IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 August 2024, accepted 1 September 2024, date of publication 4 September 2024,
date of current version 13 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3454342

RESEARCH ARTICLE

A Comprehensive Evaluation of Large Language Models for Turkish Abstractive Dialogue Summarization

OSMAN BÜYÜK
Department of Electrical and Electronics Engineering, Izmir Demokrasi University, 35140 Izmir, Türkiye
Department of Research and Development, Sestek Speech Enabled Software Technologies Inc., 34396 Istanbul, Türkiye
e-mail: osman.buyuk@idu.edu.tr

ABSTRACT Text summarization is the task of generating a short and concise summary of a source text. In an abstractive text summarization, the generated summaries may potentially contain new phrases that do not appear in the source text. Dialogue summarization is a special case of text summarization in which the source text is a dialogue between two or more people. Dialogue summarization can be a crucial step especially when the source dialogues are complex and long such as call center conversations. Large language models (LLMs) show remarkable performance in natural language generation tasks and thus they can be a suitable modeling approach for abstractive text summarization. Although LLMs are extensively studied for common languages, there are only a few studies for underrepresented languages such as Turkish. In this paper, we make a comprehensive evaluation of LLMs for Turkish abstractive dialogue summarization. For this purpose, we translated 3 datasets in English to Turkish. Additionally, we make use of a test set that contains real call center dialogues originally collected in Turkish. In the experiments, we observe that fine-tuning LLMs to the dialogue summarization task significantly improves the performance. We obtain 21% overall absolute improvement with the fine-tuning over a baseline Turkish LLM. The performance is improved in all 4 test cases. Additionally, we observe that the length of the summaries plays a crucial role in the performance.

INDEX TERMS Abstractive dialogue summarization, large language models, natural language generation, text summarization.

I. INTRODUCTION

Text summarization is the task of generating a concise summary of a source text [1]. The summary must contain all key information in the source text and must not contain any information which is not mentioned in the text. The text summarization task can be divided into two categories: extractive and abstractive summarization [2]. In the extractive summarization, the most important sentences are picked directly from the source text to form the summary. In the abstractive summarization, the generated summaries may potentially contain new phrases that do not appear in the source text. Dialogue summarization is a special case of text summarization where the source text is a dialogue between two or more people.

In call centers, customers interact with a customer support agent to complete a certain task. These call center conversations might have complex structure with multiple turns between customer and agent depending on the complexity of the customer's issue. They consist of specific greetings/farewell turns and customer identity confirmation questions. They may contain repetitions to confirm the given information and sometimes include misunderstandings. These dialogues are later analyzed by the call center company in order to evaluate the customer's experience and agent's performance. Therefore, it is crucial to make a concise summary of the long dialogue in order to make the analysis more efficient [3], [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna DUlizia.

Recently, large language models (LLMs) showed remarkable performance in natural language generation tasks in zero shot/few shot settings or when fine-tuned to the target domain [5]. Generative Pre-trained Transformer 4 (GPT-4) is a closed-source LLM and was introduced in 2023 [6]. The underlying architecture and weights of the closed-source LLMs are not publicly available [7]. They are developed and backed by large corporations [7]. The data which will be analyzed should be sent to the provider corporation in order to use the closed-source LLMs. Call center conversations contain sensitive personal information and thus they should be kept private in secure locale servers. Therefore, use of the closed source LLMs (such as GPT-4) to analyze call center conversations is largely restricted due to security concerns. On the other hand, thanks to the natural language processing (NLP) community, many open-source commercially available LLMs are trained and distributed publicly [8]. The source code and weights of the open-source LLMS can be freely accessed, used, modified, and distributed [7]. However, the open-source LLMs are mostly trained for English. There are only a few studies for underrepresented languages such as Turkish.

In order to close this research gap, we make a comprehensive evaluation of LLMs for the Turkish dialogue summarization task. In the evaluations, we make use of the best performing LLMs in the Turkish LLM leaderboard in [9]. In an on-premise application, it is not always possible to allocate required computational resources for very large LLMs. Therefore, we restrict our evaluation to LLMs with at most 8B parameters with an exception for the Orbita model [10]. Although Orbita has 14B parameters, we include it in our evaluations since it yields the best performance in the leaderboard at the time when our evaluations are performed.

Turkish lacks datasets for the dialogue summarization task, thus we translated two datasets in English, namely SamSum [11] and DialogSum [12], to Turkish. We also translated the intent induction from conversations for task-oriented dialogue track at the Eleventh Dialog System Technology Challenge (DSTC11) dataset [13] to Turkish. We chose this dataset since it contains dialogues in the finance and banking domains. The DSTC11 dataset does not contain reference summaries. We automatically generated the references using the GPT model. Lastly, we used 21 real call center conversations which were collected by Sestek Inc. The conversations were originally collected in Turkish and include several different domains. We carefully anonymized any personal information in the conversations before using them in the experiments. The reference summaries for the dialogues are created by the GPT model.

The contributions of our paper can be summarized as follows; i) We publicly distribute 4 datasets used in this study in https://github.com/obu80/TurkishDialogueSummarization in order to accelerate the research on the Turkish dialogue summarization task. ii) We compare several different Turkish LLMs for the task and present the baseline results for future studies in the field. iii) We fine-tune the baseline LLMs to the dialogue summarization task in order to investigate the effects of the fine-tuning. We show that the fine-tuning significantly improves the dialogue summarization performance.

The remainder of the paper is organized as follows. In the next section, we give a literature review of Turkish LLM and text summarization studies. In Section III, we describe our dialogue summarization datasets. We present the methodology in Section IV. Section V is devoted to experimental results. We conclude our paper with a summary of our key findings.

