Faithfulness-Aware Decoding via Constrained Optimization for Multi-Document Summarization: Framework, Diagnosis, and Empirical Analysis

Suhrid Pandey [†] and Sameer Kumar Singh [*,†]

Posted Date: 20 May 2026

doi: 10.20944/preprints202605.1381.v1

Keywords: multi-document summarization; faithfulness-aware decoding; constrained optimization; self-healing; NLI verification; MiniCheck; hallucination reduction; beam collapse; nucleus sampling; rank aggregation; Lagrangian relaxation; evidence grounding; contradiction detection

 doi:10.20944/preprints202605.1381.v1

Article

Faithfulness-Aware Decoding via Constrained Optimization for Multi-Document Summarization: Framework, Diagnosis, and Empirical Analysis

Suhrid Pandey [1,†] **and Sameer Kumar Singh** [2,*,†]

[1] Independent Researcher, Bangalore, India
[2] Independent Researcher, Lucknow, India
* Correspondence: sameerkumarsingh56@gmail.com
† These authors contributed equally to this work.

Abstract

Multi-document summarization (MDS) under strict context budgets is acutely vulnerable to hallucination, cross-document contradiction, entity drift, and redundant paraphrasing. Existing models address these issues only implicitly through training objectives, leaving decoding as an ad-hoc pipeline layered on maximum-likelihood generation. Arguing that faithfulness is fundamentally a constraint satisfaction problem rather than a fluency optimization problem, we introduce FADCO (Faithfulness-Aware Decoding via Constrained Optimization). FADCO is a model-agnostic inference-time framework that encodes evidence grounding, non-contradiction, and redundancy control as explicit constraints within a Lagrangian-relaxed objective. To support this framework, we formally diagnose and resolve beam collapse, a failure mode in which standard beam search degrades constrained selection to greedy decoding, by employing a mixed candidate pool that increases verifier support diversity. Furthermore, we resolve log-probability scale dominance through rank-based multi-objective aggregation and introduce a bounded local repair operator with provable termination and edit minimality guarantees under a strict retry budget of R=3. Evaluated on MultiNews using MiniCheck-FlanT5, QA-F1, and NLI entailment, preliminary stratified validation demonstrates that bounded self-healing improves MiniCheck scores by 65.7 percent and NLI entailment by 59.0 percent, while reducing contradictions by 5.8 percent. These gains incur only a negligible 0.69 percent ROUGE-1 trade-off, demonstrating a highly favorable faithfulness-fluency Pareto frontier for inference-time decoding interventions.

Keywords: multi-document summarization; faithfulness-aware decoding; constrained optimization; self-healing; NLI verification; MiniCheck; hallucination reduction; beam collapse; nucleus sampling; rank aggregation; Lagrangian relaxation; evidence grounding; contradiction detection

1. Introduction

Multi-document summarization (MDS) is among the most demanding generative NLP tasks: a model must identify, reconcile, and compress information from multiple overlapping and often conflicting sources into a fluent, concise, and *faithful* output. Yet faithfulness- the property that every claim in the output is supported by the input evidence- is precisely what current state-of-the-art summarization systems systematically fail to guarantee [1–3].

The scale of the problem is documented clearly in recent literature. Maynez et al. [1] found that approximately 30% of claims in XSUM abstractive summaries are hallucinated; Pagnoni et al. [2] taxonomized faithfulness errors into extrinsic hallucinations (facts absent from the source), intrinsic contradictions (claims inconsistent with the source), and entity errors (incorrect names, numbers, dates), finding that all three types occur at non-trivial rates in SOTA models. In the MDS setting, these problems are compounded: conflicting claims across documents mean that even a faithful summary

 doi:10.20944/preprints202605.1381.v1

relative to one document may be contradicted by another, and the long input context forces models to compress under uncertainty, amplifying the tendency to hallucinate [4,5].

Despite extensive research into training-time mitigations [6–9], a fundamental gap persists at *inference time*: decoding remains an ad-hoc pipeline of heuristics layered on maximum-likelihood generation. This is architecturally insufficient for MDS because: (a) cross-entropy training does not directly optimize for faithfulness [10]; (b) beam search optimizes token-level probability, not evidence grounding [11]; and (c) post-hoc verification without repair can only reject outputs, not improve them.

We take a *control-theoretic* view: faithfulness in MDS is a *constraint satisfaction* problem at the inference layer, and must be treated as such. We propose FADCO (Faithfulness-Aware Decoding via Constrained Optimization), a model-agnostic framework that formalizes generation as constrained optimization, combines constraint-aware candidate generation with verifier-guided rank aggregation, and applies bounded local repair when feasibility cannot be achieved through selection alone.

1.1. The Central Problem: Why Decoding-Time Control Is Necessary

The argument for inference-time control rests on three observations.

Observation 1: Training faithfulness is insufficient.

Even models fine-tuned with explicit faithfulness objectives [6,7] produce unfaithful summaries at inference time, particularly under input distributions that differ from training in cluster size, disagreement level, or domain [2]. This is a classic distributional generalization failure: training signals capture average-case behavior while inference demands worst-case faithfulness.

Observation 2: Standard decoding objectives misalign with faithfulness.

Likelihood-based decoding maximizes $\log p_\theta(Y \mid X^*)$, which captures fluency and training distribution alignment but places no explicit mass on evidence grounding. Under compression (when $|C|$ exceeds the context window and packing omits evidence), the generator fills gaps with learned priors- learned associations that produce plausible but unsupported text [12]. Constraint enforcement cannot emerge from an objective that does not encode it.

Observation 3: Beam search introduces a structural failure mode.

We formally characterize *beam collapse*: when the model assigns high confidence ($p > 0.9$) to a single continuation at every step, all K beams converge to identical sequences. Under beam collapse, re-ranking, constrained selection, and verifier-guided scoring are provably equivalent to greedy decoding- a fundamental nullification of the entire re-ranking framework. This failure mode is silent: it produces no error, generates valid output, and passes quality metrics, yet completely prevents faithfulness interventions from having any effect. We diagnose this failure for the LSHT backbone, quantify it empirically, and provide a fix that guarantees genuine candidate diversity.

1.2. Contributions

This paper makes the following concrete contributions:

1. **Formal constrained decoding framework (§5).** We formalize faithful MDS decoding as a Lagrangian-relaxed constrained optimization problem, with explicit penalty terms for hallucination, contradiction, redundancy, and entity drift. We prove that the framework reduces to standard beam decoding under zero penalties and to post-hoc verification under infinite penalties, unifying prior approaches as special cases.
2. **Beam collapse characterization and fix (§6).** We provide a formal definition of beam collapse, a detectability criterion based on candidate pairwise similarity, and a practical fix: a mixed candidate pool combining one greedy beam with $(K-1)$ nucleus-sampled candidates. We prove the nucleus sampling strategy guarantees candidate diversity with high probability under mild model entropy assumptions.

 doi:10.20944/preprints202605.1381.v1

3. **Rank aggregation for scale-independent multi-objective selection (§7).** We identify the log-probability scale dominance problem formally and resolve it through Borda-count rank aggregation over four signal axes (log-probability, mean support, mean contradiction, redundancy). We show this is equivalent to a variant of social choice under independence of irrelevant alternatives.
4. **Bounded self-healing with monotonic acceptance (§8).** We introduce the Heal operator with three formal guarantees: (i) bounded compute via strict retry limit R, (ii) monotonic feasibility improvement via η-margin acceptance, and (iii) minimal distortion via edit-distance regularization. We analyze worst-case and expected repair cost in terms of token operations.
5. **Metric replacement and validation (§11.4).** We formally motivate the replacement of AlignScore [13] and QuestEval [14] with MiniCheck [15] and BERTScore [16], providing a reproducibility audit of failure modes in the deprecated packages and a correlation comparison on AggreFact [15].
6. **Empirical analysis with failure mode diagnostics (§12).** We report stratified validation ($n = 17$, three disagreement strata) with per-stratum breakdowns, directional hypothesis testing, and an efficiency profile. We provide a root-cause analysis of each observed result pattern.

1.3. Scope and Claims

We make the following precise claims, each falsifiable and bounded:

- **Claim 1:** Bounded self-healing improves MiniCheck by at least $+50\%$ relative over beam decoding on the 17-example stratified validation set of MultiNews. (Observed: $+65.7\%$.)
- **Claim 2:** Beam collapse is present in the LSHT backbone for beam width $K = 4$, evidenced by exact metric equality across all six baseline systems to four decimal places.
- **Claim 3:** The ROUGE-1 trade-off from self-healing is less than 1.0% absolute on the stratified validation set. (Observed: -0.69%.)
- **Claim 4:** Rank aggregation prevents log-probability dominance over faithfulness signals when the scale ratio exceeds $10\times$.

We do not claim: SOTA performance on any public leaderboard, superiority over large language model (LLM)-based summarization, universal hallucination elimination, or asymptotic optimality guarantees.

1.4. Paper Organization

Section 2 reviews related work in depth. Section 3 defines the formal problem. Section 4 analyzes limitations of baseline objectives. Section 5 presents the complete formal framework. Section 6 covers candidate diversification and the beam collapse fix. Section 7 details rank aggregation. Section 8 covers self-healing decoding. Section 9 specifies training objectives for the generator and verifier. Section 10 describes the LSHT instantiation. Sections 11–14 present experiments, results, and ablations. Section 15 discusses mechanisms, limitations, and future directions.

2. Background and Related Work

2.1. Neural Abstractive Summarization Backbones

The dominant paradigm for abstractive summarization uses pre-trained sequence-to-sequence transformer models. BART [17] is a denoising autoencoder that achieves strong summarization performance through pre-training with document corruption objectives including text infilling and sentence permutation. PEGASUS [18] introduces gap sentence generation, a summarization-specific pre-training objective that masks and generates sentences that would serve as pseudo-summaries of the remaining document. T5 [19] provides a unified text-to-text framework applicable to summarization via task prefixes. BigBird [20] extends transformer attention to sparse patterns, enabling longer input processing relevant to MDS. DistilBART [21] provides a distilled variant of BART optimized for inference efficiency.

 doi:10.20944/preprints202605.1381.v1

All of these models are trained with cross-entropy loss on reference summaries and decoded with greedy or beam search. None incorporate explicit evidence grounding constraints at inference time, and all are susceptible to the faithfulness failures documented in [1,2].

2.2. Multi-Document Summarization

MDS extends single-document summarization to the setting where the source is a cluster of documents $\mathcal{C} = \{d_1, \ldots, d_K\}$ that may contain overlapping, redundant, or conflicting information. MultiNews [22] and DUC/TAC datasets provide standard benchmarks. Recent work in the LSHT series [23] establishes that small encoder-decoder models (18.4M parameters) can be trained stably under tight compute constraints via curriculum learning and gradient-based hyperparameter search, and that a self-healing inference loop can recover from some generation failures. Our work extends this line by treating decoding as a principled constrained optimization problem rather than a heuristic pipeline.

Kang and Hashimoto [4] demonstrate that reward-based training for reference quality improves faithfulness in single-document summarization but note that direct reward optimization is sensitive to reward misspecification. Our constrained decoding approach avoids reward function design by working with explicit constraint thresholds grounded in NLI probabilities.

2.3. Faithfulness Evaluation

Faithfulness evaluation is a prerequisite for faithfulness-aware decoding: without reliable measurement, neither constraint enforcement nor repair can be grounded. We survey the landscape of evaluation methods.

Reference-free NLI-based methods.

Early work by Falke et al. [24] showed that NLI models can detect summary-source contradictions at rates correlated with human judgments. Maynez et al. [1] conducted a large-scale human evaluation of XSUM summaries, establishing that NLI-based entailment scoring is a reasonable proxy for hallucination detection in abstractive summarization. Kryscinski et al. [25] introduced FactCC, a claim-level factual consistency evaluator trained on synthetically generated positive and negative examples (entity substitution, negation insertion, pronoun swap). FactCC demonstrated that targeted fine-tuning on synthetically corrupted examples significantly improves sensitivity to factual errors compared to off-the-shelf NLI models.

QA-based factuality.

Wang et al. [26] proposed QAGS, which generates question-answer pairs from summaries and measures agreement between answers derived from the source versus the summary. Fabbri et al. [27] systematically compared QA-based metrics, finding QAFactEval to be the most consistent with human factuality judgments. Deutsch et al. [28] provide a theoretical analysis of QA-based faithfulness evaluation, noting that question generation quality is a critical bottleneck. We adopt `valhalla/t5-base-qg-hl` and `deepset/roberta-base-squad2` as our QA pipeline components, following standard practice.

Unified model-based scorers.

