

# Fine-Tuning Pretrained Language Models to Enhance Dialogue Summarization in Customer Service Centers

Jiseon Yun\*  
KakaoBank Corp.  
Kyeonggi-do, Republic of Korea  
sunny.yun@kakaobank.com

Jae Eui Sohn\*  
KakaoBank Corp.  
Kyeonggi-do, Republic of Korea  
mark.sohn@kakaobank.com

Sunghyon Kyeong†  
KakaoBank Corp.  
Kyeonggi-do, Republic of Korea  
devyn.k@kakaobank.com

## Abstract

The application of pretrained language models in real-world business domains has gained significant attention. However, research on the practical use of generative artificial intelligence (AI) to address real-world downstream tasks is limited. This study aims to enhance the routine tasks of customer service (CS) representatives, particularly in the finance domain, by applying a fine-tuning method to dialogue summarization in CS centers. KakaoBank handles an average of 15,000 CS calls daily. By employing a fine-tuning method using real-world CS dialogue data, we can reduce the time required to summarize CS dialogues and standardize summarization skills. To ensure effective dialogue summarization in the finance domain, pretrained language models should acquire additional knowledge and skills, such as specific knowledge of financial products, problem-solving abilities, and the capacity to handle emotionally charged customers. In this study, we developed a reference fine-tuned model using Polyglot-Ko (5.8B) as the baseline PLM and a dataset containing a wide range of zero-shot instructions and partially containing summarization instructions. We compared this reference model with another model fine-tuned using KakaoBank's CS dialogues and summarization data as the instruct dataset. The results demonstrated that the fine-tuned model based on KakaoBank's internal datasets outperformed the reference model, showing a 199% and 12% improvement in ROUGE-L and RDASS, respectively. This study emphasizes the significance of task-specific fine-tuning using appropriate instruct datasets for effective performance in specific downstream tasks. Considering its practical use, we suggest that fine-tuning using real-world instruct datasets is a powerful and cost-effective technique for developing generative AI in the business domain.

## CCS Concepts

• **Computing methodologies** → **Machine learning**; **Natural language processing**; • **Information systems** → **Enterprise information systems**.

\*These authors contributed equally to this research

†Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

ICAIF '23, Brooklyn, NY, USA

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0240-2/23/11

<https://doi.org/10.1145/3604237.3626838>

## Keywords

fine-tuning, instruct tuning, dialogue summarization, Korean language model

## ACM Reference Format:

Jiseon Yun, Jae Eui Sohn, and Sunghyon Kyeong. 2023. Fine-Tuning Pretrained Language Models to Enhance Dialogue Summarization in Customer Service Centers. In *4th ACM International Conference on AI in Finance (ICAIF '23)*, November 27–29, 2023, Brooklyn, NY, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3604237.3626838>

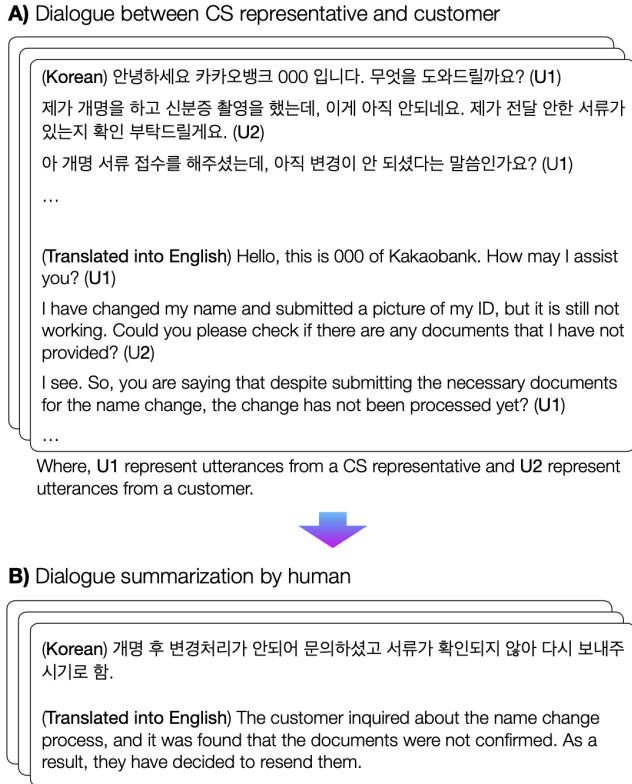
## 1 Introduction

The emergence of the generative pretrained transformer 3 (GPT-3) has gained attention for its applications in generative artificial intelligence (AI) [3, 23]. However, limited research exists on using generative for real-world problems. Recent efforts focus on applying prompt-tuning techniques to pretrained language models (PLMs) [34]. Nevertheless, the effectiveness of prompt-tuning approaches in specific domains and business processes may be limited. Hallucination can be problematic in domains like customer service (CS) in finance or healthcare [8, 25]. Overcoming these limitations can enhance the operational efficiency of a CS center (CSC) by training the model in product knowledge, problem-solving, and handling emotional customers [18]. Additionally, selecting the right model scale is crucial for cost-effective implementation in large-scale CS operations.

