JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 7, pp. 55-64, July 2026
https://doi.org/10.9708/jksci.2026.31.07.055

Integrating Emotion, Situation, and Empathy Strategies in Retrieval-Augmented Response Generation

So-Jeong Park*, Beak-Cheol Jang**

*Student, Graduate School of Information, Yonsei University, Seoul, Korea
**Professor, Graduate School of Information, Yonsei University, Seoul, Korea

[Abstract]

Empathetic Response Generation aims to produce appropriate responses by understanding a speaker's emotions and situational context, and it requires a balanced integration of affective and cognitive empathy. Prior studies have selectively addressed emotion recognition, situation/cause modeling, and empathy strategies. Recent retrieval-augmented approaches retrieve examples mainly by semantic similarity and treat strategy selection as a separately fine-tuned classifier. In this paper, we propose a RAG-based framework that extracts emotion and situation from user utterances and performs two-stage retrieval, consisting of situation-embedding-based initial retrieval and emotion-based filtering refinement; the empathy-strategy labels of the retrieved examples are used directly as few-shot context, without a separate strategy-prediction model.

▸ **Key words**: Empathetic Response Generation, Retrieval-Augmented Generation, Emotion Recognition, Empathy Strategy, Large Language Model

[요 약]

공감적 응답 생성(Empathetic Response Generation)은 화자의 감정과 상황을 이해하고 적절한 반응을 생성하는 과제로, 정서적 공감과 인지적 공감을 균형 있게 통합하는 것이 핵심이다. 기존 연구들은 감정 인식, 상황/원인 모델링, 공감 전략을 선별적으로 다루어 왔으며, 최근 연구는 의미 유사도를 중심으로 예시를 검색하고 전략 선택을 별도의 파인튜닝 모듈에 위임한다. 본 논문은 화자 발화에서 감정과 상황을 추출하고, 상황 임베딩 기반 1차 검색과 감정 필터링 기반 2차 정제 검색(Two-Stage Retrieval)을 통해 검색된 예시의 공감 전략 라벨을 별도의 전략 예측 모듈 없이 few-shot 문맥으로 직접 활용하는 RAG 기반 공감 응답 생성 프레임워크를 제안한다.

▸ **주제어**: 공감 대화 생성, 검색 증강 생성, 감정 인식, 공감 전략, 대형 언어 모델

• First Author: So-Jeong Park, Corresponding Author: Beak-Cheol Jang
*So-Jeong Park (qkrthwjd714@yonsei.ac.kr), Graduate School of Information, Yonsei University
**Beak-Cheol Jang (bjang@yonsei.ac.kr), Graduate School of Information, Yonsei University
• Received: 2026. 05. 13, Revised: 2026. 06. 29, Accepted: 2026. 07. 01.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

인간 간의 대화에서 공감은 상대방의 감정과 상황을 이해하고 적절히 반응하는 핵심적인 소통 능력이다. 심리학적으로 공감은 타인의 감정을 반영하는 정서적 공감(Affective Empathy)과 상황과 원인을 이해하는 인지적 공감(Cognitive Empathy)으로 구분되며[11], 이 두 축을 균형 있게 통합할 때 비로소 이상적인 공감 반응이 형성된다. 이러한 공감 능력을 대화 시스템에 구현하는 공감적 응답 생성은 챗봇, 고객 상담, 정신건강 지원 등 다양한 분야에서 중요성이 높아지고 있다.

초기 연구들은 감정 라벨을 직접 활용하여 감정에 부합하는 응답을 생성하는 데 집중하였으며[2, 3], 이후 감정의 원인이 되는 상황 정보를 함께 모델링하는 방향으로 발전하였다[4, 7]. 최근에는 위로, 공감, 격려 등 구체적인 공감 전략을 응답 생성에 명시적으로 반영하는 연구가 주목받고 있다[6, 8]. APTNESS [8]와 같이 RAG와 전략 선택 모듈을 결합한 공감 응답 생성 프레임워크가 등장하였으나, 의미 유사도(Semantic Similarity)를 중심으로 예시를 검색하는 구조로, 화자의 감정과 상황이 검색 조건에 직접 반영되지는 않는다. 또한 전략 선택은 별도로 파인튜닝된 분류 모듈에 위임되어 검색과 전략이 분리된 파이프라인 구조를 가진다.

본 논문은 이러한 한계를 극복하기 위해 감정, 상황, 공감 전략을 통합하는 검색 증강 기반 공감 응답 생성 프레임워크를 제안한다. 제안 프레임워크는 다음 세 단계로 구성된다. 본 연구의 주요 기여점은 다음과 같다.

- 공감 응답 생성에 감정과 상황을 동시에 고려하는 2단계 검색 구조를 도입한다.
- 검색된 예시의 공감 전략 라벨을 명시적으로 활용하는 few-shot 프롬프팅 구조를 설계하여 응답의 제어 가능성과 해석 가능성을 향상시킨다.
- EMP-EVAL 기반 다차원 평가 체계에 strategy F1, sim_response_gt를 추가한 종합 평가 체계를 구성하고, 모든 지표 표기를 표·그림·본문에서 일관되게 정리한다.
- 다섯 가지 검색 조건에 대한 ablation study를 세 가지 LLM 환경에서 수행하고, 통계적 유의성 검정을 함께 제시하여 각 구성 요소의 기여도를 실험적·통계적으로 규명한다.

II. Preliminaries

1. Related works

1.1 Empathetic Response Generation

공감적 응답 생성은 화자의 감정 상태를 인식하고 그에 적합한 반응을 생성하는 과제로, 심리학적으로 정서적 공감과 인지적 공감이라는 두 축으로 구성된다 [11]. 전자는 감정에 정서적으로 반응하는 능력을, 후자는 타인의 감정 원인과 상황을 파악하고 추론하는 능력을 의미하며, 이상적인 공감 대화는 두 축을 균형 있게 통합해야 한다.

