

SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
http://saucis.sakarya.edu.tr/

e-ISSN: 2636-8129
Publisher: Sakarya University

Vol. 9, No.4, 1127-1140, 2026
DOI:10.35377/saucis...1864801

SAKARYA UNIVERSITY

RESEARCH ARTICLE

Evaluating Large Language Models with a Unified Prompt-Based Pipeline for Turkish News Summarization and Tag Extraction

Burcu Koçak[1], **Kazım Yıldız**[2*]

[1]Department of Computer Engineering, Institute of Pure and Applied Sciences, Marmara University, İstanbul, Türkiye
[2]Computer Engineering Department, Faculty of Technology, Marmara University, İstanbul, Türkiye

*Corresponding author:
kazim.yildiz@marmara.edu.tr

SAKARYA UNIVERSITY

Article History:
Received: 16.01.2026
Revised: 14.03.2026
Accepted: 30.03.2026
Published Online: 21.09.2026

Abstract

As access to information grows ever faster, the acceleration of news flow demands the automatic and accurate analysis of texts. This study aims to measure the multitasking performance of summarizing and tag extraction tasks in Turkish news texts using a prompt-based unified pipeline with a small dataset, employing different prompt techniques. Summarization and tag extraction tasks were performed comparatively using zero-shot and few-shot learning approaches with commercial models (GPT-40 and Claude 4.5 Sonnet) via official API providers and open-source models (Qwen3-32B and Llama 3.3-70B-Instruct) on a local server using a vLLM infrastructure. Experiments were conducted with deterministic parameters and standardized system prompts. Experimental results show that the Claude 4.5 Sonnet model exhibited the highest performance in the 3-shot scenario with a BERTScore of 0.676 and an LLM-as-a-judge score of 7.67. Furthermore, it has been observed that switching to a few-shot strategy resulted in up to a 35% increase in tagging success. The Llama 3.3 model was found to be competitive with commercial models in the tagging task with an F1 score of 0.287. These results also validate the proposed unified pipeline method in terms of accuracy and semantic similarity using BertScore, F1 score, and the LLM-as-a-judge method, which simulates human judgment. The proposed framework is expected to contribute to the literature by offering a reliable automation solution that increases efficiency with small data requirements in Turkish news processing processes.

Keywords: Large Language Models (LLMs), LLM-as-a-judge, Prompt engineering, Turkish Natural Language Processing, Unified pipeline

1. Introduction

The rapid increase in access to information has accelerated news flow; the growth in the amount of data generated is crucial for extracting significant insights from large volumes of data to meet user requirements [1, 2]. News articles, which have maintained a particularly critical role from the past to the present, must remain relevant throughout constantly varying agendas. The rapid flow of news today and the high intensity of information across various fields make an accurate analysis of this data essential [2]. In this regard, news data facilitates research areas for evaluating real-world problems such as summarization, sentiment analysis, classification, tagging, and question-answering tasks.

Lately, substantial advances have been made in the field of Natural Language Processing with the development of deep learning and transformer-based architectures. Large language models are advanced artificial intelligence systems trained on large volumes of text data to comprehend, produce, and process human language. By minimizing human intervention, these models offer high performance in various tasks such as news text analysis, sentiment analysis, summarization, classification, question-answering, and text generation, leading to revolutionary advances [3–5]. The success of large language models is based on the self-attention module of the transformer architecture. Thanks to their in-context learning capability, transformer architectures can identify dependencies in long texts by establishing meaningful relationships between inputs. This

Cite as: B. Koçak, and K. Yıldız, "Evaluating large language models with a unified prompt-based pipeline for turkish news summarization and tag extraction " , *Sakarya University Journal of Computer and Information Sciences* 9 (4) 2026, 1127-1140. doi: 10.35377/saucis...1864801

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License

feature enables large language models to interact with users through natural language, producing more durable and context-based results, making them relevant for speech-based applications [6].

Users can interact with these models using diverse input formats, including text, audio, and images. This interaction process enables the implementation of prompt-based learning approaches, which provide customized cues to the model in natural language, enabling the model to achieve the demanded result without changing parameters [7]. Prompt-based learning is based on two fundamental learning approaches: zero-shot learning and few-shot learning. In contrast, zero-shot learning enables models to implement desired tasks without any examples, while few-shot learning aims to accomplish tasks by providing the model with a few relevant examples. These approaches enable large language models to adapt to new tasks without requiring plenty of data, increasing the models' efficiency and usability in a comprehensive range of applications [3].

Recent studies evaluating the performance of LLMs in text summarization tasks reveal that with proper prompt engineering, LLMs perform strongly in news texts [8]. Large language models adapt efficiently to the latest tasks with small datasets; evaluating their performance on high-volume data, such as news data, has become crucial. In this sense, studies on Turkish news data have also been raised. In 2020, large-scale MLSUM [9] with more than 1.5 million datasets and XL-SUM [10] datasets in 2021 established the essential for summarization studies, including Turkish.

In 2022, TR-News [1], a large-scale summarization and tag extraction dataset with 307,562 articles, was released as open-source. In 2023, Zhang et al. evaluated the news summarization capabilities of LLMs [11]; Çekinel, Karagöz, and Çöltekin published the Turkish FCTR dataset, which contains 3238 real-world claims for fact-checking [12].

In 2024, Üyük et al. discussed the outputs of LLMs for Turkish and multilingual neural news generation [13]. In the same year, Güven compared the zero-shot and few-shot learning capabilities of large language models for Turkish news classification tasks [5]; Yüksel et al. assessed Turkish LLMs using the TurkishMMLU [14] dataset, which has 10,032 multiple-choice questions; Kuş and Acı evaluated the hybrid use of LLMs and statistical methods in summarizing Turkish news [15].

Furthermore, Kaya & Kayalı introduced the HTV-News [16] dataset, which comprises approximately 173,000 news articles and summaries, to address the issue of data scarcity in the field of Turkish news summarization. Uluduğan et al. developed the TURNA [17] model and Melikşah et al. developed the VBART [18] model to assess the performance of Turkish LLMs. In 2025, Er et al. presented CETVEL [19], which deals with the comprehensive evaluation of LLMs comprising 23 tasks for the Turkish language; Bayram et al. improved TR-MMLU [20], which consists of 6200 multiple-choice questions to evaluate the performance of Turkish LLMs.

To clarify the scope of studies in the literature and the place of this study within the literature, key studies between 2020 and 2025 are summarized in Tables 1 and 2.

Table 1. Comparative summary of Turkish news data and LLM evaluation studies (2020–2023)

Year	Reference	Model	Scope and Key Contribution
2020	Scialom et al. [9]	MLSUM	Construction of a large-scale (1.5M+) multilingual news summarization dataset.
2021	Hasan et al. [10]	XL-SUM	A comprehensive data source presentation for multilingual summarization tasks, including Turkish.
2022	Baykara et al. [1]	TR-News	A summarizing and tagging set containing over 300,000 data points for Turkish news articles.
2023	Zhang et al. [11]	LLM Analysis	A comparative analysis of the news summarization capabilities of large language models (LLMs).
2023	Çekinel et al. [12]	FCTR	Developing datasets for fact-checking Turkish news articles.

Table 2. Comparative summary of Turkish news data and LLM evaluation studies (2024–2025)

Year	Reference	Model	Scope and Key Contribution
2024	Üyük et al. [13]	Neural Gen.	Discussing the quality and accuracy of LLM outputs in Turkish news production.
2024	Güven [5]	0/F-Shot	The prompt-based learning capacity of LLMs in Turkish news classification.
2024	Yüksel et al. [14]	TurkishMMLU	A multi-tasking evaluation of Turkish language models with over 10,000 questions.
2024	Kuş & Acı [15]	Hybrid Method	The combined use of statistical methods and LLMs in Turkish summarization.
2024	Kaya & Kayalı [16]	HTV-News	A dataset designed to address the data scarcity in the field of Turkish news summarization.
2024	Uluduğan [17, 18]	TURNA/VBART	Architectural development and testing of indigenous and national Turkish major language models (LLM).
2025	Er et al. [19]	CETVEL	A comprehensive assessment of Turkish LLM performance, including 23 different tasks.
2025	Bayram et al. [20]	TR-MMLU	A benchmark developed to assess the academic and linguistic success of Turkish language models.
2024	Gu et al. [8]	Unified Prompting	Demonstrating multi-task success in few-shot scenarios without fine-tuning.
2024	Jin et al. [21]	LLM-as-a-Judge	Validating LLMs as reliable evaluators compared to human experts.
2025	Aly et al. [22]	News Summarization	Proving LLMs reach human-level quality in news via prompt engineering.

The studies summarized in Tables 1 and 2 demonstrated that Turkish news data plays a crucial role in evaluating the performance of large language models. However, these studies generally require fine-tuning of different models for each task. Current research shows that LLMs can perform multiple tasks simultaneously, even in few-shot scenarios, thanks to combined pipeline learning [21]. Nevertheless, evaluation approaches that address both summarization and label extraction in Turkish news data with a prompt-based combined pipeline structure are quite limited. However, evaluation approaches that utilize a prompt-based unified pipeline for both summarization and tag extraction in Turkish news data are limited.

This study proposes a zero-shot and few-shot evaluation methodology capable of extracting both summaries and tags from Turkish news data using a unified pipeline. This framework utilizes both commercial LLMs and open-source models. Comprehensive evaluations are conducted using BERTScore and ROUGE metrics for summary performance and F1 score, precision, and recall metrics for tag extraction. In recent years, alongside text-based or semantic metrics, the LLM-as-a-judge method, where LLMs act as referees, has become a reliable standard in the literature due to its scalability and high correlation with human judgment [22]. Furthermore, the model evaluation processes are supported by

the LLM-as-a-judge approach, where LLMs serve as referees to evaluate the quality of the results produced by LLMs, along with traditional metrics [23].
This study aims to measure the performance of LLMs across multiple tasks, specifically summarization and tag extraction, using a prompt-based unified pipelined learning approach. Additionally, it provides a fundamental contribution to the literature on the effectiveness of prompt-based approaches in the Turkish language by enabling faster, more accurate, and highly reliable results with optimized cost-effectiveness.