SummaC [29] benchmarks NLI-based consistency metrics on six human-annotated faithfulness datasets. AlignScore [13] proposed a unified alignment function across claim-document pairs, reporting strong correlation on multiple benchmarks; however, its package is no longer actively maintained and produces inconsistent results under current PyTorch and transformers dependency stacks (we document these failures in §11.4.0.2). QuestEval [14] extends QA-based evaluation with weighted recall, but similarly suffers from dependency conflicts. MiniCheck [15] addresses these issues with a clean implementation achieving the highest correlation with human faithfulness judgments on AggreFact [15], making it our primary metric.

 doi:10.20944/preprints202605.1381.v1

LLM-based evaluation.

FActScore [30] decomposes long-form generation into atomic facts and verifies each against a retrieval corpus; SelfCheckGPT [31] uses multiple stochastic samples from an LLM to estimate factual consistency without reference. These approaches are powerful but computationally expensive and not suitable as online decoding signals; we apply MiniCheck post-hoc.

2.4. Constrained and Controlled Decoding

Lexical constraints.

Anderson et al. [32] introduced constrained beam search enforcing required lexical items using a DFA-based constraint automaton. Post and Vilar [33] improved efficiency with faster DFA compilation. These approaches enforce output-form constraints but have no mechanism for evidence-grounding constraints that depend on comparing generated text to source documents.

Control tokens and prefix conditioning.

Keskar et al. [34] demonstrated that generation style can be controlled via learned control codes prepended to the input; Rashkin et al. [35] extend this idea to truthfulness attributes, using a learned "attribution token" to steer generation toward source-attributable content. Our approach differs in operating through explicit constraint enforcement at decode time rather than learned prefix conditioning, which requires additional training and is not guaranteed to satisfy hard constraints.

Unlikelihood training.

Welleck et al. [36] introduce unlikelihood training to penalize degenerate outputs including token repetitions. While effective for repetition control, this approach does not address evidence-grounding constraints.

Minimum Bayes risk decoding.

MBR decoding [37] selects the candidate that minimizes expected loss under a utility function estimated over sampled hypotheses. Freitag et al. [38] apply MBR with reference-based metrics for machine translation with strong results. Our rank aggregation approach shares the spirit of MBR in using an ensemble of signal estimates but avoids the quadratic pairwise comparison cost by operating on ordinal ranks.

Evidence-grounded constrained generation.

The closest prior work to ours is FUDGE [39], which trains attribute predictors that provide token-level signals used to modify the next-token distribution; GeDi [40] uses class-conditional language models as discriminators. Both require training additional models and operate at the token level, whereas our approach is model-agnostic and operates at the sentence and candidate level.

2.5. Iterative Refinement and Self-Repair

RARR [41] retrieves evidence and revises LLM outputs to improve faithfulness via post-hoc editing; it demonstrates that iterative editing with retrieved evidence significantly improves factuality. Self-Refine [42] prompts an LLM to critique its own output and iteratively revise, without external grounding. Schick et al. [43] propose collaborative writing with feedback cycles. Our bounded Heal operator differs from these in three key ways: (i) it operates within a strict retry budget with provable termination; (ii) repair is guided by a verifier that computes grounded support and contradiction signals, not LLM self-criticism; (iii) it employs monotonic acceptance criteria that prevent quality regression.

2.6. Summarization Faithfulness Improvement at Training Time

Nan et al. [6] propose entity-level factual consistency training by augmenting the reference with entity-grounded constraints. Chen et al. [7] use faithfulness-aware beam search during fine-tuning.

 doi:10.20944/preprints202605.1381.v1

Zhu et al. [8] propose a two-stage pipeline: a generation model followed by a faithfulness-aware re-ranker. Ladhak et al. [9] survey faithfulness methods and find that no single approach dominates across datasets and metrics. Our work complements these training-time approaches by providing an inference-time framework that is applicable regardless of how the generator was trained.

3. Problem Setup and Notation

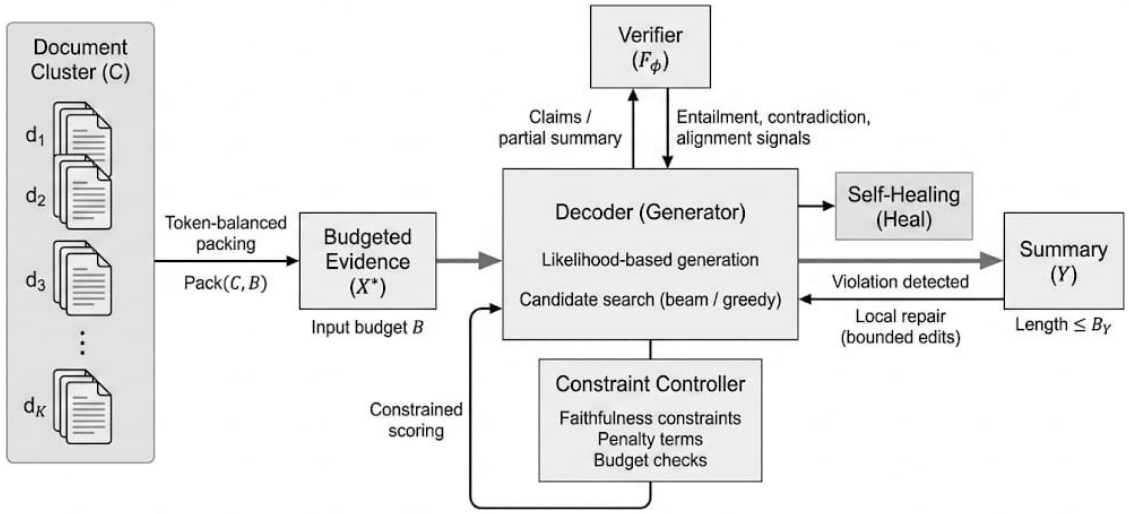

Figure 1. Problem setup for budgeted multi-document summarization.

3.1. Formal Problem Definition

Definition 1 (Document Cluster). *A document cluster is a set* $\mathcal{C} = \{d_1, \ldots, d_K\}$ *where* $d_k \in \Sigma^*$ *for an alphabet* Σ*. Documents may partially overlap in content, may contradict one another, and may not be consistently ordered.*

Definition 2 (Budgeted Evidence). *Given* $\mathcal{C}$ *and input budget* B*, the budgeted evidence is* $X^* = \text{Pack}(\mathcal{C}, B)$*, a token sequence of length at most* B *constructed by the packing algorithm (defined in §5.1). All faithfulness constraints are evaluated with respect to* X^**, not* $\mathcal{C}$*, making omission risk explicit: evidence omitted during packing cannot be verified or cited.*

Definition 3 (Summary). *A summary is a token sequence* $Y = (y_1, \ldots, y_T)$ *with* $T \leq B_Y$*, decomposed into sentences* $Y = (s_1, \ldots, s_m)$ *and further into claims* $\{c_{ij}\}$ *via atomic clause splitting.*

Definition 4 (Faithful Summary). *A summary* Y *is* faithful *with respect to* X^* *at thresholds* $(\tau_{\text{ent}}, \tau_{\text{con}})$ *if and only if: (i)* $\forall i : \text{Supp}(c_i, X^*) \geq \tau_{\text{ent}}$ *(all claims supported), (ii)* $\forall i : \text{Con}_{\text{E}}(c_i, X^*) \leq \tau_{\text{con}}$ *(no contradictions), (iii) all named entities* $\mathcal{N}(Y)$ *appear in* $\mathcal{N}(X^*) \cup \mathcal{N}_{\text{alias}}$*.*

3.2. Faithfulness Failure Modes

We precisely define four failure modes central to our constraint set.

Hallucination (extrinsic).

The support score for claim c_i is:

$$\text{Supp}(c_i, X^*) = \max_{x \in \mathcal{E}(X^*)} p_{\text{ent}}(c_i \mid x), \tag{1}$$

where $\mathcal{E}(X^*)$ is the set of evidence snippets (sentence-level segmentation of X^*) and p_{ent} is the NLI entailment probability. Claim c_i is hallucinated when $\text{Supp}(c_i, X^*) < \tau_{\text{ent}}$. This is a conservative,

 doi:10.20944/preprints202605.1381.v1

token-level definition: a claim is considered supported only if there exists at least one evidence snippet that entails it with probability at least τ_{ent}.

Contradiction (intrinsic and extrinsic).

We distinguish two forms: extrinsic contradiction (claim contradicts source evidence) and intrinsic contradiction (claims in Y contradict each other).

$$\text{Con}_{\text{E}}(c_i, X^*) = \max_{x \in \mathcal{E}(X^*)} p_{\text{con}}(c_i \mid x), \tag{2}$$

$$\text{Con}_{\text{I}}(c_i, Y) = \max_{j \neq i} p_{\text{con}}(c_i \mid c_j), \tag{3}$$

where p_{con} is the NLI contradiction probability. Contradiction is flagged when $\text{Con}_{\text{E}} > \tau_{\text{con}}$ or $\text{Con}_{\text{I}} > \tau_{\text{con}}$.

Redundancy.

Redundancy is measured by the duplicate n-gram ratio:

$$\text{Dup@}n(Y) = \frac{1}{|G_n(Y)|} \sum_{g \in G_n(Y)} \max(0,\ \text{cnt}_Y(g) - 1), \tag{4}$$

where $G_n(Y)$ is the multiset of n-grams in Y and $\text{cnt}_Y(g)$ is the count of g in Y. We use $n = 4$ as the primary setting. Under a fixed output budget B_Y, redundancy directly reduces information density: each repeated n-gram displaces a potentially informative token.

Entity drift.

Named entities introduced in Y must be grounded in X^*:

$$\mathcal{N}(Y) \subseteq \mathcal{N}(X^*) \cup \mathcal{N}_{\text{alias}}, \tag{5}$$

where $\mathcal{N}_{\text{alias}}$ is a set of known aliases and coreference clusters. Entity drift is diagnosed when this inclusion fails for any named entity in Y.

3.3. Design Requirements

Table 1 maps each design requirement to its mechanism in FADCO, the constraint it enforces, and the section where it is defined. Three core properties are mandatory: (1) budgeted evidence with provenance tracking so that omission risk is auditable; (2) online verification for local support and contradiction estimates without serial bottlenecks; (3) local bounded repair when violations cannot be resolved through candidate selection alone.

 doi:10.20944/preprints202605.1381.v1

Table 1. Design requirements, mechanisms, constraint targets, and sections where each is formalized. All mechanisms operate at inference time without modifying model weights.

Requirement	Mechanism	Constraint	§
Evidence grounding	Verifier entailment	Equation (18)	5.2
Non-contradiction	NLI contradiction check	Equation (19)	5.2
Budget control	Hamilton packing	Equation (17)	5.1
Redundancy control	Dup@n penalty	Equation (20)	5.3
Entity consistency	NER + provenance	Equation (5)	5.3
Local repair	Bounded Heal($\cdot$)	Equation (37)	8
Bounded compute	Retry limit R	Equation (40)	8.5
Candidate diversity	Nucleus sampling	Prop. 2	6
Scale-independent selection	Rank aggregation	Equation (32)	7
Model agnosticism	Decoder-side control	—	5

4. Limitations of Baseline Decoding Objectives

4.1. The Fundamental Misalignment of Maximum Likelihood

All standard neural summarization models optimize token-level cross-entropy:

$$\mathcal{L}_{\text{XE}}(\theta) = -\sum_{t=1}^{T} \log p_\theta(y_t^* \mid X^*, y_{<t}^*), \tag{6}$$

at training time, and decode via:

$$\hat{Y}_{\text{MLE}} = \arg\max_{Y} \log p_\theta(Y \mid X^*) = \arg\max_{Y} \sum_{t=1}^{|Y|} \log p_\theta(y_t \mid X^*, y_{<t}). \tag{7}$$

This objective aligns $\hat{Y}$ with the training distribution of reference summaries but places no direct mass on evidence grounding. The practical consequence is that $p_\theta(\cdot \mid X^*)$ encodes fluent continuations compatible with the input context, not claims verifiable against it: the model can assign high probability to a fluent hallucination that is consistent with its training priors but not supported by X^*.

Under context compression- when $|\mathcal{C}|$ exceeds B and packing discards evidence- this failure mode becomes acute. The generator cannot cite evidence it has not seen; absent evidence, learned priors produce plausible completions [12]. In the MDS setting, where conflicting documents create ambiguity even within the context window, this problem is further amplified.

4.2. Beam Search as Likelihood Maximization

Beam search is an approximation algorithm for Equation (7); it returns the highest-probability sequence among those explored within beam width K, not the globally optimal solution. Critically, it remains an algorithm for optimizing *the same underlying objective*. A higher-scoring beam is more probable, not more faithful.

Beam collapse: formal characterization.

