IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version Aug 16, 2024.

Digital Object Identifier 10.1109/ACCESS.2024.1120000

PETIS: Intent Classification and Slot Filling for Pet Care Services

NAMRAH ZAMAN [1,†], SEONG-JIN PARK [2,†], HYUN-SIK WON [1], MIN-JI KIM [1], HEE-SU AN [1], and KANG-MIN KIM [1,3]

[1] Department of Artificial Intelligence, The Catholic University of Korea, Bucheon-si 14662, Republic of Korea
[2] Department of Mathematics, The Catholic University of Korea, Bucheon-si 14662, Republic of Korea
[3] Department of Data Science, The Catholic University of Korea, Bucheon-si 14662, Republic of Korea
[†] These authors contribute equally to this work.

Corresponding author: Kang-Min Kim (kangmin89@catholic.ac.kr).

This work was supported by the Basic Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1C1C1010317) and the Research Fund, 2023, of The Catholic University of Korea.

ABSTRACT During the COVID-19 pandemic, the surge in online pet care services led to an increased demand for conversational AI systems specifically designed for the veterinary domain. However, traditional natural language understanding (NLU) tasks and datasets often fall short due to domain-specific terminology, the descriptive nature of user utterances, and the high cost of expert annotations. To fill this gap, we introduce PETIS, a novel dataset comprising 10,636 annotated utterances specifically designed for intent classification and slot filling in pet care domain, featuring 10 unique intent classes and 11 slot classes. PETIS addresses the scarcity of annotated data in this domain and serves as a challenging benchmark for evaluating NLU models. We demonstrate its effectiveness through experiments using state-of-the-art models, achieving 93.32 accuracy in intent classification and a Micro $F1^C$ score of 91.21 in slot filling using multi-task AdapterFusion. Furthermore, domain adaptation significantly enhanced performance, showcasing the potential of PETIS to drive research and development in conversational AI for online pet care services, offering a valuable resource for advancing the field.

INDEX TERMS Conversational AI, intent classification, Korean language understanding, natural language understanding, parameter-efficient fine-tuning, pet care services, slot filling

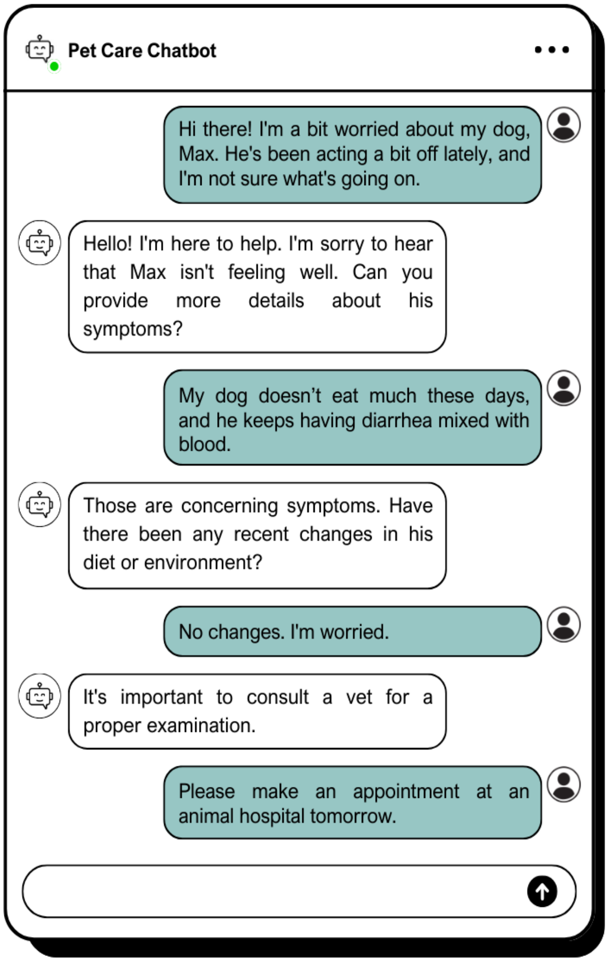
I. INTRODUCTION

The pet care market has expanded due to the increasing number of pet-owning households [1], as well as a significant surge in online pet care and veterinary services during the COVID-19 pandemic. This led to an increase demand for conversational AI systems that focus on prevention, diagnosis, and treatment [1] specifically for pet care services. Although task-oriented dialogue systems for healthcare services [2] have been explored, research on pet care services has been limited due to the unique challenges of this domain, such as the complexity of interpreting non-standardized, descriptive language used by pet owners and the necessity for domain-specific knowledge to accurately understand the user intent and respond to user utterances, which presents significant challenges for natural language understanding systems.

Natural language understanding tasks (e.g., intent classification and slot filling) are the core components of task-oriented dialogue systems [3], [4], [5], [6]. However, traditional NLU datasets, while effective in general contexts, are often insufficient for handling the specific veterinary terms and needs of pet care services. These datasets typically lack the focus on domain-specific terminology and struggle with the descriptive nature of user utterances, which is common in pet care services. For instance, existing datasets that excel in tasks like booking a flight [7], making a restaurant reservation [8], or playing a song [9], do not perform well in understanding the complexities of pet care services. This is because understanding users utterances in pet care services may be more difficult than in these existing conversational services. This difficulty can be attributed to two factors:

1) When users refer to information about their pets (e.g., My dog does not eat much these days and he keeps having diarrhea mixed with blood), they usually use descriptive expressions in the free text rather than standard veterinary terms, as shown in Figure 1.
2) Annotation for pet care services is expensive because it

[1] https://daxueconsulting.com/south-koreas-pet-industry/



Hi there! I'm a bit worried about my dog, Max. He's been acting a bit off lately, and I'm not sure what's going on.

Hello! I'm here to help. I'm sorry to hear that Max isn't feeling well. Can you provide more details about his symptoms?

My dog doesn't eat much these days, and he keeps having diarrhea mixed with blood.

Those are concerning symptoms. Have there been any recent changes in his diet or environment?

No changes. I'm worried.

It's important to consult a vet for a proper examination.

Please make an appointment at an animal hospital tomorrow.

FIGURE 1. Example of a user expressing symptoms in free-text format.

requires domain experts, such as veterinarians, to correctly identify and label the symptoms mentioned.

To tackle these challenges, we introduce **PETIS** (**PET** care services **I**ntent classification and **S**lot filling), a novel dataset designed specifically for conversational AI in the pet care domain. PETIS comprises 10,636 annotated utterances, featuring 10 unique intent classes and 11 slot classes, providing a focused and effective resource for enhancing the accuracy of NLU tasks in pet care services. The dataset is particularly valuable as it captures the complexity of user language in this domain, offering a more relevant benchmark for evaluating and improving NLU models.

The creation of PETIS involved a careful process, beginning with the collection of data from real pet owners and online pet communities. This data was then carefully filtered, annotated, and validated by domain experts to ensure its relevance and accuracy. Details of the data construction and annotation processes are provided in Sections IV and V, respectively. This approach ensures that PETIS accurately reflects the specific veterinary terms and needs of pet care services, making it a robust tool for advancing research in this field.

To demonstrate the effectiveness of PETIS, we conducted extensive experiments using state-of-the-art models, including both large and base models, across various techniques. Our experiments, which utilized a multi-task framework based on AdapterFusion [10], [11], achieved a notable accuracy of 93.32 in intent classification and a Micro F1^C score of 91.21 in slot filling. These results highlight the robustness and superior performance of models trained on PETIS, proving its potential to significantly improve conversational AI systems in the pet care domain. In summary, our contributions are threefold:

1. We have introduced PETIS, an innovative dataset specifically designed for conversational pet care services. PETIS includes intent classification, slot filling, and symptom matching tasks, providing a comprehensive real-world benchmark for evaluating natural language understanding models in the pet care domain.
2. Our extensive evaluation of PETIS with state-of-the-art models demonstrates robustness and superior performance, offering a unified solution for handling multiple tasks simultaneously. Utilizing a multi-task framework based on AdapterFusion, we significantly improved the efficiency and effectiveness of conversational pet care services.
3. We release our dataset for research and educational purposes[2]. We hope that our dataset and code will be a starting point for future research and applications in conversational pet care services.

The PETIS is a dataset composed of Korean. For clarity and convenience, the paper presents all examples in both English and Korean.

II. RELATED WORK

A. PRE-TRAINED LANGUAGE MODELS FOR INTENT CLASSIFICATION AND SLOT FILLING

Our work on PETIS, focusing on intent classification and slot filing in pet care services, aligns with broader research in conversational AI and natural language understanding (NLU). The use of pre-trained language models like BERT [12] and RoBERTa [13], has revolutionized NLU, particularly in tasks requiring deep contextual understanding. BERT has become a cornerstone in NLU research because of its deep bidirectional transformers. However, RoBERTa robustly optimized the pretraining process of BERT, enhancing its performance in various NLU tasks.