1.1. Research contribution

This study offers the following contributions to the existing approaches in the literature:

- A system was established in which the model can execute summarization and tag extraction tasks simultaneously within a unified pipeline. This pipeline provides an efficient process by enhancing the model's performance across multiple tasks.
- Models were adapted to the tasks using a prompt-based approach that does not require large amounts of data, thus reducing the need for extensive data collection processes. This leads to lower data collection overhead.
- The model's performance on tasks in both zero-shot and few-shot learning approaches was evaluated. This evaluation provides insight into the model's learning process with and without task-specific examples.
- Using commercial and open-source models, the performance of different model architectures on Turkish language tasks was demonstrated. This evaluation provided insights into the cost, availability, and accuracy outcomes of various models.
- Models were evaluated not only with traditional metrics but also with the LLM-as-a-judge approach, where the models assess themselves as judges. Thus, model performance is also evaluated in terms of semantic accuracy, increasing reliability.
- It contributes to the Turkish natural language processing literature by investigating the multitask performance of LLMs in the Turkish language with a prompt-based unified pipeline.

1.2. Paper organization

This study consists of four main sections. The first section explains the main problem of the research, highlights the importance of the topic, and the source of motivation. In line with the increase in LLM-based studies in recent years, the literature on LLM studies on the Turkish language and news articles is reviewed in detail. The Materials and Methods section includes the dataset used in the analysis, displays the dataset distribution figure, and describes the data preprocessing steps. Then, the research methodology is also described, including the infrastructure and implementation of the models employed. In the third section, the experimental results are demonstrated, and the model performance is evaluated in detail using ROUGE and BertScore metrics for the summarization task, and F1 score, precision, and recall metrics for the tag extraction task. In the final section, the overall results of the study are summarized, the contribution to the literature is highlighted, and recommendations for future research are presented.

2. Materials and Methods

2.1. Dataset description

In this study, a specialized dataset of Turkish news texts compiled from the TRT Haber [24] News website was employed. The dataset consists of 2000 rows and 5 columns, containing samples from different news categories between January 2025 and November 2025. The dataset incorporates essential news features such as the main news text, title, category, and date. Moreover, these basic features, the dataset also includes target variables used in model evaluation, such as gold summaries and tags for news summarization and tag extraction tasks. The reference summaries and tags used consist of spot texts and tags prepared by the news site's editors. To ensure the reproducibility of the study and to present the dataset within a scientific framework, the lengths of the news texts, reference summaries, and reference tags were analyzed. The statistical details of the dataset are presented in Table 3, showing that the average length of the news texts was 322.62 words, and the average length of the reference summaries was 24.48 words. Furthermore, it was observed that each news text contained an average of 2.77 tags for the tag extraction task.

Table 3. Descriptive statistics of the dataset

Feature	News Article (Source)	Ref. Summary (Target)	Reference Tags
Average Word Count	322.62	24.48	–
Average Character Count	2,554.30	193.03	–
Average Number of Tags	–	–	2.77
Total Number of Samples	1500	1500	1500
Number of Categories	15	–	–
Language	Turkish	Turkish	Turkish

A comprehensive overview of the distribution of the datasets collected from Turkish news texts is illustrated by Figure 1, and the dataset has a total of 15 categories. As seen in Figure 1, news is structured with a balanced distribution among the following categories: World (7.7%), Turkey (7.7%), Economy (7.3%), Education (6.9%), and others. This structure aims to measure the adaptability of the models to different types of news.

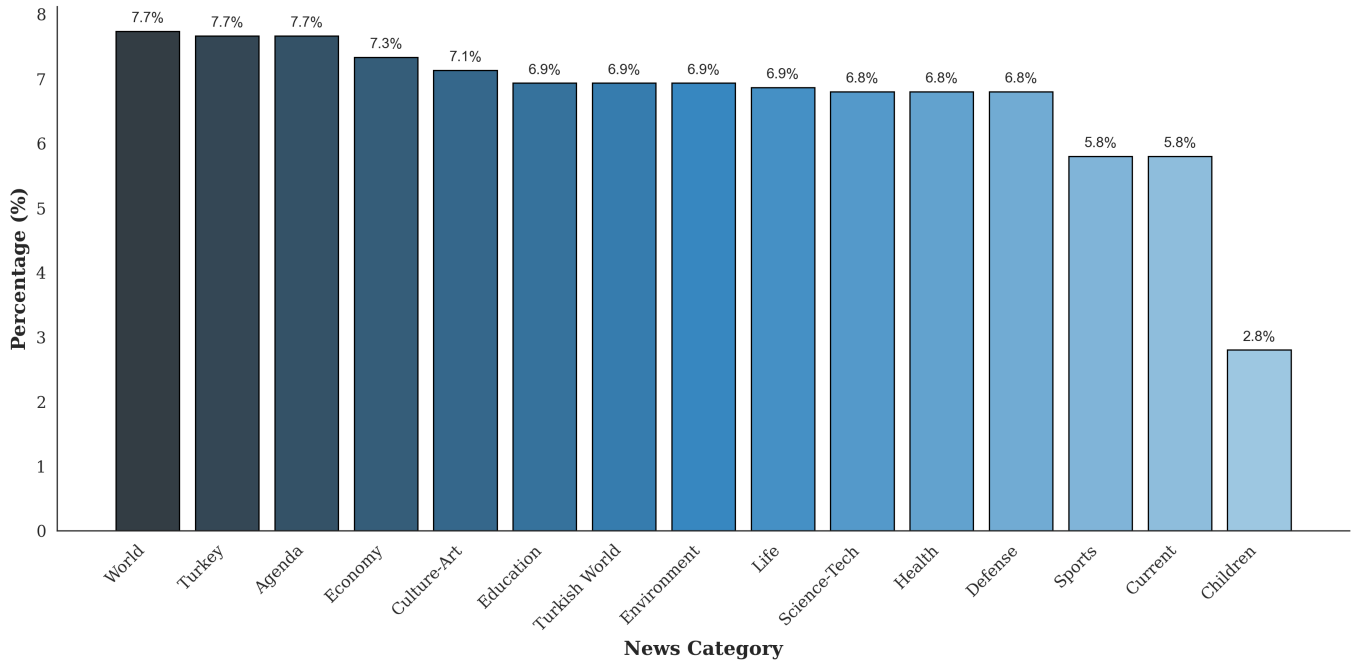

Figure 1. End-to-end pipeline for prompt-based summarization and tag extraction

2.2. Data pre-processing

Pre-processing operations were executed on the dataset so that the models could better understand the news texts and produce meaningful outputs after the learning process.

2.2.1. Content and HTML cleaning

The HTML tags and URLs within the tags of the raw news texts in the data set were cleaned.

2.2.2. Text normalization

After the cleaning process, unnecessary spaces originating from escape characters remaining in the news texts were cleaned and converted into a single space character.

2.2.3. Data splitting

Finally, the dataset was split into an evaluation set of 1500 news texts to evaluate the model performance and a sample set of 400 news texts to be presented to the models in the few-shot learning approach. In addition, care was taken to ensure that no familiar news texts were shared between the two sets.

2.3. Rationale for large language model selection

In this study, the large language models evaluated were strategically selected to compare the capabilities of current SOTA models in the literature, as well as commercial and open-source models. Among commercial models, GPT-4o, considered ground truth in the literature and demonstrating high consistency in the LLM-as-a-judge approach, and Claude Sonnet 4.5, notable for its multitasking capabilities, were preferred. For open-source models, Llama 3.3 70B Instruct and Qwen3-32B were chosen to assess their competitiveness against commercial models due to their high parameter efficiencies and superior performance capabilities.

2.4. Commercial (closed-source) models

2.4.1. GPT

GPT-4o (GPT-4 omni) is a multilingual, multimodal LLM developed by OpenAI and released in May 2024 [25] that can receive various types of data from the user, such as text, audio, images, and video, and produce text, audio, and visual outputs in the same way. The autoregressive model, called "Omni," is established to process various types of data, such as audio, images, text, and video, through the same neural network [25, 26]. The knowledge base of the model consists of information acquired from various data sources until October 2023 [26].

GPT-4o can execute tasks such as text summarization and content generation, in addition to managing user interactions. Furthermore, it has a context window of up to 128,000 tokens and an output limit of 4096. This provides the model to quickly process extensive documents or conversations in a single interaction, understand them contextually, and provide relevant answers [26, 27]. Due to the o200k tokenization tool, GPT-4o can process different languages very rapidly [28]. One study demonstrated that the GPT-4o model consumes 1.3 times fewer tokens than the GPT-4 model when processing the Turkish language. This feature of the model enables applications to execute better at cost effectiveness [26, 28]. Moreover, the GPT-4o model is 50% cheaper and 2 times faster than the GPT-4 Turbo model when used via the API [28]. In this study, the GPT-4o model was applied to evaluate the performance of commercial LLM in summarization and tag extraction tasks from Turkish news texts.

2.4.2. Claude

Claude Sonnet 4.5 is Anthropic's state-of-the-art LLM model, released in September 2025, with strong reasoning capabilities in areas such as coding, complex agentic-based tasks, and system operation [29, 30]. The model's knowledge base acquired non-public data from third parties,

publicly available internet data, datasets provided by paid contractors, data from users who have agreed to have their data used for training, and Anthropic-specific data until July 2025 [29].