The growth in generative AI model size has surpassed the capabilities of graphic process units. Fine-tuning the latest PLMs and storing all parameters is expensive and impractical [19]. The financial industry must consider privacy and protect customer information when using AI [2]. It is practical to develop AI systems internally, leveraging publicly available PLMs as baseline models and fine-tuning them to business objectives. Parameter-efficient fine-tuning is the norm in language model utilization [7, 21], and we aim to advance AI-driven CSC by fine-tuning PLMs with billions of parameters.

Our goal is to improve CS representatives' performance in the finance industry for abstractive summarization tasks based on client conversations as illustrated in Fig. 1. We fine-tuned a Korean-based PLM using a real dataset from KakaoBank's CSC. Summarizing dialogues poses unique challenges compared to summarizing general knowledge or procedural documents. It requires understanding customer issues in finance and incorporating information about issue resolution during the conversation. AI models trained on non-dialogue documents struggle with these aspects, leading to limited research on dialogue summarization.

As of the first quarter of 2023, KakaoBank serves around 21.2 million customers, accounting for 73% of the working population



**Figure 1: Example of KakaoBank’s CS datasets: A) dialogue between CS representative (U1) and customer (U2), B) dialogue summarization by a human.**

in South Korea (<https://eng.kakaobank.com>). KakaoBank’s services include fee-based banking, such as deposits, loans, and debit cards, as well as platform-related businesses like cobranded credit cards, loan referrals, securities brokerage, and advertising. KakaoBank handles a significant volume of transactions, reaching 4.8 billion in 2022. Operating as an internet-only bank, customer inquiries are directed to a dedicated CSC, which handles an average of 15,000 CS calls daily. After the completion of the CS calls, the CS representatives are responsible for organizing the key customer complaints from the dialogues with clients and summarizing how they were resolved. Additionally, they must identify the specific area of the complaint or inquiry, such as deposits, loans, remittances, or debit cards, and select the appropriate category for logging into the system to finalize the CSC process.

The objective of this study is to improve the efficiency of various tasks within CSCs by fine-tuning a PLM using a dataset acquired from KakaoBank’s CSC. To the best of our knowledge, this is the first research endeavor in the financial sector’s CSC to apply generative AI to enhance operational efficiency. In summary, the following contributions were made. First, we developed a language model that can effectively summarize CS dialogues. This was achieved by fine-tuning a Korean PLM model with a few billion parameters using KakaoBank’s CSC data. Second, considering the potential impact of the contents in the instruct dataset on the model’s performance, we

**A) KoAlpaca - Korean translation of instruct datasets of Alpaca (D1)**

```
"instruction": "가정에서 전기 사용량을 줄이기 위한 계획을 세우세요.",
"input": "",
"output": "교통수단 대신 도보나 자전거를 이용하고, 전기제품을 꺼놓으 소등하며, 에너지 효율적인 수압조를 사용하는 등의 방법으로 전기 사용량을 줄일 수 있습니다."
```

**B) Utterances only from CS representative plus summarization (D2)**

```
"instruction": "Write a summary of the inquiries and the outcomes achieved during the customer service described below.",
"input": "Hello, this is 000 of Kakaobank. How may I assist you? (U1) I see. So, you are saying that despite submitting the necessary documents for the name change. But, it has not been processed yet? (U1)",
"output": "The customer inquired about the name change process, and it was found that the documents were not confirmed. As a result, they have decided to resend them."
```

**C) Comprehensive dialogue plus summarization (D3)**

```
"instruction": "Write a summary of the inquiries and the outcomes achieved during the customer service described below.",
"input": "Hello, this is 000 of Kakaobank. How may I assist you? (U1) I have changed my name and submitted a picture of my ID, but it is still not working. Could you please check if there are any documents that I have not provided? (U2) I see. So, you are saying that despite submitting the necessary documents for the name change, the change has not been processed yet? (U1)",
"output": "The customer inquired about the name change process, and it was found that the documents were not confirmed. As a result, they have decided to resend them."
```

**Figure 2: Example of instruct datasets: A) KoAlpaca instruct datasets that are Korean translations of instruct datasets of Alpaca (D1), B) Instruct datasets with inputs as utterances only from CS representatives (finds U1 sentences in Fig. 1A) and outputs as dialogue summarization (D2), C) Instruct datasets with inputs as whole dialogue (finds U1 and U2 sentences in Fig. 1A) between CS representatives and customer and outputs as dialogue summarization (D3).**

conducted experiments to evaluate the summarization capabilities of the model. Specifically, we investigated how effectively the fine-tuned model (FTM) could summarize CS dialogues when trained solely based on the CS representatives’s utterances (Fig. 1B) and the entire CS dialogue along with the summaries (Fig. 1C) compared with that of the FTM trained with the KoAlpaca dataset (Fig. 1A).

The remainder of this paper is organized as follows: Section 2 presents a comprehensive review of the relevant literature. In Section 3, the proposed methodology is outlined, including a description of the Korean baseline model and the fine-tuning process using instruct datasets. Section 4 describes the evaluation framework of the fine-tuned Korean PLMs. The experimental setups and results are then presented in Sections 5 and 6, respectively. Finally, Section 7 summarizes the findings.

## 2 Related Works

PLMs represent a crucial technology in the realm of AI, which is essential for solving a wide array of downstream tasks. Notably, PLMs with an extensive number of parameters, such as GPT-3 [3], PaLM [6], and OPT [35], exhibit remarkable performance across

various general-purpose tasks. However, their training has predominantly relies on knowledge-based resources, such as books, web documents, and academic papers, and inadequately incorporates data derived from private conversations between individuals [10, 36].

