JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 12, pp. 61-72, December 2025
https://doi.org/10.9708/jksci.2025.30.12.061

A Study on the Application and Quality Comparison of LLM-Based Task-Oriented Dialogue Data Generation Methodology in the Korean Cafe Domain

Changgou Kang*, Namgyu Kim**

*Graduate Student, Graduate School of Business IT, Kookmin University, Seoul, Korea
**Professor, Graduate School of Business IT, Kookmin University, Seoul, Korea

[Abstract]

HR-MultiWOZ methodology has recently emerged as an efficient approach for generating Task-Oriented Dialogue (TOD) data with Large Language Models (LLMs). It introduced the first synthetic TOD dataset in the Human Resources (HR) domain, highlighting cost-effectiveness through schema-based design and minimal human input. However, the influence of prompt design on data quality and its applicability to non-English languages or domains beyond HR remain underexplored. Accordingly, this study is conducted as a case study to empirically examine the applicability of the HR-MultiWOZ methodology to Korean and café ordering. Results show that simple prompt adjustments can effectively control the characteristics of LLM-generated dialogue (LGD), underscoring the methodology's scalability and the pivotal role of prompt design in shaping dialogue data quality.

▸ **Key words**: Task-Oriented Dialogue, Synthetic Data Generation, LLM, HR-MultiWOZ, Prompt

[요 약]

HR-MultiWOZ 방법론은 최근 대규모 언어 모델(LLM)을 활용한 작업 지향 대화(Task-Oriented Dialogue, TOD) 데이터 생성의 효율적인 접근법으로 주목받고 있다. 이 방법론은 인사(HR) 도메인에서 최초의 합성 TOD 데이터셋을 제안하였으며, 스키마 기반 설계와 최소한의 인적 개입을 통해 비용 효율성을 주장하였다. 그러나 해당 연구는 프롬프트 설계가 데이터 품질에 미치는 영향과 비영어권 언어 및 HR 이외 도메인에서의 적용 가능성을 충분히 다루지는 않았다는 한계를 갖는다. 이에 따라 본 연구는 HR-MultiWOZ 방법론의 적용 가능성을 한국어와 카페 도메인 중심으로 확인하는 사례 연구 형태로 진행된다. 본 연구에서는 실험을 통해 간단한 프롬프트 조정만으로도 LLM이 생성하는 대화의 특성을 효과적으로 제어할 수 있음을 확인하였으며, 결과적으로 HR-MultiWOZ 방법론의 확장성과 대화 데이터 품질 형성에 있어 프롬프트 설계가 핵심적 역할을 수행할 수 있음을 입증하였다.

▸ **주제어**: 작업지향대화, 가상데이터생성, 초거대언어모델, HR-MultiWOZ, 프롬프트

• First Author: Changgou Kang, Corresponding Author: Namgyu Kim
*Changgou Kang (changgou.kang@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
**Namgyu Kim (ngkim@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
• Received: 2025. 10. 17, Revised: 2025. 11. 22, Accepted: 2025. 12. 08.

Copyright © 2025 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945
www.kci.go.kr

I. Introduction

인공지능 대화 시스템은 최근 급격히 발전하고 있으며, 다양한 산업에 폭넓게 적용이 가능할 것으로 예측된다. Allied Market Research[1]는 2020년 57.8억 달러로 평가된 글로벌 대화형 인공지능 시장 규모가 연평균 성장률(CAGR) 20.0%를 기록하며 2030년에는 326.2억 달러에 도달할 것으로 전망했다. 그리고, 해당 보고서는 BSFI(금융), 소매 및 이커머스, 헬스케어 및 바이오, 여행 및 호텔, 미디어 및 엔터테인먼트, 그리고 통신 등의 다양한 산업 분야에 대화형 인공지능이 도입될 것으로 예측했다.

이러한 인공지능 대화 시스템은 일반적으로 개방형 도메인 대화(Open Domain Dialogue, ODD) 시스템과 작업 지향 대화(Task Oriented Dialogue, TOD) 시스템의 두 가지 범주로 구분된다[2]. ODD 시스템이 사용자와의 상호작용 자체를 목표로 하는 반면, TOD 시스템은 특정 도메인에서 사용자가 특정 작업을 완수하도록 돕는다. 예를 들어, 제품 검색, 영화 예약, 그리고 버스 일정 확인 등의 작업에 TOD 시스템이 활용되어 실생활에 직접적인 도움을 줄 수 있다.

하지만, TOD 시스템이 다양한 산업에 적극적으로 도입되기에는 두 가지 주요 한계가 있다. 첫 번째는 TOD 시스템이 가지는 도메인의 특수성이다. 최근 OpenAI의 GPT-3[3], Google의 PaLM[4], 그리고 Meta의 LLaMA[5] 등과 같은 거대 언어 모델(LLM)이 텍스트 이해, 텍스트 생성, 그리고 감정 분석 등과 같은 범용적인 자연어 처리 과제에서 우수성을 입증했다.

하지만, LLM만으로 효과적인 TOD 시스템을 구축하기는 충분하지 않으며[6], 특히 특정 도메인에 대한 훈련 데이터 세트의 부족은 TOD 시스템의 구현을 어렵게 한다. 두 번째 한계는 TOD 데이터 구축에 필요한 비용 및 시간의 측면에서 찾을 수 있다. 대화형 데이터 세트를 구축하는 주요 방법론은 클라우드 소싱(Cloud Sourcing)으로, 인간 작업자가 제공된 지침에 따라 데이터를 직접 생성한다. 하지만, 이 접근방식은 비용과 시간이 많이 소요되어 새로운 도메인으로 확장 적용하기가 어렵다는 한계를 갖는다[7].

이러한 한계를 극복하기 위해, 최근 합성(Synthetic) 대화 데이터를 생성하는 방법론이 클라우드 소싱의 대안으로 주목받고 있다. 이 방법론은 텍스트 데이터(예: 문서, 표, 지식 그래프)를 대화 형식으로 변환하거나, 기존의 대화 데이터를 새로운 사례로 보강하는 것을 포함한다. 이러한 데이터 생성(Generation) 방식은 새로운 데이터를 만들어 데이터 세트를 확장하거나 향상하는 과정이며, 기존 데이터에 변형을 적용하는 데이터 증강(Augmentation)과는 다르다. 데이터 생성은 기존 데이터 세트에는 없지만 관련성이 높은 새로운 데이터 포인트를 만들어, 다양하고 강력한 학습 샘플을 제공한다[7]. 최근에는 LLM에 대한 미세 조정(Fine-tuning) 없이, In-context learning 기법을 활용해 비용 효율적으로 합성 대화 데이터를 생성하는 연구가 활발히 이루어지고 있다[8-12].

특히 최근에는 기존 데이터 없이도 비용 효율적으로 HR 도메인에 특화된 TOD 데이터 세트를 구축할 수 있는 HR-MultiWOZ[13]가 공개되어 주목받고 있다. 해당 연구의 방법론은 크게 세 가지 프로세스로 구성된다. 첫째, 스키마 기반 접근법(Schema Approach)[14]에 따라 의도(Intent)와 슬롯(Slot)을 자연어로 설명하는 스키마를 설계한다. 둘째, LLM을 활용하여 대화 상태(Dialogue State) 및 자연스러운 발화(Utterance)를 생성한다. 그리고 마지막으로 경량 언어 모델과 인간 작업자가 품질을 평가한다.

해당 논문의 저자는 제안 방법론이 다른 도메인에도 적용이 매우 용이하다고 주장하지만, 해당 방법론을 다른 도메인, 특히 다국어 언어 환경에서 적용한 사례는 아직 충분히 보고되지 않았다. 또한 LLM을 활용하여 생성한 대화의 경우 생성에 사용된 프롬프트에 따라 그 품질이 크게 영향을 받을 수 있음에도[15], 해당 연구에서는 이러한 영향을 면밀하게 다루지 않았다는 한계를 갖는다.