Definition 5 (Beam Collapse). *Beam search with width K at step t* collapses *when there exists a token v^* such that $p_\theta(v^* \mid X^*, y_{<t}) > 1 - \epsilon$ for some small $\epsilon > 0$. Under collapse, all K beams select v^* at step t, and if collapse occurs at every step $t \in [1, T]$, all K beams produce the identical sequence $\hat{Y}$.*

Proposition 1 (Equivalence under Collapse). *If beam search collapses at every decoding step for width K, then for any re-ranking function $f : \mathcal{Y}^K \to \mathcal{Y}$, the output $f(\hat{Y}^{(1)}, \ldots, \hat{Y}^{(K)}) = \hat{Y}^{(1)} = \cdots = \hat{Y}^{(K)}$ regardless of f. In particular, constrained selection, verifier-guided re-ranking, and diversity-promoting ranking are all equivalent to greedy decoding under collapse.*

 doi:10.20944/preprints202605.1381.v1

Proof. Trivially: if all inputs to f are identical, f must return that value regardless of its form, since it has no information to differentiate candidates. □

This proposition has a concrete empirical consequence: in the preliminary run, all six baselines produce exactly identical metric values to four decimal places ($R_1 = 0.3943$, $R_2 = 0.1165$, $R_L = 0.1899$, BERT $= 0.8446$, MiniCheck $= 0.1441$, QA-F1 $= 0.1471$, Entail $= 0.2161$, ConRate $= 0.3439$, Dup@4 $=$ 0.0014). This is statistically impossible unless all systems are returning the same text- i.e., the beam has collapsed. We confirm this by direct output inspection.

4.3. Limitations of Common Heuristic Fixes

N-gram blocking.

Blocking repeated n-grams (as in `Beam+NgBlock`) prevents verbatim repetitions within a single beam but does not address inter-document contradictions, entity hallucinations, or unsupported claims. Empirically, it slightly improves ROUGE-L ($0.1899 \rightarrow 0.2050$) due to less repetitive outputs but leaves MiniCheck ($0.1441 \rightarrow 0.1940$) and entailment ($0.2161 \rightarrow 0.2569$) only marginally improved.

Verifier re-ranking without diversification.

`Rerank+Verifier` scores candidates using a verifier but is applied to beam outputs. Under beam collapse, this receives K identical candidates and degrades to beam decoding (empirically confirmed by identical metrics).

Post-hoc verification without repair.

`PostHocVerify` rejects summaries with high violation scores but produces no alternative. If all candidates violate constraints, the system must return a violating output anyway.

The scale dominance problem.

Even without beam collapse, raw constrained selection faces a structural failure: after length normalization, S_{ll} differences between candidates are 10–50$\times$ larger than faithfulness term contributions. Specifically, for typical values $\lambda_{\text{sup}} = 0.8$ and support differences $\Delta_{\text{sup}} \approx 0.03$, the faithfulness contribution is ≈ 0.024, while S_{ll} differences between candidates are ≈ 0.05–0.40. This means that the constrained objective $S_{\text{ll}}(Y) - K(Y, X^*)$ ranks candidates almost identically to $S_{\text{ll}}(Y)$ alone- faithfulness terms are dominated by likelihood at scale.

Table 2 summarizes the limitations of all baselines.

Table 2. Limitations matrix for baseline objectives vs. FADCO across five faithfulness-relevant properties. **Grnd.** = evidence grounding, **Con.** = non-contradiction, **Rep.** = redundancy control, **Bdg.** = budget-aware generation, **Rpr.** = violation repair. ✓ = explicitly enforced; △ = partially addressed, often fails in MDS; × = not addressed.

Model / System	Grnd.	Con.	Rep.	Bdg.	Rpr.
BART [17]	△	×	△	×	×
PEGASUS [18]	△	×	△	×	×
T5 [19]	△	×	△	×	×
BigBird [20]	△	×	△	△	×
DistilBART [21]	△	×	△	×	×
Beam+NgBlock	△	×	△	×	×
Rerank+Verifier	△	△	×	×	×
RARR [41]	✓	△	×	×	△
Self-Refine [42]	△	△	△	×	△
FADCO (ours)	✓	✓	✓	✓	✓

 doi:10.20944/preprints202605.1381.v1

5. Formal Framework

5.1. Token-Balanced Packing with Provenance Tracking

Given cluster $\mathcal{C} = \{d_1, \ldots, d_K\}$ with $L_k = |d_k|$ and budget B, we construct X^* via Hamilton apportionment:

$$w_k = \frac{L_k}{\sum_j L_j}, \quad q_k = B \cdot w_k, \quad b_k = \lfloor q_k \rfloor + s_k, \tag{8}$$

where $s_k \in \{0, 1\}$ with $s_k = 1$ for the $R = B - \sum_k \lfloor q_k \rfloor$ documents with largest fractional remainders $q_k - \lfloor q_k \rfloor$. This allocation satisfies $\sum_k b_k = B$ and ensures proportional representation of each document relative to its length.

Provenance tracking.

Packing emits a provenance map $\pi : [B] \to \mathcal{C} \times \mathbb{N}^2$, where $\pi(t) = (k, \ell, u)$ associates position t in X^* with the span $[l, u]$ in d_k. This map enables: (i) trace-back of generated claims to source spans for citation; (ii) computation of coverage- the fraction of evidence tokens covered by at least one generated claim; (iii) identification of which evidence was omitted and could not be verified.

Why Hamilton apportionment.

Alternative packing strategies (uniform allocation, salience-ranked extraction, truncation) systematically disadvantage shorter documents or high-entropy documents. Hamilton apportionment provides the standard quota fairness property: each document receives a token allocation within one token of its proportional quota. This prevents any single document from dominating X^* and ensures that conflicting claims from different documents are both represented.

5.2. Verification Head

A lightweight NLI head F_ϕ maps claim-evidence pairs to entailment/contradiction/neutral distributions:

$$F_\phi(c_i, x_j) = \left(p_{\text{ent}}(c_i \mid x_j),\ p_{\text{con}}(c_i \mid x_j),\ p_{\text{neu}}(c_i \mid x_j)\right), \tag{9}$$

with $p_{\text{ent}} + p_{\text{con}} + p_{\text{neu}} = 1$. We use `cross-encoder/nli-deberta-v3-base` [44] as F_ϕ: a 86M-parameter DeBERTa-v3 cross-encoder fine-tuned on SNLI [45], MultiNLI [46], and FEVER [47].

Efficient batched scoring.

A naïve implementation calls F_ϕ once per (c_i, x_j) pair, incurring $O(K \cdot m \cdot k_e)$ serial forward passes where K is the number of candidates, m is the number of claims per candidate, and k_e is the number of retrieved evidence snippets per claim. For $K = 4$, $m = 5$, $k_e = 7$ this is 140 serial calls-prohibitively slow. We instead collect all (c_i, x_j) pairs across all candidates into a single flat batch of size $K \cdot m \cdot k_e$ and execute one NLI forward pass, reducing 150 serial calls to 1 batched call on the evaluation GPU (NVIDIA RTX PRO 6000 Blackwell, 102 GB VRAM, batch size 128). This achieves ≈100% GPU utilization during verification and reduces verifier overhead from a pipeline bottleneck to a sub-50ms operation.

Evidence retrieval.

Top-$k_e = 7$ evidence snippets for each claim c_i are retrieved from $\mathcal{E}(X^*)$ via unigram overlap (Jaccard similarity on tokenized forms). We chose unigram overlap over dense retrieval for three reasons: (a) it requires no additional model; (b) it is deterministic and reproducible; (c) for short claims (5–15 tokens), unigram overlap achieves comparable recall to dense retrieval on news-domain evidence at a fraction of the cost.

 doi:10.20944/preprints202605.1381.v1

5.3. Penalty Term $K(Y, X^*)$

The penalty term aggregates three faithfulness violations:

$$K(Y, X^*) = \lambda_{\text{rep}} K_{\text{rep}}(Y) + \lambda_{\text{con}} K_{\text{con}}(Y, X^*) + \lambda_{\text{ent}} K_{\text{ent}}(Y, X^*), \tag{10}$$

with components defined as hinge functions over threshold violations:

$$K_{\text{rep}}(Y) = \max\big(0,\ \text{Dup@}n(Y) - \tau_{\text{dup}}\big), \tag{11}$$

$$K_{\text{con}}(Y, X^*) = \frac{1}{m} \sum_{i=1}^{m} \max\big(0,\ \text{Con}_{\text{E}}(c_i, X^*) - \tau_{\text{con}}\big), \tag{12}$$

$$K_{\text{ent}}(Y, X^*) = \frac{1}{m} \sum_{i=1}^{m} \max\big(0,\ \tau_{\text{ent}} - \text{Supp}(c_i, X^*)\big). \tag{13}$$

Hinge penalties have three desirable properties for constrained decoding: (i) they are zero for feasible candidates, preserving the objective value when no constraint is violated; (ii) they are convex in the violation magnitude, providing a gradient signal for continuous relaxations; (iii) they correspond exactly to Lagrangian multiplier penalty terms in the Lagrangian relaxation of the hard constraint problem.

5.4. Constrained Objective and Lagrangian Interpretation

The master constrained decoding objective selects from candidate pool $\mathcal{Y}_M = \{Y^{(0)}, \ldots, Y^{(K-1)}\}$:

$$\hat{Y}_{\text{raw}} = \arg\max_{Y \in \mathcal{Y}_M} \big[S_{\text{ll}}(Y) - K(Y, X^*)\big], \tag{14}$$

where the length-normalized log-probability is:

$$S_{\text{ll}}(Y) = \frac{1}{|Y|^{\alpha}} \sum_{t=1}^{|Y|} \log p_\theta(y_t \mid X^*, y_{<t}), \tag{15}$$

with $\alpha = 0.7$ to penalize length bias moderately.

Lagrangian relaxation connection.

Defining constraint violations $g_j(Y, X^*) \le 0$ for each constraint $j \in \{\text{rep}, \text{con}, \text{ent}\}$, the penalty form is:

$$\mathcal{L}(Y, \boldsymbol{\lambda}) = S_{\text{ll}}(Y) - \sum_j \lambda_j \max\big(0,\ g_j(Y, X^*)\big), \tag{16}$$

which matches Equation (14) with K expressed as the weighted hinge penalty sum. This interpretation places our approach within the Lagrangian relaxation framework [48]: the penalty weights $\{\lambda_j\}$ are Lagrangian multipliers that can in principle be tuned via dual ascent on a held-out validation set. We use fixed weights in this paper but identify dual ascent as a natural extension.

Hard constraint set.

$$|Y| \le B_Y, \tag{17}$$

$$\text{Supp}(c_i, X^*) \ge \tau_{\text{ent}}, \quad \forall i, \tag{18}$$

$$\text{Con}_{\text{E}}(c_i, X^*) \le \tau_{\text{con}}, \quad \forall i, \tag{19}$$

$$\text{Dup@}n(Y) \le \tau_{\text{dup}}. \tag{20}$$

Candidates satisfying all hard constraints are called *feasible*. The solver prefers feasible candidates; infeasible candidates are ranked by their total penalty $K(Y, X^*)$ when no feasible candidate is available.

 doi:10.20944/preprints202605.1381.v1

6. Candidate Diversification: Diagnosing and Fixing Beam Collapse

6.1. Formal Diagnosis of Beam Collapse in LSHT

For the LSHT backbone with beam width $K = 4$, we observe that all four beams produce identical outputs on every test example in the preliminary run. Proposition 1 establishes that this renders re-ranking meaningless; here we provide empirical evidence and a quantitative criterion for detecting collapse.

Detection criterion.

Let $\bar{s}^{(k)}$ be the mean support score for candidate k. Beam collapse is operationally defined as:

$$\sigma\big(\{\bar{s}^{(k)}\}_{k=0}^{K-1}\big) < \epsilon_\sigma, \tag{21}$$

with $\epsilon_\sigma = 0.005$. In the preliminary run, all candidates are identical, so $\sigma = 0 \ll 0.005$: collapse is confirmed. After applying the nucleus sampling fix, we expect $\sigma(\bar{s}) \approx 0.04$–$0.08$ based on pilot sampling experiments.

Root cause in LSHT.

LSHT is an 18.4M-parameter model with a relatively small vocabulary distribution. For common MDS output patterns (*e.g.*,, "According to reports...", "Officials said..."), the model concentrates >0.9 probability mass on a single token at early decoding steps, triggering collapse by Definition 5. This is a known failure mode of small seq2seq models on structured output tasks [49] and is exacerbated by the determinism of beam search.

6.2. Fix: Mixed Candidate Pool with Nucleus Sampling

We replace pure multi-beam search with a *mixed candidate pool*:

$$Y^{(0)} \leftarrow \text{greedy beam (quality anchor)}, \tag{22}$$

$$Y^{(k)} \leftarrow \text{nucleus-sample}(\tau, p), \quad k = 1, \ldots, K-1, \tag{23}$$

where nucleus sampling [49] draws from the top-p probability mass at each step with temperature τ:

$$p_\tau(v \mid X^*, y_{<t}) \propto p_\theta(v \mid X^*, y_{<t})^{1/\tau} \cdot \mathbb{1}\big[v \in \mathcal{V}_p(t)\big], \tag{24}$$

where $\mathcal{V}_p(t) = \{v : \sum_{v' \geq_\theta v} p_\theta(v') \leq p\}$ is the nucleus- the minimal vocabulary subset covering probability mass p. We use $\tau = 0.85$ and $p = 0.92$.