Furthermore, no instances in which PLMs have been trained on conversational data involving interactions between CS representatives and clients have been identified. Nonetheless, several studies have explored the impact of generative AI on chat-based CS operations, investigated fine-tuning techniques that utilize human feedback, and examined prompt-tuning approaches aimed at summarizing conversational content.

## 2.1 AI-powered CS center

CS operations are vital as a gateway for resolving customer inquiries and concerns, which significantly impacts a company's reputation. Although most CS representatives undergo training to interact with customers based on predetermined guidelines and procedures, substantial variation in productivity exists among them [29]. Efforts have been made to leverage the AI assistants to facilitate more efficient handling of these CS tasks. In one study, a generative AI model was fine-tuned based on actual chat conversations involving approximately 5,000 CS representatives, and its application within a CS system was investigated [4]. The findings revealed that AI considerably aided low-skilled workers in improving their task performance and elevating their ability to engage in chat-based CS tasks. In particular, employees in the low-skill group exhibited significant improvements in their ability to engage in chat conversations, reaching levels comparable to those of high-skill employees when using AI assistants. However, it was observed that skilled workers, who were already proficient in their tasks, experienced positive effects in terms of chat volume per hour but did not demonstrate improvements in average handle time [4].

## 2.2 Fine-tuning

Generative AI models encounter persistent challenges when generating harmful or toxic content. To address this issue, researchers have explored the utilization of instruction-tuning techniques with human feedbacks [23] or machine-generated datasets [26, 32]. Acknowledging the limitations of large-scale language models, such as GPT-3, which has over 100 billion parameters and may struggle with specific tasks such as mathematical problem-solving, a recent study investigated the fine-tuning of an LLaMA baseline model using specialized datasets designed for proficient mathematical problem-solving [21]. Their findings showed an exceptional performance that surpassed existing large language models, particularly in operations involving large numbers, such as addition and subtraction [21].

Furthermore, researchers found that fine-tuning language models using instruction datasets significantly enhances their ability to effectively perform unfamiliar tasks, even without prior exposure to them [33]. To enhance sentiment classification performance in the finance domain, researchers applied the fine-tuning methodology using a domain-specific corpus [1]. In contrast, the effectiveness of fine-tuning using imitation datasets was experimentally examined and found to be ineffective in enhancing model performance. Furthermore, fine-tuning using imitation datasets showed subpar

performance even in specific downstream tasks [11]. These results emphasize the significance of fine-tuning using high-quality real-world datasets.

## 2.3 Conversation summarization

During the process of summarizing conversational content, the generation of prompts is vital in determining result stability, as it can lead to unstable outcomes characterized by factual inconsistencies and hallucinations [12, 15, 22]. To address these challenges, one study introduced few-shot learning into GPT-3-based models [28]. By employing an evaluation system rooted in the dialogue structure, researchers enhanced the performance of abstractive summarization through prompt tuning based on high-scoring conversations. Another study applied meta-transfer learning to perform an abstractive summarization task [5]. The other study employed the transformation of nondialogue (*i.e.*, document) data to bolster the effectiveness of dialogue summarization models [24]. Although studies on abstractive summarization exist, research specifically targeting language models tailored to summarizing CSC within the context of financial domains (Fig. 1) is lacking.

## 3 Methodology

This section describes the characteristics of the baseline PLM and the details of the fine-tuning methodology.

### 3.1 Baseline model

This study aims to develop a fine-tuned language model that can effectively summarize dialogues between clients and CS representatives. For this purpose, we chose the open-source-based Polyglot-Ko as our baseline model owing to its strong performance in Korean. Polyglot-Ko models were specifically designed for the Korean language. They were developed and trained by the EleutherAI polyglot team [13, 14] and trained by using a vast dataset of 863 GB of Korean text (1.2 TB before preprocessing) obtained from TUNiB (<https://tunib.ai>). The datasets consisted of Korean text data from various sources, including blog posts, news articles, fictional texts, patents, questions and answers, and Wikipedia. The Polyglot-Ko models were prepared in four different sizes: 1.3B, 3.8B, 5.8B, and 12.8B. In this study, we selected the 5.8B model as the baseline and utilized the 12.8B model to assess whether there was any performance enhancement. The Polyglot-Ko (5.8B) model consists of 28 transformer layers with model and feedforward dimensions of 4,096 and 16,384, respectively.