II. LITERATURE REVIEW FOR TURKISH LLM AND TEXT SUMMARIZATION

Only very recently, a few LLMs are made publicly available for the Turkish language. Trendyol-LLM-7b-chat [14] is one of the most successful Turkish models. It is trained by the Trendyol group and has 7B parameters. The model is based on the Mistral-7B model [15]. The Trendyol-LLM-7b-chat is fine-tuned on 180K chat instructions with the low-rank adaptation (LoRA) technique [16]. Different from the Trendyol-LLM-7b-chat, the direct preference optimization (DPO) [17] was applied in the Trendyol-LLM-7b-chat-dpo model [18]. The Trendyol models might be well suited to the dialogue summarization task since they are fine-tuned with chat instructions. Turkcell group also trained a LLM for the Turkish language [19]. Turkcell-LLM-7b has 7B parameters and was trained on a cleaned Turkish raw dataset containing 5B tokens. The training process initially used the weight-decomposed low-rank adaptation (DORA) technique [20]. Then, several different Turkish instruction sets are used to fine-tune the model with LoRA. Orbita is another LLM for the Turkish language with 14B parameters [10]. The Orbita was trained/fine-tuned on a cleaned/annotated Turkish datasets to perform multiple tasks such as coding, math problem solving, and others.

The number of studies for text/dialogue summarization task is limited for the Turkish language. In [21], a news summarization (TR-News) dataset is collected for Turkish and monolingual/multilingual Bidirectional Encoder Representations from Transformers (BERT) based models are compared for the task. In [22], in addition to the TR-News, the Turkish subset of the multilingual news summarization (MLSum) dataset [23] is used for evaluations. In the study, pre-trained sequence-to-sequence language models, Bert2Bert [24], Multilingual Bidirectional and Auto-Regressive Transformers (mBart) [25] and Multilingual Text-to-Text Transfer Transformer (mT5) [26], are evaluated. In [27], a new evaluation measure based on semantic similarity between the input and corresponding summary is proposed for abstractive text summarization. The proposed method is tested on the Turkish MLSum subset. In the study, the authors fine-tuned the multilingual mT5 model to the summarization task. Turkish news summarization is also studied in [28] and [29].

There are only a few studies focusing on the use of LLMs for Turkish natural language understanding (NLU) and

generation (NLG) tasks. In [30], an encoder-decoder LLM, named as TURNA, with 1.1B parameters is developed for Turkish and is compared to several other multilingual models. We did not include TURNA in our evaluations since it did not yield a comparable performance to the other larger Turkish LLMs in our preliminary tests. In [31], an in-depth analysis is conducted to evaluate the impact of training strategies, model choices, and data availability on the performance of Turkish LLMs.

To the best of our knowledge, our paper is the first study that investigates the use of LLMs for the Turkish dialogue summarization. To significantly contribute to the field, we publicly distribute 4 Turkish dataset for the task and provide the baseline and fine-tuning results for several different Turkish LLMs in different test cases and fine-tuning scenarios.

III. DATASETS

A. SAMSUM-TR AND DIALOGSUM-TR DATASETS

The SamSum [11] and DialogSum [12] datasets are commonly used for abstractive dialogue summarization in English. The SamSum contains natural messenger-like conversations created and written down by linguists fluent in English. The DialogSum contains face-to-face spoken dialogues that cover a wide range of daily-life topics, including schooling, work, medication, shopping, leisure and travel. The conversations mostly take place between friends and colleagues or service providers and customers.

We translate the datasets to Turkish using Helsinki NLP's open source neural machine translation model [32], [33]. We prefer the Helsinki NLP toolkit since its English to Turkish translation performance is quite good [34]. Number of samples, average number of words in the dialogues and summaries are presented in Table 1 for the datasets. As observed in the table, the DialogSum-TR contains longer dialogues and summaries compared to the SamSum-TR. We use the same train/test splits as in the original English dataset for the evaluations in Turkish. We performed the evaluations with the first 200 samples from each test set in order to keep the duration of the evaluations reasonable.

In order to obtain reference summaries with a LLM, we asked GPT-3.5-turbo to summarize the dialogues in the Turkish datasets. We preferred the GPT-3.5 since it is cheaper than the GPT-4. We used the Turkish prompt in the Appendix A for summarization. Average number of words in the GPT-summaries is provided in Table 2. When we compare Table 1 and Table 2, we observe that the summaries of the GPT-3.5 are approximately 2.5 times longer than the human annotations.

Throughout this study, we use GPT-4 to assess the accuracy of the summaries. We ask GPT-4 to assess whether a summary is accurate or inaccurate according to the following four criteria; i) The summary must contain the information which is key to the conversation. ii) The summary must be concise iii) The summary must not contain unnecessary information. iv) The summary must not include information which is not contained in the text. Complete prompt for the summary assessment is provided in the Appendix A.

We utilize GPT-4 in order to assess the quality of translated dialogue/summary pairs. For this purpose, we use 42 test sentences from the SamSum-TR dataset. The GPT-4 evaluates 33 of the human annotation summaries as accurate in the Turkish dataset. The number of accurate evaluations are 36 for the corresponding 42 sentences in the original English dataset. The relatively small degradation might be attributed to the translation inaccuracies. The GPT-4 evaluates 33 summaries generated by GPT-3.5 as accurate in the SamSum-TR. As a result, although GPT-3.5 generates much longer summaries compared to the human annotators, GPT-3.5 summaries are found to be as accurate as the human annotations by GPT-4.

B. DSTC11-TR DATASET

The DSTC11 dataset [13] is originally collected for intent induction from conversations for task-oriented dialogue systems. The dataset is designed to emulate natural call center conversations between customers and agents. Each conversation emulates a two-party spoken-form customer support scenario and complex conversational phenomena to encourage diversity and naturalness [13]. Three domains are provided in the dataset; insurance (used as development data), banking and finance. We use banking and finance as training data. 200 sentences from the development set (insurance domain) are used for testing.

The DSTC11 dataset is translated to Turkish using Helsinki NLP's open source neural machine translation model [32], [33]. Originally, the dataset does not contain reference summaries for the dialogues. We created the reference summaries using GPT-3.5 with the summarization prompt in the Appendix A. We ask GPT-4 to assess the accuracy of the summaries. GPT-4 finds 40 out of 42 summaries as accurate.

Statistics of the DSTC11-TR dataset are presented in Table 3. As observed in the table, the DSTC11-TR contains much longer dialogues compared to the SamSum-TR and DialogSum-TR. The length of the dialogues in the DSTC11-TR is closer to the length of the dialogues in the RealCall-TR dataset which will be discussed in the next subsection.

C. REALCALL-TR DATASET

We use 21 real-life call center conversations collected by Sestek for the evaluations. The conversations are originally in Turkish. The domain of the conversations are quite diverse including banking, insurance, automotive, white goods and online shopping. However, approximately half of the conversations are about personal banking issues. All the personal information in the dialogues are carefully anonymized before they are used in the evaluations. Since the number of dialogues are limited in the dataset we use GPT-4 to generate the reference summaries. All the reference summaries are found to be accurate by the GPT-4 evaluation.