초기 연구들은 주로 감정 라벨을 직접 활용하는 방향에 집중하였다. MoEL [2]은 감정별로 특화된 복수의 리스너 디코더를 구성하고 화자 발화의 감정 분포를 추적하여 이를 선택적으로 결합해 응답을 생성한다. 이후 MIME [3]은 감정의 극성(Polarity)에 따라 군집을 분리하고 감정 모방(Mimicry)을 도입해 응답 다양성을 높였다. 그러나 두 모델 모두 감정 라벨에만 의존하여 해당 감정이 왜 발생했는지를 이해하지 못한다는 한계를 가진다.

이를 개선하기 위해 SDAM [7]은 상황 정보를 대화 맥락과 연관 짓는 양방향 필터링 인코더를 도입하여 인지적 공감을 강화하였고, CEM [4]은 COMET 기반 상식 추론을 통해 화자의 암묵적 상황을 추가로 이해하는 방향으로 나아갔다. Qian 등 [5]은 감정 원인 간의 전이 관계를 그래프로 모델링하였으며, Cai 등 [13]은 상식 지식을 대화 흐름에 따라 동적으로 주입하는 방식을 제안하였다. CASE [6]는 상식 인지 그래프와 감성 개념 그래프를 구축하고 이를 거칠고 세밀한 두 수준에서 정렬하는 방법을 제안하였다. 그러나 이들 모두 파인튜닝 의존적이며, 공감 전략을 명시적으로 선택하거나 외부 지식 검색을 통해 생성을 제어하는 구조는 갖추지 못한다.

1.2 Retrieval-Augmented Empathetic Response Generation

Retrieval-Augmented Generation(RAG)은 질의와 관련된 외부 정보를 검색하여 응답 생성에 활용하고 응답 품질을 향상시키는 기법이다 [9]. 최근 APTNESS [8]를 중심으로 공감 대화 생성에 RAG를 적용하여 유사한 공감 예시를 few-shot 문맥으로 활용하는 시도가 등장하기 시작했다.

APTNESS [8]는 평가 이론에 기반한 감성 팔레트(7개 대범주, 23개 소범주)를 구성하고 이를 활용해 공감 응답 데이터베이스를 구축한다. 응답 생성 시 의미적으로 유사한 예시를 검색하여 LLM 입력에 통합하고, 파인튜닝 모듈을 통해 정서 지원 전략을 선택하는 구조를 사용한다.

APTNESS는 RAG와 전략 선택을 결합한 선도적 시도이나, 검색 키로 의미 유사도만을 사용하며 전략 예측을 위해 별도의 파인튜닝 모듈에 의존한다는 한계를 가진다.

본 논문은 APTNESS의 RAG 기반 전략 활용 방식에서 착안하였으나 두 가지 측면에서 명확히 구별된다. 첫째, 검색 키로 의미 유사도 대신 상황 임베딩을 1차로 사용하고 감정 필터링을 2차로 적용하는 검색 구조를 도입하여 상황적 맥락과 감정 정보를 모두 검색 조건에 반영한다. 둘째, APTNESS가 별도로 파인튜닝된 분류기로 전략을 예측하는 것과 달리, 본 연구는 별도의 전략 예측 모델을 두지 않고 검색된 예시에 이미 부여된 전략 라벨을 few-shot 문맥으로 그대로 활용한다. 이를 통해 검색된 예시의 전략을 명시적으로 노출하여 응답의 해석 가능성을 높인다.

1.3 Automatic Evaluation of Empathetic Dialogue and Comparison with Previous Studies

공감은 주관적이고 다차원적인 특성으로 인해 자동 평가가 어렵다. EMP-EVAL [10]은 감정 공감, 인지적 공감, 감정적 반응의 세 축을 정의하고, 인간 주석 없이 각 차원의 점수를 자동으로 산출하는 프레임워크를 제안하였다. 본 논문은 EMP-EVAL의 평가 체계를 기반으로 감정 반영·상황 해석·감정 일치 지표를 적용하며, 검색된 예시의 전략 라벨과 GT 전략 라벨 간의 일치도를 측정하는 strategy F1, 임베딩 유사도(sim_response_gt)를 추가로 구성한다. 다만, EMP-EVAL 역시 LLM 기반 평가라는 점에서 자기 선호·문체 선호·길이 편향 등의 한계를 가질 수 있으며, 본 연구는 이를 4,5절에서 명시적으로 논의하고 이종 평가자(Claude Haiku 4.5) 교차 검증을 수행한다.

Table 1은 주요 선행 연구들을 감정, 상황, 전략, 검색 기반 여부를 기준으로 비교한 것이다. 기존 연구들이 이 중 일부 요소만 선별적으로 다루었던 것과 달리, 본 연구는 세 요소를 모두 검색 키·문맥에 통합한 검색 증강 프레임워크를 제안한다. 다만 본 연구는 별도의 strategy prediction 모델을 학습하지 않으며, 검색된 예시에 부여된 전략 라벨을 그대로 활용한다는 점을 강조한다.