### 3.2 Fine-tuning using instruct datasets

Fine-tuning with instruct datasets is a commonly used approach in recent language research that leverages PLMs to enhance performance on specific downstream tasks. This technique entails the initialization of a pretrained model, such as LLaMA, with extensive linguistic knowledge obtained from a substantial corpus of text data. Subsequently, the model is subjected to further training on instruct datasets tailored to the task or domain at hand. These instruct datasets encompass labeled examples, furnishing explicit guidance to the model for executing the desired task. During the fine-tuning phase, the parameters of the model optimized using

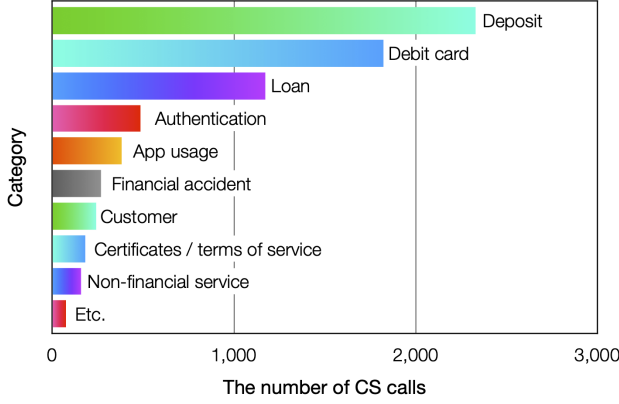


Figure 3: Number of CS calls in each CS category.

techniques such as backpropagation and gradient descent, aligning them more closely with the requirements of the specific task. The utilization of instruct datasets during fine-tuning enables the model to transfer its acquired knowledge from pretraining to the target task, thereby facilitating the capture of task-specific patterns and nuances. By synergistically incorporating both the general linguistic knowledge acquired through pretraining and the task-specific guidance from instruct datasets, fine-tuned models exhibit heightened performance and a more comprehensive contextual understanding, providing valuable implications for downstream AI applications in the realm of finance.

## 4 Evaluation Framework

This section describes the evaluation metrics for assessing the performance of the fine-tuning method using instruct datasets for customer dialogue summarization. We utilized two evaluation metrics: the recall-oriented understudy for gisting evaluation (ROUGE) [9, 20] and the reference and document aware semantic score (RDASS) [16]. In text summarization, two factors play a pivotal role: 1) effective selection of content that involves choosing the key information from the document and 2) accurate representation of the same meaning using different expressions (*i.e.*, paraphrasing). Although ROUGE is widely used as an evaluation metric for text summarization models, it has limitations in capturing the essence of the summary because it compares the similarity between expert-crafted reference summaries and model-generated sentences. In contrast, RDASS is a comprehensive evaluation metric that considers the relationships among the original document, reference summary, and model-generated summary. Compared to ROUGE, RDASS performed better in terms of relevance, consistency, and fluency of sentences in Korean. Therefore, we employed both ROUGE and RDASS as evaluation metrics, considering their respective strengths and weaknesses of each metric. The mathematical description of each metric is as follows:

### 4.1 ROUGE

ROUGE-N can be computed as the overlap of n-grams between the generated summary and reference summary. Specifically, ROUGE-1

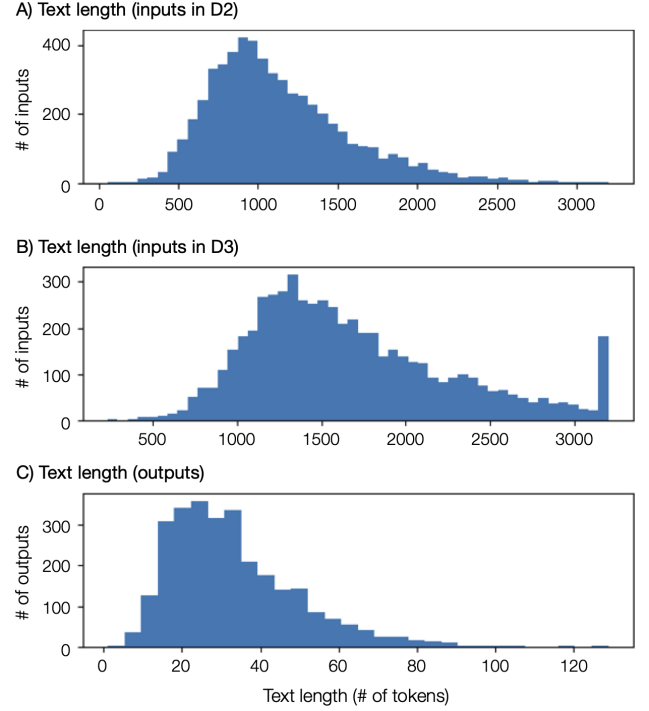


Figure 4: Distributions of the text length in the instruct datasets of KakaoBank’s CSC for (A) inputs in D1, (B) inputs in D2, and (C) outputs.

can be computed as the overlap of unigram (each word) between a generated summary and reference summary and is used to assess the quality of automatic summarization systems. In addition, ROUGE-2 can be computed as the overlap of bigrams (pairs of consecutive words) between a generated summary and reference summary and is also used to assess the quality of automatic summarization systems.

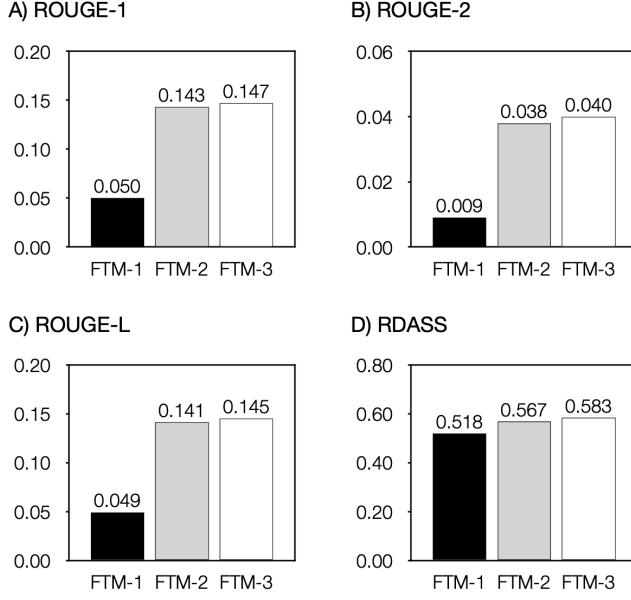
ROUGE-L is an evaluation metric that calculates the longest common subsequence (LCS) between a generated summary and reference summary, which captures the overall content overlap and order of words. It is also commonly used to assess the quality of automatic summarization systems.

