. We use this dataset to evaluate whether global sentence selection is particularly beneficial when relevant information is distributed across multiple sources.

FaithBench FaithBench (Bao et al., 2025) is a hallucination benchmark designed to evaluate summaries produced by modern LLMs. It contains 750 source–summary pairs constructed from 75 source passages and outputs from 10 LLMs spanning eight model families. The benchmark emphasizes challenging examples on which widely used hallucination detectors disagree. Expert annotators identify unsupported spans and categorize them as *benign*, *questionable*, or *unwanted*, in addition to fully consistent cases. We use the source passages as summarization inputs and evaluate whether the proposed selector improves source grounding on examples associated with difficult and diverse hallucination patterns.

TofuEval TofuEval (Tang et al., 2024b) evaluates factual consistency in topic-focused dialogue summarization. It is constructed from 100 long dialogues, including 50 MediaSum interview transcripts and 50 MeetingBank city council meetings. Three topics are identified for each dialogue, and five LLMs generate a summary for each topic. After invalid outputs are removed, the benchmark contains 1,479 summaries comprising 3,966 sentences. Professional linguistic annotators provide sentence-level factuality labels and explanations for unsupported statements, together with summary-level relevance and completeness assessments. We use each dialogue–topic pair as a summarization instance to examine whether the framework transfers beyond news articles to long, multi-speaker, and conversational inputs.

CNN/DailyMail and Multi-News provide standard reference-based evaluation, whereas FaithBench and TofuEval offer complementary tests of source grounding under challenging LLM-generated hallucinations and domain transfer.

B.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** (Gemma Team, 2024; Google DeepMind, 2026).

B.3 Implementation Details

All experiments are implemented using PyTorch and the Hugging Face Transformers library. For each input, the generation model first produces multiple candidate summaries, which are segmented into sentence-level candidates to construct the optimization pool $\mathcal{C}$. Unless otherwise specified, all optimization-based methods operate on the same cached candidate pools and sentence-level scores, ensuring that comparisons among MMR, ILP, and DPP-inspired selection isolate the effect of the selector.

Default Experimental Configuration

The main optimization-based experiments use a candidate-generation width of 12 and a target sentence budget of $B = 3$. All selector comparisons in Table 3 use identical candidate pools, scoring functions, and sentence budgets. The sentence-budget analysis varies $B \in \{2, 3, 4, 5\}$ while holding the beam-12 candidate pools and candidate scores fixed in Table 7. The candidate-pool analysis varies the generation width over $\{4, 8, 12\}$, while the weight-sensitivity analysis uses beam size 8 and the configurations reported in Table 8.

Unless otherwise specified, the main experiments use the balanced weighting configuration

$$(w_{\text{cov}}, w_{\text{fact}}, w_{\text{red}}) = (0.33, 0.33, 0.34),$$

where the three components control ROUGE-based coverage, MiniCheck-based factuality, and pairwise redundancy, respectively.

Generation Models For CNN/DailyMail, we use facebook/bart-large-cnn (Lewis et al., 2020). The direct-generation baseline follows the checkpoint’s standard summarization configuration. The main optimization-based experiments retain all generated hypotheses for candidate-pool construction. Other decoding settings follow the checkpoint configuration, including length normalization, no-repeat trigram blocking, and early stopping.

For Multi-News, 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 language models, we use concise, dataset-specific summarization prompts. The evaluated checkpoints are Qwen/Qwen3.5-9B (Team, 2026), meta-llama/Llama-3.1-8B-Instruct (AI@Meta, 2024), and google/gemma-4-E4B-it (Gemma Team, 2024; Google DeepMind, 2026). For compactness, we refer to meta-llama/Llama-3.1-8B-Instruct as Llama-3-8B in the tables and discussion. Direct-generation baselines use one returned summary, whereas the main optimization-based runs retain 12 generated hypotheses per input before sentence-level selection.

For Multi-News inputs, source documents are concatenated in their original dataset order using explicit document delimiters. Inputs are truncated only when necessary to satisfy the maximum context length of the corresponding generation model.

