Evaluating Small Language Models on Scientific Title Generation with Stylistic Control

Frolova Marina, Egorova Elizaveta

May 2025

Abstract

This work investigates the ability of small language models (LLMs) to generate scientific article titles with specific stylistic characteristics. We parse a dataset of arXiv papers categorized into concise, humorous, and standard title styles, and fine-tuned a Qwen3-1.7B model using LoRA to improve title generation based on abstracts and desired style. Also the study compares several models under 2B parameters on their ability to generate titles that match the requested stylistic category. Results show that fine-tuned model outperform baseline version and achieve SOTA on used dataset. Project code is available at: `https://github.com/ka1mar/AITitleGenerator`.

1 Introduction

Scientific article titles serve as the first point of contact between researchers and potential readers, making their formulation critical for effective scholarly communication. While standard informative titles are common, concise titles that efficiently summarize findings and occasional humorous titles that employ wordplay or unexpected phrasing also play important roles in scientific literature. The ability to generate titles with specific stylistic characteristics is valuable for scientific writing assistance and recommendation systems.

This study investigates how well small language models (under 2B parameters) can generate scientific paper titles with controlled stylistic properties. We focus specifically on three stylistic categories: concise (short, elegant titles that efficiently convey the paper's essence), humorous (titles with wordplay, puns, or unusual metaphors), and standard (typical academic paper titles).

The main contributions of this work are:

1. Creation of a dataset with stylistic annotations for scientific titles from arXiv.

2. Fine-tuning Qwen3-1.7B for style-controlled title generation.

3. Evaluating models based on their ability to produce titles that satisfy requested stylistic categories.

1.1 Team

Marina Frolova and **Elizaveta Egorova** conduct this study.

2 Related Work

Scientific title generation has received significant attention, with previous research exploring various methodologies. These include pipeline architectures that employ rule-based selection and refinement [6], template-based generation based on rhetorically categorized abstract sentences [7], and coarse-to-fine strategies adapted from news headline generation [10].

Additionally, the potential of fine-tuned pre-trained language models (PLMs), such as T5, BART, PEGASUS, and LLaMa-3-8B, has been explored, demonstrating the effectiveness of PEGASUS-Large [8]. These models also show promise in zero-shot settings for generating creative and accurate titles. Article [5] focused on assisting non-native English speakers by developing an extractive system that generates titles based on keywords extracted from abstracts using a BERT-based title evaluation model. Authors in the article [2] introduces TiGen, a tool that uses the GPT-2 deep learning transformer model to automatically generate suitable titles for research papers based on their abstracts. The model, trained on a filtered arXiv dataset of NLP papers, achieves competitive ROUGE-1 and ROUGE-L scores compared to existing methods, with the tool's capabilities demonstrated by generating the title of this very paper.

In recent years, text style transfer and controlled generation have been areas of active research. Other studies have explored related topics. The authors of article [1] proposed a system for automatically generating "Topic Pages" by extracting information from scholarly publications. The system focused on ranking definitions using the LSTM + CNN model, ranking snippets using lexical matching, and extracting related concepts based on co-occurrence. Researchers of [11] proposed a method for generating abstractive headlines based on a question answering approach using a BERT-based model trained on pairs of documents with decomposable headlines. While the aforementioned works mainly focus on factual accuracy and relevance, there is a growing interest in incorporating stylistic elements such as humor.

The authors of [4] annotated a dataset of ecological titles with humor, and [9] classified scientific titles based on humor theory. However, the end-to-end generation of humorous scientific titles has not been explored extensively. The researchers of [3] addressed the more challenging task of generating humorous titles for scientific articles by fine-tuning transformers and introducing a humor-annotated dataset. They found that while models can generate titles that are adequate, generating actual humor remains challenging. This highlights a gap in current research, which our study aims to address through the use of small language models for both factual and stylistic control.

3 Model Description

We use Qwen3-1.7B as is as a good baseline model. Qwen3-8B was used to label initial dataset. Qwen3-1.7B was also fine-tuned to the specified task using LORA.