따라서, 본 연구의 목표는 크게 두 가지로 요약된다. 첫째, 프롬프트의 변화에 따른 대화 데이터 세트의 품질을 평가하여, 프롬프트 설계가 데이터 품질에 미치는 영향을 분석하고자 한다.

둘째, HR-MultiWOZ 방법론이 타 언어 및 도메인에도 확장 적용 가능한지 여부를 살펴보기 위해, HR 도메인이 아닌 카페 도메인과 영어가 아닌 한국어 환경에서의 적용 가능성을 사례를 통해 확인하고자 한다.

구체적으로 HR-MultiWOZ의 스키마 기반 파이프라인을 카페 주문 도메인의 한국어 대화 세트 생성에 그대로 적용하고, 생성된 LLM 기반 대화 데이터(LGD)와 클라우드 소싱(AI-Hub)으로 구축된 동일 도메인 인간 생성 데이터(HGD)의 품질을 정량·정성적으로 비교 분석하고자 한다.

본 논문의 이후 구성은 다음과 같다. 2장에서는 TOD 시스템과 TOD 데이터 세트 생성에 관한 선행 연구를 살펴보고, 3장에서는 HR-MultiWOZ에서 제안한 방법론에 기반을 둔 본 연구의 실험 설계를 소개한다. 4장에서는 실험 결과와 분석을 제시하고, 5장에서는 결론 및 향후 연구 방향을 논의한다.

www.kci.go.kr

II. Preliminaries

1. TOD System

작업 지향 대화(TOD) 시스템은 사용자가 특정 도메인(호텔 예약, 식당 문의 등)에서 목표 달성을 돕는 대화형 시스템으로, 목표 지향성·도메인 특화성·구조화된 정보 처리·모듈형 아키텍처가 특징이다[16]. 전통적인 TOD 시스템은 Fig. 1과 같이 자연어 이해(NLU), 대화 상태 추적(DST), 정책 학습(PL), 그리고 자연어 생성(NLG)의 네 모듈로 구성된다. 중앙 제어 역할의 대화 관리자(DM)가 DST와 PL을 담당하며, DST가 사용자 의도·슬롯 정보를 추출하는 핵심적 역할을 수행한다[17].

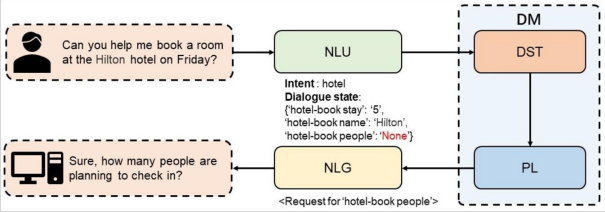


Fig. 1. Example of Pipeline based TOD System

최근에는 파이프라인의 각 모듈을 모두 통합한 종단 간(EtoE) 방식이 등장했다[16]. EtoE 시스템은 모듈화와 완전 통합형으로 구분되며, 특히 완전 EtoE는 하나의 모델이 입력부터 응답 생성까지 수행한다. 파이프라인 방식의 모듈 간 오류 전파 및 높은 개발 비용이 EtoE 방식이 대두된 배경이다[18].

초기 규칙 기반[19] 및 검색 기반[20] 시스템은 제한된 시나리오만 가능했지만, 딥러닝 기반의 다중 턴 시스템[21]으로 발전했다. 최근 GPT-4[22]와 같은 사전학습 LLM은 단일 모델로 다중 작업 처리가 가능한 것이 입증되었지만, 도메인 특화 데이터 부족으로 성능의 제한이 있다[18].

2. TOD Data

작업 지향 대화(TOD) 데이터는 항공권이나 식당 예약 등 특정 목표 달성을 위해 구조화된 대화로, 고객 서비스와 가상 비서에서 중요한 역할을 한다. TOD 시스템은 사실 기반 정확성 및 논리적 제약을 엄격히 준수해야 한다. 예를 들어 식당 예약의 경우 위치, 음식 종류, 인원 등의 제약 조건을 모두 충족해야 한다. 이에 따라, TOD 데이터는 다양한 사용자 목표, 시스템 행동, 그리고 도메인 제약을 구조적으로 반영하며, 실제 환경에서 신뢰성 있게 작동할 수 있도록 설계되어야 한다[7].

TOD 시스템 연구에 있어 TOD 데이터 세트는 핵심적인 역할을 담당하므로, TOD 시스템의 개발과 평가를 위해 다양한 데이터 세트가 구축되어 왔다. 초기에는 단일 도메인에 국한된 소규모 데이터 세트가 주로 사용되었으나, 최근에는 다중 도메인을 포괄하는 대규모 데이터 세트가 개발되고 있다[23].

MultiWOZ(Multi-Wizard of Oz)[24]는 레스토랑, 호텔, 관광, 택시, 그리고 기차 등 7개 도메인에 걸친 8,438개의 다중 턴 대화를 포함하는 대표적인 TOD 벤치마크 데이터 세트로, 각 대화는 평균 14개의 턴으로 구성되어 있다. MultiWOZ는 지속적인 개선을 통해 현재 2.4 버전까지 발전했으며, 주로 대화 상태 라벨의 오류를 수정하는 방향으로 개선되었다[25].

스키마 기반 대화 (Schema-Guided Dialogue, SGD)[26] 데이터 세트는 Table. 1과 같이 16개 도메인에 걸친 16,000개 이상의 다중 도메인 TOD 대화를 포함하는 대규모 데이터 세트이다. 이 데이터 세트는 은행, 이벤트, 미디어, 캘린더, 여행, 그리고 날씨 등 다양한 도메인을 아우르며, 의도 예측, 슬롯 채우기, 그리고 대화 상태 추적 등 다양한 작업에 활용될 수 있다.

Table 1. Characteristic Comparison of TOD Datasets

Dataset→ / Metric↓	DSTC2	WOZ-2.0	FRAMES	M2M	Multi-WOZ	SGD
No.of domains	1	1	3	2	7	**16**
No. of dialogues	1,612	600	1,369	1,500	8,438	**16,142**
Total no. of turns	23,354	4,472	19,986	14,796	113,556	**329,964**
Avg. turns of dialogue	14.49	7.45	14.60	9.86	13.46	**20.44**
Avg. tokens per turn	8.54	11.24	12.60	8.24	13.13	9.75
Total unique token	986	2,142	12,043	1,008	23,689	**30,352**
No. of slots	8	4	61	13	24	**214**
No. of slot values	212	99	3,871	138	4,510	**14,139**

비영어권에서도 CrossWOZ[27] 및 JMultiWOZ[28] 등의 데이터 세트가 개발되었으나, 여전히 영어 외 언어에서의 대화 데이터 세트는 규모와 다양성 측면에서 한계가 있다. 또한 언어별 특성과 문화적 맥락을 반영한 대화 데이터 생성은 추가적인 연구가 필요한 영역이다.

전통적으로 대화 데이터 세트는 클라우드 소싱을 통해 구축되어 왔다. 이 방법은 인간 작업자가 제공된 지침에

www.kci.go.kr

따라 직접 대화를 생성하거나, 특정 대화 시나리오를 연기하는 방식으로 진행된다[7]. 그러나 이 방법은 비용과 시간이 많이 소요되며, 대규모 데이터 세트를 구축하거나 새로운 도메인으로 확장하는 데 어려움이 있다는 한계를 갖고 있다[7]. 또한 클라우드 소싱으로 구축된 데이터는 작업자의 숙련도와 지침 이해도에 따라 품질 편차가 크게 나타날 수 있으며, 다국어 환경에서는 언어별로 고품질 작업자를 확보하는 것이 추가적인 도전 과제가 된다[7].

