Divide-Prompt-Refine: a Training-Free, Structure-Aware Framework for Biomedical Abstract Generation

Sylvey Lin[1], Joe Menke[1], Shufan Ming [1], Dongin Nam[1], Neil Smalheiser[1,2], Halil Kilicoglu[1],

[1] School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL
[2] Department of Psychiatry, University of Illinois College of Medicine, Chicago, IL

Correspondence: yuhsinl2@illinois.edu

Abstract

Biomedical abstracts play a critical role in downstream NLP applications, such as information retrieval, biocuration, and biomedical knowledge discovery. However, a non-trivial number of biomedical articles do not have abstracts, diminishing the utility of these articles for downstream tasks. We propose DPR-BAG (Divide, Prompt, and Refine for Biomedical Abstract Generation), a training-free, zero-shot framework that generates coherent and factually grounded abstracts for biomedical articles with full text but no abstract. DPR-BAG decomposes full-text documents into structured rhetorical facets following the Background-Objective-Methods-Results-Conclusions (BOMRC) schema, performs parallel LLM-based summarization for each facet, and applies a final refinement stage to restore global discourse coherence. On PMC-MAD, a distribution-aligned dataset of 46,309 biomedical articles, DPR-BAG improves abstractive novelty over strong extractive and fine-tuned baselines, while maintaining factual consistency. Our ablation study reveals a counterintuitive finding: increasing prompt complexity or explicitly injecting entity-level guidance can degrade factual alignment, highlighting the importance of controlled prompting strategies. These findings underscore the potential of training-free, structure-aware frameworks for scalable biomedical abstract generation in low-resource settings. Our data and code are available at https://huggingface.co/datasets/pmc-mad/PMC-MAD and https://github.com/ScienceNLP-Lab/MultiTagger-v2/tree/main/DPR-BAG.

1 Introduction

Many biomedical NLP tasks rely heavily on abstracts, due to their accessibility and information density. Abstracts provide an author-written summary of core scientific findings, making them a useful proxy for full-text articles in downstream applications. For example, Luo et al. (2022) showed that pre-training tasks designed around the title-abstract structure improve biomedical information retrieval; Wiegers et al. (2025) used abstracts as an initial data source for biocuration. Beyond content, the structured organization of abstracts also benefits downstream tasks: Ueda et al. (2021) leverage abstract-level structure to refine retrieval; PubMedQA (Jin et al., 2019) is based on structured abstracts to support high-fidelity biomedical knowledge discovery. Moreover, abstracts alone can serve as a stronger training signal than full texts in some settings (Gu et al., 2021).

However, the absence of abstracts in a significant portion of biomedical articles creates a bottleneck for these tasks. As of April 2026, 11,603,796 out of 40,414,072 (~29%) PubMed articles were missing abstracts, with this volume continuing to rise despite a declining overall proportion, driven by the growth of publication types such as case reports, editorials, and letters. These publication types carry substantial scientific value. For instance, Gurulingappa et al. (2012) and Fan et al. (2020) utilized case reports for adverse drug event detection; Magnet and Carnet (2006) and Nuzzo (2021) assessed letters to characterize post-publication scientific discourse, including patterns of critique, rhetorical features of disagreement, and trends in authorship; and Waaijer et al. (2011) and Ioannidis and Schippers (2025) analyzed editorials to study how journals shape scientific discourse, including the distribution of topics, the framing of policy issues, and the presence of systematic biases.

The absence of abstracts in these articles motivates the Biomedical Abstract Generation (BAG) task, which aims to automatically generate abstracts from full-text biomedical articles. Although related to standard document summarization, BAG differs in important ways. It must adhere to scientific reporting conventions, including structured

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 770–790

presentation of methods, results, and conclusions, while preserving fine-grained biomedical entities, quantitative findings, and explicit argumentative relationships that are often critical for scientific interpretation. Early BAG work by Chachra et al. (2016) utilized extractive sentence selection, which can lead to fragmented coherence and poor lexical flow. Moreover, because full-length biomedical articles often exceed the context limits of standard models, BAG is inherently a long-context task, making it vulnerable to extractive bias and factual fidelity issues. For example, Wang et al. (2025) demonstrates that even state-of-the-art models like GPT-4 suffer from hallucinations and information omission when extracting from non-decomposed scientific full texts, emphasizing the inherent fidelity risks in long-document processing required for BAG task. Beyond fidelity risks, recent analysis also reveals that when forced to process complex long full texts, even specialized models like LongT5 exhibit a strong extractive bias, relying on simple heuristics to copy verbatim snippets rather than synthesizing information (Chernyshev and Dobrov, 2024). As a result, these models can suffer from the same core issue as traditional extractive summarizers: they produce fragmented text that lacks the natural flow and cohesion of human-written summaries (Giarelis et al., 2023).

To address these limitations, we propose the Divide, Prompt and Refine Biomedical Article Generation (DPR-BAG) framework. Drawing on prior work showing that divide-and-conquer decomposition reduces intermediate errors in LLMs (Zhang et al., 2025), DPR-BAG decomposes full-text articles along their rhetorical structure, performs parallel summarization on each resulting facet, and applies a modular refinement stage to reconcile fragmented outputs and restore discourse coherence. We target six rhetorical facets: Background, Objectives, Methods, Results, Conclusions (BOMRC), and Others. BOMRC is adopted as it represents the standard discourse structure validated in the PubMed 200k RCT dataset (Dernoncourt and Lee, 2017), while the “Others" facet retains any unclassified content. Using this design, we focus on three research questions:

1. Can we develop a training-free approach for the BAG task?
2. Does structure-aware decomposition of full-text articles improve the quality of generated abstracts compared to naive prompting?
3. To what extent does increasing prompting complexity (from detailed instructions to entity guidance) improve generation quality?

Our main contributions are as follows:

1. We propose DPR-BAG, a training-free, structure-aware method for BAG.
2. We release a dataset of more than 46K biomedical full-text publications for BAG task.
3. We compare DPR-BAG to strong extractive and abstractive baselines.
4. We systematically evaluate the effect of various prompting and splitting strategies as well as entity guidance within DPR-BAG.

2 Dataset

We constructed a BAG dataset based on PubMed publications from 1987 to 2023. To ensure a representative sample, we first calculated the publication type (PT) distribution of articles lacking abstracts using PT queries adapted from prior work (Menke et al., 2024). We then performed stratified sampling based on this distribution to retrieve 130,000 candidate XML files from the PubMed Central (PMC) Open Access subset, ensuring that the sampled articles reflect the publication type distribution of abstract-less PubMed records. For data processing, we adapted the extraction pipeline from the Longsummarization framework (Cohan et al., 2018) to parse structured sections and abstracts from the raw XML files. After filtering out records that were unparseable or lacked extractable abstracts, the final dataset, hereafter referred to as PMC-MAD (Missing-Abstract Distribution-aligned PMC), consists of 46,309 articles.

3 Methods

DPR-BAG follows a modular pipeline designed to generate structure-aware abstracts from biomedical full-text articles (Figure 1). The process begins by decomposing the document into five distinct facets based on BOMRC, plus an "Others" facet for unclassified content. For each facet, we perform parallel LLM-based summarization, which can optionally be augmented with entity guidance extension. The resulting sectional summaries are then concatenated and passed to a final LLM-based refinement stage to restore discourse coherence. DPR-BAG

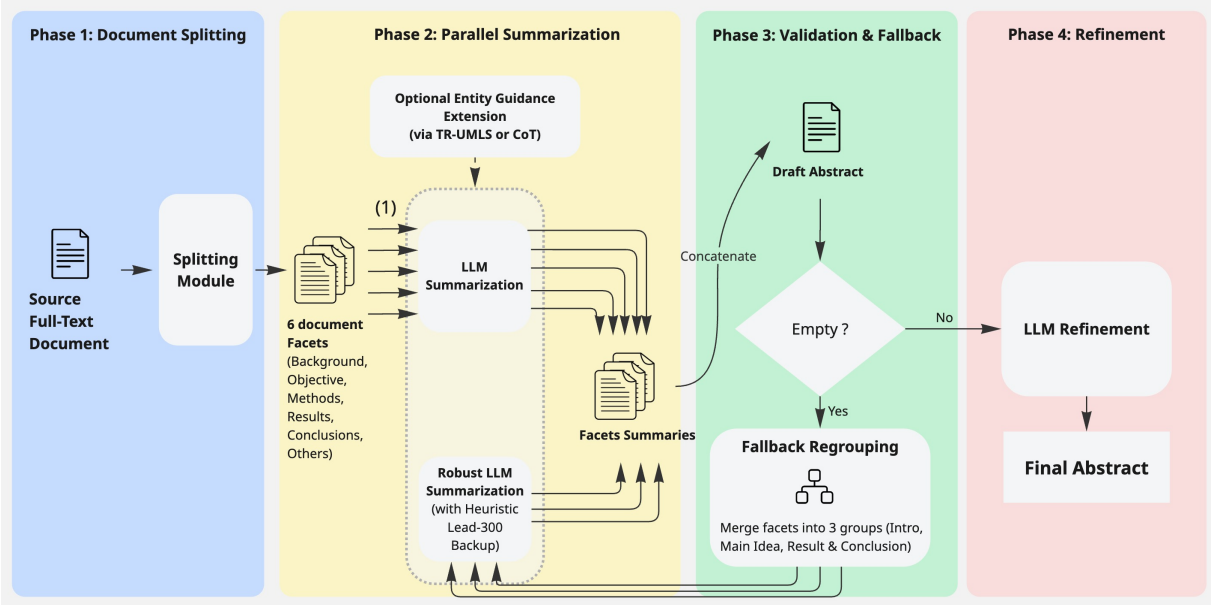

Figure 1: Overview of the DPR-BAG framework for biomedical abstract generation.

requires no task-specific training or fine-tuning; all components operate in a zero-shot manner using pre-trained, off-the-shelf models.