Table 1. Comparison with Previous Studies

Study	Core Approach	Emotion	Situation	Strategy	Retrieval Based
MoEL (EMNLP 2019)	Mixture of Emotion-Specialized Listeners	✓	✗	✗	✗
SDAM (2024)	Situation-Dialogue Relation Modeling	✓	✓	✗	✗
CASE (ACL 2023)	Commonsense-Affection Dual Alignment	✓	△**(commonsense)**	✗	✗
APTNESS (CIKM 2024)	Semantic Similarity-based RAG	✓	✓	✓	△**(semantic only)**
Proposed Method	Integrated RAG + Retrieved Strategy Prediction	✓	✓	✓	✓

III. The Proposed Scheme

1. Research Overview

본 연구의 전체 파이프라인은 Table 2와 같이 감정·상황 추출(Step 1), 2단계 검색(Step 2), 응답 생성(Step 3)의 세 단계로 구성된다.

Table 2. Framework pipeline

Stage	Component	Description
Step 1	Emotion and Situation Extraction	Automatically extract emotion labels and situation sentences from speaker utterances
Step 2	Two-Stage Retrieval	① Situation embedding-based FAISS retrieval (k=15) ② Emotion matching filtering refinement (k=3)
Step 3	Response Generation using Two-Stage Retrieval	Generate an empathetic response by integrating emotion, situation, and the strategy labels + texts of the 3 retrieved few-shot examples into the prompt

1.1 Dataset and Splits

본 연구는 한국어 공감 대화 데이터셋을 사용한다. 데이터셋은 AI Hub의 공감형 대화(웰니스 대화) 계열 말뭉치를 기반으로 구축되었으며, 각 샘플은 (화자 발화, 리스너 발화, 감정 라벨, 상황 문장, 공감 전략 라벨)로 구성된다.

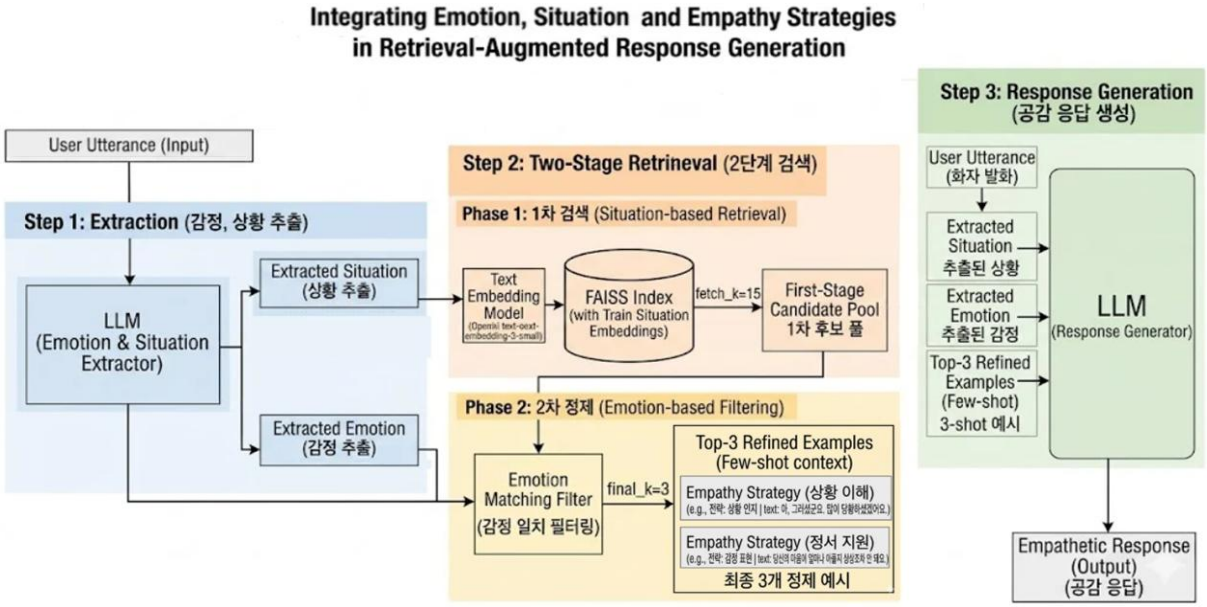

Fig. 1. Overall Pipeline of the Proposed Retrieval-Augmented Empathetic Response Generation Framework

총 4,801개의 공감 대화 쌍을 사용하며, 각 라벨의 성격과 생성 방식은 다음과 같다.

• 감정 라벨: 원본 말뭉치의 감정 주석을 Plutchik 감정 이론 [12] 기반의 6개 범주(분노, 슬픔, 불안, 상처, 당황, 기쁨)로 매핑하여 사용한다. 이 6개 범주는 영문 [anger, sadness, anxiety, hurt, embarrassment, joy]와 일대일로 대응한다.

• 상황 문장: 말뭉치에 상황 주석이 존재하는 경우 이를 사용하며, 평가 시점의 입력 발화에 대해서는 Step 1의 LLM 추출 결과(한 문장 요약)를 사용한다

• 공감 전략 라벨: 리스너 발화에 부여된 공감 전략 주석(예: 상황 인지, 감정 표현, 조언/제안, 격려 등)을 그대로 사용한다. 전략 라벨은 검색 저장소의 각 리스너 발화에 부착되어 few-shot 문맥으로 사용된다.

학습/평가 분할은 대화 단위(conversation-level)로 수행하여 동일하거나 유사한 발화가 train과 test에 중복되지 않도록 하였다. 전체 4,801개를 Train(80%, 3,840개) / Test(20%, 961개)로 분할하였으며, 분할은 고정 난수 시드(seed=42)로 수행한다. Train 데이터의 리스너 발화와 공감 전략 라벨을 FAISS 벡터 데이터베이스에 인덱싱하여 검색 저장소로 활용하고, Test 데이터의 화자 발화를 입력으로 하여 전체 파이프라인을 평가한다. 별도의 검증(validation) 세트는 하이퍼파라미터(k=15, final k=3)가 사전 고정되어 있어 사용하지 않았으며, 하이퍼파라미터 탐색은 수행하지 않았다.