Claude Sonnet 4.5 is a hybrid model that can switch between the default mode, which allows users to get quick answers, and the "extended thinking mode". The extended thinking mode is designed to handle more complex, challenging problems by enabling the model to process information for extended periods. The latter, the model presents the thinking process as an output called a "thought process" (also known as a "chain-of-thought") [29]. Sonnet 4.5 has a context window of up to 200K tokens and an output limit of 64K tokens [31]. Thanks to its high token and output capacity, the model can focus on multi-step, complex problems for more than 30 hours and sustain semantic consistency in long text processing tasks [29, 32]. Claude Sonnet 4.5 is also approximately five times cheaper than Opus 4.1 [32].

In this study, the Claude Sonnet 4.5 model was applied to evaluate the performance of commercial LLM in summarization and tag extraction tasks from Turkish news texts.

2.5. Open source models

2.5.1. Llama

Meta Llama 3.3 70B Instruct is a multimodal LLM that uses an optimized transformer autoregressive architecture with 70 billion parameters [33], released by Meta on December 6, 2024. The model was trained on a dataset of nearly 15 trillion tokens acquired from publicly available sources, licensed data, and Meta products, and had its knowledge cut in August 2023 [33].

Meta Llama 3.3 70B Instruct proved high performance in text-based tasks such as reasoning, coding, mathematics, and multilingual conversational tasks [33]. The model has a context window of up to 128K tokens, which enables the capability to process long contexts in complex tasks [33]. The model is available under the Llama 3.3 "Community License Agreement", which affords users the right to create, modify, and distribute new works, and requires compliance with the Acceptable Use Policy for model deployment [33]. In this study, the Llama 3.3 70B Instruct model was applied to evaluate the performance of an open-source LLM in summarization and tag extraction tasks from Turkish news texts. To increase the efficiency of the study and avoid hardware limitations, a 4-bit quantized version of the model optimized using the Activation-aware Weight Quantization (AWQ) method was employed.

2.5.2. Qwen

Qwen3-32B is a multilingual LLM structured as a dense architecture, released by Alibaba Cloud in May 2025, featuring 32.8 billion total parameters [34, 35]. The model has the ability to switch between "thinking mode" and "non-thinking mode" within a single architecture and demonstrates high performance in areas such as text comprehension, coding, instruction following, reasoning, and mathematics [34, 35]. The model was trained on a dataset composed of approximately 36 trillion tokens, including coding, STEM, books, multilingual texts, synthetic data, and text extracted from PDF-like documents using the Qwen 2.5-VL model [35]. Qwen3-32B implements two main techniques, Dual Chunk Attention (DCA) and YaRN (RoPE scaling), to process long contexts. This enhances the model's context processing capacity to approximately 33K tokens and extends its ultra-long context processing capacity to 132K tokens [35]. The model is accessible to users under the Apache 2.0 license, as with all Qwen 3 models [34].

In this study, the Qwen3-32B model was applied to evaluate the performance of an open-source LLM on the task of extracting summaries and tags from Turkish news texts.

2.6. Proposed method

The methodological framework established in this research focuses on evaluating the performance of LLMs in multiple tasks, specifically the summarization and tag extraction of Turkish news texts, via a prompt-based unified pipeline. This evaluation is executed using a single prompt and measures performance in terms of semantic integrity, accuracy, and consistency. The overall process is presented in Figure 2 and starts from the processing of news texts, combines the summary and tag extraction tasks with a prompt-based unified pipeline, and concludes with the generation of the summarization and tags for these tasks. The prompt-based pipeline stage is structured into two fundamental approaches

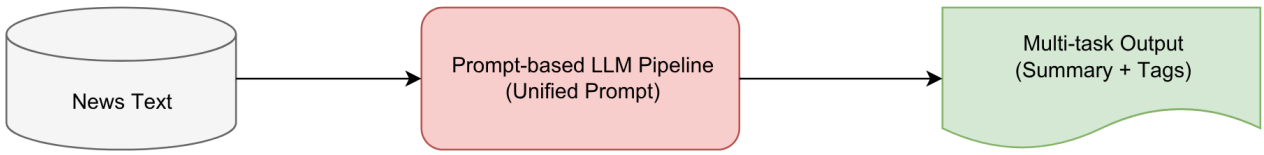

Figure 2. End-to-end pipeline for prompt-based summarization and tag extraction

and consists of two scenarios, Zero-Shot Evaluation and Few-Shot Evaluation, as presented in Figure 3 and Figure 4, respectively. The first evaluation approach of the methodology, the zero-shot learning approach, is illustrated in Figure 3. This approach demonstrated the internal structure of the "Prompt-based LLM Pipeline" shown in Figure 2. The methodology aims to measure how effectively the models execute the desired task through an instruction prompt that only describes the task, without providing any examples for summarization and tag extraction tasks.

In contrast, the zero-shot learning approach and the few-shot learning approach are illustrated in Figure 4. This methodology aims to condition models to execute a task by providing them with a few examples. In this approach, in addition to the instruction prompt explaining the task to the model, two separate datasets, comprising several additional examples, are provided. This approach aims to measure how task examples affect the model's performance.

This study aims to evaluate the multitasking performances of both open-source and commercial LLMs with a unified prompt-based pipeline using zero-shot and few-shot learning approaches, and evaluate the effectiveness of these two learning methodologies.

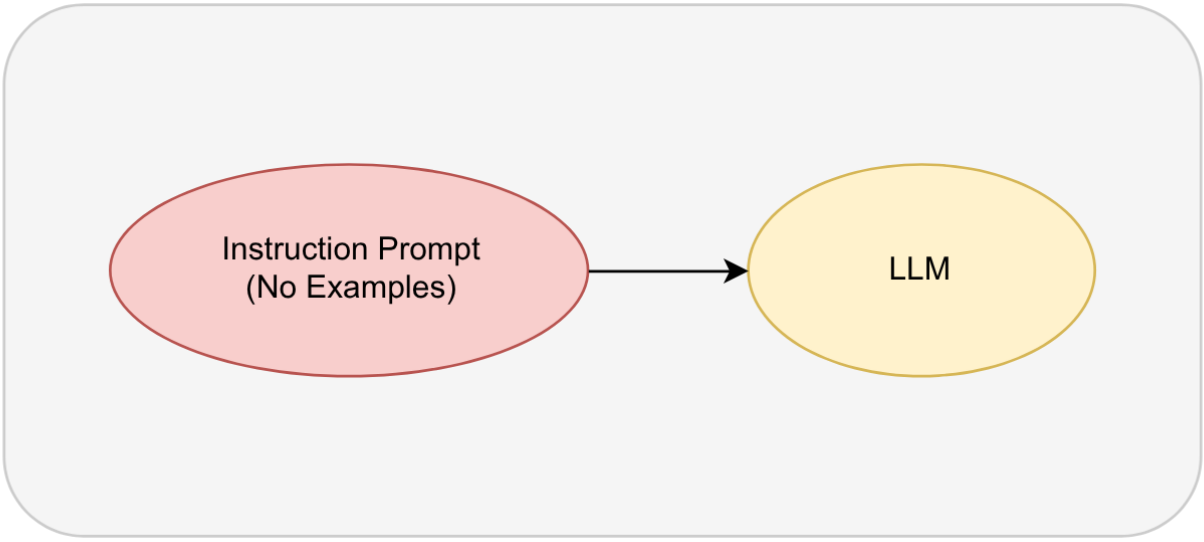

Figure 3. Zero-shot evaluation approach

3. Results and Discussion

3.1. Experimental design

In this study, a comprehensive experimental setup was built with two proposed strategic approaches to evaluate the effectiveness of the prompt-based unified pipeline architecture and to assess the multitasking performance of LLMs in Turkish news summarization and tag extraction tasks. The Zero-shot strategy in Figure 2 and the Few-shot strategy in Figure 3 were applied to four different models. The GPT-4o and Claude 4.5 Sonnet models were performed on official API providers.

The Qwen3-32B and the 4-bit quantized version of the Llama 3.3 70B-Instruct -optimized with the AWQ method- models were run on OpenAI-inference local servers with NVIDIA RTX A6000 GPUs using tensor parallelism configuration with 80% GPU memory usage on 4 GPUs using the vLLM library. To ensure deterministic behavior during the inference process, the temperature value was set to 0.1, the batch size to 10, and the maximum output length was set to a maximum of 512 tokens to avoid unnecessarily long outputs.

A customized system prompt was designed to ensure all models exhibited the same behavior and that the experiments were conducted fairly. In this prompt, the models were assigned the role of expert with the statement ”You are an expert NLP system specialized in analyzing Turkish news articles” in the system prompt. For the summarization task, considering that the ground truth summaries in the dataset consisted of spot paragraphs, the models were asked to produce a one-sentence, unbiased summary describing the main idea of the news. In the tag extraction task, the models were required to extract between 2 and 5 keywords related to the news content. To ensure the structural consistency of the model outputs, the models were required to provide the summaries and tags in JSON format.

In the Few-shot scenarios, short events, complex content, and examples from different domains were added to the system prompt to improve the model’s adaptability to different news types. In the evaluation phase, a hybrid evaluation approach combining traditional and semantic metrics was executed.

