# Enhancing Scientific Title Generation via Optimized Sentence Ordering

**Thanh-Thien Khuu[1,2], Thien-Thuan Huynh[1,2], Nam Van Chi[1,2], Tung Le[1,2,*]**
[1]Faculty of Information Technology, University of Science, Vietnam
[2]Vietnam National University, Ho Chi Minh city, Vietnam
{ktthien22,htthuan22}@clc.fitus.edu.vn
{vcnam, lttung}@fit.hcmus.edu.vn
[*]Corresponding author: Tung Le - lttung@fit.hcmus.edu.vn

## Abstract

Generating concise titles for machine learning abstracts is essential for navigating complex literature but challenging due to specialized terminology and dense text structures. We propose a novel sentence ordering method that uses a two-stage sequence-to-sequence BART framework, augmented by an auxiliary model that extracts sentences with the highest keyword overlap to the title and ranks candidate sentence permutations. The optimal ordering guides BART to produce coherent and concise titles. We evaluate our method on the test split of a large dataset of over 21,000 machine learning title–abstract pairs from Springer journals. Results show that structured input via optimized sentence ordering improves title quality compared to baseline models. These findings highlight sentence ordering as an underexplored yet effective strategy for enhancing scientific text generation.

## 1 Introduction

Machine learning (ML) research has surged in recent years, with Springer journals publishing thousands of abstracts that distill cutting-edge advances, from neural architecture search to reinforcement learning. Titles play a pivotal role in this ecosystem, serving not only as entry points for researchers but also as essential elements for indexing and retrieval across academic platforms like SpringerLink. Crafting a title that concisely conveys complex technical contributions remains a non-trivial task, especially when abstracts contain dense, domain-specific terminology (e.g., “attention mechanisms”) and exhibit structural dispersion, where key ideas are embedded across multiple, non-adjacent sentences. These characteristics pose significant challenges for automated title generation systems, which must distill and reorganize salient content into coherent and informative summaries.

In this work, we utilize BART, a denoising autoencoder (Lewis et al., 2019), for this title generation task. While previous methods typically treat abstracts as flat, unordered text, our approach introduces a two-stage sequence-to-sequence pipeline that incorporates sentence-ordering cues to improve the semantic coherence and informativeness of generated titles. By doing so, we propose a novel strategy that bridges document structure awareness with abstractive summarization, paving the way for more effective automated indexing of scientific literature.

Our method begins with input structuring, where a RoBERTa-based auxiliary model (Liu et al., 2019) is trained to identify important sentences from each abstract using heuristic labels derived from keyword overlap with the gold title. A permutation scoring module subsequently ranks up to `max_permutation` shuffled sentence orderings according to their similarity to the gold title embedding. The permutation that receives the highest score, which reflects coherence and relevance to the title, is selected as the input. This refined input is then used in the title generation stage, where a BART model is trained to produce scientific titles. By conditioning the generator on input that is more coherent and focused, the model learns to generate concise and accurate titles. Training and evaluation are carried out on a curated dataset of machine learning title and abstract pairs obtained from Springer journals. Our results demonstrate that incorporating sentence reordering not only improves standard metrics such as ROUGE and BERTScore, but also enhances semantic alignment with reference titles. This two-stage approach highlights the value of structural preprocessing in abstractive title generation and offers a practical framework for improving title quality in scientific publishing workflows.

Our contributions are as follows:

- We present a novel two-stage method that integrates sentence selection and ordering using RoBERTa and BART to improve scientific title generation in the machine learning domain.

- We construct and release a new dataset of over 21k machine learning research title-abstract pairs collected from Springer journals.[1] [2]

- We demonstrate that our approach improves over a strong BART baseline in both ROUGE and BERTScore, showing consistent gains in title relevance and coherence.

The paper is organized as follows. Section 2 reviews related work, Section 3 details our methodology, Section 4 presents evaluations, Section 5 analyzes input properties, Section 6 discusses the results, and Section 7 concludes.

## 2 Related Work

### 2.1 Scientific Title Generation

Scientific title generation is often framed as an extreme summarization task, aiming to distill the essence of a research article or abstract into a concise and informative title. Early approaches relied on statistical and rule-based methods, such as feature-based classifiers that extracted keywords from abstracts to construct titles (Kupiec et al., 1995). These methods, while computationally efficient, struggled with capturing nuanced semantic relationships and often produced formulaic outputs lacking domain-specific precision.