3.1 Task Formulation

Given a full-text biomedical article $D = (p_1, p_2, \ldots, p_n)$, where each p_i denotes a paragraph, the goal is to generate an abstract A that covers a predefined set of rhetorical facets $\mathcal{F} = \{\text{Background, Objectives, Methods, Results, Conclusions, Others}\}$ (BOMRC+).

We reformulate this as a facet-conditioned summarization problem, where the model generates content for each rhetorical facet separately. This decomposition allows the model to address each rhetorical component independently and capture discourse structure. Specifically, the document is partitioned into $K = 6$ facet-specific sub-documents $\{D_{f_k}\}_{k=1}^{K}$, where each D_{f_k} aggregates paragraphs rhetorically aligned with facet f_k. Each sub-document is independently summarized to produce a facet summary $\hat{a}_{f_k}$, and the concatenation $\hat{A} = \bigoplus_{k=1}^{K} \hat{a}_{f_k}$ is subsequently refined into the final abstract $R(\hat{A}) = A$. Facets absent from the source document yield empty strings.

3.2 Document Splitting

To segment full-text documents into rhetorically coherent texts, we use LLM-SSC (Lan et al., 2024), an LLM-based sequential sentence classification framework that assigns rhetorical labels (BOMRC) to sentences using in-context learning (performance details in Appendix E). While the model was trained on structured abstracts, we assume that the underlying rhetorical intent (e.g., methodological description vs. result reporting) remains consistent within full-text paragraphs. Specifically, we leverage the role of the first sentence of the paragraph as a topic sentence that typically encapsulates the paragraph's functional purpose. We assign the label of the first sentence of each paragraph as the global label for that paragraph, subsequently concatenating all paragraphs with matching labels to form the input document facets. We refer to this approach as the **First Sentence Labeling (FS)** strategy, and empirically validate it against naive splitting (NS) and section-header (SH) ablation variants below.

Naive Splitting Approach (NS): This approach distributes paragraphs into six segments, aiming for an approximately even distribution while maintaining paragraph integrity. Including this baseline allows us to assess whether a semantic-aware division (e.g., LLM-SSC) offers advantages over a purely structural, length-based partition.

Section-header Normalization (SH): This strategy serves as a coarse-grained semantic baseline. Utilizing the Transformer model developed in Lin et al. (2025), this approach categorizes existing section headers into standard BOMRC categories and concatenates paragraphs within the same facet. This comparison helps determine if the fine-grained, sentence-level classification used in LLM-SSC provides additional utility beyond

simple section-level organization.

3.3 Parallel Summarization

After the documents are divided into six document facets (BOMRC and Others), each is input into the LLM to generate a corresponding facet summary; facets that are not present in the source document are represented as empty summaries. These summaries are subsequently concatenated to form the draft abstract. The following subsections detail the prompting strategies and optional entity guidance extensions used during summarization.

3.3.1 Prompting Strategies

To investigate the effect of prompt complexity on generation quality, we adopt a **Basic Concise (BC)** prompting strategy as the baseline, and ablate prompting complexity by evaluating two more elaborate variants, **Detailed Instruction (DI)** and **Structural Instruction (SI)**. Full prompt templates are described in Appendix A.

Basic Concise Prompting (BC): BC is a minimal prompting strategy with coarse-grained focal points for each rhetorical facet (e.g., directing the model to "prioritize key findings and data" for the Results section) without further elaboration or explicit formatting structure.

Detailed Instruction Prompt (DI): DI is a more detailed prompting strategy modeled after the abstract submission guidelines of JMIR Publications[1] whose five-part BOMRC structured guideline aligns with the target rhetorical categories that DPR-BAG uses. By shifting the LLM persona to a "biomedical synthesis assistant," this prompt aims to enforce the extraction of granular details and mandates the inclusion of specific research designs, sample sizes, response rates, and statistical metrics (such as p-values and confidence intervals) to ensure adherence to multifaceted reporting standards.

Structural Instruction Prompt (SI): SI extends the basic prompt by introducing an explicit structural schema using Markdown formatting strategy, inspired by He et al. (2024). Compared to DI, which focuses on detailed content guidance, SI organizes the prompt into a more structured format, which aims to improve instruction adherence.

[1] https://support.jmir.org/hc/en-us/articles/37982552280987-Submitting-Your-Manuscript-to-JMIR-Publications-A-Guide-for-Authors

3.3.2 Entity Guidance

We additionally introduce an optional knowledge-grounding extension to further enhance semantic fidelity during parallel summarization. This component extracts key biomedical entities to guide the LLM in summarizing each facet. We consider two instantiations of this extension, TR-UMLS and CoT, detailed below.

TextRank and UMLS normalization (TR-UMLS): We extract key phrases from each facet using TextRank, a graph-based unsupervised method that requires no additional training and is well-suited to our zero-shot setting, and link them to UMLS (Bodenreider, 2004) concepts using scispaCy's UMLS entity linker (Neumann et al., 2019), retaining only phrases with valid UMLS mappings. When multiple phrases map to the same UMLS concept, they are grouped together and represented by a single term to avoid redundant anchoring. The top-n entities, ranked by TextRank centrality, are incorporated into the summarization prompt as anchor terms to guide the model toward the most structurally significant medical information in the text. We ablate $n \in \{5, 10\}$ in Section 5.2.

Chain-of-Thought (CoT): As an alternative, we employ a two-stage prompting strategy within the LLM summarization module when the extension is activated. In the first stage, we prompt the model to list the important entities of the facet, and in the second stage, the model is prompted to synthesize the information based on the facet and the entities it listed in the first stage.

We ablate these two entity guidance strategies independently. TR-UMLS pairs with DI, while CoT pairs with SI, whose structured format makes it well-suited for CoT's two-stage reasoning.

3.4 Validation and Fallback

To ensure the summarization pipeline robustness, we implement a validation and fallback mechanism. If all six facet summaries are empty due to insufficient context or LLM output format violations, the facets are regrouped into three broader categories (Intro, Main Idea, and Results & Conclusions) and sent back to the parallel summarization module. If regrouping fails, a first 300 characters (Lead-300) heuristic backup is used to guarantee non-empty output (details in Appendix B).

3.5 Refinement

Upon the formation of the draft abstract, we prompt the LLM to perform a global refinement to ensure structural coherence and stylistic consistency. This final processing step synthesizes the concatenated facets into a unified Final Abstract (the refinement prompt is detailed in Appendix C).

4 Experimental Setup

DPR-BAG is implemented with `Llama-3.2:3B` deployed via Ollama in instruction-tuned mode. Hardware details, token-limit constraints, and fine-tuning hyperparameters are provided in Appendix D.

4.1 Baseline Models

To establish robust baselines for our framework, we compared our approach against several standard long-document summarization models. We utilized two off-the-shelf variants of the Longformer Encoder-Decoder (LED) architecture (Beltagy et al., 2020)—pretrained on arXiv[2] and PubMed[3], respectively—to evaluate their zero-shot transferability to our task. Additionally, we included an off-the-shelf LongT5 model (Guo et al., 2022) pretrained on PubMed[4] to broaden our baseline comparisons across different architectures. Finally, to ensure maximal adaptation to our corpus, we evaluated a supervised fine-tuned version of the PubMed-pretrained LED and LongT5 using our 80% training split, using the 10% validation split for early stopping. All evaluations were performed on the remaining 10% (test split).

4.2 Evaluation Metrics

To assess the generated abstracts, we employ a multi-dimensional suite of metrics, with particular emphasis on abstractiveness and factuality. Detailed formulations and implementation details of each metric are provided in Appendix I. Paired bootstrap significance tests for the main comparisons are provided in Appendix J.

Abstractiveness: **Bigram and trigram novelty** measure the proportion of tokens absent from the source text, serving as a proxy for abstractive synthesis. We additionally adopt **Density** from Newsroom (Grusky et al., 2018), which quantifies the length of verbatim copying from the source, offering a complementary view of the model's extractive behavior. For both, smaller absolute deviations from the human-written reference indicate closer stylistic alignment.

Factuality: Given that our target abstracts are expected to be highly abstractive, factuality metrics must remain reliable under heavy paraphrasing. We therefore adopt **AlignScore** (Zha et al., 2023) as our primary factuality measure, as it evaluates alignment against the source full-text across a broad range of dimensions (notably paraphrasing), making it more robust than purely entailment-based alternatives. We additionally report **SummaC** (Laban et al., 2022) and **MiniCheck** (Tang et al., 2024), both NLI-based metrics, as cross-checks.

Semantic Alignment: We adopt **BERTScore** (Zhang et al., 2020) to evaluate semantic similarity between generated abstracts and the reference abstracts via contextual embeddings. To further assess discourse coherence, we adopt **DiscoScore** (Zhao et al., 2023), reporting DS_SENT_NN (sentence-level structural alignment) and DS_FOCUS_NN (shared noun semantic alignment).

Supporting Metrics: We additionally adopt **ROUGE-L** (Lin, 2004) to measure n-gram overlap with the reference. **UMLS Recall** quantifies the proportion of UMLS (Bodenreider, 2004) concepts from the original abstract retained in the generated output. **Coverage** and **Compression** from Newsroom (Grusky et al., 2018) serve as complementary metrics to measure how source content is preserved and condensed. For these two metrics, as well, smaller absolute deviations from the human-written reference are preferred.