Candidate Pool Construction Generated summaries are segmented into sentence-level candidates using standard sentence tokenization. Sentences from all retained hypotheses are combined into a single pool, and exact duplicates are removed before scoring. The resulting deduplicated pool is shared by all downstream selectors. We retain the originating hypothesis and within-summary posi-

tion of each candidate for analysis and final summary realization.

Scoring Configuration Candidate coverage is computed using a ROUGE-based relevance score, while sentence-level factuality is estimated using MiniCheck (Tang et al., 2024a). Pairwise redundancy is computed using semantic similarity between candidate sentences. For each selected candidate, source-aligned realization identifies its most semantically similar source sentence and uses the corresponding source position for ordering.

Coverage, factuality, and redundancy scores are normalized independently within each candidate pool. Sentence-level coverage and factuality scores are normalized using min–max scaling. For redundancy, min–max normalization is applied to the off-diagonal entries of the pairwise matrix, while diagonal entries are retained as self-similarity values. This places the three scoring components on comparable scales while preserving their relative ranking within each candidate pool.

Optimization MMR greedily selects candidates by balancing normalized utility against similarity to the current subset until the sentence budget is filled. The integer-programming variants are implemented in PuLP and solved using the CBC solver. The hard formulation selects exactly B candidates and introduces a mutual-exclusion constraint for each candidate pair whose redundancy exceeds the configured threshold. The tri-metric formulation uses binary auxiliary variables to linearize the pairwise redundancy penalty and permits $1 \leq \sum_i z_i \leq B$. In the reported tri-metric experiments, the redundancy coefficient uses per-edge scaling:

$$\alpha = \frac{w_{\text{red}}}{\max(1, B - 1)}.$$

For the DPP-inspired selector, candidate utilities are clipped below at $\epsilon_q = 0.01$ to construct the diagonal quality matrix Q . Pairwise redundancy scores are scaled according to w_{red} and clipped to $[0, 1]$ to form K . We then construct

$$L = QKQ + 10^{-6}I$$

and apply greedy log-determinant maximization until the target budget is filled. Candidate additions with non-positive determinant sign are treated as invalid. If no valid candidate remains, the selector adds the candidate with the largest remaining diagonal entry of L . A numerical exception triggers fallback selection by descending candidate quality. When $|\mathcal{C}| \leq B$, all candidates are retained.

B.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, ROUGE-L, and ROUGE-Lsum, 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. Finally, **FaithLens** (Si et al., 2026) is an 8B-parameter explainable faithfulness evaluator that determines whether a generated claim is supported by the source document and produces a corresponding rationale for its prediction.

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

C Case Study

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 5 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.

D Additional Results

This appendix provides sensitivity analyses that complement the main findings.

D.1 Sensitivity to Optimization Weights

Table 6 analyzes the effect of varying the relative weights assigned to coverage, factuality, and diversity while keeping candidate generation and DPP selection fixed. We compare balanced, coverage-heavy, faithfulness-heavy, and diversity-heavy configurations.

Faithfulness-heavy weighting generally produces the strongest factuality results. On CNN/DailyMail, the faithfulness-heavy setting improves Llama+DPP from 48.81 to 50.01 on FactCC, from 86.23 to 87.81 on MiniCheck, and from 81.26 to 82.00 on AlignScore relative to balanced weighting. For BART+DPP, the same configuration obtains the highest FactCC, MiniCheck, and AlignScore.

A similar pattern appears on Multi-News. For PRIMERA+DPP, faithfulness-heavy weighting im-

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 5: 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.

proves FactCC from 42.31 to 43.75, MiniCheck from 79.37 to 80.43, and AlignScore from 65.59 to 66.41. For Llama+DPP, it improves FactCC from 52.60 to 53.63, MiniCheck from 78.22 to 79.25, and AlignScore from 68.94 to 69.80.

Diversity-heavy weighting often yields the highest ROUGE-Lsum, whereas faithfulness-heavy weighting generally achieves the strongest factuality scores. This result further demonstrates that the framework supports explicit control over the trade-off between reference coverage, source grounding, and redundancy reduction.