3.2. Performance metrics

In this study, the performance of a unified pipeline architecture was evaluated using ROUGE and Bertscore metrics (syntactic and semantic similarity metrics) for the summarization task, and F1, Precision, and Recall metrics for the tag extraction task. Moreover, the LLM-as-a-judge approach was implemented for both summarization and tag extraction tasks. Additionally, the meaning of the metrics, their mathematical definitions, and the reasons for their preference in this study have been explained in detail. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a metric that indicates the syntactic success of the model by measuring the quality of the generated abstracts, the amount of overlap between units such as n-grams, word sequences and word pairs with the reference abstracts. ROUGE-L calculates a weighted harmonic mean by dividing the length of the Longest Common Subsequence (LCS) between the reference abstract and the candidate abstract by the length of the reference and candidate abstracts [36]. Equation 1 shows the calculation of the ROUGE-L metric:

$$
\begin{aligned}
R_{lcs} &= \frac{LCS(X,Y)}{m} \\
P_{lcs} &= \frac{LCS(X,Y)}{n} \\
F_{lcs} &= \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}
\end{aligned}
\tag{1}
$$

BertScore is a metric that calculates the cosine similarity of context-sensitive vectors derived from pre-trained BERT models instead of n-gram-based word matches [37]. This metric, which measures how semantically close the generated summary is to the reference summary, is

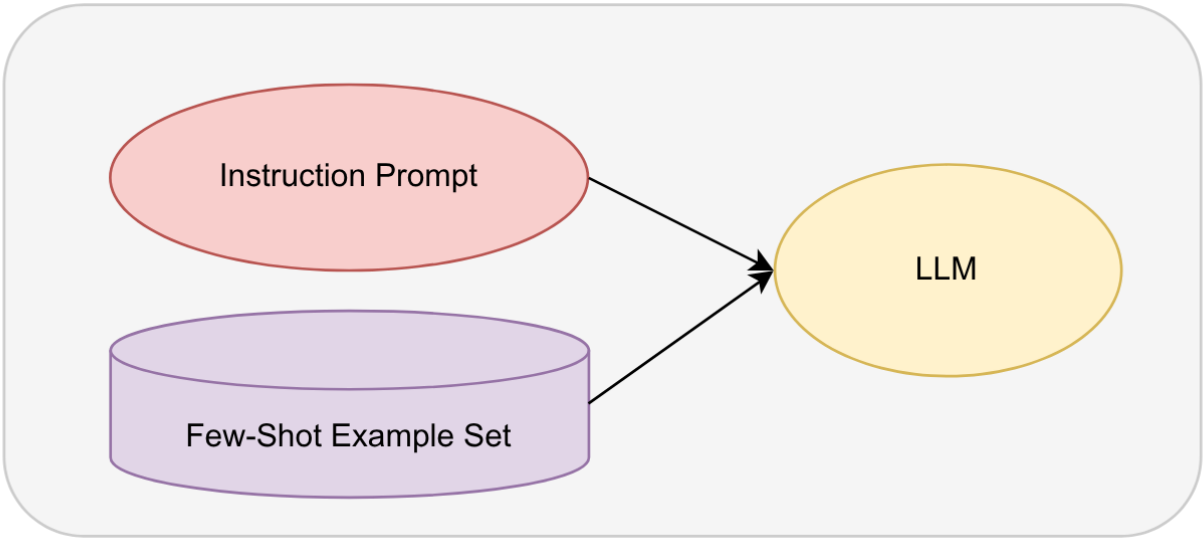

Figure 4. Few-shot evaluation approach

shown in Equation 2:

$$
\begin{aligned}
R_{\text{BERT}} &= \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \\
P_{\text{BERT}} &= \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \\
F_{\text{BERT}} &= 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}
\end{aligned}
\tag{2}
$$

In this study, the dbmdz/bert-base-turkish-cased model was applied as the basis for BertScore calculation to measure the semantic similarity of the generated summary with the reference summary.
Precision is the metric that measures how many of the positive predictions made by the model are actually correct [38]. Equation 3 shows how to calculate the precision value:

$$
\text{Precision} = \frac{TP}{TP+FP} \tag{3}
$$

This study executed the model to test whether it generated unnecessary or erroneous tags.
Recall is a metric that shows the percentage of true positive cases detected by the model [38]. Equation 4 shows how to calculate the recall value:

$$
\text{Recall} = \frac{TP}{TP+FN} \tag{4}
$$

F1 score is a metric that is the harmonic mean of Precision and Recall and indicates the balanced performance of the model [38]. Equation 5 shows how to calculate the F1 score:

$$
\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}
$$

In this study, it was applied to express overall success in cases where the number of tags varied.
LLM-as-a-judge was used to measure semantic depth. To ensure fair, reliable, and consistent comparisons between models, 150 randomly selected samples were determined from the test dataset, and the models' responses to these inputs were analyzed. In this evaluation phase, the GPT-4o model was assigned the role of an impartial judge in the system prompt, stating, "You are an expert NLP evaluator acting as an impartial linguistic judge." The judge model was presented with the news text, ground truth summaries, ground truth tags, and the outputs produced by the models simultaneously, ensuring that the evaluation was source-sensitive.
For summarizing performance, the source news text was considered the primary true source. The model was asked to evaluate the summaries in terms of accuracy and semantic meaning, and to keep the scores low if hallucinations were present in the summaries. For tag extraction performance, scoring was again based on semantic relevance. All evaluations were rated on a scale of 1-10. Furthermore, to ensure transparency, the model was required to provide justification for each score it assigned and to present its outputs in JSON format.

3.3. Experimental results

Table 4. Comparative analysis of summarization and tagging performance across different shot strategies

	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Judge (Sum)	Tag F1	Tag Prec.	Tag Rec.	Judge (Tag)
Claude 4.5 Sonnet (0-Shot)	0.407	0.239	0.331	0.662	7.610	0.245	0.199	0.319	6.840
Claude 4.5 Sonnet (1-Shot)	0.415	0.249	0.342	0.670	7.570	0.316	0.260	0.403	6.880
Claude 4.5 Sonnet (3-Shot)	**0.423**	**0.257**	**0.351**	**0.676**	**7.670**	**0.333**	**0.283**	**0.405**	**6.990**
GPT-4o (0-Shot)	0.383	0.216	0.313	0.654	7.350	0.216	0.172	0.293	6.570
GPT-4o (1-Shot)	0.387	0.223	0.317	0.658	7.370	0.252	0.199	0.345	6.560
GPT-4o (3-Shot)	0.383	0.220	0.315	0.657	7.370	0.265	0.222	0.326	6.680
Llama 3.3 (70B) (0-Shot)	0.372	0.217	0.310	0.644	6.650	0.239	0.197	0.304	6.040
Llama 3.3 (70B) (1-Shot)	0.384	0.227	0.316	0.650	6.810	0.254	0.203	0.336	6.030
Llama 3.3 (70B) (3-Shot)	0.385	0.229	0.321	0.652	6.850	0.287	0.244	0.348	6.300
Qwen3 (32B) (0-Shot)	0.370	0.211	0.302	0.645	6.970	0.181	0.140	0.252	6.240
Qwen3 (32B) (1-Shot)	0.374	0.212	0.303	0.647	7.050	0.231	0.180	0.324	6.290
Qwen3 (32B) (3-Shot)	0.371	0.210	0.302	0.649	6.910	0.266	0.232	0.310	6.430

The experiments were conducted on four different language models and three different learning strategies: zero-shot, one-shot, and three-shot. As shown in Table 1, which presents a comparative analysis of model performance, the Claude 4.5 Sonnet model outperformed the other models in all metrics, including ROUGE, BertScore, F1, and LLM-as-a-judge.

Among the commercial models, Claude 4.5 Sonnet achieved the best model in the summarization task with a BertScore of 0.676 and an LLM-as-judge score of 7.67 in the 3-shot scenario, while it was the most successful model in the tag task with an F1 score of 0.333 and an LLM-as-judge score of 6.9. GPT-4o followed closely behind with a BertScore of 0.657 in the summarization task and an LLM-as-judge score of 7.37 in the tag task, and an LLM-as-judge score of 6.68 in the tagging task. Both models also demonstrated, through their outputs, that commercial models are effective not only in generating text but also in extracting summaries and tags from news texts.

In contrast, among open-source models, the Llama 3.3 model in the 3-shot scenario showed performance close to commercial models with a BertScore of 0.652 in the summarization task. It also outperformed the GPT-4o model in the tagging task with an F1 score of 0.287. On the other hand, the Qwen3 model achieved remarkable success in the summarization task with an LLM-as-a-judge score of 7.05 in the one-shot scenario, while in the tagging task, it struggled with complex tasks with an F1 score of 0.181 in the zero-shot scenario, but its performance improved when the number of samples was increased. These results show that open-source models, with efficient system prompts, can compete with commercial models in complex information extraction tasks.

From the presented data in Table 1, it was observed that the models showed a linear increase in summarizing tasks when switching from the zero-shot strategy to the few-shot strategy. In the tagging task, a performance increase of up to 35% in F1 scores was observed, indicating that the models better adapted to the expected format and understood the semantic meanings of the tags thanks to the given examples.

3.3.1. Summarization task analysis

In evaluating the summarizing performance of the models, syntactic and semantic similarity metrics such as ROUGE and Bertscore were applied, as well as the LLM-as-a-judge method, which simulates human judgment.