2. Emotion and Situation Extraction

본 연구의 첫 단계는 화자 발화에서 감정과 상황을 추출하는 것이다. 감정 라벨은 Plutchik의 감정 이론 [12]을 참고하여 분노(anger), 슬픔(sadness), 불안(anxiety), 상처(hurt), 당황(embarrassment), 기쁨(joy)의 6개 범주로 제한하며, 상황은 감정이 발생한 맥락을 한 문장으로 요약한 텍스트로 정의한다. 본문·표·프롬프트의 감정 범주는 영문 [anger, sadness, anxiety, hurt, embarrassment, joy]와 국문 [분노, 슬픔, 불안, 상처, 당황, 기쁨] 순서로 일관되게 표기한다. 추출은 LLM 프롬프트를 통해 수행되며, 출력 형식을 JSON으로 고정하여 파싱 안정성을 보장한다.

Prompt 1. Emotion and Situation Extraction Prompt Structure

```
You are an expert in analyzing Korean empathetic dialogue.
Extract the following two items from the speaker's utterance below.

1. situation: Summarize the cause of the emotion in one sentence.
Example: "The grades improved." / "They do not understand my feelings."
2. emotion: [anger, sadness, anxiety, hurt, embarrassment, joy]

[Speaker Utterance] {speaker_text}
Output Format(JSON): {"situation": "...", "emotion": "..."}
```

2.1 Validation of the Extraction Module

추출 모듈의 출력(감정 라벨)은 2단계 검색의 필터링 조건으로 직접 사용되므로, 추출 정확도는 검색 품질과 최종 응답 품질 모두에 영향을 미치는 핵심 요소이다. 따라서 본 절에서는 추출 모듈의 성능을 파이프라인과 분리하여 독립적으로 평가한다. 정답 감정 라벨이 존재하는 Test 화자 발화 전체에 대해 추출 감정과 정답 라벨의 일치도를 측정한 결과, 정확도(Accuracy) 0.634, Macro-F1 0.633을 얻었다(6개 범주 무작위 기준 약 0.17 대비 크게 우수). 클래스별로는 분노(F1 0.69)·기쁨(0.65)이 높고 슬픔(0.57)이 가장 낮았으며, 혼동은 주로 분노↔기쁨, 슬픔↔상처 등 정서가가 인접한 범주에서 발생하였다. 상황 추출 품질은 무작위 100건에 대한 GPT-4o-mini 충실도 평가(1-5점)에서 평균 4.64점(5점 64건·4점 36건)을 얻어, 실용 수준의 신뢰도를 가짐을 확인하였다.

추출 오류가 검색·응답 품질에 미치는 영향: 감정 추출이 오류일 경우 2차 감정 필터링이 부적절한 예시를 선택하여 검색 품질이 저하되고, 이는 최종 응답 품질로 전파될 수 있다. 본 연구는 이러한 오류 전파를 완화하기 위해, 필터링 결과 감정 일치 예시가 3개 미만일 때 1차 상황 유사도 순위로 폴백(fallback)하여 3개를 채우는 규칙을 두었다(3.2절 참조). 이를 통해 감정 추출이 틀리더라도 상황 유사도가 높은 예시가 확보되므로, 추출 단계의 오류가 최종 응답의 실패로 직결되지 않도록 설계하였다.

3. Two-Stage Retrieval

검색 단계는 상황 임베딩 기반 1차 검색과 감정 필터링 기반 2차 정제로 구성된 2단계 구조를 채택한다.

3.1 First-Stage Retrieval: Situation Embedding Based Similarity Search

Train 데이터의 각 리스너 발화에 대응하는 상황 문장을 OpenAI의 text-embedding-3-small 모델로 임베딩하여 FAISS 인덱스를 구축한다. 이 모델은 다국어 의미 표현에서 검증된 성능을 보이면서도 비용 효율적이어 짧은 상황 요약 문장 수준의 유사도 검색에 적합하다.

구현 세부사항(재현성): 임베딩 차원은 1,536차원(text-embedding-3-small의 기본 차원)이며, FAISS 인덱스는 IndexFlatL2를 사용한다. 거리 척도는 L2 거리를 사용하며, 임베딩은 OpenAI API가 반환하는 정규화(L2-normalized) 벡터를 그대로 사용한다. 1차 후보 수 k=15는 감정 필터링 후에도 최종 3개를 안정적으로 확보하기 위한 경험적 선택값으로, k∈{5,10,15,20}에 대한 예비 실험에서 k=15가 필터링 후 3개 확보 실패율을 충분히 낮추면서 계산 비용을 과도하게 늘리지 않는 지점으로 확인되었다. 입력 발화에서 추출된 상황 문장을 동일 모델로 변환한 뒤 가장 유사한 k=15개 후보를 검색한다.

3.2 Second-Stage Refinement: Emotion Matching Filtering

1차 검색으로 수집된 k개의 후보 중 추출된 감정 라벨과 동일한 감정을 가진 예시를 우선적으로 선별한다. 후보 중 감정 일치하는 예시를 선별하여 최종 3개를 구성하며, 각 예시는 공감 전략 라벨과 해당 리스너 발화 텍스트를 포함하여 응답 생성 단계의 few-shot 문맥으로 활용된다. 이를 통해 기존의 단순 유사도 검색이나 감정 단독 필터링 대비 더 정교한 공감 예시 선별이 가능하다. 감정 일치 예시가 3개 미만인 경우, 부족한 개수를 1차 검색 후보(감정 불일치 포함) 중 상황 유사도(L2 거리) 상위 순으로 보충하여 항상 3개를 구성한다.

