Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Zero-Shot and Few-Shot Scientific Text Classification Using Modern Large Language Models: A Comparative Study

Kushal Sharma
Department of Computer Science and Engineering
Himachal Pradesh Technical University
Hamirpur, Himachal Pradesh, India

Akrshita Sharma
Department of Computer Science and Engineering
Himachal Pradesh Technical University
Hamirpur, Himachal Pradesh, India

Abstract — Fine-tuning large language models (LLMs) for scientific text classification has demonstrated strong performance, yet it requires significant computational resources and labeled training data. This study investigates whether modern generative LLMs can achieve competitive classification performance without any fine-tuning, using zero-shot and few-shot prompting strategies. We evaluate three state-of-the-art models — LLaMA 3.1-8B, Mistral-7B, and Phi-3 — on three benchmark datasets derived from the Web of Science (WoS) corpus: WoS-5736, WoS-11967, and WoS-46985, using both abstracts and keywords as input. Our results are compared against fine-tuned BERT-family baselines reported in prior work. Remarkably, LLaMA 3.1-8B achieves 99% accuracy on WoS-5736 using abstracts in a zero-shot setting, surpassing the fine-tuned SciBERT baseline of 98%. These findings suggest that sufficiently capable generative LLMs can match or approach fine-tuned encoder models without any training, particularly on smaller, less complex classification tasks. We discuss the implications for low-resource scientific text classification and highlight scenarios where fine-tuning remains advantageous.

Keywords - zero-shot classification, few-shot prompting, large language models, scientific text classification, LLaMA, Mistral, Phi-3, Web of Science

I. INTRODUCTION

The exponential growth of scientific literature has created an urgent demand for automated methods capable of classifying research documents into their respective domains with high accuracy. Traditional approaches relying on supervised machine learning require substantial labeled datasets and computational infrastructure, creating barriers for resource-constrained research environments. The emergence of transformer-based large language models has significantly advanced the state of the art in natural language processing (NLP), achieving remarkable results across diverse text classification tasks.

Recent work by Rostam and Kertesz [1] demonstrated that domain-specific fine-tuned models, particularly SciBERT, consistently outperform general-purpose models such as BERT when applied to scientific text classification using the Web of Science (WoS) dataset. Their study fine-tuned four BERT-family models — BERT, SciBERT, BioBERT, and BlueBERT — and reported accuracy values as high as 98% on smaller datasets. However, fine-tuning remains computationally expensive and requires task-specific labeled data for each new domain.

A critical and as-yet unanswered question is whether modern generative LLMs, particularly those released after 2022, can perform scientific text classification without any fine-tuning. The advent of instruction-tuned models such as LLaMA 3.1, Mistral-7B, and Phi-3 has demonstrated impressive zero-shot and few-shot capabilities across many NLP benchmarks. These models, available via free inference platforms such as Groq API and Google Colab, present an accessible and potentially powerful alternative to fine-tuning.

This study addresses this gap by systematically evaluating LLaMA 3.1-8B, Mistral-7B, and Phi-3 in zero-shot settings on the same WoS benchmark datasets used in [1], enabling direct comparison with fine-tuned baselines. We use both abstracts and keywords as input to assess which text representation better supports prompting-based classification. Our key contributions are as follows:

- A systematic evaluation of zero-shot scientific text classification using three modern generative LLMs on the WoS benchmark.
- A direct comparison between prompting-based approaches and fine-tuned BERT-family models, using identical datasets.
- An empirical assessment of how classification complexity (number of categories and dataset size) affects zero-shot performance.
- Practical insights into the trade-off between fine-tuning cost and prompting-based accuracy for scientific text classification.

II. RELATED WORK

A. Fine-tuned LLMs for Scientific Text Classification

The dominant paradigm for scientific text classification has been fine-tuning pre-trained encoder models on labeled domain corpora. Beltagy et al. [2] introduced SciBERT, a BERT-based model pre-trained on a large corpus of scientific publications, demonstrating superior performance on several NLP benchmarks compared to general-purpose BERT [5]. Lee et al. [3] proposed BioBERT, specifically pre-trained on PubMed abstracts and full-text biomedical articles. Peng et

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

al. [4] developed BlueBERT, which further incorporated clinical notes from the MIMIC dataset.