D.2 Effect of Sentence Budget and Output Length

To assess budget controllability and determine whether the observed factuality gains depend on a particular output length, we vary the target sentence budget over $B \in \{2, 3, 4, 5\}$. For each setting, we report average output length in words, exact-budget adherence, reference-overlap scores, and factuality-oriented metrics.

Table 7 shows that the DPP-inspired selector satisfies the requested sentence budget for every evaluated example, whereas direct generation becomes increasingly unreliable as B grows. The exact-budget hit rate of direct BART decreases

from 99.76% at $B = 2$ to 13.73% at $B = 5$, while direct PRIMERA decreases from 100.00% to 66.36%. In contrast, the DPP-inspired selector maintains a 100% hit rate across all budgets and both datasets.

Across budget settings, DPP-inspired selection consistently improves most factuality-oriented metrics, including FactCC, MiniCheck, AlignScore, and the percentage of fully faithful summaries, although these gains are accompanied by lower ROUGE-Lsum and BERTScore. The source-grounding improvements therefore persist across multiple sentence budgets rather than arising from a single length configuration.

Because the constraint is defined over sentences rather than words, summaries with the same target budget may still differ substantially in realized word count. Notably, the DPP-inspired outputs are often longer than the corresponding direct-generation outputs at the same value of B , particularly on CNN/DailyMail. Their factuality improvements therefore cannot be attributed simply to producing shorter summaries. Reporting both exact-budget adherence and average word count provides a more complete view of controllability under a sentence-level budget.

Dataset	Model	Weight	R-Lsum	FactCC	MiniCheck	AlignScore	FactKB	FaithLens
CNN/DM	BART+DPP	Bal.	23.43	74.33	97.51	92.09	99.59	98.60
		Cov.	21.01	73.25	96.96	91.45	99.46	98.40
		Faith.	23.12	74.67	97.63	92.60	99.50	98.40
		Div.	26.24	73.37	97.35	91.99	99.35	98.40
	Llama+DPP	Bal.	19.54	48.81	86.23	81.26	99.43	98.00
		Cov.	19.00	46.57	80.98	78.96	99.52	98.00
		Faith.	19.02	50.01	87.81	82.00	99.43	98.60
		Div.	20.48	48.42	85.55	80.87	99.22	98.00
Multi-News	PRIMERA+DPP	Bal.	33.43	42.31	79.37	65.59	97.80	78.40
		Cov.	33.64	39.32	75.71	62.58	98.36	78.80
		Faith.	32.83	43.75	80.43	66.41	97.48	77.00
		Div.	34.64	42.06	79.08	65.28	97.71	76.60
	Llama+DPP	Bal.	30.41	52.60	78.22	68.94	97.98	93.20
		Cov.	30.85	50.61	73.79	66.93	97.80	92.60
		Faith.	29.59	53.63	79.25	69.80	98.24	94.00
		Div.	31.26	52.23	77.16	68.61	97.51	92.00

Table 6: Trade-offs between coverage, faithfulness, and diversity under different optimization weight profiles. All results use DPP selection with beam size 8. Weight abbreviations denote: Bal. = (0.33, 0.33, 0.34), Cov. = (0.60, 0.20, 0.20), Faith. = (0.20, 0.60, 0.20), and Div. = (0.20, 0.20, 0.60), where the tuple corresponds to (w_R, w_M, w_D) for ROUGE-based coverage, MiniCheck-based faithfulness, and redundancy/diversity control. Best values within each model block are shown in bold.

D.3 Effect of Candidate Pool Size

Table 8 examines the effect of candidate pool size by varying the beam size used during candidate generation. Beam size 4 serves as the smallest-pool setting, while beam sizes 8 and 12 expose larger sets of sentence candidates to the DPP-inspired selector.

Larger candidate pools generally improve factuality-oriented performance. On CNN/DailyMail, increasing the beam size from 4 to 12 raises BART MiniCheck from 94.96 to 97.73 and AlignScore from 91.55 to 94.63. For Llama-3-8B, MiniCheck increases from 76.00 to 87.56, while AlignScore increases from 75.08 to 79.56. The corresponding FactCC scores also improve from 75.91 to 79.93 for BART and from 46.78 to 51.57 for Llama-3-8B. However, the Llama-based system does not improve monotonically on FactKB or FaithLens, whose highest values occur in the beam-4 setting.