이러한 클라우드 소싱의 한계를 극복하기 위해 최근에는 합성 대화 데이터 생성 방법론이 대안으로 주목받고 있으며, 구체적으로 최근에는 LLM을 활용하여 대화 데이터를 생성하는 연구가 활발히 진행되고 있다[8-12].

특히 In-context learning 기법을 활용하면 미세 조정 없이도 양질의 대화 데이터를 생성할 수 있다는 장점이 있다. 또한, 이러한 접근법은 다양한 시나리오에서 자연스러운 대화를 생성할 수 있는 유연성을 제공한다. 최근 LLM을 활용하여 비용 효율적으로 기존의 벤치마크 데이터 세트에 포함되지 않은 HR 도메인에 특화된 TOD 데이터 세트를 구축하는 방법론이 제안되었다[13].

3. HR-MultiWOZ

HR-MultiWOZ는 10개 HR 도메인에 걸쳐 총 550개의 대화를 포함한, 최초의 HR 특화 공개 대화 데이터 세트이다. 이 데이터 세트는 8,910 턴과 181,363 토큰으로 구성되어 있으며[13], 다양한 HR 시나리오 및 이에 상응하는 대화 상태를 포괄하고 있다. 이러한 대화는 HR 가상 에이전트와 직원 간의 상호작용을 시뮬레이션하는 데 활용될 수 있다.

HR-MultiWOZ는 기존 MultiWOZ, M2M[29] 등과 비교해 더욱 풍부하고 상세한 대화를 포함하고 있다[13]. 이 데이터 세트는 HR 업무에 실제로 필요한 맥락, 긴 엔티티, 공감적 언어 사용 등 HR 대화의 특수성을 반영하였다. Table 2와 같이, HR-MultiWOZ 데이터 세트는 고유 토큰 비율 (0.0156), 고유 바이그램(Bigram) 비율 (0.1177), 평균 턴 수(16.2), 그리고 응답의 평균 토큰 수(14.53) 등에서 기존 데이터 세트 대비 우수한 특성을 갖고 있다.

또한 HR-MultiWOZ 데이터 세트는 MTurk (Amazon Mechanical Turk) 작업자 평가에서 직원 응답의 자연스러움, HR Assistant의 질문의 명확성과 공감성에서 모두 매우 긍정적인 평가를 받았다. 해당 논문의 저자는 이런 평가 결과를 바탕으로, HR-MultiWOZ 데이터 세트가 기존 데이터 세트에 비해 더 복잡하고 풍부하며 자연스러운 대화를 제공한다고 주장한다.

Table 2. Characteristic Comparision in MultiWOZ 2.2, M2M Restaurants and HR-MultiWOZ

Dialogue Set	MultiWOZ 2.2	M2M Restaurant	HR-MultiWOZ
# Dialogues	8,437	1,116	550
# Utterances	113,552	6,188	8,910
Avg. Utterances / Dialogue	13.46	11.09	16.2
# Tokens	1,742,157	1,742,157	181,363
Avg. Tokens / Utterance	15.34	8.07	20.35
Avg. Tokens / Response	13.46	5.56	14.35
Unique Token Ratio	0.0103	0.0092	0.0156
Unique Bigram Ratio	0.0634	0.067	0.1177

또한 해당 논문의 저자는, HR-MultiWOZ의 제안 방법론이 LLM 기반 자동화를 통해 인간 작업자의 개입을 최소화하여, 대화 데이터 세트 구축의 시간과 비용을 절감할 수 있다고 주장한다. 실제로 HR-MultiWOZ 데이터 세트 구축 사례의 경우, LLM 추론 비용 약 $38, 그리고 MTurk 비용 약 $50의 매우 적은 비용이 발생하였다.

HR-MultiWOZ의 데이터 생성 파이프라인은 Fig. 2에 나타난 바와 같으며, 기존 MultiWOZ의 Wizard-of-Oz 방식과 의도 및 슬롯을 자연어로 설명하는 스키마 기반 접근법(SGD Approach)에서 발전된 자동화 중심의 파이프라인을 따른다.

HR-MultiWOZ 방법론의 절차는 다음과 같다. 첫째, HR 전문가가 실제 업무를 반영한 도메인별 태스크 스키마를 설계한다. 이 과정에서 질문, 답변 유형, 그리고 제약조건 등이 포함된다. 둘째, LLM(Claude)[30]을 활용해 다양한 직원 프로파일을 생성하여 시나리오의 다양성을 극대화한다. 셋째, 태스크 스키마와 사용자 프로필을 결합하여 LLM을 통해 실제 상황에 가까운 시나리오(질문-답변 쌍)를 자동 생성한다. 넷째, 시나리오의 발화 순서를 랜덤하게 변형 및 병합하여 데이터의 다양성 확보한다. 다섯째, LLM을 활용해 대화를 자연스럽고 공감적으로 재구성한다. 이후, DeBERTa[31] 모델을 사용해 대화 내에서 라벨(대화 상태)을 추출하고 MTurk를 통한 검증 및 정제 과정을 거쳐 고품질 라벨을 확보한다. 마지막으로 인간 평가자들이 대화의 자연스러움, 질문의 명확성 및 공감성을 평가하여 데이터의 품질을 보장한다.

www.kci.go.kr

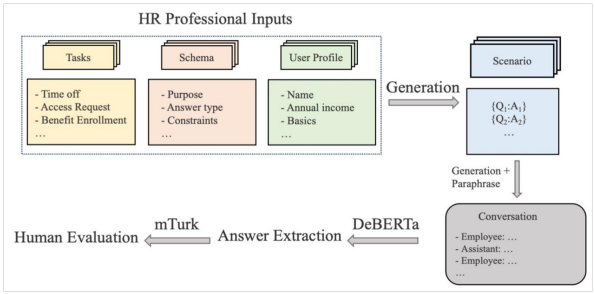


Fig. 2. Data Generation Pipeline of HR-MultiWOZ

HR-MultiWOZ 방법론이 HR 도메인과 영어 중심으로 제안되었으나, 해당 방법론에서 제시한 스키마 기반 대화 생성 방식은 다른 언어와 다른 도메인에서도 적용될 수 있는 잠재적인 가능성을 갖고 있다. 예를 들어, 태스크 스키마를 타 도메인에서 지역별 문화적 맥락에 맞게 재정의하고 LLM의 사전 학습된 지식과 다국어 생성 능력을 활용한다면, 다양한 도메인에서 현지화된 대화 데이터를 비용 효율적으로 생성할 수 있을 것이다.