TABLE 1. Statistics of the SamSumTR and DialogSumTR datasets.

Dataset name	# of samples (train / test)	Average # of words in dialogues (train / test)	Average # of words in summaries (train / test)
SamSum-TR	14731 / 200	72.79 / 73.44	15.52 / 14.95
DialogSum-TR	12460 / 200	96.23 / 99.12	19.09 / 16.67

TABLE 2. Average number of words in the GPT3.5-summaries.

Dataset name	Average # of words in summaries (train / test)
SamSum-TR (GPT3.5 Summaries)	37.00 / 36.84
DialogSum-TR (GPT3.5 Summaries)	45.96 / 44.22

TABLE 3. Statistics of DSTC11-TR and RealCall-TR datasets.

Dataset name	# of samples (train / test)	Average # of words in dialogues (train / test)	Average # of words in summaries (train / test)
DSTC11-TR	3000 / 200	637.89 / 553.03	58.96 / 61.16
RealCall-TR	0 / 21	- / 369.28	- / 41.80

Statistics of the RealCall-TR dataset are presented in Table 3. As shown in the table, the dialogues in the dataset contain approximately 370 words on average which is approximately 4-5 times longer than the dialogues in the SamSum-TR and DialogSum-TR datasets. The longest and shortest dialogues in the RealCall-TR dataset contain 1004 words and 115 words, respectively. The average dialogue length in the RealCall-TR dataset is closer to the DSTC11-TR. We can also say that the domains of the RealCall-TR and DSTC11-TR are better matched.

IV. METHODOLOGY

Block diagram for our methodology is shown in Figure 1. As shown in the figure, the baseline LLMs are trained using a huge unsupervised dataset with the next word prediction task [6]. Then, the baseline model is fine-tuned to the specific task with a relatively small domain-specific supervised dataset. The supervised dataset contains input-output pairs such as instruction-response or dialogue-summary pairs. As shown at the bottom of Figure 1, the user sends the dialogue to the LLM to get its concise summary. Our methodology is described in more detail in the following subsections.

A. PRE-TRAINED LARGE LANGUAGE MODELS

There are various commercially available LLMs such as Gemma [35], Llama3 [36], Mistral [15], Qwen [37], Bloom [38], Falcon [39], MPT-30B [40] which are mostly trained for English. These models can follow instructions in Turkish since their training data might be including a small percentage of Turkish text. However, their performances in Turkish are expected to be much lower compared to English. Very recently, a few Turkish LLMs were trained and made publicly available. Trendyol group recently released two LLMs with 7B parameters, namely Trendyol-LLM-7b-chat (Trendyol-7B-chat) [14] and Trendyol-LLM-7b-chat-dpo (Trendyol-7B-dpo) [18]. Both models are trained based on the Mistral-7B model [15]. They are fine-tuned using 180K chat instructions with the LoRA technique. Additionally, DPO is applied in Trendyol-7B-dpo using a 11K prompt-chosen-reject set [18]. Turkcell, the biggest telecommunication company in Turkey, also released a Turkish LLM with 7B parameters [19]. Turkcell-7B is also based on the Mistral-7B. It was initially trained on a cleaned Turkish dataset containing 5B tokens using the DORA technique [20]. Then, the base model is fine-tuned using several different open-source and closed-source Turkish instruction sets with the LoRa [16]. Orbita is another Turkish LLM with 14B parameters and is based on Qwen-14B model [37]. It is trained on a Turkish dataset annotated to carry out Turkish instructions in an accurate and organized manner [10].

In an on-premise application, it is not always possible to allocate required graphics processing unit (GPU) resources for very large scale LLMs. Therefore we restrict our evaluation to LLMs with at most 8B parameters with an exception for Orbita-14B. In addition to the Orbita-14B, the performances of three other Turkish LLMs (Trendyol-7B-chat, Trendyol-7B-dpo, Turkcell-7B) are evaluated for the dialogue summarization task. We also include the very recently released Llama3 in the evaluation to compare the performances of the Turkish models to one of the best commercially available English models. We use the instruction fine-tuned version of Llama3 with 8B parameters [36] in the experiments.

B. FINE-TUNING LARGE LANGUAGE MODELS

LLMs show remarkable text generation capabilities even in zero-shot setting scenarios. On the other hand, their performances are improved when fine-tuned to the target domain especially when the domain requires a particular knowledge base. In this paper, we fine-tune the baseline LLMs using the dialogue-summary pairs in the

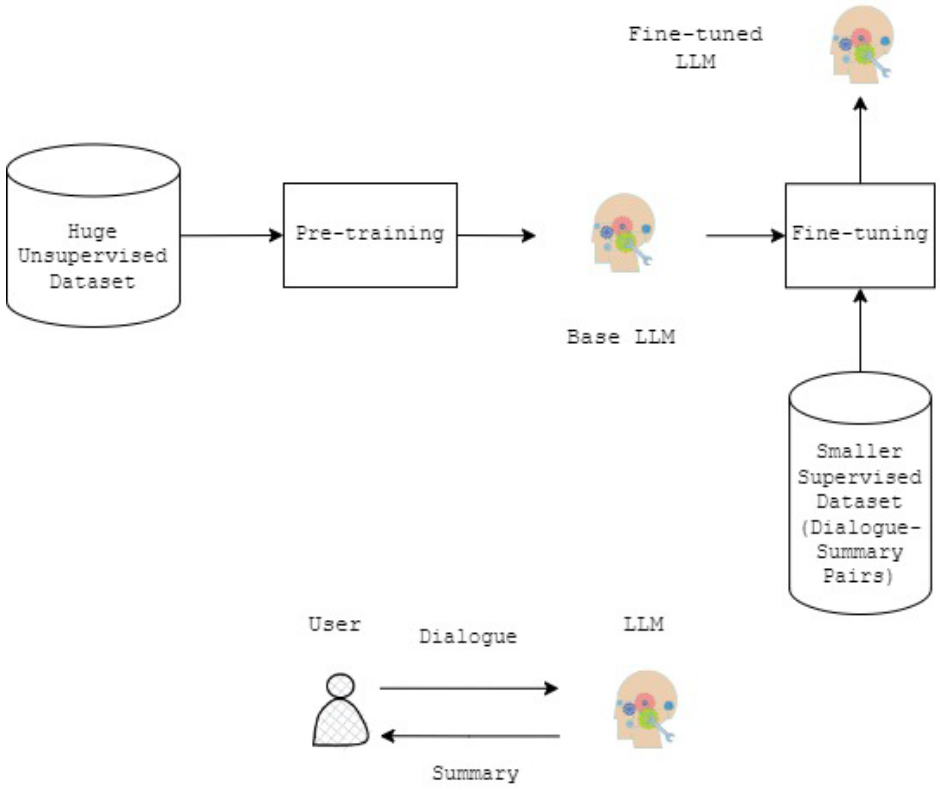


FIGURE 1. Block diagram for LLM pre-training, fine-tuning and inference.