5 Results

5.1 RQ1: Effectiveness of the Training-Free Approach

We first evaluate whether the proposed training-free approach can achieve competitive performance on the BAG task. As shown in Table 1 and Table 2, DPR-BAG (BC prompt, no entity guidance extension) achieves competitive performance against fine-tuned baselines. It generates abstracts that

[2] https://huggingface.co/allenai/led-large-16384-arxiv

[3] https://huggingface.co/patrickvonplaten/led-large-16384-pubmed

[4] https://huggingface.co/Stancld/longt5-tglobal-large-16384-pubmed-3k_steps

Model	Abstractiveness			Factuality		
	Bi-g	Tri-g	Dens.	AS	MC	SummaC
Original abstract	0.495	0.663	6.650	–	–	–
LED-Arxiv (Base)	0.140 (-0.355)	0.222 (-0.441)	22.698 (+16.048)	0.720	0.856	**0.735**
LED-Pubmed (Base)	0.167 (-0.328)	0.261 (-0.402)	19.233 (+12.583)	0.667	0.845	0.693
LED-Pubmed (FT)	0.309 (-0.187)	0.453 (-0.210)	11.369 (+4.719)	0.511	0.653	0.503
LongT5 (Base)	0.174 (-0.321)	0.280 (-0.383)	15.492 (+8.842)	0.668	0.875	0.711
LongT5 (FT)	0.252 (-0.243)	0.379 (-0.284)	12.989 (+6.339)	0.540	0.725	0.565
DPR-BAG	**0.397 (-0.098)**	**0.605 (-0.058)**	**4.690 (-1.960)**	**0.762**	**0.890**	0.570

Table 1: Performance comparison on primary evaluation dimensions. For Abstractiveness metrics, parenthetical values indicate the difference from the original abstract; best = smallest absolute deviation. Factuality scores are computed against the source full-text; higher is better. Best results bolded. (Bi-g: Bigram novelty; Tri-g: Trigram novelty; Dens.: Density; AS: AlignScore; MC: MiniCheck)

Model	Semantic Alignment			Supporting			
	BS	SENT	FOCUS↓	R-L	U-R	Cov.	Comp.
Original abstract	–	–	–	–	–	0.743	23.827
LED-Arxiv (Base)	0.635	0.907	0.702	0.234	0.340	0.950 (+0.217)	26.916 (+3.089)
LED-Pubmed (Base)	0.652	**0.908**	**0.670**	0.262	0.354	0.944 (+0.201)	27.387 (+3.560)
LED-Pubmed (FT)	0.634	0.868	0.791	0.264	**0.364**	**0.868 (+0.125)**	**20.901 (-2.926)**
LongT5 (Base)	**0.664**	0.876	0.861	0.274	0.312	0.948 (+0.205)	46.449 (+22.622)
LongT5 (FT)	0.650	0.895	0.713	**0.277**	0.361	0.911 (+0.168)	27.002 (+3.175)
DPR-BAG	0.617	0.868	0.828	0.183	0.266	0.894 (+0.151)	50.037 (+26.210)

Table 2: Performance comparison on semantic alignment and supporting metrics. For Coverage and Compression, parenthetical values indicate the difference from the original abstract; best = smallest absolute deviation. BS, SENT, R-L, and U-R: higher is better. FOCUS: lower is better. Best results bolded. (BS: BERTScore F1; SENT: DS_SENT_NN; FOCUS: DS_FOCUS_NN; R-L: ROUGE-L; U-R: UMLS Recall; Cov.: Coverage; Comp.: Compression)

more closely resemble human-written abstracts (more abstractive), while maintaining factual consistency with the full text (Table 1).

DPR-BAG outperforms the fine-tuned baselines in both abstractiveness and factuality (all paired bootstrap $p < 0.001$ on Bigram and Trigram novelty, Density, AlignScore, and MiniCheck; Appendix J.2), while the fine-tuned baselines themselves perform better than other baselines in abstractiveness but their generations are less factual.

DPR-BAG yields mostly lower scores for semantic alignment and other supporting metrics compared to the baselines, reflecting a known bias in these metrics toward extractive outputs (Table 2). Baseline models exhibit high density and low novelty relative to human-written abstracts (Table 1), indicating extractive behavior that likely inflates their scores. At the same time, DPR-BAG yields the highest compression rate (50.037), potentially leading to over-simplification of key information. We provide qualitative examples of generated abstracts in Appendix K.

5.2 RQ2: Impact of Structure-Aware Decomposition

We next evaluate whether structure-aware decomposition improves generation quality compared to naive prompting. As shown in Table 3, FS achieves the best overall performance. Compared to NS, FS achieves significantly higher AlignScore (paired diff = +0.006, $p < 0.05$), with no significant difference on MiniCheck or SummaC. Both approaches have similar abstractiveness, with NS achieving scores marginally closer to the human-written abstracts on Trigram novelty. Comparing FS with SH further underscores the necessity of fine-grained local context: SH significantly degrades AlignScore (-0.126), MiniCheck (-0.099), and SummaC (-0.141), all $p < 0.001$, indicating that broad semantic boundaries provided by section headers fail to provide sufficient contextual anchoring for faithful generation.

5.3 RQ3: Impact of Prompting Strategy and Entity Guidance

Next, we examine the effect of increasing prompt complexity.

Prompting Strategy: While DI and SI yield marginal but significant improvements in BERTScore ($p < 0.001$), both suffer from degradation in AlignScore, MiniCheck, SummaC, and UMLS Recall compared to BC (all $p < 0.001$; Table 4). This suggests that dense, multi-faceted

Splitting	Abstractiveness		Factuality		Semantic Alignment		
	Tri-g	Dens.	AS	MC	BS	FOCUS↓	U-R
FS	0.605 (-0.058)	**4.690 (-1.960)**	**0.762**	0.890	0.617	**0.828**	**0.266**
NS	**0.613 (-0.050)**	4.508 (-2.142)	0.756	**0.892**	0.615	0.856	0.262
SH	0.746 (+0.083)	2.885 (-3.765)	0.636	0.791	**0.636**	0.920	0.229

Table 3: Ablation on document splitting strategies. Splitting strategies are Naive Splitting (NS), Section Header Normalization (SH), and First Sentence Labeling (FS). (Abbreviations as in Table 1 and Table 2)

Prompt	Abstractiveness		Factuality		Semantic Alignment		
	Tri-g	Dens.	AS	MC	BS	FOCUS↓	U-R
BC	**0.605 (-0.058)**	**4.690 (-1.960)**	**0.762**	**0.890**	0.617	**0.828**	**0.266**
DI	0.725 (+0.063)	3.131 (-3.519)	0.642	0.806	**0.638**	0.885	0.247
SI	0.736 (+0.073)	3.049 (-3.601)	0.630	0.805	0.636	0.836	0.251

Table 4: Ablation on summarization prompt strategies. Prompt variants are Basic Concise (BC), Detailed Instruction (DI), and Structural Instruction (SI). (Abbreviations as in Table 1 and Table 2)

instructions might have introduced a distraction effect, where the model struggles to simultaneously satisfy formatting constraints and maintain source-grounded factual alignment, motivating the need for explicit reasoning pathways and external grounding.

Entity Guidance: To assess whether external grounding can recover the factual consistency degradation observed in DI and SI, we ablate two entity guidance approaches: TR-UMLS integrated into DI, and CoT integrated into SI, with BC (no entity guidance) serving as the reference baseline.

As shown in Table 5, applying TR-UMLS to DI does not improve the overall performance. Top-5 and top-10 variants show no significant change in AlignScore relative to the base DI prompt, and both remain significantly below the BC baseline ($p < 0.001$). DS-Focus score also increases under top-5 entities ($p < 0.01$), suggesting that explicit entity conditioning might cause the model to overprioritize the provided terms, exacerbating the distraction effect rather than serving as an effective grounding mechanism.

Table 6 shows that integrating CoT into the SI prompt improves semantic alignment with the original abstract compared to BC (BERTScore +0.015, $p < 0.001$) while yielding overall higher abstractiveness and lower factuality (AlignScore -0.166, MiniCheck -0.098, SummaC -0.158, all $p < 0.001$). This shows that CoT encourages more abstractive and novel phrasing, at the expense of reducing the factual overlap with the source.

6 Discussion

Our results show that the proposed training-free approach generates abstracts that more closely resemble human-written abstracts in terms of abstractiveness and factuality. Notably, models fine-tuned on PMC-MAD do not match this performance on these dimensions. However, these fine-tuned models achieve slightly better semantic alignment, lexical/semantic overlap, and compression than DPR-BAG. Since the baseline models do not explicitly model rhetorical structure, we attribute the stronger performance of DPR-BAG on abstractiveness and factuality to its structure-aware design. Specifically, the instruction-tuned LLM backbone provides a baseline tendency toward natural, paraphrased output over verbatim copying. This tendency is amplified by the decompose-then-refine design: partitioning full-text articles into facet-specific subdocuments allows each summarization step to operate over shorter, topically coherent contexts, reducing verbatim copying under long-context pressure. The subsequent refinement stage then resynthesizes the concatenated facet summaries, further encouraging paraphrasing over fragment assembly.

The distraction effect observed under DI and SI prompts indicates that instruction complexity can actively harm factual grounding in small LLMs. Similarly, entity guidance via UMLS only seemed to increase instruction complexity, yielding little positive effect. The inconsistent gains of SI+CoT further suggest that publication-type distribution can influence generation behavior. This motivates publication-type-aware prompting as a future direction. Additional BC-prompt ablations corroborate these findings for TR-UMLS (Appendix H).

