BioTitleGen: Title Generator for Medical Abstracts

Naresh Kumar, Deepak Srivastava, Xinyi Zhu
{naresh_kumar, dsri, xinyi_zhu}@ischool.berkeley.edu
GitHub: https://github.com/UC-Berkeley-I-School/datasci-266-2025-summer-project

Abstract

We present **BioTitleGen**, a hybrid summarization framework for generating coherent and semantically accurate titles from medical abstracts using large language models (LLMs). Starting with an encoder–decoder Transformer baseline, we evaluate decoder-only LLMs with larger parameter counts, enhancing performance through prompt engineering and fine-tuning on a curated medical dataset. We further benchmark against a domain-specific BART model fine-tuned for biomedical literature. Results show that general-purpose decoder-only LLMs, when fine-tuned and guided with optimized prompts, can match or surpass specialized models, achieving superior ROUGE and SBERT scores. These findings underscore the potential of adapting general LLMs for domain-specific summarization tasks such as medical title generation.

1 Introduction

As the biomedical literature continues to grow at an unprecedented rate, healthcare professionals struggle to efficiently extract relevant information. With over 38 million citations indexed in PubMed, the need for effective summarization tools are more critical than ever. This study investigates automated title generation for clinical research abstracts using state-of-the-art natural language processing models.

We conducted a comparative analysis of various summarization models—including PEGASUS, Llama 3.1, Qwen3, Gemma 2 and PubMed BART—evaluating their performance against an extractive summarization baseline using both automatic (ROUGE, SBERT) and human-centric evaluation metrics. We trained our models using three distinct training datasets and generated inferences on a held-out test set. The generated titles were evaluated using a combination of lexical and semantic similarity metrics, including ROUGE-1, ROUGE-2, ROUGE-L, cosine similarity from an all-round sentence transformer tuned for many use-cases and PubMed-specific sentence transformer. We are computing the cosine similarity between two sentences which is the reference title and the generated title using a sentence transformer model from the sentence-transformers library.

Our findings show that fine-tuned LLMs—particularly LLaMA 3.1 and Qwen3—produce titles that are more informative and semantically faithful than those generated by extractive methods. On average, cosine similarity and PubMed cosine scores improved over extractive baselines, indicating greater semantic relevance. These semantic scores were also consistently higher than ROUGE scores, which prioritize surface-level word overlap, further highlighting the models' ability to capture meaning beyond lexical similarity. Prompt optimization and lightweight fine-tuning further enhanced overall performance. While challenges such as title truncation, topic drift, and hallucinations persist, human evaluation confirms the overall superiority of LLM-generated titles in relevance and readability. This work demonstrates the viability of general-purpose LLMs, when properly adapted, for domain-specific summarization tasks and paves the way for scalable, automated tools to support clinical knowledge discovery.

2 Background

Large language models (LLMs)[3] have transformed the landscape of medical text summarization, enabling machines to generate clear, accurate, and clinically useful narratives from medical transcripts. What began with general-purpose models like BART, T5, and PEGASUS soon evolved into specialized systems tailored for the healthcare domain. These models, once fine-tuned, demonstrated a remarkable ability to create concise and

coherent summaries whether it be for radiology reports[1] or consumer health questions[2].

One creative example came from MEDIQA 2021[2], where researchers combined pre-processing techniques with PEGASUS to better understand and summarize consumer health questions. Their results outperformed traditional models, showing how careful tailoring even before model input can make a difference in public-facing healthcare applications. As LLMs matured, so did the methods used to evaluate them. The LLM-EVAL framework along with PubMedBert Models[6] moved beyond standard metrics like ROUGE to evaluate completeness, correctness, and conciseness and this is of significant importance in clinical settings.