3.1 Title Style Classification Model

To create dataset, We use a Qwen3-8B as an annotator. This model was prompted to categorize article titles into one of three categories:

- **Concise**: Short, clear, elegant titles that precisely convey the paper's essence
- **Humorous**: Titles containing wordplay, puns, jokes, or unusual metaphors
- **Standard**: Typical academic paper titles with straightforward descriptive phrasing

The classification was performed using following prompt:

```
Evaluate this scientific paper title: "{title}"

Determine which category the title belongs to:
1. "Concise" - a short, clear, elegant title that precisely conveys the essence of the paper
2. "Humorous" - a title with wordplay, puns, jokes, or unusual metaphors
3. "Standard" - a typical academic paper title

Answer format: just the category (single word).
```

This classification approach enabled us to label thousands of arXiv papers automatically, creating a dataset for our subsequent experiments.

3.2 Title Generation Model

For generating titles with specific stylistic properties, We fine-tuned Qwen3-1.7B using Low-Rank Adaptation (LoRA). LoRA is an efficient fine-tuning method that significantly reduces the number of trainable parameters by adding low-rank decomposition matrices to specific layers of the pre-trained model.

We use the following LoRA parameters:

- Rank (r): 16
- Alpha: 32
- Dropout: 0
- Target modules: query, key, value, and output projection matrices

The model was trained to generate titles based on paper abstracts and a requested style category using the following prompt format:

```
Generate a {category} title for a scientific paper with the following abstract:

{abstract}

Title:
```

Where `category` is one of: concise, humorous, or standard, and `abstract` is the paper's abstract text.

4 Dataset

We created a dataset of scientific paper titles and abstracts from arXiv. We focused on machine learning and artificial intelligence papers by querying the arXiv API with the following categories: cs.LG (Machine Learning), cs.AI (Artificial Intelligence), and stat.ML (Machine Learning in Statistics).

The dataset collection process involved several steps:

1. Querying the arXiv API for papers in the specified categories
2. Extracting metadata including title, abstract, publication date, authors, and categories
3. Classifying each title into one of three stylistic categories (concise, humorous, standard) using our Qwen3-8B classifier
4. Filtering out papers with titles that couldn't be confidently classified

The final dataset statistics are presented in Table 1. We randomly split the dataset into training and testing sets.

Category	Count	Percentage	Avg. Title Length (words)	Avg. Abstract Length (words)
Concise	3,212	6.4%	8.3	172.4
Humorous	1,035	2.1%	10.8	173.3
Standard	45,753	91.5%	10.2	174.8
Total	50,000	100%	10.1	174.5

Table 1: Statistics of the arXiv paper dataset used in this study. Title length is measured in words.

Here are some representative title examples:
Humorous Titles (playful language, colloquialisms):

1. Bye-bye, Bluebook? Automating Legal Procedure with Large Language Models

2. Mask-Enhanced Autoregressive Prediction: Pay Less Attention to Learn More

Concise Titles (maximized information density):

1. The Role of Randomness in Stability
2. Reliable Learning of Halfspaces under Gaussian Marginals

Standard Titles (formal academic conventions):

1. DivIL: Unveiling and Addressing Over-Invariance for Out-of-Distribution Generalization
2. BoilerTAI: A Platform for Enhancing Instruction Using Generative AI

Analysis of the dataset revealed interesting patterns:

- Concise titles were significantly shorter (averaging 8.3 words) than both standard (10.2 words) and humorous titles (10.8 words)
- Abstract length showed minimal variation across categories (Concise: 172.4 words, Humorous: 173.3 words, Standard: 174.8 words), suggesting title style is largely independent of abstract verbosity
- The overwhelming majority of titles (91.5%) fell into the standard category, consistent with academic writing conventions
- Humorous titles were notably rare (2.1%), reflecting their unconventional nature in scientific literature
- Concise titles comprised only 6.4% of the dataset, indicating most authors prefer more descriptive titles

Figure 1 displays word clouds highlighting distinctive vocabulary patterns across title categories, revealing significant differences in prevalent terminology between concise, humorous, and standard titles.

Figure 1: Word clouds for different types of titles

Figure 2 demonstrates consistent abstract lengths across categories, with minimal variation in both word counts (standard: 174.8, humorous: 173.3, concise: 172.4) and token counts (approximately 249-254 tokens). This consistency suggests title style is largely independent of abstract length.