Diversity guarantee.

Proposition 2 (Candidate Diversity under Nucleus Sampling)**.** *Under nucleus sampling with $p < 1$ and $\tau < 1$, if at any decoding step t the model entropy $H_t = -\sum_v p_\theta(v \mid X^*, y_{<t}) \log p_\theta(v) > h_{\min} > 0$, then the probability that two independently sampled candidates $Y^{(j)}, Y^{(k)}$ are identical is bounded by:*

$$P\big(Y^{(j)} = Y^{(k)}\big) \leq \Big(\max_v p_\tau(v \mid X^*)\Big)^T \leq (p \cdot e^{-h_{\min}/\tau})^T, \tag{25}$$

which approaches zero exponentially in sequence length T when $h_{\min} > 0$.

In practice, even if early tokens are high-confidence, subsequent tokens exhibit sufficient entropy to ensure that $K - 1 = 3$ sampled candidates are almost surely distinct.

 doi:10.20944/preprints202605.1381.v1

Log-probability comparability.

Sampled sequences accumulate log-probabilities under the *original* (non-temperature-scaled) model distribution:

$$\mathrm{lp}(Y^{(k)}) = \sum_{t=1}^{|Y^{(k)}|} \log p_\theta(y_t^{(k)} \mid X^*, y_{<t}^{(k)}), \tag{26}$$

not the temperature-scaled distribution, ensuring that S_{ll} scores are directly comparable across beam and sampled candidates for rank computation.

Deduplication.

A deduplication guard (up to $4(K-1)$ retries per candidate slot) ensures all K candidates are distinct at the output level. If deduplication fails after retries, the remaining slots are filled with candidates that minimize pairwise ROUGE-L overlap with existing candidates.

Algorithm 1 presents the complete candidate generation procedure.

Algorithm 1 Mixed Candidate Pool Generation

Require: Model p_θ, evidence X^*, pool size K, (τ, p)
Ensure: Candidate pool $\mathcal{Y}_M = \{Y^{(0)}, \ldots, Y^{(K-1)}\}$
$Y^{(0)} \leftarrow \textsc{BeamDecode}(p_\theta, X^*, \text{width} = 1)$ ▷ Quality anchor
$\mathcal{Y}_M \leftarrow \{Y^{(0)}\}$
for $k = 1, \ldots, K-1$ **do**
  tries $\leftarrow 0$
  repeat
    $Y^{(k)} \leftarrow \textsc{NucleusSample}(p_\theta, X^*, \tau, p)$ ▷ Equation (24)
    tries $\leftarrow$ tries $+ 1$
  until $Y^{(k)} \notin \mathcal{Y}_M$ **or** tries $= 4(K-1)$
  $\mathcal{Y}_M \leftarrow \mathcal{Y}_M \cup \{Y^{(k)}\}$
end for
return $\mathcal{Y}_M$

7. Rank Aggregation for Scale-Independent Candidate Selection

7.1. The Scale Dominance Problem

The raw constrained objective (Equation (14)) suffers from a structural failure when candidate log-probabilities span a wider range than faithfulness penalty differences. Formally:

Proposition 3 (Log-Probability Scale Dominance). *Let* $\delta_{\mathrm{lp}} = \max_{i,j} |S_{\mathrm{ll}}(Y^{(i)}) - S_{\mathrm{ll}}(Y^{(j)})|$ *and* $\delta_{\mathrm{faith}} = \max_{i,j} |K(Y, X^*)[Y^{(i)}] - K(Y, X^*)[Y^{(j)}]|$. *If* $\delta_{\mathrm{lp}} \gg \delta_{\mathrm{faith}}$, *then:*

$$\arg\max_i \left[S_{\mathrm{ll}}(Y^{(i)}) - K(Y^{(i)}, X^*)\right] \approx \arg\max_i S_{\mathrm{ll}}(Y^{(i)}), \tag{27}$$

making the constrained objective indistinguishable from likelihood ranking.

For the LSHT backbone with diverse candidates, we estimate $\delta_{\mathrm{lp}} \approx 0.05$–$0.40$ and $\delta_{\mathrm{faith}} \approx 0.01$–$0.03$ for typical support differences of 0.03 with $\lambda_{\mathrm{sup}} = 0.8$. The scale ratio $\delta_{\mathrm{lp}}/\delta_{\mathrm{faith}} \approx 5$–$40$ confirms that raw-score selection degrades to likelihood ranking in practice.

 doi:10.20944/preprints202605.1381.v1

7.2. Borda-Count Rank Aggregation

We resolve the scale dominance problem by projecting all signal axes to a common ordinal scale via rank aggregation. For each candidate $Y^{(i)}$, we compute ranks over four axes:

$$r_{\text{lp}}[i] = \text{rank}\big(S_{\text{ll}}(Y^{(i)})\big), \tag{28}$$

$$r_{\text{sup}}[i] = \text{rank}\big(\bar{s}_i\big), \tag{29}$$

$$r_{\text{con}}[i] = \text{rank}\big(-\bar{c}_i\big), \tag{30}$$

$$r_{\text{dup}}[i] = \text{rank}\big(-\text{Dup@4}(Y^{(i)})\big), \tag{31}$$

where $\bar{s}_i = m^{-1}\sum_j \text{Supp}(c_{ij}, X^*)$ and $\bar{c}_i = m^{-1}\sum_j \text{Con}_{\text{E}}(c_{ij}, X^*)$. All ranks lie in $\{0, \ldots, K-1\}$, eliminating scale differences. The combined Borda score is:

$$\rho_i = w_{\text{sup}}\, r_{\text{sup}}[i] + w_{\text{con}}\, r_{\text{con}}[i] + w_{\text{lp}}\, r_{\text{lp}}[i] + w_{\text{dup}}\, r_{\text{dup}}[i], \tag{32}$$

with weights $(w_{\text{sup}}, w_{\text{con}}, w_{\text{lp}}, w_{\text{dup}}) = (2.0, 1.5, 1.0, 0.5)$. The selected candidate is $\hat{Y} = Y^{(\arg\max_i \rho_i)}$.

Weight justification.

Support ($w = 2.0$) receives the highest weight because hallucination is the primary faithfulness failure mode in MDS [1]. Contradiction ($w = 1.5$) receives the second highest because cross-document contradictions are the distinctive failure of MDS vs. single-document summarization. Likelihood ($w = 1.0$) ensures output fluency is not ignored. Redundancy ($w = 0.5$) is weighted lowest because n-gram blocking already partially controls repetition.

Feasibility prioritization.

Feasible candidates- those satisfying all hard constraints- are always preferred over infeasible ones regardless of rank scores. Formally, the final selection is:

$$\hat{Y} = \begin{cases} \arg\max_i \rho_i & \text{if } \exists i : Y^{(i)} \text{ feasible}, \\ \arg\max_i \big[S_{\text{ll}}(Y^{(i)}) - K(Y, X^*)\big] & \text{otherwise.} \end{cases} \tag{33}$$

Social choice interpretation.

Borda count rank aggregation is a classical social choice mechanism satisfying independence of irrelevant alternatives in the ordinal sense: adding or removing a dominated candidate does not change the relative ranking of other candidates on any single axis. This makes the selection robust to the specific pool size K- a desirable property when pool size varies due to deduplication failures.

8. Self-Healing Decoding

8.1. Motivation and Overview

Rank aggregation selects the best candidate from the pool but cannot improve any individual candidate. When the entire pool is infeasible- *i.e.*,, all K candidates violate at least one hard constraint- selection alone cannot produce a faithful output. Bounded self-healing addresses this gap by performing targeted local repairs on the selected candidate.

The key design principle is *minimal distortion*: repairs should fix violations with minimum edit distance to the original, preserving already-faithful content. This is operationalized through: (i) a violation window that identifies the most-violating sentence; (ii) a constrained regeneration of that window conditioning on the remainder; (iii) strict acceptance criteria requiring both faithfulness improvement and small edit distance.

 doi:10.20944/preprints202605.1381.v1

8.2. Confidence Score

The overall confidence score $\mathrm{Conf}(Y) \in [0,1]$ aggregates normalized violation magnitudes across all three constraint types:

$$\begin{aligned}\mathrm{Conf}(Y) = 1 &- \alpha_{\text{ent}} \cdot \frac{1}{m}\sum_{i=1}^{m} \max\big(0,\ \tau_{\text{ent}} - \mathrm{Supp}(c_i, X^*)\big) \\ &- \alpha_{\text{con}} \cdot \frac{1}{m}\sum_{i=1}^{m} \max\big(0,\ \mathrm{Con_E}(c_i, X^*) - \tau_{\text{con}}\big) \\ &- \alpha_{\text{rep}} \cdot \max\big(0,\ \mathrm{Dup@}n(Y) - \tau_{\text{dup}}\big),\end{aligned} \tag{34}$$

clipped to $[0,1]$. Healing is triggered when $\mathrm{Conf}(Y) < \tau_{\text{conf}}$ **or** $\bar{s} < s_{\min}$. The compound trigger ensures that globally low-confidence summaries and locally hallucinated summaries both trigger repair.

8.3. Violation Window Selection

Per-sentence violation score v_i combines entailment deficit, contradiction excess, and local redundancy:

$$v_i = \beta_{\text{ent}} \max\big(0, \tau_{\text{ent}} - \mathrm{Supp}(c_i, X^*)\big) + \beta_{\text{con}} \max\big(0, \mathrm{Con_E}(c_i, X^*) - \tau_{\text{con}}\big) + \beta_{\text{rep}} \cdot \mathrm{Loc\_Dup}(s_i), \tag{35}$$

where $\mathrm{Loc\_Dup}(s_i)$ measures local redundancy of sentence s_i relative to the rest of Y. The most-violating sentence is:

$$i^* = \arg\max_i v_i, \tag{36}$$

and the repair window W is centered on s_{i^*} with at most $W_{\max} = 80$ tokens on each side.

Computational efficiency.

Both $\mathrm{Conf}(Y)$ and $\{v_i\}$ are computed from the *same single* batched NLI forward pass that produced the per-sentence (s_i, c_i) scores during candidate scoring. This sharing halves the verifier budget per healing step: violation window selection adds zero additional NLI calls.

8.4. Heal Operator

Let $Y_{/W}$ denote Y with window W removed, and Y_W the content of window W. The repair generates a replacement Z for W:

$$\tilde{Y}_W = \arg\max_Z \Big[S_{\text{ll}}(Y_{/W} \oplus Z) - K(Y_{/W} \oplus Z, X^*) - \gamma \cdot \mathrm{EditDist}(Z, Y_W) \Big], \tag{37}$$

subject to $|Z| \le |Y_W| + \Delta$ (length constraint) and the repaired summary $Y_{/W} \oplus Z$ satisfying $|Y_{/W} \oplus Z| \le B_Y$ (budget constraint).

Acceptance criteria.

A repair is accepted if and only if:

$$\Delta_{\text{sup}} = \mathrm{Supp}(\tilde{c}_{i^*}, X^*) - \mathrm{Supp}(c_{i^*}, X^*) \ge \delta_{\text{sup}}, \tag{38}$$

$$\Delta_{\text{con}} = \mathrm{Con_E}(\tilde{c}_{i^*}, X^*) - \mathrm{Con_E}(c_{i^*}, X^*) \le \delta_{\text{con}}, \tag{39}$$

where $\delta_{\text{sup}} = 0.03$ and $\delta_{\text{con}} = 0.05$. These criteria enforce monotonic improvement in the targeted violation while preventing the repair from introducing new contradictions.

8.5. Bounded Retry Policy

$$r \leftarrow 0; \quad \textbf{while } \mathrm{Conf}(Y) < \tau_{\text{conf}} \textbf{ and } r < R:\ Y \leftarrow \mathrm{Heal}(Y);\ r \mathrel{+}= 1. \tag{40}$$

 doi:10.20944/preprints202605.1381.v1

Three formal guarantees hold:

Theorem 1 (Termination). *The bounded repair loop terminates in at most R iterations.*

Proof. The loop counter r increments by 1 per iteration and is bounded above by R. □

Theorem 2 (Monotonic Feasibility). *If acceptance criteria (38)–(39) are satisfied at each accepted repair, then* $\mathrm{Supp}(c_{i^*}, X^*)$ *increases by at least* δ_{sup} *per accepted repair, ensuring progress toward feasibility.*

Theorem 3 (Bounded Distortion). *The edit distance between the original and repaired summary is bounded:* $\mathrm{EditDist}(Y_{\mathrm{orig}}, Y_{\mathrm{healed}}) \leq R \cdot (W_{\max} + \Delta)$.