To evaluate the generalizability of DPR-BAG, we applied it to the PubMed Summarization dataset (Cohan et al., 2018). Unlike PMC-MAD, this dataset is not stratified to reflect the publica-

Config	Abstractiveness		Factuality		Semantic Alignment		
	Tri-g	Dens.	AS	MC	BS	FOCUS↓	U-R
BC	0.605 (-0.058)	**4.690 (-1.960)**	**0.762**	**0.890**	0.617	**0.828**	**0.266**
DI	0.725 (+0.063)	3.131 (-3.519)	0.642	0.806	**0.638**	0.885	0.247
DI+Top-5	0.725 (+0.062)	3.138 (-3.512)	0.644	0.807	0.637	0.933	0.245
DI+Top-10	**0.677 (+0.014)**	2.938 (-3.712)	0.644	0.811	0.595	0.915	0.231

Table 5: Ablation on UMLS entity guidance variants. Top-n denotes the number of top-ranked UMLS entities injected into the DI prompt. (Abbreviations as in Table 1 and Table 2)

Config	Abstractiveness		Factuality		Semantic Alignment		
	Tri-g	Dens.	AS	MC	BS	FOCUS↓	U-R
BC	**0.605 (-0.058)**	**4.690 (-1.960)**	**0.762**	**0.890**	0.617	0.828	**0.266**
SI	0.736 (+0.073)	3.049 (-3.601)	0.630	0.805	**0.636**	0.836	0.251
SI + CoT	0.749 (+0.086)	2.940 (-3.710)	0.596	0.792	0.631	**0.794**	0.249

Table 6: Ablation on Chain-of-Thought guidance for the SI prompt. (Abbreviations as in Table 1 and Table 2)

tion type distribution of abstract-less PubMed articles. Overall, the trends observed on PMC-MAD, particularly with respect to semantic alignment, abstractiveness, and factuality, largely persist on this dataset. Some differences are also observed (lower MiniCheck scores and better performance of SI+CoT relative to BC). Despite these variations, the results indicate that DPR-BAG maintains robust performance across datasets, supporting its generalizability (Appendix F).

To investigate whether larger model sizes can improve generation quality, we also evaluated `Qwen2.5:7B` and `Qwen2.5:14B` under the BC prompt. These models achieve improvements in semantic alignment and compression rates; however, they also yield excessive abstractiveness and lower factuality, overall yielding little advantage over the base 3B model (Appendix G).

6.1 Limitations

Several limitations apply. First, our evaluation relies primarily on automated metrics, which vary in their robustness. Human evaluation is needed for complementary validation. Second, the BOMRC schema may be suboptimal for articles with non-standard discourse structure. More adaptive decomposition strategies remain an open direction. Finally, DPR-BAG's summaries exhibit substantially higher compression than human-written abstracts, indicating considerably terser outputs. This excessive compression risks omitting auxiliary but informative content such as background, secondary findings, or caveats. Calibrating facet-level length targets is a natural direction for future work.

7 Related Work

For the BAG task, Chachra et al. (2016) integrated domain-specific classifiers and entailment graphs for extractive sentence selection, inheriting the disjointed flow typical of pure extractive approaches. Sybrandt and Safro (2021) proposed CBAG, which generates abstracts conditioned on author-provided MeSH keywords rather than full-text articles, leaving long-context challenges unaddressed.

To generate abstractive summaries from full-text articles, DANCER (Gidiotis and Tsoumakas, 2020) and IDCUOT (Shen and Lam, 2022) pioneered the strategy of breaking long documents into manageable sections to bypass context window limitations. However, these supervised approaches rely on heuristic alignment algorithms to map abstract sentences back to source sections and generate summaries for each section in isolation, making them prone to error propagation and fragmented coherence across sections.

GenCompareSum (Bishop et al., 2022) leverages similar divide-and conquer principles but remains extractive-heavy: abstractive fragments serve only as anchors for sentence selection, resulting in discontinuous summaries that lack the narrative transitions typical of human-written abstracts.

8 Conclusion

We presented DPR-BAG, a training-free, rhetorical structure-aware, divide-and-conquer framework for biomedical abstract generation. The method decomposes full-text articles into semantic facets and applies parallel LLM-based summarization followed by refinement. Across both the PMC-MAD and PubMed Summarization datasets, DPR-BAG produces abstracts that most closely match human-

written abstracts in terms of abstractiveness and factual consistency, without task-specific training and within standard hardware constraints. The framework can be integrated as a preprocessing component into pipelines that require biomedical abstracts, thereby improving downstream performance.

Acknowledgement

This work was supported by the National Library of Medicine of the National Institutes of Health under the award number R01LM14292. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funder had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Jennifer Bishop, Qianqian Xie, and Sophia Ananiadou. 2022. GenCompareSum: a hybrid unsupervised summarization method using salience. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 220–240, Dublin, Ireland. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270.

Suchet Chachra, Asma Ben Abacha, Sonya Shooshan, Laritza Rodriguez, and Dina Demner-Fushman. 2016. A hybrid approach to generation of missing abstracts in biomedical literature. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1093–1100, Osaka, Japan. The COLING 2016 Organizing Committee.

Daniil Chernyshev and Boris Dobrov. 2024. Investigating the pre-training bias in low-resource abstractive summarization. *IEEE Access*, 12:47219–47230.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Brandon Fan, Weiguo Fan, Carly Smith, and Harold "Skip" Garner. 2020. Adverse drug event detection and extraction from open data: A deep learning approach. *Information Processing & Management*, 57(1):102131.

Nikolaos Giarelis, Charalampos Mastrokostas, and Nikos Karacapilidis. 2023. Abstractive vs. extractive summarization: An experimental review. *Applied Sciences*, 13(13).

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontañón, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on LLM performance? *Preprint*, arXiv:2411.10541.

John PA Ioannidis and Michaéla C Schippers. 2025. In-house editorials and journalistic pieces comprise a massive corpus in the scientific literature that can be improved. *European Journal of Clinical Investigation*, 55(8):e70061.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Mengfei Lan, Lecheng Zheng, Shufan Ming, and Halil Kilicoglu. 2024. Multi-label sequential sentence classification via large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16086–16104, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sylvey Lin, Joseph Menke, Arthur Holt, Halil Kilicoglu, and Neil Smalheiser. 2025. Section header normalization in biomedical articles using transformers. In *AMIA Annual Symposium Proceedings*. Poster P116.

Man Luo, Arindam Mitra, Tejas Gokhale, and Chitta Baral. 2022. Improving biomedical information retrieval with neural retrievers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11038–11046.

Anne Magnet and Didier Carnet. 2006. Letters to the editor: Still vigorous after all these years?: A presentation of the discursive and linguistic features of the genre. *English for Specific Purposes*, 25(2):173–199.

Joe D Menke, Halil Kilicoglu, and Neil R Smalheiser. 2024. Publication type tagging using transformer models and multi-label classification. *AMIA Annual Symposium Proceedings*, 2024:818–827.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

James L. Nuzzo. 2021. Letters to the editor in exercise science and physical therapy journals: an examination of content and "authorship inflation". *Scientometrics*, 126(8):6917–6936.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Xin Shen and Wai Lam. 2022. Improved divide-and-conquer approach to abstractive summarization of scientific papers. In *2022 4th International Conference on Natural Language Processing (ICNLP)*, pages 395–398.

Justin Sybrandt and Ilya Safro. 2021. CBAG: Conditional biomedical abstract generation. *PLoS One*, 16(7):e0253905.

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

Alberto Ueda, Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2021. Structured fine-tuning of contextual embeddings for effective biomedical retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2031–2035, New York, NY, USA. Association for Computing Machinery.

Cathelijn J. F. Waaijer, Cornelis A. van Bochove, and Nees Jan van Eck. 2011. On the map: Nature and science editorials. *Scientometrics*, 86(1):99–112.

Tairan Wang, Xiuying Chen, Qingqing Zhu, Taicheng Guo, Shen Gao, Zhiyong Lu, Xin Gao, and Xiangliang Zhang. 2025. New paradigm for evaluating scholar summaries: A facet-aware metric and a meta-evaluation benchmark. *ACM Trans. Inf. Syst.*, 43(4).

Thomas C Wiegers, Allan Peter Davis, Jolene Wiegers, Daniela Sciaky, Fern Barkalow, Brent Wyatt, Melissa Strong, Roy McMorran, Sakib Abrar, and Carolyn J Mattingly. 2025. Integrating AI-powered text mining from PubTator into the manual curation workflow at the Comparative Toxicogenomics Database. *Database*, 2025:baaf013.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Yizhou Zhang, Defu Cao, Lun Du, Qiang Fu, and Yan Liu. 2025. When splitting makes stronger: A theoretical and empirical analysis of divide-and-conquer prompting in LLMs. In *Second Conference on Language Modeling*.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating text generation with BERT and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.

A Summarization Prompt Templates

All prompting strategies share a `facet_guidelines` dictionary that maps each facet label to a facet-specific instruction. We use two versions: a concise version (Table 7) used in BC and SI prompts, and a detailed version adapted from JMIR author guidelines (see footnote 1) (Table 8 used in DI prompts). In all prompt templates, `<facet_text>` denotes the input facet text, `<facet_type>` denotes the rhetorical facet label, and `<facet_guide>` denotes the corresponding facet-specific instruction.

A.1 Basic Concise Prompt (BC)

System Message

You are a biomedical summarization assistant. 1. Use a formal, objective, scientific tone. 2. Never use meta-phrases like 'the authors state' or 'this section describes'. Respond ONLY with a JSON object: {"summary": "...", "reasoning": "..."}. No markdown, no talk.

User Message

Summarize this `<facet_type>` section
Specific focus: `<facet_guide>`

Paragraph text:
`<facet_text>`

A.2 Detailed Instruction Prompt (DI)

System Message

You are a biomedical synthesis assistant. 1. Use a formal, objective, scientific tone.
2. Never use meta-phrases like 'the authors state' or 'this section describes'.
Define your output strictly as:
-'summary': The synthesized biomedical text.
-'reasoning': A brief explanation.
Format: {"summary": "...", "reasoning": "..."}. No markdown, no talk.

User Message

Synthesize the critical information from this `<facet_type>` section provided in the 'Paragraph text'
`<facet_guide>`

Paragraph text:
`<facet_text>`

A.3 Structural Instruction Prompt (SI)

System Message

ROLE
You are an expert Biomedical Summarization Assistant.

