(Online) ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 14, No. 4, Oct. 2026, pp.87-114

Scaling transformer-based abstractive multi-document summarization to Arabic language

Amer Al-Dulame, Mohsen Kahani, Mohamed Allahbakhsh
Department of Computer Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Iran

Corresponding author E-mail: kahani@um.ac.ir

ABSTRACT

The abstractive multi-document summarization (AMDS) of the Arabic language is a daunting task because of the morphological richness of its language, the syntactic diversity, and the lack of task-specific benchmarks. We present Ara25, a Modern Standard Arabic multi-document dataset, built upon more than a hundred news sources, and divided into 700+ real-world events, in this work, wherein each event contains several related articles, each with a single human-written abstractive description. Based on this resource, we construct and comparatively analyze five transformer-based summarizers, mT5, Longformer EncoderDecoders (LED), AraT5, AarBERT and multilingual BERT, within a common long-context fine-tuning protocol that encodes and decodes multi-article inputs with explicit article-boundary markers. We include Arabic conscious preprocessing and a full evaluation suite, consisting of ROUGE, BLEU, METEOR, chrF, BERTScore, MoverScore and human judgments of fluency, coherence, faithfulness, and coverage. According to experimental results, long-context abstractive models, especially those initialized with an XLSum-tuned checkpoint, significantly outperform encoder-decoder models on short contexts with respect to syntactic consistency and semantic adequacy, making them the best current baselines to date on Arabic AMDS on Ara25. The combination of the dataset, modeling recipe and evaluation protocol gives a realistic benchmark and a reproducible reference point to future study on the abstractive multi-document summarization in low-resource languages.

Keywords: Arabic multi-document summarization, mT5, multilingual pretraining, SentencePiece, subword tokenization, long-context inputs, Arabic news dataset, cross-document summarization.

1. Introduction

The explosive nature of growth in the contents of digital texts has rendered the tracking of evolving stories in the outlets more challenging to readers, analysts, and decision-makers, which has further intensified the need of multi-document summarization (MDS) systems that can filter through a large number of sources and bring forward the most important stories [1]. Abstractive MDS, in which fluent summaries are produced by paraphrasing and reorganizing input instead of duplicating it, is useful in particular when stakeholders need one, consistent account based on heterogeneous reports and high-speed news events, in responding to crises, in policy briefs, or literature reviews where timeliness and synthesis are vital [2].

Although there has been significant advancements in English and other high-resource languages, Arabic continues to be under-served, a phenomenon not only due to the lack of data but also due to the language characteristics that make it harder to model, templatic morphology and cliticization, flexible word order, diacritic optionality, orthographic variation, rich derivational patterns, and high levels of dialectal diversity and

© The Author 2026. This work is licensed under a Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/) that allows others to share and adapt the material for any purpose (even commercially), in any medium with an acknowledgement of the work's authorship and initial publication in this journal.

Modern Standard Arabic [3]. These properties render tokenization, sentence segmentation, and cross-document coreference more susceptible, and increase the chance of error propagation in long-context generation. The confrontation of abstractive summarization of Arabic is thus twofold; to handle the syntactic discrepancy and discourse coherence across documents and retain the factual fidelity and stylistic naturalness.

This is further exacerbated by the fact that large, high-quality, task-specific annotations that encode cross-source agreement, contradiction, and temporal ordering information, which are the focus of multi-document synthesis, are seldom represented in current sources [4]. Architectures like BERT and T5 have revolutionized summarization in other languages, and Arabic versions including AraBERT, mT5 and AraT5 have started to bridge the divide. The majority of previous Arabic literature, however, focuses on single-document environments or extractive regimes, with little comparatively known about full-abstractive, multi-document generation [5, 6].

In this work, this gap is dealt with in three fronts. We start by presenting an Arabic abstractive multi-document summarization (AMDS) dataset of news clusters, each containing several articles and a human-written summary that is incentivized to reward grammatical well-formedness, cross-document coherence, and factual consistency. The data set represents real-life issues such as redundancy, salience trade-offs, cross-source event and entity coreferences, and inconsistencies that need to be addressed or cautiously worded in the final summary [7, 8].

Second, we leverage an adapted five transformer based pretrained models with Arabic-aware preprocessing: Unicode normalization of common orthographic variants, light punctuation harmonization, optional clitic-segmentation for more stable sub-word vocabularies, and document-level packing with explicit article boundary markers to aid the model's notion of discourse structure. We also explore long-context handling strategies to better model cross-document dependencies [9, 10].

Third, we employ an evaluation protocol that combines automatic metrics with Arabic-sensitive tokenization (e.g., ROUGE variants), complemented by targeted human evaluation for coherence, faithfulness, fluency, and coverage. Beyond headline scores, we report qualitative analyses of typical failure modes in Arabic AMDS hallucinated attributions, timeline drift, brittle handling of named entities and numerals, dialectal code-switching, and study the impact of our preprocessing and input-structuring choices through ablations. Taken together, these contributions advance the state of Arabic abstractive MDS by providing a realistic benchmark, a practical modeling recipe grounded in Arabic linguistic considerations, and a transparent evaluation suite [11, 12].

To the best of our knowledge, this work represents the first systematic attempt to investigate long-context fine-tuning for Arabic abstractive multi-document summarization (AMDS). While earlier studies have addressed Arabic summarization either at the single-document level (e.g., AraSum, AraBART) or through extractive multi-document pipelines (e.g., LDA- or TextRank-based methods), no prior research has directly explored abstractive MDAS under extended input contexts. Our contribution therefore lies not only in providing an abstractive model capable of handling heterogeneous multi-source Arabic news but also in empirically demonstrating that long-context optimization mitigates harmful truncation and improves recall-oriented metrics.

2. Literature review

Natural Language Processing (NLP) is one of the most important and viral applications of artificial intelligence in both industry and science. It, in short, uses computational techniques to automatically represent and process human languages. There are several NLP applications and tasks such as information extraction, machine translation, question answering, and text summarization. Nowadays, there are huge amounts of textual data, including online documents, articles, news, and reviews. They contain long strings of text that must be summarized to enable decision-makers to make decisions in a reasonable time. Text summarization, in general, is the process of summarizing a single article or a set of related ones by summing up the most important ideas, making sure the ideas sequence is coherent by ordering them chronologically [13, 14].

Text summarization has been around for more than fifty years, and there is a thriving research community in this field. To improve the quality of the output summary, researchers continue to enhance the performance of existing text summarization approaches or suggest brand-new approaches. The output of the most recent text summarization models (specially Arabic) is still only moderate level. Anyway, text summarization systems can be classified in several ways Here we going to introduce the most common classifications of automatic text (Figure 1)

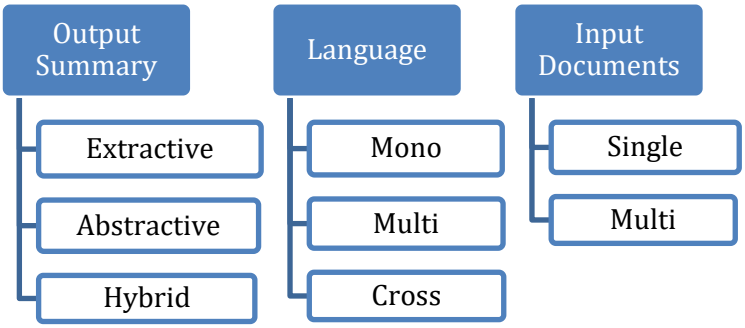


Figure 1. The most common summarization systems classification

Based on the output summary Approach, extractive, Abstractive, or Hybrid [13]. In The extractive text summarization approach, the system selects the important sentences from the input document and rewrites them in the same order as the original text [11][12]. The abstractive text summarization approach represents the input document in an intermediate representation, and the output summary is generated from this representation [10][11]. The hybrid text summarization approach merges both the extractive and abstractive approaches [16, 17].

Monolingual, Multilingual, or Cross-Lingual. A summarization system is monolingual when the language of the source and target documents is the same [18, 19]. A summarization system is multi-lingual when the source text is written in a number of languages (for example, English, Arabic, and French), and the summary is also generated in these languages. A summarization system is cross-lingual when the source text is written in one language (for example, English), and the summary is generated [9][11].

Based on the Input Type, single-document and multi-document [8]. Single-Document Summarization uses a single text document to generate a summary, and the target is to shorten the input document while keeping the important information [20, 21]. In Multi-Document Summarization, a set of input documents and the target is to remove repetitive content in the input documents [22, 23]. Multi-document is more complex than Single-document and has some prominent issues like redundancy, coverage, temporal relatedness, compression ratio, etc. [7][11].

Arabic multi-document summarization (MDS) seeks to distill a coherent single summary from clusters of related Arabic texts, most commonly news [24, 25], while balancing topical coverage, non-redundancy, and fluency under rich morphology, clitics, and stylistic variation; early Arabic MDS established clustering-driven extractive baselines that group sentences by topical similarity and prune duplicates on Arabic adaptations of DUC-style corpora, demonstrating that simple cluster-then-select pipelines can yield competitive summaries for this language family [1, 26]. Subsequent Arabic work extended those results to show that explicit optimization of salience in addition to redundancy yields better coverage at minimal length inflation, and that measures of recall improve when selection pressure is balanced with diversity [27, 28].

Hybrids that are topic-aware, e.g., fuzzy C-means + LDA, also assisted in aligning the sentence clusters with the latent event structure in multi-article contexts, particularly where stories grow over time, and which share named entities; cross-source consolidation was also achieved effectively by graph-based methods that combine PageRank with distributed embeddings [29, 30]. In the abstractive space, Arabic modeling with Transformer pretraining and Arabic-aware tokenization has achieved more progress on summarization tasks, and offers

powerful decoders that can be combined with extractive preselection to combine evidence across documents and paraphrase fluently in Modern Standard Arabic an increasingly popular design that uses extractive signals to plan contents and Transformer generators to generate surfaces [31, 32].

In addition to Arabic-only work, coarse-to-fine, multi-granularity models based on the rest of the MDS literature suggest initially encoding document-scale interactions and then refining to sentence- and phrase-scale relations; these structures can be easily adapted to the case of Arabic using morphology-sensitive preprocessing, length control, and decoding constraints (e.g., coverage penalty, repetition control) to cope with Together, the trend in Arabic MDS is towards hybrid schemes that (i) cluster and rank to cover and de-duplicate [1,28], (ii) inject topic and graph structure to handle cross-source agreement and contradiction [17,29], and (iii) use Transformer-based abstraction to paraphrase reliably and faithfully [32], and (iv) apply multi-granularity reasoning

Arabic text summarization has received a long history of interest as Arabic text spreads into newsrooms, platforms, and archives, and the methodological space has expanded correspondingly between the initial neural baselines and hybrid pipelines and cross-lingual transfer, but still lacks significant gaps to fully abstractive, multi-document scenarios. On the neural front, [9] author provided one of the first Arabic Single-document abstractive summarization sequence-to-sequence systems, summarizing the formulation of the problem, and revealing the issue of exposure bias, the lack of coverage of salient content, and the shortage of the long-input resistance. Their data is constructed on the basis of single Arabic news stories along with their headlines. Their task is more or less headline generation based on a single article.

Graph-based centrality algorithms that are central to Arabic offer simple, efficient baselines to sentence selection, clustering with distributional semantics, extractive summarization of Arabic, where the key of the approach is the cautious use of linguistic preprocessing and morphological decoders to enhance the ranking of sentences. Their work is unique in that it uses noun-based ranking with cosine similarity to capture grammatical significance and semantic similarity, and also tries alternative morphological analyzers to more effectively cope with the Arabic complexities. Such a combination of more profound language characteristics demonstrates evident extractive summarization advantages. The approach was found to be good in the single-document summarization task, and it performed well on the EASC (Essex Arabic Summaries Corpus) which has 153 documents.

The researchers of [2] presented multi-document Arabic summarization, which is unsupervised, one step further than those of a single text. They used Word2Vec embeddings and k-means clustering to fight redundancy followed by the utilization of the Weighted Principal Component Analysis (W-PCA) to rank sentences more sensibly. Their system performs a layered selection: first clustering related documents, then isolating key sentences based on semantic similarity and weighted importance via W-PCA, no new sentences are generated, so it's extractive text summarization. Analysis with the Essex Arabic Summaries Corpus (EASC) with ROUGE-1 and ROUGE-2 showed encouraging progress in comparison with previous algorithms.

To close that rift, hybrid pipelines have incorporated structured semantics in systems that are learning-centric. In [4], the authors suggested an innovative hybrid design of single-document Arabic summarization, which integrates the advantages of both statistical-based and semantic-based methods in a graph-based design. Their approach is notable due to its application of hierarchical graph based on ontology, which integrates classic statistical properties with more profound semantic links, which are more effective in capturing interdependence between words. The system scored well on the EASC (Essex Arabic Summaries Corpus), and there was significant improvement in ROUGE-1 scores.

In [26], the authors explored the abstractive Arabic text summarization by fine-tuning two state-of-the-art pretrained transformers, mT5 and AraBART. Both models were modified to produce consistent summaries of Arabic texts, and the task was set as a single-document summarization. Article-headline pairs of the Arabic Gigaword dataset were used to run the experiments and measure the performance in terms of the typical ROUGE

metrics. The findings indicated mT5 and AraBART generated competitive summaries, indicating the power of transformer-based pretraining to learn semantic and contextual subtleties over previous neural processes.