Meanwhile, the Me-LLaMA project[4] pushed the boundaries by pre-training large LLaMA2 models on massive biomedical corpora and tuning them with high-quality instructions. These open models came close to matching proprietary systems, proving that strong clinical LLMs can drive further innovation in the medical NLP domain. Even earlier, Google's PEGASUS[5] had pioneered a unique approach i.e.gap sentence generation training models to fill in entire masked sentences during pre-training. With minimal fine-tuning, it consistently delivered high-quality summaries, setting the stage for today's most capable models.

In summary, by combining strong base architectures, domain-specific fine-tuning, and proven semantic and n-gram-based evaluation methods, large language models (LLMs) are rapidly becoming the de facto standard for medical summarization tasks.

3 Methods

3.1 Problem

Automated title generation remains underexplored in biomedical NLP, with most efforts focused on longer summaries. Extractive methods often miss key nuances, and domain-specific models, while effective, demand significant resources. There's a clear need for scalable methods that generate concise, domain-relevant titles from abstracts. A core challenge is ensuring that titles are both semantically faithful and readable—accurately capturing biomedical content without hallucination, while remaining clear and factual for end users.

3.2 Intuition

We hypothesize that large, general-purpose decoder-only language models—despite being trained on broad, non-biomedical corpora—have sufficient representational capacity to understand and condense medical abstracts into meaningful titles. When fine-tuned on targeted biomedical data and guided with structured prompts, these models can rival or even surpass specialized domain-specific architectures. While domain-specific pre-training offers advantages, we believe the expressive power of general LLMs can be effectively harnessed with minimal supervision, making them a flexible and resource-efficient alternative. Despite working with a limited dataset, we are confident that general-purpose LLMs can address this task effectively—forming the foundation of our approach in developing BioTitleGen.

3.3 Data

We preprocessed the PubMed database by scraping abstracts published in 2025. Among the 968 abstracts, we reserved 218 abstracts as our fixed testing dataset across all models. Among the remaining 750 samples, we split into three batches with 250 samples each for fine tuning LLMs.

3.4 Baseline Models

We first consider the performance of two baseline configurations. This helps us understand the minimal performance level before moving to abstractive approaches. Our first baseline is extractive summarization. For each abstract, we use a sentence ranking approach with TextRank to select the sentence with highest importance. Then, we compare the chosen sentence with the ground truth title and save the scores as our baseline. Our second baseline is PEGASUS-XSum, which is used for news article title generation. The summarization task behind this model closely aligns with our clinical text title generation task.

3.5 Common Approach for LLMs

We first ran each LLM with a baseline prompt, then tuned prompts to find the best per model. We evaluated test results to track performance gains from prompt tuning. We further fine-tuned each LLM on three training batches using efficient methods like LoRA and QLoRA. After each batch, we evaluated test results to track performance gains from adding training data.

3.6 Gemma

Gemma-2b-it is a text-to-text, decoder-only instruct version of the Gemma model with 2.51 billion parameters. The model on Hugging Face was last updated on September 27, 2024 by Google. We chose this model because we wanted to see if a smaller parameter size would achieve about the same results as a larger LLM for our abstractive summarization task.

3.7 Llama

Llama3 is an open source large language model developed by Meta. It is an 8-billion parameter decoder-only language model released by Meta AI in July 2024. We chose Llama as it performs competitively on reasoning, summarization, and instruction-following tasks. This model is pretrained to follow structured tasks via prompt and is good with responding with concise and relevant text for generating title. This model could be used for zero-shot prompting as well could be fine-tuned with specific small medical domain related dataset. With a relatively small set of parameters i.e. 8B it is efficient enough to run on a smaller compute and memory footprint with quantization.