Recent studies continue to explore the capabilities of BERT-based models, especially in the context of multi-task learning. The study [14] demonstrated the effectiveness of a BERT variant for joint intent classification and slot filling. The study [15] explored BERT-based encoders for sequence classification and joint intent-slot filling in dialogue acts, demonstrating the versatility and efficiency of these models

[2] Dataset and Code will be available at https://github.com/cuknlp22/Intent-Classification-and-Slot-Filling-for-Pet-Care-Services.git

in multi-task learning. Similarly, another study [16] leveraged BERT for next-generation spoken language understanding, focusing on joint intent classification and slot filling, which is essential for applications requiring simultaneous handling of multiple tasks. The effectiveness of multi-task learning approaches is further exemplified by recent advances like AdapterFusion. The study [11] introduced an adapter-based unified model that facilitates multiple tasks, showcasing how adapter-based methods can efficiently handle multi-task learning without requiring extensive retraining of the entire model.

Additionally, SimCSE model [17] has proven to be a robust approach for improving sentence embeddings in similarity-based tasks. It uses contrastive learning to improve the quality of sentence embeddings, which helps it perform better on tasks that need to evaluate semantic similarity. Another study [18] developed a pre-trained joint model JMBSF that integrates BERT with semantic fusion, showing improved results on benchmark datasets by effectively utilizing contextual semantic features.

In the domain of healthcare, BERT-based models have shown significant promise. The study [1] explored the integration of chatbots into healthcare services, emphasizing their role in improving medical diagnostics and patient interactions. Similarly, another study [2] proposed a BERT-based medical chatbot that enhances the precision and reliability of healthcare conversations. Beyond healthcare, the flexibility of pre-trained models has been demonstrated in diverse domains, such as study [19], which applied similar techniques to analyze privacy policies, highlighting the versatility of BERT-based approaches across different domains.

The growing interest in multi-tasking and multilingual learning approaches is also evident in the study [20], who proposed a multilingual multi-task approach for intent classification and slot filling, achieving notable performance improvements across different languages. These advancements underscore the critical role that advanced pre-trained models and learning techniques play in specialized domains. Understanding user intent and accurately extracting relevant information are essential for delivering effective conversational AI services.

B. DATASETS FOR INTENT CLASSIFICATION AND SLOT FILLING

Intent classification and slot filling tasks are crucial for various NLU applications, requiring datasets that accurately capture the specific challenges of different domains. Traditional benchmark datasets such as ATIS [21], Snips [22], MultiATIS++ [23], TOPv2 [24], and MTOP [25], provide comprehensive resources for developing and evaluating models in these areas. However, the increasing complexity of domain-specific applications has led to the development of more specialized datasets that address the unique needs of particular domains.

Recent efforts have focused on creating datasets that address the unique needs of specific domains. For instance, the study [19] introduced a dataset designed for analyzing privacy policies, which presents distinct challenges due to the legalistic language involved. Similarly, another study [26] developed a dataset focused on the agricultural sector, addressing the need for NLP tools that can interpret and process agricultural terminology effectively. Similarly, the PETIS dataset we introduce in this paper is tailored for the pet care domain, specifically designed to handle the unique vocabulary and descriptive language used by pet owners in this domain.

In addition to domain-specific datasets, there has been significant progress in developing datasets for various languages, particularly those that are underrepresented in NLU research. The study [27] introduced a dataset focused on intent classification and slot tagging for Indic languages within the agricultural domain, expanding the linguistic diversity of resources. Further contributing to the linguistic diversity, study [28] developed MIntRec, a dataset for multimodal intent recognition, which integrates multiple data modes for intent classification. Another, study [29] proposed a home assistant dataset in Bangla and Sylheti languages, while study [30] introduced ArBanking77, a dataset in modern and dialectical Arabic. Additionally, study [31] created a dataset for intent classification and slot filling in Vietnamese, demonstrating ongoing efforts to enhance NLU resources across different languages and cultural contexts.

Despite these advancements, there remains a need for more datasets that address the specific challenges of particular domains and languages. Our work with PETIS contributes to this effort by providing a novel dataset designed for the pet care domain in Korean. This dataset is intended to bridge the gap between general-purpose NLU models and the specific needs of pet care services, offering a valuable resource for both research and practical applications in this field.

III. TASK DESCRIPTION

For the conversational pet care service to accurately understand the user's utterance and respond appropriately, the following tasks are required:

1) Intent classification: Classifying user utterance intents.
2) Slot filling: Analyzing user utterances to extract meaningful information. The conversational system should continue to interact with users until the required information is filled.
3) Symptom matching: Matching symptom names described by users to names within a veterinary ontology.

A. MAIN TASK 1: INTENT CLASSIFICATION

Conversational pet care services aim to assist users to accomplish tasks related to pet care. Understanding user intent is crucial for generating appropriate responses because the nature of the response varies according to the user's intent. Intent classification is the process of categorizing user utterances according to a specific intent. For instance, when the user inputs a sentence like "Please make an appointment at an animal hospital tomorrow" ("내일 동물병원 예약해주세요"),

the service needs to recognize the intent, such as scheduling a *Vet appointment*.

Intent classification involves assigning a user utterance to a relevant intent from a predefined list. In developing intent for conversational pet care services, we align with the three categories proposed in study [1] for conversational systems in healthcare provision. These categories include "Diagnosis," which involves identifying specific conditions similar to consulting a medical specialist; "Prevention," aimed at tracking health, building awareness, and preventing health declines through habit-building; and "Therapy," which provides assistance or treatment for specific health declines or conditions. We constructed an intent to align the three roles. In addition to these three categories, we added the "Others" category to handle *Vet appointment* and *Vet clinic recommendation* intents. The "Others" category includes the intents related to "add details" to the conversational system's question (i.e., *Positive*, *Negative*, and *Uncertain* intents). Finally, we added a *Fallback* intent to handle out-of-range utterances not included in the previously mentioned intent. Therefore, the following is the list of intents covered in this study, along with their respective descriptions:

1) Diagnosis
   - *Diagnostic test feature*: An utterance describes the condition of a pet and requests assistance.
   - *Vet appointment*: An utterance referring to a veterinarian's appointment.
2) Prevention
   - *Disease prevention inquiry*: An utterance asking for information about a disease or how to prevent it.
   - *Vet clinic recommendation*: An utterance requiring veterinarian recommendation that meets certain conditions.
3) Therapy
   - *Disease treatment inquiry*: An utterance asking about the treatment and surgical procedure of the disease.
   - *Disease treatment cost inquiry*: An utterance asking about the cost of treating a disease.
4) Others
   - *Positive*: An affirmative answer to a question.
   - *Negative*: A negative answer to a question.
   - *Uncertain*: An answer given when the user is unsure about a question.
   - *Fallback*: An out-of-range utterance.

B. MAIN TASK 2: SLOT FILLING

If the utterance "Please make a reservation at an animal hospital tomorrow" ("내일 동물병원 예약해주세요") was classified as a *Vet appointment*, it can be seen that this is intended to make a veterinarian's appointment. However, information such as the name of the veterinary clinic and reservation time cannot be known only by utterance and intent. Therefore, a process called slot filling is required in which the conversational system continues to ask questions until all the necessary information is filled.

To deal with the 10 intents mentioned above, we constructed a slot list as follows:

1) *Symptom*: Indicates the symptom that has appeared
2) *No symptom*: Indicates the difference between statements that describe the presence of symptoms and those that indicate there are no symptoms
3) *Species*: Indicates a pet species
4) *Name*: Indicates the name of a pet
5) *Date*: Indicates date information
6) *Time*: Indicates time information
7) *Disease*: Indicates the disease mentioned by the user
8) *Location*: Indicates the region and location information mentioned by the user
9) *Hospital*: Indicates the name of the veterinarian hospital that the user mentioned
10) *Hospital information*: Indicates the desired veterinarian conditions mentioned by the user
11) *Age*: Indicates the age of a pet

Unlike those who have experience in veterinary care, ordinary people who raise pets tend to describe the animal's condition in natural language (e.g., My dog has diarrhea mixed with blood)(우리 강아지가 피가 섞인 설사를 해요). Therefore, symptom mentions in utterances are not easily coded within the existing ontology. Therefore, we dealt with the symptoms described in natural language through text-based annotation.

The list of slots that must be filled to perform the final action for each intent is different. The slots that must be filled for each intent are given in Table 1. If the intent "*Vet appointment*" is identified in the utterance example, the conversational pet care service and the user continue communicating until the slot *Date*, *Time*, *Hospital*, *Name*, *Species*, *Age* are filled.

TABLE 1. Required slots for each intent.

Intent	Required Slots
Diagnostic test feature	Symptom
Disease prevention inquiry	Disease
Disease treatment inquiry	Disease, Species
Disease treatment cost inquiry	Disease, Species, Age
Vet appointment	Date, Time, Hospital, Name, Species, Age
Vet clinic recommendation	Location, Hospital information
Positive, Negative, Uncertain, Fallback	No slots required

C. AUXILIARY TASK: SEMANTIC MATCHING

As mentioned in Section III-B, pet owners explain their symptoms in their native language (e.g., bloody stool → stool mixed with blood). We used a semantic matching task and measured the accuracy of the top-k predictions to see how often the right symptoms were in the top predictions. This task is critical for increasing the accuracy of conversational services in comprehending and responding to user inquiries. By properly matching user symptom within a veterinary ontology, the system may provide more relevant and precise guidance. The semantic matching task also helps in finding frequent variances in symptom, allowing for more effective training of the language model.