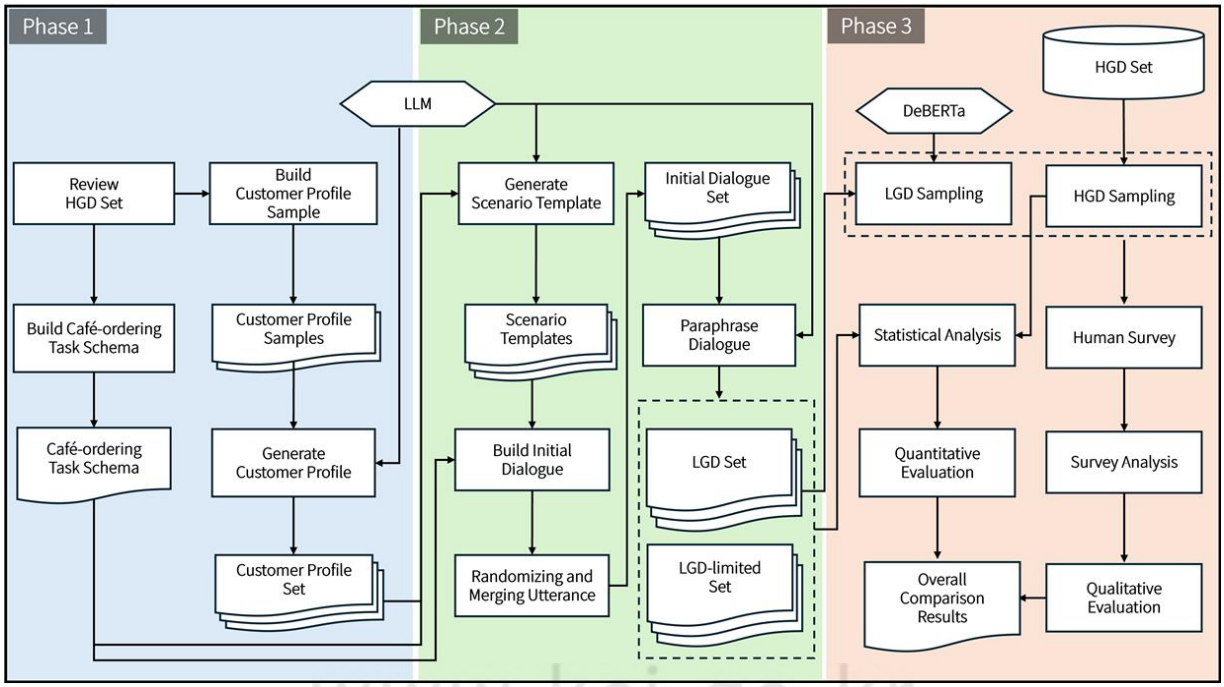
하지만 현재까지 HR-MultiWOZ 방법론을 다른 언어 및 다른 도메인으로 확장한 사례는 보고되지 않아 추가적인 검증이 필요하다. 또한 LLM을 활용하여 생성한 대화의 경우 생성에 사용된 프롬프트에 따라 그 품질이 크게 영향을 받을 수 있음에도, 해당 연구에서는 이러한 영향을 면밀하게 다루지 않았다는 한계를 갖는다.

III. The Proposed Method

1. Proposed Method Overview

본 장에서는 LLM을 활용하여 카페 도메인의 한국어 대화 세트를 생성하고, 이를 전통적인 클라우드 소싱 방식으로 구축한 대화 데이터와 비교 평가하는 방법론을 소개한다. 우선 대화 세트 생성은 HR-MultiWOZ의 데이터 생성 파이프라인을 기반으로 하되, 카페 도메인과 한국어 환경에 적합하게 수정 및 확장하여 사용한다. 그리고, 선행 연구에서 수행한 대화 데이터 세트의 평가 방법을 기반으로, LLM이 생성한 대화 세트(LLM Generated Dialogue Set, LGD)와 클라우드 소싱으로 구축한 대화 세트 (Human Generated Dialogue Set, HGD) 간의 비교 평가를 수행한다. HGD는 3절에서 구체적으로 소개하도록 한다.

제안 방법론의 전체적인 과정은 Fig. 3과 같다. 전체 프로세스는 크게 Phase 1(태스크 스키마 및 사용자 프로파일 생성), Phase 2(대화 데이터 생성), 그리고 Phase 3(평가 비교 분석)으로 구분된다. Phase 1에서는 이미 구축된 HGD를 검토하고, 유사한 카페 주문 태스크 스키마와 고객 프로파일을 구현한다. Phase 2에서는 대화 시나리오를 생성하고, LLM을 활용하여 최종적인 대화 세트를 생성한다. Phase 3에서는 정량평가와 정성평가 방식을 통해 LGD와 HGD 간의 상호 비교 분석을 수행한다.



www.kci.go.kr

Fig. 3. Proposed Mehtod Overall Process

2. Task Schema and User Profile Generation

본 절에서는 Fig. 3의 Phase 1에서 이루어지는 과정을 설명한다. Phase 1은 선행 연구의 데이터 생성 파이프라인을 따른다. 첫 단계로 카페 주문 태스크 스키마 설계를 수행한다. 카페 주문 도메인에 특화된 태스크 스키마를 설계하기 위해, 선행 연구 방법론은 SGD 데이터 세트에서 제안한 스키마 기반 접근법을 채택한다. 이 접근법은 의도와 슬롯을 자연어로 설명하여, LLM이 이전에 접하지 않은 도메인에 대해서도 효과적으로 대응할 수 있게 한다. 구체적으로는 실제 카페 주문 환경에서 발생할 수 있는 다양한 상황을 이해하기 위해, 기존의 HGD를 참조하여 태스크 스키마를 설계한다. 태스크 스키마는 Table 3과 같이, Key와 Description을 포함한다. Key는 스키마의 슬롯을 의미하며, Description은 각 Key가 가질 수 있는 값과 제약조건을 정의한다. 즉, 카페 주문을 완료하기 위한 다양한 슬롯(주문 음료, 음료 크기, 테이크-아웃 여부, 그리고 맞춤 요청 등)을 정의하고 설명한다.

Table 3. Example of Task Schema in Cafe Domain

Key	Description
주문 음료	"원하는 음료가 무엇인가요?","문자열"
음료 크기	"원하는 음료 사이즈는 무엇인가요?","문자열"
테이크-아웃 여부	"테이크-아웃인가요? 네, 아니오. 중에 선택해 주세요.","범주"
영수증 출력 여부	"영수증 출력을 원하면, 네 또는 아니오 중에 선택하세요.","범주"
…	…

다음으로, Fig. 3의 Phase 1에서 이루어지는 두 번째 과정은 고객 프로파일 생성이다. 이 프로세스는 다양한 시나리오의 대화 데이터를 생성하기 위해, 다양한 취향과 배경을 가진 고객 프로필을 생성한다. 이를 위해, 우선 Table 4와 같은 현실적인 고객 프로필 샘플을 수작업으로 제작한다. 그리고, 이를 LLM에 예시로 제공하여, 다양한 고객 프로필을 대량으로 자동 생성하도록 프롬프트 한다. 생성된 고객 프로필은 검토 과정을 거쳐, 비현실적이거나 부적절한 내용이 있는 경우 수정 또는 제거한다.

Table 4. Example of Customer Profile

Key	Description
선호 음료	"캐러멜 마키아토"
음료 크기	"벤티"
방문 이유	"원격 근무 공간으로 활용"
할인 방법	"멤버십 포인트"
…	…

3. Dialogue Data Generation

본 절에서는 Fig. 3의 Phase 2에서 이루어지는 과정을 설명한다. Phase 2 역시 선행 연구 방법론의 데이터 생성 파이프라인을 따른다. 첫 번째 단계는 시나리오 템플릿 생성이다. 카페 주문 상황에서 발생할 수 있는 다양한 시나리오를 생성하기 위해, 사용자 프로파일 세트와 태스크 스키마를 LLM의 입력으로 하여 In-context learning 기법으로 시나리오 템플릿을 생성한다(Fig. 4).

Instruction
User: {user}
Template: {template}
You are User.
Fill out all questions in template based on experience.
Generated dictionary should contain key name and generated answer.
All keys from Template are in generated dictionary.
Make the answer extremely short (within 5 words).
Put the generated dictionary in <answer></answer>XML tags.

Fig. 4. Example of Scenario Generation Instruction

Fig. 4에 나타난 LLM 지시문의 구성은 다음과 같다. LLM은 사용자 프로파일 세트의 개별 프로파일을 참고하여, 태스크 스키마에 정의된 Description, 즉 자연어 질문에 대해 응답을 생성한다. 결과적으로 최종 응답의 형태는 Fig. 5와 같이, 태스크 스키마의 Key와 이에 상응하는 응답이 하나의 쌍으로 구성된 하나의 템플릿이 된다.

