A Comprehensive Benchmarking Framework for Evaluating Traditional and Large Language Model-Based Text Summarization Techniques

Ann Baby[1,3], Basant Kumar[1,2]
[1] Lincoln University College, Malaysia;
[2] Modern College of Business and Science, Muscat, Sultanate of Oman
[3] Rajagiri College of Social Sciences, India
Email ID – annbaby2k@yahoo.co.in

Abstract
The surge of unstructured textual data on various platforms such as recruitment websites, legal databases, and enterprise management systems has further driven the need for automation in text summarization. There are a variety of summarization techniques that fall into one of three categories: extractive, abstractive, and large language model (LLM) based summarization. This paper introduces a unified and deployable framework that unifies and places extractive, abstractive and LLM-based summarization methods in a single standardized evaluation pipeline. It uses a common input repository, a domain-specific resume corpus, the same pre-processing for all paradigms, and benchmarks each approach using ROUGE-1, ROUGE-2 and ROUGE-L F1 scores on a CPU-based deployment served via a Hugging Face Spaces interface. Three progressive model variants – an abstractive summarizer, a multi-paradigm summarizer mode, and an advanced resume intelligence model – were implemented and assessed in terms of model performance, resource consumption, cost-effectiveness, and interpretability. The macro-average ROUGE F1 score ranges around 0.55 for paradigms, and extractive summarization has the best lexical overlap between summary and reference, while LLM-based summarization has the most contextually fluent summary but the least lexically anchored one. The proposed framework enables the reproducible comparative evaluation and also brings practical guidelines for choosing summarization paradigms in a real world deployment scenario, e.g., resume screening and recruitment analytics in a resource constrained setting.

Keywords*: Text Summarization, Extractive Summarization, Abstractive Summarization, Large Language Models, ROUGE Evaluation, Resume Intelligence, Deployable NLP Systems*

1. INTRODUCTION

In the era of natural language processing (NLP), automatic text summarization has emerged as an essential tool for transforming vast amounts of unstructured text into succinct, informative summaries. In the last 20 years, there has been a constant interest in automated summarization techniques due to the increasing number of digital documents that are being published on web-based recruitment platforms, legal documentation systems, customer support logs, and enterprise knowledge bases that make manual summarization increasingly impractical.

Summarization research has, in general, developed in three major paradigms. Extractive summarization is the process of finding and capturing the most important sentences or phrases from the source text, without introducing too much fluency or coherence to the summary. Abstractive summarization relies on sequence-to-sequence transformer architectures, producing novel sentences that are paraphrases and shorten the source, resulting in more human-like summaries at higher computational cost. More recently, a third paradigm, known as large language models (LLMs), has emerged which can generate fluent, context-aware, domain-generalized summaries, but requires much more resource and comes without clear interpretability. Although both paradigms are well established, the research field is still fragmented. The studies usually compare one summarization strategy with a limited number of baselines and employ evaluation methods unique to the individual studies. This fragmentation will hinder informed and

evidence-based decisions for practitioners on which summarization strategy to use for a particular deployment scenario, especially when computational resources, latency, and interpretability are competing constraints.

2. PROBLEM STATEMENT AND RESEARCH GAP

Existing research on summarization has failed to keep up with the pervasive and fast growth of unstructured textual data in recruitment, legal, and enterprise applications. Based on the problem context given above, specific gaps that this work addresses are:

- Lack of ability to have a unified system that supports extractive, abstractive summarization, and LLM-based summarization in a single deployable pipeline.
- Study-specific evaluation criteria make it difficult to compare the results obtained from different studies.
- Most previous work has been done in a lab setting, typically offline and in notebook-like experiments, rather than in interactive, user-fronted systems.
- Lack of research on summarization methods on domain-specific, high-stakes tasks like summarizing resumes and candidate-profiles, where interpretability and factual correctness are essential.

3. RESEARCH OBJECTIVES

Considering the research gap reviewed above, the following objectives for this study are set:

- To capture existing gaps in the summarization research, such as the prevalence of single paradigms, non-uniform evaluation processes, and the absence of deployable and reproducible systems indicated by publication and collaboration trends.
- To create a consistent and deployable platform to run extractive, abstractive and LLM-based summarization models and provide a single standardized testing procedure.