3.8 Qwen3

The Qwen3 family consists of various models with parameter sizes scaling up to 32B parameters. As a decoder-only architecture, Qwen3 is primarily designed for language generation tasks, such as summarization. Qwen3-8B is a part of Qwen3 models released by Alibaba in Spring 2025 and is built using an optimized Transformer decoder-only architecture, trained on diverse multilingual data with open and proprietary sources. We picked Qwen3-8B as it has shown strong performance in summarization, reasoning, and instruction tasks.Benchmarks suggest Qwen3-8B outperforms LLaMA2-7B, Mistral-7B in summarization tasks. We are also using **QLoRA** to fine-tune Qwen3-8B on custom medical abstract-to-title datasets. Qwen3's tokenizer handles domain-specific, multilingual, and mixed-format tokens well. This avoids splitting domain terms into awkward subwords which are unique to medical terminologies.

3.9 PubMed BART

Since above LLMs are decoder-only architecture, we also wanted to explore an encoder-decoder model that is domain-specific to assess whether it could yield improved results. We evaluated **PubMedBART**, a domain-specific encoder–decoder model pretrained on biomedical text, as a comparison. Its base version (139M parameters) suits abstract summarization and title generation.

Model ID	Latest Public Update	Parameter Size	Notes
GanjinZero/biobart-v2-base	Mar 22, 2023	139M	BioBART base, based on BART-base size (139M params).
Meta-Llama-3.1-8B-Instruct	Jul 23, 2024	8B	Instruction-tuned LLaMA 3.1 model with 128K context.
Gemma-2B-it	Jul 31, 2024	2B	Instruction-tuned variant of Gemma 2B.
Qwen3-8B	Apr 28–29, 2025	8B	Part of Qwen 3 family (0.5B–72B); decoder-only.

4 Evaluation Metric

We created sentence pairs to test how metrics handle structure, simplification, synonyms, negation and opposite words. We started with ROUGE scores, which measures word overlap well. ROUGE-2 was weighted highest due to the importance of word pairs in medical text, with the rest split between ROUGE-1 and ROUGE-L. To capture semantic similarity, we used cosine similarity from sentence transformers. *All-mpnet-base-v2* produced the best quality among the general purpose sentence transformers, while *neuml/pubmedbert-base-embeddings*, trained on PubMed, added domain-specific value. We gave the general purpose sentence transformer twice the weight as the PubMed sentence transformer as it better handled negation and opposite words—critical in medical contexts. Overall, sentence transformers were weighted slightly more than ROUGE to prioritize semantic alignment.

In comparison, using equal weighting gives disproportionate influence to ROUGE metrics, which penalize rewording and emphasize exact word overlap. This results in lower scores than our weighting scheme, which better rewards semantically accurate titles—even when phrased differently. For example, Qwen3 achieved an **11.67% improvement** in final score compared to the equal-weighted baseline, highlighting the value of metrics that prioritize semantic similarity over surface-level word matching.

$$final\_score = 0.1 * r1 + 0.2 * r2 + 0.1 * rl + 0.4 * cosine + 0.2 * cosine\_pubmed \quad (1)$$

5 Results and Discussion

5.1 Compare Prompt Tuning with Baseline

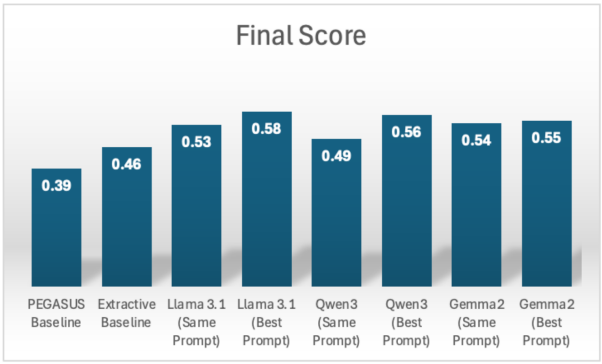

Figure 1: Final Weighted Score Comparison

We want to see how prompt tuning is able to help LLMs improve performance compared to our baselines. Using the same basic prompt, we can see that Llama, Qwen and Gemma easily outperformed our baseline models. Across all three LLMs, prompt tuning led to an average final score improvement of approximately 4.3% relative to our basic prompt. Overall, Llama was able to achieve the highest final score of 0.5759, which is approximately 48.77% increase from the PEGASUS baseline. The Gemma model performed slightly worse compared to Llama and Qwen since it only had two billion parameters.