The authors of [32] published SumArabic, an abstractive Arabic news dataset, consisting of a single document, constructed on Common Crawl pages of Emarat Al Youm and AlMamlaka TV, to assess Arabic abstractive summarization using RNN-based and transformer-based models, comparing RNN and Transformer configurations (including pretrained models) to Arabic abstractive summarization, and measure them on ROUGE-.

In [30], the authors introduced GEATS, a single-document, extractive Arabic summarizer, which constructs a sentence graph based on cosine similarity of Word2Vec embeddings and ranks sentences using PageRank; the Farasa stemmer is used to preprocess.. The system is evaluated on the EASC corpus (153 docs; five human summaries per document based on Wikipedia, Alwatan and Alrai) and evaluated in terms of precision, recall, F-measure, and also reports ROUGE-N and BLEU formulations; the system obtains competitive scores at 40 percent compression and even scores above the nearest baseline by more than 7.5 percent F-measure. GEATS, in general, exhibits a robust graph-plus-embeddings recipe to the Arabic extracted summarization.

The authors of [23] introduced AraTSum, a pipeline for Arabic Twitter trend summarization that treats each trend as a multi-document (multi-tweet) extractive task: they first discover topics with LDA/NMF, classify tweet sentiment with a modified AraBERT model, extract aspects via TF-IDF, then cluster by topic × sentiment and select one representative tweet per cluster to form the final summary (typically 10 tweets). They evaluate on five Arabic Twitter datasets spanning different dialects, using ROUGE against human (manual) summaries, and report gains over baselines; code/data are linked in their repository.

An example of this work is Al-Taani and Al-Sayadi, who fuse fuzzy C-means clustering and Latent Dirichlet Allocation (LDA) to sample the collection of related articles based on topic, and then ranking the sentences within each topic cluster: experiments on TAC-2011 Arabic news demonstrate that the two-step topic→selection pipeline is effective at surfacing

A second strand addresses social media MDS based on AraTSum in [23] summarises Arabic Twitter trends, extracting topics/aspects and using AraBERT-based sentiment to create topicpolarity clusters, and then sample representative posts; this study identifies the phenomenon of aspect- and stance-conscious selection as a means to balance coverage in high-noise, high-velocity social streams.

The authors of [24] created thematically segmented documents with LDA-driven annotation, build per-topic graphs, and use a PageRank-like centrality measure to eliminate redundancy and select high-value sentences, which further supports the trend of segment-then-rank architecture in Arabic news.

Being an influential baseline on the boundary of the five-year window, the authors of [2] suggested a clustering- and Word2Vec-led framework (with further dimensionality reduction/feature weighting) to regulate redundancy and enhance ranking in Arabic multi-document context; though old, it is also actively used in later surveys on Arabic MDS. Conversely, truly abstractive Arabic MDS is still uncommon: beyond isolated experiments (such as a textual-graph based, so-called minimal language-dependent, model that fuses related sentences together using graph traversal rules), most recent Arabic attempts do not go all the way to fuse-generation across documents; modern abstractive work in Arabic is generally single-document as opposed to multi-document.

Studies shifted to longer sequence-to-sequence and graph extractor models and newer transformer models with longer context, although we have yet to have a robust, fully abstractive Arabic MDAS that de-duplicates, orders events, and maintains entities and numbers. It requires improved Arabic multi-doc data with multiple human annotations, wider Arabic pretraining, effective fine-tuning, and tests that verify cross-source coherence and factual grounding, not word overlap.

In order to have a closer look the accuracy of automatic text summarization depends on the data collection, the size of the text and the number of sentences. The Arabic language, however, lacks benchmarks as the rest of the

languages and the majority of the data sets that are available on Arabic are in an unfortunate state. Arabic standard datasets) (summarized in Table 1) include text consisting of short sentences, inadequate to produce a true summary.

Table 1. Comparative standard datasets for Arabic text summarization

Developer	Year ↓	Dataset Name	Number of Documents	Summary Type	Summary Length
Our Dataset	2025	Ara25	Multi-Document	Abstractive	Multi-Sentence
[3]	2023	SumArabic	Single-Document	Abstractive	Single-Sentence
[21]	2013	KALIMAT	Single-Document	Extractive	Multi-sentence
[27]	2010	EASC	Single-Document	Extractive	Multi-sentence

SumArabic [26] is a single-document abstractive dataset focused on generating single-sentence summaries from news sources, files of the articles of the following two Arabic news websites: www.maratalyoum.com and www.almamlakatv.com. KALIMAT [21] stands for extractive dataset for single or multiple documents. It contains 20,291 single documents for texts with summaries and another 2,057 for multi-documents. The data is news documents [34]. (EASC) [33] stands for popular dataset for single-document extractive summarization research. It has 150 articles and 765 human-generated summaries of those articles. (DUC2004) [36] is a dataset for abstractive single-document summarization. It contained 100 Arabic news articles with four human-written summaries apiece. Each cluster has 10 documents on average. However, the summary is written in English, whereas the text is written in Arabic. These datasets are usually used to evaluate the ATS systems, because they don't have enough data to train neural network models. Other datasets, (OSAC) Popular dataset for single document abstractive summarization. It contains 22,429 news documents. (SANAD) is a large collection of Arabic news articles and can be used for single-document extractive summarization. (The RTAnews Dataset) is a collection of multi-label Arabic texts, collected from Russia Today in the Arabic news for single-document extractive summarization. (NADA) contains a set of newswire texts which are taken from two existing corpora, OSAC, and DAA, for single-document abstractive summarization. (Multi-Document Summaries Corpora) It can be used for generic extractive summarization of both Arabic and English multi-documents. They translated the English gold-standard summary into Arabic by using Google translate [28]. (Xl_Sum) Single document dataset for abstractive summarization. This dataset was created from the BBC website's news using a set of well-designed algorithms. The dataset contains 44 languages. (WikiHow) Contains 18 languages. It includes Arabic texts and a summary of a single abstractive document. (AHS) Dataset can be used for a single document for abstractive summarization. It contains 300k texts and their titles without any summary. Consider the title as a summary of it. (AMN) Dataset can be used for a single document for abstractive summarization. It contains 265 k news texts and their summaries [28].

2. Methodology

The present study is based on a methodology framework that merges the existing practices in terms of data preparation, model adjustment, and evaluation with recent developments in long-context transformers.Our initial design decision was to favor realism over excessive curation: we do not restrict our experiments to a few newsrooms with homogenous style, but chose to sample a broadly with a corpus that reflects what real readers see on the Arabic web, heterogeneous prose, and uneven structure of articles and multiple editorial voices covering the same story with slightly different angles. The overall goal, then, was twofold; to build a high-quality dataset with a rich enough linguistic and topical coverage to allow strong generalization, and to do so in

a manner that is easy to reconstruct or expand by other researchers who may have an interest in stress-testing their models under similar conditions.

At the instance level, we consider a multi-document event to be a collection of news stories, which give a clear description of the same underlying real-world event (say, the same attack, election result, court ruling or diplomatic move, over a short time frame). Beginning with the raw crawl, we sorted articles into candidate events based on headline hints, publication dates, and topical similarity and analyzed ambiguous cases by hand to eliminate articles where only shared background topics were present without reporting on the identical concrete event. This cleaning gives 715 multi-article events and 2,266 source articles, which have a mean/median of 3.17 / 3.00 articles per event. The complete article texts are stored in each event record, and their headlines, thus retaining sufficient context to re-group them in the event that other researchers may re-assemble or extend Ara25 (Our dataset name is Arabic language and year of implementation 2025).

In each event, a human annotator reads all the articles that make the event and then composes a single abstractive summary in the Modern Standard Arabic which indicates the aggregated perception of the story. Summaries are written with transparent in-house policies that place emphasis on (i) fluency in MSA, (ii) coherence and logical flow of ideas across multiple points of view, (iii) faithfulness to the original reports (no hallucinated actors, places or numbers) and (iv) moderate compression over telegraphy. The annotators are also asked to synthesize shared information among the articles and also to de- duplicate duplicated sentences but not to verbatim copy the information unless it is of vital named entity or brief quote. To act as annotators of our dataset, we selected five native Arabic speakers, with different educational levels, including undergraduate studies up to a PhD in Arabic language and literature. Their main responsibility was to write the reference human summaries on the training set, and the test set. In each instance, they read all the source articles in the cluster, and they generated one abstractive summary as a group that summarizes the main facts, the storyline, and the most important details of the context, instead of simply copying or slightly modifying the sentences of the inputs. This process resulted in that the reference summaries adhere to a logical editorial style and a similar degree of informative content across splits, and are indicative of the knowledge base of the annotators in Arabic language, discourse organization and news-writing standards. By so doing, every reference summary is the synthesis of multiple documents, as opposed to the rewritings of one particular source article, and this gives it a suitable gold standard to abstractive multi-document models.

In order to ensure quality, we conducted light weight systematic checks during construction. Random spot-checks (around 10 percent of the events in each crawl window) were conducted to ensure that grouping, normalization, and deduplication were functioning as desired and that the human summaries were accurate to the articles being grouped. Clearly unsuccessful efforts, like broken markup, wire copy, or summaries that strayed off the point, were either fixed or deleted. We used publicly available news websites as sources of all content, and we stored canonical URLs and timestamps to allow downstream consumers to trace individual items to their origin and abide by the terms of the publisher; where we were required to restrict distribution to metadata and scripts that re-fetch the articles.

Altogether, these design decisions render Aar25 an intentional tough challenge to Arabic multi-document abstractive summarization. Input is represented as a number of heterogeneous articles per event, authored by different newsrooms with different styles, and joined with clear article-boundary hints, compelling models to do cross-document aggregation as opposed to compressing documents individually. The tokenization statistics indicate that merged inputs are usually multiple thousands of subword tokens and may even reach 6k tokens in size, whereas reference summaries are typically more than 10 times shorter with an average compression ratio of about 3-3.5x. Models should thus be able to deal with long-context reasoning, solve small contradictions between sources, eliminate redundancy, and yet generate fluent, concise MSA, in a morphologically rich language with flexible word order. It is this combination of long inputs, many viewpoints, and non-trivially compressed that we wish modern transformer architectures to be tested on.

The rich morphology and orthographic variability of Arabic pose familiar but non-trivial preprocessing problems. We want to apply sufficient normalization to ensure that learning is consistent across models, but still maintain the natural variability that systems will encounter at inference time. To this end, we use one, model-agnostic preprocessing pipeline, the results of which are subsequently used by all five summarizers (mT5, LED, AraT5, AraBERT, mBERT).

We delete non-linguistic artefacts first which include HTML scaffolding, boilerplate navigation text, ad remnants, inline scripts, tracking fragments, hyperlinks, email addresses. We then regularize orthography, eliminating diacritics (infrequently found in headlines and captions), tatweel and harmonizing common hamza / alif patterns (ا → أ/إ/آ) to make sparsity less sparse without merging truly different forms. Multidigits are normalized through a mapping of Arabic-indic digits to Western digits, making downstream tokenization easier and preventing the sparsity of mixed headlines.

Dialectal material is approached conservatively. In cases where dialectal tokens can be found in otherwise MSA-style reports and where the contrast is simply a matter of orthography (such as simple clitic spellings), we normalise to an MSA-like form; in cases where the contrast bears lexical or stylistic content, we do not over-edit the actual usage. This is especially necessary in opinion pieces and quotes within news articles, where dialect may indicate voice or stance.

Importantly, we adhere to a principle of realism: we do not use aggressive content filters that would artificially sterilize the input (e.g. exclamation marks, small sub-titles or emphasis marks are banned). Rather, we include such phenomena in such a way that models can be subjected to the same edge cases that practitioners regularly see at deployment time, noisy punctuation, sub-headings within the article body, and mixed-script named entities.

The corpus keeps all the events after preprocessing (more than 700 multi-article events) to be used to model. Aggregate input statistics, calculated with the mT5 SentencePiece tokenizer, are that the average input length is in the range of hundreds of subword tokens (meaning around 270 on average, with a median a bit smaller), and that the largest events go up to the low-thousands of tokens and create a small long tail of very detailed stories. The average length of human-written summaries of references (one per event, written after reading all the articles that form it) is about three times shorter, with an empirical compression ratio of the ≈3.1–3.4× range. We do some basic quality control: on a random sample of events per crawl window (approximately 10 percent), we spot-check date extraction, paragraph delimitting, and deduplication, and in the few cases where problems are found, we make corrections or drop the problem item.

Since our experimental design involves five transformers, we have to trade comparability with model-specific tokenizers. To analyze data and make length-sensitive training choices, we use the canonical reference of the SentencePiece tokenizer of mT5, as it operates with subword units that are compositional and trained to work with multilingual news tasks. This is especially significant with the Arabic language, in which the cliticization, active affixation and a high frequency of numeral, named entity and foreign string mixing (e.g. 19-كوفيد, F-16, etc.) would otherwise run the vocabulary out of hand.