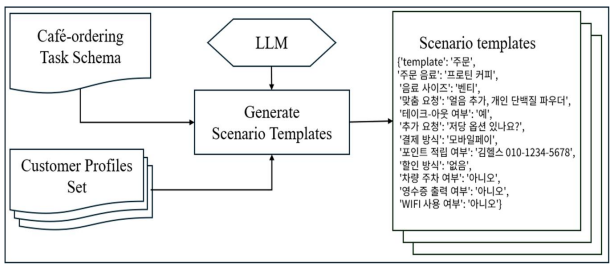


Fig. 5. Example of Scenario Generation Process

Fig. 3의 Phase 2에서 이루어지는 두 번째 과정은 초기 대화 세트 생성이다. 해당 프로세스는 최종적인 대화를 생성하기 이전의 과정으로, 두 개의 하위 프로세스로 구성된다. 먼저, 시나리오의 템플릿의 응답 값과 태스크 스키마의 Description에 정의한 자연어 질문을 조합한다. 다음으로, 질문-응답 쌍의 순서를 변경하거나 두 개의 질문-응답 쌍을 병합하는 프로세스를 랜덤하게 수행한다. 이 프로세스는 태스크 스키마의 정의에 따라 생성된 단조롭고 반복적일 수 있는 대화 흐름을 재구성하여 대화를 복잡하고 다양하게 만드는 과정으로, 그 예는 Fig. 6과 같다.

www.kci.go.kr

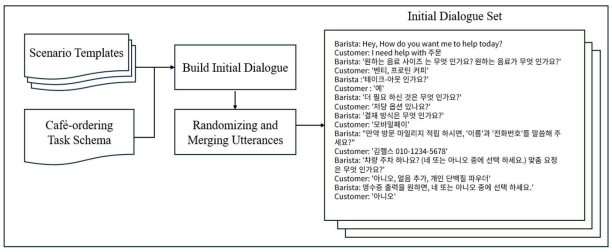


Fig. 6. Example of Initial Dialogue Generation Process

Fig. 3의 Phase 2에서 이루어지는 마지막 과정은 최종 대화 생성 프로세스이다. 해당 프로세스는 LLM을 활용하여 대화를 자연스럽게 재구성(Paraphrase)하는 과정을 진행한다. LLM 재구성 과정에서는 (i) 한국어의 자연스러운 표현과 문화적 맥락 반영, (ii) 공감적 언어와 카페 주문 상황에 고객 중심적인 말투, (iii) 이전 대화 활용하여 대화 흐름의 논리적 일관성 유지, (iv) 원래 정보와 의미 유지의 기준을 고려하여 Fig. 7과 같이 대화를 재구성한다.

바리스타: 안녕하세요! 어떻게 도와드릴까요?
고객: 음료 주문을 하고 싶은데요. 도와주실 수 있을까요?
바리스타: 네, 물론이죠! 어떤 음료를 원하시는지, 그리고 사이즈는 어떻게 해드리면 좋을까요?
고객: 벤티 사이즈로 프로틴 커피 하나 부탁드릴게요.
바리스타: 프로틴 커피 벤티 사이즈로 준비해 드리겠습니다. 혹시 적용 가능한 할인이 있으신가요? 그리고 와이파이 사용을 원하시나요?
고객: 할인은 없을 것 같아요. 와이파이도 필요 없을 것 같습니다.

----------< 중략 >----------

바리스타: 네, 얼음 추가하고 고객님의 개인 단백질 파우더도 함께 넣어드리겠습니다. 영수증은 필요하신가요?
고객: 영수증은 안 필요할 것 같아요. 감사합니다.

Fig. 7. Example of Paraphrased Dialogue

추가로 본 연구에서는 LLM 프롬프트의 변화가 대화 데이터의 품질에 영향을 주는 양상을 확인하기 위한 실험을 수행한다. 이를 위해, 대화 재구성 단계의 프롬프트에 '발화 당 토큰 수를 10개 내외로 제한한다.'는 지시사항을 추가하는 방식으로 LGD-*limited* 데이터를 생성한다. 이러한 추가 실험에 대한 분석 결과는 이후 장에서 소개한다.

4. Comparison Dataset and Evaluation Criteria

본 절에서는 제안 방법론에 따라 LLM 기반으로 생성한 대화 세트, 즉 LGD의 품질을 상대적으로 평가하기 위한 방안을 소개한다. 구체적으로 품질 평가의 비교 대상이 되는 데이터 세트, 즉 클라우드 소싱을 통해 수작업으로 생성한 데이터인 HGD를 소개하고, 이를 대상으로 평가가 이루어지는 Phase 3의 과정을 설명한다.

Table 5. Example of Cafe domain Dialogue in AI-Hub Korean Dialogue Dataset

Dialogue (10 Utterances)	Dialogue (27 Utterances)
고객: 아이스 아메리카노 하나요 **점원:** 테이크아웃하실 건가요? **고객:** 네 텀블러에 넣어주세요 **점원:** 텀블러 할인 300원 해드릴게요 **고객:** 그란데 사이즈로 주세요 **점원:** 적립카드 있으세요? **고객:** 네 보고 쿠폰으로 결제할게요 **점원:** 현금영수증 해드릴까요? **고객:** 괜찮습니다 **점원:** 저쪽에서 잠시만 대기해주세요	**고객:** 테이크아웃은 더 싼가요? **점원:** 아니요 저희는 똑같아요 **고객:** 아메리카노 뜨거운 걸로 하나 주세요 **점원:** 네 잠시만요 **고객:** 더치커피도 있어요? **점원:** 더치커피는 없어요 **고객:** 텀블러 가져오면 할인해주시나요? **점원:** 저희는 텀블러 할인이 안됩니다 **고객:** 카카오페이도 돼요? **점원:** 카카오페이로는 결제가 안됩니다 **고객:** 삼성페이는 되고요? **점원:** 네 삼성페이는 됩니다 **고객:** 그냥 결제해주세요 **점원:** 네 영수증 드리겠습니다 --------< 중략 >-------- **고객:** 카페 생긴 지 얼마나 됐나요? **점원:** 일년 좀 넘었어요 **고객:** 원두는 로스팅 된 거 사오시나요? **점원:** 네 저희는 원두 주문하는 곳이 있습니다

우선, HGD에 대한 개요는 다음과 같다. HGD는 한국지능정보사회진흥원(NIA)이 AI-Hub 이니셔티브의 하나로 구축한 한국어 대화 데이터 세트이다[32]. 이 대규모 한국어 대화 데이터 세트는 10개의 도메인으로 구성되어 있으며, 본 연구에서는 카페 도메인의 대화 데이터를 성능 평가에 사용한다(Table 5). 해당 데이터 세트는 챗봇 서비스 개발 시 카페 운영에 필요한 대화 시스템의 구축에 활용하기 위해 구축되었으며, 총 대화 524건, 총 발화 7,859개, 대화 당 평균 발화 15개로 구성되어 있다.

Fig. 3의 Phase 3에서 이루어지는 성능 평가 과정은 다음과 같다. 우선 첫 번째 단계는 정량평가(Quantitative Evaluation)이다. 이 과정은 선행 연구에서 사용한 정량적인 평가 방법을 채택하되, 본 연구에서는 범위와 항목을 조정하여 적용하였다. 정량평가 분석 프로세스의 하위 단계는 다음과 같다. 우선 HGD 데이터는 대화 당 발화 수가 최소 10개부터 최대 27개까지 매우 다양한데, 이 가운데 대화 당 평균 발화 수가 15개 내외인 대화를 샘플링하여 사용한다. 또한, 이와 비슷한 특성을 갖는 LGD와 LGD-*limited*를 생성한다. 이 과정을 통해 준비된 데이터의 특성이 Table 6에 요약되어 있다.

www.kci.go.kr

Table 6. Data samples for Quantitative Evaluation

Dialogue Set	HGD	LGD	LGD-*Limited*
# Dialogues	46	46	46
# Utterances	692	654	656
Avg. Utterances / Dialogue	15	14	14

다음으로, 각 샘플 데이터 세트를 대상으로 통계 분석과 정량평가를 수행한다. 통계 분석의 경우, 대화 데이터를 한국어 어절 단위의 토큰으로 계산하여 수행한다. 정량평가는 분석 데이터를 바탕으로 발화의 문맥적 복잡성과 다양성을 평가하기 위하여 수행되며, Table 7과 같이 토큰의 길이 및 고유성 등을 주요 평가 요소로 채택한다.