The advent of neural networks marked a significant shift in title generation. Encoder-decoder architectures, particularly recurrent neural networks (RNNs), were employed to generate abstractive summaries and titles (Nallapati et al., 2016). These models improved fluency but were limited by their dependence on sequential processing, which often failed to capture long-range dependencies in complex scientific texts. To address this, retrieval-based methods, such as $k$-nearest-neighbors ($k$-NN) approaches, leveraged word co-occurrence patterns in abstracts to propose candidate titles (Putra and Khodra, 2017). While effective for general-purpose texts, these methods often generated generic or overly broad titles when applied to scientific domains, where precision and specificity are paramount.

More recently, pre-trained transformer models, such as GPT-2 and T5, have been fine-tuned for title generation tasks (Riku and Masaomi, 2022). These models generate multiple candidate titles, which are then ranked or refined using heuristic or learned scoring mechanisms. For instance, fine-tuned GPT-2 models have been used to propose diverse title candidates, with post-processing steps to select the most contextually relevant option. However, transformer-based approaches often struggle with domain-specific scientific terminology and may produce titles that lack the innovative or precise phrasing required in academic contexts. Additionally, these models typically treat the input text as a fixed sequence, ignoring the potential benefits of restructuring the input for better coherence or informativeness.

Recent advances have explored the incorporation of domain knowledge into title generation. For example, some approaches integrate scientific ontologies or citation networks to enhance the relevance of generated titles. Others have experimented with hybrid models that combine neural generation with rule-based constraints to ensure adherence to domain-specific conventions. Despite these advances, a key limitation persists: existing methods do not explicitly optimize the input structure (e.g., sentence ordering) to enhance the quality of generated titles. Our work addresses this gap by introducing a novel pipeline that ranks sentence permutations using an auxiliary RoBERTa scorer to guide BART title generation, ensuring that the input structure maximizes coherence and informativeness for scientific title generation.

### 2.2 Sentence Ordering in NLP

Sentence ordering is a critical task in Natural Language Processing, aimed at arranging sentences to maximize coherence and logical flow, which is particularly important for text generation, summarization, and question answering. Early approaches to sentence ordering relied on heuristic methods, such as ranking sentences based on lexical cohesion or syntactic patterns. However, these methods often failed to capture deep semantic relationships, leading to suboptimal arrangements in complex texts like scientific abstracts.

Neural approaches have significantly improved sentence ordering. For instance, Gong et al. (2016) introduced an end-to-end pointer network to model

[1] Source: https://link.springer.com

[2] Dataset: https://www.kaggle.com/datasets/tiamatt/springerjournal-450tk-0-5cosine

sentence sequences for summarization, achieving better coherence than traditional methods. Similarly, Logeswaran et al. (2017) used RNNs to treat sentence ordering as a coherence optimization problem, leveraging sequential dependencies. More recently, transformer-based models like BERT have employed sentence-level representations, such as their dedicated [CLS] embeddings, to predict optimal orderings through pairwise comparisons or sequence modeling (Devlin et al., 2019). However, to our knowledge, no prior work has directly used [CLS] embeddings to score and rank sentence permutations specifically for sentence ordering.

Our work proposes a novel application of sentence ordering tailored specifically for scientific title generation. Unlike previous approaches, which mainly focus on coherence in summarization or narrative tasks, we use `roberta-base`'s [CLS] embeddings to rank sentence permutations, optimizing the input structure for a `bart-base` title generation model. This allows the generated titles to be not only coherent but also better aligned with the key contributions and domain-specific content of the abstract, addressing an important gap in the literature.

## 3 Methodology

We present a novel method to enhance title generation by optimizing the sentence order within scientific abstracts. Our approach consists of three key stages: (1) training a `roberta-base` model to identify salient sentences, (2) generating multiple sentence permutations and ranking them , and (3) fine-tuning a `BART-base` model on the highest-ranked permutations to produce coherent and informative titles. This section outlines the full pipeline (see Figure 1), including model architectures, data pre-processing, and training strategies.

### 3.1 Sentence Selection

To estimate the relevance of a sentence with respect to the title, we train the `roberta-base` (Liu et al., 2019) encoder as a regression model. Each sentence $S_i$ in the abstract is independently encoded using RoBERTa, with a special [CLS] token prepended to the input sequence for sentence-level representation. A scalar relevance score is predicted via a linear head applied to the [CLS] embedding.