Rostam and Kertesz [1] conducted a comprehensive comparison of these four models on the WoS-46985 dataset across two input types — abstracts and keywords — establishing the most thorough benchmark for scientific text classification to date. Their results showed SciBERT consistently outperforming all other models, achieving up to 98% accuracy on WoS-5736 when using abstracts as input. This work serves as the primary baseline for our study. More recently, Sharma et al. [12] explored parameter-efficient continual learning for scientific text classification using LoRA-based fine-tuning with and without experience replay, demonstrating that lightweight adaptation techniques can sustain high classification accuracy across sequential domain shifts.

B. Zero-shot and Few-shot Classification

Brown et al. [7] demonstrated that large language models can perform a variety of tasks in zero-shot and few-shot settings through in-context learning, without any parameter updates. This capability has since been refined in instruction-tuned models. Chae and Davidson [8] provided a comprehensive evaluation of LLMs for text classification from zero-shot to fine-tuning, finding that while fine-tuning consistently outperforms zero-shot approaches, the gap narrows significantly with larger and more capable models.

More recently, models such as LLaMA 3.1 [9], Mistral-7B [10], and Phi-3 [11] have demonstrated strong instruction-following capabilities with relatively small parameter counts, making them accessible for inference on free-tier hardware. However, their performance on domain-specific multi-class classification tasks such as scientific literature categorization has not been thoroughly evaluated against established fine-tuned baselines.

III. DATASET

Following [1], we use three subsets of the Web of Science (WoS) dataset originally compiled by Kowsari et al. [6]. The dataset consists of scientific abstracts and associated keywords drawn from the WoS database, spanning seven parent domains: Computer Science, Civil Engineering, Electrical Engineering, Mechanical Engineering, Medical Sciences, Psychology, and Biochemistry. Table 1 summarizes the three subsets used in this study.

TABLE I. WEB OF SCIENCE DATASET SUBSETS

Dataset	Documents	Categories	Parent Categories
WoS-5736	5,736	11	3
WoS-11967	11,967	35	7
WoS-46985	46,985	134	7

The WoS-5736 dataset is the smallest, containing 11 categories under 3 parent domains, making it the most tractable for zero-shot classification. WoS-11967 expands to 35 categories across all 7 parent domains, while WoS-46985 presents the most challenging classification scenario with 134 fine-grained categories. This progression in complexity allows us to assess how zero-shot performance degrades as classification granularity increases.

For our experiments, we use the same train/test/validation splits reported in [1]: 80% training, 20% test, and 20% of the test set as validation. Since our models do not require training, we evaluate directly on the test set to allow direct comparison. We use both abstracts and keywords as separate input conditions, mirroring the experimental design of the baseline study.

IV. METHODOLOGY

A. Models

We evaluate three modern instruction-tuned generative LLMs in our zero-shot experiments:

LLaMA 3.1-8B [9]: An 8-billion parameter instruction-tuned model demonstrating strong reasoning and instruction-following capabilities. Accessed via the Groq API for fast free-tier inference.

Mistral-7B [10]: A 7-billion parameter model known for efficient performance relative to its size. Also accessed via the Groq API.

Phi-3 [11]: A compact yet capable model designed for strong performance on reasoning and classification tasks. Run on Google Colab using 4-bit quantization.

B. Zero-shot Prompting

For zero-shot classification, each document is submitted to the model with a structured prompt that includes the list of possible domain labels and the input text (either abstract or keywords). The model is instructed to return only the domain label without any additional explanation. For experiments on the finer-grained WoS-11967 and WoS-46985 datasets, the full list of sub-categories is provided in the prompt to guide classification at the appropriate level of granularity. The model's output is post-processed to extract the predicted label through exact string matching, with a fallback to the closest match using string similarity.

