A Unified and Deployable Framework for Comparative Evaluation of Extractive, Abstractive, and LLM Based Text Summarization

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: Large scale unstructured text data used in enterprise systems, job markets, and online communication channels has enhanced the need to develop effective text summarization methods. Classical summarization models fall into extractive and abstractive, whereas more recent models based on large language models (LLM) offer significant context-sensitive and highly fluent summarization functions. Although research has been made in each paradigm, the majority of the studies contrast these methods and are mostly in controlled offline settings. This diversity reduces the possibility of a systematic comparison of the summarization paradigms and their adequacy in the real-world application. A deployable framework of the comparative assessment of extractive, abstractive, and LLM-based text summarization systems is proposed in this paper. The structure combines several models of summarization in a stratified architecture and compares them with each other through conventional metrics and statistical analysis. The system proposed is oriented to real-world applications like recruitment and resume analysis, where the summarization of the profiles of the candidates is essential. By using the regular evaluation metrics and comparative statistical analysis, the framework will seek to offer reproducible and deployment-oriented insights on the performance of various summarization paradigms.

Keywords: **Text Summarization, large language model (LLM), natural language processing (NLP), Extractive Summarization, Abstractive Summarization, Large Language Models**

Introduction

Digitalization of organizations has led to a new amount of textual information being produced in various fields never before. The human resource systems, the legal documentation systems, the academic repositories and enterprise knowledge bases keep on generating enormous amounts of loose text. Making significant conclusions on such large textual datasets is a significant challenge that poses significant challenges to both organizations and researchers. Text summarization has become one of the important natural language processing (NLP) tasks that are used to summarize big textual documents into shorter forms without losing important information[1]. There have been two distinct types of text summarization

methods namely extractive and abstractive methods. Extractive summarization locates and picks the most relevant sentences or phrases that are directly related to the source document and creates summaries that are not faithful to the original wording [2]. Conversely, abstractive summarization produces new sentences based on the original text, which is a paraphrase of the original text, and in most cases, the summaries are more coherent and concise [3].

Due to the introduction of transformer-based architecture and large language models (LLMs) the summarization possibilities have grown very large. Current studies determine summarization methods separately on various datasets, evaluation criteria, and experimental set-ups. Moreover, most of the studies are only restricted to experimental prototypes without the implementation of deployable architecture that can be incorporated in real life systems [4]. Such absence of assessment models constrains the capability of researchers and practitioners to identify the most appropriate summarization paradigm that can be applied under particular applications. To overcome this issue, this paper suggests a single and deployable system that would allow the comparative analysis of extractive, abstractive, and LLM-based summarization strategies under one experimental setting.

Problem Statement and Research Gap.

Despite the extensive applications and literature on text summarization in natural language processing community, the assessment of the summarization models is still disjointed. The present research is usually confined to one paradigm, i.e. extractive summarization or transformer-based abstractive summarization, and the comparison of these two paradigms is not done systematically in the same framework. Such a strategy has various constraints. First, due to the lack of standard evaluation pipelines, it is hard to compare across different summarization paradigms fairly. The results of different datasets, preprocessing plans, and metrics of evaluation cannot be compared directly [5].

Second, the majority of the research in the field of summarization is carried out in laboratory settings; models are tested on benchmark data under the assumption that their implementation does not affect scalability or consumption of computing resources, or their interaction with various enterprise applications. Third, large language models have presented a genre of summarization methods that are generative and based on reasoning contextually. Although these models have proven to be incredibly performing, they are usually evaluated independently of the usual summarization methods [6].

These weaknesses demonstrate a strong gap in the research: the absence of a single system that would combine a variety of summarization paradigms and allow comparing them to each other in the system of equal evaluation. The proposed study fills this gap by creating a consolidated deployable framework to support extractive, abstractive, and LLM based summarization models on a shared architecture. The standardized evaluation metrics and methods of statistical analysis are also integrated into the framework in order to provide reproducible and meaningful comparisons.

Research Objectives

The primary objective of this work is to make a framework that makes it possible to compare the achievements of various text summarization paradigms. The objectives of this study are to:

- Examine literature in text summarization, specifically the recent shift in their traditional extractive models, and how abstractive text models are designed as transformers and more recent text models that are based on LLMs.
- Determine current gaps in the research of summarization such as inconsistency in evaluation practice, isolated comparisons of models, and lack of deployable architectures to support real-life applications.
- Create a deployable evaluation system that will enable a standardized comparison of the summarization techniques based on standardized metrics and statistical validation measures.