IV. DATA CONSTRUCTION

A. DATA SELECTION

To create user utterances, we initially recruited pet keepers with real-life experiences related to companion animal symptoms or diseases. Following this, we conducted web scraping in a pet-related community with 200,000 users to gather inquiries from pet keepers. Web scraping is the process of programmatically extracting and structuring web data. Despite similarities in health-related queries between the community questionnaire and the conversational system, variations exist in question expression, symptoms, and text length.

1) INITIAL COLLECTION

Two types of utterances were collected. The first type consisted of data generated from actual pet owners. We achieve this by recruiting workers who have owned pets for over seven years. We provided workers with a table listing frequently occurring symptoms or diseases in animals, and they generated user utterances based on the table and their own experiences, resulting in 5,578 user utterances. The second type of data was from the pet community, as described in Section IV-A. We collected health-related questions from the pet community using various community forums related to pets in the pet community (for example,. Veterinary pet supplements and nutraceuticals[3], Best Pet Supply Sites[4]). However, when we conducted web scraping to collect data on various expressions for symptoms, we only considered health-related community forums that described pet symptoms and asked for help, resulting in an additional 42,803 user utterances.

2) FILTERING

From the web scraping data, we eliminated utterances that were not related to pet medical counseling, which is our scope. We also filtered out the utterances that were too short ($<$ 3 words) or too long ($>$ 240 words). After this process, the collected utterances ratio between actual pet owners and web scraping sources became approximately 4:6.

3) POST-PROCESSING

The process involves reviewing whether an utterance is suitable for intent purposes before the annotation proceeds. We asked domain experts (who had been working in the pet care domain for more than three years) to examine the selected utterances. The goal of the examination is to ensure that the collected utterances cover the following four types of intent (i.e., "Diagnosis", "Prevention", "Therapy" and "Other"). These four categories indicate the chatbot's role in providing healthcare.

V. DATA ANNOTATION

After collecting the raw user utterances, we labeled each utterance using both the intent and slot labels.

A. ANNOTATION PROCEDURE

In addition to annotating sentences with intents, we aimed to identify text spans in sentences that explain the specific details of the utterances. For example, in the sentence "My dog's fur has become dull and he is vomiting" ("강아지 털이 푸석해지고 구토를 해요"), the underlined text span conveys the purpose of the data collection. In our annotation schema, we refer to the identification of such text spans as slot filling. As mentioned in Section III-B, there are 11 slot labels. These slots are associated with a list of attributes, for example, *"Diagnostic test feature"* and *"Disease treatment inquiry"* have the attributes of *Disease, Species, Name,* etc. In Section V-B2 several examples are presented.

General crowdworkers, such as Amazon Mechanical Turk, are unsuitable for annotating utterances in pet care services due to their lack of domain-specific knowledge and expertise required for accurately understanding and interpreting pet-related contexts. We hired three annotators who have more than seven years of experience raising pets. We wrote a detailed annotation guideline and pre-tested it through multiple rounds of pilot studies. The guidelines were updated with notes to resolve complex or corner cases during the annotation process. The annotators were presented with one sentence and asked to perform the annotation. Detailed guidelines are provided in Section V-B. The domain expert closely monitored the annotation process. The annotators worked for 10 weeks, with an average of 7 hours per week, and completed annotations for 10 intents. Each annotator was paid $10 per hour.

B. ANNOTATION GUIDELINE

We performed the following annotation process for our interactive pet care service. These guidelines describe the rules and examples used to annotate a pet care service corpus dataset. For the annotators, read each sentence in the pet care service corpus, identify the appropriate intent and slot, and mark it in the given CSV file.

1) INTENT

The intent refers to the reason or purpose of the user's utterance. The response of an interactive pet care service may vary depending on the user's utterance intention. Therefore, it is very important to properly grasp the utterance intention. List of intent covered in the commentary presented in Table 2.

[3] https://www.mediacityvets.com/post/supplements-and-nutraceuticals
[4] https://www.bestpethouse.com/

TABLE 2. Annotation Schema for Intent.

Intent	Intent Examples
Diagnostic test feature: When a user is talking about pet symptoms.	• My dog suddenly stopped eating since yesterday and vomited this morning. (우리 강아지가 갑자기 어제부터 밥을 안 먹더니 오늘 아침에는 토를 했어요.) • He has been eating a lot in a hurry, but strangely, he doesn't vomit. (요즘 밥을 급하게 많이 먹는데 신기하게 토는 안 하더라고요.) • He drinks a lot of water and pees a lot. (물을 엄청 많이 마시고, 오줌을 많이 싸요.) • His legs get swollen and walks awkwardly. (다리가 팅팅 부어 오르고 어정쩡하게 걸어요.) • My dog keeps peeing because he is an old dog. (우리 강아지가 노견인데 자꾸 오줌을 지려요.)
Vet appointment: When a user wants to make a hospital appointment.	• Please make a reservation at an animal hospital near Gangnam Station. (강남역 근처 동물병원으로 예약해줘요.) • Please make a reservation for Sarang Animal Hospital at 11:30 a.m. tomorrow. (사랑 동물병원 내일 오전 11시 30분으로 예약해주세요.) • Please make a reservation for Happy Animal Hospital near Tanhyeon at 3 p.m. on September 15th. (탄현 근처 해피동물병원 9월 15일 오후 3시로 예약해줘.) • I will make a reservation there. (거기로 예약하겠습니다.) • I will be there at 6 p.m. today. (오늘 오후 6시에 갈게요.)
Vet clinic recommendation: When a user inquires about the location of a veterinary hospital or requests a recommendation.	• Please tell me the animal hospital near Daehwa Station. (대화역 근처동물병원 좀 알려주세요.) • Please let me know the nearest animal hospital to my house. (집에서 제일 가까운 동물병원 아무 데나 알려주세요.) • Please tell me about an animal hospital that also does dog grooming. (강아지 미용도 해주는 동물병원 알려주세요.)
Disease treatment inquiry: When a user asks about the cost of treatment and hospitalization for a disease.	• My dog has uterine infection, what should I do to get better quickly? (뽀송이가 자궁축농증인데 빨리 나으려면 어떻게 해야 하나요?) • How do you treat my dog for a cold? (강아지 감기 치료는 어떻게 해?) • Should he receive regular treatment for high blood pressure? (고혈압이면 주기적으로 치료를 받아야 하나요?)
Disease treatment cost inquiry: When a user ask about the cost of treatment and hospitalization for a disease.	• How much does it cost to operate on pyometra? (자궁축농증 수술하는데 얼마나 드나요?) • He has a fracture; does he need to be hospitalized? How much is the hospitalization fee if he is hospitalized? (골절을 당했는데, 입원도 해야하나요? 입원하면 입원비는 얼마나 나와요?) • I'm curious about the cost of neutering surgery. (중성화 수술 비용이 궁금합니다.)
Disease prevention inquiry: When a user asks about about disease information or how to prevent it.	• Please tell me about rabies. (광견병에 대해서 알려주세요.) • Is gastrinoma also a disease? (가스트린종도 질병인가요?) • What can I do to prevent my dog from catching a cold? (감기 안 걸리게 하려면 어떻게 해야 하나요?) • My dog has glaucoma, what is it? (우리 집 강아지가 녹내장에 걸렸다는데, 그게 뭐야?)
Positive: When a user expresses positively to the question of interactive pet care services.	• Yes, that's right. (예 맞습니다.) • I think so. (그런 것 같아요) • Yes. (네네 맞아요)
Negative: When a user expresses negatively to the question of interactive pet care services.	• No, it's not. (아닌데요.) • He does not have any symptoms like that. (그런 증상은 없었어요.) • No. (아니야.)
Uncertain: When a user say they are not sure about the question of interactive pet care services.	• I don't know. (잘 모르겠는데.) • Oh, I'm not sure about that. (아 그건 잘 모르겠어요.) • Uh, I don't know. (어.. 잘 모르겠어요.)
Fallback: When a user's utterance is not included in the above intent.	• Please recommend cat food. (고양이 사료 추천해주세요.) • My child's health has been so bad lately that I gave him OOO medicine, and I think he's getting better. (우리 아이가 요즘 건강이 너무 안 좋아서 OOO 약을 먹었더니 좀 나아진것 같아요.) • Please recommend eye drops. (안약 추천 좀 해주세요.)