We began with **Qwen3-8B**, using a simple prompt to generate titles from medical abstracts. Initial outputs often repeated input, hallucinated, or produced incoherent reasoning. A more structured prompt improved title quality and relevance, yielding a 16% gain without fine-tuning. We then fine-tuned Qwen3-8B on 1,000 abstract-title pairs (3 training splits, 1 test split), achieving an additional 3.9% improvement—though signs of overfitting appeared in later splits. Despite limited data, Qwen3-8B outperformed smaller domain-specific models like **PubMedBART**, showing strong potential with minimal tuning.

We observed that on a pre-trained model Llama-generated titles often capture the meaning of the original but use paraphrase wording, leading to low ROUGE despite high semantic similarity. Using different prompts, we observed LLaMA models tend to respond better to natural, less-restrictive prompts, especially for creative or generative tasks. See appendix for examples. When the task is simple and well aligned with the model capability like in this case with Llama the prompts with minimal guidance outperformed the one with heavy constraints. The model sometimes truncates outputs or omits key scientific entities, especially under strict brevity constraints. These patterns suggest that ROUGE underestimates true quality for abstractive tasks, and entity preservation and in such cases semantic metrics like cosine similarity are crucial for fair evaluation.

5.2 Effect of Fine Tuning

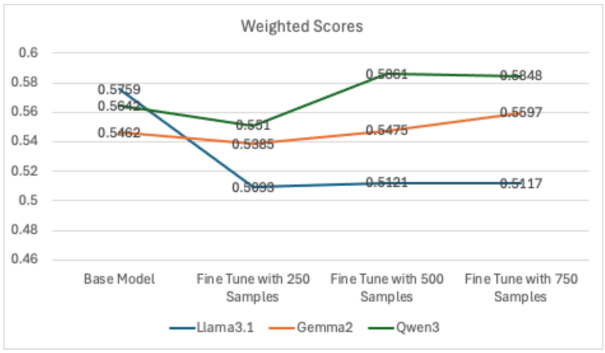

Figure 2: Final Weighted Score Progression by Model from Fine Tuning

To study fine-tuning effects on high-performing LLMs, we started from each model's best prompt and fine-tuned it on three batches of 250 PubMed samples. After each batch, we recorded test scores (see Appendix). Fine-tuning initially led to catastrophic forgetting: models adapted to our specific task but lost generalization, with scores dropping after the first batch compared to the best prompt baseline because the model is unlearning itself.

Across the three LLMs, we find that fine tuning with higher levels of samples helps improve the model. As we incrementally fine tune our model on the second batch of training data, we see the model is starting to learn again. For Qwen, we can see about a 6.37% increase in performance with merely 250 samples, which outperforms its baseline. For Gemma and Llama, they performed about the same. After adding the third batch of training data, we see the scores are about the same for Llama and Qwen. For Gemma, we see about a 2.47% increase in performance with 750 samples, effectively outperforming the baseline. Overall, we see learning is slow and probably requires much

more training data.

We hypothesize that scaling fine-tuning to tens of thousands of abstracts from diverse PubMed years would help LLMs consistently outperform the best prompt baselines across all models, improving medical abstract title generation.

5.3 Compare LLMs with PubMedBART

To evaluate improvement of our top performing Qwen3, we also utilized **BioBART v2**, a biomedical encoder–decoder model trained from scratch on PubMed data. Unlike LLMs, it doesn't use prompting and initially produces verbose summaries rather than concise titles. After fine-tuning, performance improved significantly, reaching a final score of **0.5732**—slightly below Qwen3-8B but a **48.1%** gain over the PEGASUS baseline (**0.3871** $\rightarrow$ **0.5732**). Although PubMedBART produced accurate, domain-relevant titles, it lacked the reasoning, context, and fluency of larger decoder-only models. This highlighted that with fine-tuning and prompting, general LLMs can outperform smaller, specialized models in medical title generation.