Related work

Extractive Text Summarizing: Extractive summarization techniques are among the oldest automatic summarization techniques. Such methods are used to find and pick the most relevant sentences in a document depending on their features like the location of the sentence, word frequency, similarity, and graphical ranking algorithms. The principle benefit of extractive summarization is that it is transparent and factual. Given that the chosen sentences are the direct copy of the source document, the probability of the semantic distortion is low [7]. Moreover, extractive approaches tend not to need a large amount of computational resources in contrast to generative models. Extractive summaries can be redundant and incoherent since the sentences picked do not necessarily create a coherent account of what is being discussed.

Abstractive Summarization Transformer-based models: Abstractive summarization tries to produce new sentences that reflect the text of the source as opposed to reproducing it. The abstractive summarization models were greatly enhanced with the introduction of transformer architectures. Transformer-based models use attention as a means of capturing long-range dependencies in text and therefore can generate more coherent and semantically rich summaries. These models have the ability of paraphrasing information, combining several sentences, and creating brief summaries of complicated text [8]. Although these models have these benefits, abstractive models can be expensive in terms of training data and computation. Also, they can sometimes produce hallucinated or factually wrong information.

Large Language Model Based Summarization: The most recent development of text summarization is large language models. The models are trained on large corpora and can be used to accomplish summarization tasks using prompt-based instructions or fine-tuning strategies. The systems that are based on LLM show good contextual knowledge and produce much fluent summaries. They are also able to change summarization styles and length based on user prompts [9]. Nonetheless, these models are demanding in resources and might pose a dilemma in terms of cost of computation, latency, and overhead deployment limitations.

Evaluation Measures in Text Summarization: Summarization systems are usually evaluated based on the automatic measures like ROUGE scores that assess the similarity between generated and reference summaries [10]. Although they are very popular, these measures record mostly lexical similarity and do

not necessarily capture the quality of semantics or readability. So, statistically comparing the metrics is critical to automated metrics in order to deliver sound evaluation outcomes.

Current Layered Architecture Proposal.

In order to overcome the limitations, the proposed study will present a single-layered architecture that evaluates extractive, abstractive, and large language model (LLM) based summarization methods as a single evaluation framework. The architecture is formulated in a manner that facilitates regular preprocesses, systematic evaluation as well as implementable deployment of summarization models into real life applications. The suggested architecture is in the form of a modular layered design, with every layer having a particular function in the summarization pipeline. The proposed framework is detailed in Figure 1. The different layers have different functions but still, they are interoperable throughout the system.

Input Standardization Layer.

The first layer of the framework is the Input Standardization Layer. The main goal of it is to make sure that all the summarization paradigms work with the same and uniformly processed input data and thus can be fairly compared. A layer consumes a domain-specific corpus, specifically resume datasets retrieved in the public repositories. Because resumes are usually heterogeneous in terms of formatting, punctuated differently, and industry specific terminologies, preprocessing is very essential. It is a layer that carries out a number of preprocessing tasks such as text cleaning, tokenization, and normalization. Text cleaning removes meaningless characters, HTML tags and artifacts of formatting [11]. The tokenization divides the text into sentences or words enabling the analysis of the text downstream. The standardization of the textual representation is done by normalization, which transforms text into standardized formats like lowercasing and elimination of unnecessary whitespace. This layer renders methodological consistency across the summarization paradigms by standardizing the input of all the models.

Parallel Summarization Engine Layer.

The Parallel Summarization Engine Layer is the main analytical element of the framework. Different models of summarization are applied to the standardized data in this layer. This design enables the system to produce summaries with the various paradigms but with the same input conditions. Three engines function here in parallel.

- Extractive Summarization Engine whereby significant sentences are identified and ranked by some statistical characteristics including term frequency-inverse document frequency (TF-IDF) and graph-based ranking methods. Top-k selection does the top-ranking of the sentences to create the final summary.
- Abstractive Summarization Engine, which is based on encoder decoder transformer architecture. The summaries in these models are produced by perception of contextual correlation of words

through attention mechanisms. Contrary to extractive models, abstractive models form new sentences that have the semantic meaning of the source text.

- LLM-Based Summarization Engine, which uses the pre-trained large language models, which have the ability to provide summarization using prompt-based inference. These are models that rely on large-scale pretraining on large corpora to produce fluent, coherent and context-aware summaries.