4. Empathetic Response Generation

검색된 3개의 예시를 few-shot 문맥으로 활용하여 공감 응답을 생성한다. 프롬프트는 역할 설정, 입력 정보(감정·상황·전략), few-shot 예시 3개, 생성 지시문으로 구성되며, 출력은 자연스러운 한국어 한 문장으로 제한한다. 또한 Prompt 2의'Strategy' 필드는 오직 검색된 3개 예시에 부여된 전략 라벨로부터만 유래한다. 즉, test 쌍의 정답(ground-truth) 전략 라벨은 응답 생성 입력에 결코 사용되지 않으며, 별도의 전략 예측 모듈도 존재하지 않는다. 검색된 3개 예시의 전략이 서로 다를 경우, few-shot 블록은 각 예시의 전략을 개별로 명시하고(예: 1. Strategy:…/2. Strategy:…), 상단의 대표'Strategy' 필드에는 최근접(situation 유사도 1위) 예시의 전략을 대표값으로 제공한다. 따라서 생성 단계에 사용되는 모든 입력 정보는 test 시점에 실제로 이용가능한 정보이다.

Prompt 2. Response Generation Prompt Structure

```
You are an expert in Korean empathetic dialogue.
Based on the following four elements, generate an
empathetic listener response to the speaker.

- Emotion: {emotion} #
- Situation: {situation}
- Strategy: {strategy}
- few-shot 예시3가지:
  1. Strategy: {strategy_1}  |  text: {listener_text_1}
  2. Strategy: {strategy_2}  |  text: {listener_text_2}
  3. Strategy: {strategy_3}  |  text: {listener_text_3}

Listener Response (Write one Korean sentence:
```

5. Experimental Conditions

제안 프레임워크의 각 구성 요소의 기여도를 분리하여 검증하기 위해 Table 3와 같이 다섯 가지 검색 조건을 설계하였다.

Table 3. Ablation Study on Retrieval Condition Design

Condition	Retrieval Method	Description
no_retrieval	No Retrieval	Generate responses using only emotion and situation
situation_ only	Situation Embedding Retrieval	Select top-3 examples based on situation similarity
emotion_ only	Random Sampling within Emotion Group	Randomly select 3 examples from the same emotion category (reported as mean ± sd over 5 seeds; overall sd ±0.019)
concat	Concatenated Situation+Emotion Retrieval	Retrieve using a single query in the form of "Situation [Emotion]"
filter(제안)	Two-Stage Retrieval	First-stage situation retrieval → second-stage emotion filtering refinement

IV. Experiment Result

1. Quantitative Study

본 절에서는 다섯 가지 검색 조건을 세 가지 LLM 환경에서 비교한 정량적 분석 결과를 정리한다. 실험에 사용된 모델은 GPT-4o-mini, NAVER CLOVA HCX-DASH-002, Qwen2.5 7B이다. GPT-4o-mini는 OpenAI의 범용 경량 모델로, 다국어 공감 응답 생성의 기준 성능을 대표하며, NAVER CLOVA HCX-DASH-002는 한국어 특화 상용 모델로 한국어 공감 대화의 도메인 적합성을 검증하기 위해 선정하였다. Qwen2.5 7B는 오픈소스 다국어 모델로, 파인튜닝 없이 한국어 공감 응답 생성이 가능한지를 확인하는 비교군으로 활용한다. 평가의 일관성을 위해 EMP-EVAL 점수 산출은 모든 모델의 응답에 대해 GPT-4o-mini를 단일 평가 모델로 통일하여 적용한다.

1.1 Evaluation Metrics

본 연구의 평가 체계는 공감 응답의 다양한 측면을 측정하기 위해 세 범주의 지표로 구성된다. 지표 표기는 본문·표·그림에서 다음 약어로 통일한다.

• EMP-EVAL 공감 품질 지표: EMP(score_overall, 0-5) 는 종합 공감 점수, ER(emotional_reaction, 0-5)은 정서적 반응의 적절성을 측정한다.

• 응답 유사도 SIM(sim_response_gt): 생성 응답과 정답(Ground Truth) 응답 간의 코사인 유사도.

• 전략 매칭 지표: 검색된 예시의 전략 라벨과 정답 전략 라벨 간 일치도를 F1(strategy_f1, 집합 P/R의 조화평균), Jac(strategy_jaccard, 집합 중복 비율), Hit(strategy_hit, 정답 전략 1개 이상 포함 여부)로 측정한다. 세 지표 모두 값이 높을수록 좋으며 Hit > F1 > Jac 순으로 관대하다.

종합 지표 Comp(composite): 세 범주를 통합한 점수로 Comp = (EMP/5 + SIM + F1) / 3으로 산출한다. 세 지표는 척도가 서로 다르므로(EMP는 0-5, 나머지는 0-1) EMP를 5로 나누어 [0,1]로 정규화한 뒤 동일 가중치로 평균한다. 동일 가중치는 특정 측면을 우선하지 않는 중립적 선택이며, 가중치 민감도는 1.4절에서 별도로 분석한다. Comp는 절대적 품질 척도가 아니라 조건 간 상대 비교를 위한 지표로 해석한다.

1.2 Overall Composite Score & Ablation Study

Table 4는 세 모델 × 다섯 조건에 대한 전체 지표를 수치표로 제시한 것이다.

