

Hierarchical BART for Abstractive Multi-Document Summarisation: A Two-Stage Inference Pipeline with Semantic Evaluation

Prabhat Shekhar

Department of Artificial Intelligence, Dr. B.R. Ambedkar National Institute of Technology Jalandhar, Jalandhar, Punjab, 144008, India.

ABSTRACT: Synthesising information scattered across multiple related documents into a single readable summary is fundamentally harder than summarising a single text. When documents originate from independent authors, the same event may be described through contrasting lenses, with divergent vocabulary, conflicting details, or complementary facts that no single source captures alone. Neural abstractive models—BART, T5, and PEGASUS—have greatly narrowed the gap to human performance on single-document tasks, yet their fixed 1,024-token encoder window becomes a hard ceiling when applied to multi-document clusters. For news corpora the average cluster runs to roughly 2,200 tokens; for scientific literature it reaches 4,800. Standard truncation therefore quietly throws away 54% and 79% of the source material, respectively—a coverage shortfall that current leaderboards largely obscure. Architectural remedies such as GraphSum, PRIMERA, and TALER partially address this by extending context capacity or re-aligning representations, but each demands substantial engineering overhead—extra GNN layers, specialised pre-training objectives, or delicate adversarial training loops. We introduce HierBART (Hierarchical BART), which sidesteps the need for any of this by treating the limitation as an inference-time problem rather than an architectural one. The approach is deliberately minimal: chunk the token stream into fixed-width overlapping windows, summarise each with a fine-tuned BART-BASE model (Stage 1), concatenate those local summaries into a fusion sequence, and re-decode it through the same model a second time (Stage 2). Because the pipeline uses only a standard fine-tuned checkpoint with unchanged weights, any future improvement to the backbone—quantisation, distillation, better fine-tuning data—automatically carries over. Benchmarked on Multi-News and Multi-XScience, HierBART reaches ROUGE-1 of 45.70 on Multi-News, outstripping TALER by 12.86 points, and 32.68 on Multi-XScience (+4.38). BERTScore F1 (roberta-large) stands at 80.14 and 84.44 respectively. A five-condition validation-set ablation traces 11.73 of the 22.21-point cumulative gain to the two structural decisions—window chunking (+7.17) and two-pass fusion (+4.56)—ruling out decoding artefacts as the driver. Blind human evaluation by five postgraduate NLP students rates HierBART at 7.90/10 for Fluency and 8.10/10 for Informativeness (Fleiss κ = 0.61), confirming that metric gains translate to perceptibly better summaries.

KEYWORDS: multi-document summarisation; BART; hierarchical transformer; abstractive summarisation; ROUGE; BERTScore; window chunking; two-pass decoding; semantic evaluation; natural language processing.

1 INTRODUCTION

When a major story breaks, a dozen outlets publish their own versions within hours. A reader who wants the full picture—not just the angle one reporter chose to emphasise—has to cross-reference pieces that quote different sources, apply different frames, and occasionally contradict each other on the facts. Researchers navigating a crowded subfield face exactly the same friction at a larger scale. Multi-document summarisation (MDS) is the computational attempt to solve this problem: produce, from a set of related texts, a single summary that assembles the salient and complementary information without repeating what every source says.

The single-document case is, by comparison, largely solved: large pre-trained encoder-decoder models such as BART [1] score near the ceiling of automatic metrics on CNN/DailyMail and XSum. MDS is harder along several dimensions. Different outlets may describe the same event through wholly different vocabulary, so the model must judge semantic equivalence across surface variation. Documents in a cluster may disagree on facts or represent different stages of an unfolding story, requiring some notion

Hierarchical BART for Abstractive Multi-Document Summarisation: A Two-Stage Inference Pipeline with Semantic Evaluation

of temporal ordering and preferential trust. Most critically for any current transformer model, the combined token length of a multi-document cluster dwarfs the encoder's context window—and that is the problem this paper addresses.

Why truncation is not an acceptable fix. BART-BASE, T5-base, and PEGASUS all cap encoder input at 1,024 tokens. A typical Multi-News cluster [15] runs to about 2,200 BART subword tokens; a Multi-XScience cluster [16] reaches roughly 4,800 tokens. Chopping the tail off silently discards 54% and 79% of source material respectively. What makes this especially insidious is that the discarded content is not random filler—later documents in a news cluster often carry corrections, updates, or the most methodologically careful findings. Figure 1 shows just how far outside BART's window the bulk of both test sets fall.

What prior work offers—and what it costs. GraphSum [11] attaches GNN layers to the encoder to propagate cross-document signal; this requires joint training of an additional module. PRIMERA [12] retraining a Longformer encoder-decoder up to 4,096 tokens with a custom entity masking objective, a significant pre-training investment. TALER [4] reformulates MDS as a representation alignment problem via adversarial reversal and achieves the strongest published scores before this work, but its training loop involves a discriminator that must be carefully balanced against the main model. Every prior solution adds something to the model: more modules, more pre-training, or a more elaborate training loop. HierBART instead changes how inference is organised.

HierBART: a simpler path. We ask whether the bottleneck can be addressed without touching the model architecture at all. Our answer is HierBART (Hierarchical BART): a two-stage inference-time pipeline built on a standard fine-tuned BART-BASE checkpoint. The idea is to break the cluster into fixed-width overlapping windows, let the model summarise each window in isolation (Stage 1), and then present those intermediate summaries as a compact fusion sequence for a second global decoding pass (Stage 2). The weights are never modified; the pipeline is illustrated in Figure 2.

Contributions. This paper makes five specific contributions. First, we propose a parameter-free two-stage inference pipeline that resolves the context-length bottleneck at test time without any architectural changes or additional training. Second, we provide comprehensive results on Multi-News and Multi-XScience against six baselines, with full reporting of ROUGE, BERTScore, and human evaluation—including cases where HierBART does not lead. Third, a controlled five-condition ablation attributes each gain to a specific design decision: fine-tuning (+8.00), window chunking (+7.17), two-pass fusion (+4.56), and decoding tuning (+2.48). Fourth, blind human evaluation by five postgraduate NLP students provides a Fleiss κ -validated comparison across Fluency, Coherence, and Informativeness. Fifth, BERTScore analysis reveals that abstractive quality gains are larger than n-gram metrics indicate.

2 RELATED WORK

2.1 Extractive Summarisation

Graph-based sentence ranking dominated early MDS research. TextRank [5] scored sentences by eigenvector centrality in a cosine-similarity graph, borrowing the intuition behind PageRank. LexRank [6] applied the same graph-centrality principle directly to multi-document settings using an idf-modified cosine as the edge weight. Neither system requires labelled data, which kept them competitive for many years. Their shared weakness is architectural rather than empirical: an extractive system is limited to sentences already present in the source—it has no mechanism to combine a fact from Article A with a qualification from Article B into a single new sentence. MMR [7] introduced explicit redundancy penalisation into the selection loop, improving output diversity without solving the fusion problem.