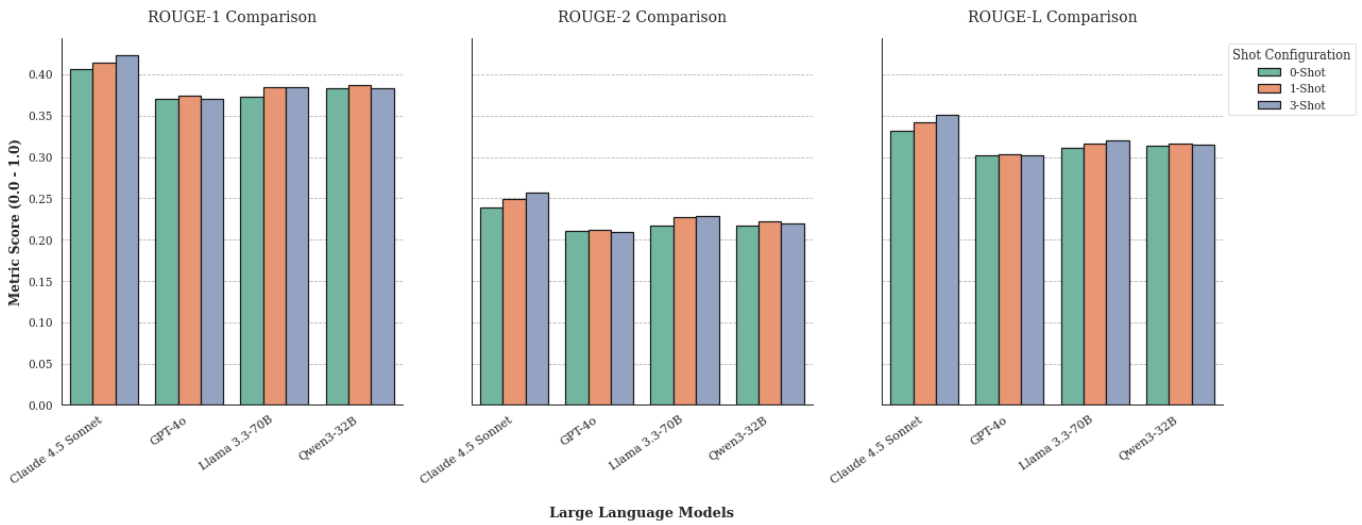

Figure 5. Syntactic summarization performance of models based on ROUGE metrics

Syntactic Performance Analysis In the first stage of the summarization task, the quality of the summaries produced by the models was measured, specifically the degree of overlap with reference summaries on units such as n-grams and word sequences. As shown in Figure 5, Claude 4.5 demonstrated its syntactic ability by achieving a ROUGE-L score of 0.351 in the Sonnet 3-shot scenario and the highest performance in all scenarios. Moreover, the open-source Llama 3.3 model, closely followed by Claude, with a ROUGE-L score of 0.321, exceeded the GPT-4o model, which had a score of 0.315. The Qwen model lagged behind with a score of 0.302. However, switching to a few-shot strategy

increased the ROUGE scores of all models. These results proved that the Llama 3.3 model offers an alternative to commercial models in the Turkish language news summarization task.

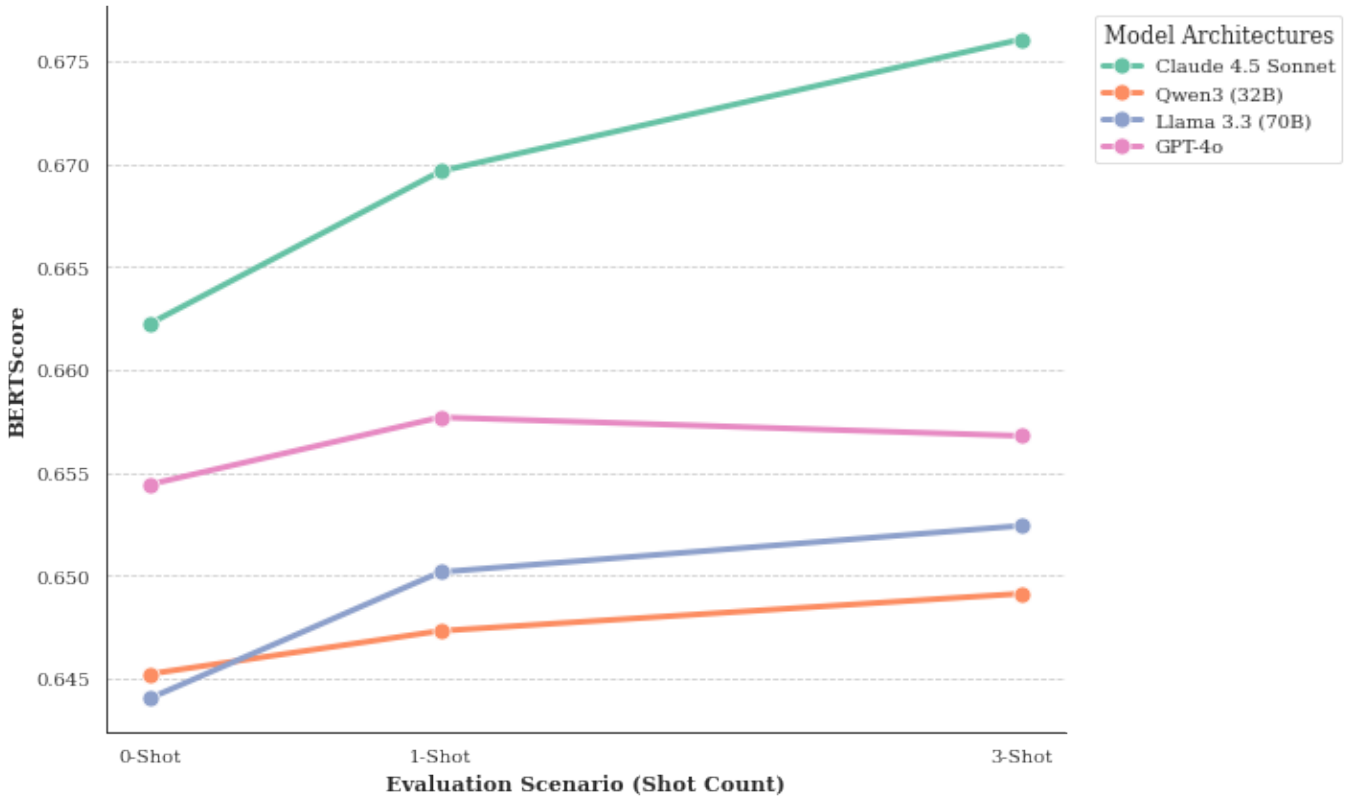

Figure 6. Semantic consistency analysis of models based on BERTScore

Semantic Consistency Analysis In the semantic evaluation, the second stage of the summarization task, as shown in Figure 6, Claude 4.5 Sonnet 3-shot achieved the best results with a BertScore of 0.676, demonstrating its superior ability to preserve the semantic integrity of the text. Unlike the ROUGE metric, the Llama 3.3 model received a BertScore of 0.652, falling slightly behind its competitor GPT-4o. Qwen3 was also the model with the lowest scores in semantic analysis. Furthermore, the graph lines shown in Figure 6 indicate that the semantic score boosts as the number of samples given to the models increases. These results proved that open-source models perform almost equivalently to commercial models in capturing the meaning of the text.

Model-Based Evaluation The LLM-as-a-judge method, which is the final phase of the summarization task, scored the accuracy and comprehensiveness of the models on a 1-10 scale, and the results shown in Figure 7 were obtained. In this stage, where GPT-4o presented as the judge, the Claude 4.5 Sonnet model produced the most reliable and accurate results with a score of 7.67 in the 3-shot scenario, followed by GPT-4o in second place with 7.37 points. Among open-source models, Qwen3 scored 6.91, exceeding the Llama 3.3 model, which scored 6.85. Furthermore, switching to a few-shot strategy increased the models' scores linearly. These results presented that commercial models produce accurate answers with a low tendency towards hallucinations, but open-source models, when proper prompting techniques are applied, can produce consistent summaries above a certain standard.

3.3.2. Tag extraction analysis

In evaluating the tag extraction performance of the models, metrics such as F1, Precision, and Recall were used, along with the LLM-as-a-judge method, which simulates human judgment.

Scenario-Based Performance Evolution In the first stage of the tag extraction task, the performance of the models in zero-shot, one-shot, and three-shot scenarios was investigated in the radar graphs shown in Figure 8. As seen in Figure 8, the effect of providing samples to the models was also numerically confirmed. In the zero-shot scenario, the lowest F1 score was 0.181 for the Qwen3 model, and the highest was 0.245 for the Claude 4.5 Sonnet model. However, in the three-shot scenario, the scores of the models increased, and the Claude 4.5 Sonnet model, with an F1 score of 0.333, engaged the largest area in the graph. Looking at the three-shot scenario, the open-source Llama 3.3 model, with an F1 score of 0.287, performed better than the commercial model GPT-4o. These results presented that the adaptability of open-source models, when the number of samples is increased, is at a level that can compete with commercial models.

Precision and Recall Analysis In the second stage of the tag extraction task, a precision-recall analysis was performed in a 3-shot scenario to measure how the models could use the provided samples and how they behaved. As shown in Figure 9, the Claude 4.5 Sonnet model displayed its ability to generalize to the given samples with precision values of 0.283 and recall values of 0.405. In contrast, the GPT-4o model lagged behind open-source models, showing poor performance in precision and recall metrics. In contrast, the Qwen3 open-source model exhibited more economical behavior, while the Llama 3.3 model showed its reliability with precision values of 0.245 and recall values of 0.350.