The benefits are more pronounced on Multi-News. For PRIMERA, increasing the beam size from 4 to 12 improves MiniCheck from 62.26 to 85.38, AlignScore from 52.24 to 71.27, and FaithLens from 64.46 to 79.70. FactKB peaks at beam size 8, indicating that not every factuality metric improves monotonically with pool size. For Llama-3-8B, all five factuality metrics improve substantially: MiniCheck rises from 60.46 to 83.15, AlignScore

from 59.04 to 72.56, FactKB from 90.99 to 98.29, and FaithLens from 68.52 to 95.14.

These results indicate that larger candidate pools provide the selector with more source-supported and complementary content. The gains are accompanied by lower ROUGE-Lsum, particularly for the Llama-based systems, as well as higher generation and scoring costs. Candidate pool size therefore acts as a practical control parameter for balancing source grounding, reference overlap, and inference efficiency.

E Human Evaluation Protocol

We conduct a blind pairwise evaluation on 100 CNN/DailyMail examples: 50 generated with BART and 50 with Llama-3-8B. Twenty third-year undergraduate annotators each evaluate five unique examples. Each comparison is assigned to one annotator.

For every example, the annotator is shown the source article, reference summary, and two anonymized outputs presented in randomized order as Summary A and Summary B. One output is produced by direct generation and the other by DPP-based selection. System identities are not shown.

Annotators first select one of four overall preference labels: *A better*, *B better*, *Same*, or *Not sure*. They may also rate both summaries on a five-point

Dataset	B	Method	Output Control		Reference Overlap		Factuality				
			Avg. words	Hit (%) $\uparrow$	R-Lsum $\uparrow$	BERTScore $\uparrow$	FactCC $\uparrow$	MiniCheck $\uparrow$	AlignScore $\uparrow$	FactKB $\uparrow$	FaithLensFF $\uparrow$
CNN/DM (BART)	2	Direct	36.08	99.76	37.19	88.14	74.76	95.24	91.85	98.51	99.46
			44.57	100.00	32.65	86.84	80.94	98.01	94.99	98.77	99.87
	3	Direct	53.44	94.26	40.35	88.22	76.13	95.12	91.69	98.57	99.23
			62.78	100.00	36.13	87.18	80.53	97.90	94.84	99.18	99.74
	4	Direct	62.39	54.67	41.01	88.23	76.08	95.02	91.62	98.48	98.51
			80.00	100.00	36.66	87.23	79.93	97.73	94.63	99.46	99.63
	5	Direct	64.48	13.73	41.04	88.22	75.94	94.96	91.56	98.47	98.08
			96.85	100.00	35.35	87.06	79.48	97.59	94.43	99.44	99.50
Multi-News (PRIMERA)	2	Direct	52.64	100.00	24.24	86.55	23.81	52.46	41.97	91.75	81.00
			51.04	100.00	19.85	84.62	47.48	92.86	78.76	91.08	96.03
	3	Direct	74.92	99.36	30.38	86.78	30.25	57.58	47.71	92.93	75.65
			72.00	100.00	24.75	84.81	48.08	92.28	77.70	93.27	93.72
	4	Direct	93.21	92.99	34.41	86.89	33.83	60.63	51.07	93.66	71.06
			92.67	100.00	28.13	84.84	47.96	91.32	76.56	94.60	91.20
	5	Direct	103.47	66.36	36.42	86.91	36.04	62.35	52.68	94.19	67.38
			113.80	100.00	30.42	84.77	47.56	90.18	75.36	95.60	88.28

Table 7: Sentence-budget and output-length analysis for direct generation and DPP-inspired selection. Avg. words reports realized output length, and Hit denotes the percentage of outputs containing exactly B sentences. FaithLensFF denotes the percentage of summaries for which every sentence is classified as source-supported by FaithLens. Comparisons are matched by target sentence budget rather than word count. Bold values indicate the better result within each dataset, budget, and metric pair.

scale according to the following dimensions:

- **Consistency:** whether the summary’s claims are supported by the source article;
- **Relevance:** whether the summary captures the most important source information;
- **Coherence:** whether the sentences form a logically organized and connected summary;
- **Clarity:** whether the summary is readable and easy to understand;
- **Conciseness:** whether the summary avoids unnecessary or repetitive information.