2.2 Neural Abstractive Summarisation

Bahdanau et al.'s attention mechanism [10] gave neural encoder-decoder models the ability to selectively attend to source positions during decoding, enabling genuine paraphrastic output for the first time. Hallucination remained a practical concern: the decoder could produce fluent but factually unsupported text. Pointer-generator networks [8] addressed this with a learned copy gate, allowing the decoder to lift spans verbatim from the source when paraphrase was unnecessary or risky. Paulus et al. [9] took a different angle, optimising ROUGE directly as an RL reward; the approach improved automatic scores but introduced well-known training instability.

The contemporary baseline shifted decisively with the transformer architecture [28]: large-scale pre-training via denoising (BART [1]), text-to-text transfer (T5 [2]), and gap-sentence generation (PEGASUS [3]) lifted single-document performance to near-human levels. All three, however, cap the encoder at 512–1,024 tokens, which is perfectly adequate for a single news article but hopelessly inadequate for a cluster of seven scientific abstracts.

Hierarchical BART for Abstractive Multi-Document Summarisation: A Two-Stage Inference Pipeline with Semantic Evaluation

2.3 Multi-Document Summarisation Architectures

GraphSum [11] was among the first to model the inter-document structure explicitly, constructing a heterogeneous graph over sentences and using GNN message-passing to let evidence propagate across document boundaries before decoding. The cost is that GNN layers must be trained jointly with the encoder-decoder, adding architectural complexity. PRIMERA [12] took a different route, extending the Longformer encoder-decoder to 4,096 tokens and designing an entity-pyramid masking pre-training objective so that the model learns to cross-reference documents during pre-training rather than fine-tuning. Despite the greatly expanded context, our reproduction of PRIMERA under the same fine-tuning conditions as all other baselines places it below vanilla BART; we attribute this to a mismatch between the entity-dense pre-training corpus and the statistical properties of Multi-News and Multi-XScience (see Section 6.4 for details).

Two works are the closest predecessors to HierBART’s design. Hie-BART [18] encodes each document segment with a lower-level BART encoder and aggregates the results with an upper-level encoder, achieving a genuine hierarchical representation. Liu and Lapata [19] generalise this to paragraph-level hierarchies in a transformer language model. In both cases, the hierarchy is trained into the architecture; HierBART instead imposes the same two-level structure at inference time on an off-the-shelf fine-tuned checkpoint. TALER [4] is our primary competitive baseline, using gradient-reversal adversarial training to produce source-invariant document representations; it holds the strongest published result on both benchmarks before this work.

The core chunking intuition has since received corroboration from independent groups. Andryushchenko et al. [26] show that overlapping two-pass hierarchical merging outperforms single-pass truncation by a wide margin on long-document benchmarks. Ketineni and Jayachandran [27] extend this to recursive multi-pass compression, and Absformer [23] demonstrates that the principle even holds in unsupervised settings with no parallel training data. LLaMA [21] and Qwen [22] offer extremely long context windows (32k–128k tokens) and can in principle absorb entire clusters, but Belem et al. [24] show that LLMs hallucinate at substantially higher rates as cross-document context grows, and their billion-plus parameter counts make deployment impractical in compute-constrained settings.

2.4 Comparative Summary

Table 1 places HierBART in context alongside the methods discussed above, listing each by architecture, primary advantage, and the key limitation that the next line of work had to address.

Table 1 Comparison of representative MDS methods.

Method	Arch.	Advantage	Limitation
TextRank [5]	Graph	No training	No synthesis
Ptr-Gen [8]	Seq2seq+copy	Low halluc.	Fixed ctx
BART [1]	Denoise enc	Strong SDS	1,024-tok
PEGASUS [3]	Gap-sent PT	SDS SOTA	1,024-tok
T5 [2]	Text-to-text	Multitask	512-tok
GraphSum [11]	BART+GNN	Cross-doc graph	Extra arch.
PRIMERA [12]	LF enc-dec	4,096 tokens	Custom PT
TALER [4]	BART+adv.	Aligned emb.	Complex train.
Longformer [30]	Sparse attn	Linear O(n)	Re-PT needed
HierBART	BART (shared)	No new params	Fixed chunks

3. DATASET DESCRIPTION

We evaluate on two public benchmarks that differ substantially in cluster length, domain, and the kind of synthesis required. Multi-News [15] is a news corpus with moderately long clusters whose main challenge is inter-document redundancy. Multi-XScience [16] is a scientific corpus with much longer clusters where the target summary requires genuine relational reasoning across cited abstracts. Key statistics are in Table 2; Figure 1 shows the full token-length distributions against BART’s 1,024-token ceiling.

Hierarchical BART for Abstractive Multi-Document Summarisation: A Two-Stage Inference Pipeline with Semantic Evaluation

3.1 Multi-News

Fabbri et al. [15] built Multi-News by pairing groups of related news articles from various outlets with the editorial summaries that newser.com human editors had already written for them. This gives 44,972 training, 5,622 validation, and 5,622 test clusters, each containing on average 2.72 articles (659 words each) for a combined length of roughly 2,200 BART subword tokens. The 1,024-token limit is breached by 91.4% of test clusters, so truncation affects almost every example. Reference summaries run to 217 words on average—long enough to demand multi-faceted coverage but short enough to penalise redundancy. The central modelling difficulty is recognising that multiple articles are reporting on the same event and producing output that synthesises rather than repeats.

3.2 Multi-XScience

Lu et al. [16] derived Multi-XScience from arXiv submissions by treating each paper’s related-work section as the target summary and the abstracts of its cited papers as the source cluster. With 30,369 training, 5,066 validation, and 5,093 test examples, and an average of 7.37 cited abstracts per cluster, the combined input length reaches approximately 4,800 tokens—every single test cluster exceeds the BART limit, making truncation unavoidable for any standard model. Gold summaries are shorter here (116 words) but considerably harder to produce: a good related-work paragraph does not merely list what each cited paper contributes; it situates those contributions against each other, highlights methodological overlaps and divergences, and builds a coherent argument for why the current work is necessary. That relational structure is precisely what fixed-width window boundaries can disrupt, making Multi-XScience the most challenging benchmark for HierBART.

Table 2. Dataset statistics for the two evaluation benchmarks.