training split of the dialogue summarization datasets in Section III.

We use the LoRA technique for fine-tuning. Our implementation is based on the code in [41]. The LoRA target modules are set to 'q_proj', 'v_proj', 'k_proj', 'o_proj', 'gate_proj', 'down_proj' and 'up_proj' in the experiments. We use the LoRa alpha, R and dropout parameters as 16, 8 and 0.05, respectively. The number of epochs is set to 2 with a batch size of 128. The initial learning rate is 3e-4 with a linear learning rate schedule. The maximum sequence length is set to 2400. Our adaptation template is provided in the Appendix A.

We use the same set of summary generation parameters for all models. The temperature and the repetition penalty are set to 0.1 and 1.1, respectively. We use the top_p, top_k and num_beams as 0.75, 40 and 4, respectively. Maximum number of new tokens is 512 in the experiments. The fine-tuning and inference parameters used in the experiments are presented in Table 4 for clarity.

C. POST-PROCESSING

It is known that the LLMs sometimes repeat the same pattern during the generation. We used a repetition penalty to overcome this problem but the repetitions still persisted for some cases. In order to obtain a more reliable comparison of the models, we post-process the generated summaries to remove the repetitions. When an LLM starts repeating a pattern, the generated summary becomes too long. If the length of the generated summary is longer than a certain threshold (1800 characters in our experiments), then we split the text into its sentences using '.' as the split character. If a sentence occurs more than once in the text, then it is removed from the summary.

In the post-process, we also removed any keywords which indicate the begin and end of the generation such as '</s>', '<s>', '##<|im_end|>'.<|end_of_text|>' and others. We also applied some additional post-processing procedures for each model when needed. For example, Llama3 sometimes generates an evaluation of its own summary in addition to the summary of the dialogue. The evaluation text usually begins with a particular phrase such as ''Özet, k1sa ve öz (summary, short and concise)'' or ''Özet, görüşmedeki önemli (summary, dialogue important)''. We removed the redundant parts in the summaries during the post-process. These model-specific post-processing procedures are described in more detail in Appendix B.

V. EXPERIMENTAL RESULTS

We evaluate the performance of the LLMs using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric. ROUGE-N measures the overlap of N-grams between the reference and generated texts. We provide ROUGE-1 (denoted by R-1) and ROUGE-2 (R-2) scores for each model.

TABLE 4. Fine-tuning and inference parameters used in the experiments.

Fine-tuning parameters	Value	Inference paramaters	Value
Number of epochs	2	Temperature	0.1
Batch size	128	Repetition penalty	1.1
Initial learning rate	3e-4	top_p	0.75
Alpha	16	top_k	40
LoRA R	8	num_beam	4
Dropout	0.05	Max. number of new tokens	512
Max. sequence length	2400		

Additionally, we provide the ROUGE-L (R-L) score which is based on the longest common subsequence (LCS) between the reference and generation. We also present the length ratio metric which shows the ratio of the length of the generated text to the length of the reference text. In addition to these statistical metrics, GPT-4 is used to evaluate the quality of the generated summaries. We ask GPT-4 to evaluate the summaries as accurate or inaccurate according to the evaluation prompt in the Appendix A. The experiments in this paper are run on a NVIDIA A100 40GB GPU server with the exception for the Orbita-14B model. The Orbita-14B model experiments are run on a NVIDIA A100 80GB GPU.

A. BASELINE EXPERIMENTS

1) THE SAMSUM-TR RESULTS

In Table 5, baseline LLM results for SamSum-TR dataset are presented. In the first four columns, the original human annotations are used as the reference summaries. In the following four columns, we use GPT-3.5 summaries as references. In the table, the best performing model is indicated as bold.

We can make several observations from Table 5. First of all, all LLMs tend to generate much longer summaries compared to the human annotations as observed in the 'Length Ratio' of the 'Human References' column. The best performing Turkcell-7B model generates 2.5 times longer summaries than the human annotations. On the other hand, the length of the generated references are better matched with the GPT references. For example, Turkcell-7B summaries have almost the same length (length ratio of 1.04) with the GPT references. The mismatch between the length of the generated and reference summaries significantly affects the performance. All the ROUGE scores with human references are much lower than the ROUGE scores with the GPT references. Overall, in the experiments, Turkcell-7B yielded the best performance with human annotations. The Trendyol-7B-chat and Trendyol-7B-dpo performed the best with the GPT references.

2) THE DIALOGSUM-TR RESULTS

In Table 6, baseline LLM results for the DialogSum-TR dataset are presented. Here, we make similar observations as in the SamSum-TR case. All LLMs generate much longer summaries when compared to the human annotations with a best length ratio of 2.4. The performance is significantly degraded due to the length mismatch between the reference and generated summaries. Overall, in the experiments, Trendyol-7B-chat and Turkcell-7B yielded the best performance with the human references. The Orbita-14B achieves the best performance with the GPT references.

3) THE DSTC11-TR AND REALCALL-TR RESULTS

In Table 7, we provide results for the DSTC11-TR and RealCall-TR datasets. In the first four columns, DSTC11-TR dataset results are presented. In the following four columns below, RealCall-TR results are provided. As observed in the table, all the length ratios are below 2 for all LLMs since reference summaries are generated by GPT for both datasets. Overall, Orbita-14B outperforms other models in the DSTC11-TR test case. Trendyol-7B-dpo slightly performs better than the other models for the RealCall-TR set.