Practically, mT5, LED, and AraT5 can use SentencePiece-style subword vocabularies that have the same compositional benefits, whereas AraBERT and mBERT use WordPiece tokenizers that are trained on Arabic or multilingual data. The reason is that we avoid adding an extra morphological segmenter (e.g. Farasa-style pre-segmentation) to make preprocessing straightforward, model-neutral and similar among all five systems; rather, we use the subword tokenizers as-is to deal with inflectional and clitic variation.

The tokenization under the mT5 tokenizer is to an average of approximately a few hundred subword tokens per event with a tight central mass and a long-tail mass of events which stretches well over a thousand tokens. Reference summaries are roughly 1/3 the length of the input, which is in line with the compression ratios in the above. In addition to their descriptive utility, such statistics are used in a number of training decisions typically shared by all models, including maximum input lengths, gradient accumulation options (to allow long sequences

occasionally), and generation length limits optimally set to the observed compression range. Notably, we do not artificially cut the sources to cause a specific length distribution rather we simply impose large limits on each architecture (e.g., long-context windows in mT5/LED and larger hierarchical buffers in AraBERT/mBERT) and have the natural variation in Ara25 push the learning signal.

After normalization and tokenization, the data was split into training and testing events at the event level: 90% of the total events were used to build the training set, and 10% of the total events were used to build the test set, which resulted in around 648 / 72 events in the training and test sets respectively. The lengths of the input (number of characters and tokens) are distributed similarly in each split to the global corpus and the length distribution summary is similarly stationary, which is favourable to tuning generation parameters on the validation set without accidentally over-fitting to a non-representative subset. In contrast to some previous literature that builds a special out-of-domain (OOD) split by clustering the sources by topic, we decided, since we already have a wide range of sources covered, to maintain the topical heterogeneity within the standard splits and leave the OOD analysis to later work; this choice allows the current research to remain targeted to the main modeling question without further assumption about topic boundaries.

Transformer models have been demonstrated to be useful in single- and multi-document summarization (e.g., BART, PEGASUS, mT5). As Reda et al. (2022) and Elsaid et al. (2023), we follow a sequence-to-sequence design, namely fine-tuning mT5, as it has an efficient SentencePiece vocabulary in the Arabic language and the capability to work with multilingual environments. In order to overcome the input length constraint of previous MDAS studies (Li et al., 2020; Wang et al., 2022), we explore long-context fine-tuning to a maximum of 6,144 tokens, enabling the model to learn more cross-document dependencies without being adversely truncated.

Though long-context models or Arabic-specialized models might, on the face of it, seem the correct solution to multi-document abstractive summarization, our exploratory experiments indicated that a practical trade-off is determined not so much by architectural slogans as by a delicate balance between (a) the nature of our data, with its stylistic heterogeneity, topical diversity, and deliberately realistic textual noise, (b) the The use of LED architecture in the context of long-term context shortly caused cognitive overload, with learning plateaus. As a result, automatic measures did not significantly change and no credible indicators of meaningful change were observed. Such stagnation seems to be a result of compound factors: that is, the need to have carefully designed global-attention patterns and anchors in merged inputs to be able to successfully exploit the increased context window. [1][2]; mismatch between the pretraining (which is mostly on texts/languages not reflective of Arabic morphology or the noisiness of real-world news); and high computational cost that requires severe tradeoffs in batch size and gradient accumulation, resulting in a insufficient number of effective updates to overcome the initial stalling of the loss.

AarT5, which is Arabic-trained, acted in a similar way of non-learning in our environment. The most likely reason is that, being linguistically compatible, it is not always tuned to multi-source abstractive summarization (which requires a lot of evidence integration across temporally closely related reports); in combination with context constraints imposed by memory, this forced us to extreme trade-offs between length of input and training stability, flattening curves and loss of gain on reference measures.

In our case, with the Wavelet variant we tested, it appeared unsuited to the task, on two counts: its architecture, in the form we tested it, was not particularly willing to demonstrate strong compositional paraphrasing across multiple documents compared to existing sequence-to-sequence generators, and its sensitivity to fine-tuning hyperparameters made the range of reliable operation very narrow, with results collapsing to near-zero ROUGE/B By contrast, mT5 offered a practical blend of three decisive advantages: (1) broad multilingual pretraining that includes Arabic within a SentencePiece framework without over-zealous normalization assumptions, reducing distributional gaps when moving from stylistically varied news to representation space without paying the cost of source-specific linguistic engineering [17][29]; (2) a robust subword scheme that handles Arabic clitics and derivation well, shrinking vocabulary fragmentation and keeping semantic content accessible to the model through compositional subunits, observed in practice as faster loss stabilization and

reduced stalling relative to other models [17]; and (3) stable, fine-grained tunability under realistic compute (Adafactor, bf16/FP16, gradient accumulation, checkpointing), allowing us to maintain input lengths suitable for our case, where a merged event input can reach up to 6,144 tokens, without fully sacrificing batch size or the number of effective update steps [31][32][20].

This in turn allowed the model to observe enough diversity at each epoch to be able to learn a uniform compression policy at about 3x average. Besides, the fact that mT5 was paired with the properties of our dataset, where we intentionally maintained the "real-world texture of news and did not over-sanitize it, was particularly useful: models with the assumptions of clean and in-domain text are likely to trip on mixed numeral encodings, minor dialectal differences, or other atypical editorial decisions, but mT5 showed more structural tolerance to such subtle irregular

In practice, where other models had not even demonstrated any statistically significant improvement, near-zero metrics, non-improving loss, and non-summarizing outputs, mT5 and LED had reached a consistent level of learning and started generating concise, coherent summaries to track event timelines and project high-information-density phrases across documents, which is exactly the kind of behavior desired in real-life MDAS. Overall, the decision to use mT5 was an experiment-based, methodological choice, rather than a strictly theoretical one: its multilingual pretraining, subword strength, and fine-tuning stability were both appropriate based on our computational resources and the lengthy and mixed nature of our inputs, whereas AraBERT, AraT5, and BERT, as implemented in our experiments, could not pass the initial learning bottleneck of this

In line with the tradition of finetuning with transformers, we use methods like Adafactor with warm-up scheduling, label smoothing, and gradient accumulation. Gradient checkpointing and mixed-precision training make training within constraints of GPU memory viable, akin to recent Arabic transformer work. Early validation loss stopping avoids overtraining, and increases the generalizability of the model across a variety of event clusters.

Similar to the previous Arabic summarization research (El-Gaaly et al., 2021; Elsaid et al., 2023), the assessment is done using both automatic and human-centered metrics. ROUGE, BLEU, chrF, MOVER-Score, BERT-Score, and METEOR are the automatic measures intended to be interpreted as indicative of success because of the morphological variability of Arabic and the abstractive nature of the task. Human evaluation supplements these metrics with an evaluation of fluency, coherence, faithfulness and coverage, which surface-level n-gram matching commonly underrepresents.

This algorithm puts our work between earlier extractive Arabic multi-document systems and more modern abstractive, transformer-based single-document (e.g. AraBART, AarT5) systems. The amalgamation of Arabic-conscious preprocessing, long-context input structuring, and stringent multi-metric evaluation places the work on the path towards a viable framework of abstractive Arabic MDAS that is linguistically and technically scalable.

We will openly publish all the artifacts of this work,dataset, code and trained models under the copyright and licensing terms of the news sources they were trained on, to maximize reproducibility and downstream reuse. The release includes three parts (i) the Aar25 dataset, (ii) modeling and evaluation code, and (iii) trained checkpoints and configuration files.

Ara25 dataset. We will make Aar25 available in the form of JSON/JSONL, with the entire train/test splits at the event level, the packed multi-article inputs with article-boundary marks, and the human-written abstractive summaries. In the case of the 72-event test split, we will have independent reference summaries done on each event to facilitate multi-reference assessment. A brief data card will be included with the dataset that provides information on coverage of the source, preprocessing choices, length statistics, and identified shortcomings.

Code and experimental pipeline. The entire preprocessing, training, and evaluation pipeline will be published as a series of Python scripts and Jupyter notebooks, and these match the implementation on Google Colab that

was used in this study. This repository will contain: (i) scripts to normalize and pack data, (ii) training scripts of all five models (mT5, LED, AraT5, AraBERT, mBERT) with their hyperparameters, (iii) evaluation scripts of ROUGE, BLEU, chrF, METEOR, BERTScore, and MoverScore, and (iv) the scripts to perform decoding ablations and An environment setup (library version, required GPUs), stepwise instructions to reproduce the main tables and figures, and instructions to add new events to Aar25 will be documented in a top-level README.

Long-term access, models, and configurations. In the case of the most effective system (long-context mT5 starting with the XLSum checkpoint), we will release the fine-tuned model weights, tokenizer, and the exact generation_config with which we report headline results. We will also offer LED checkpoints where possible to complementary long-context baselines. Artifacts will be versioned, and stored on a stable public platform (e.g., GitHub + a model/dataset hub with DOI-backed snapshots) to make them permanently accessible. In anonymous review the reviewers will be given an anonymized repository link; when acceptance is received the repository will be de-anonymized and the final DOIs and URLs will be inserted into the camera-ready version of the paper.

3. Experiment

We construct a Modern Standard Arabic (MSA) multi-document summarization dataset (Ara25) by aggregating news coverage of the same event from 100+ official Arabic news outlets, yielding 715 multi-article items (events) and 2,266 source articles (mean/median 3.17/3.00 articles per event). Each item pairs a packed multi-document input with a single human-written abstractive summary; 713 items contain both sources and references. A light language-quality screen flags only 2 low-Arabic summaries, and we remove exact and near-duplicate titles (63 exact groups; 92 near-duplicate clusters) to reduce topical leakage. Inputs are concatenated with explicit markers <ARTICLE_i> and a task prefix ("لخص بالعربية الفصحى بدقة واختصار ووضوح: …").

Tokenization analysis (SentencePiece) shows median/90th/95th/99th input lengths of 2,024/4,521/5,237/6,538 tokens, max 7,964; importantly, 97.76% of packed inputs fit within 6,144 tokens, which we adopt as our context ceiling. Reference summaries are much shorter (median 246, 90th 365, max 611 tokens). Figure 2 shows words distribution of input texts and output summary.

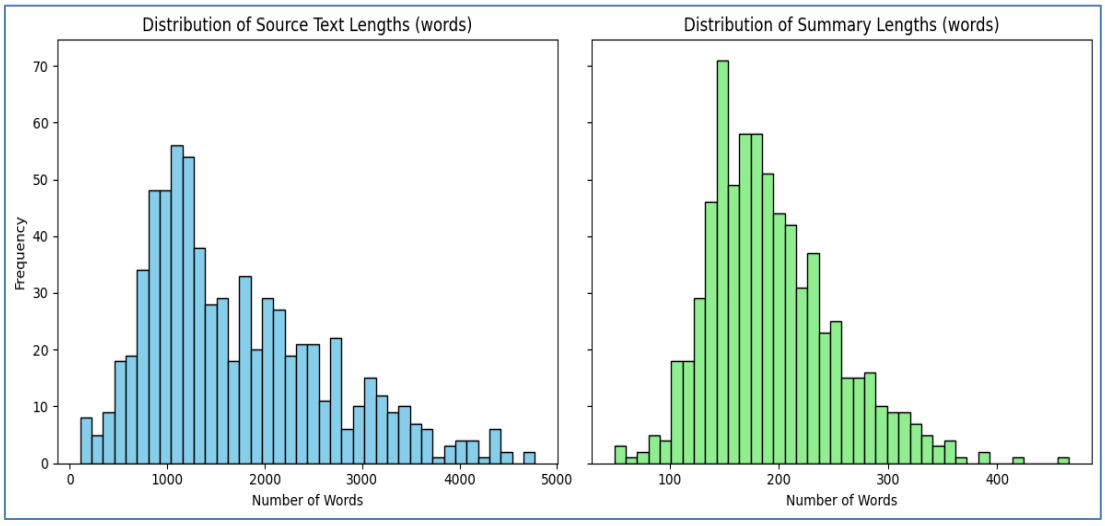


Figure 2. Words distribution of input texts and output summary

We fine-tune five pretrained transformer-based summarizers that jointly cover Arabic-specific and long-context architectures: (1) mT5, initialized from the XLSum-tuned checkpoint (csebuetnlp/mT5_multilingual_XLSum), (2) Longformer Encoder–Decoder (LED-base-16384) adapted for long-input summarization in Arabic, (3) AraT5 as an Arabic encoder–decoder baseline, (4) AraBERTv02 in a hierarchical encoder-only abstractive configuration, and (5) multilingual BERT (mBERT) in an analogous hierarchical setup. All models are trained

on Ara25, where each instance consists of several news articles describing the same event paired with a human-written Arabic summary.

For the encoder–decoder systems (mT5, LED, AraT5), the concatenated multi-article input is provided directly as a single long sequence. For AraBERT and mBERT, we employ a hierarchical encoder that first encodes article segments and then aggregates their representations before decoding. Training across models follows a shared protocol: a lightweight hyperparameter search over learning rate, batch size, and number of epochs; AdamW optimization with mixed precision where beneficial; and early stopping on a held-out split using ROUGE-Lsum as the selection criterion. At inference time we use beam search with length and repetition controls, and we evaluate all models under a matched decoding configuration unless otherwise noted. The overall multi-document abstractive pipeline is summarized procedurally in Algorithm 1 and depicted schematically in Figure 3.