Table 7. Metrics for Quantitative Evaluation

Metric	High	Low
Avg. Tokens / Utterance	Detailed or complex utterances	Concise utterances
Unique Token & Bigram Ratio	Diverse vocabulary and patterns	Repetitive vocabulary and patterns

Fig. 3의 Phase 3에서는 두 번째 분석으로 정성평가(Quantitative Evaluation)를 수행한다. 이 과정은 선행 연구에서 사용된 휴먼평가 방법을 채택하되, 본 연구에서는 범위와 항목을 조정하여 적용한다.

정성평가 프로세스는 다음과 같이 수행된다. 첫 번째는 데이터 샘플링이다. 선행연구의 샘플링 방법에 따라, 앞선 정량평가 단계에서 추출한 데이터 세트로부터 정성평가 대상의 발화들이 추출된다. 구체적으로, LLM 대화 세트를 생성한 후, 대화 내의 질문-응답 발화의 쌍을 입력으로 대화 상태 값을 추출한다. 이때 사용하는 사전학습 언어모델은 대화 상태 값과 이에 해당하는 신뢰도 점수(Confidence Score)를 반환한다. 일정 수준의 품질이 보장된 데이터를 대상으로 선별하기 위해, 신뢰도 점수가 60% 이상인 질문-응답 발화의 쌍만을 평가대상으로 선정한다. 이 과정을 통해 LGD로부터 신뢰도 점수가 60% 이상인 질문-응답 발화 쌍 42개가 추출된다. 반면, HGD의 경우는 클라우드 소싱을 통해 이미 품질 관리 프로세스를 통과한 것으로 간주하여, LGD에서 추출한 샘플 수와 동일한 수의 질문 및 응답 발화를 랜덤하게 42개 추출하여 사용한다. 부가적으로 LGD의 총 발화 대비 샘플의 비율은 선행 연구의 비율과 유사한 12~13% 수준임을 확인했다.

다음으로, 데이터 품질에 대한 정성평가를 위해 설문 조사를 수행한다. 설문에서는 선행 연구의 지표와 동일하게 답변의 자연스러움, 질문의 명확성, 그리고 질문의 공손함을 평가한다. 평가자의 경우, 선행 연구에서는 MTurk 작업자 한 명이 하나의 지표에 대해 응답을 했지만, 본 연구에서는 자연어 처리 분야 전문가 3명이 모든 지표에 대한 설문에 고르게 응답한다. 평가 척도는 5점 척도를 사용하였으며, 예를 들어 '자연스러운 답변' 지표에서 1점은 가장 부자연스러운 답변을, 5점은 가장 자연스러운 답변을 의미한다. 또한 선행 연구는 해당 방법론으로 생성된 대화 데이터를 대상으로만 설문 조사를 수행했지만, 본 연구에서는 LGD와 HGD에 대한 조사를 동시에 수행하여 생성 데이터의 상대적인 품질을 평가한다. 평가의 일관성 및 신뢰도 향상을 위해, 평가자에게는 각 지표별 지침을 사전에 제공한다. 설문에는 LGD와 HGD의 구분 표시가 없으며 질문의 순서도 무작위로 설정하였다.

설문이 완료되면, 응답 결과를 집계하여 평가 및 분석을 수행한다. 구체적으로, 지표별 응답 평균 점수를 산출하고, 표본 t-검정(One-sample t-test) 분석을 통해 응답 결과의 신뢰 수준을 확인한다.

Fig. 3의 Phase 3의 마지막 단계에서는, 앞선 정량평가와 정성평가를 종합적으로 분석하여 HGD 대비 LGD의 품질과 한계를 확인한다.

IV. Experiment

1. Experiment Overview

본 장에서는 앞서 제안한 방법론에 따라 수행한 실험 결과를 소개한다. 실험 환경은 Python을 통해 구축하였으며, 자세한 SW환경은 Table 8과 같다. 본 실험에서 사용한 LLM과 추출모델 (Extractive Model)은 모두 선행 연구의 실험 환경을 따랐으며, LLM은 본 실험의 수행 시점 기준으로 가장 최신 버전의 Claude 모델을 사용하였다.

Table 8. Experiment Environment

SW	LLM	claude-3-7-sonnet-20250219
	DeBERTa	deepset/ deberta-v3-large-squad23.8.10

2. Quantitative Evaluation Analysis

본 절에서는 통계 분석을 통해 도출한 정량평가 결과를 분석한다. 정량평가는 LLM을 활용하여 생성한 대화 데이터 세트(LGD)와 클라우드 소싱으로 구축된 대화 데이터 세트(HGD), 그리고 제한된 프롬프트로 생성한 대화 데이

www.kci.go.kr

터 세트 (LGD-*limited*)의 세 가지 데이터 세트를 대상으로 수행하였다.

Table 9에 나타난 바와 같이 세 데이터 세트는 총 발화 수와 응답 발화 수에서 유사한 규모를 보였지만, 발화당 토큰 수에서는 큰 차이를 나타냈다. LGD는 발화당 평균 10.35개의 토큰 수로 가장 높은 수치를 보였으며, 이는 HGD의 평균 4.29개보다 두 배 이상 높았다. LGD-*limited*는 발화당 평균 6.62개의 토큰으로 중간 수준을 유지했다. 응답 발화에서도 유사한 패턴이 관찰되었는데, LGD가 평균 8.16개로 가장 높았고, HGD가 4.56개, 그리고 LGD-*limited*는 5.78개로 나타났다.

Table 9. Utterance Analysis Result

Metric	HGD	LGD	LGD-*limited*
# Utterances	692	654	656
# Response Utterances	320	327	328
Avg. Tokens / Utterance	4.29	10.35	6.62
Avg. Tokens / Response	4.56	8.16	5.78

Table 10의 토큰 분석에서도 세 데이터 세트는 뚜렷한 차이를 보였다. LGD는 총 6,767개의 토큰으로 가장 많은 토큰 수를 가지고 있으며, HGD는 2,966개, LGD-*limited*는 4,345개였다. 어휘 다양성 측면에서는 HGD의 고유 토큰 비율이 44.0%로 가장 높게 나타났고, LGD와 LGD-*limited*는 각각 16.3%와 17.9%로 비교적 낮은 고유 토큰 비율을 보였다.

Table 10. Token Analysis Result

Metric	HGD	LGD	LGD-*limited*
# Tokens	2,966	**6,767**	4,345
# Unique Tokens	**1,304**	1,101	777
Unique Token Ratio	**0.440**	0.163	0.179

가장 빈번한 토큰 TOP 5 분석 결과, HGD에서는 "네"가 139회(20.1%)로 가장 많이 등장했으며, "주세요"가 76회(11.0%)로 그 뒤를 이었다(Table 11). LGD에서는 "네"가 215회(31.7%), "혹시"가 209회(31.2%)로 높은 빈도를 보였다. LGD-*limited*에서는 "혹시"가 130회(19.7%), "네"가 102회(15.5%)로 가장 많이 등장했다.