### 4.2 RDASS

RDASS measures the similarity between the vectors of the original document and reference summary. Moreover, it measures the similarity between the vectors of the original document and generated summary. Finally, RDASS can be obtained by computing their average.

## 5 Experiments

This section provides an overview of the experimental setup, which encompasses various components, such as the preparation of instruct datasets and the development and evaluation of fine-tuned PLMs.



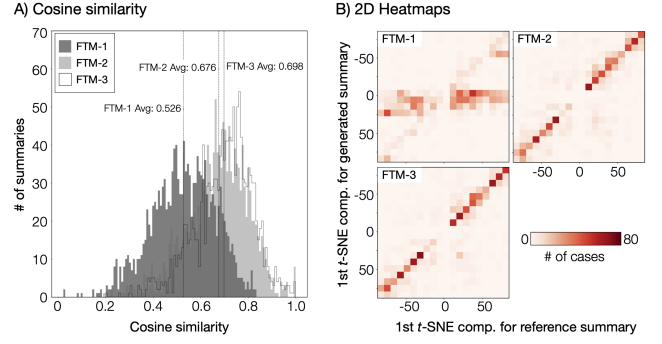
**Figure 5: Summary of model performance experimented through three major fine-tuning models (FTMs): FTM-1 represents fine-tuning Polyglot-Ko (5.8B) using the Koalpaca instruct datasets (see D1 in Fig. 2); FTM-2 represents fine-tuning Polyglot-Ko (5.8B) using the instruct datasets composed of the utterances of CS representatives and summarization (see D2 in Fig. 2); FTM-3 represents fine-tuning Polyglot-Ko (5.8B) using the instruct datasets composed of the comprehensive dialogues and summarization (see D3 in Fig. 2).**

### 5.1 KoAlpaca dataset

The KoAlpaca dataset is a translation of instruction-following data employed to fine-tune the Alpaca model, which is based on the LLaMA model [30]. It consists of 52K instruction-following instances organized in a JSON file and structured as a list of dictionaries (Fig. 2). Within each dictionary, three keys, namely instruction, input, and output, are associated with string values. The instruction entails comprehensive task descriptions for guiding the model’s task, resulting in 52K distinct instructions. The input exhibits variable content that potentially contains the value of an empty string. Finally, the output corresponds to the instruction’s answer generated using the utilization of the text-davinci-003 model.

### 5.2 KakaoBank’s CSC dataset

In this study, we used a subset of KakaoBank’s CS telephone inquiries dataset. We examined the daily CS records in October 2020 and selected a day characterized by a lack of concentrated inquiries on a specific CS category. The audio data was manually transcribed into text, with sensitive information masked for privacy. The CS engagements covered various topics, including financial services (deposits, debit cards, loans) and other areas like authentication, app usage, and financial accidents. Fig. 1A shows an example of a CS dialogue exchange. The CS session typically starts with a salutation and restatement of the customer’s inquiry for confirmation.



**Figure 6: Similarity measures between the reference and generated summaries. (A) Cosine similarity of the embedding vectors between the reference summaries and the generated summaries. (B) 2D heatmaps of the first component of  $t$ -SNE between the reference and generated summaries for the fine-tuning model 1 (FTM-1) (top-left), FTM-2 (top-right), and FTM-3 (bottom-left).**

This iterative conversation continues until the CS representative resolves the issue or provides potential solutions. We labeled the data to create instruct datasets for dialogue summarization in CS dialogues. Fig. 4 depicts the distribution of text lengths for complete dialogues, CS representatives’ utterances, and reference summaries.

### 5.3 Fine-tuning training

We divided the KoAlpaca and KakaoBank’s CSC instruct datasets into training and test datasets in a ratio of 8:2, respectively. We then used Polyglot-Ko (5.8B) as the baseline model and performed fine-tuning using various instruct datasets. The first FTM (FTM-1) involved fine-tuning using KoAlpaca instruct datasets (see Fig. 2A). This model served as a baseline for evaluating whether a Korean language model trained with general knowledge-based Korean instruct datasets could effectively perform dialogue summarization. The second FTM (FTM-2) involved fine-tuning using the utterances of the CS representatives and summarization (see Fig. 2B). The third FTM (FTM-3) involved fine-tuning using the entire CS dialogue and summarization (see Fig. 2C). Furthermore, an additional FTM (FTM-4) was developed by fine-tuning Polyglot-Ko (5.8B) by integrating the KoAlpaca and KakaoBank’s CSC instruct datasets.