Algorithm 1: Arabic Multi-Document Abstractive Summarization

Require: A set of news documents ($D = \{d_1, d_2, \ldots, d_m\}$) related to the same event
Ensure: Abstractive summary (SD)

1: Collect multi-source documents ***D***.
2: Preprocess each document: normalize Arabic text, remove noise, and tokenize.
3: Concatenate all documents into a single sequence ***M*** with segment markers
⟨ARTICLE$_1$⟩…⟨/ARTICLE$_1$⟩.
4: Load long-context transformer model M_θ.
5: Tokenize ***M*** with maximum context length ***L***.
6: Encode input sequence: H=Mθ.encode(M)H = M_{θ}.encode(M)H=Mθ.encode(M).
7: Apply long-context attention mechanism to capture cross-document relations.
8: Decode abstractive summary using fine-tuned decoder:
SD=Mθ.generate(H,decoding_params)SD = M_{θ}.generate(H, decoding_params)SD=Mθ .generate(H,decoding_params).
9: Post-process output summary: detokenize, normalize Arabic script, remove repetitions.
10: Return ***SD***.

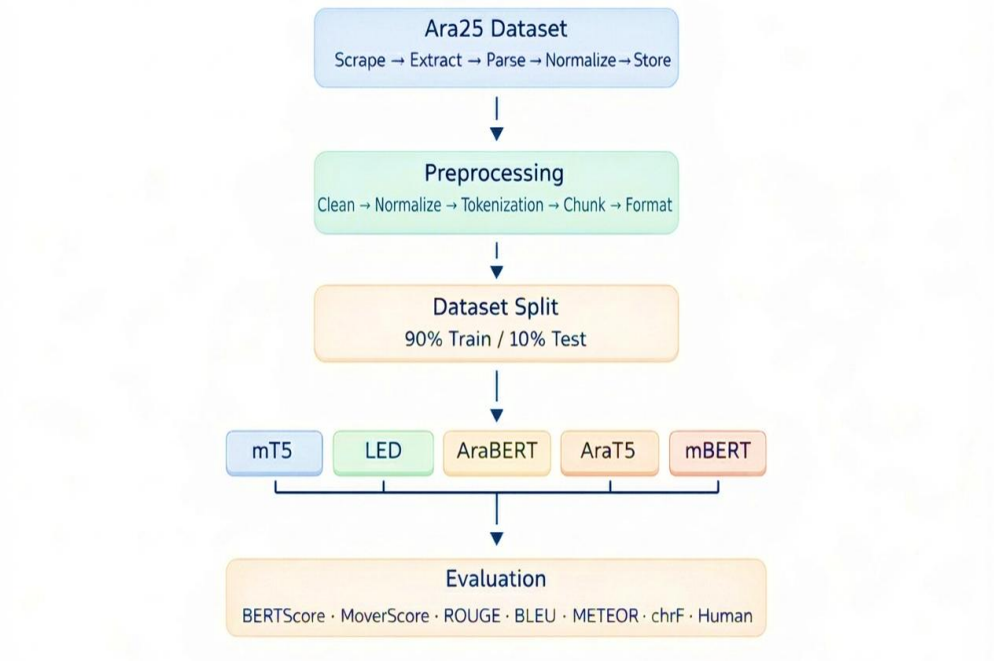


Figure 3. Schematic overview of the Ara25 training and evaluation pipeline

The Ara25 dataset is preprocessed and split into train and test sets, then used to fine-tune five models (mT5, LED, AraBERT, AraT5, mBERT), whose outputs are subsequently assessed with automatic and human evaluation.

The first system we evaluated was a baseline built on top of the publicly available AraT5-base model. We fine-tuned this model for eight full epochs without early stopping, using Adafactor with a 2e-5 learning rate and a cosine scheduler, on Google Colab with a single GPU in mixed precision (bf16). The input to the model was a long-merged document that concatenates multiple news articles about the same event, cleaned and standardized, with an explicit Arabic instruction prefix ("summarize in Modern Standard Arabic with accuracy, brevity, and clarity"). We allowed up to 6,144 tokens on the encoder side and 512 tokens on the decoder side, and accumulated gradients (2 × 16) to compensate for Colab memory limits. During training, the model appeared to converge in terms of loss, and validation loss decreased across epochs. However, when we evaluated the model on the held-out test set, the actual summarization quality was extremely weak: ROUGE-1, ROUGE-2, ROUGE-L, and BLEU were effectively zero, and METEOR was also close to zero. The produced summaries were relatively long (about 168 words on average), but they did not reliably capture the factual content of the reference summaries, and they exhibited very low lexical overlap with the human-written targets. In practice, the model tended to generate generic high-level statements rather than event-specific content. We believe this failure is caused by a combination of factors: (i) AraT5-base was not pre-trained for very long multi-document inputs, yet we forced it to consume multi-thousand-token contexts; (ii) the model was asked to perform high-level cross-article abstraction rather than sentence compression, which is outside its typical use; and (iii) the decoder tended to drift into fluent but ungrounded Arabic text that does not align with the gold summaries. The significance of this experiment was that it demonstrated a lower bound: naïvely fine-tuning a typical Arabic T5-style model on long, merged multi-source inputs does not necessarily result in usable multi-document summaries, even when the loss during training appears reasonable. The second one is multilingual BERT encoder-decoder (mBERT Abstractive Hier). This architecture is not a native seq2seq architecture; we directly constructed it by training bert-base-multilingual-cased as both encoder and decoder, and then make the decoder architecture cross-attention whereby the decoder BERT is able to attend to the encoder BERT. Google Colab was used to train in two steps. Phase A: We froze the encoder and only fine-tuned the decoder in two epochs, and the practical encoder context was 448 tokens. Phase B: Phase B involved unfreezing the encoder and training both sides an additional 6 epochs, with a maximum of 512 tokens as input to the encoder. When bf16 is available, we used gradient accumulation (2 × 16), cosine scheduling, label smoothing, and mixed precision. The validation loss steadily decreased throughout the epochs in both phases, indicating that the model was not going astray but was adapting to the task. The pipeline also had considerate Unicode-safe decoding, vocab-specific clipping of the vocab, and constraints on generation, like no repeat ngram size=3, and an optional hierarchical inference process, which breaks a long 6K-token event into windows, summarizes each window, and summarizes the summaries. But this hierarchical step was not allowed in final validation scoring and therefore, only single-pass generations were evaluated. Lexical overlap with the human summaries was effectively zero when we tested on the held-out test set: ROUGE-1/2/L and BLEU both reduced to 0.0, and METEOR was very low. The model generated outputs of approximately 129 words and in largely grammatical Modern Standard Arabic, but they read like generic background paragraphs on the topic as opposed to a focused event summary. However, semantic similarity scores like BERTScore (~7273 F1) or MoverScore (~0.95) had higher scores than AraT5, indicating the model is at least discussing the appropriate general topic. We interpret this to mean that this architecture can communicate about the story, yet it still can not generate an anchored, factual event summary that matches the human-written gold summaries. The key bottleneck is structural: mBERT has a natural max position embedding of only 512 tokens and thus, after unfreezing and joint training, the model does not actually look at the entire multi-document context (≈6K tokens). It only looks through a cut window and then fantasizes that there is some connecting prose that it is unable to perceive. The third of our variants substitutes multilingual BERT with an Arabic-specialized backbone: bert-base-arabertv02 as the encoder and decoder, and linked by cross-attention in a BERT2BERT EncoderDecoder through an architecture.

The training schedule is structurally identical to that of the multilingual model: the first two phases of training are a freeze-thaw cycle in Colab, where we freeze the encoder (Phase A) and optimize the decoder to speak in an abstractive, news-like voice, then thaw the encoder (Phase B) and train the entire stack together. We impose the same generation constraints as in the previous case (no repeat ngram size=3, 4-beam decoding, label smoothing, accumulation of gradients (2 x 16)) to live within memory constraints. The concept of AraBERT was straightforward: perhaps a monolingual Arabic model is more likely to comprehend Arabic news discourse and headlines, than a generic mBERT, and thus be able to generate more native-sounding, information-dense summaries. In Phase A we had weak yet non-zero ROUGE and BLEU on validation (e.g., ROUGE-1 of 0.02-0.03 in early epochs) which was an indication that the decoder, alone, was beginning to align with the target summaries. Once the encoder was unfreezed in Phase B, though, the behavior proved to be unstable: validation loss kept decreasing, but more often than not the lexical overlap metrics would revert to zero in subsequent epochs, and the improvements we observed during early steps were not always maintained. The resulting final AarBERT system had long outputs (approximately 285 words on average), written in grammatical Modern Standard Arabic, and was marginally more lexical and structural aligned than mBERT: ROUGE-1 best over refs was about 1.7 and METEOR and chrF were higher than pure mBERT, as well. Concurrently, BLEU and ROUGE-2 were also very low and the summaries continued to be largely high-level narratives as opposed to narrowly based event synopses. That is, the model sounds more like Arabic news than the multilingual baseline, yet it also has a hard time summarizing multiple articles into a narrow, factually accurate, event summary. We take this to indicate that monolingual Arabic pretraining assists stylistic fluency and register, but alone does not address two fundamental issues (i) the 512-token maximum context of BERT, which does not allow the model to effectively see a 6K-token multi-document input, and (ii) the challenge of learning to control abstraction in the context of an encoderdecoder model that has never been pre The fourth is a long-context abstractive summarizer that is based on allenai/led-base-16384 and modified to Arabic. Contrary to the BERT-style style models, LED is a real encoder-decoder model with sparse attention on very long documents, and is thus inherently multi-article in nature. We fine-tuned LED in Colab using a two-phase length curriculum. During Phase A, we simply trained the model by only providing inputs that are truncated at around 2K tokens (2,048 tokens) once, primarily because we wanted the decoder to learn how to generate concise Arabic summaries without depleting the attention budget at once.

In Phase B, we increased the maximum encoder length to 6,144 tokens and continued training for five epochs with gradient accumulation (1 × 16 effective batch), mixed precision, and label smoothing. We also created a world attention mask that literally indicates the beginning of each of the <ARTICLE_i> segments such that the model informs about which spots are structurally significant in the combined news articles. In training, validation ROUGE-1 and ROUGE-L immediately jumped to the range of 0.16-0.20 after only one or two epochs of Phase B, which is a big leap relative to AarBERT and mBERT. With the held-out test set, the fine-tuned LED model had ROUGE-1 best over refs of approximately 11.4, ROUGE-L best over refs of approximately 11.1 and BLEU of approximately 3.28. chrF reached ~22.8, and METEOR was around 11–12%. The summaries were on average approximately 80 words, and this implies that LED was not merely copying the lengthy spans but actually reducing the input of the multi-document into one short event-level paragraph. Qualitatively, the outputs of LEDs begin to resemble actual news summaries: they refer to specific actors and actions related to a particular incident, rather than fading into generic background. Nonetheless, the difference in faithfulness and lexical overlap still remains present, particularly on ROUGE-2 (around 2.7%), indicating that the model occasionally picks up on who/what but fails to provide finer relational information (who did what to whom when). On the whole, this experiment is a significant milestone: as soon as we switched 512-token bottleneck models to a long-context encoder-decoder explicit global attention across article boundaries, we finally got recognizably on-topic, concise, and structurally similar summaries to our human-written multi-document references. This preconditions the final system, which enhances the contents selection even more than only the fluency.