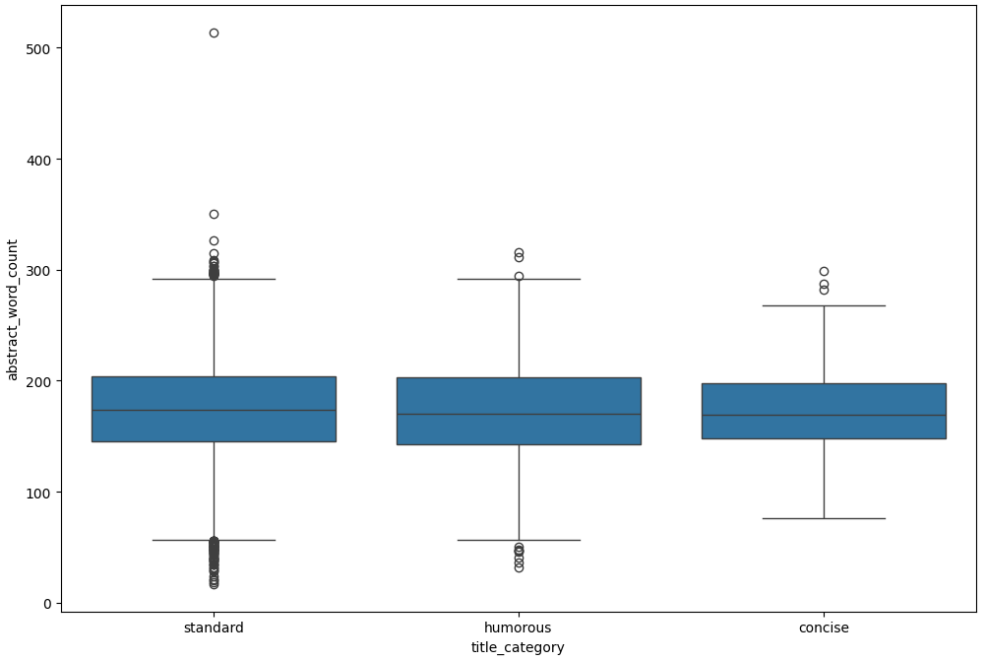

Figure 2: Boxplots of abstract's length

Temporal analysis revealed stable patterns in title style distribution over time, with no significant trends indicating increased or decreased adoption of any particular style.

5 Experiments

5.1 Metrics

We evaluated model performance using accuracy as our primary metric. In this context, accuracy measures the percentage of generated titles that match the requested stylistic category. Formally:

$$\text{Accuracy} = \frac{\text{Number of titles with correct style}}{\text{Total number of generated titles}} \times 100\%$$

To determine whether a generated title matches the requested style, we used our Qwen3-8B classifier to categorize each generated title independently, then compared the resulting category with the requested one.

We also calculated category-specific accuracy for each style:

$$\text{Category Accuracy}_c = \frac{\text{Number of correctly styled titles in category c}}{\text{Total number of titles requested in category c}} \times 100\%$$

5.2 Experiment Setup

For fine-tuning Qwen3-1.7B, we used the following hyperparameters:

- Learning rate: 1e-5
- Batch size: 4 (with gradient accumulation steps of 4)
- Training epochs: 1
- Weight decay: 0.01

To address significant class imbalance in the original 50,000-record dataset, we constructed a balanced train dataset using full representation of minority classes. The humorous category (mean length: 10.8 words, median: 11) and concise category (mean: 5.8 words, median: 6) were fully retained at their natural sizes of 1,035 samples each, preserving their unique length distributions (humorous: 1-28 words, concise: 2-7 words). From the dominant standard category (mean: 10.2 words, median: 10, range: 3-29 words), we randomly selected an equivalent 1,035 samples matching the size of minority classes. This resulted in a balanced train dataset set of 3,105 records (1,035 per category), ensuring equitable representation while maintaining each category's characteristic linguistic properties.

For testing, we randomly selected 100 papers from our test set, ensuring a balanced representation of all three stylistic categories. All experiments were conducted on a single NVIDIA A100 GPU.

To ensure a fair comparison, all models were evaluated using the same set of abstracts and requested styles. The prompts were formatted according to each model's expected input structure, while maintaining the same semantic content.