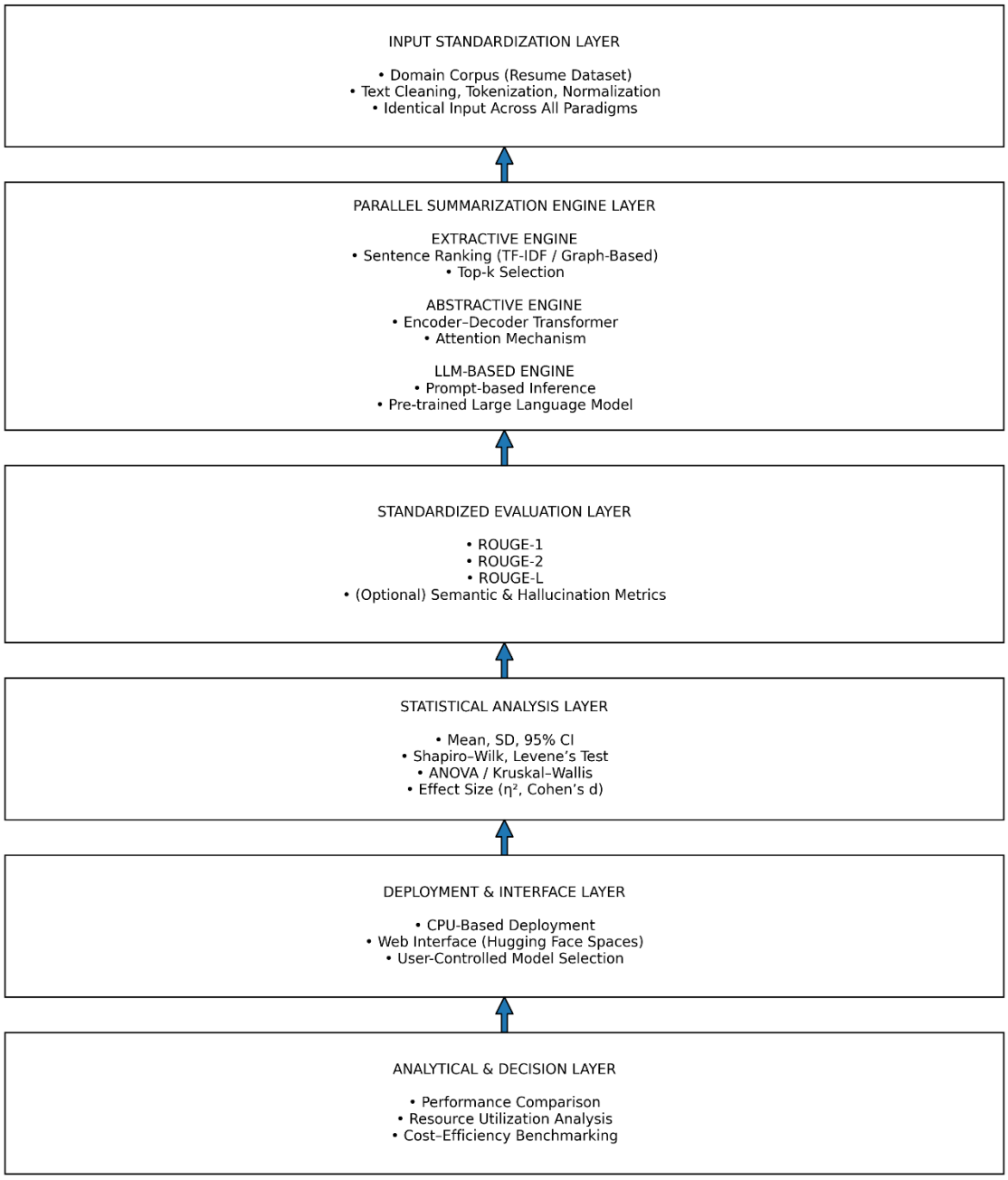

Figure 1. Architecture of the proposed unified text summarization framework integrating extractive, abstractive, and LLM-based approaches with standardized ROUGE evaluation.

Standardized Evaluation Layer

The Standardized Evaluation Layer determines the quality of the summaries that are produced by the various summarization engines. The similarity between system-generated summaries and reference summaries is determined by automated metrics. The evaluation measurements used mostly in the framework are ROUGE-based evaluation measures, such as ROUGE-1, ROUGE-2, and ROUGE-L [12]. These measures are an overlap of unigrams, bigrams and longest common sequence similarity respectively. Also, optional semantic evaluation measures can be added to measure semantic similarity and hallucinating detection, particularly in generative summarization models.

Statistical Analysis Layer

Statistical Analysis Layer provides stringent statistical testing of the results of the evaluation. Rather than taking the raw metric scores and interpreting the results only, this layer will make sure that the observed differences in performance between models are statistically significant. The layer determines descriptive statistics such as standard deviation, mean values and 95% confidence interval. The Shapiro-Wilk test is used to check the normality of the evaluation data whereas the Levene test is used to check the homogeneity of variance. Analysis of Variance (ANOVA) or the Kruskal-Wallis test are used to compare the performance of the models depending on the distribution of the data. Also, the measures of the effect size like e2 (eta squared) and Cohen d are computed to determine the scale of differences between models [13].