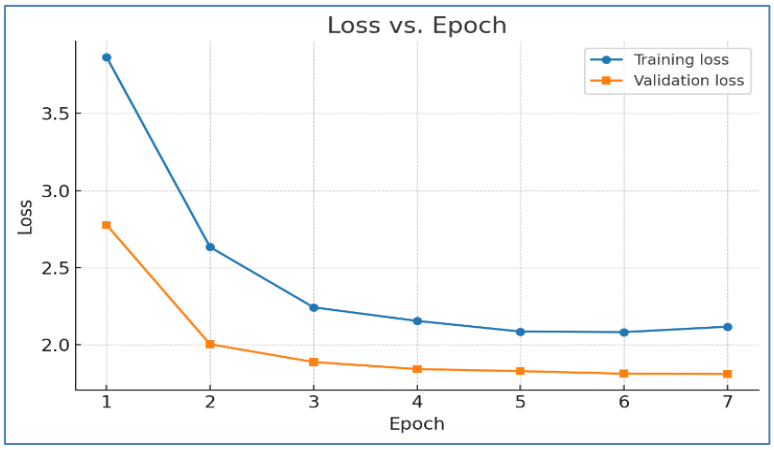


Figure 4. Loss curves showing optimization stability and convergence speed across epochs under the final training configuration

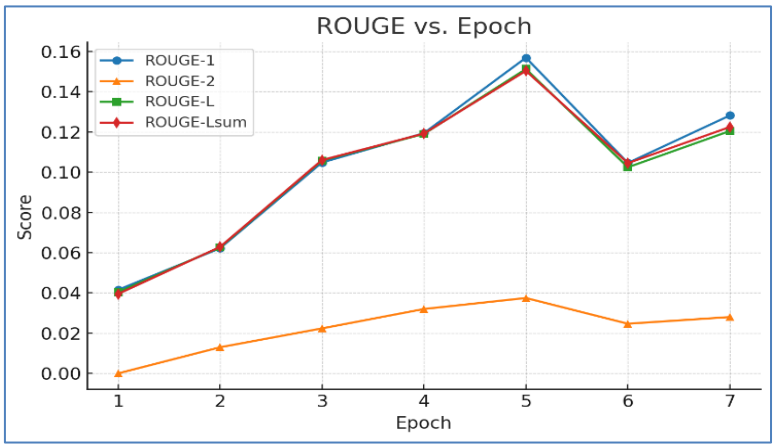


Figure 5. ROUGE curves viewing optimization across epochs

The last system is a sequence-to-sequence model, initialized with csebuetnlp/mT5_multilingual_XLSum that is already pre-trained to do abstractive news summarization in many languages. We then optimized it on our Arabic multi-document dataset in a long-input regime. In practice, we fed the model concatenated multi-article context (a maximum of 6,144 tokens fed on the encoder side) and requested it to generate a single event-level summary that is limited to 512 tokens. Each input was preceded with a clear Arabic brief (summarize in Modern Standard Arabic with accuracy, brevity, and clarity) such that the task form had a consistent format. As Figure 4 indicates, the training loss declines gradually over the epochs, which means that optimization is stable and converts to the same point under the last configuration. In Colab, training was performed with gradient accumulation (2 × 16 effective batch), with Adafactor and a 2e-5 learning rate and cosine decay, mixed precision, and early stopping on ROUGE-Lsum to avoid overfitting. In contrast to the BERT-style encoder-decoders, this model is inherently generative and resistant to abstractive long-form goals, thus it is conditioned to generate in a summarization register rather than generating random background text. We also saw an apparent jump in the validation quality after the third epoch (Figure 5) where ROUGE-1 and ROUGE-L increased approximately twofold, and the model did not collapse. This system performed best on the held out test set, with ROUGE-1 best over refs = -17.6%, ROUGE-L best over refs = -16.9%, ROUGE-2 best over refs =-2.97 and BLEU =-6.37. chrF was ~41.8, METEOR reached ~21.3, and BERTScore F1 was ~84.9, with MoverScore around 0.97. The model also generated terse summaries (one paragraph instead of several hundred words) and the summaries were event-oriented: they specifically identified actors, actions, and consequences of a particular event rather than wandering into the abstract history or politics.

Relative to LED, mT5 also had better lexical alignment and semantic fidelity to the human-written targets, particularly in METEOR and chrF, which measure adequacy on the surface n-gram overlap. We interpret this as meaning that two things simultaneously make this model particularly helpful: (i) inductive bias provided by pretraining on news-style summarization (XLSum) makes it sound more like a summary, and (ii) our long-context fine-tuning helps to apply that capacity to multi-article Arabic input, which is structurally more complicated than single-article summarization. Practically, this is the earliest system in our pipeline that consistently delivers something that a human editor would conceivably would publish, with only a small amount of editing.

We use mT5 as its SentencePiece unigram vocabulary is able to compress Arabic effectively to capture common clitics and affixes (e.g., "و", "الـ", "b") and thus stores fewer subword tokens than a byte-level or English-based BPE tokenizer. This compression is used to enlarge the context under a fixed token budget, which is critical in multi-document inputs where truncation is detrimental to recall. Moreover, the relative position bias of mT5 levels off behavior at longer sequence lengths (>6k tokens), and this is where we take advantage of this in this work.

All the experiments were run on Google Colab Pro+ with Nvidia A100 GPU (40GB VRAM), Intel Xeon CPU, and 85GB of RAM. It was implemented using Python 3.12 and PyTorch (2.1.0+cu121) deep learning framework. In the case of sequence-to-sequence modeling, we used Hugging Face Transformers (v4.39) and Datasets (v2.18) libraries. The packages used to evaluate were Evaluate (v0.4.1) along with sacreBLEU to evaluate BLEU/chrF, as well as custom to evaluate ROUGE. Experimental Environment and Training configuration sum up are shown in Table 2.

Table 2. Experimental Environment and Training Configuration

Category	Specification
Environment	Google Colab Pro+
Hardware	NVIDIA A100 GPU (40GB VRAM), Intel Xeon CPU, 85GB RAM
Software	Python 3.12, PyTorch 2.1.0+cu121, Transformers v4.39, Datasets v2.18, Evaluate v0.4.1, sacreBLEU, ROUGE
Optimizer	Adafactor (learning rate = 2e-5, weight decay = 0.01)
Regularization	Label smoothing = 0.1, Warm-up ratio = 0.2
Batching	Per-device batch size = 2, Gradient accumulation = 16 (effective batch = 32)
Memory Handling	Gradient checkpointing, mixed-precision (bf16/fp16)
Input/Output	MAX_IN = 6144 tokens, MAX_OUT = 512 tokens
Training	Up to 8 epochs, Early stopping with patience = 2
Validation	10% split with fixed seed = 42
Decoding	Beam search (num_beams = 4, max_new_tokens = 512)
Ablation Sweeps	Beam size, length penalty, repetition penalty, minimum new tokens

We fix random seed 42 for dataset split and training; preserve the best checkpoint, tokenizer, and the generation configuration; and maintain a lightweight data audit (token histograms, duplicate reports). The training/evaluation code paths are identical except for metrics logging. We release the decoding sweep script used to save the generation_config.json alongside the best model to ensure future evaluations replicate the Strict setting by default as seen by Figure 6.

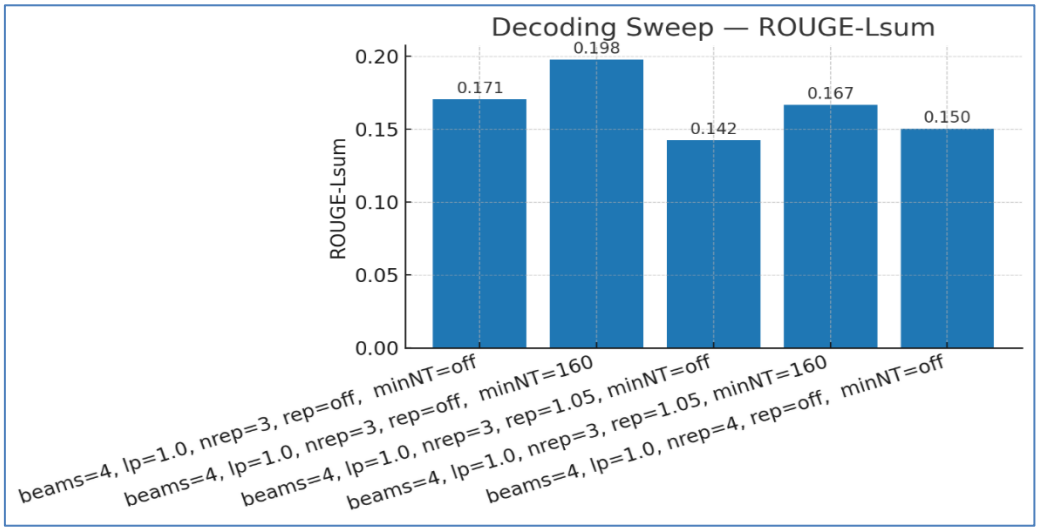


Figure 6. ROUGE-Lsum scores across different decoding configurations

We sweep num_beams ∈ {4,5}, length_penalty ∈ {1.00, 1.05, 1.10}, no_repeat_ngram_size ∈ {3,4}, repetition_penalty ∈ {off, 1.05}, and min_new_tokens ∈ {off, 160}. Enforcing a minimum summary length (e.g., min_new_tokens=160) tends to raise ROUGE-Lsum by improving coverage of salient content, for instance, the setting [beams=4, lp=1.0, nrep=3, rep=off, minNT=160] produced higher in-sweep ROUGE-Lsum, at the cost of modest redundancy and occasional templated phrasing. On the other hand, applying a repetition penalty reduces redundancy but can dampen legitimate lexical reuse in Arabic news discourse, slightly lowering ROUGE. So, headline results under a strict cab be reported, length-agnostic setting and treat the minimum-length control as a deployment-time knob for recall-oriented use cases.

5. Results and discussion

The model was evaluated with the help of several measures to give a detailed analysis:

Automated System Evaluation is a crucial role in assessing the performance and effectiveness of abstractive multi-document Arabic text summarization systems. Measurements of evaluation are in the form of numbers. They help to objectively evaluate the quality of the summaries generated against human-generated reference summaries. The following are some of the commonly used evaluation metrics of automated summarization systems:

To assess the performance of our Arabic multi-document abstractive summarization models, we employed a set of widely used automatic evaluation metrics, including ROUGE, BLEU, chrF , BERTScore, MoverScore, and METEOR. Each of these metrics is grounded in the comparison between automatically generated summaries and human-written references, yet they differ in their underlying mechanisms.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) computes overlaps between system outputs and reference summaries in terms of n-grams, longest common subsequences, or skip-bigrams. In particular, ROUGE-1 and ROUGE-2 measure unigram and bigram overlap respectively, while ROUGE-L and ROUGE-Lsum assess the longest common subsequence between candidate and reference. ROUGE measures the overlap

of nnn-grams between the generated and reference summaries. It is typically reported as precision, recall, and F1 scores:

$$Precision_ROUGE = \frac{|G_ref \cap G_gen|}{|G_gen|}$$

$$Recall_ROUGE = \frac{|G_ref \cap G_gen|}{|G_ref|}$$

$$F_1, ROUGE = \frac{2 \cdot Precision_ROUGE \cdot Recall_ROUGE}{Precision_ROUGE + Recall_ROUGE}$$

BLEU, by contrast, was originally designed for machine translation evaluation and focuses on precision of n-gram matches, penalizing excessively short outputs through a brevity penalty. The formula is:

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \cdot ln\, p_n)$$

with brevity penalty BP:

BP = { 1 (if c > r), exp(1 − r/c) (if c ≤ r) }

chrF: chrF computes the F-score based on character nnn-gram matches, making it more sensitive to morphological variations, especially in morphologically rich languages:

$$chrF = \frac{(1 + \beta^2) \cdot P_char \cdot R_char}{\beta^2 \cdot P_char + R_char}$$

METEOR: Evaluates unigram matches by considering exact matches, stem matches, and synonyms, with an additional fragmentation penalty to account for word order. Its components are defined as:

$$Precision_M = \frac{m}{w_t}$$

$$Recall_M = \frac{m}{w_r}$$

$$F_mean = \frac{10 \cdot Precision_M \cdot Recall_M}{Recall_M + 9 \cdot Precision_M}$$

Penalty = $0.5 \cdot (ch/m)^3$

$$METEOR = (1 - Penalty) \cdot F_mean$$

BERTScore uses contextual embeddings to compute semantic similarity between tokens in the generated and reference summaries.

$$P_{\text{BERT}} = \frac{1}{|\, G_{\text{gen}} \,|} \sum_{i \in G_{\text{gen}}} \max_{j \in G_{\text{ref}}} \cos(x_i, y_j)$$

$$R_{\text{BERT}} = \frac{1}{|\, G_{\text{ref}} \,|} \sum_{j \in G_{\text{ref}}} \max_{i \in G_{\text{gen}}} \cos(y_j, x_i)$$

$$F_{1,\text{BERT}} = \frac{2 \cdot P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

BERTScore is well-suited for Arabic because contextual embeddings capture meaning beyond surface form, handling paraphrasing and morphological variation. MoverScore builds upon embeddings by computing semantic transport cost between summaries using Earth Mover's Distance (EMD).

$$\text{MoverScore} = -\text{EMD}(x, y)$$

A smaller EMD (i.e., smaller semantic distance) yields a higher score.

MoverScore is particularly effective for abstractive and multi-document summarization because it tolerates paraphrasing, sentence reordering, and content synthesis, all common in Arabic news narratives.

Despite their ubiquity, these metrics exhibit important limitations when applied to Arabic abstractive summarization. They are surface-level and overlap-based, assuming lexical similarity with the reference as the primary indicator of quality. This is problematic in abstractive settings where valid summaries may use entirely different lexicalizations or structures from the reference while still being accurate, coherent, and faithful to the source. Moreover, for Arabic specifically, the high degree of morphological richness, orthographic variation, and flexible word order introduces mismatches that depress scores, even when the generated summary is semantically sound.

In practice, BERTScore which relies on contextual embeddings from transformer encoders and judges similarity at the level of meaning rather than exact word overlap. Earlier studies suggest that such embedding-based metrics align better with human judgments in morphologically complex languages, which makes BERTScore a reasonable fit for Arabic summarization. We further include MoverScore which combines these contextual embeddings with Earth Mover's Distance to estimate how much "semantic mass" has to be moved to align the candidate summary with the reference. Intuitively, this allows MoverScore to reward summaries that paraphrase, merge, or reorder information across documents while still staying close in meaning to the human-written references.