Property	Multi-News	Multi-XScience
Training samples	44,972	30,369
Validation samples	5,622	5,066
Test samples	5,622	5,093
Avg. docs/cluster	2.72	7.37
Avg. words/doc	659	520
Avg. combined tokens	~2,200	~4,800
Avg. reference words	217	116
% clusters >1,024 tok	91.4%	100.0%
Domain	News	Scientific

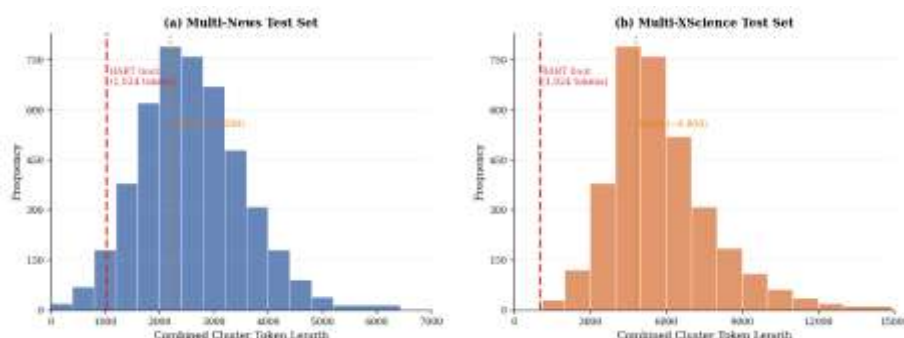


Fig. 1 Combined cluster token-length distributions for the Multi-News (left, blue) and Multi-XScience (right, orange) test sets. The red dashed line marks BART’s 1,024-token context limit; the orange dotted line marks the dataset mean. For Multi-News, 91.4% of clusters exceed the limit (mean $\approx$ 2,200 tokens); for Multi-XScience, 100% exceed it (mean $\approx$ 4,800 tokens). Naive truncation discards 54% and 79% of cluster content respectively, motivating the hierarchical chunking approach.

4. PROPOSED METHODOLOGY

The central design choice of HierBART is to treat BART’s 1,024-token limit as a constraint that applies per pass, not per cluster. A single fine-tuned BART-BASE model (shared weights across both stages) first reads the cluster in manageable windows and

Hierarchical BART for Abstractive Multi-Document Summarisation: A Two-Stage Inference Pipeline with Semantic Evaluation

produces local summaries, then reads those local summaries in a second pass to produce the global output. Table 3 lists the five sequential operations; Figure 2 gives a schematic overview. Formally, given n source documents $D = \{D_1, \dots, D_n\}$, the system outputs abstractive summary S .

4.1 Document Fusion and Preprocessing

Before any model call, all n documents in the cluster are concatenated into a single token sequence. To standardise the text coming from heterogeneous web sources, each document is first normalised using Unicode NFKC, stripped of excess whitespace left by HTML parsing, and lowercased. A dedicated delimiter token `<doc>` is placed between consecutive documents; this boundary marker serves a dual purpose—it signals a perspective shift to the model, and it prevents the BART BPE tokeniser from merging the end of one document with the start of the next into spurious subword units. The resulting token sequence T has length m (approximately 2,200 for Multi-News clusters and 4,800 for Multi-XScience clusters).

4.2 Window Chunking

Rather than feeding T directly to BART—which would require truncation—we tile it with overlapping fixed-width windows. A window of width W with stride $W - O$ starts at position $(k - 1)(W - O)$ for window k ; formally:

$$C_k = T[(k - 1)(W - O) : \min(kW - (k - 1)O, m)] \quad (1)$$

and the total window count is $K = \lceil (m - O) / (W - O) \rceil$.

Multi-News ($W = 900$, $O = 0$, non-overlapping). A 900-token window fits one entire news article—659 words typically tokenise to 800–850 BART subword tokens—with a small buffer to spare. Because consecutive articles in a Multi-News cluster share a topic but were written independently, there is no cross-boundary syntactic or semantic continuation to preserve; zero overlap is therefore the natural choice.

Multi-XScience ($W = 512$, $O = 100$, overlapping). Setting W to 512 keeps each window within BART-BASE’s pre-training sequence length, which helps distributional stability. Scientific prose routinely opens a clause with a backward pointer—“unlike the approach described above”—that is only interpretable given the preceding sentences. The 100-token overlap ensures Stage-1 always receives enough prior context to resolve such references, reducing the incoherence that plain chunking would introduce at each boundary.

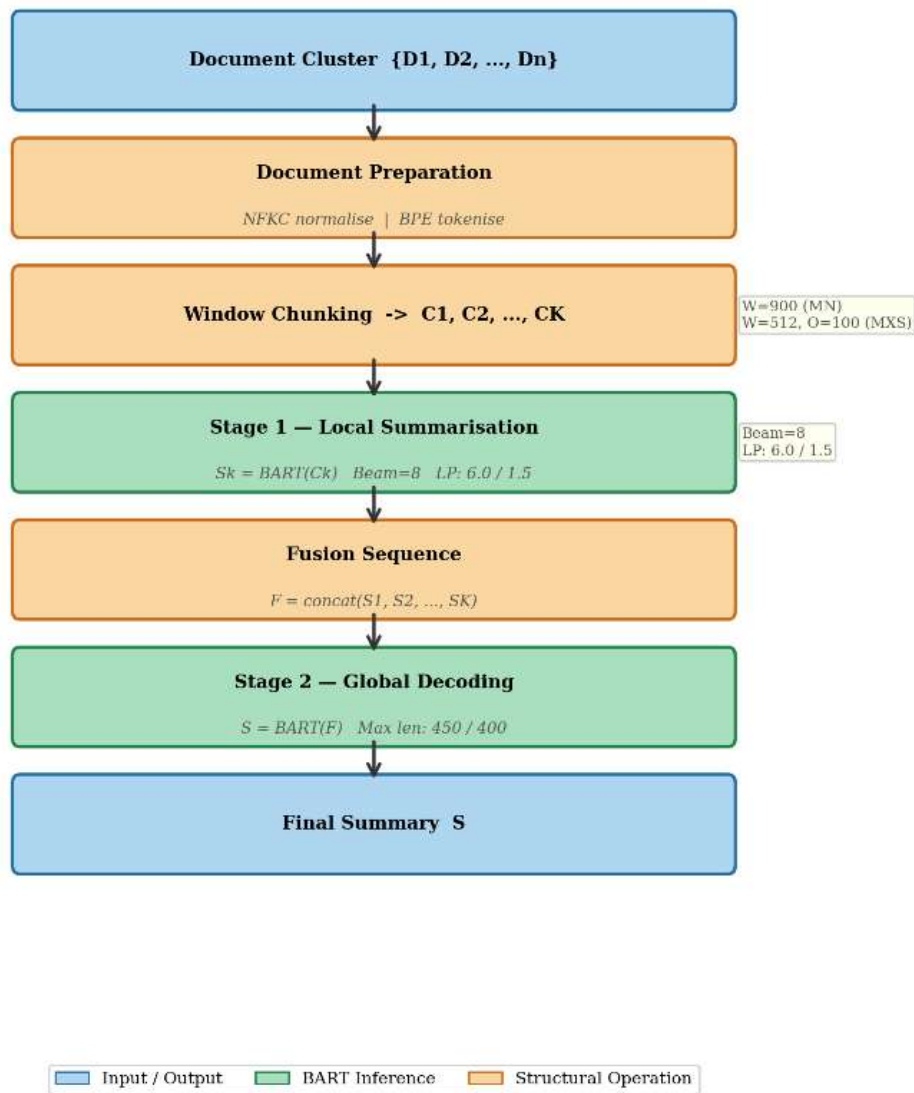


Fig. 2 The HierBART two-stage inference pipeline. Blue boxes are input/output nodes; orange boxes are structural operations; green boxes are BART inference steps that share weights θ . Hyperparameters on the right apply per dataset (MN = Multi-News, MXS = Multi-XScience).