STRICT GUIDELINES
- **Tone**: Formal, objective, and academic.
- **No Meta-Talk**: Do NOT use phrases like 'The authors state' or 'This section describes'.
- **Output Format**: Respond **ONLY** with a valid JSON object. No Markdown blocks, no preamble, no postscript.
```json
{"reasoning": "Brief explanation of how your summary fulfills the given instructions...", "summary": "Final summary..."}
```

Facet	Guideline
Background	Focus on research gap and motivation.
Objective	State the primary aim or hypothesis.
Methods	Detail study design and procedures.
Results	Prioritize key findings and data.
Conclusions	Summarize implications and take-home messages.
Others	Focus on the main point of the paragraph.
Intro	Introduce the main topic and context.
Main Idea	Focus on the central concept or hypothesis.
Results & Conclusions	Focus on key findings and their implications.

Table 7: Concise facet-specific guidelines (used in BC and SI prompts).

Facet	Guideline
Background	Briefly describe the context and significance of the research.
Objective	State the specific aim(s) of the study in a complete sentence.
Methods	Outline the research design, study sample, data collection, and analysis procedures.
Results	Present key findings, including relevant statistics (sample sizes, response rates, P values, confidence intervals). Be specific.
Conclusions	Summarize the main findings and their implications.
Others	Synthesize the core biomedical information, focusing on the primary argument, concept, or supplementary context presented.
Intro	Describe the research context and significance, and clearly state the specific aims or hypotheses.
Main Idea	Outline the research design and procedures, while capturing any supplementary methodological context or core concepts.
Results & Conclusions	Present key findings with relevant statistics, and summarize their broader implications and take-home messages.

Table 8: Detailed facet-specific guidelines (used for DI prompts), adapted from JMIR author guidelines.

User Message

TASK: SUMMARY GENERATION
Target Focus: `<facet_guide>`

Instructions: Generate a professional biomedical summary.
—
INPUT TEXT (Reference)
`<facet_text>`

A.4 BC for Naive Splitting (BC-NS)

When paired with the Naive Splitting baseline, the system message remains unchanged. As no facet label is assigned, the user message uses a generic focus instruction:

User Message

Summarize this section.
Specific focus: Summarize the main topic.

Paragraph text:
`<facet_text>`

A.5 BC with TR-UMLS Entity Guidance (BC+TR-UMLS)

System Message

You are a biomedical summarization assistant. 1. Use a formal, objective, scientific tone. 2. Never use meta-phrases like 'the authors state' or 'this section describes'. Respond ONLY with a JSON object: {"summary": "...", "reasoning": "..."}. No markdown, no talk.

User Message

Summarize this `<facet_type>` section
Specific focus: `<facet_guide>`
Ensure the core meanings of these key biomedical entities are preserved or synthesized accurately:

Paragraph text:<top_entities>
`<facet_text>`

A.6 DI with TR-UMLS Entity Guidance (DI+TR-UMLS)

User Message

Synthesize the critical information from this <facet_type> section provided in the 'Paragraph text'
<facet_guide>

Ensure the core meanings of these key biomedical entities are preserved or synthesized accurately: <top_entities>

Paragraph text:
<facet_text>

A.7 SI with Chain-of-Thought (SI+CoT)

System Message

ROLE
You are an expert Biomedical Summarization Assistant.

—
OPERATIONAL FRAMEWORK
You must follow this 2-Stage Chain-of-Thought process:
1. Stage 1: Element Extraction (entities, parameters, methodologies, statistics).
2. Stage 2: Summary Generation (synthesize into scientific narrative).

—
STRICT GUIDELINES
- **Tone**: Formal, objective, and academic.
- **No Meta-Talk**: Do NOT use phrases like 'The authors state' or 'This section describes'.
- **Output Format**: Respond **ONLY** with a valid JSON object. No Markdown blocks, no preamble, no postscript.
```json
{"reasoning": "Stage 1 extraction results...", "summary": "Stage 2 final summary..."}
```

User Message 1 (Stage 1)

TASK 1: ELEMENT EXTRACTION
Section Type: <facet_type>
Instructions: Extract key elements from the text below, including:
- **Entities:** Diseases, genes, drugs, proteins.
- **Parameters:** Sample sizes, dosage, duration.
- **Methodology:** Study design, assays, equipment.
- **Statistics:** P-values, confidence intervals, effect sizes.

—
INPUT TEXT
<facet_text>

User Message 2 (Stage 2)

TASK 2: SUMMARY GENERATION
Target Focus: <facet_guide>

Instructions: Using the elements extracted in Task 1, generate a professional biomedical summary.
Ensure the summary is dense with information but remains readable and scientifically accurate.

—
INPUT TEXT (Reference)
<facet_text>

B Validation and Fallback Details

When the fallback process is triggered, the original six facets are regrouped into three broader categories: *Intro* (Background and Objective), *Main Idea* (Methods and Others), and *Results & Conclusions* (Results and Conclusions). Background and Objective are merged as both establish the research context and motivation. Results and Conclusions are paired as both convey findings and their implications. The remaining Others facet, which retains paragraphs not assigned to any BOMRC category by the sentence classifier, is grouped with Methods by elimination, as the other four facets form more natural rhetorical pairs. These regrouped facets are then sent back to the parallel summarization module. If the LLM still fails to summarize a specific

regrouped facet, the first 300 characters (Lead-300) of that facet are used as the facet summary to capture core information via lead bias without adding granular noise. The fallback mechanism was rarely triggered in practice, suggesting that this pairing has minimal impact on overall generation quality. In a random sample of 300 test articles (6.5% of the test set), the fallback mechanism was never triggered, suggesting with 95% confidence that fewer than 1% of articles require fallback intervention.

C Refinement Prompt

This prompt is used in the final stage to smooth the concatenated facets, ensuring structural coherence and stylistic consistency across the unified abstract. `<draft_abstract>` denotes the place holder for the concatenated draft abstract.

System Message

You are a biomedical abstract refinement assistant. Refine the abstract based on the abstract draft.
CRITICAL INSTRUCTION:
Respond ONLY with a valid JSON object.
Do NOT use Markdown code blocks (like “‘json).
Do NOT provide any conversational text.

Format:
{
"abstract": "your abstract text here",
"reasoning": "your reasoning here"
}

User Message

abstract draft: `<draft_abstract>`

D Extended Implementation Details and Token Distribution

All fine-tuning and evaluation procedures were conducted on an NVIDIA Tesla V100 GPU (32GB VRAM). While the underlying LED architecture theoretically supports sequences up to 16,384 tokens, processing such lengths on standard hardware is computationally prohibitive, inevitably leading to out-of-memory (OOM) errors even with minimal batch sizes. To fit within this memory budget, the fine-tuned LED-Pubmed used gradient accumulation with an effective batch size of 8, halted at

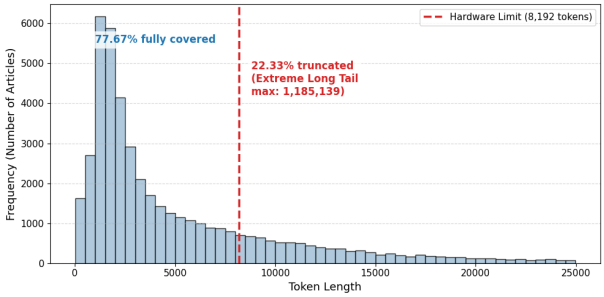

Figure 2: Distribution of document token lengths in the dataset. The red dashed line denotes the 8,192-token hardware limit for standard baselines.

500 steps based on validation performance.

This constraint restricts standard baselines to 8,192-token inputs. As illustrated in Figure 2, our empirical analysis of the 46,309 articles in the dataset reveals a median of 2,959 tokens and an average length of approximately 6,018 tokens. While the 8,192-token capacity successfully accommodates 77.67% of the dataset, the length distribution exhibits a severe long-tail characteristic. Specifically, 22.33% of the articles exceed this limit, with the longest document reaching an extreme 1,185,139 tokens. For these extensive studies, standard baselines operating within memory limits are forced to truncate critical information, such as discussion and conclusion sections, which underscores the necessity of the partitioned approach introduced in our DPR-BAG framework.

E LLM-SSC

LLM-SSC (Lan et al., 2024) is evaluated on the BIORC800 dataset, a manually annotated multi-label SSC dataset of biomedical abstracts using the BOMRC schema. Under task-specific fine-tuning, LLM-SSC achieves a micro F1 of 0.907 and macro F1 of 0.912 on BIORC800, outperforming prior SSC baselines (Cohan et al., 2019). We adopt LLM-SSC for our document splitting module as its label schema (Background, Objective, Methods, Results, Conclusions, and None) directly aligns with the BOMRC+ facets in DPR-BAG, where the None label corresponds to our Others facet.

F PubMedSum Dataset Validation

To evaluate the generalizability of our pipeline, we also tested DPR-BAG on the PubMed Summarization dataset (Cohan et al., 2018), hereafter PubMedSum. As shown in Figure 4, unlike PMC-MAD, which is stratified to reflect the publication type dis-

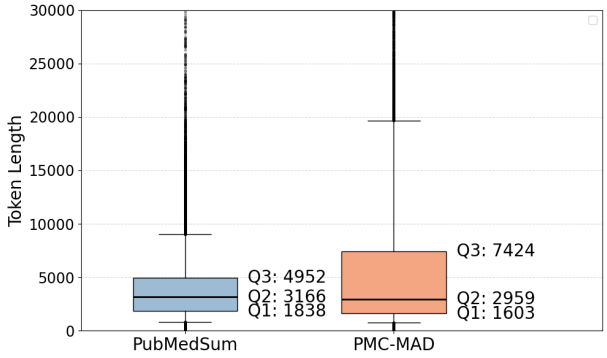

Figure 3: Token Distribution Comparison

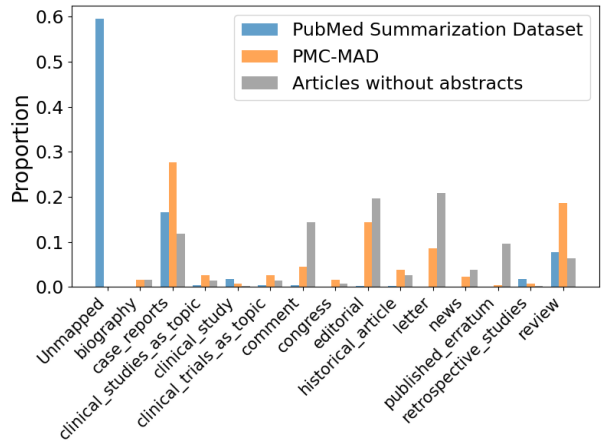

Figure 4: Publication type distribution of PMC-MAD, PubMed Summarization Dataset, and PubMed articles without abstracts.

tribution of abstract-less PubMed manuscripts, the majority of articles in the PubMedSum could not be mapped to a specific publication type. This allows us to benchmark the performance of DPR-BAG against established standards in the biomedical domain and ensure that our findings are not limited to the specific characteristics of the PMC-MAD corpus.