Table 11. TOP 5 Tokens in Token Frequency

Rank	HGD	LGD	LGD-*limited*
1	네: 139 (20.1%)	네: 215 (31.7%)	혹시: 130 (19.7%)
2	주세요: 76 (11.0%)	혹시: 209 (31.2%)	네: 102 (15.5%)
3	하나: 38 (5.5%)	그리고: 181 (27.7%)	것: 84 (12.5%)
4	따뜻한: 34 (4.9%)	준비해: 135 (19.6%)	어떤: 75 (11.3%)
5	아메리카노: 32 (4.6%)	것: 125 (18.5%)	있으신가요:69 (10.5%)

이러한 분석은 세 데이터 세트가 서로 다른 발화 패턴을 나타내고 있음을 시사한다. 구체적으로 HGD는 비교적 간결하고 다양한 패턴, LGD는 비교적 자세하고 반복적인 패턴을 갖는 것으로 해석할 수 있다, 그리고 LGD-*limited*의 경우, 고유 토큰 비율은 LGD와 유사하게 측정되었으나 발화 당 토큰 수는 중간 수준에 위치하는 등, 프롬프트의 변경에 따라 발화 특성이 조절될 수 있는 가능성을 보였다.

바이그램 분석에서는, HGD가 66.2%의 고유 바이그램 비율을 나타내며 가장 다양한 언어 패턴을 보였다(Table 12). 한편 LGD와 LGD-*limited*는 각각 38.9%와 38.2%로 유사한 수준의 고유 바이그램 비율을 나타냈다. 각 데이터 세트의 가장 빈번한 바이그램 TOP 5를 살펴보면, HGD는 "하나 주세요"(19회), "한 잔"(17회)과 같은 직접적인 주문 표현이 많았다(Table 13). 한편, LGD는 "것 같아요"(98회), "준비해 드리겠습니다"(73회)와 같은 정중하고 상세한 표현이 두드러졌으며, LGD-*limited*는 "것 같아요"(66회), "준비해 드릴까요"(40회)와 같은 질문 중심의 서비스 표현이 많이 나타났다.

Table 12. Bigram Analysis Result

Metric	HGD	LGD	LGD-*limited*
# Bigrams	2,274	6,113	3,689
# Unique Bigrams	1,963	2,630	1,661
Unique Bigram Ratio	**0.662**	0.389	0.382

Table 13. TOP 5 Bigrams in Bigram Frequency

Rank	HGD	LGD	LGD-*limited*
1	하나 주세요 : 19회	것 같아요 : 98회	것 같아요 : 66회
2	한 잔 : 17회	준비해 드리겠습니다 : 73회	준비해 드릴까요 : 40회
3	수 있어요 : 10회	수 있을까요 : 55회	수 있을까요 : 39회
4	따뜻한 거 : 9회	없을 것 : 54회	차량주차하셨나요 : 37회
5	영수증 주세요 : 8회	필요하신 것이 : 42회	사이즈는 어떻게 : 35회

www.kci.go.kr

위 실험에서 LGD가 가장 높은 바이그램 빈도수를 나타냈으며, 이는 반복적인 표현의 특성을 반영하는 것으로 보인다. 반면 HGD는 비교적 간결한 발화 패턴에 일치하는 낮은 빈도수를 보였다. LGD-*limited*는 바이그램 빈도 측면에서 다른 두 데이터 세트의 중간에 위치하는 것으로 나타났다. 이러한 결과는 앞선 토큰 분석 결과와 일치한다. HGD는 직접적이고 간결한 발화 패턴을 나타내고, LGD는 형식적이고 상세한 표현을 보이며, LGD-*limited*는 두 데이터 세트의 중간 수준에 위치하여 프롬프트의 변경에 따라 발화의 통계적 특성이 조절될 수 있음을 보였다.

3. Qualitative Evaluation Analysis

본 절에서는 3장에서 설계한 설문조사 방식으로 3명의 평가자의 응답을 집계하여 정성평가 결과를 분석한다. 정성평가 역시 LGD와 HGD 모두를 대상으로 수행하였으며, 두 데이터 세트를 '명확한 질문', '자연스러운 답변', 그리고 '공손한 질문'의 세 가지 지표로 LGD와 HGD 각각 126개의 응답을 집계하였다. (Table 14, 16).

각 지표는 5점 척도(1점: 매우 낮음, 3점: 보통, 5점: 매우 높음)로 측정되었으며, one-sample t-test를 통해 통계적 검증을 시행하였다. one-sample t-test는 단일 집단의 평균이 특정 기준값과 다른지 확인하기 위해 사용하는 분석으로, 본 연구에서는 각 데이터 세트의 지표별 응답 평균 점수와 기준값(3점)의 차이가 통계적으로 유의미한지 여부를 판단하기 위해 해당 분석을 수행하였다(Table 15).

Table 14. Average Score Comparsion by Metrics

Metric	Avg. Response Score	
	LGD	HGD
Clear Questions	4.21	3.79
Natural Questions	4.15	3.86
Polite Questions	4.28	2.82

Table 15. One-sample t-test Result

Data	Metric	One-sample t-test (3-point scale)	
		t-stat.	p-value
LGD	Clear Questions	18.87	≤ 0.000000001
	Natural Questions	12.76	≤ 0.000000001
	Polite Questions	19.55	≤ 0.000000001
HGD	Clear Questions	6.95	≤ 0.000000001
	Natural Questions	8.21	≤ 0.000000001
	Polite Questions	-1.95	0.026192

Table 16. survey response counts by Metrics

Metric	# Evaluators	# Items		# Resp.	
		LGD	HGD	LGD	HGD
Clear Questions	3	42	42	126	126
Natural Questions	3	42	42	126	126
Polite Questions	3	42	42	126	126

Table 14에서, LGD는 모든 평가 지표에서 우수한 평가를 받은 것으로 나타났다. 구체적으로 '명확한 질문'에서 4.21점, '자연스러운 답변'에서 4.15점, '공손한 질문'에서 4.28점을 기록하며 모든 영역에서 4점 이상의 높은 점수를 획득하였다. 또한 이러한 결과는 Table 15에서 $p \leq 0.05$ 기준 통계적으로 유의한 것으로 나타났다. 즉, LLM이 고객 서비스 맥락에서 요구되는 공손함과 정중함 측면에서, 이상적인 형태를 잘 표현할 수 있음을 시사한다.

가장 흥미로운 점은 '공손한 질문' 영역에서 두 데이터 간의 점수 차이가 1.44점으로 가장 크게 나타났다는 것이다. 이는 HGD는 실제 카페 현장에서의 대화를 그대로 반영하여, 때로는 덜 정중하거나 직접적인 표현이 포함되었기 때문인 것에서 그 원인을 찾을 수 있다.

V. Conclusions

본 연구에서는 작업 지향 대화(TOD) 시스템의 고비용·도메인 한정적인 데이터 구축 문제를 해결하기 위해, LLM을 활용한 합성 대화 데이터 생성 방법론을 소개하고 분석하였다. HR-MultiWOZ 방법론을 한국어 카페 도메인에 적용해 클라우드 소싱 데이터와 비교한 결과, 프롬프트 설계가 대화 데이터 품질과 특성에 직접적인 영향을 미칠 수 있음을 확인하였다. 또한 합성 데이터는 정중하고 상세하나 어휘 및 패턴 다양성이 부족하다는 특징을 보였으며, 이러한 한계를 향후 프롬프트 전략 개선을 통해 보완할 수 있는 가능성도 확인하였다.

본 연구의 기여는 크게 세 가지로 요약된다. 첫째, 프롬프트 엔지니어링의 중요성을 실증적으로 입증하였다. 둘째, HR-MultiWOZ 방법론의 도메인 및 언어 확장성을 한국어 카페 도메인에 적용해 검증하였다. 셋째, 합성·실제 데이터 간의 특성을 정량·정성적으로 분석하여 장단점을 명확히 규명하였다.