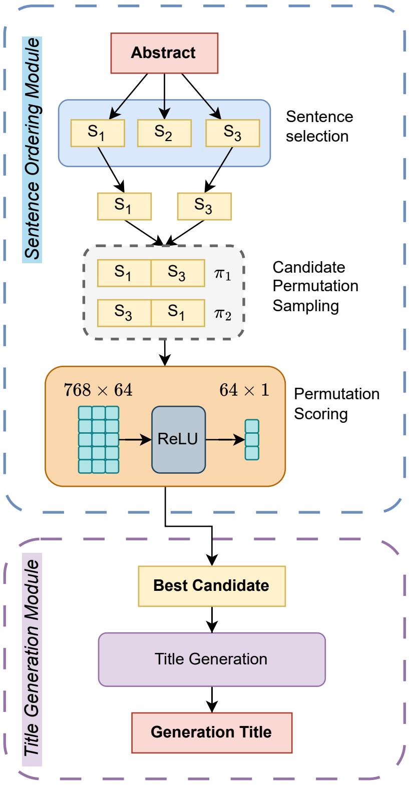

Figure 1: Overview of the title generation architecture with sentence extraction and permutation scoring.

Let $h_i \in \mathbb{R}^d$ be the [CLS] embedding vector extracted from the last hidden state of RoBERTa for sentence $S_i$ and $y_i \in [0, 1]$ the normalized keyword overlap score (see Equation 8). The model employs a linear head with parameters $W \in \mathbb{R}^{768}$ and $b \in \mathbb{R}$, computing:

$$\hat{y}_i = W^\top h_i + b \tag{1}$$

where the output is a single scalar score without activation to constrain it to $[0, 1]$. The model is trained to minimize the Mean Squared Error (MSE) loss function:

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{2}$$

where $\hat{y}_i$ is the predicted score and $y_i$ is the ground truth label, and $N$ is the number of sentences in the batch.

We demonstrate our model's sentence selection process using the abstract from Logeswaran et al. (2017) as input:

### 3.2 Dual Encoder Architecture

Building on the extracted sentences from the previous stage, the dual encoder framework reorders

Table 1: Example of model input and output for sentence extraction in scientific abstracts. The yellow box highlights the original abstract, while the green box contains the sentences selected by the model.

**Title:**
Sentence Ordering and Coherence Modeling using Recurrent Neural Networks

**Original abstract:**

Modeling the structure of coherent texts is a key NLP problem. The task of coherently organizing a given set of sentences has been commonly used to build and evaluate models that understand such structure. We propose an end-to-end unsupervised deep learning approach based on the set-to-sequence framework to address this problem. Our model strongly outperforms prior methods in the order discrimination task and a novel task of ordering abstracts from scientific articles. Furthermore, our work shows that useful text representations can be obtained by learning to order sentences. Visualizing the learned sentence representations shows that the model captures high-level logical structure in paragraphs. Our representations perform comparably to state-of-the-art pre-training methods on sentence similarity and paraphrase detection tasks.

**Sentences selected by the model:**

1. Modeling the structure of coherent texts is a key NLP problem.
2. We propose an end-to-end unsupervised deep learning approach based on the set-to-sequence framework to address this problem.
3. Our model strongly outperforms prior methods in the order discrimination task and a novel task of ordering abstracts from scientific articles.
4. Visualizing the learned sentence representations shows that the model captures high-level logical structure in paragraphs.
5. Our representations perform comparably to state-of-the-art pre-training methods on sentence similarity and paraphrase detection tasks.

these inputs to improve title generation. The architecture integrates a main encoder-decoder for generation and an auxiliary encoder for ranking sentence permutations.

**Main Encoder-Decoder.** A `bart-base` model serves as the primary sequence-to-sequence component. It takes a linearized sequence of selected sentences as input and generates the corresponding scientific title.

**Auxiliary Encoder and Scoring Head.** To guide the sentence ordering, a `roberta-base` encoder is paired with a trainable scoring head, which is a two-layer multilayer perceptron (MLP) implemented as $\text{Linear}(768 \rightarrow 64) \rightarrow \text{ReLU} \rightarrow \text{Linear}(64 \rightarrow 1)$, where the input is the `[CLS]` embedding of each permutation. The output is a scalar score indicating the estimated quality of the sentence ordering.