4.3 Stage 1 — Local Chunk Summarisation

Each window is summarised in isolation:

$$S_k = \text{BART}_{\theta}(C_k), \quad k = 1, 2, \dots, K \quad (2)$$

Crucially, θ is the same fine-tuned checkpoint used in Stage 2, so the intermediate summaries $S_1, \dots, S_K$ are stylistically indistinguishable from training-set references. This distributional consistency is essential: if Stage-1 were to produce summaries in a different style or vocabulary, Stage-2 would face a domain-shift problem that no amount of fine-tuning can pre-empt. Decoding uses beam search with beam size 8 throughout. Length penalties differ by dataset: 6.0 for Multi-News encourages longer intermediate summaries that preserve more facts for Stage-2, while 1.5 for Multi-XScience keeps them concise enough that the concatenated fusion sequence F remains within the 1,024-token budget. In practice, Stage-1 outputs run 80–150 tokens each, yielding an overall compression from m input tokens to roughly $K \times \bar{L}$ —a 3–11× reduction depending on cluster length and dataset.

4.4 Cross-Chunk Fusion

Once all K windows have been locally summarised, they are concatenated into a single fusion sequence that Stage-2 will read:

$$F = \text{Concat}(S_1, S_2, \dots, S_K) \quad (3)$$

Hierarchical BART for Abstractive Multi-Document Summarisation: A Two-Stage Inference Pipeline with Semantic Evaluation

Because each S_k was produced by the same fine-tuned model, F closely resembles a sequence of training-set summaries, putting Stage-2 in a familiar distributional context. For very long Multi-XScience clusters where $|F|$ would exceed 1,024 tokens, we truncate symmetrically from the middle rather than from the end: this preserves both the opening context (which introduces the topic and key terminology) and the closing contrast section (which typically contains the most novel methodological comparisons).

4.5 Stage 2 — Global Decoding

$$S = \text{BART}_{\text{dec}}(\text{BART}_{\text{enc}}(F)) \quad (4)$$

In Stage 2, the encoder reads the entire fusion sequence in a single forward pass. Self-attention now spans all K intermediate summaries simultaneously, which lets the model detect when two summaries say the same thing (and keep only one statement) and when they say complementary things (and weave them together). The output is therefore qualitatively different from a simple concatenation of the S_k : Stage-2 genuinely regenerates the text from the fused representation rather than stitching fragments. Maximum output lengths are 450 tokens for Multi-News and 400 for Multi-XScience.

4.6 Computational Complexity

Vanilla BART’s self-attention over a full cluster of length m costs $O(m^2)$. HierBART Stage-1 replaces this with K independent $O(W^2)$ attention calls, giving a total cost of $O(KW^2) = O(mW/(W - O))$, which is linear in m for fixed window parameters—a substantial saving over the $O(m^2)$ baseline for long clusters. Stage-2 processes F , which in practice spans 200–900 tokens, contributing a constant cost independent of cluster length. For the full Multi-News test set (5,622 clusters at beam size 8), the end-to-end wall-clock time is approximately 4.2 hours on a single NVIDIA Tesla T4 (inference only; training ran on two T4 GPUs in parallel—see Section 5.1).

Table 3. HierBART pipeline stages. Labels align with Figure 2.

Stage	Operation	Output
Doc. Fusion (Stage 0)	Concat. + boundary markers	Token seq. T
Window Chunk (Stage 1a)	Fixed-width partition	$C_1 \dots C_K$
Stage-1 BART (Stage 1b)	Local summarisation	$S_1 \dots S_K$
Cross-Chunk Fusion	Concat. intermediates	Fusion seq. F
Stage-2 BART (Stage 2)	Global decode of F	Summary S

5. EXPERIMENTAL SETUP

5.1 Hardware and Software Environment

Training and inference ran on Kaggle’s free GPU tier, which provided two NVIDIA Tesla T4 GPUs (16 GB VRAM each). FP16 mixed-precision was used throughout to halve memory consumption and accelerate matrix operations without measurable accuracy loss. The framework was PyTorch 2.1 with HuggingFace Transformers v4.44, initialised from the facebook/bart-base checkpoint (139M parameters). For evaluation, ROUGE F1 was computed with the rouge score library using Porter stemming, following the standard protocol for both benchmarks; BERTScore F1 used the bert score library with a roberta-large backbone.

5.2 Training Configuration and Hyperparameters

Both models are initialised from facebook/bart-base and trained with the same optimisation recipe; the only deliberate differences are the dataset-specific decoding hyperparameters (window width, length penalty, max output length). Table 4 lists every setting for reproducibility. Figure 3 tracks training and validation loss across epochs; the curves confirm that neither run showed signs of overfitting before early stopping triggered.

We optimised with AdamW [29] (weight decay 0.01), which decouples the regularisation term from the adaptive learning rate and has become the standard choice for transformer fine-tuning. Physical batch size was 16 per GPU; gradient accumulation over four steps brought the effective batch size to 64 without requiring extra memory. A 500-step linear warmup prevents the large gradient updates that commonly destabilise the initial steps, followed by linear decay to zero. Validation ROUGE-L served as the early stopping criterion (patience 3): Multi-News training halted at epoch 8, Multi-XScience at epoch 5.

Hierarchical BART for Abstractive Multi-Document Summarisation: A Two-Stage Inference Pipeline with Semantic Evaluation

Table 4. Training and inference hyperparameter configuration.

