BDA-UC3M @ BioLaySumm: Efficient Lay Summarization with Small-Scale SoTA LLMs

Ilyass Ramzi[1] **and Isabel Segura Bedmar**[2]
[1]Graduate School of Engineering and Basic Sciences, Universidad Carlos III de Madrid
[2]Computer Science and Engineering Department, Universidad Carlos III de Madrid
100510978@alumnos.uc3m.es, isegura@inf.uc3m.es

Abstract

The growing need to make biomedical research accessible to non-expert audiences has motivated the development of effective lay summarization systems. While large language models (LLMs) have set recent benchmarks, their computational demands limit widespread adoption. This paper explores the use of small-scale, state-of-the-art LLMs (4B–7B parameters) for biomedical lay summarization in the BioLaySumm 2025 shared task. Leveraging dynamic 4-bit quantization, extractive preprocessing, prompt engineering, data augmentation, and Direct Preference Optimization, our system, based on Gemma3 4B, Qwen3 4B, and GPT-4.1-mini, ranked second in its category, showing that compact models can deliver high-quality, factually accurate summaries.

1 Introduction

Recent advances in large language models (LLMs) have demonstrated exceptional performance in generating lay summaries of biomedical literature, supporting the critical goal of making complex scientific content accessible to non-expert audiences (Goldsack et al., 2024, 2023). However, most state-of-the-art approaches rely on very large models—often with tens of billions of parameters—posing significant barriers for practical deployment and reproducibility due to hardware and computational requirements.

The BioLaySumm 2025 Shared Task challenges participants to develop automated systems for lay summarization of biomedical research articles, with evaluation based on relevance, readability, and factuality across established benchmark datasets (eLife and PLOS) (Xiao et al., 2025; Goldsack et al., 2022). The official baselines for this task, such as Llama3 8B and Qwen2.5 7B, set a high standard for both scale and performance.

This paper presents the approach developed by BDA-UC3M for BioLaySumm 2025, aiming to demonstrate that small-scale, state-of-the-art LLMs (4B–7B parameters), when carefully optimized and fine-tuned, can achieve competitive—and in some cases, superior—performance to larger baselines. Our system leverages recent advances in LLM efficiency, notably:

- **Parameter-efficient models and training:** Utilizing compact LLMs, including Gemma3 4B (Dynamic 4-bit Instruct) (Team, 2025), Qwen3 4B (Dynamic 4-bit Safetensor, fine-tuned) (Yang et al., 2025), and GPT-4.1-mini (via OpenAI API), all selected for their strong performance-to-size ratio.

- **Accessible compute and deployment:** All model training and inference is performed on consumer-grade GPUs[1], with deployment streamlined using RunPod pods[2].

- **Advanced pipeline building on prior SoTA:** Our approach systematically integrates and improves strategies from the top BioLaySumm 2024 systems (You et al., 2024; Zhao et al., 2024; Kim et al., 2024)—combining robust extract-then-summarize frameworks, advanced prompt engineering, targeted data augmentation, and factuality-aware fine-tuning (DPO).

While retrieval-augmented generation (RAG) has shown promise in biomedical summarization by enriching model inputs with external knowledge sources such as Wikipedia, this work does not incorporate RAG due to timeline constraints. Future iterations may revisit lightweight retrieval solutions for even greater factuality.

[1]https://docs.unsloth.ai/
[2]https://docs.runpod.io/pods/overview

BioNLP 2025 Shared Tasks, pages 249–255

2 Methods

2.1 Datasets

We used the official BioLaySumm 2025 task datasets: **eLife** and **PLOS** (Xiao et al., 2025; Goldsack et al., 2022). Table 1 summarizes the dataset splits:

Dataset	Train	Validation	Test
eLife	4346	241	142
PLOS	24773	1376	142

Table 1: Dataset splits for BioLaySumm 2025.

Tokenization (using `cl100k_base`) revealed substantial variation in article lengths, consistent with previous dataset reports (Goldsack et al., 2022):

- **eLife**: Train articles averaged 14,140 tokens (max 46,150), summaries averaged 428 tokens.
- **PLOS**: Train articles averaged 8,925 tokens (max 32,623), summaries averaged 233 tokens.

Test sets do not include reference summaries. eLife summaries are typically longer and more abstracted, while PLOS summaries are shorter and more closely tied to the article content (Luo et al., 2022).