Semantic Accuracy and Qualitative Evaluation In the final stage of the tag extraction task, the tag quality and semantic accuracy of the models were scored on a scale of 1-10 using the LLM-as-a-judge approach, yielding the results shown in Figure 10. In this stage, where GPT-4o

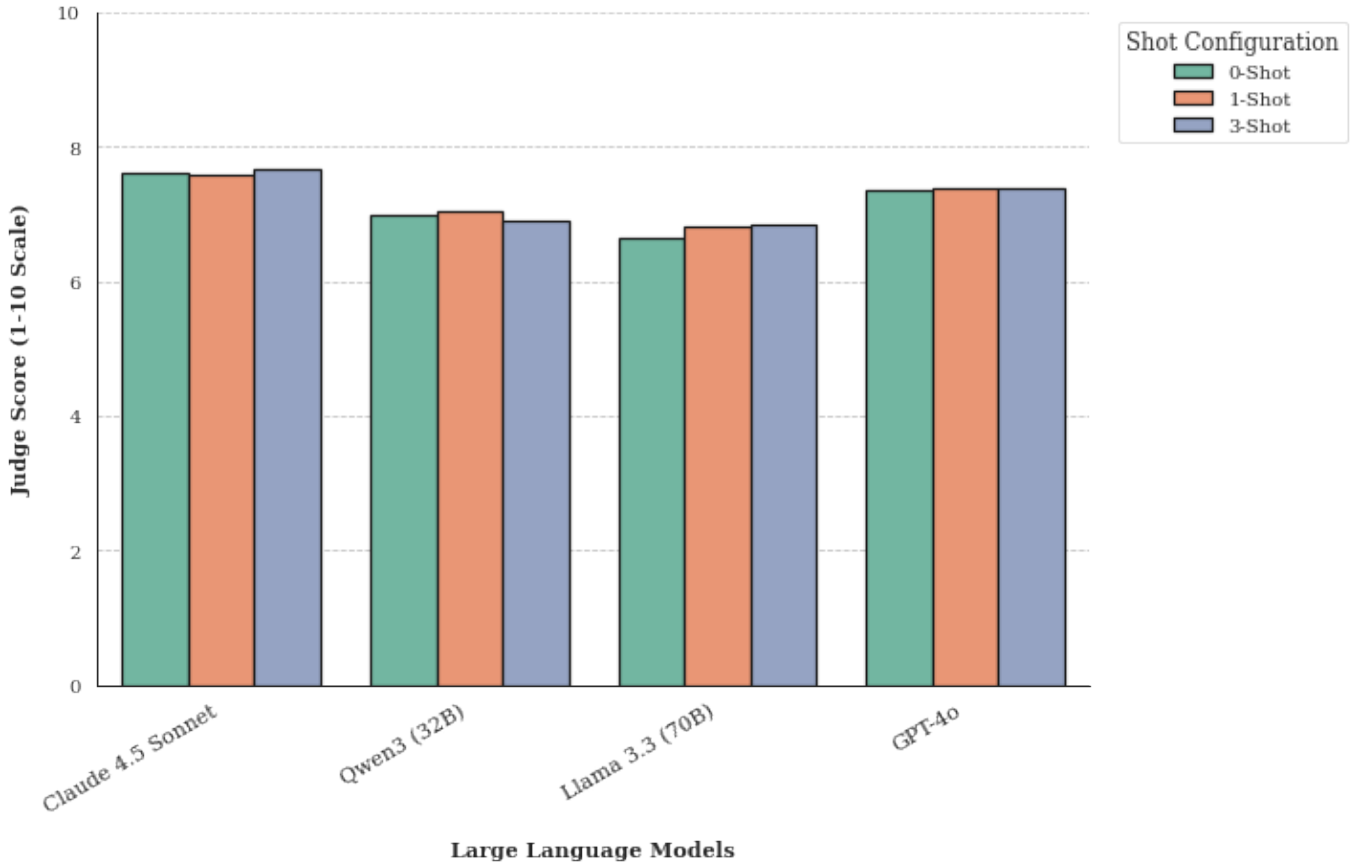

Figure 7. Qualitative summarization analysis of models based on LLM-as-a-Judge

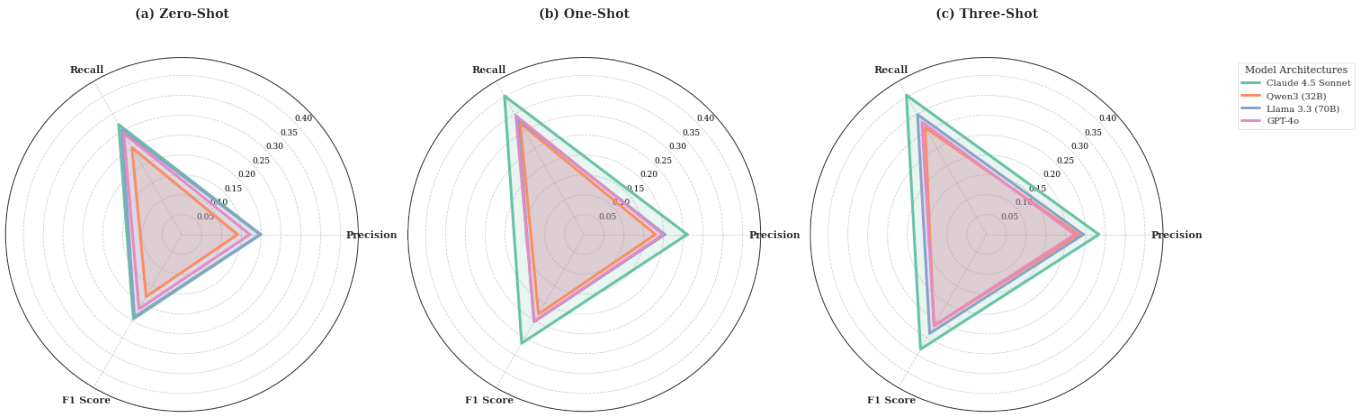

Figure 8. Evaluation of tag extraction performance across different shot scenarios

acted as the judge, the Claude 4.5 Sonnet model, in the 3-shot scenario, was the best tag-producing model with a score of 6.99, as well as in other metrics. The GPT-4o model followed with a score of 6.68. Among open-source models, the Llama 3.3 model scored 6.30, while the Qwen3 model, with a score of 6.43, encouragingly exceeded the Llama 3.3 model in semantic tag generation quality. However, as the number of examples given to the models increased, the scores of all models improved. These results offered an important insight into observing the differences between automated metrics and semantic evaluation.

3.3.3. Cost-performance and efficiency analysis

In addition to measuring the models with technical metrics, the relationship between cost and quality was analyzed to evaluate their applicability in real-world scenarios. The cost values shown in Figure 11 present the evaluation cost of the GPT-4o model, which is the peer reviewer, depending on the length of the summaries and tags produced by the models. In this respect, as seen in the graph, the Claude 4.5 Sonnet model offered the highest quality but also incurred the highest cost. However, the GPT-4o model lagged behind despite being high-cost like Sonnet. Among open-source models with zero operational cost, Qwen3, despite being a small model, surpassed the Llama 3.3 model with high quality. Especially in the tag generation task, the Qwen3 model produced higher-quality tags than the Llama 3.3 model, demonstrating its usability in the operational field. These results showed that the competition between open-source models and commercial models is becoming increasingly important. On the other hand, it shows that they are becoming an alternative in situations where cost is important.