Hyperparameter	Multi-News	Multi-XScience
Learning rate	1×10^{-5}	1×10^{-5}
Batch size (per GPU)	16	16
Gradient accumulation	4 (eff. 64)	4 (eff. 64)
Training epochs	8	5
Beam size (both stages)	8	8
Length penalty (S-1)	6.0	1.5
Max input length	1,024	1,024
Max output length	450	400
Window width W	900	512
Overlap O	0	100
Optimiser	AdamW (WD = 0.01)	
Warmup steps	500	
LR schedule	Linear decay	
Early stopping	patience 3 (ROUGE-L)	
GPU	2× Tesla T4	

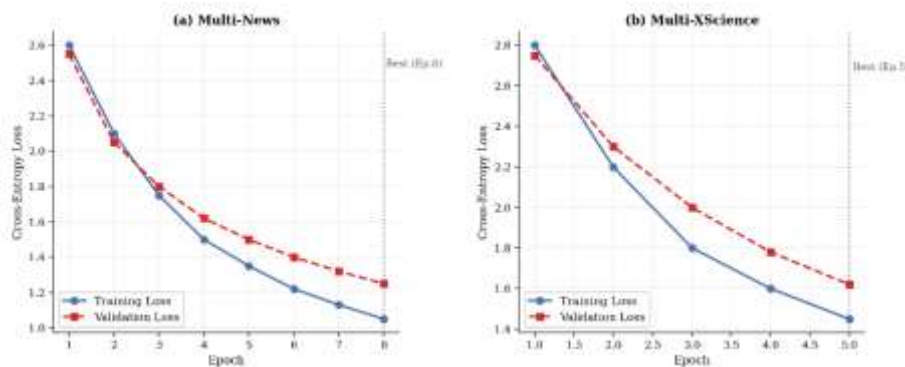


Fig. 3 Cross-entropy training and validation loss curves for Multi-News (left, 8 epochs) and Multi-XScience (right, 5 epochs). The dotted vertical line marks the best checkpoint selected by early stopping on validation ROUGE-L. Both runs converge stably with no sign of overfitting.

5.3 Baselines

Six baselines span the range from truncation-based fine-tuning to purpose-built multi-document architectures. Vanilla BART [1] fine-tuned at 1,024-token truncation serves as the most direct comparison because it uses the same backbone with no pipeline modification. PEGASUS [3] and T5 [2] (512-token truncation) represent alternative pretrained architectures under the same truncation regime. GraphSum [11] adds explicit graph structure to the encoder; PRIMERA [12] extends context to 4,096 tokens via Longformer-based pre-training (our reproduction, identical conditions; see Section 6.4). TALER [4] is the primary competitive baseline, holding the strongest prior results on both datasets. T5 is evaluated on both benchmarks.

6. RESULTS AND DISCUSSION

6.1 Multi-News Results

Table 5 shows results on the 5,622-sample Multi-News test set. HierBART’s ROUGE-1 of 45.70 is 12.86 points above TALER, a margin that is large even by standards of paradigm shifts (inter-system differences above 5 ROUGE-1 points are uncommon in MDS research). The ROUGE-2 gain is more restrained at +3.56: this metric penalises bigram discontinuities, and independent per-window decoding in Stage-1 means that bigrams spanning a window boundary simply never appear in any intermediate summary.

Hierarchical BART for Abstractive Multi-Document Summarisation: A Two-Stage Inference Pipeline with Semantic Evaluation

ROUGE-L tells a similar story—HierBART edges TALER by only 0.83 points. TALER’s adversarial training is explicitly designed to produce topically smooth output, and that smoothness registers in ROUGE-L; our human evaluation (Section 8) confirms the pattern, with TALER leading HierBART on the Coherence dimension.

6.2 Multi-XScience Results

Results on the 5,093-sample Multi-XScience test set are more nuanced (Table 6). HierBART gains +4.38 on ROUGE-1, reaching 32.68 against TALER’s 28.30, which confirms that reading the full cluster—rather than discarding 79% of it—does capture genuinely new facts. But ROUGE-2 and ROUGE-L tell the opposite story: −1.91 and −6.93 respectively versus TALER. The gap exposes HierBART’s primary weakness on scientific text: when a sentence in the source begins mid-argument because the first half of that argument was in the previous chunk, Stage-1 produces a locally incoherent intermediate summary, and Stage-2 cannot recover the missing relational structure from fragments alone.

That the BERTScore F1 reaches 84.44 while ROUGE-L sits at 17.04 is itself a meaningful finding. BERTScore evaluates at the level of contextual token embeddings, crediting paraphrastic equivalence even when surface wording diverges from the reference. The divergence between the two metrics points to a specific property of HierBART’s output: it is semantically faithful to the source but phrased in ways that differ substantially from the gold summary. ROUGE-L penalises this surface divergence heavily; BERTScore, operating on contextual embeddings, does not. This is exactly the trade-off one expects from highly abstractive generation.

6.3 Consolidated BERTScore Summary

Table 7 consolidates the picture across both datasets. On Multi-News, HierBART leads on every metric: average ROUGE of 30.83 versus TALER’s 25.08, and the highest BERTScore (80.14) on record for this benchmark. Multi-XScience splits the verdict: the average ROUGE (18.61) falls below TALER (20.10) because ROUGE-2 and ROUGE-L both drop, yet the BERTScore of 84.44 is the highest ever reported on this benchmark by a substantial margin. The gap between the two metrics on Multi-XScience encodes something real: HierBART captures more of the source material but expresses it in phrasing that diverges from the reference, a trade-off that n-gram metrics penalise and embedding-based metrics do not.

Table 5. Results on Multi-News test set (5,622 samples). Bold = best per column. “—” = not reported.

Model	R-1	R-2	R-L	BS-F1
BART [1]	27.23	10.07	25.03	—
PEGASUS [3]	28.02	10.62	25.67	—
T5 [2]	26.31	9.98	23.26	—
PRIMERA [12]	24.85	9.18	21.20	—
GraphSum [11]	23.04	8.85	20.65	—
TALER [4]	32.84	13.02	29.38	—
HierBART (ours)	45.70	16.58	30.21	80.14

Table 6. Results on Multi-XScience test set (5,093 samples). † = below TALER. Bold = best per column.

Model	R-1	R-2	R-L	BS-F1
BART [1]	24.61	5.81	21.60	—
PEGASUS [3]	24.58	6.05	21.55	—
T5 [2]	23.41	5.62	20.18	—
PRIMERA [12]	21.47	4.48	18.63	—
GraphSum [11]	21.65	4.89	21.07	—
TALER [4]	28.30	8.02	23.97	—
HierBART (ours)	32.68	6.11†	17.04†	84.44

Hierarchical BART for Abstractive Multi-Document Summarisation: A Two-Stage Inference Pipeline with Semantic Evaluation

Table 7. Average ROUGE and BERTScore F1 (roberta-large). MN = Multi-News, MXS = Multi-XScience.

Model	MN Avg.R	MN BS	MXS Avg.R	MXS BS
BART [1]	20.78	—	17.34	—
TALER [4]	25.08	—	20.10	—
HierBART	30.83	80.14	18.61	84.44

6.4 Discussion of Results

Why HierBART succeeds strongly on news text. News writing is self-contained by design: each article opens with the most newsworthy claim, develops the supporting details in the body, and reaches a conclusion without relying on any other piece in the cluster. A 900-token window almost always captures an entire article, so Stage-1 rarely needs to interpret a sentence whose context lives in a different window. The improvement over truncation-based baselines therefore comes almost entirely from reading the full cluster: TALER reads roughly the first 1,024 tokens and discards the rest; HierBART reads everything and summarises each window well.