Table 4. Full Quantitative Results across Three LLMs and Five Retrieval Conditions

Model	Condition	EMP	SIM	ER	F1	Jac	Hit	Comp
GPT-4o-mini	no_retrieval	3.903	0.374	4.254	0.000	0.000	0.000	0.385
	situation_ only	4.137	0.382	4.464	0.315	0.255	0.461	0.508
	emotion_ only*	4.048	0.380	4.372	0.341	0.278	0.497	0.510
	concat	4.129	0.381	4.461	0.327	0.268	0.465	0.511
	filter (proposed)	4.133	0.382	4.466	0.328	0.267	0.473	0.512
CLOVA HCX-DASH-002	no_retrieval	3.268	0.254	3.495	0.000	0.000	0.000	0.303
	situation_ only	3.399	0.258	3.585	0.123	0.099	0.183	0.354
	emotion_ only*	3.334	0.258	3.506	0.121	0.096	0.182	0.348
	concat	3.370	0.256	3.548	0.113	0.091	0.164	0.347
	filter (proposed)	3.391	0.254	3.574	0.122	0.099	0.179	0.351
Qwen 2.5 7B	no_retrieval	3.553	0.360	3.719	0.000	0.000	0.000	0.357
	situation_ only	3.602	0.356	3.755	0.308	0.250	0.451	0.461
	emotion_ only*	3.534	0.353	3.674	0.302	0.241	0.453	0.454
	concat	3.584	0.357	3.733	0.304	0.247	0.441	0.459
	filter (proposed)	3.594	0.356	3.752	0.308	0.249	0.451	0.461

모든 검색 조건은 no_retrieval 대비 유의미하게 우수하였다(filter vs no_retrieval: Wilcoxon $p < 0.001$, 중앙 차이 +0.050). 반면 filter와 situation_only 사이의 차이는 통계적으로 유의하지 않았다($p = 0.488$, 중앙 차이 0.000). 따라서 본 결과는 '검색 기반 조건이 무검색 대비 일관되게 개선되었음'을 뒷받침하며, filter의 우위는 일부 모델·지표에서 나타나는 제한적·국소적 우위로 해석해야 한다.

1.3 Per-Metric Analysis and Retrieval Quality

Fig. 2는 다섯 가지 검색 조건의 종합 점수(Comp)를 세 모델별로 비교한 막대그래프이다. 모든 모델에서 검색을 사용하지 않은 no_retrieval이 가장 낮고, 네 가지 검색 조건(filter, situation_only, concat, emotion_only)은 서로 근접한 상위 구간에 분포한다. 이는 검색 증강 자체의 기여가 검색 키 설계보다 크다는 점을 보여준다.

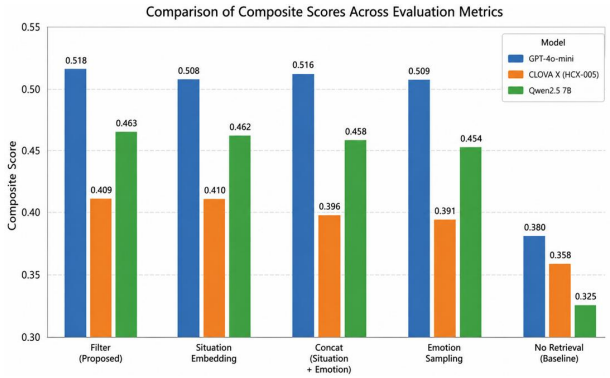

Fig. 2. Comparison of Composite Performance across Retrieval Conditions

또한 세 모델 모두 no_retrieval의 Comp가 가장 낮아 검색 증강의 일관된 기여가 확인되며, baseline(situation_only·concat·emotion_only) 대비로 해석한다. 검색 품질 지표에서 filter는 CLOVA 환경의 strategy F1(0.253)과 Jaccard(0.205)에서 단독 1위를 기록하여 감정 필터링이 전략 매칭 정합성에 기여함을 보인다. 반면 emotion_only는 GPT-4o-mini 환경에서 F1·Jac·Hit가 높으나 EMP·SIM이 낮아 종합 점수에서는 filter·situation_only에 미치지 못한다. 즉 감정 일치 검색은 전략 매칭에 유리하지만 상황 정합성이 약하면 응답 품질로 이어지지 않으며, filter는 상황 검색(1차)과 감정 필터(2차)를 결합하여 두 축의 균형을 꾀한다. 다만 종합 점수 기준으로 filter와 situation_only의 우열은 모델·지표에 따라 갈리며, 평균적으로는 통계적으로 구분되지 않는다. 응답 유사도(SIM)와 전략 제어 가능성(strategy F1) 측면에서도 filter와 situation_only는 사실상 동등하며, no_retrieval만 F1=0으로 뚜렷이 구분된다.

1.4 Effect of Retrieval Augmentation

filter의 no_retrieval 대비 세 모델 평균 향상폭은 EMP +0.129, SIM +0.007, ER +0.101, F1 +0.296, Jac +0.234, Hit +0.432, Comp +0.109로, 전략 검색 관련 지표(F1·Jac·Hit)에서 향상 폭이 가장 크다. 이는 전략 라벨이 부착된 예시의 few-shot 활용이 검색 기반 지표를 직접 끌어올림을 보여준다.

1.5 Sensitivity Analysis of Composite Weights

종합 지표 Comp의 동일 가중치 타당성을 점검하기 위해 EMP·SIM·F1의 가중치를 균등·EMP 중심·SIM 중심·F1 중심으로 변화시키며 세 모델 평균 순위를 재계산한 결과, 모든 가중치 조합에서 네 검색 조건이 no_retrieval보다 상위라는 핵심 결론은 유지되었으며, filter와 situation_only의 1·2위만 가중치에 따라 근소하게 교체되었다. 따라서 동일 가중치 Comp는 검색 증강의 효과를 요약하는 데 견고하나, 두 조건의 미세 우열 판정 근거로는 사용하지 않는다.