In our experiments, we therefore treat these metrics as indicative signals rather than absolute judgments of summarization quality. They serve as consistent benchmarks for model tuning and comparison across experiments, but they cannot by themselves determine the true effectiveness of a summarizer. Human evaluation remains indispensable, especially in abstractive multi-document summarization, where fluency, coherence, and factual faithfulness are not adequately captured by surface-level n-gram matching. Accordingly, our reported scores should be interpreted as relative indicators within the controlled setting of our experiments, rather than definitive measures of the generated summaries' quality.

We report ROUGE-1/2/L/Lsum, BLEU, and chrF on the validation set (n=72) using evaluate (HuggingFace). For comparability, we use two settings: (i) Strict, the same decoding as during training (beams=4, max_new_tokens=512, no extra constraints); and (ii) Decoding sweeps a small grid over decoding hyperparameters to probe the extractive–abstractive trade-offs. Texts are lightly normalized (whitespace compaction; punctuation tightening) prior to scoring.

Table 3. Overview of automatic evaluation metrics used in this study

Metric	Basis of Computation	Strengths	Limitations (esp. for Arabic abstractive summarization)
ROUGE	Recall-oriented n-gram and subsequence overlap (ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum)	Simple, widely adopted; good for extractive summaries	Penalizes valid paraphrases; sensitive to word order and morphology

Metric	Basis of Computation	Strengths	Limitations (esp. for Arabic abstractive summarization)
BLEU	Precision of n-gram overlap with brevity penalty	Established in MT; good for measuring precision	Overly harsh for abstractive summaries; poor correlation with human judgments on long texts
chrF	Character-level F-score based on n-gram matches	Robust to morphological variation; handles rich morphology better	Still surface-based; ignores semantic adequacy and discourse structure
METEOR	Alignment with stemming, synonym, and paraphrase matching	Higher correlation with human evaluation; less rigid than ROUGE/BLEU	Computationally heavier; limited Arabic resources for synonyms; may misalign due to word order flexibility
BERTScore	Contextual embedding similarity using transformer encoders (token-level cosine similarity)	Captures semantic meaning beyond surface overlap; handles paraphrasing and lexical variation; well-suited to Arabic due to contextual embeddings	Dependent on quality of pretrained Arabic or multilingual encoders; may inflate scores for fluent but partially incorrect summaries; higher computational cost
MoverScore	Semantic alignment via contextual embeddings + Earth Mover's Distance	Measures semantic "distance" between candidate and reference; robust to paraphrasing, reordering, and summarization-style compression; effective for multi-document synthesis	Requires strong contextual embeddings for Arabic; computationally expensive; may still struggle with hallucinations that preserve semantic similarity but distort factuality

Table 3 reports the performance of the five models, mT5, AraBERT, AraT5, mBERT, and LED, on the held-out evaluation set under the multi-reference setup. We include ROUGE-1/2/L/Lsum (best over the two references), SacreBLEU and chrF (multi-reference), as well as METEOR, BERTScore, and MoverScore (best over references). Overall, mT5 is clearly the strongest system: it achieves the highest ROUGE-1, ROUGE-2, and ROUGE-L/Lsum (≈0.18/0.03/0.17), together with the best SacreBLEU (≈6.4), chrF (≈41.8), and METEOR (≈0.21), indicating that it captures more of the reference content and wording than the competing models.

Based on Table 4, LED is the best choice among the other systems. It has a ROUGE-1/L/Lsum score of about 0.11, a chrF score of more than 22, and a METEOR score of more than 0.11. Both mT5 and LED get very high embedding-based scores, with BERTScore F1 over 0.83 and MoverScore close to 0.97. This means that their summaries keep the meaning of the human references better than just n-gram overlap. The hierarchical AraBERT, AraT5, and mBERT baselines, on the other hand, give very low ROUGE and SacreBLEU values. However, their BERTScore and MoverScore are still fairly high (≈0.71–0.73 F1 and ≈0.95–0.96), which means that the summaries are semantically related but not as well aligned. In Arabic multi-document, these results show that long-context encoder–decoder architectures (mT5 and LED) are clearly better than encoder-only baselines.

Table 4. Automatic evaluation results on the validation Set (n = 72)

Metric / Model	mT5	AraBERT	AraT5	mBERT	LED
ROUGE-1	**0.176**	0.017	0.000	0.000	0.114
ROUGE-2	**0.029**	0.000	0.000	0.000	0.027
ROUGE-Lsum	**0.169**	0.015	0.000	0.000	0.111
SacreBLEU	**6.37**	0.05	0.00	0.03	3.28
chrF	**41.80**	10.29	13.94	11.50	22.76
METEOR	**0.213**	0.063	0.000	0.021	0.117
BERTScore	**0.849**	0.727	0.710	0.729	0.833
MoverScore	**0.970**	0.957	0.957	0.954	**0.970**

mT5 summaries usually include all the main facts about each event, put information from several articles into a single, well-organized story, and keep the references to entities and numbers consistent throughout the text. In contrast, LED often makes shorter, more compressed outputs that are still fluent but leave out important details or break the story into loosely connected parts. To put it another way, LED's metrics show a good level of semantic overlap, but they don't show how much better mT5 actually is at content selection, cross-document aggregation, and discourse coherence. This shows that when comparing strong abstractive models, automatic scores should be backed up by human inspection. Human evaluation means having people judge how good the summaries are. Judges can give the summaries scores based on how well they are coherent, informative, and fluent. This subjective evaluation offers significant insights into human perceptions of the summaries. We did a manual check of the outputs from the five systems—mT5, LED, AraBERT, mBERT, and AraT5—along four standard dimensions: fluency and grammaticality, global coherence, informativeness/coverage, and faithfulness to the source. This was in addition to the automatic metrics. In order to get accurate human judgments, we hired five native Arabic speakers with different levels of education, from an undergraduate to a PhD holder (Dr. Asmaa, asmaalazim@iku.edu.iq), to be the team leader (annotation/evaluation) and to work as annotators and evaluators. For each document cluster, these annotators looked at both the gold human-written summary and the system-generated summary together and judged how well the system-generated summary matched the human-written one. They were specifically asked to evaluate how well each system-generated summary maintained the essential information, overall structure, and natural flow of the human-written summary, as well as to identify any missing content, nuanced shifts in meaning, or stylistic issues that could diminish its readability or fidelity to the original. By doing this, our annotators not only made sure that the human summaries were consistent and accurate, but they also gave us a detailed, expert-informed view of the model outputs that speaks to both the formal linguistic quality and how proficient Arabic readers actually experience the summaries. Overall, mT5 scored the highest on all counts: its summaries were always well-formed, logically structured, and covered the main event information with only a few minor mistakes or drift, and there were no systematic hallucinations. LED came in second place. Its outputs were mostly fluent and accurate, but they were a little more extractive and less complete than mT5's, which made them seem a little less informative and

coherent. On the other hand, the hierarchical AraBERT, mBERT, and AraT5 baselines often produced text that was very bad, with sentences that were only partially broken and long strings of repeated or nonsensical tokens. This led to very low scores for fluency, coherence, and coverage, and in some cases, the summaries were completely unusable. In summary, the human evaluation backs up the automatic ranking: mT5 is the only model that consistently gives high-quality abstractive summaries in practice. LED is a weaker but still useful option, and the other models don't meet basic requirements for readability and faithfulness. These three examples—mT5 with ROUGE-Lsum = 1.0, mT5 with ROUGE-Lsum = 0.0, and LED with ROUGE-Lsum ≈ 0.67—show that mT5 usually makes fluent and accurate summaries, even when there is no n-gram overlap. LED, on the other hand, makes reasonable but less complete outputs. This shows the difference between lexical overlap scores and true human-perceived quality.

First: mT5 :rougeLsum_best = 1.0

Title:

محكمة العدل الدولية تقول إن الاحتلال الإسرائيلي للأراضي الفلسطينية غير قانوني

Prediction:

قالت محكمة العدل الدولية إن سياسة الاستيطان الإسرائيلية في الضفة الغربية والقدس الشرقية تنتهك القانون الدولي. ورغم أن هذه سياسات تتعارض مع المادة 49 من اتفاقية جنيف الرابعة، فإن هذا القرار يشكل ضررا على محاولات حل الصراع بين إسرائيل والفلسطينيين. كما رفضت إسرائيل دعوة الجمعية العامة للأمم المتحدة لإصدار رأي استشاري يدعو إلى انسحاب فوري أو غير مشروط من الأراضي الفلسطينية المحتلة، حيث وصفت المحكمة قرارها بأنها محاولة من الفلسطينيين لإملاء نتائج التسوية السياسية للنزاع بين إسرائيل وحركة حماس. وفي الوقت نفسه، طالبت الأمم المتحدة بتقديم تعويضات عن الأضرار التي تسببها هذه السياسات تجاه الفلسطينيين، حيث أدان رئيس الوزراء الإسرائيلي بنيامين نتنياهو بأنها ترحب بهذا القرار، مما يؤثر على السياسات الإسرائيلية تجاه القدس الشرقية. طالبت إسرائيل بتقديم تعويضات عن الأضرارا التي سببها احتلال إسرائيل للأراضي الفلسطينية. وفي هذا الصدد، أدان الجيش الإسرائيلي العثور على عشرات من مسارات الأنفاق على امتداد الحدود مع مصر، مما أدى إلى تدمير الفلسطينيين تدميرا شاملا. وقد رفضت السلطات الإسرائيلية دعوة الأمم المتحدة لدعوة جمعية استشارية تدعو إلى الانسحاب فوري او غير مشروط. كما أدانت إسرائيل الدعوة لجمعية عامة للامم المتحدة لإسقاط رأي الاستشاري الذي دعا إليه رئيس الحكومة الإسرائيلية، قائلة إن هذه الخطوة تعد محاولة من حماس لإملاء النتائج تسوية سياسية للحرب على غزة. وقد رفعت جنوب أفريقيا دعوة للانسحاب فوري يا غير مشروط، حيث تعرضت إسرائيل للدعوة لإنسحاب فوري من الأراضي المحتلة. وفيما يلي أبرز القضايا التي يواجهها الفلسطينيون. كما طالبت منظمة العفو الدولية بتقديم تعوضات عن هذه الأضرار، حيث طلبت الجمعية الوطنية للأمم المتحدة دعوة استشارية يدعو الفلسطينيين إلى إنسحاب فوري، وذلك في ظل مطالبة إسرائيل بإص

REF1:

أصدرت محكمة العدل الدولية رأيًا استشاريًا يفيد بأن سياسة الاستيطان الإسرائيلية في الضفة الغربية والقدس الشرقية تعد انتهاكًا للقانون الدولي. رغم أن هذا الرأي غير ملزم قانونيًا، إلا أنه قد يؤثر على المواقف الدولية تجاه بناء المستوطنات الإسرائيلية، ويمكن أن يؤدي إلى تغييرات في السياسات الإسرائيلية. وأوضحت المحكمة أن نقل إسرائيل للسكان إلى الضفة الغربية والقدس الشرقية، وكذلك استمرار وجودهم، يتعارض مع المادة (49) من اتفاقية جنيف الرابعة. كما أكدت المحكمة أن السياسات الاستيطانية لإسرائيل واستغلال الموارد الطبيعية في الأراضي الفلسطينية تعتبر خرقًا للقانون الدولي. وكانت الجمعية العامة للأمم المتحدة قد طلبت من المحكمة تقييم الأثر القانوني لـالاحتلال المطول والاستيطان والضم الإسرائيلي للأراضي الفلسطينية، بما فيها القدس الشرقية. في حين أن إسرائيل تعتبر القدس بالكامل عاصمتها الموحدة، إلا أن هذا الادعاء لا يحظى بقبول غالبية المجتمع الدولي. وأعربت المحكمة عن قلقها من استمرار توسع سياسة الاستيطان الإسرائيلية، خاصة بعد فتواها الاستشارية السابقة حول جدار الفصل.

REF2:

أصدرت محكمة العدل الدولية رأيًا استشاريًا يؤكد أن سياسة الاستيطان الإسرائيلية في الضفة الغربية والقدس الشرقية تشكل انتهاكًا واضحًا للقانون الدولي. ورغم أن هذا الرأي غير ملزم، إلا أنه قد يترك أثرًا ملموسًا على مواقف المجتمع الدولي تجاه الاستيطان، وقد يدفع إلى تغييرات محتملة في السياسات الإسرائيلية. وأوضحت المحكمة أن قيام إسرائيل بنقل سكانها إلى الضفة والقدس الشرقية، واستمرار وجودهم هناك، يتعارض مع المادة (49) من اتفاقية جنيف الرابعة، كما شددت على أن استغلال إسرائيل للموارد الطبيعية في الأراضي الفلسطينية يُعد خرقًا للقانون الدولي. وجاء هذا الرأي بعد طلب من الجمعية العامة للأمم المتحدة لتقييم الأبعاد القانونية للاحتلال الطويل الأمد والاستيطان والضم الإسرائيلي، بما في

ذلك القدس الشرقية. وفي حين تتمسك إسرائيل باعتبار القدس كاملة عاصمتها الموحدة، إلا أن هذا الموقف لا يحظى باعتراف دولي واسع. وأعربت المحكمة عن قلقها إزاء استمرار التوسع الاستيطاني، خصوصًا في ظل تجاهل إسرائيل لفتواها السابقة بشأن جدار الفصل.