Deployment Layer and Interface Layer.

The Deployment and Interface Layer make sure that the suggested framework is capable of working in practice. The system can be deployed on a CPU-based as well as can be deployed anywhere even in low-computation capacity environments. The interface is a web-based interface, which may be done through Hugging Face Spaces, and in which the user can interact with the system and create summaries in real-time [14]. The interface also allows selecting the model controlled by the user, which gives a possibility to compare the techniques of summarization in real time.

Decision Layer and Analytical.

Analytical and Decision Layer is the last phase of the architecture that offers information based on the results of the evaluation. This layer does comparative performance analysis and looks at the performance of various summarization paradigms on various evaluation criteria. It also performs resource usage analysis, which evaluates computation needs like processing time and memory usage. In addition, the layer conducts benchmarking of cost-efficiency, which assists in saying what summarization method is the most efficient in relation to performance and computational cost.

Collectively, these layers form a full, repeatable, and implementable architecture of assessing the current audio termination methods in practical use instances like staffing analytics and processing records.

Performance Analysis of the Model

The approach taken in this study is aimed at implementing systematic and statistically valid comparisons between extractive, abstractive and LLM-based summarization processes. The process of evaluation will start with the creation of summaries of the individual documents in the dataset with the various models of summarization applied in the framework. Automated evaluation metrics are then used to compare the generated summaries with reference summaries. The performance scores obtained undergo statistical methods and the dissimilarities among summarization models assess whether the dissimilarities are statistically significant.

First, the Shapiro-Wilk test is used to establish whether the scores of the evaluation follow a normal distribution or not [15]. This is the step that is needed to determine whether the parametric or non-parametric statistical tools are to be used. When the evaluation scores are normally distributed, then a One-Way Analysis of Variance (ANOVA) is applied to test the difference in the mean performance scores of the various summarization models [16]. ANOVA is used to find out the statistically significant differences between models being tested. Individual paring of models across each other of the significantly different pairs are identified using Tukey Honest Significant Difference (HSD) test which is a post-hoc analysis, in case the ANOVA test shows significant difference.

When the Shapiro-Wilk test shows that the data is not normally distributed, then Kruskal-Wallis H test is applied as a non-parametric alternative to ANOVA. The test is used in comparison between the median performance measure of the various summarization methods. In the case where significant differences are detected using Kruskal-Wallis test, the Dunn test is conducted to conduct a pair-wise comparison between the summarization models [17]. This is a combination of the statistical tests, which guarantee the rigor, reliability, and scientific validity of the evaluation of the summarization techniques.

Dataset Description

The study will use publicly available resume datasets as retrieved by open data repositories. These datasets include textual data derived out of candidate resumes on various professional fields. Every resume has normally seen sections describing work experience, education, technical grades, certification and project description. This is because such documents with a lot of information will give a perfect scenario of assessing the systems of summarization. In this study, the main datasets are derived on publicly available platforms, i.e., the Kaggle and open-source GIT repositories. Such data sets comprise hundreds or thousands of resumes having different document lengths, styles of writing and structure.

Resume datasets make the research on summarization have a realistic application environment. When hiring employees, hiring managers and human resource experts are usually required to go through hundreds of profiles and select a few in a short period. They can be helped with automated summarization systems that will create concise summaries with focus on the most relevant qualifications and experiences. The heterogeneity of the documents in these datasets also guarantees that the evaluation

structure will have received test when evaluating systems under real-world conditions in which the summarization systems will be required to process heterogenous textual documents.

Expected Contributions

The suggested study will make a number of significant contributions to the sphere of natural language processing and automated summarization. First, the paper presents a common evaluation system that allows making a systematic comparison of extractive, abstractive, and LLCM-based summarization methods. This is a case of the significant weakness of current studies that various summarization paradigms are mostly tested separately. Second, a layered structure suggested by the author enhances the reproducibility and scalability of the suggested architecture by offering a modular system design which enables the integration of new models, datasets and evaluation metrics. Third, the study includes a strict level of statistical testing to justify the difference in performance of the summarization models, thus enhancing the credibility of the experimental findings. Fourth, the addition of a deployment layer helps bridge the gap between the research and practical applications, showing the ways of how summarization systems can be put into practice into such real-life application as recruitment analytics systems. Lastly, the paper offers information about the trade-offs among the paradigms of summarization such as summary accuracy, linguistic fluency, cost of computation, and scalability.