4. LITERATURE REVIEW

The literature is categorized into five related fields:

4.1.1 Extractive Text Summarization Methods

The extractive approaches are those that generate a summary by directly picking out and stitching together salient sentences or phrases from the source document, often either through statistical scoring or by ranking based on graphs or lexical-frequency heuristics. These methods are appreciated because they are clear, accurate, and inexpensive, as the output that they produce is assured to be a text that matches the source verbatim. An extractive summary, however, may suffer a lack of inter-sentence coherence, redundancy, and may not be able to compress information beyond sentence selection (Nallapati et al., 2016).

4.2 Abstractive Summarization with Transformer Models

Transformer-based sequence-to-sequence models such as BART, T5 and distilled versions have significantly pushed the envelope of abstractive summarisation by training to produce concise, semantically rich paraphrases of the source text rather than just extract the text. These are trained on large text corpora and fine-tuned on summarization-specific datasets, allowing them to generate human-like and coherent summaries (Lewis et. al., 2020).

4.3 Large Language Model-Based Summarization

With the advent of large language models, there has emerged a third summarisation paradigm that generates fluent, context-appropriate, and very generalizable summarisation. While LLM summarizers are flexible about the domains and prompts they can handle with little fine-tuning, they have higher memory, computational, and latency costs, and interpretability is lower than extractive and even traditional abstractive summarizers. Despite the fact that more and more researchers are aware that ROUGE is not

the only reliable metric for assessing factual consistency or semantic fidelity, most studies to date in the literature are still based on ROUGE evaluation (Shleifer et. al., 2020).

4.4 Evaluation Metrics for Text Summarization

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is still the most widely-used family of metrics for evaluating summarization, which include ROUGE-1 for the overlap of unigrams, ROUGE-2 for the overlap of bigrams and ROUGE-L for the longest common subsequence between generated and reference summaries (Lin, 2004)). Though ROUGE is simple to compute and commonly used, it is a lexical overlap metric and fails to directly measure semantic adequacy, factual correctness and interpretability, leading to the addition of complementary resource utilisation and interpretability data to the suggested framework (Zhang et. al., 2020).

4.5 Summarization Systems for Recruitment and Resume Analysis

Summarization in the area of recruitment and resume analysis has its own set of problems, such as semi-structured formatting, heavy factual content and the high cost of fact errors in the hiring process (El-Kassas et. al., 2021). Previous research has been carried out in this domain using a single summarization paradigm to resume a text, with no clear, comparative evidence from which to draw conclusions for this use case, such as when recruiting or researching a text (Zhang & Zhang, 2023).

5. PROPOSED METHODOLOGY AND SYSTEM ARCHITECTURE

In this study, we come up with a new text summarization framework that unifies extractive, abstractive, and LLM-based text summarization methods into a single, unified evaluation and deployment pipeline to solve the identified research gap. The architecture follows three design principles, namely: input normalization, paradigm agnostic deployment and multi-dimensional and consistent evaluation.

5.1 Standardized Input Repository

The framework is based on a common input repository, which in this case is a domain-specific text corpus (in this case a resume text corpus) that means all summarization paradigms are working on the same set of input documents. This design feature allows input-variability to not be a confounding variable when comparing paradigms. All documents have been preprocessed before summarization in the same way, namely by tokenization, text cleaning and text normalization, to ensure that the differences in ROUGE scores that are measured in the downstream part of the pipeline are due to the summarization method and not due to inconsistencies in how the documents are prepared as input to summarization.

5.2 Deployment Architecture

Deployed on CPUs-based infrastructure, the framework is an achievable performance baseline that is not dependent on specialized GPU infrastructure, thus representing the resources constraints of small organizations and individual researchers. The system is presented using a Hugging Face Spaces interface, which offers an interactive and user-friendly interface where a user can pick a summarization paradigm and get inference results in real time. The deployment design directly solves the research gap identified for the lack of real world deployment evidence in previous summarization research. In addition to raw summarization output, the automated system enables four types of comparative analyses: performance comparison between the paradigms based on the ROUGE standard, analyses of resource utilization (such as memory footprint and inference latency), cost-efficiency benchmarking of the system as a function of model size and hardware requirements, and qualitative interpretability evaluation of the generated summaries.

5.3 Three-Stage Model Progression

The framework was designed and tested in three stages of progressive changes in the capability of the model. Baseline Abstractive Summarization Model: A single-paradigm pipeline based on a distilled transformer (DistilBART-CNN-12-6) as a baseline model that serves as the baseline to compare with the multi-paradigm variants.