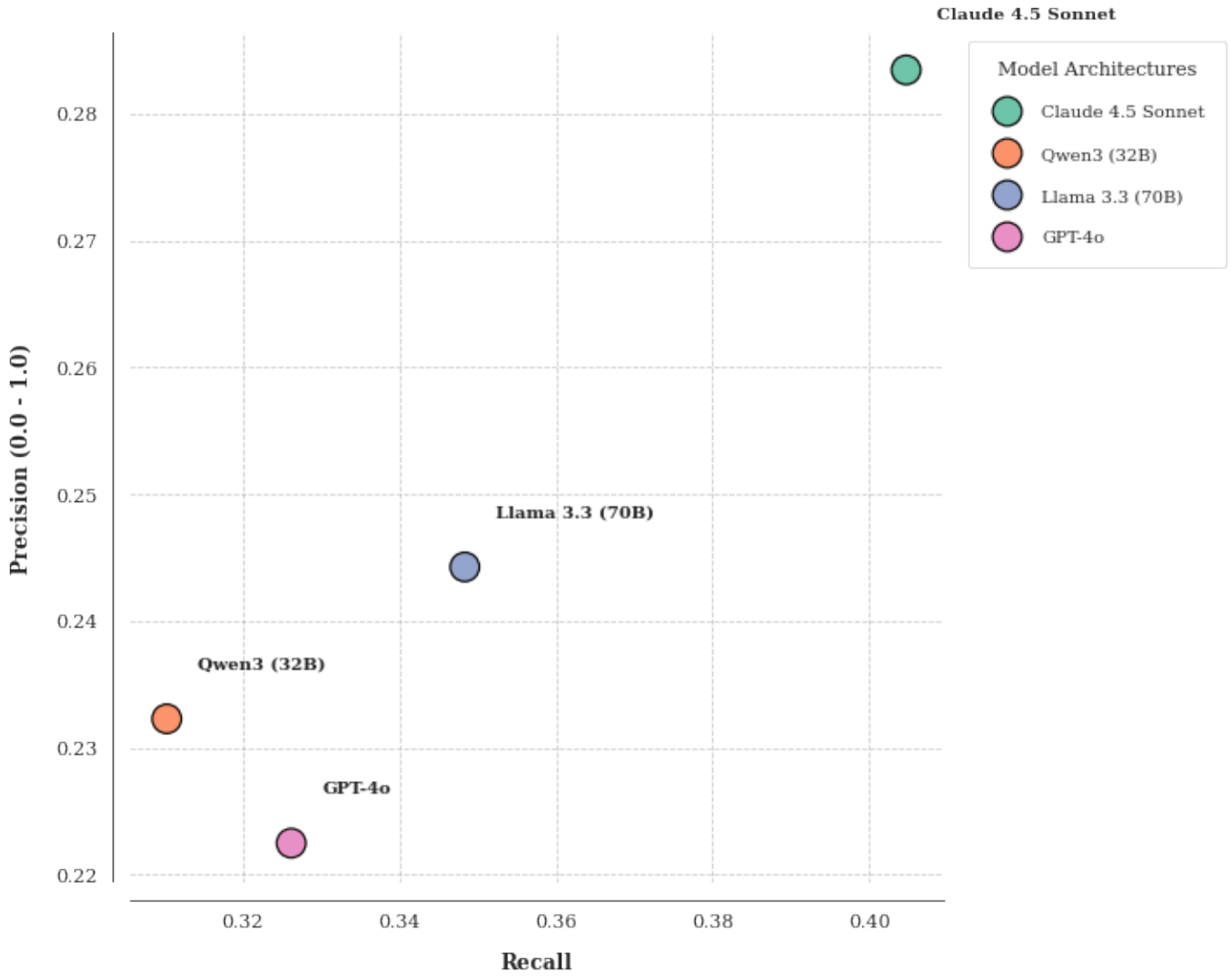

Figure 9. Precision - Recall trade-off of models in 3-shot

3.4. Discussion

The experimental results obtained in this study present both significant similarities and original contributions to the literature when compared with existing studies. The increase in the accuracy rate of all models when switching from the zero-shot strategy to the few-shot strategy coincides with the learning findings in Güven's [5] study on evaluating the zero-shot and few-shot learning performance of LLMs for the Turkish news classification task. In particular, the 35% increase in F1 score in the tag extraction tasks showed that the LLMs adapt the models to the task better with the examples given in the Turkish language.

The success of commercial models such as Claude 4.5 Sonnet and GPT-4o in the summarization task is consistent with the emphasis on the success of commercial models in Turkish news summarization tasks by Zhang *et al.* [11]. In addition, the fact that the Llama 3.3 model outperformed the GPT-4o model in the tagging task shows that open-source models optimized for languages can be a strong alternative to commercial models.

Many studies in the literature treat news summarizing and tagging tasks as separate processes. In this study, the proposed unified pipeline enables the summarizing and tagging tasks in news data using a prompt-based approach without requiring large amounts of data. This provides a unique contribution to the literature in terms of efficiency and cost for LLMs to perform multitasking.

In model evaluation processes, the positive correlation between classical metrics and LLM-as-a-judge scores supports the comprehensive evaluation approaches proposed in the CETVEL study presented by Er *et al.* [19]. This demonstrates that using human-like assessment metrics together with classical metrics provides a more reliable and holistic evaluation framework.

Finally, the cost-performance analysis presented in the study shows that open-source models like Qwen3-32B can be a strong alternative to commercial models in tasks requiring high accuracy. This demonstrates that open-source models are a powerful and cost-effective solution for those with cost constraints.

Overall, through its methodological, experimental results and their relationship with the existing literature, this study offers significant theoretical and practical contributions to the effective use of LLMs in the field of Turkish news reporting.

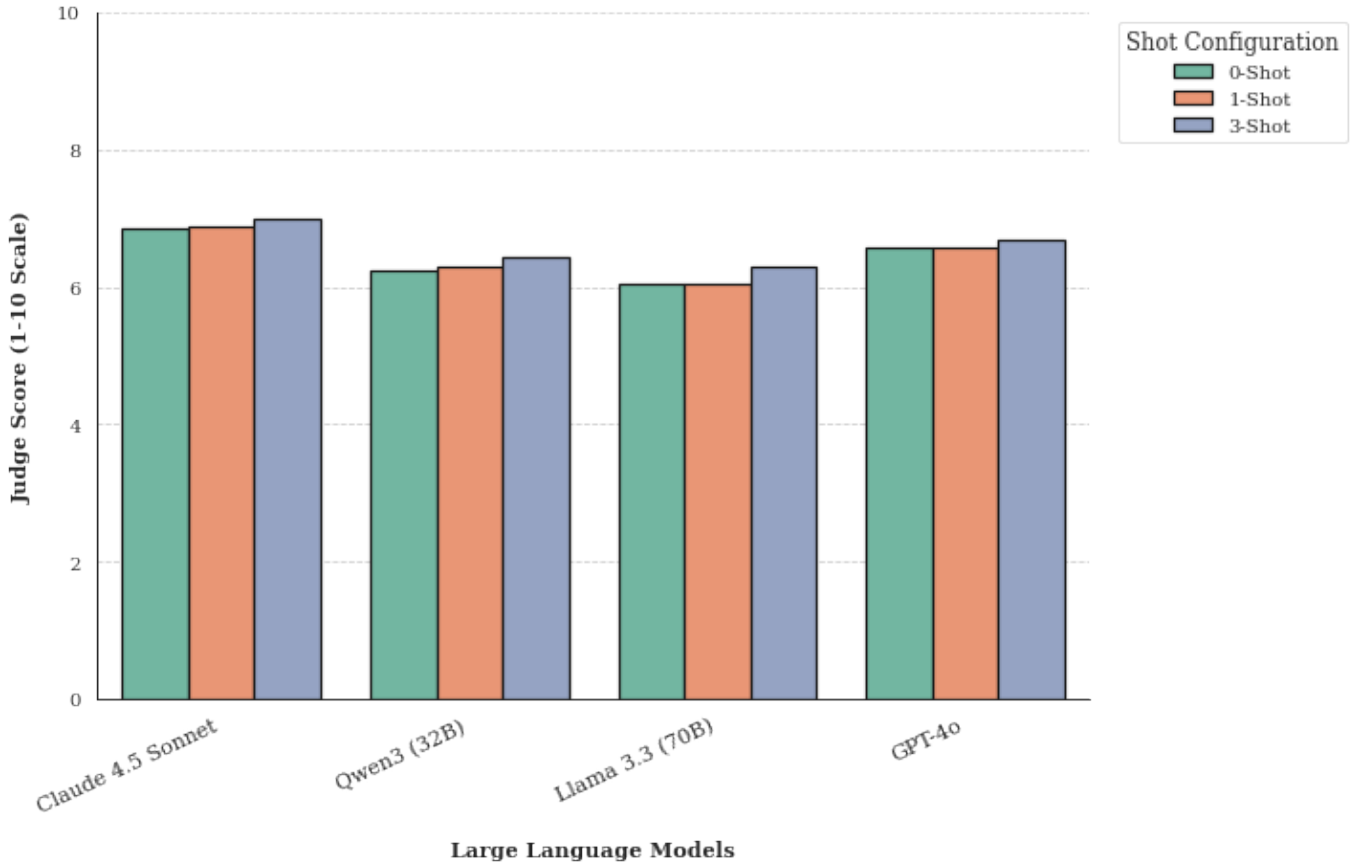

Figure 10. Analysis of tag quality and semantic accuracy via the judge model

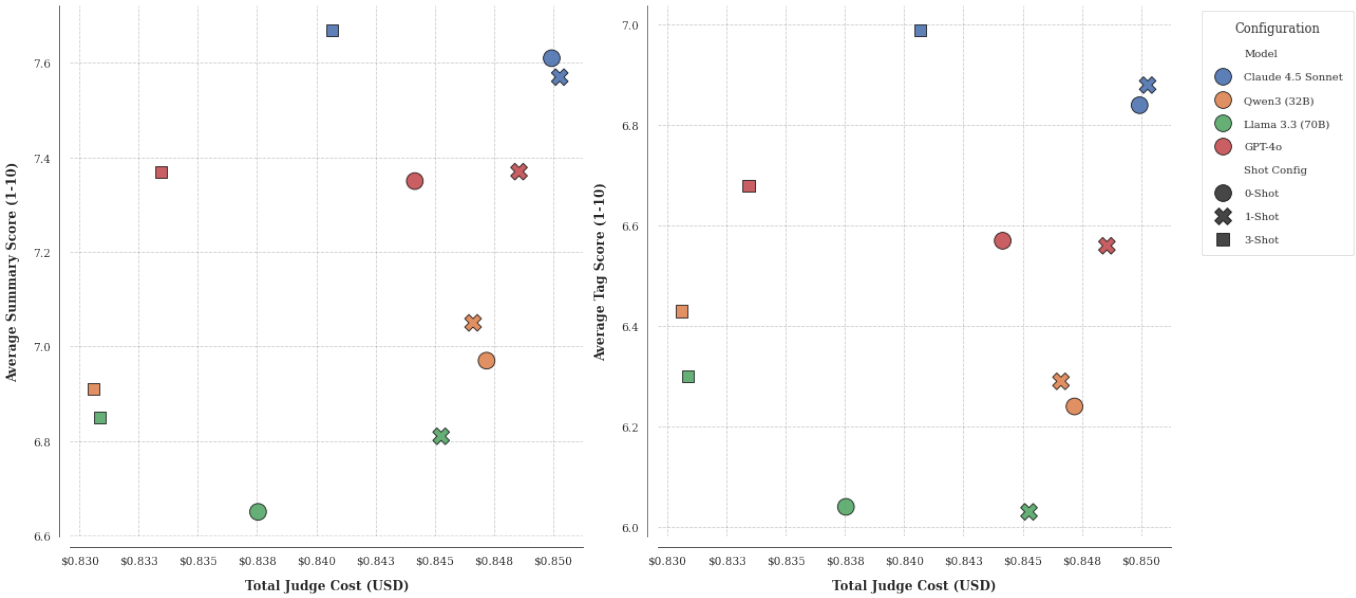

Figure 11. Cost vs. quality evaluation of models

4. Conclusion

This study presents a comparative analysis evaluating the multitasking performance of LLMs with a prompt-based unified pipeline in Turkish news summarization and tag extraction tasks. Unlike existing studies in the literature, this approach, which combines summarization and tag extraction tasks with a single prompt, minimizes high data requirements and increases efficiency in operational processes. The presented methodological framework offers a unique solution for producing high-accuracy output in low-resource scenarios.