### 5.4 Model performance evaluation

Each FTM model was subjected to inference procedures on the test dataset, followed by an assessment of the similarity between the reference and generated summaries using the ROUGE and RDASS metrics. In addition, the cosine similarity of the embedding vectors between the reference and generated summaries was measured to determine which FTM exhibited the best performance. Furthermore, the embedding vectors of the reference and generated summaries were transformed using  $t$ -distributed stochastic neighbor embedding ( $t$ -SNE) [31] and visualized as a 2D heatmap of the first  $t$ -SNE component to assess the model’s performance. Measurement of certain aspects of the FTM using ROUGE or RDASS proved to be



**Table 1: Examples of the reference and generated summaries using fine-tuning models (FTMs). The customer’s situation is highlighted in cyan, whereas the primary inquiry is highlighted in orange within the given context.**

<b>Reference summary</b>
Inquired about expediting the process because it was taking time after applying for the removal of the transfer limit, Informed about errors in the documents and provided guidance on resubmitting the required documents.
<b>Generated summary by FTM-1</b>
Because you are registered with an account that has a daily limit.
<b>Generated summary by FTM-2</b>
Applied the removal of the limit on the account. Requested urgent processing.
<b>Generated summary by FTM-3</b>
Requested expedited processing because it wasn’t being handled after applying for the removal of the transfer limit. Provided instructions for document submission and guidance on the process.

difficult, necessitating the validation of its effectiveness through qualitative methodologies.

## 5.5 Multiple instruction templates

To improve the zero-shot ability of the fine-tuned language model [33], we conducted fine-tuning using multiple instructions to enable the versatile utilization of the summarization task. For example, it can be applied to a task involving the detailed summarization of CS inquiries and the resolution process, as well as a task involving the brief summarization task of CS inquiries. To obtain various forms of summarization results for the same CS dialogue, we performed fine-tuning using multiple instruction templates to apply the fine-tuned model to various summarization tasks. Examples of the instruction templates are as follows:

- **Instruction 1:** Write a detailed summary of the inquiry and the resolution process in the following CS dialogue.  
Additional requirements of instruction 1: The summary must include the specific words and have a token length greater than or equal to 45.
- **Instruction 2:** Write a brief summary of the following CS dialogue.  
Additional requirements of instruction 2: The summary must include the specific words and have a token length of less than or equal to 25.
- **Instruction 3:** Write a short summary of the following CS inquiry within five words.  
Additional requirements of instruction 3: The summary must not include the specific words and must have a token length between 5 and 15.

In the above instructions, the specific words that should and should not be included in the summaries are as follows: guide, answer, progress, closed, emergency, resolved, and completed.

## 6 Experimental Results

In this section, we describe the evaluation results of the FTMs for CS dialogue summarization; we present the ROUGE and RDASS metrics for the various FTMs. Additionally, an assessment of the similarity between the reference and generated summaries by each FTM using cosine similarity and *t*-SNE methods are provided.

### 6.1 Performance of fine-tuning models

The summarization performances of the FTMs were evaluated by using two metrics: ROUGE and RDASS (Fig. 5). The reference fine-tuning method (FTM-1) yielded ROUGE-1, ROUGE-2, ROUGE-L, and RDASS scores of 0.050, 0.009, 0.049, and 0.518, respectively. Subsequently, the FTM-2 demonstrated improved performance, with ROUGE-1, ROUGE-2, ROUGE-L, and RDASS scores of 0.143, 0.038, 0.141, and 0.567, respectively. FTM-3 showed further enhancement, with ROUGE-1, ROUGE-2, ROUGE-L, and RDASS scores of 0.147, 0.040, 0.145, and 0.583, respectively. Among the three major FTMs, fine-tuning by incorporating comprehensive dialogue data and summarization delivered the most favorable outcomes.

We visually explored the summaries generated by the FTMs compared to the reference summaries (Table 1). FTM-1 produced a summary unrelated to the reference, while FTM-2 partially contained essential details. FTM-3, although slightly different, concisely summarized the customer’s inquiry and the CS representative’s guidance, closely aligned with the reference summary. Lee *et al.* (2021) reported that training a masked language model (MLM) with financial domain data led to an approximate enhancement of 3.3% in ROUGE-1 and 5.7% in ROUGE-L compared to the absence of non-financial data [17]. This observation implies that training substantial knowledge in the financial domain is imperative for achieving commendable performance in document summarization tasks within the financial context. Our research findings are in line with those of previous research, indicating that higher performance on a summarization task in the financial domain necessitates training models with appropriate financial data.

To investigate the hypothesis regarding the potential improvement in FTM performance through training with multiple sources of instruct datasets, additional experiments were conducted. The evaluation of the summarization performance for the model trained using fine-tuning with both the dialogue and KoAlpaca instruct datasets (FTM-4) resulted in ROUGE-1, ROUGE-2, ROUGE-L, and RDASS scores of 0.136, 0.035, 0.133, and 0.576, respectively. Nevertheless, these results revealed a lower performance compared with FTM-3, which was trained solely using the dialogue instruct dataset. Specifically, FTM-4 demonstrated a decrease in performance relative to FTM-3, with reductions of 7.6%, 14.6%, 8.1%, and 1.1% in ROUGE-1, ROUGE-2, ROUGE-L, and RDASS, respectively. These findings indicate that task-specific instruct datasets have a beneficial effect on enhancing the performance in fine-tuning for specific downstream tasks. However, the inclusion of a substantial number of instruct datasets unrelated to the task can have a detrimental effect on the model’s performance.