1.6 Cross-Evaluator Reliability Check

핵심 결론이 단일 LLM 기반 자동 평가(EMP-EVAL, GPT-4o-mini)에 의존한다는 우려를 보완하기 위해, 생성 모델과 다른 계열의 평가자(Claude Haiku 4.5)를 도입한 교차 평가를 수행하였다. GPT-4o-mini가 생성한 응답 400건(다섯 조건 각 80건)을 동일한 EMP-EVAL 기준으로 재채점하여 두 평가자의 점수 일치도를 분석하였다.

Table 5. GPT-4o-mini-Claude Haiku 4.5 Cross-Evaluation (400 Responses)

Condition	GPT-4o-mini (EMP)	Claude Haiku 4.5 (EMP)	Δ (GPT − Claude)
filter (proposed)	4.119	3.266	+0.853
situation_only	4.184	3.309	+0.875
emotion_only	4.059	3.175	+0.884
concat	4.072	3.244	+0.828
no_retrieval	3.919	3.281	+0.638

응답 단위에서 두 평가자의 점수는 양의 상관(Spearman ρ=0.62, Pearson r=0.63)을 보여, 어떤 응답이 더 공감적 인지에 대한 판단은 평가자 계열이 달라도 대체로 일치하였다. 다만 GPT-4o-mini의 평균 점수

(4.07)가 Claude(3.25)보다 일관되게 높아, 생성과 평가가 동일 계열일 때 절대 점수가 다소 후하게 매겨지는 자기선호 경향이 관찰되었다. 다섯 조건의 점수 차이는 두 평가자 모두 0.2점 내외로 작아 조건 간 미세 순위는 표본 노이즈에 민감하며(조건 평균 ρ=0.40), 이는 §1.3의 통계 분석과 일관된다. 따라서 자동 평가는 '검색 vs 무검색' 같은 큰 경향 해석에는 견고하나, 미세한 조건 우열의 단독 근거로는 사용하지 않으며, LLM-as-a-judge의 한계와 인간 평가의 필요성은 한계 절에서 명시한다.

1.7 Positioning Relative to Prior Methods

MoEL·MIME·CEM·CASE·SDAM등은 Empathetic Dialogues를 대상으로 파인튜닝된 영어 전용 모델로, 본 연구의 한국어·파인튜닝 없는 RAG 설정과 데이터·언어·학습 패러다임이 달라 동일 수치의 직접 비교는 공정하지 않다. 가장 가까운 비교 대상인 APTNESS [8] 역시 한국어 공감 대화에 맞춰 제공되지 않아 동일 조건 재현이 어렵다. 따라서 본 연구는 APTNESS의 핵심 설계인 '의미 유사도 단일 검색 + 별도 파인튜닝 전략 분류기'를 본 프레임워크 내에서 근사 재현한 조건(situation_only가 의미 유사도 단일 검색에 해당)을 강한 baseline으로 두고, 여기에 감정 필터링 단계를 추가한 filter와 비교하는 방식으로 기여를 검증하였다. 그 결과는 앞서 보인 대로 filter가 일부 모델·지표(특히 CLOVA의 전략 매칭)에서 우위를 보이나 종합 성능에서는 강한 검색 baseline과 통계적으로 동등한 수준이다. 즉 본 연구의 기여는 파인튜닝된 전략 분류기 없이도 검색된 예시의 전략 라벨만으로 경쟁력 있는 전략 제어가 가능함을 보인 것과 검색 키에 감정·상황을 명시적으로 결합하는 해석 가능한 구조를 제시한 것에 있다.

2. Qualitative Study

본 절에서는 동일한 화자 발화에 대해 다섯 가지 검색 조건별로 선택된 3-shot 예시, 생성 응답, 정답 응답을 Table 6에 나란히 비교하여, 3-shot 예시가 생성 응답에 실제로 어떻게 반영되는지, 그리고 검색 조건에 따라 생성 응답과 정답 간의 괴리가 어떻게 달라지는지를 직관적으로 보인다. 사례로 사용한 화자 발화는 "일상생활을 하면서 문득 쌓인 감정이 터져나와"이며, 추출 감정은 분노, 추출 상황은 "일상생활 중 쌓인 감정이 터져나와 힘든 상황"이다. 정답(GT) 감정은 분노, 정답 전략은 '동조', 정답(GT) 리스너 응답은 "네 감정이 나한테까지 전해져서 나도 마음이 좋지 않다. 분한 감정은 “스스로를 갉아먹게 될까봐 걱정돼."이다(GPT-4o-mini 실제 로그).

Table 6. Condition-wise Retrieved Examples and Generated Responses for the Same Utterance

Condition	Retrieved 3-shot examples (strategy | text)	Generated response
no_retrieval	(검색없음– 감정·상황만입력)	정말 괴롭겠어요.
situation_only	슬픔·위로,동조/ 슬픔·조언,격려/ 슬픔·조언,격려 (상황유사, 감정불일치)	일상에서 쌓인 감정을 터뜨리다니 정말 힘들었겠다.
emotion_only	분노·동조/ 분노·격려/ 분노·조언,격려 (동일감정 무작위, 상황불일치)	그 마음 충분히 이해돼, 정말 힘들었겠다.
concat	분노·동조/ 슬픔·위로,동조/ 분노·조언,격려 (상황+감정 결합질의)	그동안 쌓인 감정이 결국 터져 나온 거였구나. 정말 많이 힘들었겠다.
filter (proposed)	분노·위로/ 분노·동조/ 분노·동조(상황1차+감정2차)	그런 상황에서 쌓인 감정이 한꺼번에 터져 나왔구나. 얼마나 답답하고 힘들었을지 충분히 이해돼.