**Permutation Sampling and Scoring.** For each abstract, a set of permutations is sampled from the previously selected sentences by randomly shuffling their order. To maintain computational feasibility, only a fixed number of unique permutations are generated per abstract. Each permutation is then encoded and scored by the auxiliary model to estimate its informativeness and coherence. The permutation receiving the highest score is selected as input to the main encoder-decoder for title generation.

**Proxy-Based Ranking Supervision.** To supervise the scorer, we use a weak proxy signal based on semantic similarity to the gold title. Each permutation $\pi_i$ and the gold title $t$ are encoded using the frozen auxiliary encoder, and their `[CLS]` embeddings $h_{\pi_i}$ and $h_t$ are extracted. The proxy score for each permutation is computed as the cosine similarity:

$$\text{proxy}(\pi_i) = \cos(h_{\pi_i}, h_t) = \frac{h_{\pi_i} \cdot h_t}{\|h_{\pi_i}\| \, \|h_t\|} \tag{3}$$

Despite containing the same content, different sentence orderings lead to distinct contextualized embeddings due to the encoder's positional encodings and attention patterns. Thus, the cosine similarity reflects the degree to which a permutation semantically aligns with the target title.

Prior to computing pairwise supervision signals, both the predicted scores $\phi(h_{\pi_i})$ and the proxy scores $\text{proxy}(\pi_i)$ are sorted in descending order to obtain their respective ranking positions. This ensures that relative pairwise preferences are computed consistently.

To train the scorer, we align its predicted scores with the proxy scores using a differentiable approximation of Kendall's tau rank correlation (Kendall, 1938). We define the pairwise target label for each permutation pair $(\pi_i, \pi_j)$ as:

$$y_{ij} = \frac{\text{sign}(\text{proxy}(\pi_i) - \text{proxy}(\pi_j)) + 1}{2} \tag{4}$$

which equals 1 if $\pi_i$ should be ranked above $\pi_j$, and 0 otherwise. The soft prediction of the scorer is:

$$\hat{y}_{ij} = \sigma\left(\frac{\phi(h_{\pi_i}) - \phi(h_{\pi_j})}{\tau}\right) \tag{5}$$

where $\sigma$ is the sigmoid function, $\tau$ is a temperature hyperparameter, and $\phi(h_\pi)$ denotes the scorer's scalar output for permutation $\pi$.

The final ranking loss is computed using the binary cross-entropy (BCE) between the predicted and proxy-based pairwise preferences:

$$\mathcal{L}_{\text{kendall}} = \frac{1}{N(N-1)} \sum_{i \neq j} \text{BCE}(\hat{y}_{ij}, y_{ij}) \quad (6)$$

This formulation softly penalizes misalignments between the predicted permutation order and the proxy-induced ranking while remaining fully differentiable.

**Overall Training Objective.** The final loss combines the main sequence-to-sequence generation objective with the auxiliary ranking supervision. Let $\mathcal{L}_{\text{seq2seq}}$ denote the cross-entropy loss between the generated and gold titles. The complete training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{seq2seq}} + \lambda \mathcal{L}_{\text{kendall}} \quad (7)$$

where $\lambda$ is a tunable hyperparameter balancing the two components.

# 4 Experiments and Evaluations

## 4.1 Dataset

We created a custom dataset for training and evaluating our model by scraping title–abstract pairs from SpringerLink journals in the machine learning domain. The dataset includes 21,545 English-language articles, focusing on subfields like deep learning, neural networks, and natural language processing.

To ensure quality and consistency, we limited abstracts to 128–450 tokens and titles to 8–32 tokens. Poorly aligned title–abstract pairs were filtered out based on cosine similarity between their embeddings (generated via a Sentence-Transformer model), retaining only pairs with a similarity score above 0.5.

For vocabulary and topic consistency across splits, we used stratified sampling based on key machine learning terms (e.g., 'CNN', 'transformers', 'reinforcement learning'). The dataset was split into training (80%, 17,236 samples), validation (10%, 2,154 samples), and test (10%, 2,155 samples) subsets, proportional to each group's size.

Data were sourced from open access metadata and abstracts, with no restrictive licenses from Springer journals prohibiting their use for research at the time of collection.