Multi-Paradigm Summarization Mode: This is an extension of the baseline where extractive and LLM based summarization models are integrated with the abstractive model, using a common input pipeline, and a common ROUGE based evaluation harness for easy side-by-side comparison of all three paradigms on the same document. The most complete variant that adds domain-specific resume-analysis functionality — structured fields extraction and summarization of candidate's profiles — to the multi-paradigm summarization core, making it suitable to be practically used in recruitment and HR-analytics workflows. From an architectural perspective, the pipeline takes each input document to a standard, pre-processing phase, and then sends each to the extractive selector, the abstractive transformer, and the LLM-based summarizer, all of which run in parallel. The outputs of all three paradigms then pass through a common ROUGE evaluation module, which provides performance, resource and interpretability metrics to be captured and compared in the same experimental conditions.

6. DATASETS

The system was tested with publicly available resumes datasets that were chosen to have the semi-structured and information-rich nature of the text commonly found in the recruitment domain which included:

- Kaggle – Resume Dataset (https://www.kaggle.com/datasets/andrewmvd/resume-dataset)
- Kaggle – HR Analytics Resume Dataset (https://www.kaggle.com/datasets/mohammedmostafa/resumes)
- GitHub – Resume Text Collections (supplementary domain text sources)

The use of publicly available, resume-specific datasets guarantees the repeatability of the results presented, and enables other researchers to follow the same preprocessing and multi-paradigm evaluation protocol outlined in Section 5 on the same data base.

7. EXPERIMENTAL SETUP

All experiments were performed on CPU-based infrastructure to mimic realistic scenarios with limited resources as opposed to idealized scenarios on GPUs for benchmarking. In the extractive paradigm, a statistical sentence-selector was utilized to rank and extract salient sentences without any paraphrasing was done (Mihalcea & Tarau, 2004). The DistilBART-CNN-12-6 model was used for the abstractive paradigm, as it is a compact and distilled transformer that can be used with CPU inference. In the LLM-based paradigm, we chose TinyLLaMA-1.1B, a lightweight LLM that can generate fluent and context-aware summaries within the computational budget of the deployment environment (Zhang et. al., 2024). The 3 models were tested on the same group of pre-processed resume documents from the same repository of standardized input documents (as described in Section 5.1). Summaries generated by the three paradigms were then evaluated against reference summaries, using the ROUGE-1, ROUGE-2, and ROUGE-L F1 metrics. ROUGE scores were calculated the same way for all three paradigms so that the evaluation could provide a fair and controlled comparison (Erkan et. al., 2004). For each paradigm, in addition to ROUGE scoring, inference latency and memory usage were recorded, for the purpose of resource-utilization and cost-efficiency analyses, as described in Section 5.2.

8. RESULTS AND DISCUSSION

Table 1 summarises the ROUGE-1, ROUGE-2 and ROUGE-L F1 scores achieved with each summarisation paradigm and a qualitative description of the characteristics of the output for each summarisation paradigm.

Table 1. Comparative ROUGE F1 performance across summarization paradigms

Summarization Paradigm	Target Model Used	ROUGE-1 (F1)	ROUGE-2 (F1)	ROUGE-L (F1)	Output Characteristics
Extractive	Statistical Selector	0.584	0.542	0.569	Stable lexical overlap; zero paraphrasing
Abstractive	DistilBART-CNN-12-6	0.561	0.528	0.549	Balanced compression; concise phrasing
LLM-Based	TinyLLaMA-1.1B	0.535	0.501	0.524	Contextually fluent; highly generalized
Macro Average	Framework Baseline	0.560	0.523	0.547	Overall uniform key-phrase tracking

8.1 Quantitative Benchmarking

The results of the experiment indicate that the macro-average ROUGE F1 score ranges around 0.55 on all three paradigms. As its verbatim-selection approach maximizes the lexical overlap with the reference text it obtained the highest scores for all three ROUGE variants (ROUGE-1 = 0.584, ROUGE-2 = 0.542, ROUGE-L = 0.569). The DistilBART-CNN-12-6 model obtained scores that were very close to the best, with ROUGE-1 = 0.561, ROUGE-2 = 0.528 and ROUGE-L = 0.549, indicating that it is able to capture the overall theme of the article while retaining the most essential information. Although LLM-based summarization is qualitatively the most fluent and generalized in terms of context, it achieved the lowest ROUGE scores of the three paradigms (ROUGE-1 = 0.535, ROUGE-2 = 0.501, and ROUGE-L = 0.524). As summarization paradigms progress from extractive selection to abstractive paraphrasing to open-ended generation by LLM, the ROUGE scores decrease in a predictable way even while the contextual fluency and readability seem to increase along the way. The identification of this finding has practical applications for paradigm selection: applications that value factual traceability and auditability, such as legal or compliance summarization, may benefited more from extractive or abstractive approaches, while applications that value readability and narrative coherence over strict lexical fidelity may benefit more from LLM-based summarization, allowing for additional computational cost.