As illustrated in Figure 3, the two datasets exhibit distinct token length profiles. PMC-MAD has a median document length of 2959 tokens (IQR: 1603–7424), while PubMedSum has a median of 3,166 tokens (IQR: 1838–4952).

Table 9 and Table 10 present results on the PubMed Summarization dataset for the two best PMC-MAD configurations: BC (without entity guidance) and SI+CoT. Consistent with PMC-MAD findings, DPR-BAG underperforms all baseline models on ROUGE, UMLS Recall, DiscoScore, and SummaC, while achieving better coverage, density, and novel n-gram scores closer to the original human-written abstract. Both configurations outperform all baselines on AlignScore, confirming that the abstractiveness gains do not compromise factuality.

However, DPR-BAG's MiniCheck advantage on PMC-MAD does not transfer here, attributable to PubMedSum's inherent higher abstractiveness (lower density, higher novelty), which penalizes NLI-based metrics (MiniCheck) even when factual content is preserved — as evidenced by DPR-BAG's consistently superior AlignScore across both datasets.

Unlike on PMC-MAD where SI+CoT only improved BERTScore, on PubMedSum SI+CoT significantly improves AlignScore, ROUGE-L, UMLS Recall, DiscoScore, and Compression while showing no significant degradation on factuality (MiniCheck, SummaC) or abstractiveness. This suggests that CoT's benefit varies with document characteristics, potentially driven by publication-type distribution differences.

G Effect of Backbone Model Size

To investigate whether scaling up the backbone LLM improves generation quality, we evaluated `Qwen2.5:7B` and `Qwen2.5:14B` under the BC prompt without entity guidance. As the Llama3.2 series does not offer a model beyond 3B, we use Qwen 2.5 for the scaling analysis.

As shown in Table 11 and 12, Qwen2.5:14B improves BERTScore, UMLS Recall, and DS-Focus relative to the 3B backbone, recording the best DS-Focus, and UMLS Recall among all DPR configurations, alongside a compression rate most closely aligned with the original abstracts. However, both larger backbones exhibit excessively higher abstractiveness (lower Density, higher Novel n-grams) and declining factuality (SummaC, AlignScore and MiniCheck) relative to the 3B backbone, suggesting that larger models promote more abstractive phrasing and precise semantic alignment of key noun foci at the cost of reduced factual overlap with the source.

H BC Entity Guidance Ablations

To isolate the effect of entity guidance from prompt complexity, we additionally apply TR-UMLS to the BC prompt and compare against BC without entity guidance.

We integrate TR-UMLS with top-5 UMLS entities into the BC prompt (BC+Top-5). As shown in Table 13, BC+Top-5 significantly degrades fac-

Model	Abstractiveness			Factuality		
	Bi-g	Tri-g	Dens.	AS	MC	SummaC
Original abstract	0.478	0.680	4.917	–	–	–
LED-Arxiv (base)	0.121 (-0.358)	0.205 (-0.475)	25.146 (+20.228)	0.755	**0.872**	**0.712**
LED-Pubmed (base)	0.156 (-0.322)	0.260 (-0.420)	19.565 (+14.648)	0.640	0.807	0.661
LongT5 (base)	0.157 (-0.322)	0.271 (-0.409)	15.827 (+10.910)	0.605	0.826	0.645
DPR-BAG (BC)	0.507 (+0.029)	0.733 (+0.053)	2.966 (-1.951)	0.765	0.794	0.414
DPR-BAG (SI+CoT)	**0.503 (+0.024)**	**0.723 (+0.043)**	**3.153 (-1.765)**	**0.772**	0.787	0.411

Table 9: Performance comparison on the PubMed Summarization dataset across primary evaluation dimensions. Conventions and abbreviations as in Table 1

Model	Semantic Alignment			Supporting			
	BS	SENT	FOCUS↓	R-L	U-R	Cov.	Comp.
Original abstract	–	–	–	–	–	0.878	16.084
LED-Arxiv (base)	0.641	0.913	1.398	0.236	0.348	0.968 (+0.091)	**17.739 (+1.655)**
LED-Pubmed (base)	0.664	**0.929**	**1.100**	**0.272**	**0.415**	0.960 (+0.082)	14.235 (-1.849)
LongT5 (base)	**0.664**	0.894	1.585	0.265	0.345	0.962 (+0.085)	25.643 (+9.560)
DPR-BAG (BC)	0.651	0.814	2.116	0.192	0.217	0.876 (-0.002)	43.962 (+27.878)
DPR-BAG (SI+CoT)	0.652	0.839	1.927	0.197	0.245	**0.878 (+0.000)**	40.068 (+23.984)

Table 10: Performance comparison on the PubMed Summarization dataset across semantic alignment and supporting metrics. Conventions and abbreviations as in Table 2

tuality relative to BC ($p < 0.001$) and decreases UMLS Recall ($p < 0.001$), indicating that the injected UMLS terms fail to anchor concept reproduction in the output. These results confirm that TR-UMLS entity guidance remains ineffective in this zero-shot setting regardless of the underlying prompt strategy. Significance details are reported in Appendix J.5.

I Evaluation Metric Details

I.1 Factual Consistency (AlignScore, MiniCheck and SummaC)

AlignScore (Zha et al., 2023): AlignScore trains a unified alignment model on 4.7M examples spanning 7 tasks—NLI, fact verification, paraphrase, semantic textual similarity, question answering, information retrieval, and summarization—producing a single factual consistency score. At inference time, the source document is split into overlapping chunks of approximately 350 tokens, and each sentence in the generated abstract is evaluated against all chunks; the highest alignment score per sentence is averaged to yield the final score. We use the base variant (RoBERTa-base, 125M parameters).

MiniCheck (Tang et al., 2024): MiniCheck trains a small fact-checking model on synthetically constructed data generated by GPT-4, where each instance is designed to require verifying multiple atomic facts against multi-sentence evidence. It produces a binary supported/unsupported prediction per sentence. In our evaluation, each sentence in the generated abstract is treated as an individual claim and verified against the source full-text document. We use the Flan-T5-Large variant (770M parameters).

SummaC (Laban et al., 2022): SummaC segments the source document into individual sentences and scores each generated sentence by aggregating sentence-level NLI entailment probabilities against all source sentences. The $\text{SummaC}_{\text{Conv}}$ variant learns a convolutional layer over the full distribution of these entailment scores, rather than relying only on the maximum, making it more robust to outliers. We use the `SummaCConv` implementation with a VitaminC-trained NLI backbone and sentence-level granularity.

I.2 DiscoScore (DS_SENT_NN & DS_FOCUS_NN)

DiscoScore(Zhao et al., 2023) evaluates discourse coherence by modeling focus transitions across sentences. In this work, we use nouns (NN) as the focus, one of several focus choices supported by the metric, to compare the discourse coherence between the reference and generated abstracts.

DS_SENT_NN: Constructs a sentence graph where edges are drawn between any two sentences (not just adjacent ones) that share at least one noun focus. Edge weights are inversely proportional to the distance between the two sentences ($1/(j-i)$), capturing both local and long-range co-

Backbone	Abstractiveness			Factuality		
	Bi-g	Tri-g	Dens.	AS	MC	SummaC
Original abstract	0.495	0.663	6.650	–	–	–
Llama 3.2:3B	0.397 (-0.098)	**0.605 (-0.058)**	**4.690 (-1.960)**	**0.762**	**0.890**	**0.570**
Qwen 2.5:7B	0.594 (+0.099)	0.806 (+0.143)	2.380 (-4.270)	0.635	0.793	0.432
Qwen 2.5:14B	**0.568 (+0.073)**	0.778 (+0.115)	2.651 (-3.999)	0.647	0.812	0.443

Table 11: Scale-up comparison on primary evaluation dimensions. All configurations use the BC prompt without entity guidance. Conventions and abbreviations as in Table 1

Backbone	Semantic Alignment			Supporting			
	BS	SENT	FOCUS↓	R-L	U-R	Cov.	Comp.
Original abstract	–	–	–	–	–	0.743	23.827
Llama 3.2:3B	0.617	0.868	0.828	0.183	0.266	0.894 (+0.151)	50.037 (+26.210)
Qwen 2.5:7B	0.631	**0.876**	0.670	0.177	0.279	**0.846 (+0.103)**	32.273 (+8.446)
Qwen 2.5:14B	**0.633**	0.866	**0.596**	**0.189**	**0.337**	0.857 (+0.114)	**23.234 (-0.593)**

Table 12: Scale-up comparison on semantic alignment and supporting metrics. Conventions and abbreviations as in Table 2

herence. Sentence embeddings are then aggregated according to this graph structure, and the cosine similarity between the resulting graph-level embeddings of the generated and reference texts is used as the final score.

DS_FOCUS_NN: Measures how closely the frequency and semantics of shared noun foci match between the generated and reference texts. For each focus shared by both texts, it computes the distance between their embeddings (derived by summing the contextualized token embeddings of all associated tokens), and averages these distances as the final score.