For sentence extraction, we split each abstract into individual sentences using the Natural Language Toolkit (Loper and Bird, 2002). To determine sentence importance, we adopt a keyword-based scoring method. Specifically, we use KeyBERT (Grootendorst, 2020) to extract keywords from both the gold title and each sentence in the abstract. We define an overlap score between a sentence and the title as:

$$\text{score}(S_i) = \frac{|KW(S_i) \cap KW(T)|}{KW(T)} \quad (8)$$

where $KW(S_i)$ and $KW(T)$ are the sets of keywords extracted for sentence $S_i$ and title $T$, respectively.

These scores serve as soft labels to supervise a RoBERTa as a scoring model, allowing it to learn to predict sentence relevance in alignment with the title semantics (see Figure 2).

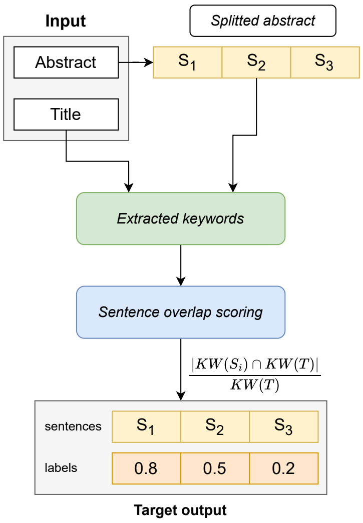

Figure 2: Overview of the data preprocessing pipeline for sentence extraction and labeling.

## 4.2 Experimental Setup

We perform a manual hyperparameter search using a dataset of 21,545 abstract–title pairs, split into 80% training (17,236), 10% validation (2,154), and 10% testing (2,155), preserving the original distribution. The title generation model is initialized from BART-base, and the sentence scoring model uses a RoBERTa-base encoder. The model is trained for 3 epochs with a batch size of 4, a learning rate of 2e-5, and gradient accumulation over 8 steps, giving an effective batch size of 32.

For each abstract, the top 7 sentences are selected using the sentence importance module. Up to 30 unique permutations of these sentences are then sampled by random shuffling and scored by the auxiliary encoder. The highest-scoring permutation is used as input to the main encoder-decoder for title generation.

The scorer is trained with soft Kendall's tau loss (see Equation 6), using a temperature $\tau = 0.5$ and a ranking loss weight $\lambda = 1.5$. All experiments are run on a Kaggle-provided Tesla P100 GPU, taking about 6 hours for training the sentence extraction model and 8 hours for the title generation model. To ensure statistical reliability under permutation variability, we repeat each experiment five times without fixed seeds and report the mean and standard deviation of the evaluation metrics.

### 4.3 Evaluation Metrics

We evaluate generated titles using both lexical overlap and semantic similarity metrics. For lexical evaluation, we report ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum (Lin, 2004). ROUGE-1 and ROUGE-2 measure unigram and bigram overlap, respectively, while ROUGE-L captures the longest common subsequence between the generated and reference titles. ROUGE-Lsum is designed for summarization tasks and computes ROUGE-L at the sentence level, which better aligns with abstractive generation tasks such as ours.

To assess semantic similarity, we use BERTScore (Zhang et al., 2020), which computes token-level similarity using contextual embeddings from a pretrained BERT model. We report the precision, recall, and F1 scores, with F1 as the primary semantic metric. BERTScore captures meaning beyond surface form and correlates well with human judgment in generation tasks.

### 4.4 Results and Analysis

We evaluate the effectiveness of our proposed method by comparing it against both a strong encoder-decoder baseline and several recent instruction-tuned large language models (LLMs). Table 2 reports performance across ROUGE and BERTScore metrics.

**Standard BART Baseline.** We first compare with a `facebook/bart-base` model trained on full abstracts in their original sentence order, without any sentence filtering or reordering. This model serves as a conventional baseline for scientific title generation. As shown in Table 2, our proposed method consistently outperforms this baseline, demonstrating the benefit of explicitly modeling sentence selection and coherence.

**LLM-Based Generation.** We compare our method with three instruction-tuned large language models: GPT-4.1 Mini (OpenAI, 2025), Gemini 2.5 Flash (Gemini Team, 2025), and a fine-tuned LLaMA 3.2 (1B) (Grattafiori et al., 2024). For GPT and Gemini, we use a zero-shot approach, prompting each model to create a title from the full abstract with the instruction: *"Write a short, formal and clear title for this scientific research. Return ONLY the title: <abstract>"*.