2) SLOT

Of these 10 slots, only "*Symptom*" and "*No symptom*" are the parts that the user describes about the symptoms of the pet, so you should tie the predicate with the symbol <>. Please specify that the rest of the slots are made up of nouns or phrases and refer to the following example to comment:

- My <Bbosongi:*Name*> has been <vomiting:*Symptom*> for a while.
  (우리 <뽀송이:*Name*>가 얼마전부터 <구토를 해요:*Symptom*>.)
- My <puppy:*Species*> <does not have diarrhea:*No symptom*>, <but his anus is swollen:*Symptom*>.
  (우리 <강아지:*Species*>가 <설사는 안 하는데:*No symptom*>, <항문이 부어있어요:*Symptom*>.)
- While I was brushing his teeth, I noticed <bleeding:*Symptom*> and his <gums were swollen:*Symptom*>.
  (<양치하다가 피가 나서:*Symptom*> 보니까 <잇몸이 부어있었어요:*Symptom*>.)
- My baby's <eyes are very swollen:*Symptom*>, and his <pee is sticky:*Symptom*>.
  (우리 애기 <눈 쪽이 엄청 부어오르고:*Symptom*>, <오줌이 끈적끈적했어요:*Symptom*>.)
- Please make a reservation at an animal hospital near <Gangnam Station:*Location*.>
  (<강남역:*Location*> 근처 동물병원 예약해줘.)
- Please make a reservation at <Sarang Animal Hospital:*Hospital*> <tomorrow:*Date*> at <11:30 a.m.:*Time*>.
  (<사랑 동물병원:*Hospital*> <내일:*Date*> <오전 11시 30분:*Time*>으로 예약해주세요.)
- Is it okay if I go around <10:30 a.m.:*Time*>?
  (<오전 10시 30분:*Time*> 쯤에 가도 괜찮을까요?)
- Please tell me the animal hospital where <parking lot: Hospital Information> is located.
  (<주차장:*Hospital Information*> 장소 있는 동물병원 알려줘요.)
- Please tell me about the animal hospital that also offers <puppy:*Species*> <grooming:*Hospital Information*>.
  (<강아지:*Species*> <미용:*Hospital Information*>도 해주는 동물병원 알려주세요.)
- Please go to a hospital near the <house:*Location*> with a long history of animal surgery.
  (동물 수술 경력이 오래된 <집:*Location*> 근처 병원으로 부탁합니다.)
- Please recommend a hospital that is good at operating on <pyometra:*Disease*>.
  (<자궁축농증:*Disease*> 수술 잘하는 병원 추천해주세요.)
- If it's <high blood pressure:*Disease*>, should I receive regular treatment?
  (<고혈압:*Disease*>이면 주기적으로 치료를 받아야 하나요?)
- What is <gastrinoma:*Disease*>?
  (<가스트린종:*Disease*>이 뭐야?)
- <Lulu:*Name*> ate chocolate yesterday.
  (<루루:*Name*>가 어제 초콜릿을 먹었어요.)
- Please recommend <cat:*Species*> food.
  (<고양이:*Species*> 사료 추천해줘.)
- Tell me about the famous <dog:*Species*> cafe.
  (유명한 애<견:*Species*>카페 알려주세요.)

3) MORE RULES/NOTES OF ANNOTATIONS

- No slots exist in any sentence. For such sentences, we can move on to the next sentence.
- *Vet Appointment* and *Vet Clinic Recommendation* intents can also appear at the same time, such as "Please make a reservation at an animal hospital that also provides dog grooming" ("개 손질도 해주는 동물병원으로 예약 부탁드립니다"). In addition to these two combinations, more than one intent can simultaneously appear in a sentence. However, for our dataset, only one intent should be annotated, please mark this as the intent that is considered meaningfully closer. If it is too ambiguous, please contact us, and we will help you.
- There should be no space between ":" and slot names when displaying a slot, such as <vomit:*Symptom*>. In addition, there should be no space between slot names and ">".
- When displaying a slot, include punctuation marks such as "." "," and so on. For example:
  - "My Bbosongi keeps vomiting." → "My <Bbosongi:*Name*><keeps vomiting:*Symptom*>."
- When displaying a slot, it is not necessary to italicize it, however the case must be kept.
  - <Bbosongi:*name*> → x
  - <Bbosongi:*Name*> → o
- In the case of *Date* and *Time* slots, it can also appear in hospital reservations and recommended utterances, but it can also appear if the user describes the symptoms of the pet. At this time, the *Date* can be displayed on the date of symptom onset and *Time* can be displayed on the time of symptom onset.
  - Since <yesterday:*Date*>, my<Bbosongi:*Name*> has been having <diarrhea:*Symptom*>.
  - Since <two days ago:*Date*>, she <has not been eating:*Symptom*> and <has not been going outside:*Symptom*>.
- In the case of the slot of *Hospital Information*, it may be confusing to annotate it in units of nouns or phrases because it describes the conditions of the animal hospital that the user wants. To prevent this, the *Hospital Information* only annotates what can be objectively divided.
  - Recommend an animal hospital that performs the surgery first. → "Surgery first" is not objectively divided, so it is not annotated.
  - Please recommend a veterinary hospital with a doctor with an extensive history of surgery. → "extensive history of surgery" cannot be objectively divided, so no comment is made.
  - Tell me the animal hospital with <parking space:*Hospital Information*>.

  - Please recommend a veterinary hospital with a <female doctor:*Hospital Information*>.
- As mentioned above, *Symptom* and *No symptom* slots must also be marked with a predicate. When annotating symptoms, they should divided as much as possible. If you can not even mark the predicate while annotating, please refer to the following examples:
  - Our Rossi has <vomiting:*Symptom*> and <diarrhea:*Symptom*>.
  - <Broken teeth:*Symptom*>, <bleeding from the inside of the mouth:*Symptom*>
  - My dog's <eyes and mouth were swollen:*Symptom*>, and he kept <shaking his head:*Symptom*>.

We obtained the dataset by crawling from an actual large-scale pet community, which may contain information that is challenging to annotate according to the guidelines. We update the guidelines to provide more accurate annotations when we encounter ambiguous data.

C. POST-EDITING AND QUALITY CONTROL

To find the inter-annotator agreement for each segment of the annotated utterances, we used Krippendorff's alpha [32]. The annotators were asked to discuss their annotations and re-annotate those sections using token-level α_k below 0.75.

Table 3 presents Krippendorff's alpha scores for both slot and intent annotations before and after the agreement process. For slot annotations, the average Krippendorff's alpha score increased from 0.6653 to 0.7900 after the annotators discussed and re-annotated. However, for intent annotations, Krippendorff's alpha score improved from 0.7053 to 0.8004 after the agreement process. The process involved three annotators working together to resolve discrepancies in their intent tags. Specifically, the process included the following steps:

- Annotators identified data where the intent tags from the three annotators did not match.
- Each annotator reviewed and re-tagged the data, particularly focusing on tags that were significantly different.
- A consensus discussion was conducted for ambiguous tags to reach a unified decision.

Overall, the post-editing and quality control processes, including re-annotation and expert review, substantially enhanced the quality and consistency of the annotations, as evidenced by the improved Krippendorff's alpha scores.

TABLE 3. Krippendorff's Alpha Scores for Slot and Intent Annotations.

Metric	Before Agreement	After Agreement
Slot Krippendorff	0.6650	0.7900
Intent Krippendorff	0.7053	0.8004

D. DATA STATISTICS AND FORMAT

We compiled 10,636 annotated utterances in the PETIS corpus after performing annotation, post-editing, and quality control. Table 4 presents the statistics for the PETIS corpus. The total number of examples per intent and slot is presented in Tables 5 and 6, respectively.

TABLE 4. Statistics of the PETIS Corpus.

Dataset	Train	Dev	Test
# Sentences	8,508	1,065	1,063
# Slots	153,725	20,284	19,337
Avg. sentence length	32.075	33.745	32.246
Avg. # slot / sent.	11.728	12.514	11.498
Avg. slot length	2.990	3.013	2.992

TABLE 5. Total number of examples per intent.

Intents	Number of examples
Diagnostic test feature	3771
Vet appointment	493
Disease prevention inquiry	511
Vet clinic recommendation	443
Disease treatment inquiry	326
Disease treatment cost inquiry	552
Positive	136
Negative	136
Uncertain	260
Fallback	4008

TABLE 6. Total number of examples per slot.

Slots	Number of examples
Symptom	6869
No symptom	70492
Species	1566
Name	2682
Date	394
Time	483
Disease	1165
Location	2665
Hospital	559
Hospital information	1364
Age	3114

VI. MODELS AND SETUP

PETIS is a conversational pet care service dataset consisting of user input, intent, and slot labels. It provides annotations of the user's utterance intent and the corresponding text spans in the user's utterance. The intent classification task identifies a user's utterance intent, while the slot filling task extracts information to provide services appropriate for each intent.

Furthermore, we perform an auxiliary task, the symptom matching task, when the user asks for the name of the disease in the companion animal. For effective disease prediction, it extracts symptoms from a dataset that are semantically similar to the input symptom.