8.2 Resource Utilization and Cost-Efficiency

As in the other paradigms, the extractive selector had the lowest computational overhead because it does not depend on deep learning for inference. The abstractive DistilBART model was a moderate computational cost model as it was distilled when compared to full-scale transformer summarizers. While LLM-based summarization is a relatively small LLM, it still had the highest memory usage and inference latency of the three paradigms, highlighting the resource cost tradeoff that comes with using LLM-based summarization even when working at smaller LLM sizes. In this context, these observations advocate for the importance of a unified and deployable framework to evaluate these three paradigms: the differences concerning resource and cost-efficiency that are often neglected in the offline comparison of the paradigms are directly visible when all three paradigms are compared under the same CPU-based deployment conditions.

8.3 Interpretability Assessment

The three paradigms were distinguished by qualitative interpretability assessment. Extractive summaries had the best interpretability because each word of the summary was explicitly related to a word in the source document and no paraphrasing was done. Abstractive summaries were typically short and constrained in length and expressed in a way that maintained some sense of connection to the source

text, but not to a high degree. The LLM-based summaries were the most fluent and generalized of the three paradigms, but were also the least interpretable, since the generative flexibility of LLM-based summaries made it more difficult to identify specific phrases in the output that correspond to evidence in the source materials, which is a key factor in contexts like recruitment and resume analysis where factual traceability becomes directly relevant to the fairness and defensibility of hiring decisions.

10. LIMITATIONS, CONCLUSION AND FUTURE WORK

The present study has several limitations. There are a number of limitations of the present study. There are some restrictions that should be noted for the conclusions drawn from this study. First, the evaluation is based mainly on ROUGE-based metrics that reflect lexical overlap, but not in the sense of being factually consistent, semantically adequate or free from bias in the generated summary — a property that is much more relevant in the context of recruitment and hiring.

This paper introduced a unified and deployable framework for comparative evaluation of extractive, abstractive and LLM-based text summarization developed and tested with a domain specific resume dataset. This work tackles the research gap relating to the absence of a unified, reproducible and deployable summarization-evaluation system highlighted in the literature by introducing a unified input repository, preprocessing pipeline and evaluation protocol, and by deploying the framework on CPU-based infrastructure via an interactive Hugging Face Spaces interface (Fabbri et. al., 2021). The three-stage model progression from a baseline abstractive summarizer to a multi-paradigm summarization mode to an advanced resume intelligence framework is a practical way of paving the way to deployable multi-paradigm summarization systems starting from single-paradigm summarization research. With a ROUGE-F1 macro-average of approximately 0.55, our results demonstrate the trade-off between lexical fidelity and semantic fluency across extractive, abstractive, and LLM-based summarization, providing practical guidance for model selection. Future work will integrate semantic and factual evaluation metrics and extend validation to diverse domains, such as legal and enterprise documents, to enhance generalizability. (Koh et. al., 2022).

REFERENCES

[1] Nallapati, R., Zhou, B., dos Santos, C., Gùlçehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, 280–290.

[2] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BERT: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7871–7880.

[3] Shleifer, S., & Rush, A. M. (2020). Pre-trained summarization distillation. arXiv preprint.

[4] Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 74–81.

[5] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. International Conference on Learning Representations.

[6] El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. Expert Systems with Applications, 165, 113679.

[7] Zhang, P., & Zhang, T. (2023). Large language models for text summarization: A survey. arXiv preprint.

[8] Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). TinyLlama: An open-source small language model. arXiv preprint.

[9] Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 22, 457–479.

[10] Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 404–411.
[11] Kaggle. Resume Dataset. https://www.kaggle.com/datasets/andrewmvd/resume-dataset
[12] Kaggle. HR Analytics Resume Dataset. https://www.kaggle.com/datasets/mohammedmostafa/resumes
[13] Hugging Face. Hugging Face Spaces Documentation. https://huggingface.co/docs/hub/spaces
[14] Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). SummEval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9, 391–409.
[15] Koh, H. Y., Ju, J., Liu, M., & Pan, S. (2022). An empirical survey on long document summarization: Datasets, models, and metrics. ACM Computing Surveys, 55(8), 1–35.