After annotation, the randomized A/B labels are mapped back to the corresponding systems. Because each comparison is evaluated by one annotator, inter-annotator agreement cannot be estimated. The results should therefore be interpreted as a targeted preference study rather than a reliability analysis.

F Statistical Significance

We evaluate each beam-12 +DPP system against its same-backbone direct-generation baseline using paired document-level scores. For every metric, we report the mean paired difference $\Delta = +\text{DPP} - \text{baseline}$, a 95% paired-bootstrap confidence

interval based on 10,000 resamples, and a Holm-adjusted p value from a two-sided paired sign-flip permutation test with 10,000 permutations. Holm correction is applied over the ten reported metrics within each system comparison.

Table 10 confirms the trade-off observed in the main results. All four +DPP systems show significant decreases in ROUGE and BERTScore, together with significant improvements on nearly all factuality metrics. On CNN/DailyMail, BART+DPP improves FactCC by 4.02 points, AlignScore by 3.07 points, and FaithLens by 1.60 points. The gains are substantially larger on Multi-News: PRIMERA+DPP improves MiniCheck by 22.32 points and AlignScore by 17.97 points, while Llama-3-8B+DPP improves MiniCheck by 22.67 points and FaithLens by 26.63 points. The only nonsignificant differences are FactKB and FaithLens for Llama-3-8B+DPP on CNN/DailyMail. Overall, the paired analyses support the conclusion that global sentence-level selection reliably improves automatic source-grounding metrics, particularly in multi-document summarization.

Dataset	Method	Beam	R-Lsum	FactCC	MiniCheck	AlignScore	FactKB	FaithLens
CNN/DM	BART	4	41.05	75.91	94.96	91.55	98.47	98.02
		8	36.68	79.21	97.60	94.28	99.23	99.48
		12	36.30	79.93	97.73	94.63	99.46	99.63
	Llama-3-8B	4	35.28	46.78	76.00	75.08	98.51	99.09
		8	23.65	51.05	85.60	78.48	98.35	98.73
		12	22.91	51.57	87.56	79.56	98.50	98.89
Multi-News	PRIMERA	4	37.30	37.37	62.26	52.24	94.31	64.46
		8	33.74	43.35	80.51	67.39	97.68	75.49
		12	32.79	45.68	85.38	71.27	97.54	79.70
	Llama-3-8B	4	34.81	38.35	60.46	59.04	90.99	68.52
		8	29.43	52.68	79.40	70.23	98.14	93.92
		12	28.18	53.67	83.15	72.56	98.29	95.14

Table 8: Effect of candidate pool size. Increasing the beam size enlarges the candidate pool available to the DPP-inspired selector. Larger candidate pools generally improve factuality-oriented metrics, particularly FactCC, MiniCheck, and AlignScore, while reducing ROUGE-Lsum. Best values within each model block are shown in bold.

Design Component	Description
Evaluation type	Blind pairwise comparison
Dataset	CNN/DailyMail
Compared systems	Direct generation vs. DPP selection
Backbones	BART and Llama-3-8B
Comparison pairs	100 total; 50 per backbone
Annotators	20 third-year undergraduate students
Pairs per annotator	5 unique pairs
Assignment	One annotator per comparison pair
Displayed content	Source, reference, and anonymized outputs
Preference labels	A better, B better, Same, Not sure
Quality ratings	Five dimensions on a 1–5 scale

Table 9: Protocol for the blind pairwise human evaluation.