A. METHODOLOGY FOR INTENT CLASSIFICATION AND SLOT FILLING

In the context of multi-tasking, full fine-tuning is parameter-inefficient and requires excessive computational resources, including GPU memory and storage [33]. It may not optimally utilize limited data, leading to a potential knowledge loss when adapting to new tasks. Parameter-Efficient Fine-Tuning (PEFT) [34], on the other hand, solves these problems by learning only a few extra parameters of the model with fully fine-tuned weights. By freezing the original model weights, PEFT prevents catastrophic forgetting, ensuring that previously acquired knowledge is retained while adapting to new tasks. This implies a reduction in the required storage and computing power. Adapter tuning [33], Prefix-tuning [35], P-tuning v2 [36], LoRA [37], and AdapterFusion [10], [11] are notable PEFT techniques.

1) ADAPTER TUNING

We first employ PEFT with an adapter-based perspective to explore our dataset. Adapters encapsulate the backbone model, making it easy to add or remove them for different tasks like intent classification and slot filling. For this purpose, we configured and activated adapter modules tailored to each task within the base model. The shared layers of the pre-trained model captures the general meaning of language, while the adapter modules fine-tuned this knowledge to the specific tasks at hand.

During training, the appropriate adapter for each task is activated and trained, with losses from both intent classification and slot filling backpropagated through their respective adapter modules. For multi-task learning, a multi-task adapter is activated and trained. For single-task learning, specific adapters for intent classification (intent adapter) and slot filling (slot adapter) are activated and trained. We used gradient accumulation to handle larger effective batch sizes, ensuring the efficient use of computational resources. This approach leverages the capabilities of the pre-trained model while allowing focused training on task-specific requirements. By managing the active adapters per task, we achieved efficient multi-task training and yielded promising results for our dataset.

2) P-TUNING V2

We also examined our data using two-stage P-tuning v2 [36], focusing on both domain-specific and task-specific adaptations for intent classification and slot filling.

First, we applied domain-specific P-tuning to help our model understand the unique language and terms used in pet care services. This step made the model better at understanding user intent related to pet care services.

Next, for two main tasks: intent classification and slot filling, we performed task-specific P-tuning. For intent classification, the model learned to identify the specific intent behind user utterances, such as scheduling a *Vet appointment*. In slot filling, the model was trained to extract key slots of information from the user's utterance, such as *Date*, *Time*, *Hospital*, *Name*, *Species*, and *Age*. This ensured the model could accurately identify the user's intent and extract relevant details from their utterance, making it highly effective for practical applications in the pet care domain.

Overall, this two-stage P-tuning process significantly improved the model's performance by combining the benefits of domain-specific and task-specific adaptations, leading to an accurate and reliable system for understanding and responding to user queries in the context of pet care services.

3) MULTI-TASK ADAPTERFUSION

We implemented a multi-task framework using AdapterFusion [10], [11] to enhance conversational AI systems, particularly in the pet care domain. Initially, we trained separate adapters for intent classification and slot filling tasks. After training these distinct adapters, we combined them using the AdapterFusion method, allowing the model to leverage knowledge from both tasks. This fusion enhances the model's efficiency and effectiveness in understanding and responding to user queries, making it a robust solution for multi-task learning in conversational services.

B. METHODOLOGY FOR SEMANTIC MATCHING

The methodology for semantic matching involves using a pre-trained language model enhanced with already trained task-specific adapters to understand and compare sentences. To effectively conduct semantic matching task, the model first generate embeddings for extracted symptom description from user utterance. By calculating the cosine similarity between these embeddings, the model identifies the most similar symptoms for the given input sentences. The performance was measured using top-k accuracy, which checks how often the correct symptoms appear in the top predictions. This approach effectively matched the semantic meanings by leveraging advanced language models and fine-tuning techniques.

C. SETUP

1) IMPLEMENTATION

We implemented the models and baselines using PyTorch [38] and HuggingFace [5]. The models were trained on a single machine equipped with two Intel Xeon 16-core processors, 512 GB of RAM, and three NVIDIA A100 processors with 80 GB of RAM. We conducted a series of experiments with various hyperparameters of the models, focusing on enhancing their accuracy and F1 scores. The hyperparameters were carefully selected based on experiments with the following ranges: batch size [8, 16, 32, 64], learning rate [1e-3, 1e-4, 1e-5, 1e-6], epochs [20, 30, 60, 80, 100], intent loss alpha and slot loss alpha ranging from 0.1 to 1.0, and reduction factor [4, 8, 16, 32, 64]. After evaluating these ranges, we selected hyperparameters that produce best results, as listed in Table 7. Specifically, we found that a reduction factor of 4 for adapter tuning and a pre-sequence length of 50 for P-tuning

[5] https://huggingface.co/

TABLE 7. Hyperparameters.

PEFT	Task	Batch size	Epochs	Learning Rate	Intent Loss Alpha	Slot Loss Alpha
Adapter Tuning	multi-task adapter	32	20	1e-4	0.5	0.5
	intent adapter	16	20	1e-5	1.0	0.0
	slot adapter	32	100	1e-4	0.0	1.0
P-tuning v2	multi-task	16	100	1e-3	0.5	0.5
AdapterFusion	multi-task	32	20	1e-4	0.5	0.5
Domain Adaptation	all tasks	16	20	1e-4	0.5	0.5

v2 provided the optimal balance between model performance and computational efficiency.

2) MODELS

We investigate various BERT and RoBERTa-based models to evaluate their performance on the PETIS dataset:

- **KLUE-BERT-base**[6] is a pre-trained BERT model for a Korean dataset. It is part of the KLUE (Korean Language Understanding Evaluation) benchmark [39].
- **KLUE-RoBERTa-large**[7] and **KLUE-RoBERTa-base**[8] were pre-trained RoBERTa models in Korean. These are also part of the KLUE benchmark.
- **KoSimCSE-BERT**[9] and **KoSimCSE-BERT-multitask**[10] are pre-trained SimCSE-BERT models, designed to enhance sentence embedding and semantic understanding tasks in Korean.
- **KoSimCSE-RoBERTa**[11] and **KoSimCSE-RoBERTa-multitask**[12] are pre-trained SimCSE-RoBERTa models designed for better understanding of sentence embedding tasks in Korean.

3) EVALUATION METRICS

We evaluated the model using specific metrics for different tasks:

- **Intent classification:** To accurately identify the intent of user utterances, accuracy was used.
- **Slot filling:** We utilized Macro F1^E and Micro F1^E for entity level, and Macro F1^C and Micro F1^C for character level, to assess the model's ability to accurately extract relevant slot from user utterances.
- **Symptom Matching:** We used the top-k accuracy to assess how well the model ranked and matched symptoms in pet care services.

These metrics provided a comprehensive assessment of model performance across different aspects of the PETIS dataset, ensuring that our experiments met the desired standards of accuracy and reliability. To further facilitate reproducibility, we will make our dataset and code publicly available.

VII. EXPERIMENTAL RESULTS

We conducted experiments with a multi-task adapter, a single-task adapter, P-tuning v2, AdapterFusion, and domain adaptation using the PETIS dataset to demonstrate its effectiveness in significantly improving intent classification and slot filling tasks within the pet care domain. Additionally, we conducted two ablation studies: one focusing on semantic matching to analyze the frequency of correct symptoms in predictions, and the other evaluating the performance of large language models.

A. MAIN RESULTS

1) MULTI-TASK ADAPTER

In our study, we employed the multi-task adapter with various Korean RoBERTa and BERT base models to demonstrate the effectiveness of our proposed dataset in training and evaluating NLU models tailored for the pet care domain. As shown in Table 8:

- **KLUE-RoBERTa-large:** Achieved outstanding results, with a top accuracy of **93.23** in intent classification. The model also performed exceptionally well in slot filling, achieving high scores in all metrics: Macro F1^E (73.96), Micro F1^E (74.74), Macro F1^C (83.38), and Micro F1^C (91.32).
- **KLUE-RoBERTa-base:** Had the second-best performance in intent classification, with an accuracy of 92.85 and good slot filling results.
- **KLUE-BERT-base:** Demonstrated poorer efficacy than KLUE-RoBERTa-base, with lower scores in both intent classification and slot filling.
- **KoSimCSE Models:** Exhibited diverse performance, achieving intent classification accuracy between 90.12 and 91.91. The slot filling results were quite satisfactory, yet they were generally average.

The results highlight the exceptional capability of KLUE-RoBERTa-large in effectively managing both intent classification and slot filling tasks, positioning it as an excellent choice for applications in the pet care domain.

[6] https://huggingface.co/klue/bert-base
[7] https://huggingface.co/klue/roberta-large
[8] https://huggingface.co/klue/roberta-base
[9] https://huggingface.co/BM-K/KoSimCSE-bert
[10] https://huggingface.co/BM-K/KoSimCSE-bert-multitask
[11] https://huggingface.co/BM-K/KoSimCSE-roberta
[12] https://huggingface.co/BM-K/KoSimCSE-roberta-multitask

TABLE 8. Multi-task adapter: Evaluation results of intent classification and slot filling. Underline shows the second-best performance.