Finally, additional experiments were conducted to examine the impact of the baseline model size on the performance of the downstream task of dialogue summarization. In this study, instruct datasets identical to those used in FTM-3 were employed, with a

A) Task Instruction 1		B) Task Instruction 2		C) Task Instruction 3	
Instruction		Instruction		Instruction	
(Korean) 다음 상담 내용중 문의 사항과 처리결과를 자세히 요약해주세요		(Korean) 다음 상담 내용을 짧게 요약해주세요		(Korean) 다음 상담 내용 중 의사항을 다섯단어 이내로 짧게 요약해 주세요	
(Translated into English) Write a detailed summary of the inquiry and the resolution process in the following CS dialogue.		(Translated into English) Write a brief summary of the following CS dialogue.		(Translated into English) Write a short summary of the following CS inquiry within 5 words	
Output		Output		Output	
(Korean) 후불교통카드 사용금액 출금시간 한도금액 문의로 출금시간은 6시 이후로 안내하고 이체한도 확인할 수 있게 안내함		(Korean) 후불교통카드 사용금액 출금시간 한도 금액 문의로 안내함		(Korean) 후불교통카드 사용금액 출금시간 한도금액 문의	
(Translated into English) The inquiry was about the usage amount, withdrawal time, and limit amount of the postpaid transportation card. The withdrawal time was guided to be after 6 p.m., and instructions were provided on how to check the transfer limit.		(Translated into English) The inquiry was about the usage amount, withdrawal time, and limit amount of the postpaid transportation card. Information was provided in response to the inquiry.		(Translated into English) The inquiry was about the usage amount, withdrawal time, and limit amount of the postpaid transportation card.	

**Figure 7: Illustrative inference outcomes of the fine-tuning model using multiple instruction templates. The task instructions and their corresponding outputs are demonstrated for (A) an instance of instruction 1 (Write a detailed summary of the inquiry and the resolution process in the following CS dialogue), (B) instruction 2 (Write a brief summary of the following CS dialogue), and (C) instruction 3 (Write a short summary of the following CS inquiry within 5 words).**

modification of the baseline model to Polyglot-Ko (12.8B) for fine-tuning training (FTM-5). A performance evaluation was conducted using the ROUGE and RDASS metrics. FTM-5 model exhibited performance scores of 0.149, 0.042, 0.148, and 0.584 for ROUGE-1, ROUGE-2, ROUGE-L, and RDASS, respectively. Compared to FTM-3, these values displayed improvements of 1.8%, 4.4%, 1.9%, and 0.1% for ROUGE-1, ROUGE-2, ROUGE-L and RDASS, respectively. This observation highlights that larger parameters in the baseline PLM can help improve the performance in the dialogue summarization task.

## 6.2 Similarity measures

To assess the quality of generated summaries by the FTMs, we evaluated the cosine similarity of the embedding vectors between the reference summaries produced by humans and the generated summaries produced by the FTMs. As illustrated in Fig. 6A, the average cosine similarity values of the embedding vectors between the reference and generated summaries by FTM-1, FTM-2, and FTM-3 are 0.526, 0.676, and 0.698, respectively.

Additionally, we projected the embedding vectors of the summaries onto  $t$ -SNE and visualized the 2D heatmaps depicting the first  $t$ -SNE components alongside the reference and generated summaries (Fig. 6B). Within the 2D heatmap, an elevated value in the diagonal bins indicates a heightened degree of similarity between the sentences of the reference and generated summaries. Conversely, a higher value in the off-diagonal bins implies a reduced level of similarity between the sentences of the reference and generated

summaries. Specifically, in the case of FTM-1, the off-diagonal bins are predominantly populated, indicating a lower degree of similarity between the generated and reference summaries. Conversely, for FTM-2 and FTM-3, the distribution tends to concentrate more within the diagonal bins, suggesting a greater similarity between the generated and reference summaries. Notably, FTM-3 exhibits values in the off-diagonal bins that are closer to 0 compared to FTM-2, which suggests that the sentences of the generated summaries by FTM-3 are most similar to the reference summaries.

## 6.3 Fine-tuning with multiple instruction templates

Examples of inference results of zero-shot unseen data for the fine-tuning using multiple instruction templates, which are composed of three different types based on the length of the summary and whether specific words (see Section 5.5) are included in the summary, are shown in Fig. 7. Fine-tuning using a single instruction template produces almost similar outputs (*i.e.*, similarly generate summaries) even if various instructions are given during inference. However, fine-tuning using multiple instruction templates obtains different outputs depending on the instruction during inference. Specifically, Instruction 1 generated a detailed summary, whereas Instruction 2 produced a more concise summary. Instruction 5 provided a summary in five Korean words. Although it was expanded to more than five words during the translation process into English, it performed well in Korean in terms of the task.

## 7 Conclusions

Although GPT-3 has features such as multilingual support, there are several constraints to consider when applying it to solving specific business problems that aligns with the culture, policies, and regulations of each country. In this study, we proposed a method for addressing the downstream task of dialogue summarization by fine-tuning using real-world instruct datasets. We developed a reference FTM using Polyglot-Ko (5.8B) as the baseline PLM and the KoAlpaca instruct dataset containing various zero-shot and partially document summarization instructions. We compared this model with the FTM-3, which was fine-tuned using KakaoBank’s CS dialogues and summarization as the instruct dataset. The results demonstrated that FTM-3 based on KakaoBank’s internal datasets outperformed the reference model, showing a 199% and 12% improvement in ROUGE-L and RDASS, respectively.