B. FINE-TUNING EXPERIMENTS

1) COMPARISON OF THE MODELS

We combine the training splits of the SamSum-TR (14731 samples), DialogSum-TR (12460 samples) and DSTC11-TR (3000 samples) datasets for fine-tuning. As a result, we obtain 30191 dialogues and their summaries. We chose to fine-tune the Trendyol-7B-dpo model since it yielded slightly better overall performance in the baseline experiments.

The fine-tuning results are presented in Table 8, Table 9 and Table 10 for the SamSum-TR, DialogSum-TR and DSTC11-TR/RealCall-TR datasets, respectively. In the first row of the tables, the baseline Trendyol-7B-dpo model results are presented. In the second row with the label Trendyol-7B-dpo(FT-Human), the human annotations in the SamSum-TR and DialogSum-TR are used for fine-tuning. We use GPT references for the DSTC11-TR in this experiment. In the previous results, we realized that the length of the human and LLM-generated references are very different. In order to perform the fine-tuning with a more unified data, we fine-tune the baseline model with the GPT references for all three datasets. This experiment is named Trendyol-7B-dpo(FT-GPT) and its results are provided in the third row of the tables.

We can make several observations from the tables. First of all, when we fine-tune the baseline model with the short human annotations, the fine-tuned model generates much shorter summaries. The length of these shorter summaries match with the human annotations. For example, in the human references test case of Table 8, the length ratio for

TABLE 5. Baseline LLM results for the SamSum-TR dataset. R-1, R-2 and R-L represent ROUGE-1, ROUGE-2 and ROUGE-L scores, respectively.

	Human References			
Model Name	**R-1**	**R-2**	**R-L**	**Length Ratio**
Trendyol-7B-chat	0.263	0.106	0.214	2.99
Trendyol-7B-dpo	0.249	0.101	0.200	3.09
Orbita-14B	0.249	0.102	0.197	3.77
Turkcell-7B	**0.288**	**0.116**	**0.235**	**2.55**
Llama3-8B	0.232	0.094	0.186	3.93
	GPT References			
Model Name	**R-1**	**R-2**	**R-L**	**Length Ratio**
Trendyol-7B-chat	0.437	**0.212**	**0.348**	1.22
Trendyol-7B-dpo	**0.439**	0.209	0.339	1.26
Orbita-14B	0.422	0.203	0.330	1.54
Turkcell-7B	0.426	0.198	0.331	**1.04**
Llama3-8B	0.393	0.184	0.311	1.61

TABLE 6. Baseline LLM results for the DialogSum-TR dataset.R-1, R-2 and R-L represent ROUGE-1, ROUGE-2 and ROUGE-L scores, respectively.

	Human References			
Model Name	**R-1**	**R-2**	**R-L**	**Length Ratio**
Trendyol-7B-chat	**0.250**	0.107	**0.203**	**2.45**
Trendyol-7B-dpo	0.236	0.101	0.184	2.61
Orbita-14B	0.231	0.106	0.187	3.24
Turkcell-7B	0.248	**0.110**	0.200	2.80
Llama3-8B	0.237	0.103	0.186	3.06
	GPT References			
Model Name	**R-1**	**R-2**	**R-L**	**Length Ratio**
Trendyol-7B-chat	0.395	0.188	0.305	**1.15**
Trendyol-7B-dpo	0.416	0.199	0.313	1.22
Orbita-14B	**0.427**	**0.216**	**0.330**	1.52
Turkcell-7B	0.381	0.181	0.290	1.31
Llama3-8B	0.404	0.201	0.315	1.43

TABLE 7. Baseline LLM results for the DSTC11-TR and RealCall-TR datasets. R-1, R-2 and R-L represent ROUGE-1, ROUGE-2 and ROUGE-L scores, respectively.

	DSTC11-TR			
Model Name	**R-1**	**R-2**	**R-L**	**Length Ratio**
Trendyol-7B-chat	0.356	0.154	0.257	0.79
Trendyol-7B-dpo	0.379	0.153	0.258	**1.10**
Orbita-14B	**0.380**	**0.159**	**0.267**	1.34
Turkcell-7B	0.339	0.126	0.226	1.52
Llama3-8B	0.345	0.135	0.238	1.69
	RealCall-TR			
Model Name	**R-1**	**R-2**	**R-L**	**Length Ratio**
Trendyol-7B-chat	0.318	0.126	0.231	**1.20**
Trendyol-7B-dpo	**0.329**	0.134	**0.235**	1.54
Orbita-14B	0.320	0.126	0.234	1.84
Turkcell-7B	0.308	0.124	0.232	1.51
Llama3-8B	0.307	**0.138**	0.234	1.96

the baseline Trendyol-7B-dpo model is 3.09. The length ratio becomes 0.87 with fine-tuning in the Trendyol-7B-dpo(FT-Human) model. The performance is also significantly improved compared to the baseline. Similarly in Table 9, the length ratio is improved from 2.61 to 1.29 in human references test case after fine-tuning with a significant performance gain.

On the other hand, we observe a performance degradation when we fine-tune the model with short human annotations but test the fine-tuned model with longer GPT

TABLE 8. Fine-tuning results for the SamSum-TR dataset. R-1, R-2 and R-L represent ROUGE-1, ROUGE-2 and ROUGE-L scores, respectively.

	Human References			
Model Name	**R-1**	**R-2**	**R-L**	**Length Ratio**
Trendyol-7B-dpo	0.249	0.101	0.200	3.09
Trendyol-7B-dpo(FT-Human)	**0.467**	**0.249**	**0.400**	**0.87**
Trendyol-7B-dpo(FT-GPT)	0.276	0.115	0.222	2.77
	GPT References			
Model Name	**R-1**	**R-2**	**R-L**	**Length Ratio**
Trendyol-7B-dpo	0.439	0.209	0.339	1.26
Trendyol-7B-dpo(FT-Human)	0.293	0.125	0.235	0.36
Trendyol-7B-dpo(FT-GPT)	**0.513**	**0.291**	**0.425**	**1.13**

TABLE 9. Fine-tuning results for the DialogSum-TR dataset. R-1, R-2 and R-L represent ROUGE-1, ROUGE-2 and ROUGE-L scores, respectively.