Second: mT5: rougeLsum_worest = 0.0

Title:

إسقاط التهم عن عشرات الطلاب الأميركيين الرافضين للحرب على غزة

Prediction:

أسقطت محكمة في نيويورك، الخميس، جميع التهم التي وُجهت سابقا للعشرات من طلاب جامعة كولومبيا، الذين شاركوا في اعتصامات ضد الدعم الأميركي للعدوان الإسرائيلي على قطاع غزة، مما أدى إلى مقتل ما يقرب من 30 ألف شاب، بينهم أطفال ونساء، ودمارا هائلا، حسب بيانات أممية وأممية فلسطينية وأممية. يواجه جو بايدن اتهامات جديدة له بعدم الإنجاز في الانتخابات التمهيدية للحزب الديمقراطي، وذلك بعدم الإنسحاب من الشركات الداعمة للاحتلال، مما دفع الناخبين إلى وضع العلامة أمام اختيار غير ملتزم في بطاقة اقتراع انتخابات الرئاسة المقرر إجراؤها في أغسطس/آب المقبل، التي يواجهها الرئيس الأمريكي جو بايدن بعدم الإنتخاذ لصالح الحزب الديمقراطي. في هذه الاحتجاجات أثارت موجة من الاضطرابات في أنحاء الولايات المتحدة، حيث تجمع مئات الآلاف للاحتجاجات، مطالبين بإلغاء خطة لبناء صالة رياضية في حي هارلم، حيث أطلقت قوات الحرس الوطني بولاية أوهايو، حيث دفعت القوات الحكومية إلى إطلاق النار في جامعة كينت الحكومية، حيث أسقطت المحكمة جميع الاتهامات التي وجهت سابقًا لعشرات طلاب كولومبيا الذين اعتصاما ضد العدوان الأميركي على غزة، إلى أن الحزب لا يزال بعيدا عن التعامل مع الشركات الدعمية للاحتلال الأمريكي، إلى تراجع تأييد الانتخاب التمهيدية لحزبه الديمقراطي، إلى إغلاق خطة للبناء الصالة الرياضية التي كانت محطة فارقة في مسار النضال من أجل الحقوق المدنية والعدالة الاجتماعية، حيث كان الناخبون يواجهون وضع علامة أمام الاختيار غير الملتزم في البطاقة اقتراع الانتخابات الرئاسية المقرر إجريؤها، كما يواجه بايدن تهم جديدة له بسبب عدم الإنجاز، حيث لم يواجه الرئيس جو بايدن أي اتهامات بعدم الإنضمام إلى حزبه، لكنه أشار إلى أن الحرب الإسرائيلية كانت محطات فارقة في المسار

REF 1:

أسقطت محكمة في نيويورك التهم الجنائية عن 46 من طلاب جامعة كولومبيا الذين اعتُقلوا خلال اعتصام ضد الدعم الأميركي للعدوان الإسرائيلي على غزة. الاحتجاجات الطلابية ضد الحرب الإسرائيلية على غزة انتشرت من جامعة كولومبيا إلى أكثر من 70 جامعة أميركية، وصولاً إلى جامعات في أوروبا. طلاب جامعة كولومبيا نظموا اعتصامًا في 17 أبريل 2024 ضد الدعم الأميركي لإسرائيل، مطالبين بإغلاق تعاون الجامعات مع الجامعات الإسرائيلية. الحراك الطلابي المناهض للحرب في غزة في الولايات المتحدة أدى إلى اعتقال أكثر من 2000 من المحتجين. الطلاب في الجامعات الأميركية يطالبون بوقف الحرب على غزة وسحب استثمارات الجامعات من الشركات الداعمة لإسرائيل. الاحتجاجات الطلابية الحالية تذكرنا بحركة الاحتجاجات ضد حرب فيتنام التي نشأت في السبعينيات. في 1970، استخدمت قوات الأمن في جامعة كينت الوطنية الأسلحة ضد الطلاب الذين كانوا يحتجون على حرب فيتنام، مما أسفر عن مقتل 4 طلاب. الاحتجاجات الطلابية في السبعينيات ساعدت في تشكيل الحركة المناهضة للحرب وأثرت على السياسة الأميركية. بعد 54 عامًا من احتجاجات فيتنام، لا تزال الاحتجاجات الطلابية في الجامعات الأميركية تحمل تأثيرًا سياسيًا كبيرًا، خاصة في الانتخابات الرئاسية. طلاب الجامعات الأميركية الحالية يعبرون عن غضبهم من سياسة بايدن تجاه إسرائيل والقتال في غزة. احتجاجات طلابية في جامعة كولومبيا تعرضت للقمع من قبل الشرطة، مع تعرض العديد من الطلاب للاعتقال. الاحتجاجات المناهضة للحرب في غزة لم تقتصر على الجامعات الأميركية فقط، بل امتدت إلى فرنسا وبريطانيا وألمانيا وغيرها.

REF 2:

قامت محكمة في نيويورك باسقاط التهم الجنائية عن 46 من طلاب جامعة كولومبيا الذين اعتُقلوا خلال اعتصامٍ نظموه احتجاجًا على الدعم الأميركي للعدوان الإسرائيلي على غزة، في تطوّرٍ يجيء فيما كانت موجة الاعتصامات والاحتجاجات الطلابية قد خرجت من أسوار كولومبيا لتتمدّد إلى أكثر من سبعين جامعة أميركية ثم إلى جامعات أوروبية، حيث بدأ الحراك في 17 أبريل/نيسان 2024 باعتصامٍ يطالب بإغلاق قنوات التعاون مع الجامعات الإسرائيلية وبسحب استثمارات الجامعات من الشركات الداعمة لإسرائيل، بينما تشير التقديرات إلى أنّ هذا الحراك المناهض للحرب في الولايات المتحدة أفضى إلى اعتقال أكثر من ألفي محتج؛ هذا وتستدعي هذه التعبئة الواسعة ذاكرة السبعينيات، إذ تُذكّر, على نحوٍ مُلح, بحركة الاحتجاج على حرب فيتنام وما رافقها من قمعٍ بلغ ذروته في جامعة كينت الحكومية عام 1970 حين استخدمت قوات الأمن الرصاص ضد طلابٍ محتجين فَقُتل أربعة، وهو الإرث الذي ساهم في تشكيل حركة مناهضةٍ للحرب أثّرت في السياسات الأميركية آنذاك، وإلى اليوم, بعد أربعةٍ وخمسين عامًا على احتجاجات فيتنام, لا تزال التعبئة الطلابية تحمل أثرًا سياسيًا معتبرًا، وخصوصًا في مواسم الانتخابات الرئاسية، حيث يعبّر طلابٌ كثر عن غضبهم من سياسة إدارة بايدن حيال إسرائيل والقتال الدائر في غزة؛ وإلى جانب ذلك، شهد حرم كولومبيا تدخلًا أمنيًا قوبل بانتقاداتٍ واسعة بعد تعرّض عددٍ كبير من الطلبة للاعتقال، فيما لم يقتصر المشهد على الجامعات الأميركية، إذ امتدّ إلى فرنسا وبريطانيا وألمانيا وغيرها، بما يعكس تحوّل الجامعات إلى ساحاتٍ مركزية للنقاش والضغط السياسي إزاء الحرب ومساراتها.

Third: LED: rougeLsum_best = 0.6667

بأكبر مشاركة دولية في تاريخه.. معرض الدوحة الدولي للكتاب ينطلق في مايو

Prediction:

أعلنت وزارة الثقافة القطرية أمس الأحد، أن فعاليات الدورة 33 من معرض الدوحة الدولي للكتاب ستنطلق في مركز الدحة للمعارض والمؤتمرات من 9 إلى 18 مايو/ أيار المقبل. وتحت شعار بالمعرفة تبنى الحضارات سيشارك أك [SEP] □أبرز الفرق المسرحية والفنية العربية من سلطنة عمان ولبنان وغيرهما. فقد أضاف العلي يحظى كذلك سيتم من خلال الص أطفال أفق التعلم والتفكير والإبداع بطرقة تجني ثمار التطور الكبير الذي تاريخية وعلماء لهم بصمة اهتماما خاصا بفع اليات أنشطة تفاعلية عدة. ويشهد الجناح أيضا تدشين

REF 1:

أعلنت وزارة الثقافة القطرية عن انطلاق فعاليات الدورة 33 من معرض الدوحة الدولي للكتاب في مركز الدوحة للمعارض والمؤتمرات من 9 إلى 18 مايو. يُعد هذا الحدث الأكبر في تاريخ المعرض، حيث يشارك فيه أكثر من 515 دار نشر من 42 دولة. تحل سلطنة عُمان ضيف شرف الدورة، بمشاركة جناح خاص وبرنامج ثقافي يُبرز التراث العماني الأصيل. ويشمل برنامج المعرض مجموعة واسعة من الفعاليات، بما في ذلك جلسات أدبية وفكرية، ندوات ومحاضرات، أمسيات شعرية، عروض مسرحية، وفعاليات مخصصة للأطفال. تتميز الدورة الحالية بمشاركة دولية متنوعة، من بينها المجلس الوطني للثقافة والفنون والآداب بدولة الكويت، هيئة الشارقة للكتاب، ووزارات الثقافة في المغرب والجزائر واليمن ومصر. ويضم المعرض المنطقة المركزية التي تركز على شخصيات تاريخية وعلماء بارزين، فيما يستضيف المسرح الرئيسي عروضًا مسرحية، منها مسرحية جميل وبثينة لفرقة كركلا من لبنان، بالإضافة إلى عروض يومية. كما يقدم المعرض ورش عمل، ومعرضًا للصور الفوتوغرافية، وبرنامج مرشد القراءة للمبتدئين، مما يجعله حدثًا ثقافيًا شاملاً يلبي اهتمامات الزوار من مختلف الفئات.

REF 2 :

أعلنت وزارةُ الثقافةِ القطرية انطلاقَ فعاليات الدورة الثالثة والثلاثين من معرض الدوحة الدولي للكتاب في مركز الدوحة للمعارض والمؤتمرات خلال الفترة من 9 إلى 18 مايو، في نسخةٍ تُعد الأكبر في تاريخ المعرض بمشاركةٍ تتجاوز 515 دار نشر من 42 دولة، فيما تحلّ سلطنةُ عُمان ضيفَ شرفٍ ببرنامجٍ ثقافي وجناحٍ خاص يُبرز ملامح التراث العُماني الأصيل ويعرّف بروّاده وفنونه؛ وتزخرُ أجندةُ المعرض بطيفٍ واسعٍ من الفعاليات يمتدّ من الجلسات الأدبية والفكرية إلى الندوات والمحاضرات والأمسيات الشعرية والعروض المسرحية وفعاليات الأطفال، إلى جانب حضورٍ دوليٍّ متنوّع يضمّ المجلس الوطني للثقافة والفنون والآداب بدولة الكويت، وهيئة الشارقة للكتاب، ووزارات الثقافة في المغرب والجزائر واليمن ومصر، في حين تتوسّط المعرضَ «المنطقةُ المركزية» التي تركز على شخصيات تاريخية وعلماء بارزين وتقدّم لهم قراءاتٍ معرفيةً معمقة، بينما يستضيف المسرحُ الرئيسي عروضًا مسرحيةً من بينها «جميل وبثينة» لفرقة كركلا اللبنانية إضافةً إلى عروضٍ يومية متجددة، هذا ويقدّم المعرضُ ورشَ عملٍ متخصّصة ومعرضًا للصور الفوتوغرافية وبرنامج «مرشد القراءة» للمبتدئين، بما يجعل من هذه الدورة حدثًا ثقافيًا شاملًا يلبّي اهتمامات الزوّار على اختلاف فئاتهم ويعكس حيوية المشهد الثقافي في الدوحة وتنامي حضوره الإقليمي والدولي.

Table 5 summarizes how our work compares to the most closely related studies in Arabic summarization. Earlier systems are almost exclusively single-document and either extractive or hybrid, typically built on pre-existing corpora such as EASC, NADA, SumArabic, , or AraSum with the dataset creators and model designers being distinct. By contrast, our study both creates a new multi-document, fully abstractive benchmark (Ara25) and provides the first long-context transformer baselines for it, using mT5, LED, AraT5, AraBERT, and mBERT. This joint contribution new event-centric dataset plus unified long-context baselines positions Ara25 as a more realistic and challenging testbed for Arabic multi-document abstractive summarization than any of the prior resources.