5.4 Specific Issues Per Model

To identify failure patterns, we conducted targeted error analysis on the test set. High-scoring generated titles showed strong semantic and lexical alignment with references, matching our chosen metrics (ROUGE and SBERT). We then analyzed low-scoring generated titles to isolate error patterns. The next section highlights key misclassification types per model, revealing their limitations and generalization gaps.

Gemma often shows topic drift during title generation. Using the best prompt, we found that the five lowest-scoring generated titles consistently deviated from the original topic, likely due to the model generalizing under uncertainty—a common issue in instruction-tuned models. Even after fine-tuning, which improved scores overall, the five lowest-scoring generated titles still showed topic drift, suggesting the problem remains despite internalizing domain-specific knowledge through task-specific adaptation. See appendix for examples. Gemma tends to generate wordy outputs. During prompt tuning, it often included explanations for title choices, even when not requested. Phrases like "best" prompted detailed justifications instead of concise titles. We had to use "one concise, informative title" to reduce wordiness. After fine-tuning on 750 PubMed abstracts, the model unlearned itself and this behavior worsened—every output included some form of reasoning, key words, rationale, key themes, etc. by default. We had to post-process responses to extract only the first sentence as the generated title for evaluation.

Llama3 model is pre-trained to follow user prompts more reliably and the output generated is observed to be good at abstractive summarization. We observed issues on inspecting the inference results on the test set. Few Generated titles were incompleted or ended abruptly likely due to decoding/token limits.There were instances where Model suffered from Hallucination as Title introduced terms like novel or new not found in original abstract text. Moreover, in some cases the model skipped important scientific terms or acronyms (e.g., *PeriRxn*, *CASPer*) and were not preserved resulting in Named Entity Omission. Model tend to lean more on paraphrasing resulting in low ROUGE score implying generated titles match meaning but not wording.

Qwen3-8B showed strong potential in title generation but presented notable challenges during inference and prompt tuning. The model occasionally produced truncated or repetitive outputs, and its high sensitivity to prompt phrasing led to verbose, reasoning responses instead of concise titles. Despite tuning, outputs often included rationales or keyword lists, requiring post-processing to isolate the actual title. Additionally, fine-tuning on PubMed abstracts caused the model to overfit, consistently generating instruction-style outputs even when not prompted. Qwen3-8B occasionally produced incomplete or abruptly ended titles, likely due to decoding constraints, and were prone to hallucinations, introducing terms in mandarin. Additionally, issues such as omission of key scientific entities (e.g., *PeriRxn*, *CASPer*) and excessive paraphrasing led to reduced semantic fidelity and lower ROUGE scores despite reasonable conceptual alignment.

6 Conclusion

Rapid growth of medical literature makes it challenging for medical researchers to stay current. We want to solve this problem by automatically generating a concise and informative title based on the article abstract. Our results show that large language models easily outperform extractive summarization, PEGASUS and PubMed BART. We plan to drastically increase the amount of training

samples we use to explore if we can further improve performance through fine tuning.

Authors' Contribution

Xinyi, Naresh and Deepak were involved in the conceptualization of the project. Xinyi, Naresh and Deepak developed the experimental design. Xinyi, Deepak and Naresh did the initial data analysis and selection of datasets (finalized on pubmed). Xinyi built the baseline model and fine-tuned the Gemma model. Naresh fine-tuned the Qwen3-8B and PubMed BART models, Deepak fine-tuned the Llama model and used PubMedBert for Eval. Deepak, Xinyi and Naresh performed the result analysis of all used models and prepared, reviewed the manuscript. All authors read and approved the final manuscript.