In the summarizing and tag extraction tasks, the zero-shot learning strategy was implemented without providing any examples to the models, while in the one-shot and three-shot scenarios, examples were provided to the models to facilitate their adaptation to the task. In the summarization task, Claude scored 7.67 on a 1-10 scale in the judge's evaluation in the 4.5 Sonnet 3-shot scenario. In the tagging task, it exhibited the best

performance with an F1 score of 0.333. The fact that switching to the few-shot strategy in the tag extraction and summarization tasks linearly increased the scores of all models confirms the critical role of few-shot learning in the adaptation of LLMs to Turkish news data.
In open-source modeling, Qwen3 outperformed Llama 3.3, which has approximately twice as many parameters, in tagging and summarization tasks. Qwen3's ability to produce more specific outputs with fewer tokens, its cost-free operation on local servers, and its as a commercially superior Claude model in cost-effective situations demonstrate its strength. This data serves as a guide for model selection for institutions and researchers with hardware constraints for various needs.
The positive correlation identified between classical metrics and LLM-as-a-judge scores demonstrates the reliability of assessment metrics that simulate human judgment. This approach shows that LLM-based peer review models can be used as a holistic validation tool in conjunction with traditional metrics in future Turkish NLP projects.
In conclusion, the proposed framework, with its unified pipeline, offers a prompt-based approach for summarizing and tag extraction tasks in Turkish news data without requiring large amounts of data, thus making a consequential contribution to the multitasking performance of LLMs. Furthermore, this study provides a comprehensive methodological foundation for similar multitasking scenarios in the field of Turkish natural language processing using LLMs. The study offers a reliable, fast, and scalable automation solution for news agencies and media monitoring centers. Integration of the proposed framework into real-time news analysis systems has the potential to optimize manual workload in content processing processes. Future studies can evaluate the generalizability of this framework by applying it domain-specifically across different news verticals, such as economics, sports, and world news, and by supporting it with Retrieval-Augmented Generation (RAG) architectures. The aim can be to increase the interpretability and accuracy of the models by reducing hallucination rates.

Article Information Form

Author's Contributions: B.K. performed data retrieval, processing, implementation, and article writing; K.Y. supervised the research and provided technical guidance. Both authors approved the final version.

Acknowledgments: The authors would like to thank TRT Haber for the news data and the open-source community for providing the Llama and Qwen models. Special thanks to the research supervisor for the technical guidance and support throughout this work.

Conflict of Interest: The authors declare no conflict of interest.

Ethical Approval: Scientific and ethical principles were followed. No ethical committee approval was required as the study used scraped public news data without human/animal subjects.

Availability: The dataset was specifically compiled by scraping public news.

Artificial Intelligence Statement: No artificial intelligence tools were used while writing this article.

Plagiarism Statement: This article has been scanned by iThenticate™.

References

[1] B. Baykara and T. Güngör, "Abstractive text summarization and new large-scale datasets for agglutinative languages Turkish and Hungarian," *Language Resources and Evaluation*, vol. 56, no. 3, p. 973–1007, 2022. [Online]. Available: https://doi.org/10.1007/s10579-021-09568-y
[2] M. Akashvarma, G. Yashaswini, E. Keerthana, S. Siddharth, K. Saipooja, and G. Prasanna Kumar, "A comprehensive review of large language models in abstractive summarization of news articles," in *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*. IEEE, Conference Proceedings, p. 1–6.
[3] M. Cate, "The role of zero-shot and few-shot learning in enhancing llms for real-world applications," 2023.
[4] I. Borhan and A. Bajaj, "The effect of prompt types on text summarization performance with large language models," *Journal of Database Management (JDM)*, vol. 35, no. 1, p. 1–23, 2024.
[5] Z. A. Guven, "Performance comparison of large language models, gpt and gemini on turkish news classification task," 08 May 2024.
[6] Y. Chang *et al.*, "A survey on evaluation of large language models," *ACM transactions on intelligent systems and technology*, vol. 15, no. 3, p. 1–45, 2024.
[7] K. Chang *et al.*, "Efficient prompting methods for large language models: A survey," *arXiv preprint arXiv:2404.01077*, 2024.
[8] J. Gu *et al.*, "A survey on llm-as-a-judge," *The Innovation*, 2024.
[9] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano, "Mlsum: The multilingual summarization corpus," *arXiv preprint arXiv:2004.14900*, 2020.
[10] T. Hasan *et al.*, "Xl-sum: Large-scale multilingual abstractive summarization for 44 languages," *arXiv preprint arXiv:2106.13822*, 2021.
[11] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," *Transactions of the Association for Computational Linguistics*, vol. 12, p. 39–57, 2024.
[12] R. F. Cekinel, P. Karagoz, and C. Coltekin, "Cross-lingual learning vs. low-resource fine-tuning: A case study with fact-checking in turkish," *arXiv preprint arXiv:2403.00411*, 2024.
[13] C. Üyük, D. Rovó, S. Kolli, R. Varol, G. Groh, and D. Dementieva, "Crafting tomorrow's headlines: Neural news generation and detection in english, turkish, hungarian, and persian," *arXiv preprint arXiv:2408.10724*, 2024.
[14] A. Yüksel, A. Köksal, L. K. Şenel, A. Korhonen, and H. Schütze, "Turkishmmlu: Measuring massive multitask language understanding in turkish," *arXiv preprint arXiv:2407.12402*, 2024.
[15] A. Kuş and Ç. İ. Acı, "A hybrid approach to automatic text summarization of turkish texts: Integrating extractive methods with llms," in *2024 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, Conference Proceedings, p. 1–6.
[16] N. Z. Kayalı and S. I. Omurca, "Htv-news: A new dataset with high novelty rate for turkish text summarization," in *2024 9th International Conference on Computer Science and Engineering (UBMK)*. IEEE, Conference Proceedings, p. 1–6.
[17] G. Uludoğan, Z. Y. Balal, F. Akkurt, M. Türker, O. Güngör, and S. Üsküdarlı, "Turna: A turkish encoder-decoder language model for enhanced understanding and generation," *arXiv preprint arXiv:2401.14373*, 2024.
[18] M. Turker, M. E. Ari, and A. Han, "Vbart: The turkish llm," *arXiv preprint arXiv:2403.01308*, 2024.
[19] Y. A. Er, I. Kesen, G. G. Şahin, and A. Erdem, "Cetvel: A unified benchmark for evaluating language understanding, generation and cultural capacity of llms for turkish," *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1052–1085, 2026.
[20] M. A. Bayram, A. A. Fincan, A. S. Gümüş, B. Diri, S. Yıldırım, and Ö. Aytaş, "Setting standards in turkish nlp: Tr-mmlu for large language model evaluation," *arXiv preprint arXiv:2501.00593*, 2024.
[21] F. Jin, J. Lu, and J. Zhang, "Unified prompt learning makes pre-trained language models better few-shot learners," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Conference Proceedings, p. 1–5.
[22] W. M. Aly, T. H. A. Soliman, and A. M. AbdelAziz, "An evaluation of large language models on text summarization tasks using prompt engineering techniques," *arXiv preprint arXiv:2507.05123*, 2025.
[23] A. Odabaşı and G. Biricik, "Unraveling the capabilities of language models in news summarization," *arXiv preprint arXiv:2501.18128*, 2025.
[24] TRT Haber, "Trt haber web portal," https://www.trthaber.com/, 2025, accessed: December. 15, 2025.

[25] OpenAI, "Gpt-4o system card," 2024, (accessed 29 October, 2025). [Online]. Available: https://openai.com/index/gpt-4o-system-card/
[26] E. Glover, "What is gpt-4o?" 2024, (accessed 29 October, 2025). [Online]. Available: https://builtin.com/artificial-intelligence/what-is-gpt4o
[27] R. Islam and O. M. Moushi, "Gpt-4o: The cutting-edge advancement in multimodal llm," in *Intelligent Computing-Proceedings of the Computing Conference*. Springer, Conference Proceedings, p. 47–60.
[28] OpenAI, "Hello gpt-4o," 2024, (accessed 30 October, 2025). [Online]. Available: https://openai.com/index/hello-gpt-4o/
[29] Anthropic, "Claude sonnet 4.5 system card," Report, 2025, Accessed: 30 October 2025. [Online]. Available: https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf
[30] Anthropic, "Introducing claude sonnet 4.5," 2025, (accessed: 30 October 2025.). [Online]. Available: https://www.anthropic.com/news/claude-sonnet-4-5
[31] Anthropic, "Models overview," 2025, (accessed: 31 October 2025.). [Online]. Available: https://docs.claude.com/en/docs/about-claude/models/overview#model-comparison-table
[32] G. Ojeda, "Claude sonnet 4.5: Highest-scoring claude model yet on swe-bench," 2025, (accessed: 31 October 2025.). [Online]. Available: https://caylent.com/blog/claude-sonnet-4-5-highest-scoring-claude-model-yet-on-swe-bench
[33] I. Meta Platforms, "Llama 3.3-70b-instruct," 2024, Accessed: 31 October 2025. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
[34] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, and C. Lv, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.
[35] A. Cloud, "Qwen/qwen3-32b," 2025, Accessed: 4 January 2026. [Online]. Available: https://huggingface.co/Qwen/Qwen3-32B
[36] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
[37] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.
[38] K. M. Sujon, R. Hassan, K. Choi, and M. A. Samad, "Accuracy, precision, recall, f1-score, or mcc? empirical evidence from advanced statistics, ml, and xai for evaluating business predictive models," *Journal of Big Data*, vol. 12, no. 1, p. 268, 2025.