I.3 UMLS Recall

This metric measures how well the generated abstract preserves biomedical concepts present in the reference abstract. We extract unique UMLS concepts from both the original abstract (C_{ref}) and the generated abstract (C_{gen}) using a biomedical entity linker (scispaCy, Neumann et al., 2019) with the UMLS knowledge base). UMLS Recall is then computed as:

$$\text{UMLS Recall} = \frac{|C_{\text{gen}} \cap C_{\text{ref}}|}{|C_{\text{ref}}|} \quad (1)$$

Newsroom (Grusky et al., 2018): Coverage, Density, and Compression: We adopt the extractive fragment measures from Grusky et al. (2018). Given a summary S and source document D, extractive fragments are the set of longest common substrings shared between S and D. **Coverage** measures the proportion of summary tokens belonging to an extractive fragment. **Density** measures the average length of the extractive fragments, reflecting how verbatim the summary is. **Compression** is the ratio of source document length to summary length.

N-gram Novelty Following See et al. (2017), we compute the proportion of bigrams and trigrams in the generated abstract that do not appear in the source full-text document. Higher novelty indicates greater abstractive synthesis relative to the source.

J Statistical Significance Analysis

J.1 Method

For each comparison between two configurations A and B, we align per-document scores by article ID and resample with replacement over 10,000 iterations. For metrics where higher (or lower) scores indicate better performance (AlignScore, MiniCheck, SummaC, BERTScore, DiscoScore variants, ROUGE-L, UMLS Recall), we test $H_0 : \mathbb{E}[s_A] = \mathbb{E}[s_B]$ on raw scores. For abstractiveness metrics where the goal is proximity to the reference distribution (Bigram novelty, Trigram novelty, Density, Coverage, Compression), we instead test $H_0 : \mathbb{E}[\,|s_A - s_{\text{ref}}|\,] = \mathbb{E}[\,|s_B - s_{\text{ref}}|\,]$, where s_{ref} is the corresponding score of the human-written abstract on the same document. Under this formulation, a negative difference indicates that A is closer to the reference than B. All tests are two-sided, with significance markers * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$). Comparisons not marked are not significant at $p < 0.05$. The per-document scores used for testing are extracted from the same evaluation pipeline that produces the average numbers in Section 5; only articles with

Config	Abstractiveness		Factuality		Semantic Alignment		
	Tri-g	Dens.	AS	MC	BS	FOCUS↓	U-R
BC	**0.605 (-0.058)**	**4.690 (-1.960)**	**0.762**	**0.890**	0.617	**0.828**	**0.266**
BC+Top-5	0.733 (+0.070)	2.994 (-3.656)	0.637	0.798	0.639	0.931	0.238

Table 13: Ablation of TR-UMLS entity guidance applied to the BC prompt. (Abbreviations as in Table 1 and Table 2)

valid scores from both configurations are included in each pairwise test.

In all tables below, mean differences are computed as $A - B$. Arrows next to metric names indicate whether higher (↑) or lower (↓) values are preferred.

J.2 DPR-BAG vs. Baselines (PMC-MAD)

Table 14, Table 15, and Table 16 report significance for the comparisons between DPR-BAG (BC) and the baseline models on PMC-MAD.

Metric	vs. LED-arXiv	vs. LED-PubMed
Bi-g↓	-0.152^{***}	-0.131^{***}
Tri-g↓	-0.217^{***}	-0.187^{***}
Dens.↓	-12.121^{***}	-9.048^{***}
AS↑	$+0.041^{***}$	$+0.095^{***}$
MC↑	$+0.033^{***}$	$+0.044^{***}$
SummaC↑	-0.167^{***}	-0.124^{***}
BS↑	-0.019^{***}	-0.036^{***}
SENT↑	-0.039^{***}	-0.040^{***}
FOCUS↓	$+0.126^{***}$	$+0.158^{***}$
R-L↑	-0.052^{***}	-0.080^{***}
U-R↑	-0.075^{***}	-0.089^{***}
Cov.↓	-0.017^{***}	-0.011^{***}
Comp.↓	$+20.499^{***}$	$+20.720^{***}$

Table 14: Mean differences (DPR-BAG minus baseline) between DPR-BAG (BC) and the pretrained LED baselines on PMC-MAD.

Metric	vs. LED-PubMed-FT	vs. LongT5
Bi-g↓	-0.032^{***}	-0.128^{***}
Tri-g↓	-0.042^{***}	-0.170^{***}
Dens.↓	-2.383^{***}	-5.565^{***}
AS↑	$+0.251^{***}$	$+0.093^{***}$
MC↑	$+0.236^{***}$	$+0.013^{***}$
SummaC↑	$+0.067^{***}$	-0.142^{***}
BS↑	-0.018^{***}	-0.048^{***}
SENT↑	-0.000	-0.008^{***}
FOCUS↓	$+0.033$	-0.036
R-L↑	-0.082^{***}	-0.093^{***}
U-R↑	-0.099^{***}	-0.047^{***}
Cov.↓	$+0.052^{***}$	-0.015^{***}
Comp.↓	$+22.307^{***}$	$+6.213^{*}$

Table 15: Mean differences (DPR-BAG minus baseline) between DPR-BAG (BC) and the fine-tuned LED-PubMed and LongT5 baselines.

Metric	vs. LongT5-FT
Bi-g↓	-0.059^{***}
Tri-g↓	-0.081^{***}
Dens.↓	-3.048^{***}
AS↑	$+0.221^{***}$
MC↑	$+0.164^{***}$
SummaC↑	$+0.005$
BS↑	-0.034^{***}
SENT↑	-0.027^{***}
FOCUS↓	$+0.111^{***}$
R-L↑	-0.095^{***}
U-R↑	-0.097^{***}
Cov.↓	$+0.021^{***}$
Comp.↓	$+19.467^{***}$

Table 16: Mean differences (DPR-BAG minus baseline) between DPR-BAG (BC) and the fine-tuned LongT5 baseline.

J.3 Splitting Strategies

Table 17 reports significance for the comparisons between FS and other splitting strategies.

Metric	vs. NS	vs. SH
Bi-g↓	$+0.005^{*}$	$+0.045^{***}$
Tri-g↓	$+0.006^{*}$	$+0.046^{***}$
Dens.↓	$+0.150^{**}$	$+0.698^{***}$
AS↑	$+0.006^{*}$	$+0.126^{***}$
MC↑	-0.003	$+0.099^{***}$
SummaC↑	$+0.003$	$+0.141^{***}$
BS↑	$+0.001$	-0.019^{***}
SENT↑	$+0.004^{*}$	$+0.015^{***}$
FOCUS↓	-0.033	-0.098^{***}
R-L↑	$+0.002^{*}$	$+0.004^{***}$
U-R↑	$+0.004$	$+0.037^{***}$
Cov.↓	$+0.000$	$+0.057^{***}$
Comp.↓	$+1.875$	$+1.897$

Table 17: Mean differences between FS and other splitting strategies.

J.4 Prompting Strategies

Table 18 reports significance for the prompting strategy ablation.

J.5 TR-UMLS Entity Guidance

Table 19 and Table 20 report significance for the TR-UMLS entity guidance ablation.

Metric	BC vs. DI	BC vs. SI	DI vs. SI
Bi-g↓	$+0.041^{***}$	$+0.037^{***}$	-0.004^{**}
Tri-g↓	$+0.045^{***}$	$+0.040^{***}$	-0.005^{***}
Dens.↓	$+0.662^{***}$	$+0.616^{***}$	-0.045^{*}
AS↑	$+0.121^{***}$	$+0.132^{***}$	$+0.012^{***}$
MC↑	$+0.084^{***}$	$+0.086^{***}$	$+0.002$
SummaC↑	$+0.133^{***}$	$+0.139^{***}$	$+0.006^{**}$
BS↑	-0.022^{***}	-0.019^{***}	$+0.003^{***}$
SENT↑	$+0.007^{***}$	$+0.004^{**}$	-0.003
FOCUS↓	-0.065^{**}	-0.013	$+0.050^{**}$
R-L↑	$+0.001$	$+0.001$	$+0.001$
U-R↑	$+0.018^{***}$	$+0.014^{***}$	-0.004^{*}
Cov.↓	$+0.052^{***}$	$+0.061^{***}$	$+0.009^{***}$
Comp.↓	$+3.895$	$+5.627$	$+1.732^{*}$

Table 18: Mean differences across prompting strategies.

Metric	vs. DI+Top-5	vs. DI+Top-10
Bi-g↓	$+0.000$	-0.021^{***}
Tri-g↓	-0.000	-0.030^{***}
Dens.↓	-0.011	-0.137^{***}
AS↑	-0.003	-0.002
MC↑	-0.001	-0.006
SummaC↑	$+0.000$	-0.002
BS↑	$+0.001$	$+0.043^{***}$
SENT↑	$+0.002$	$+0.001$
FOCUS↓	-0.048^{**}	-0.026
R-L↑	-0.001	$+0.011^{***}$
U-R↑	$+0.002$	$+0.016^{***}$
Cov.↓	-0.002^{**}	-0.043^{***}
Comp.↓	$+0.395$	-0.069

Table 19: Mean differences between DI and TR-UMLS-augmented DI.