2.2 Preprocessing

TextRank Extraction. To efficiently compress long articles and highlight salient content, we used a custom TextRank implementation (adapted from methods described in (You et al., 2024)):

- Articles were segmented into sentences using spaCy (`en_core_web_sm`), with only sentences over 20 characters retained.
- TF-IDF vectors and cosine similarity were used to construct a similarity matrix.
- Sentences were ranked with PageRank over the similarity graph, and the top N ($N = 50$) were selected.

Chunking + TextRank. For models with smaller context windows (notably Qwen3 4B, 32K tokens (Yang et al., 2025)), we applied chunking:

- Articles were split into chunks of ∼40 sentences.
- TextRank was run independently within each chunk, and the top sentences (e.g., 10 per chunk) were extracted.
- If the total number of selected sentences across all chunks exceeded the limit, we applied a global re-ranking step: all previously selected sentences were pooled and TextRank was run again on this subset to select the final top 50, ensuring the most salient content across the full article was retained. This step was used when the combined top sentences from all chunks could not fit in the model's input context.

For GPT-4.1-mini and Gemma3 4B (Team, 2025), chunking and re-ranking were not required due to their larger context capabilities.

Data Augmentation. Training diversity was enhanced by using GPT-4.1-mini to paraphrase and augment summaries, especially where extractive strategies omitted lay-relevant detail. This data augmentation step follows insights from previous top systems (Zhao et al., 2024).

2.3 Model Setup and Training

We focused on small yet state-of-the-art LLMs for efficiency and reproducibility:

- **Gemma3 4B** (Dynamic 4-bit Instruct)
- **Qwen3 4B** (Dynamic 4-bit Safetensor)
- **GPT-4.1-mini** (via OpenAI API)

Fine-tuning and inference for Gemma3 and Qwen3 models leveraged the Unsloth framework, which combines two key techniques for maximal efficiency:

- **Dynamic 4-bit quantization** reduces memory usage by compressing model weights to 4 bits on-the-fly, enabling large LLMs to run on consumer-grade GPUs (Han et al., 2024).
- **LoRA (Low-Rank Adaptation)** introduces lightweight, trainable adapter layers, allowing only a small subset of parameters to be fine-tuned while the core model weights remain frozen (Hu et al., 2021).

Together, these methods allowed efficient training and adaptation of large models on standard hardware (RTX 3090, 24GB VRAM). For comparison, GGUF format is intended only for inference.

2.4 Fine-Tuning and Hyperparameters

- **Gemma3 4B:** Fine-tuned with Unsloth using LoRA adapters (Hu et al., 2021) and default settings: temperature=1.0, top_k=64, top_p=0.95. Training used per-device batch size 2, gradient accumulation 4, max_steps 30, learning rate 2×10^{-4}, weight decay 0.01, AdamW 8-bit optimizer.

- **Qwen3 4B:** Followed Unsloth's effective setup: rank=32, lora_alpha=32, dropout=0, "unsloth" gradient checkpointing. Training used the same batch, learning rate, and optimizer setup as above, with memory optimized for 32K context (Yang et al., 2025).

- **GPT-4.1-mini:** Utilized OpenAI API with recommended temperature and top_p settings. SFT used standard instruction-following templates; context window up to 32K tokens. Prompt design followed OpenAI's best practices[3].

2.5 Prompt Engineering

We systematically developed and tested a suite of prompts, evaluating both zero-shot and few-shot settings as well as dataset-specific refinements. Our approach was influenced by prior competition leaders (You et al., 2024; Zhao et al., 2024; Kim et al., 2024) and included:

- **Baseline Prompts (V1):** Focused on clarity and accessibility for lay readers.

- **Structured/Prescriptive Prompts (V2):** Provided numbered guidelines for better output organization.

- **Competition-Optimized Prompts (V3):** Explicitly referenced BioLaySumm metrics (ROUGE, BLEU, METEOR, BERTScore, LENS, AlignScore, SummaC, FKGL, CLI, DCRS) (Xiao et al., 2025), instructing models to optimize relevance, readability, and factuality.

- **Refined Prompts (V4):** Further emphasized factuality, accuracy, and discouraged speculative language or fabricated author names.

Model-Specific Prompts:

- *Qwen3 4B and Gemma3 4B:* Used instruction-tuned prompts with explicit, structured guidance for one-paragraph, factually accurate lay summaries (Yang et al., 2025; Team, 2025).