For LLaMA 3.2 (1B), we evaluate two settings. First, we fine-tune the model on full abstracts without sentence selection or reordering. This improves over zero-shot prompting but remains weaker than BART and our method. Second, we integrate the same sentence selection and ordering strategy as in our proposed method. This variant (*LLaMA 3.2 + Our Method*) yields substantial gains over the plain fine-tuned version, especially in ROUGE and recall-oriented BERTScore, showing that sentence-level control is crucial even for LLMs. However, it still trails the improved BART model, suggesting that encoder-decoder architectures remain better suited for compact title generation under structured input.

As shown in Table 2, the zero-shot LLMs perform significantly worse than the supervised baselines, especially in ROUGE metrics. Their outputs often miss key technical terms or include generic phrasing, which lowers both precision and recall. The fine-tuned LLaMA model performs better than the zero-shot models, showing that task-specific training helps. However, it still lags behind our proposed method in all metrics. This highlights the importance of selecting relevant sentences and presenting them in a coherent order before generation. Our method benefits from this sentence-level control, achieving stronger coverage (high recall) and more accurate phrasing (high precision and F1).

**Overall Performance.** The results highlight that task-specific supervision and input control (through sentence selection and ordering) are more effective than prompting general-purpose LLMs.

Table 2: Performance comparison on the test set. Results are reported as the mean ± standard deviation over 5 runs. We evaluate using ROUGE-1/2/L and BERTScore (Precision, Recall, and F1).

| **Model** | **ROUGE-1** | **ROUGE-2** | **ROUGE-L** | **Precision** | **Recall** | **F1** |
|---|---|---|---|---|---|---|
| BART-base (baseline) | 56.76 | 36.67 | 49.33 | 91.66 | 90.84 | 91.23 |
| GPT-4.1 mini | 13.41 | 7.16 | 10.37 | 87.83 | 80.71 | 84.12 |
| Gemini 2.5 Flash | 10.86 | 6.22 | 9.00 | 88.52 | 80.38 | 84.24 |
| LLaMA 3.2 1B Instruct | 30.39 | 17.95 | 26.33 | 83.00 | 85.14 | 83.94 |
| LLaMA 3.2 1B (w/ Method) | 44.19 | 24.89 | 35.22 | 87.56 | 91.28 | 89.37 |
| **Proposed Method** | **59.69 ± 0.33** | **38.99 ± 0.24** | **51.66 ± 0.23** | **92.01 ± 0.09** | **91.06 ± 0.06** | **91.52 ± 0.04** |

While integrating our method into LLaMA narrows the gap with BART, the encoder-decoder model still achieves the strongest balance of precision, recall, and F1. Our method therefore demonstrates both the utility of structured input and the architectural advantage of supervised sequence-to-sequence learning for scientific title generation.

To complement the quantitative results, we present a qualitative comparison using the same abstract previously introduced in the sentence selection analysis (Section 3.1). The example, shown in the box below (see Table 3), includes the gold title and titles generated by different models. All titles were generated using a maximum length of 32 tokens to ensure a fair comparison across models.

The **gold title** explicitly conveys both the core task (sentence ordering and coherence modeling) and the methodological framework (recurrent neural networks). The **BART baseline** generates a fluent but generic title that lacks task specificity and fails to mention sentence ordering. **Our method (Dual Encoder)** improves on this by directly referencing sentence ordering, thus more accurately reflecting the research focus, albeit in a simpler phrasing.

Among the LLMs, **GPT-4.1 mini** produces the most faithful and specific title. It correctly identifies both the set-to-sequence modeling approach and the sentence ordering task, resulting in a well-structured and informative title that closely aligns with the abstract. **Gemini 2.5 Flash**, while fluent and coherent, shifts focus toward sentence representation learning and omits the sentence ordering aspect, partially reflecting the abstract. **LLaMA-3.2 1B Instruct** captures the main task, sentence ordering, and introduces the concept of coherence, but phrases it in broader terms. While its title is shorter and more abstract than the others, it still reflects key ideas and avoids hallucinating unrelated terms.

The **LLaMA-3.2 1B Instruct + Our Method** variant shows how structured input control influences generation. Its output is more detailed and task-relevant than the plain LLaMA version, explicitly highlighting abstract ordering and positioning the method as unsupervised deep learning. However, by adding elements like “sentence similarity” and “paraphrase detection” not present in the abstract, it becomes informative but less precise than the gold reference.