Metric	vs. BC+Top-5	vs. DI+Top-5	vs. DI+Top-10
Bi-g↓	$+0.040^{***}$	$+0.041^{***}$	$+0.021^{***}$
Tri-g↓	$+0.044^{***}$	$+0.045^{***}$	$+0.016^{***}$
Dens.↓	$+0.685^{***}$	$+0.650^{***}$	$+0.525^{***}$
AS↑	$+0.126^{***}$	$+0.118^{***}$	$+0.119^{***}$
MC↑	$+0.092^{***}$	$+0.083^{***}$	$+0.078^{***}$
SummaC↑	$+0.134^{***}$	$+0.133^{***}$	$+0.130^{***}$
BS↑	-0.022^{***}	-0.021^{***}	$+0.022^{***}$
SENT↑	$+0.014^{***}$	$+0.008^{***}$	$+0.008^{***}$
FOCUS↓	-0.107^{***}	-0.113^{***}	-0.087^{***}
R-L↑	$+0.002$	$+0.000$	$+0.012^{***}$
U-R↑	$+0.028^{***}$	$+0.021^{***}$	$+0.035^{***}$
Cov.↓	$+0.051^{***}$	$+0.050^{***}$	$+0.009^{***}$
Comp.↓	$+3.045$	$+4.291$	$+3.827$

Table 20: Mean differences between BC and TR-UMLS-augmented variants.

Metric	BC vs. SI+CoT	SI vs. SI+CoT
Bi-g↓	$+0.032^{***}$	-0.005^{**}
Tri-g↓	$+0.037^{***}$	-0.003
Dens.↓	$+0.591^{***}$	-0.030
AS↑	$+0.166^{***}$	$+0.034^{***}$
MC↑	$+0.098^{***}$	$+0.012^{**}$
SummaC↑	$+0.158^{***}$	$+0.019^{***}$
BS↑	-0.015^{***}	$+0.005^{***}$
SENT↑	-0.000	-0.004^{***}
FOCUS↓	$+0.031$	$+0.042^{*}$
R-L↑	$+0.005^{***}$	$+0.004^{***}$
U-R↑	$+0.017^{***}$	$+0.003$
Cov.↓	$+0.070^{***}$	$+0.009^{***}$
Comp.↓	$+7.488^{*}$	$+1.860^{***}$

Table 21: Mean differences for the SI+CoT ablation.

J.6 Chain-of-Thought Guidance

Table 21 reports significance for the SI+CoT ablation.

J.7 PubMedSum Generalization

Table 22 reports significance for DPR-BAG (BC) against the baseline models on PubMedSum, and Table 23 compares BC against SI+CoT on the same dataset.

J.8 Backbone Scale

Table 24 reports significance for the backbone scaling experiments.

K Case Study: Qualitative Comparison of Generated Abstracts

As shown in Table 25, both LED and LongT5 models exhibit severe verbatim copying from the source full text, with LongT5 further producing a lexical hallucination. In contrast, DPR-BAG demonstrates stronger topic identification and more abstractive generation for the BAG task, though it still occasionally exhibits entity relation confusion (e.g., misattributing target genes as lncRNAs).

Metric	vs. LED-arXiv	vs. LED-PubMed	vs. LongT5
Bi-g↓	-0.226^{***}	-0.190^{***}	-0.190^{***}
Tri-g↓	-0.333^{***}	-0.279^{***}	-0.269^{***}
Dens.↓	-17.806^{***}	-12.261^{***}	-8.710^{***}
AS↑	$+0.010^{**}$	$+0.125^{***}$	$+0.160^{***}$
MC↑	-0.078^{***}	-0.013^{**}	-0.032^{***}
SummaC↑	-0.299^{***}	-0.248^{***}	-0.231^{***}
BS↑	$+0.011^{***}$	-0.012^{***}	-0.013^{***}
SENT↑	-0.099^{***}	-0.115^{***}	-0.080^{***}
FOCUS↓	$+0.719^{***}$	$+1.015^{***}$	$+0.532^{***}$
R-L↑	-0.044^{***}	-0.081^{***}	-0.073^{***}
U-R↑	-0.131^{***}	-0.198^{***}	-0.127^{***}
Cov.↓	-0.035^{***}	-0.024^{***}	-0.027^{***}
Comp.↓	$+21.934^{***}$	$+23.572^{***}$	$+16.835^{***}$

Table 22: Mean differences (DPR-BAG minus baseline) between DPR-BAG (BC) and baseline models on PubMedSum.

Metric	BC vs. SI+CoT
Bi-g↓	-0.002
Tri-g↓	-0.001
Dens.↓	-0.016
AS↑	-0.007^{*}
MC↑	$+0.007$
SummaC↑	$+0.003$
BS↑	-0.000
SENT↑	-0.025^{***}
FOCUS↓	$+0.189^{***}$
R-L↑	-0.006^{***}
U-R↑	-0.028^{***}
Cov.↓	$+0.001$
Comp.↓	$+2.999^{***}$

Table 23: Mean differences between DPR-BAG (BC) and DPR-BAG (SI+CoT) on PubMedSum.

Metric	vs. Qwen-7B	vs. Qwen-14B
Bi-g↓	$+0.006$	$+0.043^{***}$
Tri-g↓	-0.005	$+0.046^{***}$
Dens.↓	$+0.519^{***}$	$+0.736^{***}$
AS↑	$+0.131^{***}$	$+0.115^{***}$
MC↑	$+0.098^{***}$	$+0.078^{***}$
SummaC↑	$+0.139^{***}$	$+0.128^{***}$
BS↑	$+0.048^{***}$	-0.015^{***}
SENT↑	-0.008^{***}	$+0.003$
FOCUS↓	$+0.145^{***}$	$+0.242^{***}$
R-L↑	$+0.023^{***}$	-0.006^{***}
U-R↑	$+0.013^{***}$	-0.070^{***}
Cov.↓	$+0.012^{***}$	$+0.068^{***}$
Comp.↓	$+15.960^{***}$	$+21.741^{***}$

Table 24: Mean differences (Llama-3.2:3B minus Qwen) between Llama-3.2:3B and the scaled Qwen2.5 backbones (denoted Qwen-7B and Qwen-14B).

Model	Generated Abstract (truncated)
Original Abstract	Long non-coding RNAs (lncRNAs) comprise a sizeable class of non-coding RNAs with a length of over 200 base pairs. Little is known about their biological function, although over 20,000 lncRNAs have been annotated in the human genome. Through a diverse range of mechanisms, their primary function is in the regulation of the transcription of protein-coding genes. lncrna transcriptional activation can result from a group of nucleus-retained and chromatin-associated lncrnas, which function as scaffolds in the cis/trans recruitment of transcription factors, co-activators or chromatin remodelers, and/or promoter enhancers. their synthesis is initiated by various processes including ceramide synthesis, release of intracellular ca2+ or acid-base balance disorders. prior to vesicle creation, selective cargo loading occurs in the endosomal sorting complex required for transport. participation of endosomal sorting proteins such as tetraspanins or specific sumoylated proteins required for transport has been indicated in research. the endosomal-sorting complex consists of four components, these induce the formation of multivesicular bodies and the induction of membrane deformation to form exosomes. nanovesicles could be formed inside multivesicular bodies to allow transport outside the cell or digestion in lysosomes. the molecular content of exosomes is more heterogenic than its synthesis process, with different cargoes being examined inside vesicles with regard to the type or stage of cancers. this paper will review the importance of lncrnas as crucial molecular content of exosomes, indicating its involvement in tumour suppression, pro-tumorigenic events and the development of novel therapeutic approaches in the near future. further studies of their mechanisms of function are essential, as well as overcoming several challenges to gain a clearer insight to the approaches for the best clinical application.
LED-Pubmed (FT)	The exosome is a type of extracellular vesicle involved in cell-to-cell communication and disease transmission. Exosomes can be found in urine, sperm, amniotic fluid and other biological fluids, and have been implicated in various biological processes, such as cancer development, immune responses, invasion, invasion and invasion, cell proliferation, invasion or invasion, migration and invasion of tumour cells, as well as in the regulation of gene expression in cancer cells. the lncrna (non-coding rna) is a class of transcripts that do not encode proteins and whose length exceeds 200 bp and therefore play a vital role in regulating signaling pathways and gene expression, and therefore plays a pivotal role in cancer development and cancer-related diseases such as metastasis, angiogenesis and tumour cell invasion, and is involved in the formation of exosomal lipid bilayers, migration, proliferation and invasion in tumour microenvironment, which in turn are involved in cancer progression and progression, as evidenced by their involvement in the secretion of cytokines and growth factors, lipids or proteins from endosomes, and in the activation of the transcription factor hypoxia inducible factor (hif-β) pathway, among others, and their role in the induction of pre-metastatic niche formation in cancerous diseases, including lung cancer. the purpose of this review is to summarize the role of various exosomally-derived lnas in carcinogenesis and in other biological processes and to highlight the importance of these molecules in cancer biology and cancer development.
LongT5 (Base)	Circulating long non-coding rnas are a class of transcripts that do not encode proteins and whose length exceeds 200 bp. They play a significant role in diverse cellular processes from enabling normal development to disease progression, as well as cell communication. A valuable example of a biological carrier that can transport abundant quantities of long non - codons is that of the exosome. its role has been documented in both physiological conditions and pathological changes such as cancer development, where they are responsible for the regulation of various processes including immunosuppression, proliferation and induction of pre - metastatic niche formation. in this review, the role of exosomal long non
DPR-BAG (BC)	LncRNA has greater tissue specificity than mRNA, making it a potential diagnostic biomarker for specific tumour types. Circulating RNA in body fluids can resist ribonucleases, providing stability and potential as cancer prognostic markers. A study found two groups of long non-coding RNAs (lncRNAs) with opposite effects on cancer cells. Gas5 was shown to induce apoptosis and inhibit tumour growth, while Particl enhanced DNA methylation and gene silencing. Another lncRNA, CCND1, bound to a modulator and blocked transcription factors. Exo1-4 lncRNAs increased recipient cell viability by interacting with proteins involved in metabolism and nucleosomal architecture. Exosomal long non-coding RNAs(lncRNAs) have shown potential as effective nanocarriers for regulating lncRNA expression in cancer cells, offering a promising tumour therapy approach.

Table 25: Qualitative comparison of generated abstracts for a sample article (PMCID: PMC6625196) from the PMC-MAD test set. Orange: verbatim copy from source full text. Red: factual error.