	Human References			
Model Name	**R-1**	**R-2**	**R-L**	**Length Ratio**
Trendyol-7B-dpo	0.236	0.101	0.184	2.61
Trendyol-7B-dpo(FT-Human)	**0.401**	**0.228**	**0.344**	**1.29**
Trendyol-7B-dpo(FT-GPT)	0.241	0.096	0.194	2.59
	GPT References			
Model Name	**R-1**	**R-2**	**R-L**	**Length Ratio**
Trendyol-7B-dpo	0.416	0.199	0.313	1.22
Trendyol-7B-dpo(FT-Human)	0.310	0.142	0.246	0.60
Trendyol-7B-dpo(FT-GPT)	**0.497**	**0.294**	**0.413**	**1.21**

TABLE 10. Fine-tuning results for the DSTC11-TR and RealCall-TR datasets. R-1, R-2 and R-L represent ROUGE-1, ROUGE-2 and ROUGE-L scores, respectively.

	DSTC11-TR			
Model Name	**R-1**	**R-2**	**R-L**	**Length Ratio**
Trendyol-7B-dpo	0.379	0.153	0.258	**1.10**
Trendyol-7B-dpo(FT-Human)	0.469	0.241	0.353	1.03
Trendyol-7B-dpo(FT-GPT)	**0.479**	**0.255**	**0.366**	**1.00**
	RealCall-TR			
Model Name	**R-1**	**R-2**	**R-L**	**Length Ratio**
Trendyol-7B-dpo	0.329	0.134	0.235	1.54
Trendyol-7B-dpo(FT-Human)	0.378	0.171	0.274	1.23
Trendyol-7B-dpo(FT-GPT)	**0.392**	**0.187**	**0.289**	**1.23**

references. This result might be attributed to the fact that the fine-tuned model generates summaries according to the shorter fine-tuning references which do not match with the test case. In the GPT references test case of Table 8, the length ratio becomes 0.36 (1.26 with the baseline model) after fine-tuning which results in a severe performance degradation. Similarly, the length ratio reduces from 1.22 to 0.60 with the Trendyol-7B-dpo(FT-Human) model in Table 9 with a significant degradation in performance.

There are no human annotations for the DSTC11-TR/ RealCall-TR datasets and all the reference summaries are generated by GPT. Both fine-tuned models, namely Trendyol-7B-dpo(FT-Human) and Trendyol-7B-dpo(FT-GPT), improve the performance in these test cases as observed in Table 10. The performance improvement is more significant in the DSTC11-TR test set compared to the RealCall-TR. This performance improvement might be mainly attributed to the DSTC11-TR fine-tuning samples. As mentioned earlier, the DSTC11-TR dataset contains dialogues between an agent and a customer in banking and finance domains similar to the RealCall-TR test set.

Another observation from Table 8 and Table 9 is that when the model is fine-tuned with the GPT references in the Trendyol-7B-dpo(FT-GPT) model, the performance is still slightly improved even if it is tested with the human annotations. The ROUGE-L improves from 0.200 to 0.222 in the SamSum-TR human reference test in Table 8 and from 0.184 to 0.194 in the DialogSum-TR human reference test in Table 9 with the Trendyol-7B-dpo(FT-GPT) model. This result implies that generating supervised fine-tuning datasets automatically with reliable LLMs might be a viable solution for the dialogue summarization task.

TABLE 11. GPT-4 evaluation of the summaries generated by the Trendyol-7B-dpo and Trendyol-7B-dpo(FT-GPT) models. We ask GPT-4 to evaluate the summaries as accurate or inaccurate. In the table, the number of accurate evaluations is shown.

Dataset name	SamSum-TR	DialogSum-TR	DSTC11-TR	RealCall-TR	Total Acc. (%)
Trendyol-7B-dpo	20/42	22/42	34/42	15/21	61.9%
Trendyol-7B-dpo(FT-GPT)	32/42	34/42	37/42	19/21	82.9%

We obtain significant performance improvements with the Trendyol-7B-dpo(FT-GPT) model in all four GPT-references test cases. ROUGE-L improves from 0.339 to 0.425 in the SamSum-TR, from 0.313 to 0.413 in the DialogSum-TR, from 0.258 to 0.366 in the DSTC11-TR, from 0.235 to 0.289 in the RealCall-TR test sets. These results show that fine-tuning a single LLM with a combined dataset yields a performance improvement in all individual test cases. Furthermore, we obtained a performance improvement in the RealCall-TR test set from which no samples are included in the fine-tuning set.

2) GPT-4 EVALUATIONS

We ask GPT-4 to evaluate the summaries of the baseline Trendyol-7B-dpo and the fine-tuned Trendyol-7B-dpo(FT-GPT) models. In the GPT-4 evaluations, we use 42 dialogues from the test sets of the SamSum-TR, DialogSum-TR and DSTC11-TR. All 21 samples of the RealCall-TR dataset are used. In total, we performed the evaluation with 147 test samples. The results are presented in Table 11. In the first four columns of the table, the dialogues which are evaluated as accurate are shown for each dataset. In the last column, overall percent accuracy is provided. As observed in the table, the Trendyol-7B-dpo(FT-GPT) model significantly improves the performance for the SamSum-TR and DialogSum-TR datasets. The performances are also improved with the fine-tuning for the DSTC11-TR and RealCall-TR sets. Overall, we obtain 21% absolute accuracy improvement with the fine-tuning in the GPT-4 evaluations.

VI. DISCUSSION OF RESULTS

As observed in the baseline experiments, different models have shown the best results in different test sets. None of the models significantly outperformed others for all test cases. Overall, the performances of the models are quite comparable. This might be expected since three of the models, namely Trendyol-7B-dpo, Trendyol-7B-chat and Turkcell-7B are derived from the same Mistral-7B model. Only Orbita-14B is derived from another LLM. Although the pre-training datasets have not been officially announced for the models, we might expect that similar open source Turkish datasets were used to pre-train the models. The models are also fine-tuned mainly using closed source internal datasets. These internal datasets might be the main cause of the varying performance of the models in different test cases. As a result, we used Trendyol-7B-dpo in the fine-tuning experiments which was the best overall performant 7B model.