		Intent	Slot			
Model	Model Size	Accuracy	Macro F1^E	Micro F1^E	Macro F1^C	Micro F1^C
KLUE-RoBERTa-base	110 M	92.85	72.90	73.37	82.43	90.67
KLUE-BERT-base	110 M	90.97	72.21	71.97	81.11	90.23
KoSimCSE-RoBERTa	111 M	91.91	71.16	72.93	81.07	90.65
KoSimCSE-BERT	111 M	90.12	67.95	71.40	79.64	90.22
KoSimCSE-RoBERTa-multitask	111 M	91.82	70.57	72.79	80.99	90.60
KoSimCSE-BERT-multitask	111 M	90.97	70.30	71.55	80.49	90.21
KLUE-RoBERTa-large	340 M	**93.23**	**73.96**	**74.74**	**83.38**	**91.32**

TABLE 9. Slot adapter: Evaluation results of slot filling. Underline shows the second-best performance.

		Slot			
Model	Model Size	Macro F1^E	Micro F1^E	Macro F1^C	Micro F1^C
KLUE-RoBERTa-base	110 M	**76.83**	**75.91**	84.76	**91.94**
KLUE-BERT-base	110 M	73.92	73.37	82.79	90.78
KoSimCSE-RoBERTa	111 M	74.03	74.96	82.48	90.59
KoSimCSE-BERT	111 M	73.97	73.99	82.82	90.48
KoSimCSE-RoBERTa-multitask	111 M	72.94	74.27	81.82	90.83
KoSimCSE-BERT-multitask	111 M	72.64	72.88	81.91	90.54
KLUE-RoBERTa-large	340 M	76.79	74.68	**84.99**	91.38

2) SINGLE-TASK ADAPTER

Table 9 and 10 summarize the performance of single-task adapters for slot filling and intent classification, respectively. These results demonstrate the effectiveness of the PETIS dataset, showcasing its robustness in training state-of-the-art models:

- **KLUE-RoBERTa-base:** Exhibited robust performance in both tasks. Slot filling adapter demonstrated exceptional performance, achieving the best scores in Macro F1^E (76.83), Micro F1^E (75.91), and Micro F1^C (91.94). The model also achieved accuracy of 91.25 in intent classification adapter, outperforming other base models and demonstrating its strong capability in detecting intents.
- **KLUE-RoBERTa-large:** Achieved a 92.10 accuracy in outperforming the intent classification task, showcasing the model's proficiency in understanding user intent.
- **KoSimCSE Models:** Demonstrated impressive outcomes in both tasks, achieving intent classification accuracies ranging from 88.99 to 90.12 and competitive slot filling results.

Analysis: When comparing the multi-task adapter (Table 8) with the single-task adapters (Table 9 and Table 10), the multi-task adapter approach generally demonstrated higher intent classification accuracy but slightly lower slot filling performance than slot adapter, indicating a trade-off between multi-task learning and task-specific optimization. Interestingly, the KLUE-RoBERTa-base model outperformed the larger KLUE-RoBERTa-large model in slot filling. We believe that the reason the base model performed better than the large model in slot filling could be its ability to generalize better on smaller datasets and avoid overfitting, a common issue with large models. Overall, the PETIS dataset demonstrated remarkable efficacy in training models for both tasks, highlighting its significance as a robust resource for advancing research in the pet care field.

TABLE 10. Intent adapter: Evaluation results of intent classification. Underline show the second-best performance.

			Intent
Model	Model Size	Eval$_{loss}$	Accuracy
KLUE-RoBERTa-base	110 M	0.3063	91.25
KLUE-BERT-base	110 M	0.3064	89.37
KoSimCSE-RoBERTa	111 M	0.3086	90.03
KoSimCSE-BERT	111 M	**0.2874**	90.12
KoSimCSE-RoBERTa-multitask	111 M	0.3197	89.09
KoSimCSE-BERT-multitask	111 M	0.2912	88.99
KLUE-RoBERTa-large	340 M	0.2905	**92.10**

3) P-TUNING V2

In the evaluation results for P-tuning v2, presented in Table 11:

- **KLUE-RoBERTa-large:** Stands out with a remarkable accuracy of 92.76 in intent classification. For slot filling, the model demonstrated impressive performance with

TABLE 11. P-tuning v2: Evaluation results of intent classification and slot filling. Underline show the second-best performance.

		Intent	Slot			
Model	**Model Size**	**Accuracy**	**Macro F1^E**	**Micro F1^E**	**Macro F1^C**	**Micro F1^C**
KLUE-RoBERTa-base	110 M	92.10	68.33	68.88	80.57	89.67
KLUE-BERT-base	110 M	90.31	67.80	66.96	81.76	89.32
KoSimCSE-RoBERTa	111 M	90.78	68.37	**69.59**	80.58	89.93
KoSimCSE-BERT	111 M	89.93	68.94	65.88	80.78	89.28
KoSimCSE-RoBERTa-multitask	111 M	90.68	67.94	69.27	80.44	89.90
KoSimCSE-BERT-multitask	111 M	88.80	66.22	64.80	79.77	88.44
KLUE-RoBERTa-large	340 M	**92.76**	**70.07**	68.13	**82.66**	**90.31**

TABLE 12. AdapterFusion: Evaluation results of intent classification and slot filling. Underline shows the second-best performance.

		Intent	Slot			
Model	**Data Size**	**Accuracy**	**Macro F1^E**	**Micro F1^E**	**Macro F1^C**	**Micro F1^C**
KLUE-RoBERTa-large	8508 (1.0)	**93.32**	73.39	**73.59**	**83.78**	**91.21**
	5956 (0.7)	91.13	**76.24**	72.27	83.48	91.14
	4254 (0.5)	91.15	71.52	72.17	82.28	90.66
	2552 (0.3)	90.57	70.49	69.87	80.46	89.63
	851 (0.1)	84.90	51.55	59.31	71.67	88.26

Macro F1^E of 70.07, Micro F1^E of 68.13, Macro F1^C of 82.66, and Micro F1^C of 90.31 at both the entity and character level, showing its robustness across both tasks.

- **KLUE-RoBERTa-base:** Also performed well, with an accuracy of 92.10 in intent classification and achieved good slot filling results.
- **KLUE-BERT-base:** Achieved an accuracy of 90.31, with a notable second-best slot filling result in Macro F1^C of 81.76.
- **KoSimCSE Models:** Showed varied performance in intent classification, with accuracies ranging from 88.80 to 90.78. In slot filling, KoSimCSE-RoBERTa achieved the highest Micro F1^E score of 69.59, and KoSimCSE-BERT secured the second-highest Macro F1^E score of 68.94.

Analysis: Comparing the results from the multi-task adapter, Single-task Adapter, and P-tuning v2 approaches, it is evident that while the multi-task adapter generally achieved higher intent classification accuracy, the slot adapter excelled in slot filling performance, and P-tuning v2 showed robust performance across both tasks.

4) ADAPTERFUSION

The evaluation results for AdapterFusion highlight the remarkable accuracy of KLUE-RoBERTa-large in intent classification and its impressive F1 score in slot filling. As shown in Table 12:

- **Intent classification:**
  - Achieved an outstanding **93.32** accuracy using the full dataset.
  - The accuracy slightly decreased with reduced data sizes:
    - 91.13 accuracy with 0.7 of the data.
    - 91.15 accuracy with 0.5 of the data.
    - Significant performance drops with smaller datasets (0.3 and 0.1), highlighting the importance of large data for maintaining high accuracy.
- **Slot filling:**
  - Micro F1^E: Highest score of 73.59 with the full dataset.
  - Macro F1^C: Achieved 83.72.
  - Micro F1^C: Highest score of 91.21.
  - The model maintained competitive slot filling performance even with reduced data sizes, though performance naturally declined with smaller datasets.

Analysis: Multi-task adapter (Table 8), single-task adapter (Table 9 and 10), and P-tuning v2 (Table 11) demonstrate varying performance in intent classification and slot filling. However, AdapterFusion achieves the highest overall accuracy in intent classification, and the single-task adapter for slot filling shows superior performance. These results underscore the effectiveness of the PETIS dataset in providing a challenging and valuable benchmark for evaluating and enhancing NLU models in the pet care domain.