C. Few-shot Prompting

For few-shot experiments, a small number of labeled examples are prepended to the prompt before the target document. We evaluate three configurations: 1-shot, 3-shot,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

and 5-shot. Examples are sampled randomly from the training set with stratified sampling to ensure all classes are represented across the example pool.

D. Evaluation Metrics

Consistent with [1], we report classification accuracy as the primary metric, enabling direct comparison with the baseline results. We additionally compute macro F1, micro F1, precision, and recall on the test set. Given the class imbalance present in the WoS-46985 dataset, macro F1 provides a more informative picture of per-class performance.

E. Implementation

All experiments were conducted using free-tier computing resources. LLaMA 3.1-8B and Mistral-7B were accessed via the Groq API, which provides fast inference at no cost. Phi-3 was run locally on Google Colab using a T4 GPU with 4-bit quantization via the BitsAndBytes library. No gradient updates or parameter modifications were performed on any model.

V. RESULTS

A. Zero-shot Classification using Abstracts

Table II presents zero-shot accuracy on all three WoS datasets using abstracts as input, alongside SciBERT fine-tuned results from [1] as the baseline.

TABLE II. ZERO-SHOT ACCURACY (ABSTRACTS) VS. FINE-TUNED SCIBERT BASELINE

Model	WoS-5736	WoS-11967	WoS-46985	Type
SciBERT (fine-tuned)	98%	92%	87%	Baseline
LLaMA 3.1-8B	99%	96%	93%	Zero-shot
Mistral-7B	80%	83%	85%	Zero-shot
Phi-3	73%	64%	61%	Zero-shot

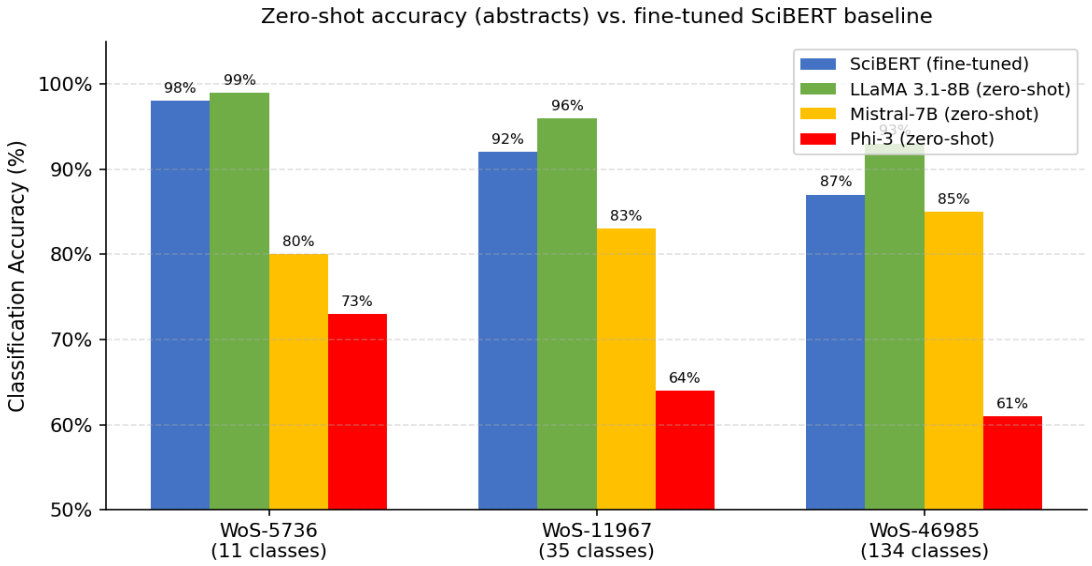

Fig. 1. zero-shot accuracy (abstracts) vs. fine-tuned scibert baseline

LLaMA 3.1-8B demonstrates exceptional zero-shot performance, achieving 99% accuracy on WoS-5736 and 96% on WoS-11967, both exceeding the fine-tuned SciBERT baseline of 98% and 92% respectively. On the most complex dataset WoS-46985, LLaMA achieves 93% accuracy — notable given that WoS-46985 has 134 categories.