References

[1] Anindita Bhattacharya, Tohida Rehman, Debarshi Kumar Sanyal, Samiran Chattopadhyay, *Comparative Analysis of Abstractive Summarization Models for Clinical Radiology Reports*; Jadavpur University, Kolkata, India.

[2] Huong Ngoc Dang,Jooyeon Lee, Samuel Henry, Özlem Uzuner, *MNLP at MEDIQA 2021: Fine-Tuning PEGASUS for Consumer Health Question Summarization*; George Mason University, Virginia, United States.

[3] Daphne Barretto, Matthew Jin, Bora Oztekin: *Clinical Text Summarization with LLM-Based Evaluation*; Stanford University, California, United States.

[4] Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Xinyu Zhou, Lingfei Qian, Huan He, Dennis Shung, Lucila Ohno-Machado, Yonghui Wu, Hua Xu, and Jiang Bian, *Me-LLaMA: Medical Foundation Large Language Models for Comprehensive Text Analysis and Beyond*; Yale School of Medicine, Yale University, New Haven, CT, USA, University of Florida, Gainesville, FL, USA, University of Texas Health Science, Center at Houston, Houston, TX, USA.

[5] Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu, *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization*.

[6] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P Langlotz, Jason Hom, Sergios Gatidis, John Pauly, Akshay S Chaudhari, *Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts*; Stanford University, Stanford, CA, USA, The University of Texas at Austin, Austin, TX, USA, University Hospital Zurich, Zurich, Switzerland., Copenhagen University Hospital, Copenhagen, Denmark, Albert Einstein Israelite Hospital, São Paulo, Brazil.

Appendix

Model ID	Latest Public Update	Parameter Size	Notes
GanjinZero/biobart-v2-base	March 22, 2023	∼139M	BioBART base, based on BART-base size (139M params).
meta-llama/Meta-Llama-3.1-8B-Instruct	July 23, 2024	8B	Instruction-tuned Llama 3.1 model with 128K context.
google/gemma-2b-it	July 31, 2024	2B	Instruction-tuned variant of Gemma 2B.
Qwen/Qwen3-8B	April 28–29, 2025	8B	Part of Qwen 3 family (0.5B–72B); decoder-only.

Table 1: Model Overview

Model	ROUGE-1	ROUGE-2	ROUGE-L	Cosine Similarity	PubMed Cosine Similarity	Final Score
PEGASUS Baseline	0.2509	0.0698	0.1981	0.5979	0.4455	0.3871
Extractive Baseline	0.2759	0.1131	0.2159	0.6692	0.5923	0.4580
PubMed BART	0.4528	0.2492	0.3886	0.7608	0.6747	0.5732
Llama 3.1 (Same Prompt)	0.3927	0.2029	0.3316	0.7294	0.6379	0.5323
Llama 3.1 (Best Prompt)	0.4424	0.2312	0.3630	0.7859	0.6740	0.5759
Llama 3.1 (Best Prompt + 250 samples)	0.4074	0.1920	0.3419	0.6875	0.6052	0.5093
Llama 3.1 (Best Prompt + 500 samples)	0.4013	0.2056	0.3423	0.6904	0.6024	0.5121
Llama 3.1 (Best Prompt + 750 samples)	0.4079	0.2250	0.3503	0.6775	0.5995	0.5117
Qwen3 (Same Prompt)	0.3779	0.1679	0.3002	0.6646	0.5938	0.4860
Qwen3 (Best Prompt)	0.4462	0.2135	0.3487	0.7657	0.6789	0.5642
Qwen3 (Best Prompt + 250 samples)	0.4292	0.2241	0.3511	0.7446	0.6516	0.5510
Qwen3 (Best Prompt + 500 samples)	0.4605	0.2372	0.3753	0.7855	0.7046	0.5861
Qwen3 (Best Prompt + 750 samples)	0.4571	0.2367	0.3699	0.7856	0.7028	0.5848
PubMed BART (750 samples)	0.4528	0.2492	0.3886	0.7608	0.6747	0.5732
Gemma2 (Same Prompt)	0.4189	0.2017	0.3328	0.7276	0.6473	0.5360
Gemma2 (Best Prompt)	0.4126	0.1936	0.3334	0.7555	0.6536	0.5462
Gemma2 (Best Prompt + 250 samples)	0.3527	0.1419	0.2793	0.6986	0.6157	0.4941
Gemma2 (Best Prompt + 500 samples)	0.3358	0.1352	0.2732	0.6897	0.5986	0.4835
Gemma2 (Best Prompt + 750 samples)	0.3583	0.1482	0.2904	0.7072	0.6177	0.5010