Table 5. Comparative summary of existing datasets and baseline models vs. our work

Work	Dataset(s) Used	Dataset Creator / Source	Input & Task Setting	Method Type / Models	Main Difference vs. Our Work
[6]	Custom Arabic article–	Created by the authors (news articles crawled	Single-document, Arabic news; article →	Early neural seq2seq encoder–	Pioneers Arabic neural abstractive summarization, but

	headline pairs	and paired with their own headlines)	short headline (headline generation).	decoder; fully abstractive but only for short inputs.	only for single short articles and one-sentence headlines; no multi-document fusion or long-context setting.
[2]	Multi-document clusters built from EASC	Base corpus: EASC by El-Haj et al.; multi-document clusters constructed by Abdulateef et al.	Multi-document, Arabic news; extractive summaries from clusters.	Unsupervised Word2Vec + k-means clustering + Weighted PCA (W-PCA) for sentence ranking.	One of the first Arabic multi-document systems, but summaries are purely extractive, with shallow embeddings and no transformer or long-context encoder–decoder models.
[4]	EASC	EASC (El-Haj et al.), reused by Alami et al. for single-document summarization.	Single-document, Arabic news; extractive/hybrid.	Hybrid graph–semantic method with an ontology-driven hierarchical graph combining statistical and semantic features.	Adds semantic structure to single-document extractive summarization, but still no multi-document input and no abstractive generation over long clusters.
[35]	EASC + NADA	EASC by El-Haj et al.; NADA introduced in prior work and reused by Reda et al.	Single-document, Arabic news; hybrid extractive→abstractive.	Two-stage pipeline: AraBERT for extractive sentence selection, then Arabic T5 (AraT5) for abstractive rewriting; plus SVM classifier.	Important hybrid extractive+abstractive system, but still single-document, moderate-length articles; does not tackle multi-article clusters or ~6k-token contexts like Ara25.
This work – (2025)	Ara25 – new Arabic multi-document abstractive MDAS dataset	Dataset created by the authors of this paper: 715 events, 2,266 articles from 100+ Arabic news outlets.	Multi-document, fully abstractive Arabic MDS: each instance = 2–5 full articles (~6k tokens) + multi-sentence MSA summary.	Unified long-context abstractive pipeline with five transformers: mT5 (XLSum-initialized), LED-base-16384, AraT5,	First realistic Arabic multi-document abstractive benchmark with long-context transformers. Unlike previous works, we both create the dataset (Ara25) and provide the first long-context abstractive baselines plus a full reproducible

				AraBERT, mBERT in encoder–decoder / hierarchical setups.	pipeline (data, code, metrics, and human evaluation).

6. Conclusion

In conclusion, this study aimed to tackle the significant challenge of generating abstractive summaries for Arabic news articles sourced from multiple, often lengthy documents. To achieve this, we developed a new dataset comprising event-based Arabic articles paired with human-written summaries and trained transformer models capable of processing longer contexts than typically feasible. The findings demonstrated that allowing the model to access more of the input data enables it to capture richer content, thereby minimizing the information loss associated with truncation. Additionally, our experiments revealed that the quality of summarization is highly sensitive to decoding choices, emphasizing the importance of ensuring alignment between training and inference processes. Beyond the quantitative results, this research contributes a novel resource and represents an initial step toward enhancing the reliability and practicality of Arabic multi-document summarization. We believe that this work can serve as a foundation for future initiatives aimed at advancing Arabic natural language processing (NLP). Key findings from our study include the feasibility of long-context Arabic multi-document abstractive summarization (MDAS) despite the linguistic complexities of the language; the positive impact of long-context fine-tuning on recall-oriented metrics such as ROUGE-L and ROUGE-Lsum; the sensitivity of performance to decoding configurations, highlighting the need for consistent training and inference strategies; the establishment of competitive baselines in Arabic MDAS compared to traditional extractive methods, despite modest absolute ROUGE/BLEU values relative to English benchmarks; and the introduction of a new dataset that serves as the first large-scale benchmark for Arabic abstractive MDAS, facilitating future reproducibility and progress in this field.

7. Future work

Expanding the dataset to include a wider range of domains can significantly enhance model generalization, as demonstrated in recent studies on low-resource languages. At the same time, exploring alternative model architectures, such as hierarchical and graph-based approaches, offers promising improvements in handling cross-document relations within multi-document summarization (MDS). Another important direction is cross-lingual summarization, particularly extending to Arabic-English multi-document abstractive summarization (MDAS), where leveraging multilingual resources could enrich performance and applicability. Finally, adopting more Arabic-specific evaluation metrics, including tailored variations of BLEU and ROUGE-AR, would provide deeper and more nuanced insights into model effectiveness, ensuring that performance assessments align closely with linguistic and cultural characteristics of the language.

Declaration of competing interest

The authors declare that they have no known financial or non-financial competing interests in any material discussed in this paper.

Authors' contributions

Amer Al-Dulame: Developed the study concept, developed research design, and helped in data analysis and interpretation. He also helped in manuscript development and revision. Mohsen Kahani and Mohamed Allahbakhsh: Helped in literature review, helped in developing the survey tool, and helped in data collection. He also made critical comments on the manuscript. All authors accepted the final version of the manuscript and approved the sequence of authorship.

Acknowledgment

We all thank our Universities for making this work possible.

Funding information

The authors declare that they have received no funding from any financial organization to conduct this research.

References

[1] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics and 11th Int. Joint Conf. Natural Lang. Process. (Vol. 1: Long Papers), 2021, pp. 7088–7105. https://doi.org/10.18653/v1/2021.acl-long.551

[2] S. Abdulateef, N. A. Khan, B. Chen, and X. Shang, "Multidocument Arabic text summarization based on clustering and Word2Vec to reduce redundancy," Information, vol. 11, no. 2, p. 59, 2020. https://doi.org/10.3390/info11020059

[3] M. Bani-Almarjeh and M.-B. Kurdy, "Arabic abstractive text summarization using RNN-based and transformer-based architectures," Inf. Process. Manage., vol. 60, no. 2, Art. no. 103227, 2023. https://doi.org/10.1016/j.ipm.2022.103227

[4] N. Alami, M. El Mallahi, H. Amakdouf, and H. Qjidaa, "Hybrid method for text summarization based on statistical and semantic treatment," Multimed. Tools Appl., vol. 80, pp. 30539–30561, 2021.

[5] I. Guellil, A. Adeel, F. Azouaou, F. Benali, A. Hacid, and A. Hussain, "Arabic natural language processing: An overview," J. King Saud Univ. - Comput. Inf. Sci., vol. 33, no. 5, pp. 497–521, 2021. https://doi.org/10.1016/j.jksuci.2019.02.006

[6] A. Alhamadani, X. Zhang, J. He, A. Khatri, and C.-T. Lu, "LANS: Large-scale Arabic news summarization corpus," in Proc. ArabicNLP 2023, pp. 89–100. https://doi.org/10.18653/v1/2023.arabicnlp-1.8

[7] M. Alhoshan and N. Altwaijry, "AUSS: An Arabic query-based update-summarization system," J. King Saud Univ. - Comput. Inf. Sci., vol. 32, no. 7, pp. 810–819, 2020. https://doi.org/10.1016/j.jksuci.2018.09.006

[8] Y. A. Al-Khassawneh and E. S. Hanandeh, "Extractive Arabic text summarization: A graph-based approach," Electronics, vol. 12, no. 2, Art. no. 437, 2023. https://doi.org/10.3390/electronics12020437

[9] M. Al-Maleh and S. Desouki, "Arabic text summarization using deep learning approach," J. Big Data, vol. 7, Art. no. 109, 2020. https://doi.org/10.1186/s40537-020-00386-x

[10] M. A. Alwan and H. M. Onsi, "A proposed textual graph based model for Arabic multi-document summarization," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 6, pp. 435–439, 2016. https://doi.org/10.14569/IJACSA.2016.070656

[11] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," arXiv, 2020. https://arxiv.org/abs/2003.00104

[12] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization, 2005, pp. 65–72. https://aclanthology.org/W05-0909/

[13] A. Ben Ayed, I. Biskri, and J.-G. Meunier, "Arabic text summarization via knapsack balancing of effective retention," Procedia Comput. Sci., vol. 184, pp. 45–52, 2021. https://doi.org/10.1016/j.procs.2021.05.100

[14] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," arXiv, 2020. https://arxiv.org/abs/2004.05150

[15] Hugging Face, "Longformer Encoder-Decoder (LED) documentation (Version 4.3.0)," 2021. [Online]. Available: https://huggingface.co/docs/transformers/model_doc/led

[16] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in Proc. 2018 Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations, 2018, pp. 66–71. https://doi.org/10.18653/v1/D18-2012

[17] L. Xue et al., "mT5: A massively multilingual pre-trained text-to-text transformer," in Proc. 2021 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., 2021, pp. 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41

[18] N. Shazeer and M. Stern, "Adafactor: Adaptive learning rates with sublinear memory cost," in Proc. 35th Int. Conf. Mach. Learn., 2018, pp. 4596–4604. doi: https://proceedings.mlr.press/v80/shazeer18a.html

[19] L. Xue et al., "mT5: A massively multilingual pre-trained text-to-text transformer," in Proc. 2021 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., 2021, pp. 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41

[20] T. Chen, B. Xu, C. Zhang, and C. Guestrin, "Training deep nets with sublinear memory cost," arXiv, 2016. https://arxiv.org/abs/1604.06174

[21] M. El-Haj and R. Koulali, "Kalimat: A multipurpose Arabic corpus," in Proc. 14th Int. Conf. Intell. Text Process. Comput. Linguistics (CICLing), 2013, pp. 398–409. https://doi.org/10.1007/978-3-642-37247-6_32

[22] A. T. Al-Taani and S. H. Al-Sayadi, "Extractive text summarization of Arabic multi-document using fuzzy C-means and Latent Dirichlet Allocation," Int. J. Syst. Assurance Eng. Manage., vol. 15, no. 2, pp. 713–726, 2024. https://doi.org/10.1007/s13198-023-02044-6

[23] E. Monir and A. Salah, "AraTSum: Arabic Twitter trend summarization using topic analysis and extractive algorithms," Int. J. Comput. Intell. Syst., vol. 17, no. 1, Art. no. 227, 2024. https://doi.org/10.1007/s44196-024-00546-0

[24] A. Merniz, A. Chaibi, and H. Ben Ghézala, "Multi-document Arabic text summarization based on thematic annotation," in Proc. 16th Int. Conf. Softw. Technol., 2021, pp. 639–644. https://doi.org/10.5220/0010557906390644

[25] M. A. Alwan and H. M. Onsi, "A proposed textual graph-based model for Arabic multi-document summarization," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 6, pp. 435–439, 2016. https://doi.org/10.14569/IJACSA.2016.070656

[26] A. H. Elsaid, A. Mohammed, L. Fattouh, and M. Sakre, "Abstractive Arabic text summarization based on MT5 and AraBART transformers," in Proc. 1st Int. Conf. Intell. Methods, Syst. Appl. (IMSA), 2023, pp. 115–120. IEEE. https://doi.org/10.1109/IMSA58542.2023.10217539

[27] M. El-Haj, U. Kruschwitz, and C. Fox, "Exploring clustering for multi-document Arabic summarisation," in Information Retrieval Technology, 2011, pp. 550–561. Springer. https://doi.org/10.1007/978-3-642-25631-8_50

[28] H. Oufaida, O. Nouali, and P. Blache, "Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization," J. King Saud Univ. - Comput. Inf. Sci., vol. 26, no. 4, pp. 450–461, 2014. https://doi.org/10.1016/j.jksuci.2014.06.008

[29] Z. Chen, L. Li, and Y. Wang, "Wavelet neural networks: A review and comparative study," Neural Comput. Appl., vol. 34, no. 23, pp. 20821–20840, 2022.

[30] G. Alselwi and T. Taşcı, "Extractive Arabic text summarization using PageRank and word embedding," Arabian J. Sci. Eng., vol. 49, no. 9, pp. 13115–13130, 2024. https://doi.org/10.1007/s13369-024-08890-1

[31] M. Zhang, J. Lu, J. Yang, J. Zhou, M. Wan, and X. Zhang, "From coarse to fine: Enhancing multi-document summarization with multi-granularity relationship-based extractor," Inf. Process. Manage., vol. 61, no. 3, Art. no. 103696, 2024. https://doi.org/10.1016/j.ipm.2024.103696

[32] M. Bani-Almarjeh and M.-B. Kurdy, "Arabic abstractive text summarization using RNN-based and transformer-based architectures," Inf. Process. Manage., vol. 60, no. 2, Art. no. 103227, 2023. https://doi.org/10.1016/j.ipm.2022.103227

[33] M. El-Haj, P. Rayson, and M. Albared, "Multi-document Arabic text summarisation: Arabic and English baseline systems," in Proc. 7th Int. Conf. Lang. Resources Eval. (LREC 2010), 2010, pp. 3404–3410. ELRA. https://aclanthology.org/L10-1493/

[34] A. Koubaa, A. Ammar, L. Ghouti, O. Najar, and S. Sibaee, "ArabianGPT: Native Arabic GPT-based large language model," arXiv, 2024. doi: https://arxiv.org/abs/2401.12345

[35] A. Reda et al., "A hybrid Arabic text summarization approach based on transformers," in Proc. 2022 2nd Int. Mobile, Intell., Ubiquitous Comput. Conf. (MIUCC), 2022, pp. 161–167. IEEE. https://doi.org/10.1109/MIUCC55081.2022.9781694

[36] C. Shang, J. Yang, and M. Li, "Arabic abstractive multi-document summarization using transformer models," Appl. Soft Comput., vol. 121, Art. no. 108771, 2022. https://doi.org/10.1016/j.asoc.2022.108771