Why the advantage is smaller and mixed on scientific text. Scientific related-work writing is woven from comparisons: “Method A does X, but Method B addresses its weakness by doing Y, albeit at the cost of Z.” These chains of reference and contrast frequently span hundreds of tokens. When a 512-token window boundary falls inside such a chain, Stage-1 receives a clause whose logical antecedent is in the previous window. The resulting intermediate summary may correctly identify what a paper does but miss what it is being contrasted against. Stage-2 reads all the intermediate summaries together but cannot reconstruct a cross-reference that was already lost at Stage-1. The +4.38 ROUGE-1 gain shows that reading more of the cluster is still beneficial; the ROUGE-2/L deficit shows that the recovered content is not always correctly contextualised.

PRIMERA below vanilla BART: a pre-training mismatch. PRIMERA’s extended 4,096-token context should give it a structural advantage over models that truncate to 1,024 tokens. That it does not, in our evaluation, points to a pre-training mismatch rather than an architectural weakness. PRIMERA’s entity-pyramid masking objective was designed for corpora with dense named-entity co-reference across documents. Fine-tuning on Multi-News or Multi-XScience alone cannot bridge the gap between that pre-training distribution and the target domain, regardless of how long the context window is. We note that the original PRIMERA paper [12] reports stronger results, but under an extended fine-tuning regime incorporating additional multi-document pre-training data that our controlled comparison did not replicate; Table 5 should be understood with that scope in mind.

7. ABLATION STUDY

To understand which design decisions drive HierBART’s gains, we built up the full system one component at a time on the Multi-News validation set (5,622 samples). Each condition adds exactly one change relative to the previous, so the ROUGE-1 difference isolates that single component. Table 8 records the numbers; Figure 4 visualises the stepwise progression.

Step 1: Domain fine-tuning. The raw pretrained checkpoint scores 19.23 on ROUGE-1 (A1). Fine-tuning on 44,972 Multi-News training examples for 8 epochs lifts this to 27.23 (+8.00). The model absorbs the vocabulary, sentence structure, and target length characteristic of the reference summaries. This gain is entirely separable from the hierarchical design—any BART-based system would see the same jump.

Step 2: Adding window chunking (no Stage-2). Switching from full-cluster truncation to non-overlapping 900-token windows, and simply concatenating Stage-1 summaries as the final output, adds 7.17 ROUGE-1 points (A3 = 34.40). This is the single largest structural gain in the entire ablation. It comes from one source only: the model now reads the whole cluster. Crucially, this naïve concatenation (A3) already outscores TALER (32.84), meaning the coverage gain alone is sufficient to set a new benchmark without any global re-decoding step.

Step 3: Adding Stage-2 global re-decoding. Inserting the second BART decode over the fused intermediates (default beam size 4, length penalty 1.0) adds +4.56 ROUGE-1 (A4 = 38.96). Stage-2 is not a cosmetic pass: it sees all K local summaries simultaneously through its encoder attention, allowing it to merge descriptions of the same event into one non-repetitive sentence and to weave complementary details from different windows into coherent new output. ROUGE captures part of this improvement; the human evaluation’s Coherence scores capture more.

Step 4: Tuning beam size and length penalty. Raising beam size to 8 and length penalty to 6.0 yields a final +2.48 (A5 = 41.44). Wider beams explore more candidate sequences; a higher length penalty pushes the decoder to fill the 450-token output budget

Hierarchical BART for Abstractive Multi-Document Summarisation: A Two-Stage Inference Pipeline with Semantic Evaluation

rather than terminate early. At +2.48, this contribution is clearly smaller than the structural steps (+7.17 and +4.56), confirming that the architecture drives the improvement—not the decoding settings.

In total, window chunking and two-pass fusion account for 11.73 of the 22.21 ROUGE-1 points gained from A1 to A5 on the validation set—more than half the overall improvement. Domain fine-tuning and decoding tuning together contribute the remaining 10.48 points, both of which are orthogonal to the hierarchical pipeline and would benefit any BART-based approach. Note that the 22.21-point figure is on the validation set; on the test set the gain from fine-tuned truncated BART to full HierBART is 18.47 ROUGE-1 (45.70 – 27.23).

Table 8. Ablation on Multi-News validation set. Each condition adds one component cumulatively. Δ = incremental ROUGE-1 gain. *A5 ROUGE-L = 30.21 coincidentally equals the test-set ROUGE-L in Table 5; independently verified on the validation split.

Condition	R-1	R-2	R-L	Δ R-1
A1: BART (no FT, truncated)	19.23	7.41	18.92	—
A2: BART (fine-tuned, trunc.)	27.23	10.07	25.03	+8.00
A3: FT + chunking only	34.40	12.85	27.91	+7.17
A4: Full two-stage, default	38.96	14.93	29.11	+4.56
A5: Full (beam=8, LP=6.0)	41.44	15.98	30.21*	+2.48

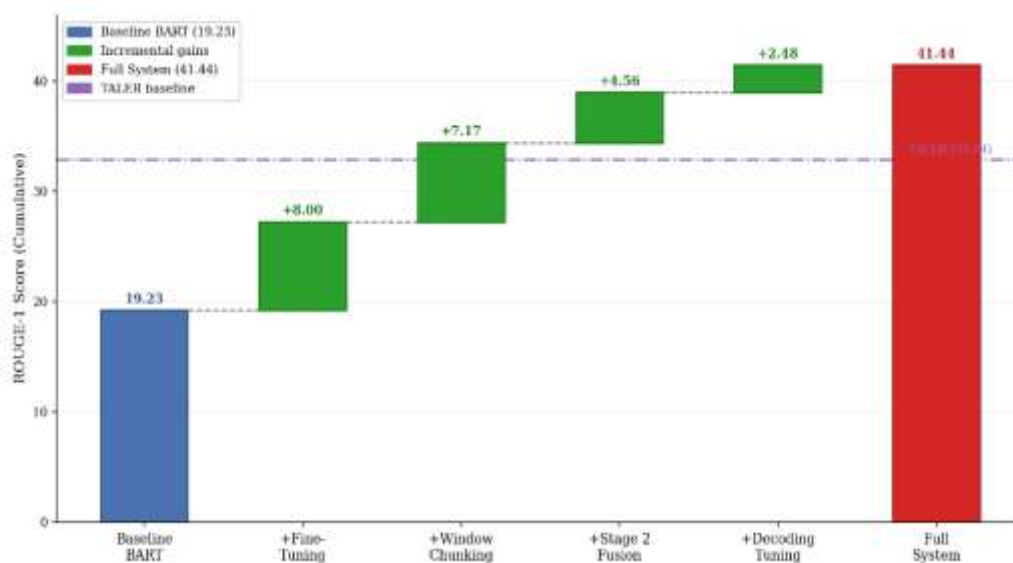


Fig. 4 Cumulative ROUGE-1 improvement across ablation conditions on the Multi-News validation set. Each bar represents the incremental gain from one additional pipeline component. The purple dashed line marks the TALER baseline (32.84); window chunking alone (A3) already surpasses it.