5) DOMAIN ADAPTATION

A comprehensive comparison of models with and without domain adaptation across various tasks, with a focus on intent classification and slot filling tasks presented in Table 13:

- **Intent classification:** Domain adaptation resulted in a slight improvement in accuracy. For example, the multi-task adapter showed a marginal increase from 92.57 to 92.66, and the intent adapter's accuracy increased from

TABLE 13. **Domain Adaptation: Evaluation results of intent classification and slot filling.**

		Intent	Slot			
Task	Eval$_{loss}$	Accuracy	Macro F1^E	Micro F1^E	Macro F1^C	Micro F1^C
Multi-task adapter w/o domain adaptation	0.3369	92.57	**72.82**	74.15	**83.72**	91.35
Multi-task adapter w domain adaptation	**0.2749**	**92.66**	71.64	**74.43**	82.72	**91.66**
Intent adapter w/o domain adaptation	**0.4004**	91.82	-	-	-	-
Intent adapter w domain adaptation	0.4378	**92.59**	-	-	-	-
Slot adapter w/o domain adaptation	0.4000	-	74.59	74.87	84.07	91.07
Slot adapter w domain adaptation	0**.2869**	-	**74.67**	**75.60**	**84.28**	**91.70**
P-tuning v2 w/o domain adaptation	**0.3513**	**87.96**	56.47	57.11	71.92	87.06
P-tuning v2 w domain adaptation	0.3539	87.11	**59.01**	**58.67**	**74.59**	**87.32**
AdapterFusion w/o domain adaptation	0.3454	92.29	**73.75**	**75.22**	**84.20**	91.53
AdapterFusion w domain adaptation	**0.2749**	**92.66**	71.64	74.43	82.72	**91.66**

91.82 to 92.59. However, AdapterFusion also demonstrated the effectiveness of domain adaptation, increasing accuracy from 92.29 to 92.66.

- **Slot filling:** Similarly, domain adaptation led to improved performance in slot filling tasks. The multi-task adapter saw a notable increase in Micro F1^C from 91.35 to 91.66. Likewise, slot adapter performance also improved, with Micro F1^C increasing from 91.07 to 91.70. Additionally, AdapterFusion benefited notably from domain adaptation, with Micro F1^C increasing from 91.53 to 91.66.

Although the enhancements were modest, they underscored the importance of domain adaptation in improving the effectiveness of pre-trained models for specific tasks.

B. AUXILIARY TASK

1) SEMANTIC MATCHING

The performance of various models on the semantic matching task is illustrated in Figure 2:

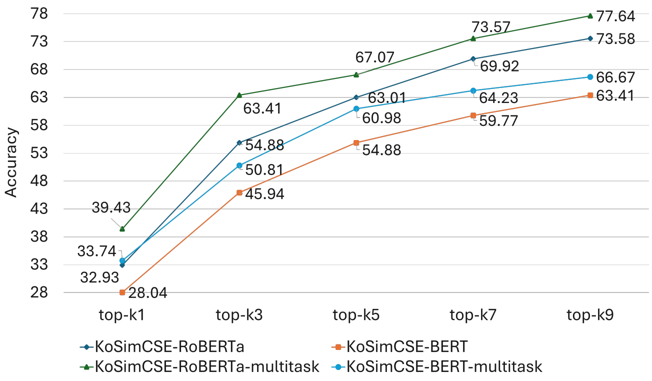


FIGURE 2. **Evaluation results for Semantic Matching with top-k accuracy.**

- **KoSimCSE-RoBERTa-multitask:** Outperformed all other models across all top-k accuracy levels, achieving the highest accuracy at each top-k level and peaking at 77.64 accuracy at top-k9.
- **KoSimCSE-RoBERTa:** Exhibited strong performance, particularly in higher top-k levels, reaching an accuracy of 73.58 at top-k9.
- **KoSimCSE-BERT and KoSimCSE-BERT-multitask:** Showed varied performance, generally trailing behind the RoBERTa-based models. However, KoSimCSE-BERT-multitask reached a top accuracy of 66.67 at top-k9, demonstrating reasonable effectiveness.

Analysis: The results demonstrate that the KoSimCSE-RoBERTa-multitask model excels at accurately capturing the semantic meaning of symptoms in the pet care domain.

2) LARGE LANGUAGE MODELS

We conducted a comprehensive analysis of our dataset using the state-of-the-art models Polyglot-Ko-1.3B [40] and Ko-GPT-Trinity 1.2B (v0.5)[13] to demonstrate the performance of the large language models with 10 epochs, the learning rate to 1e-4, and the batch size at [32, 64]. The evaluation results are summarized in Table 14:

- **Performance with larger batch size:** Both models exhibited improved performance when the batch size was increased from 32 to 64.
  - Polyglot-Ko-1.3B achieved the highest intent classification accuracy of 59.74 and a slot F1 score of 90.79 with a batch size of 64, outperforming its performance at a batch size of 32.
  - Ko-GPT-Trinity 1.2B (v0.5) also improved with a larger batch size, reaching an intent accuracy of 58.14 and a slot F1 score of 85.56.
- **Decoder model challenges:** Despite their strong performance in slot filling, both models struggled with intent classification. This difficulty is likely due to their decoder-based architecture, which is optimized for token-by-token generation rather than for understanding the global context of the sentence, which is crucial for intent classification.

[13] https://huggingface.co/skt/ko-gpt-trinity-1.2B-v0.5

TABLE 14. Large Language Models: Evaluation results of intent classification and slot filling.

Model	Batch Size	Intent Acc	Slot F1
Polyglot-Ko-1.3B	32	58.70	90.69
	64	**59.74**	**90.79**
Ko-GPT-Trinity 1.2B (v0.5)	32	57.67	85.07
	64	**58.14**	**85.56**

VIII. DISCUSSION

The PETIS dataset represents a major breakthrough in advancing conversational AI systems tailored for pet care services. Designed specifically to address the challenges of understanding the descriptive and often natural language used by pet owners, PETIS fills a critical gap left by traditional NLU datasets, which often fail to capture the nuances of domain-specific terminology. With its 10 intent classes and 11 slot classes, PETIS provides a focused and effective resource that enhances the accuracy of intent classification and slot filling. The effectiveness of this dataset is clearly demonstrated by the strong performance of state-of-the-art models applied to it.

Our experiments highlighted the effectiveness of different approaches, particularly multi-task AdapterFusion, which achieved the highest accuracy in intent classification with 93.32 and strong performance in slot filling with Micro F1^C score of 91.21. The multi-task adapter approach also delivered robust results, with the KLUE-RoBERTa-large model showing a balanced performance across both tasks. This suggests that multi-task learning, combined with parameter-efficient techniques, can be highly effective in domains where computational resources are limited. Furthermore, the single-task adapter's success in slot filling emphasizes the importance of task-specific optimization in extracting relevant information from user utterances.

However, the study also revealed challenges with large language models (LLMs) like Polyglot-Ko-1.3B and Ko-GPT-Trinity 1.2B (v0.5), which, despite their effectiveness in slot filling, struggled with intent classification. This difficulty is likely due to their decoder-based architecture, which may not be well-suited for tasks that require a deep understanding of the overall context of user utterances. These findings suggest that while LLMs hold promise, their application in intent classification requires further improvement to better capture the nuances of user intent in conversational AI systems.

IX. CONCLUSION

This study introduced the PETIS dataset, specifically designed to advance conversational AI in the pet care domain. Unlike previous datasets that often fell short in handling domain-specific language and the descriptive nature of user utterances, PETIS offers a robust resource with 10,636 annotated utterances. This dataset effectively addresses the challenges inherent in natural language understanding (NLU) tasks within pet care services, providing a valuable benchmark for evaluating NLU models in this specialized field. Our experiments demonstrated that various state-of-the-art models, such as AdapterFusion, multi-task adapters, and P-tuning v2, can effectively utilize the PETIS dataset to perform intent classification and slot filling tasks. AdapterFusion stood out for its balanced performance across tasks, while single-task adapters excelled in slot filling, emphasizing the importance of task-specific optimization. The study also revealed that while large language models are promising for certain tasks, they may struggle with nuanced intent classification, highlighting the need for models that can better capture the context in specialized domains.

In future work, we plan to expand the PETIS dataset to include multiple languages and cultural contexts, broadening its applicability on a global scale. We also aim to extend the dataset relevance to other healthcare domains, thereby contributing to the development of more versatile and sophisticated conversational AI systems. Additionally, we plan to explore large language models more extensively, focusing on few-shot and prompting techniques to further enhance their performance in intent classification and slot filling. By delving into these advanced methods, we hope to improve the ability of large language models to understand and process nuanced intent in specialized domains.

X. LIMITATIONS

Despite the successes of this study, several limitations were identified. The PETIS dataset is currently focused exclusively on the Korean language, limiting its applicability in multilingual and culturally diverse contexts. This focus, while beneficial for Korean-language applications, restricts the dataset's global relevance. Expanding the dataset to include other languages and cultural nuances would significantly enhance its applicability and usefulness in international contexts. Additionally, the dataset's narrow focus on the pet care domain, though valuable for this specific field, may limit its utility in broader healthcare domains where different intents and slot categories may be required.

Furthermore, while domain adaptation techniques were applied and showed modest improvements, these enhancements were not substantial, indicating the need for more advanced approaches to fully leverage domain-specific knowledge. The performance of large language models (LLMs) in intent classification was another limitation, as these models struggled with understanding the overall context of user utterances, particularly due to their decoder-based architecture. Addressing these limitations will be essential to fully realize the potential of PETIS and similar datasets across various domains and applications.

ACKNOWLEDGEMENTS

We acknowledge the pet owners and community members who contributed valuable experiences and insights to the creation of the PETIS dataset. We also thank our dedicated annotators and domain experts for their diligent work in ensuring dataset quality. This work was supported by the Basic

Research Program through a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1C1C1010317) and the Research Fund, 2023, of The Catholic University of Korea.