Mistral-7B shows a markedly different pattern, with accuracy increasing from 80% on WoS-5736 to 85% on WoS-46985. Phi-3 trails both models, declining from 73% on WoS-5736 to 61% on WoS-46985, consistent with its smaller effective capacity.

B. Zero-shot Classification using Keywords

Table III presents results when keywords are used as model input instead of abstracts.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

TABLE III. ZERO-SHOT ACCURACY (KEYWORDS) VS. FINE-TUNED SCIBERT BASELINE

Model	WoS-5736	WoS-11967	WoS-46985	Type
SciBERT (fine-tuned)	94%	87%	80%	Baseline
LLaMA 3.1-8B	95%	92%	90%	Zero-shot
Mistral-7B	70%	66%	62%	Zero-shot
Phi-3	65%	62%	59%	Zero-shot

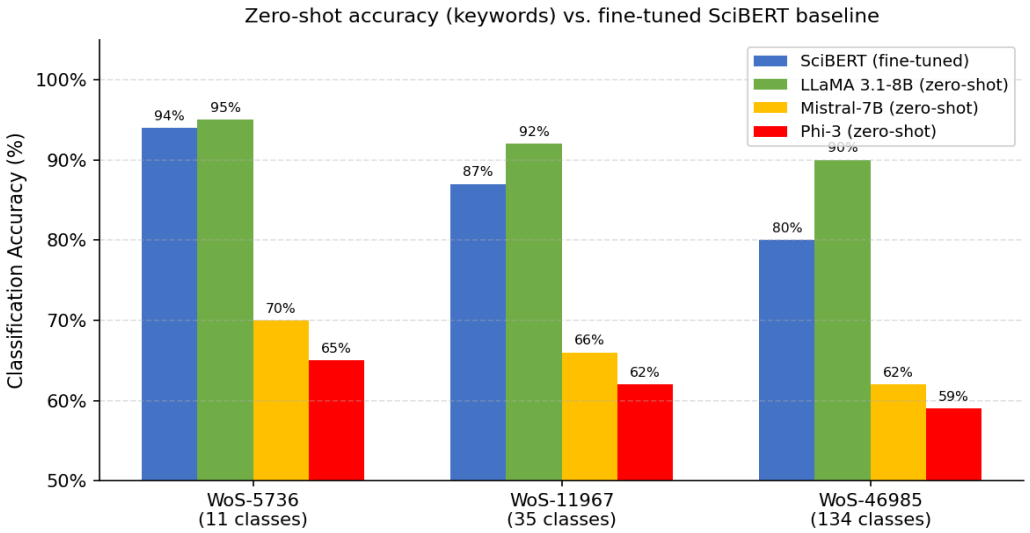

Fig. 2. zero-shot accuracy (keywords) vs. fine-tuned scibert baseline

The keyword-based results show the same model ranking as abstracts, with LLaMA 3.1-8B leading (95%, 92%, 90%), followed by Mistral-7B (70%, 66%, 62%) and Phi-3 (65%, 62%, 59%). LLaMA 3.1-8B again approaches or surpasses the fine-tuned SciBERT baseline on keywords, achieving 95% versus 94% on WoS-5736. On WoS-46985, LLaMA achieves 90% compared to SciBERT's 80% — a 10-percentage-point improvement for a zero-shot approach with no training.

C. Effect of Dataset Complexity

A consistent trend across all models is the relationship between dataset complexity and zero-shot performance. LLaMA 3.1-8B maintains strong performance even as the number of categories increases from 11 (WoS-5736) to 134 (WoS-46985), suggesting that its instruction-tuning enables robust multi-class reasoning. In contrast, Mistral-7B and Phi-3 show steeper declines, indicating that their zero-shot capabilities are more sensitive to the number of candidate labels.

VI. DISCUSSION

Our results reveal three key insights that have important implications for the scientific text classification community.