본 연구는 데이터 세트 생성에 사용되는 프롬프트의 설계가 데이터의 품질 및 특성에 영향을 미침을 확인하였지

www.kci.go.kr

만, 이 과정에서 발화 길이 제한이라는 한 가지 특성만을 고려했다는 한계를 갖는다. 향후 프롬프트 설계를 통해 더욱 다양한 특성을 엄밀하게 조절하는 실험을 수행할 필요가 있다. 또한 본 실험에서는 카페 도메인이라는 한정된 도메인의 대화 데이터 세트에 대한 성능 평가만을 수행하였기 때문에, 본 연구의 결과를 일반화하여 적용하기는 어렵다는 한계를 갖는다. 마지막으로 본 연구의 정성평가는 전반적인 경향을 확인하기 위해 수행되었으나, 평가자 간 응답의 일관성 및 LGD와 HGD 간의 직접적인 통계 비교 결과는 제시하지 못하였다. 이러한 점은 연구 해석의 엄밀성을 제한하는 요소로, 향후 연구에서는 평가자 신뢰도 분석과 집단 간 직접 비교에 대한 통계적·정량적 검증이 필수적으로 보완될 필요가 있다.

REFERENCES

[1] Allied Market Research (2021) Conversational AI Market Outlook-2030, *https://www.alliedmarketresearch.com*

[2] A. Algherairy and M. Ahmed, “Review of Dialogue Systems: Current Trends and Future Directions.”, Springer on Neural Computing and Applications, Vol. 36, pp. 6325-6351, February 2024. DOI:10.1007/s00521-023-09322-1

[3] J. Liu et al., “What Makes Good In-context Examples for GPT-3”, arXiv:2101.06804, Jan, 2021. DOI:10.48550/arXiv.2101.06804

[4] A. Chowdhery et al., “PaLM: Scaling Language Modeling with Pathways.” arXiv.2204.02311, Apr. 2022. DOI: 10.48550/arXiv.2204.02311.

[5] H. Touvron et al., “LLaMA: Open and Efficient Foundation Language Models.” arXiv.2302.13971, Feb 2023. DOI: 10.48550/arXiv.2302.13971

[6] V. Hudeček and O. Dušek, “Are LLMs All You Need for Task-Oriented Dialogue?” arXiv:2304.06556, August 2023. DOI: 10.48550/arXiv.2304.06556

[7] H. Soudani et al., “A Survey on Recent Advances in Conversational Data Generation” arXiv.2405.13003, May 2024. DOI: 10.48550/arXiv.2405.13003

[8] S. Terragni et al., “In-Context Learning User Simulators for Task-Oriented Dialog Systems” arXiv:2306.00774, June 2023. DOI: 10.48550/arXiv.2306.00774

[9] Z. Li et al., “Controllable Dialogue Simulation with In-Context Learning” arXiv:2210.04185, June 2023. DOI: 10.48550/arXiv.2210.04185

[10] Z. Ahmad et al., “INA: An Integrative Approach for Enhancing Negotiation Strategies with Reward-Based Dialogue System” arXiv:2310.18207, Oct. 2023. DOI:10.48550/arXiv.2310.18207

[11] M. Chen et al., “PLACES: Prompting Language Models for Social Conversation Synthesis” arXiv:2302.03269, Feb. 2023. DOI:10.48550/arXiv.2302.03269

[12] H. Kim et al., “SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization” arXiv:2212.10465, Oct 2023. DOI:10.48550/arXiv.2212.10465

[13] W. Xu et al., “HR-MultiWOZ: A Task Oriented Dialogue (TOD) Dataset for HR LLM Agent” arXiv:2402.01018, Feb 2024. DOI:10.48550/arXiv.2402.01018

[14] A. Rastogi et al., “Towards Scalable Multi-domain Conversational Agents: The Schema-Guided Dialogue Dataset” arXiv:1909.05855, Jan 2020. DOI:10.48550/arXiv.1909.05855

[15] G. Marvin et al., “Prompt Engineering in Large Language Models” Springer on Algorithms for Intelligent Systems, Conference paper, pp.387-402, Jan 2024. DOI: 10.1007/978-981-99-7962-2_30

[16] Z. Yi, “A Survey on Recent Advances in LLM-Based Multi-turn Dialogue Systems” arXiv:2402.18013, Feb 2024. DOI:10.48550/arXiv.2402.18013

[17 A. Rastogi et al., “Schema-Guided Dialogue State Tracking Task at DSTC8” arXiv:2002.01359, February 2020. DOI: 10.48550/arXiv.2002.01359

[18] H. Xu et al., “Rethinking Task-Oriented Dialogue Systems: From Complex Modularity to Zero-Shot Autonomous Agent” Association for Computational Linguistic, Vol. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pp.2748–2763, August 2024. DOI:10.18653/v1/2024.acl-long.152

[19] J. Weizenbaum, “ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine.” Association for Computing Machinery, Vol 9, pp.36-45, January 1966. DOI:10.1145/365153.365168

[20] D. Goddeau et al., “A form-based dialogue manager for spoken language applications” IEEE on Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, Vol 2, pp.701-704 Oct. 1996. DOI: 10.1109/ICSLP.1996.607458

[21] I. Serban et al., “Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models” arXiv:1507.04808, April 2016. DOI:10.48550/arXiv.1507.04808

[22] OpenAI et al., “GPT-4 Technical Report” arXiv:2303.08774, March 2024. 10.48550/arXiv.2303.08774

[23] Papers With Code, *https://paperswithcode.com/datasets*

[24] P. Budzianowski et al., “MultiWOZ -- A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling” arXiv:1810.00278, May 2019. DOI: 10.48550/arXiv.1810.00278

[25] F. Ye et al., “MultiWOZ 2.4: A Multi-Domain Task-Oriented Dialogue Dataset with Essential Annotation Corrections to Improve State Tracking Evaluation” arXiv:2104.00773, April

www.kci.go.kr

2021. DOI: 10.48550/arXiv.2104.00773
[26] A. Rastogi et al., “Towards Scalable Multi-domain Conversational Agents: The Schema-Guided Dialogue Dataset” arXiv:1909.05855, Sep 2019. DOI:10.48550/arXiv.1909.05855
[27] Q. Zhu et al., “CrossWOZ: A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset” arXiv:2002.11893, February 2020. DOI: 10.48550/arXiv.2002.11893
[28] A. Ohashi et al., “JMultiWOZ: A Large-Scale Japanese Multi-Domain Task-Oriented Dialogue Dataset” arXiv:2403.17319, Mar 2024. DOI: 10.48550/arXiv.2403.17319
[29] P. Shah et al., “Building a Conversational Agent Overnight with Dialogue Self-Play” arXiv:1801.04871, Jan 2018. DOI: 10.48550/arXiv.1801.04871
[30] Claude, *https://claude.com/product/overview*
[31] P. He et al., “DeBERTa: Decoding-enhanced BERT with Disentangled Attention” arXiv:2006.03654, Oct 2021. DOI: 10.48550/arXiv.2006.03654
[32] AI HUB ‘Korean Dialogue’, *https://www.aihub.or.kr*

Authors

Changgou Kang received B.S. degree in Economics from KyungSung University in 2003 and M.S. degree in Graduate School of Business IT, Kookmin University in 2025. He has over 17 years of experience in business IT industry. Currently, He works as an AI engineer at Doosan. He is interested in prompt-based learning, deep learning, and natural language processing.

Namgyu Kim received the B.S. in Computer Engineering from Seoul National University in 1998, M.S. and Ph.D. degrees in Management Engineering from KAIST, Korea, in 2000 and 2007, respectively.

Dr. Kim joined the faculty of the School of Management Information Systems at Kookmin University, Seoul, Korea, in 2007. He served as the Dean of the Graduate School of Business IT at Kookmin University and is currently a professor at the Business IT. He is interested in LLM, text mining, and deep learning.

www.kci.go.kr