The accuracy of the baseline model is 61.9% as observed in Table 11. The baseline performance is quite acceptable when we consider that the LLM is not specifically trained for the summarization task. In our internal tests, we observed that the performance of the baseline English models are better than the Turkish counterparts. This observation implies the importance of developing more robust foundation models for the Turkish language. Fine-tuning LLMs to the dialogue summarization task significantly improved the performance. The fine-tuned model yielded 82.9% overall summarization accuracy. It achieved approximately 90% accuracy in the real call center dialogues of the RealCall-TR test set. These results show that the performance of the Turkish LLMs can be suitable for a commercial dialogue summarization application when the models are fine-tuned to the task with appropriate supervised datasets.

Summarization of the call center dialogues is a complex task. Larger LLMs might be needed to accurately summarize these complex dialogues. In this paper, we limited our evaluation to 8B LLMs since, i) we want to focus on LLMs which can be deployed in a commercial application with minimum hardware requirements, ii) current mono-lingual Turkish LLMs usually have less than 8B parameters. In the future, larger multi-language models might be investigated for the task to improve the performance.

It is known that the response of the LLMs are highly dependent on the choice of the prompts. In this paper, we chose the prompts after a few preliminary trials and did not perform extensive evaluations to optimize them. Moreover, we prefer to use the same prompts for all models for a fairer comparison. It should be noted that the performance can be improved with an optimized selection of prompts for each individual model.

VII. CONCLUSION

In this paper, we evaluate the performances of Turkish LLMs for the dialogue summarization task. We performed the experiment using 4 different datasets, namely SamSum-TR, DialogSum-TR, DSTC11-TR and RealCall-TR. The RealCall-TR is originally collected in Turkish and the others are translated from English. We publicly distribute the datasets for further research. We fine-tuned Turkish LLMs to the dialogue summarization task using the dialogue-summary pairs in the datasets. We obtain significant improvement with the fine-tuning for all test sets.

LLMs is one of the most impactful research areas in machine learning recently and new open source models are frequently introduced to the community which are

trained using larger/cleaner datasets and improved training techniques. In this paper, we provide a comprehensive evaluation of the Turkish dialogue summarization task using different test cases and fine-tuning scenarios. However, the performance can be improved by using the new foundation models in the future. In the future, we also plan to synthetically generate real-like call center conversations using LLMs and fine-tune the baseline models with the synthetic dataset to further improve the performance.

APPENDIX A
LARGE LANGUAGE MODEL PROMPTS

A. DIALOGUE SUMMARIZATION PROMPT

We use the following Turkish prompt for the dialogue summarization:

''Sen bir diyalog özetleme asistan1n ve sana verilen diyaloğun özetini oluşturacaks1n. Özet, görüşmedeki tüm önemli bilgileri içermeli, k1sa ve öz olmal1. Gereksiz ve metinde yer almayan bilgiler içermemeli.''

B. SUMMARY EVALUATION PROMPT

We prefer the following English prompt for summary evaluation since it yields a slightly improved performance:

''You will be provided with a context text and its summary. Act as an evaluator and make an assessment as regards the correctness of the given summary.

Correctness criteria: 1. The summary must contain the information which is key to the conversation. 2. The summary must be concise 3. The summary must not contain unnecessary information. 4 The summary must not include information which is not contained in the text.

Generate a json response with 3 fields: Judgment: accurate or inaccurate Violation_criteria: a list of violation criteria ids from the list above Explanation: a concise reasoning for the judgment:

C. FINE-TUNING TEMPLATE FOR LoRA ADAPTATION

We use the following Turkish prompt for the LoRA adaptation:

''Sen bir diyalog özetleme asistan1s1n ve sana verilen diyaloğun özetini oluşturacaks1n. Özet, görüşmedeki tüm önemli bilgileri içermeli, k1a ve öz olmal1. Gereksiz ve metinde yer almayan bilgiler içermemeli.

Diyalog: {instruction}

Özet:''

APPENDIX B
MODEL-SPECIFIC POST-PROCESSING PROCEDURES

A. Llama3-8B POST-PROCESSING

Llama3-8B sometimes generates an evaluation of its own summary in addition to the summary of the dialogue. The evaluation text begins with a particular phrase such as ''Özet, görümedeki tüm ''Önemli bilgileri içermeli (The summary should contain all important information in the dialogue)'', ''Özet, k1sa ve öz (The summary short and concise)'' or ''Özet, görümedeki önemli (The summary dialogue important)''. In order to remove the redundant evaluation text, all the text right after the evaluation phrase is not included in the summary.

B. TURKCELL-7B POST-PROCESSING

Turkcell-7B sometimes asks a question and answers it in addition to the dialogue summary. The question/answer text begins with a particular phrase such as ''Soru: (Question:)'' or ''Aşağ1daki soruyu cevaplay1n: (Answer the question below:''. In order to remove the redundant question/answer text, all the text right after the question/answer phrase is not included in the summary.

C. ORBITA-14B POST-PROCESSING

Orbita-14B sometimes generates multiple lines in the summary. We observed that the lines after the first line are redundant and usually contains html codes. We don't include them in the final summary.

REFERENCES

[1] L. Liu, Y. Lu, M. Yang, Q. Qu, J. Zhu, and H. Li, ''Generative adversarial network for abstractive text summarization,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–24.

[2] Y. Huang, X. Feng, X. Feng, and B. Qin, ''The factual inconsistency problem in abstractive text summarization: A survey,'' 2021, *arXiv:2104.14839*.

[3] Y. Zhang, A. Ni, T. Yu, R. Zhang, C. Zhu, B. Deb, A. Celikyilmaz, A. H. Awadallah, and D. Radev, ''An exploratory study on long dialogue summarization: What works and What's next,'' 2021, *arXiv:2109.04609*.

[4] X. Feng, X. Feng, and B. Qin, ''A survey on dialogue summarization: Recent advances and new frontiers,'' 2021, *arXiv:2107.03175*.

[5] T. Brown, ''Language models are few-shot learners,'' 2020, *arXiv:2005.14165*.

[6] J. Achiam, ''Gpt-4 technical report,'' 2023, *arXiv:2303.08774*.

[7] *ChatGPT on Open Vs. Closed Source Large Language Models (LLMs) for Internal AI Projects*. Accessed: Aug. 16, 2024. [Online]. Available: https://www.forcepoint.com/blog/insights/chatgpt-open-vs-closed-source-large-language-models

[8] H. Touvron et al., ''Llama 2: Open foundation and fine-tuned chat models,'' 2023, *arXiv:2307.09288*.