These examples show the trade-offs across models in terms of task relevance, specificity, and fluency. LLMs like GPT and Gemini generate polished and expressive outputs, but may emphasize secondary elements or reframe the task. LLaMA, while more concise, remains grounded in the core ideas. In contrast, our model offers a more targeted and faithful summary of the task, striking a balance between relevance and simplicity without introducing content outside the source abstract.

## 5 Input Property Analysis

To study how input characteristics affect model performance, we examined two properties of the abstracts: the number of sentences and the total token count. We used a fixed test set and measured the relationship between these properties and the performance of the model using ROUGE and BERTScore. The results are shown in Table 4.

Across all ROUGE variants, we find a weak but statistically significant negative correlation with both token count and sentence count ($r$ between $-0.048$ and $-0.071$, $p < 0.05$). This suggests that longer abstracts tend to slightly reduce lexical overlap with the reference titles. In contrast, BERTScore does not show a significant correlation with either measure ($p > 0.05$), indicating that semantic similarity is largely unaffected by abstract length.

## 6 Discussion

Our results show that adding sentence selection and ordering to a standard text generation model can

Table 3: Qualitative comparison of generated titles from various models. The yellow highlights denote outputs from task-specific transformer models (e.g., BART baseline and our dual-encoder method), while the green highlights indicate outputs from general-purpose large language models (LLMs), including GPT, Gemini, and LLaMA.

**Gold Title:**
Sentence Ordering and Coherence Modeling using Recurrent Neural Networks

**GPT-4.1 mini:**

Unsupervised Deep Set-to-Sequence Modeling for Coherent Text Structure and Sentence Ordering

**Gemini 2.5 Flash:**

Unsupervised Deep Learning for Coherent Text Structuring and Sentence Representation Learning

**LLaMA-3.2 1B Instruct:**

Sentence ordering: a new approach to model coherence

**LLaMA-3.2 1B Instruct + Our method**

Ordering abstracts from scientific articles: an end-to-end unsupervised deep learning approach based on sentence similarity and paraphrase detection tasks.

**BART Baseline:**

An end-to-end unsupervised deep learning approach for coherent sentences

**Our Method (Dual Encoder):**

An unsupervised deep learning approach to order sentences

Table 4: Pearson correlation between abstract length and evaluation metrics. "Sent." refers to the number of sentences in the abstract, and "Tok." refers to the abstract token count. $r$ is Pearson's correlation coefficient, and $p$ is the corresponding significance value.

| **Metric** | **Sent.** $r$ | **Sent.** $p$ | **Tok.** $r$ | **Tok.** $p$ |
|---|---|---|---|---|
| **ROUGE-1** | -0.048 | 0.025 | -0.051 | 0.017 |
| **ROUGE-2** | -0.070 | 0.001 | -0.060 | 0.005 |
| **ROUGE-L** | -0.071 | 0.001 | -0.065 | 0.003 |
| **BERTScore F1** | -0.037 | 0.089 | -0.034 | 0.115 |

improve the quality of generated scientific titles. The full model performs better than the baseline across all metrics, though the gains are not significant. This suggests that transformer models like BART already do a good job, but guiding them with more structured input can still help.

Looking at the generated examples, our method produces titles that are more relevant to the task and better grounded in the input abstract. In comparison, large language models (LLMs) generate fluent and polished titles, but sometimes add terms that were not mentioned in the input. This makes them less reliable in settings where accuracy matters. Among the LLMs, GPT produces the most specific and faithful output, while Gemini tends to generalize or shift focus slightly. LLaMA produces a concise and mostly relevant title, but its phrasing is more abstract.

The input property analysis shows small but consistent negative correlations between abstract length and ROUGE scores, meaning longer abstracts tend to have less lexical overlap with the reference titles. Correlations with BERTScore are weaker and not significant, indicating that semantic similarity is mostly unaffected. This suggests that longer inputs may add wording variation without reducing the ability to capture the main meaning, supporting the role of sentence selection in removing less relevant content.

### 6.1 Why BART Outperforms Larger Models and LLMs

Despite the emergence of larger and more sophisticated language models, our results consistently show that the BART-based approach with structured input processing outperforms both general-purpose LLMs (GPT-4.1, Gemini 2.5) and even LLaMA 3.2 enhanced with our proposed method. We hypothesize several key factors underlying this counterintuitive finding.