Typically, the tasks of CS representatives are well-documented in the form of guidelines. For example, detailed explanations of financial products, methods of consultation or customer interaction, ways to summarize and condense content after CS calls, and methods of selecting CS categories are often organized in documents to ensure that employees are well-versed. However, tasks such as addressing customer complaints or summarizing CS records rely heavily on individual skills, even when following guidelines. Automating these areas of work was deemed infeasible until the advent of generative AI [27], which can memorize detailed product explanations, engage in conversations by following predefined guidelines, and effectively extract key points for summarization. By utilizing a fine-tuned language model to summarize CS dialogues,

we anticipate a certain level of standardization in summarization tasks between high- and low-skilled workers.

In this study, we developed a model that summarizes customer inquiries received through telephone calls. However, this model could be expanded to summarize chat-based CS inquiries or iterative email-based CS inquiries. Furthermore, by fine-tuning using different instructions, such as 1) simply summarizing the dialogue and 2) summarizing the dialogue in five words, it can be extended to categorize CS types. Our approaches can be applied to the CS operations of most financial institutions targeting retail customers.

In conclusion, this study highlights the importance of task-specific fine-tuning using appropriate instruct datasets to achieve effective performance in specific downstream tasks. We suggest that fine-tuning using real-world instruct datasets is a powerful and cost-effective technique for developing generative AI in the business domain.

## References

- [1] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. arXiv:1908.10063
- [2] Jacqui Ayling and Adriane Chapman. 2022. Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics* 2, 3 (aug 2022), 405–429.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901.
- [4] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. 2023. Generative AI at Work. arXiv:2304.11771
- [5] Yi-Syuan Chen and Hong-Han Shuai. 2021. Meta-Transfer Learning for Low-Resource Abstractive Summarization. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 14, 12692–12700. doi:10.1609/aaai.v35i14.17503
- [6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311
- [7] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence* 5, 3 (Jun 2023), 220–235.
- [8] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. arXiv:2303.10130
- [9] Kavita Ganesan. 2018. ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks. arXiv:1803.01937
- [10] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Association for Computational Linguistics, Hong Kong, China, 70–79.
- [11] Arnab Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The False Promise of Imitating Proprietary LLMs. arXiv:2305.15717
- [12] Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2023. The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey. arXiv:2104.14839
- [13] Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Jiyoung Hyun, and Sungho Park. 2022. Polyglot-Ko: Open-Source Korean Autoregressive Language Model.
- [14] Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Jiyoung Hyun, Sungho Park, and Kyubyong Park. 2023. A Technical Report for Polyglot-Ko: Open-Source Large-Scale Korean Language Models. arXiv:2306.02254
- [15] Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or Extractive? On Mitigating the Faithfulness-Abstractiveness Trade-off in Abstractive Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 1410–1421.
- [16] Dongyub Lee, Myeong Cheol Shin, Taesun Whang, Seungwoo Cho, Byeongil Ko, Daniel Lee, EungGyun Kim, and Jaechoon Jo. 2020. Reference and Document Aware Semantic Evaluation Methods for Korean Language Summarization. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5604–5616.
- [17] Huijie Lee, Wonsuk Yang, Chaehun Park, Hoyun Song, Eugene Jang, and Jong C. Park. 2021. Optimizing Domain Specificity of Transformer-based Language Models for Extractive Summarization of Financial News Articles in Korean. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*. Association for Computational Linguistics, Shanghai, China, 611–621.
- [18] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. 2023. ChatDoctor: A Medical Chat Model Fine-tuned on LLaMA Model using Medical Domain Knowledge. arXiv:2303.14070
- [19] Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. 2023. Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning. arXiv:2303.15647
- [20] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81.
- [21] Tiedong Liu and Bryan Kian Hsiang Low. 2023. Goat: Fine-tuned LLaMA Outperforms GPT-4 on Arithmetic Tasks. arXiv:2305.14201
- [22] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1906–1919.
- [23] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155
- [24] Seongmin Park, Dongchan Shin, and Jihwa Lee. 2022. Leveraging Non-dialogue Summaries for Dialogue Summarization. In *Proceedings of the First Workshop On Transcript Understanding*. International Conference on Computational Linguistics, Gyeongju, South Korea, 1–7.
- [25] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. arXiv:2302.12813
- [26] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction Tuning with GPT-4. arXiv:2304.03277
- [27] Michael Polanyi. 1966. *The Tacit Dimension*. University of Chicago Press, Chicago, IL.
- [28] George Prodan and Elena Pelican. 2022. Prompt scoring system for dialogue summarization using GPT-3. *TechRxiv* (5 2022).
- [29] Chad Syverson. 2011. What Determines Productivity? *Journal of Economic Literature* 49, 2 (Jun 2011), 326–65.
- [30] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca).
- [31] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.
- [32] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. arXiv:2212.10560
- [33] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=gEzrGC0zdqR>
- [34] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv:2302.11382
- [35] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myale Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali



- Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. arXiv:2205.01068
- [36] Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. arXiv:2104.05938