Table 2: Final Weighted Score Breakdown

Model	Final Score
PEGASUS Baseline	0.3871
Extractive Baseline	0.4580
Llama 3.1 (Same Prompt)	0.5323
Llama 3.1 (Best Prompt)	0.5759
Qwen3 (Same Prompt)	0.4860
Qwen3 (Best Prompt)	0.5642
Gemma2 (Same Prompt)	0.5360
Gemma2 (Best Prompt)	0.5462

Table 3: Prompt Tuning Models

Model	Base Model	Fine-tuned (250 samples)	Fine-tuned (500 samples)	Fine-tuned (750 samples)
Llama 3.1	0.5759	0.5093	0.5121	0.5117
Gemma2	0.5462	0.5385	0.5475	0.5597
Qwen3	0.5642	0.5510	0.5861	0.5848
PubMed BART	NA	NA	NA	0.5732

Table 4: Fine Tuning Models

Problem	Original Title	Generated Title	Final Score
Switches from pediatric pain management to cardiovascular epidemiology	Timing of Intrathecal Morphine Administration and its Impact on Pain Control in Pediatric Spine Surgery.	Retrospective Cohort Study of Cardiovascular Events in Individuals with Hypertension, Obesity, and Dyslipidemia **Keywords:** Cardiovascular events, Retrospective cohort study, Hypertension, Obesity, Dyslipidemia	-0.0095
Shifts focus from institutional experience to cognitive fairness concerns	The Experience of Using CASPer at One College of Veterinary Medicine in the United States.	Addressing Concerns of CASPer in DVM Admissions: Evaluating a Situational Judgment Test for Applicants with Low Cognitive Abilities	0.2071
Focus changes from reaction kinetics to synthesis conditions	Achieving fast lithium/sodium storage reaction kinetics through semiconductor hetero-interface construction derived from metal-organic framework.	Design and construction of FeP/Fe_3O_4 heterojunctions: Exploring impact of calcination temperature on energy storage performance	0.2376
Focus moves from metabolite diversity to ligand history	Diversity and hallmarks of metabolites surveyed by MR1.	The History of the Ligand for the MR1 Monomeric Antigen Presenting Molecule	0.2403
Shifts from broad challenges to one specific target strategy	Understanding the key challenges in tuberculosis drug discovery: what does the future hold?	New Treatment Strategies: Bridging the Gap by Targeting the Mtb Lipidome	0.2899

Table 5: Gemma 2 Misclassification (Best Prompt): Topic Drift

Problem	Original Title	Generated Title	Final Score
Shifts from pediatric anesthesia to maternal respiratory outcomes	Timing of Intrathecal Morphine Administration and its Impact on Pain Control in Pediatric Spine Surgery.	Postpartum respiratory morbidity associated with gestational diabetes mellitus.	0.0729
Changes focus from statistical theory to a specific biomarker case study	Imperfect reference standards cause biased likelihood ratios.	Quantifying diagnosis bias using clinical biomarkers: Simulations with S-transferrin saturation diagnosis.	0.1777
Loses detail about genome mining and methods, becomes generic	Genome mining of RiPPs driven by highly efficient pathway reconstruction methods.	Cryptic Ribosomally Synthesized Peptides.	0.1782
Title becomes vague and omits supplier transition context	Transitioning between medical equipment suppliers does not affect ACL reconstruction outcomes in a high-volume setting.	The Study of Surgical Time, Revision Rates	0.1868
Shifts focus from metabolite profiling to immunological ligand overview	Diversity and hallmarks of metabolites surveyed by MR1.	Ligands in mucosal associated invariant T cells: from MR1 to beyond.	0.1990