[9] *Open LLM Turkish Leaderboard*. Accessed: May 25, 2024. [Online]. Available: https://huggingface.co/spaces/malhajar/OpenLLMTurkishLeaderboard

[10] *Orbina LLM-14B Model*. Accessed: May 25, 2024. [Online]. Available: https://huggingface.co/Orbina/Orbita-v0.1

[11] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, ''SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization,'' 2019, *arXiv:1911.12237*.

[12] Y. Chen, Y. Liu, L. Chen, and Y. Zhang, ''DialogSum: A real-life scenario dialogue summarization dataset,'' 2021, *arXiv:2105.06762*.

[13] J. Gung, R. Shu, E. Moeng, W. Rose, S. Romeo, Y. Benajiba, A. Gupta, S. Mansour, and Y. Zhang, ''Intent induction from conversations for task-oriented dialogue track at DSTC 11,'' 2023, *arXiv:2304.12982*.

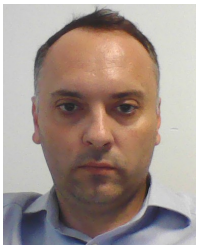
[14] *Trendyol LLM-7B Chat Model*. Accessed: May 25, 2024. [Online]. Available: https://huggingface.co/Trendyol/Trendyol-LLM-7b-chat-v1.0

[15] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, ''Mistral 7B,'' 2023, *arXiv:2310.06825*.

[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, ''LoRA: Low-rank adaptation of large language models,'' 2021, *arXiv:2106.09685*.

[17] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, ''Direct preference optimization: Your language model is secretly a reward model,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–24.

[18] *Trendyol LLM-7B Chat DPO Model*. Accessed: May 25, 2024. [Online]. Available: https://huggingface.co/Trendyol/Trendyol-LLM-7b-chat-dpo-v1.0
[19] *Turkcell LLM-7B Model*. Accessed: May 25, 2024. [Online]. Available: https://huggingface.co/TURKCELL/Turkcell-LLM-7b-v1
[20] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, "DoRA: Weight-decomposed low-rank adaptation," 2024, *arXiv:2402.09353*.
[21] B. Baykara and T. Güngör, "Abstractive text summarization and new large-scale datasets for agglutinative languages Turkish and Hungarian," *Lang. Resour. Eval.*, vol. 56, no. 3, pp. 973–1007, Sep. 2022.
[22] B. Baykara and T. Güngör, "Turkish abstractive text summarization using pretrained sequence-to-sequence models," *Natural Lang. Eng.*, vol. 29, no. 5, pp. 1275–1304, Sep. 2023.
[23] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano, "MLSUM: The multilingual summarization corpus," 2020, *arXiv:2004.14900*.
[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[25] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 726–742, Dec. 2020.
[26] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "MT5: A massively multilingual pre-trained text-to-text transformer," 2020, *arXiv:2010.11934*.
[27] F. Beken Fikri, K. Oflazer, and B. Yanıkoglu, "Abstractive summarization with deep reinforcement learning using semantic similarity rewards," *Natural Lang. Eng.*, vol. 30, no. 3, pp. 554–576, May 2024.
[28] F. Ertam and G. Aydin, "Abstractive text summarization using deep learning with a new Turkish summarization benchmark dataset," *Concurrency Comput., Pract. Exper.*, vol. 34, no. 9, Apr. 2022, Art. no. e6482.
[29] B. Ay, F. Ertam, G. Fidan, and G. Aydin, "Turkish abstractive text document summarization using text to text transfer transformer," *Alexandria Eng. J.*, vol. 68, pp. 1–13, Apr. 2023.
[30] G. Uludogan, Z. Y. Balal, F. Akkurt, M. Türker, O. Güngör, and S. Üsküdarlı, "TURNA: A Turkish encoder–decoder language model for enhanced understanding and generation," 2024, *arXiv:2401.14373*.
[31] E. C. Acikgoz, M. Erdogan, and D. Yuret, "Bridging the bosphorus: Advancing Turkish large language models through strategies for low-resource language adaptation and benchmarking," 2024, *arXiv:2405.04685*.
[32] J. Tiedemann and S. Thottingal, "OPUS-MT–building open translation services for the world," in *Proc. 22nd Annu. Conf. Eur. Assoc. Mach. Transl.*, 2020, pp. 479–480.
[33] J. Tiedemann, "The tatoeba translation challenge-realistic data sets for low resource and multilingual MT," 2010, *arXiv:2010. 06354*.
[34] *Helsinki-NLP English to Turkish Translation Model*. Accessed: Aug. 18, 2024. [Online]. Available: https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-en-tr
[35] G. Team et al., "Gemma: Open models based on Gemini research and technology," 2024, *arXiv:2403.08295*.
[36] *Llama3-8B Instruction Model*. Accessed: May 25, 2024. [Online]. Available: https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[37] J. Bai et al., "Qwen technical report," 2023, *arXiv:2309.16609*.
[38] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilic, D. Hesslow, and R. Castagne, "Bloom: A 176b-parameter open-access multilingual language model," 2022, *arXiv:2211.05100*.
[39] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay, "The Refined web dataset for falcon LLM: Outperforming curated corpora with web data, and web data only," 2023, *arXiv:2306.01116*.
[40] *MPT-30B Model*. Accessed: May 25, 2024. [Online]. Available: https://www.databricks.com/blog/mpt-30b
[41] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, and C. Guestrin, "Alpaca: A strong, replicable instruction-following model," *Stanford, CA, USA, Center Res. Found. Models*, vol. 3, no. 6, p. 7, 2023.



OSMAN BÜYÜK received the B.Sc. degree in electrical and electronics engineering from Bilkent University, in 2003, the M.Sc. degree in electrical and electronics engineering from Sabanc University, in 2005, and the Ph.D. degree in electrical and electronics engineering from Bogazici University, in 2011. He worked at Sestek Conversational Solution Technology Company, between 2005 and 2012. From 2013 to 2020, he worked as an Assistant Professor at the Department of Electronics and Communications Engineering, Kocaeli University. He is currently working as an Associate Professor with the Department of Electrical and Electronics Engineering, Izmir Demokrasi University. His research interests include speech/natural language processing, machine learning, and deep learning.