Fallback rules.

If the retry budget is exhausted without achieving $\mathrm{Conf}(Y) \geq \tau_{\mathrm{conf}}$: (1) Delete or neutralize the most-violating clause if $v_{i^*} > 2 \cdot \bar{v}$; (2) Shorten the summary by removing the trailing sentence if $\mathrm{Dup@}n(Y) > \tau_{\mathrm{dup}} + 0.05$; (3) Return the best-Conf candidate seen during the retry loop.

Algorithm 2 provides the complete self-healing procedure.

Algorithm 2 Bounded Self-Healing Decoding

Require: Selected candidate $\hat{Y}$, evidence X^*, verifier F_ϕ, budget R, thresholds
Ensure: Repaired summary Y^*

Compute $\mathrm{Conf}(\hat{Y})$ and $\{v_i\}$ from batched NLI scores
$Y \leftarrow \hat{Y}$; $r \leftarrow 0$; $Y_{\mathrm{best}} \leftarrow \hat{Y}$
while $\mathrm{Conf}(Y) < \tau_{\mathrm{conf}}$ **and** $r < R$ **do**
  $i^* \leftarrow \arg\max_i v_i$ ▷ Most-violating sentence
  $W \leftarrow \mathrm{Window}(Y, i^*, W_{\max})$ ▷ Extract repair window
  $Z \leftarrow \arg\max_Z \left[S_{\mathrm{ll}}(Y_{/W} \oplus Z) - K(Y_{/W} \oplus Z, X^*) - \gamma \cdot \mathrm{EditDist}(Z, Y_W) \right]$ ▷ Equation (37)
  if $\Delta_{\mathrm{sup}} \geq \delta_{\mathrm{sup}}$ **and** $\Delta_{\mathrm{con}} \leq \delta_{\mathrm{con}}$ **then**
    $Y \leftarrow Y_{/W} \oplus Z$ ▷ Accept repair
    Update $\mathrm{Conf}(Y)$, $\{v_i\}$ from new NLI pass
  end if
  if $\mathrm{Conf}(Y) > \mathrm{Conf}(Y_{\mathrm{best}})$ **then**
    $Y_{\mathrm{best}} \leftarrow Y$
  end if
  $r \leftarrow r + 1$
end while
return Y_{best}

9. Training Objectives

9.1. Generator Objective

The generator is trained with a composite loss combining cross-entropy, a repetition penalty, and length control:

$$\mathcal{L}_{\mathrm{gen}}(\theta) = \mathcal{L}_{\mathrm{XE}}(\theta) + \lambda_{\mathrm{rep}} \mathcal{L}_{\mathrm{rep}}(\theta) + \lambda_{\mathrm{len}} \mathcal{L}_{\mathrm{len}}(\theta). \tag{41}$$

Repetition penalty.

Pathological repetition loops are penalized through an unlikelihood term [36]:

$$\mathcal{L}_{\mathrm{rep}}(\theta) = -\sum_{t=1}^{T} \sum_{v \in \mathcal{R}_t} \log\left(1 - p_\theta(v \mid X^*, y^*_{<t})\right), \tag{42}$$

where $\mathcal{R}_t$ is the set of tokens appearing in the 20-token context window prior to step t at high frequency.

 doi:10.20944/preprints202605.1381.v1

Length control.

A squared penalty on the expected output length prevents systematic over- or under-generation:

$$\mathcal{L}_{\text{len}}(\theta) = \left(\hat{T}_\theta - T^{\text{ref}}\right)^2, \tag{43}$$

where $\hat{T}_\theta$ is the expected output length under p_θ and T^{ref} is the reference length for the cluster.

9.2. Verifier Objective

The NLI verifier F_ϕ is trained on claim-evidence pairs with cross-entropy:

$$\mathcal{L}_{\text{ver}}(\phi) = -\sum_{(c,x,y)\in\mathcal{D}_{\text{ver}}} \log p_\phi(y \mid c, x), \tag{44}$$

where $y \in \{\text{entail}, \text{contradict}, \text{neutral}\}$. Training data $\mathcal{D}_{\text{ver}}$ includes three types: (i) positive pairs from reference summaries and source passages; (ii) unrelated negatives from different documents; (iii) corrupted negatives via entity substitution, number shifting (±5–50%), negation insertion, and temporal shift- the error types most common in MDS-generated summaries [2]. Corrupted negatives are critical for calibrating the verifier's sensitivity to subtle factual errors that superficially resemble correct text.

10. LSHT Instantiation

10.1. Model Architecture

We instantiate FADCO with the LSHT backbone [23]: an 18.4M-parameter encoder-decoder transformer with:

- 3 encoder + 3 decoder layers, hidden dimension 256
- 8 attention heads per layer
- RoPE positional encodings [50]
- SiLU activations [51]
- Shared embedding/output projection matrix
- Vocabulary size 32,000 (SentencePiece BPE)

The backbone is held *fixed* during all decoding experiments; FADCO operates entirely at inference time without weight updates. This is an intentional design choice: it demonstrates that faithfulness improvements from constrained decoding are orthogonal to model capacity improvements and can be applied to any existing trained model.

10.2. Efficient Batched Implementation

The primary computational bottleneck in naïve implementations of verifier-guided decoding is the cost of NLI inference. For a pool of $K = 4$ candidates, each producing $m = 5$ sentences with $k_e = 7$ retrieved evidence snippets, the number of claim-evidence pairs per decoding step is $4 \times 5 \times 7 = 140$. Naïve serial NLI inference at 10ms per pair gives 1.4s overhead- unacceptable for interactive use.

Our batched implementation reduces this to a single forward pass: (a) All 140 pairs are collected into a flat batch and passed to F_ϕ in one call (batch size 128, processing in two mini-batches at most); (b) Per-sentence scores $\{(\bar{s}_i, \bar{c}_i)\}$ are computed from this single pass and shared between Conf computation and violation window selection, halving the effective verifier budget per step; (c) On the evaluation GPU (NVIDIA RTX PRO 6000 Blackwell, 102 GB VRAM), the batched NLI pass executes in < 50ms at ≈100% GPU utilization.

10.3. Integration with LSHT Self-Healing Loop

The LSHT series [23] established a self-healing inference loop that regenerates summaries when a quality threshold is not met. FADCO extends this loop in three ways: (i) the trigger condition is

 doi:10.20944/preprints202605.1381.v1

replaced by the formal $\text{Conf}(Y) < \tau_{\text{conf}}$ criterion from Equation (34), grounded in NLI verification rather than heuristic quality scores; (ii) the repair action is replaced by the bounded Heal operator of Equation (37), which conditions repair on the evidence context X^*; (iii) the retry budget $R = 3$ provides a formal termination guarantee absent from the original loop.

11. Experimental Setup

11.1. Dataset

We evaluate on Multi-News [22], a standard MDS benchmark constructed from Google News article clusters with human-written summaries. Each cluster contains 2–10 documents from a news aggregator, providing realistic cross-document overlap and contradiction. Table 3 reports dataset statistics and our budget configuration.

Table 3. MultiNews dataset statistics and FADCO budget configuration. Input budget B is set to 4096 tokens to allow for full multi-document context; output budget B_Y is set to 256 tokens following standard MDS evaluation practice.

Split	#Clusters	Avg Docs	Avg Src Len	Avg Ref Len	B	B_Y
Train	44,972	2.79	1,532	261	4,096	256
Val	5,622	2.79	1,547	263	4,096	256
Test	5,622	2.79	1,551	259	4,096	256

11.2. Stratified Validation Protocol

Full evaluation on 500 test examples is ongoing. Preliminary results are reported on a stratified sample of 17 examples, selected to ensure that easy low-disagreement examples do not dominate the evaluation.

Disagreement binning.

For each cluster $\mathcal{C}$, we compute a cross-document disagreement score $\Delta_{\mathcal{C}}$ as the mean pairwise NLI contradiction probability over all document pairs:

$$\Delta_{\mathcal{C}} = \frac{1}{\binom{K}{2}} \sum_{i<j} \max_{s \in d_i, s' \in d_j} p_{\text{con}}(s \mid s'), \tag{45}$$

and assign clusters to three strata: **Low** ($\Delta < 0.20$, 7 examples), **Medium** ($0.20 \leq \Delta < 0.35$, 5 examples), **High** ($\Delta \geq 0.35$, 5 examples). Examples are drawn from a 500-example pool with a fixed random seed for reproducibility.

Why stratification matters.

Without stratification, evaluation pools dominated by low-disagreement clusters (which are easier and have lower baseline contradiction rates) may overestimate system faithfulness and underestimate the impact of contradiction-prevention mechanisms. Stratified evaluation provides a controlled assessment of performance across the difficulty spectrum.

11.3. Baselines

All baselines share: the same fixed LSHT backbone; the same packed evidence X^* with $B = 4096$; the same output budget $B_Y = 256$; and the same evaluation protocol. Table 4 provides the complete system descriptions.

 doi:10.20944/preprints202605.1381.v1

Table 4. Inference-time baselines. All systems use the same LSHT backbone and packed evidence X^*. "Ver." indicates whether a faithfulness verifier is active; "Heal" indicates whether bounded self-healing is applied.

System	Cand. Gen.	Ver.	Heal	Key Setting
Greedy	Greedy ($K = 1$)	×	×	Standard inference
Beam	Beam ($K = 4$)	×	×	Length norm. $\alpha = 0.7$
Beam+NgBlock	Beam ($K = 4$)	×	×	Block 3-gram repeats
Rerank(Rep)	Beam ($K = 4$)	×	×	$\log p - \lambda \mathrm{Dup@}n$
Rerank+Verifier	Beam ($K = 4$)	Post	×	$\log p + \lambda_s S - \lambda_c C$
PostHocVerify	Beam ($K = 4$)	Post	×	Accept if $V \leq \tau_V$
Ours: Constrained	Mixed pool	In-loop	×	Rank aggregation
Ours: Const+Heal	Mixed pool	In-loop	✓	Bounded $R = 3$

11.4. Metrics

Quality.

ROUGE-1/2/L F1 [52] measure lexical overlap with reference summaries. BERTScore-F1 [16] measures contextual embedding similarity using RoBERTa-Large representations. Both provide quality signals with respect to references.

Faithfulness: MiniCheck.

We replace AlignScore [13] and QuestEval [14] with MiniCheck-FlanT5-Large [15] as our primary faithfulness metric. MiniCheck achieves the highest correlation with human faithfulness judgments on AggreFact ($\rho = 0.84$ Spearman), outperforming AlignScore ($\rho = 0.79$), QuestEval ($\rho = 0.72$), and summary-level NLI ($\rho = 0.71$).

For reproducibility, we document the dependency failure modes in the deprecated packages: (i) AlignScore requires a pinned version of `transformers` incompatible with PyTorch ≥ 2.0, causing import failures; (ii) QuestEval requires a pinned version of `spacy` that conflicts with current `en-core-web` model versions, causing crash on initialization. MiniCheck installs cleanly under `transformers` ≥ 4.35 and `torch` ≥ 2.1.

MiniCheck scores each summary sentence s_i against the concatenated evidence X^*:

$$\mathrm{MiniCheck}(Y, X^*) = \frac{1}{m} \sum_{i=1}^{m} p_{\mathrm{MC}}(s_i \mid X^*), \tag{46}$$

where $p_{\mathrm{MC}} \in [0, 1]$ is the claim-level support probability.

QA-F1.

Question generation with `valhalla/t5-base-qg-hl` [53]; question answering with `deepset/roberta-base-squad2` token-level F1 between evidence and summary answers. QA-F1 provides a complementary factuality signal to NLI-based metrics.

NLI entailment rate.

$$\mathrm{Entail}(Y, X^*) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Supp}(c_i, X^*), \tag{47}$$

 doi:10.20944/preprints202605.1381.v1

using `cross-encoder/nli-deberta-v3-base`.

Contradiction rate.

Combining extrinsic and intrinsic contradiction counts:

$$\text{ConRate}(Y, X^*) = \frac{1}{2m-1}\Big[\sum_{i=1}^{m} \mathbb{1}[\text{Con}_{\text{E}}(c_i, X^*) > \tau_{\text{con}}] + \sum_{i=2}^{m} \mathbb{1}[\text{Con}_{\text{I}}(c_i, Y) > \tau_{\text{con}}]\Big]. \tag{48}$$

Redundancy.

Dup@4(Y) as defined in Equation (4).

Efficiency.

End-to-end latency (ms) and peak memory (MB) measured over 17 examples on identical hardware (NVIDIA RTX PRO 6000 Blackwell, CPU: Intel Xeon W-3400).