REFERENCES

[1] M. Jovanović, M. Baez, and F. Casati, "Chatbots as conversational healthcare services," *IEEE Internet Computing*, vol. 25, no. 3, pp. 44–51, 2021.

[2] A. Babu and S. B. Boddu, "Bert-based medical chatbot: Enhancing healthcare communication through natural language understanding," *Exploratory Research in Clinical and Social Pharmacy*, vol. 13, p. 100419, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667276624000143

[3] L. Qin, X. Xu, W. Che, and T. Liu, "AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1807–1816. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.163

[4] R. Yan, S. Peng, H. Mi, L. Jiang, S. Yang, Y. Zhang, J. Li, L. Peng, Y. Wang, and Z. Wen, "Towards generalized models for task-oriented dialogue modeling on spoken conversations," *CoRR*, vol. abs/2203.04045, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2203.04045

[5] C.-S. Wu, S. C. Hoi, R. Socher, and C. Xiong, "TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 917–929. [Online]. Available: https://aclanthology.org/2020.emnlp-main.66

[6] W. Wang, Z. Zhang, J. Guo, Y. Dai, B. Chen, and W. Luo, "Task-oriented dialogue system as natural language generation," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '22. ACM, Jul. 2022. [Online]. Available: http://dx.doi.org/10.1145/3477495.3531920

[7] A.-H. Al-Ajmi and N. Al-Twairesh, "Building an arabic flight booking dialogue system using a hybrid rule-based and data driven approach," *IEEE Access*, vol. 9, pp. 7043–7053, 2021.

[8] L. Fernando and G. Ganegoda, "Resbot: A bilingual restaurant booking conversational artificial intelligence," in *2023 8th International Conference on Information Technology Research (ICITR)*, 2023, pp. 1–6.

[9] Y. Jin, W. Cai, L. Chen, N. N. Htun, and K. Verbert, "Musicbot: Evaluating critiquing-based music recommenders with conversational interaction," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 951–960. [Online]. Available: https://doi.org/10.1145/3357384.3357923

[10] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych, "AdapterFusion: Non-destructive task composition for transfer learning," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 487–503. [Online]. Available: https://aclanthology.org/2021.eacl-main.39

[11] V. Suresh, S. Aït-Mokhtar, C. Brun, and I. Calapodescu, "An adapter-based unified model for multiple spoken language processing tasks," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 676–10 680.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[14] Q. Chen, Z. Zhuo, and W. Wang, "Bert for joint intent classification and slot filling," 2019.

[15] S. Su Su Yee and K. Soe, "Exploring bert-based encoders for sequence classification and multi-task learning in dialogue acts and joint intent-slot filling," *Indian Journal of Computer Science and Engineering*, 06 2024.

[16] S. Gore, D. D. Jadhav, M. E. Ingale, S. Gore, and U. Nanavare, "Leveraging bert for next-generation spoken language understanding with joint intent classification and slot filling," in *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)*, 2023, pp. 1–5.

[17] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6894–6910. [Online]. Available: https://aclanthology.org/2021.emnlp-main.552

[18] Y. Chen and Z. Luo, "Pre-trained joint model for intent classification and slot filling with semantic feature fusion," *Sensors*, vol. 23, no. 5, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/5/2848

[19] W. U. Ahmad, J. Chi, T. Le, T. Norton, Y. Tian, and K.-W. Chang, "Intent classification and slot filling for privacy policies," 2021.

[20] M. Firdaus, A. Ekbal, and E. Cambria, "Multitask learning for multilingual intent detection and slot filling in dialogue systems," *Information Fusion*, vol. 91, pp. 299–315, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253522001671

[21] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The ATIS spoken language systems pilot corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*, 1990. [Online]. Available: https://aclanthology.org/H90-1021

[22] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril, M. Primet, and J. Dureau, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," 2018.

[23] W. Xu, B. Haider, and S. Mansour, "End-to-end slot alignment and recognition for cross-lingual NLU," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 5052–5063. [Online]. Available: https://aclanthology.org/2020.emnlp-main.410

[24] X. Chen, A. Ghoshal, Y. Mehdad, L. Zettlemoyer, and S. Gupta, "Low-resource domain adaptation for compositional task-oriented semantic parsing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 5090–5100. [Online]. Available: https://aclanthology.org/2020.emnlp-main.413

[25] H. Li, A. Arora, S. Chen, A. Gupta, S. Gupta, and Y. Mehdad, "MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Online: Association for Computational Linguistics, Apr. 2021, pp. 2950–2962. [Online]. Available: https://aclanthology.org/2021.eacl-main.257

[26] X. Hao, L. Wang, H. Zhu, and X. Guo, "Joint agricultural intent detection and slot filling based on enhanced heterogeneous attention mechanism," *Computers and Electronics in Agriculture*, vol. 207, p. 107756, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169923001448

[27] A. Gupta, P. Immadisetty, P. Rajesh, and S. G, "Joint intent classification and slot tagging on agricultural dataset for indic languages," in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, 2023, pp. 288–293.

[28] H. Zhang, H. Xu, X. Wang, Q. Zhou, S. Zhao, and J. Teng, "Mintrec: A new dataset for multimodal intent recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. ACM, Oct. 2022. [Online]. Available: http://dx.doi.org/10.1145/3503161.3547906

[29] F. A. Sakib, A. H. M. R. Karim, S. H. Khan, and M. M. Rahman, "Intent detection and slot filling for home assistants: Dataset and analysis for Bangla and Sylheti," in *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, F. Alam, S. Kar, S. A. Chowdhury, F. Sadeque, and R. Amin, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 48–55. [Online]. Available: https://aclanthology.org/2023.banglalp-1.6

[30] M. Jarrar, A. Birim, M. Khalilia, M. Erden, and S. Ghanem, "Arbanking77: Intent detection neural model and a new dataset in modern and dialectical arabic," 2023.

[31] M. H. Dao, T. H. Truong, and D. Q. Nguyen, "Intent detection and slot filling for vietnamese," 2021.

[32] K. Krippendorff, "Computing krippendorff's alpha-reliability," in *Annenberg School for Communication*, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:59901023

[33] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[34] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang, "Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment," 2023.

[35] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 4582–4597. [Online]. Available: https://aclanthology.org/2021.acl-long.353

[36] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, and J. Tang, "P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 61–68. [Online]. Available: https://aclanthology.org/2022.acl-short.8

[37] 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.

[38] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang,

Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS 2017 Workshop on Autodiff*, 2017. [Online]. Available: https://openreview.net/forum?id=BJJsrmfCZ

[39] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh *et al.*, "Klue: Korean language understanding evaluation," *arXiv preprint arXiv:2105.09680*, 2021.

[40] H. Ko, K. Yang, M. Ryu, T. Choi, S. Yang, jiwung Hyun, and S. Park, "A technical report for polyglot-ko: Open-source large-scale korean language models," 2023.

NAMRAH ZAMAN received the B.S. degree in Electrical Engineering in the department of Electrical and Computer engineering from International Islamic University, Islamabad, Pakistan, in 2023. She is currently pursuing her M.S. degree in Artificial Intelligence from The Catholic University of Korea, Bucheon, South Korea. Her research interests include natural language processing, parameter-efficient fine-tuning and deep learning for healthcare.

SEONG-JIN PARK is currently pursuing a B.S. degree in Mathematics and Computer Science Information Engineering and a B.S./M.S. integrated degree in Artificial Intelligence from The Catholic University of Korea, Bucheon, South Korea. His research interests include large language models, vision-language models, and autonomous agents.

HYUN-SIK WON received the B.S. degree in the department of psychology and the school of computer science and information engineering from The Catholic University of Korea, Bucheon, South Korea, in 2022, where he is currently pursuing the M.S. degree in Artificial Intelligence. His research interests include natural language processing, parameter-efficient fine-tuning, and prompting.

MIN-JI KIM received the B.S. degree in the school of computer science and information engineering from The Catholic University of Korea, Bucheon, South Korea, in 2023, where she is currently pursuing the M.S. degree in Artificial Intelligence. Her research interests include natural language processing, deep learning, parameter-efficient fine-tuning, and conversational artificial intelligence.

HEE-SU AN is currently pursuing a B.S. degree in Artificial Intelligence from The Catholic University of Korea, Bucheon, South Korea. Her research interests include natural language processing, biomedical research, and computer vision.

KANG-MIN KIM received the B.S. degree in Computer Engineering from Kyung Hee University, Yongin, South Korea, in 2016, and a Ph.D. degree in Computer Science and Engineering from Korea University, Seoul, South Korea, in 2021. Since 2021, he has been an Assistant Professor with the Department of Data Science and the Department of Artificial Intelligence, The Catholic University of Korea, Bucheon, South Korea. His research interests include natural language processing, large-scale text classification, commonsense reasoning, self-supervised learning, weakly supervised learning, and intelligent systems.

• • •