- *GPT-4.1-mini:* Incorporated OpenAI's prompt engineering best practices (OpenAI, 2025b), with iterative refinements based on validation.

Prompt selection was finalized for each dataset and model through ablation, guided by the best combination of metric performance and qualitative validation. **For full prompt templates, refer to Appendix A.**

3 Results and Analysis

3.1 Main Results

Table 2 reports the primary evaluation metrics for our three models—GPT-4.1-mini, Gemma3 4B, and Qwen3 4B—on both the eLife and PLOS test sets. Each metric is averaged per dataset, followed by the overall average across datasets.

All three models performed closely, with GPT-4.1-mini slightly outperforming on relevance and semantic similarity, while Qwen3 4B showed a small edge on factuality metrics (AlignScore, SummaC) (You et al., 2024; Kim et al., 2024; Zhao et al., 2024; Team, 2025; Yang et al., 2025).

3.2 Ablation and Component Analysis

We performed ablation studies to analyze the effect of prompt style, DPO training (Kim et al., 2024), and extractive chunking (You et al., 2024).

Prompt Style:

- GPT-4.1-mini achieved best results with a general, clarity-focused prompt.

- Gemma3 4B benefited from refined, constraint-driven prompts.

- Qwen3 4B excelled with explicit, stepwise prompts.

DPO: Direct Preference Optimization (DPO) improved factuality and readability metrics (AlignScore, SummaC, FKGL, DCRS) (Kim et al., 2024), but slightly reduced ROUGE/BLEU due to prioritizing factual alignment over surface-level overlap.

Chunking/Extraction: Chunking was crucial for Qwen3 4B due to its limited context window (Yang et al., 2025), ensuring representation across all article sections.

[3] https://cookbook.openai.com/examples/gpt4-1_prompting_guide

Model	Dataset	ROUGE	BLEU	METEOR	BERTS	FKGL	DCRS	CLI	LENS	Align	SummaC
GPT-4.1-mini	eLife	0.371	8.07	0.298	0.869	9.94	8.28	11.47	70.70	0.619	0.545
	PLOS	0.335	8.08	0.290	0.870	14.71	10.24	14.87	57.49	0.764	0.533
	Avg	**0.353**	**8.08**	**0.294**	0.870	**12.32**	**9.26**	**13.17**	**64.10**	0.691	0.539
Gemma3 4B	eLife	0.370	7.57	0.297	0.869	9.97	8.30	11.60	69.54	0.618	0.555
	PLOS	0.335	7.77	0.284	0.871	14.73	10.39	15.05	56.96	0.767	0.526
	Avg	0.352	7.67	0.290	**0.870**	12.35	9.35	13.32	63.25	0.693	0.541
Qwen3 4B	eLife	0.367	7.16	0.287	0.869	10.32	8.52	11.89	69.41	0.631	0.558
	PLOS	0.334	8.01	0.288	0.871	14.80	10.41	15.10	57.07	0.774	0.530
	Avg	0.351	7.59	0.288	0.870	12.56	9.47	13.49	63.24	**0.702**	**0.544**

Table 2: Performance of our models on eLife and PLOS test sets for BioLaySumm 2025. FKGL, DCRS, and CLI: lower is better (readability). All other metrics: higher is better. For BERTScore, values are rounded to three decimals; Gemma3 4B achieved the highest score at full precision.

3.3 Discussion of Findings

Our experiments confirm that carefully optimized small-scale LLMs (<7B parameters) can approach the performance of much larger models in biomedical lay summarization (Team, 2025; Yang et al., 2025; Xiao et al., 2025). While none of our models surpassed last year's BART/LED-based systems in extractive metrics such as ROUGE and BLEU (You et al., 2024; Goldsack et al., 2024), all achieved high semantic similarity and factuality, with Gemma3 4B posting the highest BERTScore among our submissions.

Ablation studies highlighted that prompt engineering and DPO training have strong, model-specific impacts, introducing a clear trade-off: optimizing for factuality and readability can reduce surface-level overlap with reference summaries, and vice versa (Kim et al., 2024; Zhao et al., 2024). Chunking strategies for models with limited context windows (e.g., Qwen3 4B) proved essential for consistent performance. As the datasets were unchanged year-on-year (Goldsack et al., 2022), our results indicate that further gains with small LLMs may require new architectures or additional external knowledge integration.