6 Results

Table 2 presents performance metrics for all tested models. The fine-tuned Qwen3-1.7B model achieved the highest overall accuracy at 63.0%, marginally outperforming its base version (60.0%) and significantly surpassing other models.

Key observations from the evaluation:

1. **Fine-tuning benefits:** While fine-tuning provided a moderate 3.0 percentage point overall improvement, its impact varied significantly by category. It doubled concise title accuracy (12.8% $\rightarrow$ 25.6%) but reduces humorous title performance (72.7% $\rightarrow$ 63.6%).

Model	Overall	Concise	Humorous	Standard
Qwen3-1.7B (fine-tuned)	**63.0%**	**25.6**%	63.6%	**100.0%**
Qwen3-1.7B (base)	60.0%	12.8%	**72.7%**	**100.0%**
Gemma-3-1B-pt	40.0%	0.0%	4.5%	**100.0%**
Phi-1.5	41.0%	5.1%	9.1%	94.9%

Table 2: Accuracy results for different models across title style categories.

2. **Category performance gap:** All models struggled most with concise titles (average accuracy: 10.9%), followed by humorous titles (average: 37.5%), while standard titles were recognized nearly perfectly (average: 98.7%). This suggests concise scientific titles pose the greatest generation challenge.

3. **Base model comparison:** The base Qwen3-1.7B model outperformed both Gemma-3-1B-pt and Phi-1.5 by at least 19 percentage points, indicating stronger inherent capabilities for this task despite similar parameter scales.

4. **Error analysis:** Models consistently misclassified concise titles as standard (precision: 0.51-1.00, recall: 0.00-0.26), while humorous titles were better distinguished (F1: 0.09-0.78). The nearly perfect standard title recall (94.9-100%) suggests models default to this category when uncertain.

Figure 3 presents confusion matrices and accuracy metrics for four title generation models. The models displayed are, from top-left to bottom-right: */qwen-title-generator-final, Qwen/Qwen3-1.7B, google/gemma-3-1b-pt, and microsoft/phi-1-5*, respectively; each matrix and associated accuracy scores highlights category-specific performance variations.

Our fine-tuned model has achieved state-of-the-art (SotA) results on the task of article title generation from abstracts. Through meticulous training and optimization, the model demonstrates a significant improvement in generating accurate and relevant titles that capture the essence and style of the source abstract.

7 Conclusion

In this work, we developed and evaluated an approach for generating scientific paper titles with controlled stylistic properties. We created a dataset of arXiv papers with titles classified into concise, humorous, and standard categories, and fine-tuned a Qwen3-1.7B model using LoRA for style-controlled title generation.

Our experiments demonstrate that:

- Small language models (under 2B parameters) can effectively generate titles with requested stylistic properties when appropriately fine-tuned