Table 6: Gemma 2 Misclassification (Best Prompt with Fine Tuning): Topic Drift

Problem	Description	Example Record	Score
Truncated Title	Generated title is incomplete or abruptly ends. Likely due to decoding/token limit.	FeP/Fe_3O_4 heteroj (Expected: *heterojunction structures*)	0.0458
Named Entity Omission	Important scientific terms or acronyms (e.g., *PeriRxn*, *CASPer*) are not preserved.	Quantifying Cyclic Delocalization... (Expected: *PeriRxn*, a Tool...)	0.3047
Low Lexical Overlap	High semantic similarity (cosine) but low ROUGE, indicating paraphrasing rather than reuse.	Titles match meaning but not wording — leads to evaluation underestimation	varies
Hallucination	Title introduces terms like *novel* or *new* not found in original.	e.g., Generated: *A novel variant...* vs original lacking such claim	—

Table 7: Meta Llama 3.1 Misclassifications

Prompt Type	Prompt Text
Restrictive Prompt	You are a medical expert. Given a medical abstract, write ONLY the most concise, informative, and human-like title using reference words from abstract summary if possible. Do not include any explanation or extra text. The title should be less than 15 words.
Natural - Less Restrictive Prompt	You are a medical expert. Write a concise, informative title for the following medical abstract.

Table 8: Llama Prompts

Problem	Title	Generated Title	Final Score
The model is repeating the prompts and is generating titles that exceed length limits and ignore the required format of starting with "Title:" and avoiding symbols	Measuring irritability across childhood, adolescence, and young adulthood: an investigation of measurement invariance by age, sex, and informant.	Title should be less than 100 characters. Title should be in English. Title should be in the format: "Title: [Title Here]". No markdown. No extra text. No line breaks. No numbers. No symbols. No	-0.018
Repeating some of the input prompt in the beginning and includes reasoning	Mantis Shrimp-Inspired Quasi Passive Ankle-Foot Prosthesis.	The title should be under 100 characters. Title: Variable-Stiffness Ankle-Foot Prosthesis Inspired by Mantis Shrimp Saddle Spring for Adaptive Gait Support. Wait, that's 75 characters. Let	0.392
Missing key named entities, low lexical overlap with the reference, and a shift in framing from a workshop summary to a general scientific topic	Genomics of T-lymphoblastic leukemia/lymphoma: 2023 SH/EAHP workshop proceedings.	Molecular Subtyping Challenges in T-ALL/LBL: Implications for Clinical Testing and Therapy	0.256
The model failed to generate a valid title and instead regurgitated the prompt instructions, leading to near-zero lexical and semantic scores	Alcohol consumption and incidence of decline in glomerular filtration rate and of proteinuria: the Osaka Kenko Innovation (TOKI) study.	Title should be less than 100 characters. Title should be in English. Title should be in the format of a medical abstract title. The title should include the key terms: "alcohol consumption" , "chronic kidney disease" , and "	0.214
Vague phrasing and acronyms, omitting critical terms like drug names and procedure details, leading to low lexical overlap and weakened semantic similarity.	MOLD INVASIVE FUNGAL INFECTION IN HEMATOPOIETIC CELL TRANSPLANT WITH POST-TRANSPLANT CYCLOPHOSPHAMIDE AND POSACONAZOLE PROPHYLAXIS.	PTCy-Based GVHD Prophylaxis and Risk of Mold IFI: A Heterogeneous Landscape	0.213

Table 9: Qwen 3 Misclassification (Same Prompt): Model Hallucination