4 Conclusion

This work shows that well-optimized, small-scale LLMs can produce high-quality biomedical lay summaries, rivaling larger models in semantic and factual metrics while remaining accessible for training on standard hardware.

Limitations

Despite these strengths, certain limitations remain:

- **Performance Gap to Large Models:** Despite competitive scores, small LLMs still lag behind last year's best large-scale (BART/LED) and generative models on overlap-based metrics (ROUGE, BLEU), which likely benefit from larger pretraining corpora and parameter capacity.
- **Resource and Timeline Constraints:** All training was performed on single consumer GPUs, restricting the scope of hyperparameter search, ablation, and deeper multi-stage fine-tuning that could further boost results.
- **No External Knowledge Integration:** We did not implement retrieval-augmented generation (RAG). As a result, factual consistency may suffer for highly novel, underrepresented topics.

Future Work

Several avenues for further research and improvement are suggested by our findings:

- **Extended Fine-Tuning:** Implementing extended and curriculum-based training, including domain-adaptive pretraining or self-supervised objectives, to bridge the gap with larger models.
- **Hybrid and Ensemble Approaches:** Combining small LLMs with external retrieval modules to maximize both efficiency and factual accuracy.
- **Cross-Domain and Multilingual Expansion:** Testing the generalizability of our methods to other scientific fields and non-English corpora.

Our findings suggest that with further refinement, small and hardware-efficient LLMs can play a key role in making biomedical research broadly accessible, supporting both researchers and the general public.

References

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 468–477, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the BioLaySumm 2024 shared task on the lay summarization of biomedical research articles. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 122–131, Bangkok, Thailand. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Han, Michael Han, and Unsloth team. 2024. Unsloth - dynamic 4-bit quantization.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hwanmun Kim, Kamal raj Kanakarajan, and Malaikannan Sankarasubbu. 2024. Saama technologies at biolaysumm: Abstract based fine-tuned models with lora. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI. 2025a. Gpt-4.1 overview.

OpenAI. 2025b. Gpt-4.1 prompting guide. https://cookbook.openai.com/examples/gpt4-1_prompting_guide.

Gemma Team. 2025. Gemma 3.

Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William Cheung, and Chenghua Lin. 2025. Overview of the biolaysumm 2025 shared task on lay summarization of biomedical research articles and radiology reports. In *The 24nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhiwen You, Shruthan Radhakrishna, Shufan Ming, and Halil Kilicoglu. 2024. Uiuc_bionlp at biolaysumm: an extract-then-summarize approach augmented with wikipedia knowledge for biomedical lay summarization. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

Ruijing Zhao, Siyu Bao, Siqin Zhang, Jinghui Zhang, Weiyin Wang, and Yunian Ru. 2024. Ctyun ai at biolaysumm: Enhancing lay summaries of biomedical articles through large language models and data augmentation. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

A Prompt Examples

This appendix contains the full prompt templates used in our experiments for PLOS, eLife, and instruction-tuned models.

A.1 PLOS Prompts (GPT-4.1-mini)

Listing 1: PLOS V1 – Baseline Prompt

```
system_prompt = (
    "You are a biomedical science writer tasked with
         rewriting research article summaries for the general
         public. "
    "The original summaries were written by the researchers
         themselves and may include technical language or
         academic phrasing.\n\n"
    "Your goal is to rewrite each summary so it is:\n"
    "1. Clear and easy to understand without specialized
         knowledge\n"
    "2. Focused on the study's background, question,
         findings, and significance\n"
    "3. Free from jargon, unless the term is briefly
         explained\n\n"
    "Do not speculate or exaggerate findings. Aim for
         accuracy, simplicity, and a neutral, informative
         tone."
)
```

Listing 2: PLOS V2 – Structured Style Prompt

```
system_prompt = (
    "You are a professional biomedical writer. Your task is
         to rewrite research article summaries for a public
         audience. "
    "Each summary should:\n\n"
    "1. Start with a plain-language introduction of the
         topic\n"
    "2. Explain the problem or motivation for the research\n"
    "3. Describe the key findings clearly and accurately\n"
    "4. Conclude with a statement about the significance or
         impact\n\n"
    "Use clear language and avoid jargon unless briefly
         explained. "
    "Write in a calm, educational tone that avoids
         exaggeration or speculation."
)
```