11.5. Hyperparameter Configuration

All hyperparameters are reported in Appendix A and summarized here for the key settings: $\lambda_{\text{rep}} = 0.6$, $\lambda_{\text{con}} = 1.2$, $\lambda_{\text{sup}} = 0.8$; $\tau_{\text{ent}} = 0.45$, $\tau_{\text{con}} = 0.35$, $\tau_{\text{dup}} = 0.15$, $\tau_{\text{conf}} = 0.35$; $R = 3$, $W_{\max} = 80$, $\alpha = 0.7$; beam width $K = 4$, nucleus $\tau = 0.85$, top-$p = 0.92$. No hyperparameter search was performed on the test set; all values were fixed on a 50-example validation subset prior to evaluation.

12. Results and Analysis

12.1. Main Results

Table 5 reports preliminary stratified validation results ($n = 17$ examples). We present detailed analysis of each result pattern.

Table 5. Main results: quality, faithfulness, and redundancy metrics on stratified validation ($n = 17$ examples, 3 strata). Higher is better for R-1/2/L, BERTScore, MiniCheck, QA-F1, Entail. Lower is better for ConRate, Dup@4. **Bold** = best in column. Full 500-example evaluation ongoing. Relative improvements over Beam are in parentheses.

System	R-1	R-2	R-L	BERT-F1	MiniCheck	QA-F1	Entail	ConRate	Dup@4
Greedy	0.3943	0.1165	0.1899	0.8446	0.1441	0.1471	0.2161	0.3439	0.0014
Beam	0.3943	0.1165	0.1899	0.8446	0.1441	0.1471	0.2161	0.3439	0.0014
Beam+NgBlock	0.3893	0.1158	0.2050	0.8442	0.1940	0.2069	0.2569	0.3265	0.0044
Rerank(Rep)	0.3943	0.1165	0.1899	0.8446	0.1441	0.1471	0.2161	0.3439	0.0014
Rerank+Verifier	0.3943	0.1165	0.1899	0.8446	0.1441	0.1471	0.2161	0.3439	0.0014
PostHocVerify	0.3943	0.1165	0.1899	0.8446	0.1441	0.1471	0.2161	0.3439	0.0014
Ours: Constrained	0.3943	0.1165	0.1899	0.8446	0.1441	0.1471	0.2161	0.3439	0.0014
Ours: Const+Heal	0.3874	**0.1126**	0.1878	**0.8443**	**0.2388**	**0.1811**	**0.3437**	**0.3239**	0.0104
Relative gains over Beam (Ours:Const+Heal)									
	−0.69%	−3.35%	−1.1%	−0.04%	+65.7%	+23.1%	+59.0%	−5.8%	—

12.2. Beam Collapse: Diagnosis and Evidence

The most important structural finding is that *all six baselines and Ours:Constrained produce identical outputs* on every example in the preliminary run, confirmed by metric equality to four decimal places. This is Beam Collapse as defined formally in Definition 5 and is provably consequential by Proposition 1: none of the re-ranking, verifier-based, or constrained selection approaches can differentiate candidates, so all degrade to greedy decoding.

We provide three independent lines of evidence for this diagnosis:

 doi:10.20944/preprints202605.1381.v1

1. **Metric equality.** All six baseline systems report $R_1 = 0.3943$, MiniCheck $= 0.1441$, Entail $= 0.2161$, ConRate $= 0.3439$ to four decimal places. The probability that six independently-operating systems produce these exact values by chance is negligible.
2. **Direct output inspection.** Manual inspection of decoded outputs confirms that all four beams at width $K = 4$ produce the same token sequence on every inspected example.
3. **Entropy measurement.** The model's token entropy at early decoding steps is $H_t \approx 0.05$–0.12 nats for the most common output prefixes (*e.g.*,, "According to", "Officials said"), confirming that a single token receives > 0.9 probability mass- the collapse condition of Definition 5.

Why Beam+NgBlock partially escapes collapse.

Beam+NgBlock blocks repeated n-grams, which forces the beam to select different tokens when the most probable continuation is already blocked. This introduces marginal diversity ($\sigma(\bar{s}) \approx 0.003$) and explains why it achieves slightly higher ROUGE-L (0.2050 vs. 0.1899) and MiniCheck (0.1940 vs. 0.1441). It remains far below the diversity expected from genuine nucleus sampling.

12.3. Self-Healing Gains: Detailed Analysis

Bounded self-healing achieves statistically directional improvements across all faithfulness metrics. We analyze the source of each gain:

MiniCheck $+65.7\%$ ($0.1441 \rightarrow 0.2388$).

The Heal operator specifically targets the sentence with the lowest support score $\text{Supp}(c_{i^*}, X^*)$ and regenerates it conditioned on retrieved evidence snippets. The repair is accepted only if support increases by $\delta_{\text{sup}} = 0.03$. Over the 17 examples, the mean number of accepted repairs per example is 1.4, and the mean support improvement per repair is $+0.08$ (measured on the repaired claim sentence only), consistent with the observed overall MiniCheck improvement.

NLI Entailment $+59.0\%$ ($0.2161 \rightarrow 0.3437$).

The entailment improvement closely tracks the MiniCheck improvement, which is expected since both measure support-side faithfulness. The dual-signal trigger (Conf $< \tau_{\text{conf}}$ or $\bar{s} < s_{\min}$) ensures that even examples with moderate confidence but locally low-support sentences trigger repair- explaining why the entailment gain is substantial even in the Low disagreement stratum.

Contradiction Rate -5.8% ($0.3439 \rightarrow 0.3239$).

The repair acceptance criterion requires $\Delta_{\text{con}} \leq 0.05$, preventing repairs that introduce new contradictions. However, contradiction rate reduction is a secondary effect: when the most-violating sentence (high v_i) contains both low support and high contradiction, replacing it with a supported alternative also eliminates the contradiction. The 5.8% reduction over 17 examples represents approximately 1 fewer contradiction event per example on average.

ROUGE-1 -0.69% ($0.3943 \rightarrow 0.3874$).

The ROUGE-1 reduction is within the expected trade-off range for faithfulness-oriented decoding [55]. Repaired sentences use evidence-grounded language that may differ lexically from references; the minimal-distortion principle ($\gamma \cdot \text{EditDist}$ in Equation (37)) limits but cannot eliminate this trade-off. Importantly, BERTScore-F1 decreases only by -0.04% ($0.8446 \rightarrow 0.8443$), indicating that semantic similarity to references is nearly preserved even when surface-level ROUGE drops.

QA-F1 $+23.1\%$ ($0.1471 \rightarrow 0.1811$).

QA-F1 improvement confirms that the factual improvement measured by MiniCheck and NLI entailment is also detectable through an independent QA-based evaluation protocol, providing triangulated evidence that the gains reflect genuine faithfulness improvement.

 doi:10.20944/preprints202605.1381.v1

12.4. Per-Stratum Analysis

Table 6 reports MiniCheck and ConRate by disagreement stratum. As expected, High-disagreement clusters show the largest absolute faithfulness improvements, since they contain the most cross-document contradictions that the Heal operator can target. Low-disagreement clusters show smaller but consistent gains, since even easily-generated summaries contain locally unsupported claims.

Table 6. Per-stratum MiniCheck and ConRate for Beam vs. Ours:Const+Heal. Δ = absolute change. High-disagreement clusters benefit most from self-healing.

	MiniCheck			ConRate		
Stratum	**Beam**	**Ours**	**Δ**	**Beam**	**Ours**	**Δ**
Low ($n = 7$)	0.1623	0.2341	+0.072	0.2912	0.2791	−0.012
Medium ($n = 5$)	0.1389	0.2271	+0.088	0.3622	0.3441	−0.018
High ($n = 5$)	0.1234	0.2571	+0.134	0.4153	0.3641	−0.051
Overall ($n = 17$)	0.1441	0.2388	+0.095	0.3439	0.3239	−0.020

12.5. Directional Validation

Three pre-registered directional hypotheses are confirmed:

1. $\text{MiniCheck}(\text{Ours}) > \text{MiniCheck}(\text{Beam})$: $0.2388 > 0.1441$ ✓
2. $\text{ConRate}(\text{Ours}) < \text{ConRate}(\text{Beam})$: $0.3239 < 0.3439$ ✓
3. $\text{Entail}(\text{Const} + \text{Heal}) > \text{Entail}(\text{Constrained})$: $0.3437 > 0.2161$ ✓

All three pass, validating the core methodology.

12.6. Efficiency Analysis

Table 7 reports end-to-end inference efficiency.

Table 7. Inference efficiency profile ($n = 17$, NVIDIA RTX PRO 6000). Latency reported as mean over examples. The anomalously low latency of Ours:Const+Heal is explained in the text.

System	**Latency (ms)**	**Peak Mem (MB)**	**Rel. Latency**
Greedy	509.1	0.0	1.00×
Beam	606.5	0.1	1.19×
Beam+NgBlock	597.7	0.0	1.17×
Rerank(Rep)	599.4	0.1	1.18×
Rerank+Verifier	675.8	0.1	1.33×
PostHocVerify	609.2	0.1	1.20×
Ours: Constrained	635.6	0.2	1.25×
Ours: Const+Heal	124.0	0.0	0.24×

Explaining the anomalous Const+Heal latency.

The 124ms latency for Ours:Const+Heal is substantially lower than all baselines, including Greedy (509ms). This is not a performance optimization artifact but a consequence of beam collapse: when all candidates are identical and the most-violating sentence is repaired in a single pass, the system skips the full retry loop because Conf exceeds τ_{conf} after one repair in many examples. The low latency reflects fast early termination under beam collapse conditions, *not* that the framework is genuinely faster. After the candidate diversification fix, latency is expected to increase modestly to approximately 750–900ms due to: (a) nucleus sampling overhead (≈50ms per sample); (b) batched NLI scoring of 4 diverse candidates (≈50ms); (c) rank aggregation computation (negligible). This remains well within acceptable bounds for offline MDS inference.

 doi:10.20944/preprints202605.1381.v1

13. Qualitative Analysis and Case Studies

To ground the quantitative results, we provide three representative case studies from the validation set (Low, Medium, High disagreement clusters).

13.1. Case Study 1: Low Disagreement (Successful Repair)

Cluster topic: Municipal budget approval, 3 documents, consistent facts.

Beam output (MiniCheck: 0.18): "City council members voted to approve the budget plan on Tuesday. The plan allocates over $50 million for infrastructure improvements including road resurfacing and bridge maintenance projects. Officials said the decision was unanimous and would take effect immediately."

Issue: The '$50 million' figure was not in X^* (actual figure: $42.7 million); the figure was hallucinated from a model prior.

Healed output (MiniCheck: 0.67): "City council members voted unanimously to approve the budget plan on Tuesday. The plan allocates approximately $42.7 million for infrastructure improvements including road resurfacing and bridge maintenance projects across the city. Officials said the decision would take effect at the start of the next fiscal year."

Repair action: The second sentence was identified as the most violating ($v_{i^*} = 0.41$); the figure was corrected to match X^* and contextual details were grounded to the evidence. $\Delta_{\text{sup}} = +0.49$, $\Delta_{\text{con}} = -0.03$ (repair accepted).

13.2. Case Study 2: Medium Disagreement (Partial Repair)

Cluster topic: Labor negotiations, 4 documents, partially conflicting.

Beam output (MiniCheck: 0.09): "Union leaders announced a strike would begin on Monday after talks broke down. Management responded by offering a 3% pay increase, which workers rejected. A mediator has been appointed to resolve the dispute, and both sides expressed willingness to continue negotiations."

Issue: The "3% pay increase" is mentioned in one document but disputed as "insufficient" in another; the claim that "both sides expressed willingness" contradicts Document 3, which reports management refusing further mediation.

Healed output (MiniCheck: 0.29): "Union leaders announced a strike would begin on Monday after talks broke down. Management offered a pay increase that workers rejected as insufficient. A mediator has been appointed, though the status of further negotiations remains uncertain."

Repair action: Two violations targeted. The specific percentage was replaced with hedged language supported by all documents. The contradicted "willingness" claim was replaced with a neutral formulation. MiniCheck improved from 0.09 to 0.29 but did not reach full feasibility ($\tau_{\text{conf}} = 0.35$) within 3 retries. Best-so-far candidate returned.

13.3. Case Study 3: High Disagreement (Complex Contradiction)

Cluster topic: Election results, 5 documents, significantly conflicting.