Conclusion & Future Work

Though several methods have been derived in summarization, lack of harmonized methods of evaluation has restricted the objective comparison of methods. In this work, a combined and deployable framework that could be used to do a comparative analysis of the extractive, abstractive, and large language model-based summarization models was proposed. The framework incorporates various paradigms of summarization into a modular layered framework and uses standardized evaluation metrics together with statistical testing. The proposed framework will allow conducting more and more comprehensive comparisons and understanding the advantages and drawbacks of various summarization methods. Moreover, the combination of the deployment capabilities illustrates the way the technologies of summarization may be implemented in practical contexts like recruitment analytics. Future studies will target application of the framework in bigger datasets, use of more sophisticated evaluation measures like semantic similarity measures, and human-based evaluations to establish the readability and interpretability of generated summaries.

References

1. Sharma, K.P., Yajid, M.S.A., Gowrishankar, J., Mahajan, R., Alsoud, A.R., Jadhav, A. and Singh, D., 2025. A systematic review on text summarization: techniques, challenges, opportunities. Expert Systems, 42(4), p.e13833.
2. Azam, M., Khalid, S., Almutairi, S., Khattak, H.A., Namoun, A., Ali, A. and Bilal, H.S.M., 2025. Current trends and advances in extractive text summarization: A comprehensive review. IEEE Access, 13, pp.28150-28166.
3. Giarelis, N., Mastrokostas, C. and Karacapilidis, N., 2023. Abstractive vs. extractive summarization: An experimental review. Applied Sciences, 13(13), p.7620.

4. Raiaan, M.A.K., Mukta, M.S.H., Fatema, K., Fahad, N.M., Sakib, S., Mim, M.M.J., Ahmad, J., Ali, M.E. and Azam, S., 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. IEEE access, 12, pp.26839-26874.
5. Koh, H.Y., Ju, J., Liu, M. and Pan, S., 2022. An empirical survey on long document summarization: Datasets, models, and metrics. ACM computing surveys, 55(8), pp.1-35.
6. Zhang, H., Yu, P.S. and Zhang, J., 2025. A systematic survey of text summarization: From statistical methods to large language models. ACM Computing Surveys, 57(11), pp.1-41.
7. Yadav, A.K., Ranvijay, Yadav, R.S. and Maurya, A.K., 2023. State-of-the-art approach to extractive text summarization: a comprehensive review. Multimedia Tools and Applications, 82(19), pp.29135-29197.
8. Kumar, S. and Solanki, A., 2023. An abstractive text summarization technique using transformer model with self-attention mechanism. Neural Computing and Applications, 35(25), pp.18603-18622.
9. Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K. and Hashimoto, T.B., 2024. Benchmarking large language models for news summarization. Transactions of the Association for Computational Linguistics, 12, pp.39-57.
10. Sanchan, N., 2024. Comparative study on automated reference summary generation using BERT models and ROUGE score assessment. Journal of Current Science and Technology, 14(2), pp.26-26.
11. Karunarathna, K.M.G.S. and Rupasingha, R.A.H.M., 2022. Learning to use normalization techniques for preprocessing and classification of text documents. International Journal of Multidisciplinary Studies, 9(2), pp.69-81.
12. Azhar, M., Amjad, A., Dewi, D.A. and Kasim, S., 2025. A Systematic Review and Experimental Evaluation of Classical and Transformer-Based Models for Urdu Abstractive Text Summarization. Information, 16(9).
13. Johnson, R.W., 2022. Alternate forms of the one-way ANOVA F and Kruskal–Wallis test statistics. Journal of Statistics and Data Science Education, 30(1), pp.82-85.
14. Šinik, B., Vake, D., Vičič, J. and Tošić, A., 2024. Interactive Tool for Tracking Open-source Artificial Intelligence Progress on Hugging Face. In 27th International Multiconference Information Society (IS 2024), Data Mining and Data Warehouses (SiKDD). Jožef Stefan Institute, Ljubljana, Slovenia (pp. 1-4).
15. González-Estrada, E., Villaseñor, J.A. and Acosta-Pech, R., 2022. Shapiro-Wilk test for multivariate skew-normality. Computational Statistics, 37(4), pp.1985-2001.
16. Okoye, K. and Hosseini, S., 2024. Analysis of variance (ANOVA) in R: one-way and two-way ANOVA. In R programming: statistical data analysis in research (pp. 187-209). Singapore: Springer Nature Singapore.
17. Okoye, K. and Hosseini, S., 2024. Mann–Whitney U test and Kruskal–Wallis H test statistics in R. In R programming: Statistical data analysis in research (pp. 225-246). Singapore: Springer Nature Singapore.