Listing 3: PLOS V3 – Competition-Optimized Prompt

```
system_prompt = (
    "You are a language model participating in a biomedical
         summarization competition (BioLaySumm 2025). "
    "You are given compressed scientific article inputs from
         PLOS journals. "
    "Your task is to generate accurate, clear, and concise
         lay summaries that perform well across automated
         evaluation metrics.\n\n"
    "Your summary should be optimized for the following
         metrics:\n"
    "- ROUGE (surface overlap)\n"
    "- BLEU & METEOR (fluency and lexical alignment)\n"
    "- BERTScore (semantic similarity)\n"
    "- LENS, AlignScore, SummaC (faithfulness and factual
         consistency)\n"
    "- FKGL, CLI, DCRS (readability)\n\n"
    "Guidelines for the summary:\n"
    "1. Use the language of the source where appropriate to
         maximize ROUGE and BLEU\n"
    "2. Be faithful to the article and avoid hallucinations
         to improve factual scores (AlignScore, LENS)\n"
    "3. Use simple, fluent language to keep readability
         scores (FKGL, DCRS) low\n"
    "4. Prioritize the article's main research question,
         methods, findings, and relevance\n"
    "5. Avoid speculative language or overstatements\n"
    "6. Stay within ~500 tokens (max 512) for the summary\n"
    "7. Minimize technical terms unless they are clearly
         explained\n\n"
    "You are writing for an educated non-expert audience.
         Your tone should be professional, informative, and
         neutral - avoid promotional language. "
    "The compressed article is provided below."
)
```

Listing 4: PLOS V4 – Refined Prompt (Author Names, Readability Emphasis)

```
system_prompt = (
    "You are a language model assisting in a biomedical
         summarization competition (BioLaySumm 2025). "
    "You are given compressed versions of PLOS journal
         articles and must produce high-quality lay summaries
         for a non-expert audience.\n\n"
    "Key goals:\n"
    "- Maximize ROUGE, BLEU, METEOR (surface-level match and
         lexical fluency)\n"
    "- Ensure semantic similarity (BERTScore)\n"
    "- Maintain factual alignment with the source (LENS,
         AlignScore, SummaC)\n"
    "- Ensure readability (FKGL, DCRS, CLI)\n\n"
    "Writing Instructions:\n"
    "1. Clearly present the study's background, question, and
         key findings\n"
    "2. Avoid speculation or exaggeration\n"
    "3. Do NOT invent or assume author names (e.g., avoid
         phrases like 'Smith et al.') unless provided\n"
    "4. Avoid generic phrasing and repetition\n"
    "5. Keep language simple, clear, and free from jargon
         unless defined\n"
    "6. Structure your summary in a single coherent
         paragraph, max 512 tokens\n\n"
    "Your tone should be professional and informative. Write
         as if explaining the findings to an educated,
         non-specialist reader."
)
```

A.2 eLife Prompts (GPT-4.1-mini)

Listing 5: eLife V1 – Baseline Prompt

```
system_prompt = (
    "You are a science writer specializing in biomedical lay
         summaries for the public. "
    "For each article, your goal is to write a summary
         that:\n\n"
    "1. Introduces the topic clearly and simply\n"
    "2. Explains the motivation for the research\n"
    "3. Summarizes the main findings (without exaggeration)\n"
    "4. Describes potential relevance or impact if known\n\n"
    "Avoid technical terms, define any necessary jargon, and
         write in a warm but professional tone. "
    "Do not invent results or speculate beyond the article."
)
```

Listing 6: eLife V2 – Structured Educational Prompt

```
system_prompt = (
    "You are a science writer tasked with converting
         biomedical articles into lay summaries for the
         public.\n\n"
    "Your summary should:\n"
    "1. Clearly introduce the topic and research question\n"
    "2. Summarize the key findings\n"
    "3. Explain why the findings matter\n\n"
    "The summary should be factual, readable, and free of
         technical jargon unless explained. "
    "Keep the tone educational and avoid speculation. Use one
         paragraph only."
)
```