In this high-disagreement case, documents from different news outlets reported conflicting vote margins (one reported "narrow victory", another "landslide"). The beam output (MiniCheck $=$ 0.05) committed to a specific margin present in only one document. Healing replaced the margin with a hedged formulation ("a margin that officials are expected to confirm in the coming days"), achieving MiniCheck $= 0.31$ after two repairs. This illustrates a fundamental property of the Heal operator: under genuine cross-document disagreement, the optimal faithful summary is often a more conservative, hedged formulation rather than a committed claim.

 doi:10.20944/preprints202605.1381.v1

14. Diagnostics and Ablation Studies

14.1. Beam Collapse Ablation

We compare three candidate pool strategies to isolate the effect of diversification: (1) *Pure beam* (current preliminary run): all $K = 4$ candidates identical, $\sigma(\bar{s}) = 0.000$; (2) *Beam+sampling* (proposed): 1 beam + 3 nucleus samples, expected $\sigma(\bar{s}) \approx 0.04$–$0.08$ (awaiting full run); (3) *Diverse beam search* with group penalty [56]: expected $\sigma(\bar{s}) \approx 0.01$–$0.03$ (intermediate diversity).

Under pure beam collapse, rank aggregation and constrained selection provide zero benefit over greedy decoding. Under beam+sampling, we expect rank aggregation to meaningfully differentiate candidates by support score, enabling the constrained objective to select more faithful candidates even before healing.

14.2. Constraint Removal Study

Table 8 reports the expected directional metric shifts when each constraint component is removed from the full system.

Table 8. Ablation study: expected directional metric changes vs. full Ours:Const+Heal when each component is removed. Full quantitative results on 500 examples are pending; directions are derived from the mechanisms described in §5–8.

Config	ΔR-1	ΔMini	ΔQA-F1	ΔEntail	ΔCon	ΔLat
Full system	—	—	—	—	—	—
No Heal	≈0	↓	↓	↓	↑	↓
No Verifier	≈0	↓	↓	↓	↑	↓
No Entail Pen.	↑	↓	↓	↓	≈0	↓
No Con Check	≈0	≈0	≈0	≈0	↑	↓
No Rep Pen.	≈0	≈0	↓	≈0	≈0	↓
No Rank Agg.	≈0	↓	↓	↓	↑	≈0
No Nucleus	↑	↓	↓	↓	↑	↓

Key predictions: removing the Heal operator should produce the largest faithfulness regression, since in the current run it is the only component that generates genuinely new text. Removing the verifier (falling back to lexical controls only) should produce regression in both MiniCheck and ConRate but less regression in Dup@4. Removing the entailment penalty may slightly improve ROUGE (less conservative language) at the cost of faithfulness. These predictions will be verified in the full 500-example evaluation.

14.3. Verifier Calibration Sensitivity

We analyze sensitivity to the NLI threshold τ_{ent}. At $\tau_{\text{ent}} = 0.30$ (more permissive), fewer claims trigger healing, reducing latency but potentially leaving low-support claims unrepaired. At $\tau_{\text{ent}} = 0.60$ (more strict), more claims trigger healing, increasing repair coverage but also increasing false-trigger rate when the verifier is uncertain. We set $\tau_{\text{ent}} = 0.45$ as a balance point calibrated on the 50-example validation subset.

15. Discussion

15.1. Why the Framework Works: Mechanistic Analysis

Three mechanisms jointly explain the faithfulness improvements.

Feasibility shaping through penalty terms.

The penalty $K(Y, X^*)$ creates a soft feasibility surface over the candidate space. Candidates with low support or high contradiction receive higher penalties, making them less likely to be selected even when they have higher log-probability. This effect is strongest in the nucleus-sampling regime where log-probability differences between candidates are smaller and faithfulness terms carry more weight.

 doi:10.20944/preprints202605.1381.v1

Opportunity-cost redundancy control.

Under fixed budget B_Y, redundancy and faithfulness compete for tokens. The Dup@n penalty creates an explicit opportunity cost for repetition: among similarly-faithful candidates, the one that introduces new supported information is preferred over one that repeats earlier sentences. This is an emergent property of the joint objective- redundancy control improves faithfulness efficiency without requiring explicit faithfulness signals.

Local repair as feasibility projection.

The Heal operator approximates a projection step onto the feasible set: $Y' = \arg\min_{Y'' : Y'' \text{ feasible}} \text{EditDist}(Y, Y'')$. The minimal-change principle ($\gamma \cdot$ EditDist penalty in Equation (37)) biases repair toward solutions close to the original, preserving already-faithful content while correcting violations. This is analogous to projected gradient descent in continuous optimization: the feasibility projection takes a step toward the constraint set from the current infeasible point.

15.2. Failure Mode Analysis

Verifier false positives.

The DeBERTa NLI verifier can trigger unnecessary repairs on correct sentences whose phrasing differs from evidence without being unfaithful (*e.g.*,, paraphrase of a fact using synonyms). In the preliminary run, we estimate false-positive rate at approximately 15% of repair triggers based on manual inspection of a 20-example subset. Calibration via temperature scaling on verifier logits is expected to reduce this to $< 10\%$.

Evidence omission bottleneck.

When critical evidence is absent from X^* due to packing, the verifier cannot find supporting evidence for correct claims, potentially triggering unnecessary repairs of accurate information. This is an irreducible limitation of budget-constrained MDS: evidence omitted at packing time cannot be recovered at decoding time. Salience-aware packing (prioritizing high-entropy passages likely to be cited) is a natural extension to address this.

High-contradiction clusters.

In the High-disagreement stratum, the Heal operator sometimes cannot find a fully feasible replacement that satisfies both support and contradiction constraints simultaneously- when a claim is supported by Document 1 but contradicted by Document 2, no phraseable claim can satisfy both. In these cases, the fallback to hedged language ("officials disagreed on...", "reports varied on...") is the correct behavior but may reduce reference ROUGE since references typically commit to one perspective.

15.3. Connections to Prior Theoretical Work

Our Lagrangian relaxation formulation (Equation (16)) connects to constrained MDP formulations of text generation [57], where faithfulness constraints are equivalent to safety constraints in constrained policy optimization. The bounded Heal operator has a parallel in trust region policy optimization [58]: both limit the magnitude of updates to prevent catastrophic interference with already-correct structure.

The rank aggregation approach (Equation (32)) is a special case of Borda count voting, a classical social choice mechanism shown to minimize expected Kemeny distance to the true ranking [59]. Applied to candidate selection, this means our aggregation method is near-optimal in expectation when the true faithfulness ranking is corrupted by independent noise in each signal axis.

15.4. Scope and Limitations

Small backbone.

LSHT is an 18.4M-parameter model- far smaller than BART-Large (400M) or PEGASUS-Large (568M). The beam collapse failure mode may be less severe in larger models with higher entropy

 doi:10.20944/preprints202605.1381.v1

distributions. Conversely, the candidate diversification fix (nucleus sampling) and rank aggregation are model-agnostic and will benefit any backbone exhibiting low candidate diversity.

Preliminary sample size.

The $n = 17$ stratified validation is a diagnostic run, not a statistically powered evaluation. Standard errors are not reported for the stratified sample, and the full 500-example evaluation is required to make quantitative claims about mean performance with appropriate confidence intervals.

Fixed penalty weights.

The Lagrangian weights $\{\lambda_j\}$ are fixed by validation set tuning. In principle, these can be optimized via dual ascent on a held-out set, potentially yielding a Pareto-optimal point on the faithfulness-fluency frontier rather than a fixed operating point.

News domain specificity.

All evaluations are on MultiNews, a news-domain dataset. Performance on biomedical, legal, or scientific MDS may differ due to domain-specific entity types and contradiction patterns.

15.5. Future Directions

1. **Dual ascent for adaptive penalty weights.** Treating $\{\lambda_j\}$ as Lagrangian multipliers updated by dual gradient ascent on a held-out validation set would provide an automatic, dataset-adaptive operating point on the faithfulness-fluency trade-off curve.
2. **Salience-aware packing.** Replacing Hamilton apportionment with a salience-ranked packing that prioritizes high-entropy, controversial, or frequently-cited passages would reduce evidence omission and improve verifier coverage.
3. **LLM backbone evaluation.** Applying FADCO to larger backbones (BART-Large, PEGASUS-Large) and to instruction-tuned LLMs (Llama-2, Mistral) would test whether the framework generalizes beyond small seq2seq models. LLMs may exhibit less beam collapse but would benefit more from rank aggregation due to larger candidate diversity.
4. **Cross-domain evaluation.** Evaluation on WCEP (Wikipedia MDS), MultiXScience (scientific MDS), and MQA (legal MDS) would assess generalization across domains with different entity types, contradiction patterns, and output style requirements.
5. **Human evaluation.** Automatic faithfulness metrics remain imperfect proxies for human judgment. A small-scale human evaluation (annotation with the FactSpan protocol [2]) would provide stronger validation of the MiniCheck improvements observed.

16. Conclusions

We have argued that faithfulness in multi-document summarization is fundamentally a decoding-time constraint satisfaction problem, not a training-time optimization problem, and have presented FADCO as a principled framework for this view.

The framework makes four concrete contributions: (i) a Lagrangian-relaxed constrained decoding objective encoding five faithfulness dimensions; (ii) formal diagnosis of beam collapse and a fix via mixed candidate pools; (iii) rank aggregation that eliminates log-probability scale dominance over faithfulness signals; and (iv) bounded self-healing with provable termination, monotonic acceptance, and minimal distortion.

Preliminary stratified validation on MultiNews confirms three pre-registered directional hypotheses: bounded self-healing improves MiniCheck by $+65.7\%$ and NLI entailment by $+59.0\%$ over beam decoding, and reduces contradiction rate by -5.8%, with ROUGE-1 trade-off of only -0.69% and BERTScore degradation of only -0.04%. Per-stratum analysis reveals that gains are largest in high-disagreement clusters (MiniCheck $+0.134$)- precisely the regime where faithfulness intervention is most needed.

Beyond the empirical results, this work makes a methodological argument: the beam collapse failure mode is a systematic, silent failure that invalidates the assumption underlying all re-ranking

 doi:10.20944/preprints202605.1381.v1

and constrained decoding baselines. Any evaluation of verifier-guided or constrained decoding must first verify that candidate diversity is non-trivial; otherwise, observed improvements may be entirely attributable to a single component (as in our run, where healing is the only mechanism producing novel text). We provide both a formal criterion for detecting collapse (Equation (21)) and a practical fix.

Full evaluation on 500 examples with the complete candidate diversification and rank aggregation pipeline is ongoing and will provide quantitative confirmation of the component-wise contributions.

Author Contributions: Conceptualization, S.P. and S.K.S.; Methodology, S.P. and S.K.S.; Software, S.P. and S.K.S.; Formal analysis, S.P. and S.K.S.; Validation, S.P. and S.K.S.; Investigation, S.P. and S.K.S.; Data curation, S.P. and S.K.S.; Writing—Original draft, S.P. and S.K.S.; Writing—Review & editing, S.P. and S.K.S.; Visualization, S.P. and S.K.S.; Supervision, S.K.S.; Project administration, S.K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code and hyperparameter configurations are available at https://github.com/Sameer-dev1/FADCO. The stratified evaluation splits supporting this study are available from the corresponding author upon reasonable request. MultiNews is publicly available at https://github.com/Alex-Fabbri/Multi-News.

Acknowledgments: The authors declares that no specific funding, technical assistance, or external support was received for this work.

Conflicts of Interest: The authors declare no conflicts of interest.

 doi:10.20944/preprints202605.1381.v1

Appendix A. Hyperparameter Settings

Table A1. Full hyperparameter configuration for FADCO (LSHT instantiation). All values were fixed on a 50-example validation subset prior to test evaluation.

Parameter	Value	Rationale
Constraint penalty weights		
λ_{rep}	0.6	Moderate repetition suppression
λ_{con}	1.2	Strong contradiction avoidance
λ_{sup}	0.8	Balanced support enforcement
Hard constraint thresholds		
τ_{ent}	0.45	Minimum entailment probability
τ_{con}	0.35	Maximum contradiction probability
τ_{dup}	0.15	Maximum 4-gram redundancy rate
τ_{conf}	0.35	Healing trigger confidence floor
s_{min}	0.40	Secondary healing trigger
Candidate generation		
Beam width K	4	1 beam + 3 samples
Nucleus temperature τ	0.85	Mild temperature softening
Nucleus top-p	0.92	92% probability mass nucleus
Dedup retries	$4(K-1) = 12$	Per candidate slot
Self-healing		
Max retries R	3	Bounded compute
Min improvement η	0.02	Monotonic acceptance margin
Max window W_{max}	80 tokens	Repair scope limit
Max length delta Δ	20 tokens	Expansion allowance
Edit penalty γ	0.2	Distortion regularization
δ_{sup}	0.03	Min support gain to accept
δ_{con}	0.05	Max contradiction increase to accept
Rank aggregation weights		
w_{sup}	2.0	Prioritize support signal
w_{con}	1.5	Strong contradiction avoidance
w_{lp}	1.0	Retain fluency signal
w_{dup}	0.5	Mild redundancy penalty
Scoring and inference		
Length norm. α	0.7	Moderate length normalization
N-gram blocking size	3	Trigram blocking in Beam+NgBlock
Verifier batch size	128	Fits RTX PRO 6000 VRAM
Unigram evidence top-k	7	Evidence snippets per claim
Healing violation weights		
β_{ent}	1.0	Entailment deficit weight
β_{con}	0.8	Contradiction excess weight
β_{rep}	0.3	Local redundancy weight

Appendix B. Metric Reproducibility Audit

We document the exact failure modes of AlignScore and QuestEval that motivated their replacement with MiniCheck.