**Task-specific architectural advantage.** BART's encoder-decoder architecture is specifically designed for text generation tasks that require distilling and restructuring information. Unlike decoder-only models (GPT, LLaMA), BART can explicitly separate the encoding and decoding phases, allowing for better control over input representation and output generation. This separation enables the model to better focus on the most relevant parts of the input during encoding while maintaining generation fluency during decoding.

**Supervised fine-tuning vs. general instruction following.** Our BART model is fine-tuned directly on the scientific title generation task with thousands of abstract-title pairs from the target domain. In contrast, general-purpose LLMs rely on instruction-following capabilities acquired during pre-training and instruction tuning across diverse tasks. While this makes LLMs more versatile, it may dilute their focus on the specific constraints and conventions of scientific title generation, such as maintaining technical precision while achieving conciseness.

**Input control and structured processing.** The combination of sentence selection and ordering creates a more focused and coherent input representation that plays to BART's strengths. Our analysis shows that even when this structured input approach is applied to LLaMA (yielding substantial improvements), it still falls short of BART's performance. This suggests that the encoder-decoder architecture is better suited to leverage structured inputs for generation tasks, as it can dedicate the entire encoder to processing the ordered sentences before generating the title.

**Precision-recall balance in constrained generation.** Scientific title generation requires a delicate balance between covering key concepts (recall) and avoiding extraneous information (precision). Our results show that while LLMs excel at fluency and creativity, they often introduce terms not present in the source abstract or generalize concepts beyond what is warranted. BART with structured input achieves a better precision-recall trade-off, generating titles that are both comprehensive and faithful to the source content.

**Training data alignment and domain specificity.** Our BART model is trained specifically on scientific abstracts from machine learning journals, allowing it to learn domain-specific patterns in terminology, structure, and style. While LLMs have seen vast amounts of text during pre-training, their knowledge is distributed across many domains and tasks, potentially making them less attuned to the specific requirements of scientific title generation in this domain.

## 7 Conclusion

In this work, we studied how to improve scientific title generation by adding sentence selection and ordering to a transformer-based model. These steps help the model focus on the most important parts of the abstract and arrange them in a more logical way. Our experiments, both with automatic metrics and sample outputs, showed that this extra structure makes the generated titles more relevant and aligned with the input.

Although the improvements over the baseline were not significant, both selection and ordering gave us more control over the content. This is especially useful in fields where accuracy and clarity matter. The input property analysis showed that abstract length impacts lexical overlap, supporting the role of sentence selection in enhancing title relevance, even if the overall impact is modest. Our results suggest that adding simple structure to input can make models more controllable without needing more compute. In the future, these ideas could be used as planning steps for LLMs, applied to other types of text, or combined with human feedback to make better decisions. This work gives useful insights into how to balance structure and fluency in text generation.

## Limitations

Although our method improves title generation by optimizing sentence ordering, it has several constraints. The proxy supervision signal—cosine similarity between abstract sentences and title embeddings—relies on pretrained encoders and may fail to capture subtle semantics, especially for metaphorical or abstract titles, leading to noisy ranking guidance. The cap of 30 sampled permutations limits the search space and risks overlooking better sentence orderings in content-rich abstracts. Our evaluation is confined to machine learning abstracts from Springer journals, whose relatively uniform rhetorical structures may not reflect the variability found in biomedical, humanities, or informal domains, raising concerns about generalization. Furthermore, the absence of human evaluation restricts interpretability, as automated metrics alone cannot fully assess fluency or informativeness. Finally, the need for auxiliary scoring and multiple forward passes adds computational overhead, which may hinder scalability to large datasets or real-time applications unless optimized further.

## Acknowledgments

This research is supported by research funding from Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

# References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Jingjing Gong, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. End-to-end neural sentence ordering using pointer network. *Preprint*, arXiv:1611.04953.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Julian Kupiec, Jan O. Pedersen, and Francine R. Chen. 1995. A trainable document summarizer. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2017. Sentence ordering and coherence modeling using recurrent neural networks. *Preprint*, arXiv:1611.02654.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *Preprint*, arXiv:cs/0205028.

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

OpenAI. 2025. Introducing gpt-4.1 in the api.

Jan Wira Gotama Putra and Masayu Leylia Khodra. 2017. Automatic title generation in scientific articles for authorship assistance: A summarization approach. *Journal of ICT Research and Applications*, 11(3):253–267.

Matsumoto Riku and Kimura Masaomi. 2022. A title generation method with transformer for journal articles. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1115–1120.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.