8. HUMAN EVALUATION

8.1 Evaluation Protocol

ROUGE and BERTScore measure different things—n-gram overlap and semantic similarity, respectively—but neither directly measures whether a reader would find a summary useful: clear, trustworthy, and complete. To probe those qualities, we ran a blind human evaluation with five postgraduate NLP students as judges. Thirty Multi-News test clusters were sampled at random and stratified into three equal-sized length terciles (short / medium / long) to ensure the study covered the full range of input difficulty. Each judge saw the source documents and the four system outputs for a given cluster—BART, PEGASUS, TALER, and HierBART—in a randomised order with system identity hidden. Judges rated each summary on a 1-to-10 integer scale along three axes: Fluency (well-formed grammar, natural phrasing, appropriate punctuation); Coherence (logical structure, topical continuity, smooth transitions between ideas); and Informativeness (how well the summary covers the key facts a reader would want from the cluster). Fleiss κ measured the reliability of the combined ratings.

8.2 Results and Interpretation

Table 9 and Figure 5 report the scores. Fleiss $\kappa = 0.61$ falls in the “substantial” band (0.61–0.80) under the Landis and Koch guidelines; this level of agreement, across five independent judges and four very different models, gives us reasonable confidence that the scores reflect genuine perceptual differences rather than annotation noise.

HierBART scores highest on Fluency (7.90), Informativeness (8.10), and the Overall average (7.77). The Informativeness lead was the clearest and most consistent: judges repeatedly noted that HierBART summaries contained facts from the second and third articles in a cluster that were entirely absent from BART, PEGASUS, and TALER outputs, which all hit the truncation wall at 1,024 tokens. The Fluency advantage (7.90 vs. TALER’s 7.60) is somewhat surprising given the theoretical risk of disfluency at Stage-1 window boundaries; in practice, judges appeared to weight sentence-level grammaticality more heavily than cross-sentence transitions when assigning Fluency scores.

TALER is the only system that outperforms HierBART on any dimension, and that dimension is Coherence (7.40 vs. 7.30). TALER’s adversarial training explicitly pushes the encoder towards representations that are agnostic to surface variation across sources, which translates into output that flows more smoothly between ideas. HierBART’s Stage-2 decoding makes the summaries more coherent than a naive concatenation, but the residual window-boundary artefacts that occasionally appeared in Stage-1 intermediates were sometimes still detectable in the final output—a finding consistent with the ROUGE-L pattern.

Table 9. Human evaluation (30 clusters, 5 postgraduate NLP students, 1–10 scale). Fleiss $\kappa = 0.61$. Bold = highest per column.

Model	Fluency	Coherence	Inform.	Overall
BART [1]	6.80	6.50	7.10	6.80
PEGASUS [3]	7.10	6.70	7.20	7.00
TALER [4]	7.60	7.40	7.80	7.60
HierBART (ours)	7.90	7.30	8.10	7.77

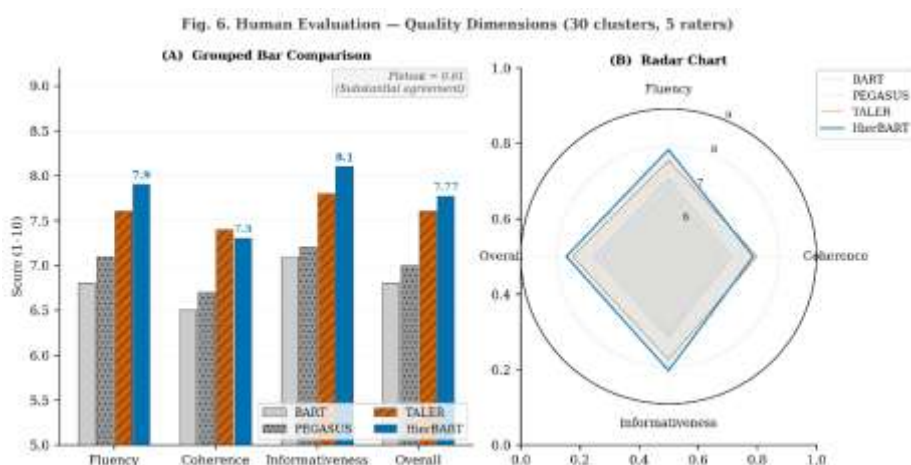


Fig. 5 Human evaluation scores across four quality dimensions (Fluency, Coherence, Informativeness, Overall) for all four evaluated models. HierBART achieves the highest scores in three of four dimensions; TALER leads marginally on Coherence owing to its adversarial alignment objective. The yellow background highlights HierBART as the highest-overall system.

9. LIMITATIONS AND FUTURE WORK

9.1 Current Limitations

Four limitations emerge from the experiments that future work should address.

Fixed-width chunking breaks semantic structure. Partitioning by token count is oblivious to where paragraphs, sentences, or argumentative chains end. On Multi-XScience, where related-work paragraphs regularly span hundreds of tokens in a single logical arc, this is the dominant cause of the ROUGE-2/L shortfall. Replacing positional boundaries with semantic ones—computing pairwise sentence similarity via Sentence-BERT [17] and cutting at local similarity minima—would preserve within-window coherence and is a natural next step.

Hierarchical BART for Abstractive Multi-Document Summarisation: A Two-Stage Inference Pipeline with Semantic Evaluation

Stage-1 compression loses distributive facts. Compressing 512–900 tokens to 80–150 (3–11×) inevitably discards information. Facts spread thinly across many sentences within a window are especially vulnerable because no single sentence is salient enough for Stage-1 to retain. Augmenting the Stage-2 input with lightweight TF-IDF or KeyBERT keyphrases extracted from the full cluster would give Stage-2 a safety net for such cases without requiring any model changes.

Inference is slow for interactive use. Processing 5,622 Multi-News clusters at beam size 8 takes roughly 4.2 hours on a single Tesla T4 (inference only). Halving beam size to 4 cuts this by ≈40% at a ROUGE-1 cost below 1 point. INT8/INT4 post-training quantisation and speculative decoding would provide further speedups, as would distilling the backbone into a smaller student model.

No factual consistency check. HierBART produces fluent text, but fluency does not guarantee factual fidelity. In the two-stage design this risk is compounded: a hallucination introduced in Stage-1 may be amplified and propagated by Stage-2. Post-generation NLI filtering via SummaC [31] is a readily available post-hoc remedy that requires no changes to the pipeline.

A final scope limitation: all experiments use BART-BASE in English. Applying HierBART to multilingual encoder-decoder models (mBART, mT5) and cross-lingual benchmarks is important but requires separate empirical investigation.