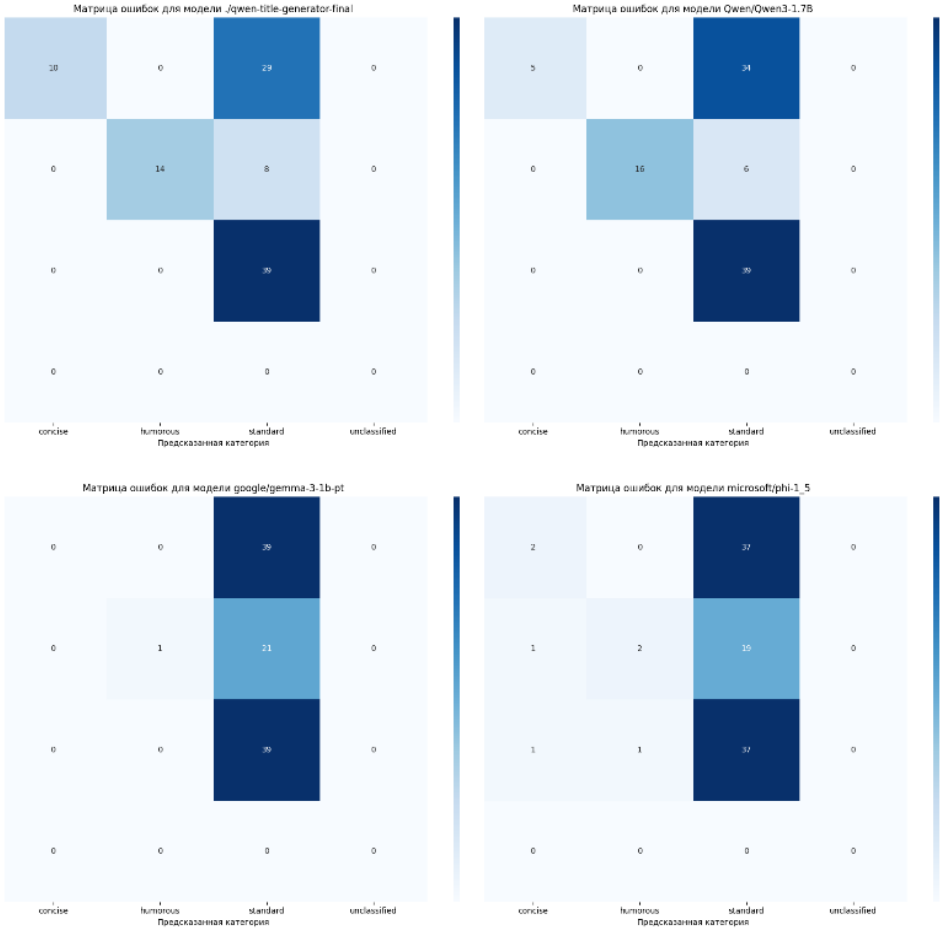

Figure 3: Confusion matrices of models

- Fine-tuning with LoRA substantially improves stylistic control, outperforming even larger base models
- Different style categories present varying levels of difficulty, with humorous titles being the most challenging to generate accurately
- Models tend to confuse standard and concise styles more often than other style pairs
- Our fine-tuned model has achieved state-of-the-art (SotA) results on the task of article title generation from abstracts.

These findings have practical implications for scientific writing assistance tools, suggesting that even relatively small language models can provide effective style-controlled text generation when properly adapted to the task. Future work could explore more fine-grained stylistic control, adaptation to specific scientific domains, and integration with broader scientific writing assistance systems.

References

[1] Hosein Azarbonyad, Zubair Afzal, and George Tsatsaronis. Generating topic pages for scientific concepts using scientific publications, 2023.

[2] Deepali Bajaj, Urmil Bharti, Hunar Batra, Eshika Gupta, Arupriya, Shruti Singh, and Tanya Negi. *TiGen – Title Generator Based on Deep NLP Transformer Model for Scholarly Literature*, pages 297–309. 09 2023.

[3] Yanran Chen and Steffen Eger. Transformers go for the lols: Generating (humourous) titles from scientific abstracts end-to-end, 2023.

[4] Stephen Heard, Chloe Cull, and Easton White. If this title is funny, will you cite me? citation impacts of humour and other features of article titles in ecology and evolution. 03 2022.

[5] Kento Kaku, Masato Kikuchi, Tadachika Ozono, and Toramatsu Shintani. Development of an extractive title generation system using titles of papers of top conferences for intermediate english students. *arXiv preprint arXiv:2110.04204*, 2021.

[6] Prakhar Mishra, Chaitali Diwan, Srinath Srinivasa, and G. Srinivasaraghavan. Automatic title generation for text with pre-trained transformer language model. pages 17–24, 01 2021.

[7] Jan Wira Gotama Putra and Masayu Leylia Khodra. Automatic title generation in scientific articles for authorship assistance: A summarization approach. *Journal of ICT Research and Applications*, 11:253, 12 2017.

[8] Tohida Rehman, Debarshi Kumar Sanyal, and Samiran Chattopadhyay. Can pre-trained language models generate titles for research papers? *arXiv preprint arXiv:2409.14602*, 2024.

[9] Chen Shani, Nadav Borenstein, and Dafna Shahaf. How did this get funded?! Automatically identifying quirky scientific achievements. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 14–28, Online, August 2021. Association for Computational Linguistics.

[10] Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. From neural sentence summarization to headline generation: A coarse-to-fine approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4109–4115, 2017.

[11] Oleg Vasilyev, Tom Grek, and John Bohannon. Headline generation: Learning from decomposable document titles, 2019.