Dataset	Backbone	Metric	Δ	95% CI	p_{Holm}	Sig.
CNN/DM $n = 11,490$	BART	R-1	-4.91	[-5.07, -4.74]	$< 10^{-3}$	‡
		R-2	-3.37	[-3.53, -3.21]	$< 10^{-3}$	‡
		R-L	-4.45	[-4.62, -4.28]	$< 10^{-3}$	‡
		R-Lsum	-4.74	[-4.91, -4.58]	$< 10^{-3}$	‡
		BERTScore	-0.99	[-1.02, -0.96]	$< 10^{-3}$	‡
		FactCC	+4.02	[+3.58, +4.44]	$< 10^{-3}$	‡
		MiniCheck	+2.77	[+2.62, +2.93]	$< 10^{-3}$	‡
		AlignScore	+3.07	[+2.87, +3.28]	$< 10^{-3}$	‡
		FactKB	+0.99	[+0.79, +1.20]	$< 10^{-3}$	‡
		FaithLens	+1.60	[+1.36, +1.85]	$< 10^{-3}$	‡
	Llama-3-8B	R-1	-12.58	[-12.76, -12.41]	0.0011	‡
		R-2	-9.67	[-9.81, -9.54]	0.0011	‡
		R-L	-7.82	[-7.94, -7.69]	0.0011	‡
		R-Lsum	-12.37	[-12.54, -12.21]	0.0011	‡
		BERTScore	-2.66	[-2.70, -2.63]	0.0011	‡
		FactCC	+4.81	[+4.16, +5.44]	0.0011	‡
		MiniCheck	+11.56	[+11.18, +11.95]	0.0011	‡
		AlignScore	+4.48	[+4.08, +4.88]	0.0011	‡
		FactKB	+0.04	[-0.18, +0.24]	0.745	n.s.
		FaithLens	-0.21	[-0.46, +0.04]	0.248	n.s.
Multi-News $n = 5,622$	PRIMERA	R-1	-4.41	[-4.65, -4.18]	$< 10^{-3}$	‡
		R-2	-4.01	[-4.19, -3.83]	$< 10^{-3}$	‡
		R-L	-2.61	[-2.78, -2.45]	$< 10^{-3}$	‡
		R-Lsum	-4.50	[-4.72, -4.27]	$< 10^{-3}$	‡
		BERTScore	-2.41	[-2.46, -2.36]	$< 10^{-3}$	‡
		FactCC	+8.28	[+7.55, +9.03]	$< 10^{-3}$	‡
		MiniCheck	+22.32	[+21.76, +22.88]	$< 10^{-3}$	‡
		AlignScore	+17.97	[+17.43, +18.53]	$< 10^{-3}$	‡
		FactKB	+3.15	[+2.71, +3.58]	$< 10^{-3}$	‡
		FaithLens	+15.24	[+13.80, +16.70]	$< 10^{-3}$	‡
	Llama-3-8B	R-1	-6.60	[-6.79, -6.40]	$< 10^{-3}$	‡
		R-2	-3.26	[-3.35, -3.17]	$< 10^{-3}$	‡
		R-L	-2.23	[-2.32, -2.14]	$< 10^{-3}$	‡
		R-Lsum	-6.62	[-6.80, -6.45]	$< 10^{-3}$	‡
		BERTScore	-1.82	[-1.86, -1.78]	$< 10^{-3}$	‡
		FactCC	+15.35	[+14.55, +16.15]	$< 10^{-3}$	‡
		MiniCheck	+22.67	[+22.12, +23.24]	$< 10^{-3}$	‡
		AlignScore	+13.52	[+12.99, +14.05]	$< 10^{-3}$	‡
		FactKB	+7.30	[+6.74, +7.87]	$< 10^{-3}$	‡
		FaithLens	+26.63	[+25.31, +27.91]	$< 10^{-3}$	‡

Table 10: Paired statistical comparisons of the four beam-12 +DPP systems against their same-backbone direct-generation baselines in Table 1. Δ is computed as +DPP minus baseline and is reported in percentage points. Confidence intervals are obtained from 10,000 paired-bootstrap resamples. Holm-adjusted p values are computed using two-sided paired sign-flip permutation tests with 10,000 permutations, with correction over the ten metrics within each system comparison. ‡ denotes $p_{\text{Holm}} < 0.01$, and n.s. denotes a nonsignificant difference. Because Δ is computed from paired document-level scores, it may differ slightly from differences between the rounded aggregate values in Table 1.