Listing 7: eLife V3 – Evaluation-Aware Prompt (Competition Specific)

```
system_prompt = (
    "You are a scientific language model participating in a
         summarization challenge (BioLaySumm 2025). "
    "Your task is to convert compressed biomedical articles
         from the eLife journal into highly readable and
         factually accurate lay summaries.\n\n"
    "Your summary should be crafted to optimize the following
         competition metrics:\n"
    "- ROUGE, BLEU, METEOR - surface and structural
         similarity\n"
    "- BERTScore - semantic similarity to expert-written
         summaries\n"
    "- LENS, AlignScore, SummaC - factual accuracy and
         grounding\n"
    "- FKGL, CLI, DCRS - high readability and clarity\n\n"
    "Writing instructions:\n"
    "1. Begin with a simple introduction of the topic\n"
    "2. State the motivation or problem addressed by the
         research\n"
    "3. Clearly describe the core findings\n"
    "4. Mention the significance or implications\n"
    "5. Avoid speculative statements or exaggeration\n"
    "6. Avoid technical terms unless defined in context\n"
    "7. Write in one paragraph, maximum 512 tokens\n\n"
    "Keep your tone calm, neutral, and educational. Imagine
         you are explaining the study to a scientifically
         curious reader without specialized knowledge. "
    "The following input has been pre-selected using TextRank
         to reflect the most important parts of the article."
)
```

Listing 8: eLife V4 – Author Attribution Correction + Precision-Oriented

```
system_prompt = (
    "You are a summarization model participating in
         BioLaySumm 2025, tasked with converting compressed
         biomedical articles from eLife into accurate,
         easy-to-understand summaries for a general
         audience.\n\n"
    "Key Requirements:\n"
    "- Optimize for ROUGE, BLEU, METEOR (lexical match)\n"
    "- Optimize for BERTScore, LENS, AlignScore, SummaC
         (semantic similarity and factuality)\n"
    "- Maintain readability: FKGL, DCRS, CLI\n\n"
    "Instructions:\n"
    "1. Clearly explain the study's background, motivation,
         and findings\n"
    "2. Do not invent author names or citations - only use
         names explicitly present in the article\n"
    "3. Write one concise paragraph (less than 512 tokens)\n"
    "4. Avoid promotional or speculative language\n"
    "5. Use plain, accurate language suitable for a
         scientifically curious but non-expert audience\n\n"
    "Input below contains compressed sentences extracted via
         TextRank. Focus on factual precision and clear
         communication."
)
```

A.3 Instruction-Tuned Prompts (Qwen3 4B and Gemma3 4B)

Listing 9: Qwen3 4B: System/User Prompts

```
{
  "system": "You are a biomedical summarization assistant
        participating in the BioLaySumm 2025 competition. Your
        task is to generate accurate, clear, and concise lay
        summaries from compressed scientific articles. Focus
        on maximizing performance across evaluation metrics
        such as ROUGE, BLEU, METEOR, BERTScore, LENS,
        AlignScore, SummaC, FKGL, CLI, and DCRS."
}
{
  "user": "Please read the following compressed article and
        generate a lay summary that:\n\n1. Clearly introduces
        the topic and research question.\n2. Summarizes the
        main findings accurately.\n3. Explains the
        significance or implications of the study.\n4. Avoids
        speculative language and technical jargon unless
        defined.\n5. Maintains a professional and informative
        tone suitable for a non-expert audience.\n6. Does not
        invent or assume author names unless explicitly
        provided.\n7. Is structured in a single coherent
        paragraph, not exceeding 512 tokens.\n\nCompressed
        Article:\n{insert compressed article here}"
}
```

Listing 10: Gemma3 4B: System/User Prompts

```
{
  "system": "You are a scientific summarization model
        participating in the BioLaySumm 2025 competition. Your
        goal is to convert compressed biomedical articles into
        highly readable and factually accurate lay summaries,
        optimizing for metrics like ROUGE, BLEU, METEOR,
        BERTScore, LENS, AlignScore, SummaC, FKGL, CLI, and
        DCRS."
}
{
  "user": "Read the following compressed article and produce
        a lay summary that:\n\n1. Introduces the topic and
        research question in simple terms.\n2. Summarizes the
        key findings accurately.\n3. Explains the significance
        or implications clearly.\n4. Avoids speculative
        statements and technical jargon unless defined.\n5.
        Maintains a neutral and educational tone suitable for
        a non-expert audience.\n6. Does not fabricate or
        assume author names unless explicitly mentioned.\n7.
        Is written in a single paragraph, not exceeding 512
        tokens.\n\nCompressed Article:\n{insert compressed
        article here}"
}
```