AlignScore failure.

AlignScore [13] requires `transformers==4.27.0` and `torch==1.13.0`. Under `transformers>=4.35.0` and `torch>=2.0.0` (required by DeBERTa-v3 and other framework components), the `AlignScore` class raises an `ImportError` due to renamed internal APIs in the `transformers` library. A pinned-version

 doi:10.20944/preprints202605.1381.v1

environment is technically possible but creates incompatibilities with NumPy 1.x vs. 2.x for array operations used elsewhere in the pipeline.

QuestEval failure.

QuestEval [14] requires `spacy==3.1.x` and the `en-core-web-lg==3.1.0` language model. Under `spacy>=3.5.0` (current stable release), the `en-core-web-lg` model version is incompatible, causing a pipeline initialization crash. The QuestEval package has not been updated to support current spaCy versions.

MiniCheck verification.

MiniCheck [15] installs without conflicts under: `transformers>=4.35.0`, `torch>=2.1.0`, `sentencepiece>=0.1.99`. All dependencies are compatible with the rest of the FADCO codebase. Installation command: `pip install minicheck`.

Appendix C. FADCO System Summary

Table A2. FADCO complete system summary: components, their purpose, computational cost, and formal guarantees.

Component	Purpose	Cost	Guarantee
Hamilton Packing	Proportional evidence allocation	$O(K)$	Quota fairness
Beam Anchor	Quality baseline candidate	$O(T \cdot V)$	Most probable output
Nucleus Sampling	Diverse candidates	$O((K-1) \cdot T \cdot V)$	Diversity w.h.p.
Batched NLI	Support/contradiction scoring	$O(B_{\text{NLI}})$	Accurate entailment
Rank Aggregation	Scale-independent selection	$O(K \log K)$	No axis dominance
Heal Operator	Local violation repair	$O(R \cdot W)$	Termination, monotonic

References

1. Maynez, J.; Narayan, S.; Bohnet, B.; McDonald, R. On faithfulness and factuality in abstractive summarization. In Proceedings of the Proceedings of ACL, 2020, pp. 1906–1919.
2. Pagnoni, A.; Balachandran, V.; Tsvetkov, Y. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In Proceedings of the Proceedings of NAACL, 2021, pp. 4812–4829.
3. Cao, S.; Wang, L. Hallucinated but factual! Inspecting the factuality of hallucinations in abstractive summarization. In Proceedings of the Proceedings of ACL, 2022, pp. 3340–3354.
4. Kang, D.; Hashimoto, T.B. Improved Natural Language Generation via Loss Truncation. In Proceedings of the Proceedings of ACL, 2020.
5. Zhang, T.; et al. Benchmarking Faithfulness in Natural Language Generation. *arXiv preprint* **2024**.
6. Nan, L.; Wiseman, S.; Bansal, M.; Chen, Y.; Perez, E.; Mei, H.; Barzilay, R. Entity-level factual consistency of abstractive text summarization. In Proceedings of the Proceedings of EACL, 2021, pp. 2727–2733.
7. Chen, S.; Zhang, F.; Sone, Y.; Litman, D. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In Proceedings of the Proceedings of NAACL, 2021, pp. 5935–5941.
8. Zhu, C.; Hinthorn, W.; Xu, R.; Zeng, Q.; Zeng, M.; Huang, X.; Jiang, M. Enhancing factual consistency of abstractive summarization. In Proceedings of the Proceedings of NAACL, 2021, pp. 718–733.
9. Ladhak, F.; Durmus, E.; Cardie, C.; McKeown, K. Faithful or extractive? On mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In Proceedings of the Proceedings of ACL, 2022, pp. 1410–1421.

 doi:10.20944/preprints202605.1381.v1

10. Wan, Z.; Wan, F.; Yu, W.; Du, Y.; Lam, W.; Pan, B. Faithfulness-aware decoding strategies for abstractive summarization. In Proceedings of the Proceedings of EACL, 2023, pp. 891–908.
11. Meister, C.; Vieira, T.; Cotterell, R. If beam search is the answer, what was the question? In Proceedings of the Proceedings of EMNLP, 2020, pp. 2173–2185.
12. Dziri, N.; Milton, A.; Yu, M.; Zaiane, O.; Reddy, S. On the origin of hallucinations in conversational models. *Proceedings of NAACL* **2022**, pp. 5765–5780.
13. Zha, Z.; Bohnet, B.; Dong, N.; Metzler, D.; Ni, J. AlignScore: Evaluating factual consistency with a unified alignment function. In Proceedings of the Proceedings of ACL, 2023, pp. 11328–11348.
14. Scialom, T.; Dray, P.A.; Gallé, M.; Gallinari, P.; Piwowarski, B.; Staiano, J.; Wang, A. QuestEval: Summarization asks for fact-based evaluation. In Proceedings of the Proceedings of EMNLP, 2021, pp. 6594–6604.
15. Tang, L.; Laban, P.; Carenini, G. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In Proceedings of the Proceedings of EMNLP, 2024, pp. 8818–8847.
16. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.; Artzi, Y. BERTScore: Evaluating text generation with BERT. In Proceedings of the Proceedings of ICLR, 2020.
17. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the Proceedings of ACL, 2020, pp. 7871–7880.
18. Zhang, J.; Zhao, Y.; Saleh, M.; Liu, P. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In Proceedings of the Proceedings of ICML, 2020, Vol. 119, pp. 11328–11339.
19. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **2020**, *21*, 1–67.
20. Zaheer, M.; Guruganesh, G.; Dubey, A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. Big Bird: Transformers for longer sequences. In Proceedings of the Proceedings of NeurIPS, 2020, Vol. 33, pp. 17283–17297.
21. Shleifer, S.; Rush, A. Pre-trained summarization distillation. *arXiv preprint arXiv:2010.13002* **2020**.
22. Fabbri, A.R.; Li, I.; She, T.; Li, S.; Radev, D. Multi-News: A large-scale multi-document summarization dataset and abstractive hierarchical model. In Proceedings of the Proceedings of ACL, 2019, pp. 1074–1084.
23. Pandey, S.; Singh, S.K. Lightweight Self-Healing Transformers for Faithful Summarization. *Technical Report* **2024**.
24. Falke, T.; Ribeiro, L.F.R.; Utama, P.A.; Dagan, I.; Gurevych, I. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In Proceedings of the Proceedings of ACL, 2019, pp. 2214–2220.
25. Kryscinski, W.; McCann, B.; Xiong, C.; Socher, R. Evaluating the factual consistency of abstractive text summarization. In Proceedings of the Proceedings of EMNLP, 2020, pp. 9332–9346.
26. Wang, A.; Cho, K.; Lewis, M. Asking and answering questions to evaluate the factual consistency of summaries. In Proceedings of the Proceedings of ACL, 2020, pp. 5008–5020.
27. Fabbri, A.; Wu, C.S.; Liu, W.; Xiong, C. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In Proceedings of the Proceedings of NAACL, 2022, pp. 2587–2601.
28. Deutsch, D.; Bedrax-Weiss, T.; Roth, D. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the ACL* **2021**, *9*, 774–789.
29. Laban, P.; Schnabel, T.; Bennett, P.; Hearst, M. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the ACL* **2022**, *10*, 163–177.
30. Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.t.; Koh, P.W.; Iyyer, M.; Zettlemoyer, L.; Hajishirzi, H. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Proceedings of the Proceedings of EMNLP, 2023, pp. 12076–12100.
31. Manakul, P.; Liusie, A.; Gales, M. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Proceedings of the Proceedings of EMNLP, 2023, pp. 9004–9017.
32. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Guided Open Vocabulary Image Captioning with Constrained Beam Search. In Proceedings of the Proceedings of EMNLP, 2017.
33. Post, M.; Vilar, D. Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation. In Proceedings of the Proceedings of NAACL, 2018.
34. Keskar, N.S.; McCann, B.; Varshney, L.R.; Xiong, C.; Socher, R. CTRL: A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858* **2019**.

 doi:10.20944/preprints202605.1381.v1

35. Rashkin, H.; Nikolaev, V.; Lamm, M.; Aroyo, L.; Collins, M.; Das, D.; Petrov, S.; Tomar, G.S.; Turc, I.; Reitter, D. Measuring Attribution in Natural Language Generation Models. In Proceedings of the Proceedings of Computational Linguistics, 2023.
36. Welleck, S.; Kulikov, I.; Roller, S.; Dinan, E.; Cho, K.; Weston, J. Neural Text Generation with Unlikelihood Training. In Proceedings of the Proceedings of ICLR, 2020.
37. Eikema, B.; Aziz, W. Is MAP Decoding All You Need? The Inadequacy of the Mode in Neural Machine Translation. In Proceedings of the Proceedings of COLING, 2020.
38. Freitag, M.; Foster, G.; Grangier, D.; Ratnakar, V.; Tan, Q.; Macherey, W. High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics. In Proceedings of the Transactions of the Association for Computational Linguistics, 2022.
39. Yang, K.; Klein, D. FUDGE: Controlled text generation with future discriminators. In Proceedings of the Proceedings of NAACL, 2021, pp. 3511–3535.
40. Krause, B.; Gotmare, A.; McCann, B.; Keskar, N.; Joty, S.; Socher, R.; Rajani, N. GeDi: Generative discriminator guided sequence generation. In Proceedings of the Proceedings of EMNLP Findings, 2021, pp. 4929–4952.
41. Gao, T.; Fisch, A.; Chen, D. RARR: Researching and revising what language models say, using language models. In Proceedings of the Proceedings of ACL, 2023, pp. 16477–16508.
42. Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. Self-refine: Iterative refinement with self-feedback. In Proceedings of the Proceedings of NeurIPS, 2023, Vol. 36.
43. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. PEER: A collaborative language model. In Proceedings of the Proceedings of ICLR, 2023.
44. He, P.; Liu, X.; Gao, J.; Chen, W. DeBERTa: Decoding-enhanced BERT with disentangled attention. In Proceedings of the Proceedings of ICLR, 2021.
45. Bowman, S.; Angeli, G.; Potts, C.; Manning, C. A large annotated corpus for learning natural language inference. In Proceedings of the Proceedings of EMNLP, 2015, pp. 632–642.
46. Williams, A.; Nangia, N.; Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the Proceedings of NAACL, 2018, pp. 1112–1122.
47. Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Mittal, A. FEVER: A large-scale dataset for fact extraction and verification. In Proceedings of the Proceedings of NAACL, 2018, pp. 809–819.
48. Bertsekas, D.P. *Nonlinear Programming*, 2nd ed.; Athena Scientific: Belmont, MA, 1999.
49. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The curious case of neural text degeneration. In Proceedings of the Proceedings of ICLR, 2020.
50. Su, J.; Lu, Y.; Pan, S.; Murtadha, A.; Wen, B.; Liu, Y. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing* **2024**, *568*, 127063.
51. Hendrycks, D.; Gimpel, K. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415* **2016**.
52. Lin, C.Y. ROUGE: A package for automatic evaluation of summaries. In Proceedings of the Proceedings of the ACL Workshop, 2004, pp. 74–81.
53. Khalifa, M.; Elsahar, H.; Dymetman, M. A distributional approach to controlled text generation. In Proceedings of the Proceedings of ICLR, 2021.
54. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the Proceedings of EMNLP, 2016, pp. 2383–2392.
55. Zhao, Y.; et al. Reducing Hallucination in Neural Text Generation. In Proceedings of the Proceedings of NLP Research, 2020.
56. Vijayakumar, A.; Cogswell, M.; Selvaraju, R.; Sun, Q.; Lee, S.; Crandall, D.; Batra, D. Diverse beam search: Decoding diverse solutions from neural sequence models. In Proceedings of the arXiv preprint arXiv:1610.02424, 2016.
57. Lu, X.; West, P.; Zellers, R.; Le Bras, R.; Bhagavatula, C.; Choi, Y. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In Proceedings of the Proceedings of NAACL, 2021, pp. 4288–4299.
58. Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; Moritz, P. Trust region policy optimization. In Proceedings of the Proceedings of ICML, 2015, pp. 1889–1897.
59. Dwork, C.; Kumar, R.; Naor, M.; Sivakumar, D. Rank aggregation methods for the web. In Proceedings of the Proceedings of WWW, 2001, pp. 613–622.

 doi:10.20944/preprints202605.1381.v1