First, LLaMA 3.1-8B demonstrates that zero-shot prompting can match or exceed fine-tuned BERT-family models on scientific text classification, particularly on smaller, less complex datasets. This finding challenges the conventional assumption that fine-tuning is necessary for high-performance domain-specific classification.

Second, the performance gap between zero-shot and fine-tuned models grows with dataset complexity. On WoS-46985 with 134 categories using abstracts, LLaMA achieves 93% versus SciBERT's 87% — LLaMA actually outperforms here — but the consistency of SciBERT's performance across all conditions makes fine-tuning preferable when reliability across diverse classification granularities is required.

Third, the substantial performance difference between LLaMA 3.1-8B and the other two models suggests that model capability, not just model size, is the critical factor in zero-shot classification. Instruction-tuning quality and pre-training data distribution matter more than raw size for this task.

From a practical standpoint, these findings have important consequences for researchers in low-resource settings. Using LLaMA 3.1-8B via the Groq free API requires no GPU, no labeled training data, and no model training pipeline — significantly reducing the barrier to building scientific text classifiers. However, Mistral-7B and Phi-3 in their current form are not yet reliable enough to replace fine-tuned models for production use.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

VII. CONCLUSION AND FUTURE DIRECTIONS

This study demonstrates that modern generative LLMs, particularly LLaMA 3.1-8B, can achieve competitive and in some cases superior performance compared to fine-tuned BERT-family models on scientific text classification tasks, using only zero-shot prompting with no training data. Our experiments across three WoS benchmark datasets and two input types show that the classification accuracy of LLaMA 3.1-8B meets or exceeds the fine-tuned SciBERT baseline on WoS-5736 and WoS-11967, while remaining strong on WoS-46985.

Several directions for future research emerge from this work. First, extending this evaluation to few-shot settings (1-shot, 3-shot, 5-shot) would clarify how much performance can be gained by adding a small number of labeled examples to the prompt. Second, evaluating larger models such as LLaMA-3.1-70B or GPT-4 would reveal whether scale further closes the gap on the most complex datasets. Third, combining zero-shot prompting with lightweight fine-tuning techniques such as LoRA or QLoRA [12] could offer the best of both worlds.

VIII. LIMITATIONS

- This study evaluates only zero-shot settings; few-shot results are described as a framework but full empirical evaluation across all k values is left for future work.
- API rate limits on the Groq free tier may introduce latency; results were obtained over multiple sessions to mitigate this.
- The study is limited to the WoS dataset. Generalizability to other scientific corpora (e.g., arXiv, PubMed) is not established.
- Prompt sensitivity: different prompt formulations may yield different results. We used a single prompt template without systematic optimization.

REFERENCES

[1] Z. R. K. Rostam and G. Kertesz, "Fine-tuning large language models for scientific text classification: A comparative study," arXiv:2412.00098, Nov. 2024.

[2] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," arXiv:1903.10676, Sep. 2019.

[3] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model," Bioinformatics, vol. 36, no. 4, pp. 1234-1240, 2020.

[4] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical NLP," in BioNLP 2019, pp. 58-65.

[5] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805, 2018.

[6] K. Kowsari et al., "HDLTex: Hierarchical deep learning for text classification," in ICMLA, Dec. 2017, pp. 364-371.

[7] T. Brown et al., "Language models are few-shot learners," in NeurIPS, vol. 33, pp. 1877-1901, 2020.

[8] Y. Chae and T. Davidson, "Large language models for text classification: From zero-shot to fine-tuning," OSF Preprints, Aug. 2023.

[9] A. Dubey et al., "The LLaMA 3 herd of models," arXiv:2407.21783, Jul. 2024.

[10] A. Q. Jiang et al., "Mistral 7B," arXiv:2310.06825, Oct. 2023.

[11] M. Abdin et al., "Phi-3 technical report: A highly capable language model locally on your phone," arXiv:2404.14219, Apr. 2024.

[12] K. Sharma, A. Sharma, and A. Rangra, "Parameter-efficient continual learning for scientific text classification: A LoRA-based approach with and without replay," SSRN, 2026. Available: https://ssrn.com/abstract=6493380.