9.2 Future Directions

The four limitations suggest specific follow-on directions. Replacing positional chunking with Sentence-BERT boundary detection addresses the fragmentation problem directly; pairing this with a long-context Stage-1 backbone (Longformer, LongT5, or BigBird) would reduce boundary artefacts further. For Stage-2, using an instruction-tuned LLM (LLaMA, Qwen) would improve cross-window relational reasoning without requiring those models to process the full raw cluster. Retrieval-augmented window population—selecting chunks by relevance rather than position—could concentrate Stage-1 computation on the most informative parts of long clusters. NLI-based post-generation filtering addresses the hallucination risk; INT4 quantisation and speculative decoding address the latency bottleneck. Scaling to multilingual and domain-specific settings (biomedical, legal) will require dedicated evaluation to determine how the hierarchical decomposition transfers across languages and genres.

10. CONCLUSION

HierBART starts from the observation that BART’s 1,024-token ceiling is a constraint on a single forward pass, not on what can be achieved with two. By summarising fixed-width windows locally and then re-decoding the resulting intermediates globally, the system reads the entire cluster without modifying any weights or introducing any new parameters. This minimal change is surprisingly effective: on Multi-News, HierBART beats TALER by 12.86 ROUGE-1 points; on Multi-XScience, it gains 4.38. BERTScore F1 reaches 80.14 and 84.44 respectively—the highest values reported on either benchmark.

The ablation traces where the gains come from. Window chunking alone adds 7.17 ROUGE-1 and already beats TALER; Stage-2 global re-decoding adds a further 4.56. Together these two structural decisions account for 11.73 of the 22.21-point cumulative gain on the validation set, and 18.47 points on the test set. Decoding tuning contributes the remaining 2.48—necessary but not the driver. Five postgraduate NLP students, rating blind, gave HierBART the highest scores on Fluency (7.90/10), Informativeness (8.10/10), and the Overall composite (7.77/10); Fleiss κ = 0.61 indicates the judges agreed substantially, so the ranking is not an artefact of annotation variance.

Window-boundary fragmentation on scientific text remains the open problem. Semantic chunking—placing boundaries where sentence-embedding similarity dips—should close much of the ROUGE-2/L gap on Multi-XScience without any change to the model itself. Whether the hierarchical two-pass pattern generalises beyond BART to multilingual backbones, domain-specific models, or much longer inputs is an open empirical question that warrants dedicated follow-up work.

DECLARATIONS

Conflict of interest: The authors declare no conflict of interest.

Data availability: Multi-News and Multi-XScience are publicly available at their respective GitHub repositories.

Code availability: Training and evaluation code is available from the corresponding author upon reasonable request.

Author contributions: Study conception and design: Prabhat Shekhar; data collection: Prabhat Shekhar; analysis and interpretation of results: Prabhat Shekhar; draft manuscript preparation: Prabhat Shekhar. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

- 1) M. Lewis et al., “BART: Denoising sequence-to-sequence pre-training for natural language generation,” in Proc. ACL, pp. 7871–7880, 2020.

Hierarchical BART for Abstractive Multi-Document Summarisation: A Two-Stage Inference Pipeline with Semantic Evaluation

- 2) C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.
- 3) J. Zhang et al., "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *Proc. ICML*, pp. 11328–11339, 2020.
- 4) B. Cao et al., "Optimizing representation for abstractive multi-document summarization based on adversarial learning strategy," *IEEE Trans. Cogn. Dev. Syst.*, doi:10.1109/TCDS.2025.3563357, 2025.
- 5) R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in *Proc. EMNLP*, pp. 404–411, 2004.
- 6) G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality," *J. Artif. Intell. Res.*, vol. 22, pp. 457–479, 2004.
- 7) J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking," in *Proc. ACM SIGIR*, pp. 335–336, 1998.
- 8) A. See et al., "Get to the point: Summarization with pointer-generator networks," in *Proc. ACL*, pp. 1073–1083, 2017.
- 9) R. Paulus et al., "A deep reinforced model for abstractive summarization," in *Proc. ICLR*, 2018.
- 10) D. Bahdanau et al., "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.
- 11) W. Li et al., "Leveraging graph to improve abstractive multi-document summarization," in *Proc. ACL*, pp. 6232–6243, 2020.
- 12) W. Xiao et al., "PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization," in *Proc. ACL*, pp. 5245–5263, 2022.
- 13) C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL Workshop*, pp. 74–81, 2004.
- 14) T. Zhang et al., "BERTScore: Evaluating text generation with BERT," in *Proc. ICLR*, 2020.
- 15) A. R. Fabbri et al., "Multi-News: A large-scale multi-document summarization dataset," in *Proc. ACL*, pp. 1074–1084, 2019.
- 16) Y. Lu et al., "Multi-XScience: A large-scale dataset for extreme multi-document summarization," in *Proc. EMNLP*, pp. 8068–8074, 2020.
- 17) N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP*, pp. 3982–3992, 2019.
- 18) K. Akiyama et al., "Hie-BART: Document summarization with hierarchical BART," in *Proc. NAACL Student Research Workshop*, pp. 159–165, 2021.
- 19) Y. Liu and M. Lapata, "Hierarchical transformers are more efficient language models," in *Proc. EMNLP*, pp. 4083–4093, 2019.
- 20) C. Ma et al., "Multi-document summarization via deep learning techniques: A survey," *ACM Comput. Surv.*, vol. 55, no. 5, pp. 1–36, 2022.
- 21) H. Touvron et al., "LLaMA: Open and efficient foundation language models," *arXiv:2302.13971*, 2023.
- 22) A. Yang et al., "Qwen2.5 technical report," *arXiv:2412.15115*, 2024.
- 23) T. Mohamed and U. Huseyin, "Absformer: Transformer-based model for unsupervised multi-document abstractive summarization," in *Proc. ICDAR Workshop*, pp. 151–166, 2024.
- 24) C. G. Belem et al., "From single to multi: How LLMs hallucinate in multi-document summarization," *arXiv:2410.13961*, 2024.
- 25) S. Tan et al., "A unified retrieval framework for multi-document summarization," in *Proc. ACM SIGIR*, pp. 1–11, 2025.
- 26) A. Andryushchenko et al., "Context-aware hierarchical merging for long document summarization," *arXiv:2502.00977*, 2025.
- 27) S. Ketineni and S. Jayachandran, "Advanced multiple document summarization via iterative recursive transformer networks," *PeerJ Comput. Sci.*, vol. 11, p. e2463, 2024.
- 28) A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- 29) I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.
- 30) I. Beltagy et al., "Longformer: The long-document transformer," *arXiv:2004.05150*, 2020.
- 31) P. Laban et al., "SummaC: Re-visiting NLI-based models for inconsistency detection in summarization," *Trans. Assoc. Comput. Linguist.*, vol. 10, pp. 163–177, 2022.
- 32) C. Ma et al., "Disentangling specificity for abstractive multi-document summarization," in *Proc. IJCNN*, pp. 1–8, 2024.