no_retrieval은 검색 예시가 없어 일반적 위로에 그치고 전략 일치가 없다. situation_only는 상황은 유사하나 감정이 어긋난(슬픔) 예시를 가져와 핵심 감정(분노)과의 정합이 약하다. emotion_only는 동일 감정(분노) 예시를 무작위로 선택하므로 전략 라벨은 정답(동조)에 부합하나, 상황 정합이 약해 다소 일반적인 응답에 그친다. concat은 상황과 감정을 한 질의로 결합해 검색하여 situation_only와 유사한 경향을 보인다. 반면 제안 방식 filter는 상황 유사성을 1차로 확보한 뒤 감정이 일치하는(분노) 예시를 선별하여 정답 감정·전략에 더 부합하는 응답을 생성하며, 생성 응답과 정답 간 괴리가 가장 작다. 다만 검색 조건 간 미세 우열은 사례에 따라 달라지며, 이는 검색 조건들이 정량적으로 통계적으로 구분되지 않는다는 결론과 일치한다.

V. Conclusions

본 연구는 화자 발화에서 감정과 상황을 자동 추출하고, 상황 임베딩 1차 검색과 감정 필터링 2차 정제로 구성된 2단계 검색을 통해 공감 전략이 명시된 예시를 선별하여, 그 전략 라벨을 별도 학습 모듈 없이 few-shot 문맥으로 활용하는 RAG 기반 공감 응답 생성 프레임워크를 제안하였다. 감정·상황·전략을 하나의 해석 가능한 검색 구조로 통합한 점이 핵심 기여이다.

세 LLM(GPT-4o-mini·CLOVA HCX-DASH-002·Qwen2.5 7B) 환경의 다섯 조건 비교에

서 검색 기반 조건은 무검색 대비 일관되게 우수하였다 (Wilcoxon $p < 0.001$). 본 연구는 filter를 '항상 최우수'가 아니라, 파인튜닝된 전략 분류기 없이 검색된 예시의 전략 라벨만으로 경쟁력 있는 전략 제어를 달성하는 해석 가능한 설계로 자리매김한다.

한계로는 (i) 평가가 주로 LLM 자동 지표에 의존하여 인간 평가를 수행하지 못하고 이종 평가자(Claude Haiku 4.5) 교차 검증으로 보완한 점, (ii) 단일 한국어 데이터셋·단일 턴에 한정된 점, (iii) filter와 situation_only의 통계적 동등성이 감정 필터링의 추가 이득이 제한적일 수 있음을 시사하는 점을 들 수 있다. 향후 인간 평가를 포함한 평가 다각화, 감정 강도·양가성 기반 가중 필터링, 다중 턴 및 다양한 데이터셋으로의 확장을 통해 감정 필터링의 기여를 더 명확히 규명할 계획이다.

REFERENCES

[1] H. Rashkin, E. M. Smith, M. Li, and Y. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 5370–5381, Florence, Italy, 2019. https://doi.org/10.18653/v1/P19-1534

[2] Z. Lin, A. Madotto, J. Shin, P. Xu, and P. Fung, "MoEL: Mixture of empathetic listeners," in Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 121–132, Hong Kong, China, 2019. https://doi.org/10.18653/v1/D19-1012

[3] N. Majumder, P. Hong, S. Peng, J. Lu, D. Ghosal, A. Gelbukh, R. Mihalcea, and S. Poria, "MIME: MIMicking emotions for empathetic response generation," in Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8968–8979, Online, 2020. https://doi.org/10.18653/v1/2020.emnlp-main.721

[4] S. Sabour, C. Zheng, and M. Huang, "CEM: Commonsense-aware empathetic response generation," in Proc. of the 36th AAAI Conference on Artificial Intelligence, Vol. 36, No. 10, pp. 11229–11237, 2022. https://doi.org/10.1609/aaai.v36i10.21373

[5] Y. Qian, X. Zheng, T. Hashimoto, J. Zhou, Y. Zhu, J. Qian, M. Kan, and H. T. Ng, "Empathetic response generation via emotion cause transition graph," in Proc. of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, Rhodes Island, Greece, 2023. https://doi.org/10.1109/ICASSP49357.2023.10095652

[6] J. Zhou, C. Zheng, B. Wang, Z. Zhang, and M. Huang, "CASE: Aligning coarse-to-fine cognition and affection for empathetic response generation," in Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Vol. 1, pp. 8223–8237, Toronto, Canada, 2023. https://doi.org/10.18653/v1/2023.acl-long.457

[7] Z. Yang, C. Wang, Y. Sun, X. Zhu, Z. Chen, T. Cai, Y. Wu, Y. Su, S. Ju, and X. Liao, "Situation-aware empathetic response generation," Information Processing and Management, Vol. 61, No. 5, 103795, 2024. https://doi.org/10.1016/j.ipm.2024.103795

[8] Y. Hu, M. Tan, C. Zhang, Z. Li, X. Liang, M. Yang, C. Li, and X. Hu, "APTNESS: Incorporating appraisal theory and emotion support strategies for empathetic response generation," in Proc. of the 33rd ACM International Conference on Information and Knowledge Management (CIKM), pp. 889–898, Boise, Idaho, USA, 2024. https://doi.org/10.1145/3627673.3679687

[9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," Advances in Neural Information Processing Systems (NeurIPS), Vol. 33, pp. 9459–9474, 2020. https://doi.org/10.48550/arXiv.2005.11401

[10] B. Amjad and S. A. Rauf, "EMP-EVAL: A framework for measuring empathy in open domain dialogues," arXiv preprint arXiv:2301.12510, 2023. https://doi.org/10.48550/arXiv.2301.12510

[11] M. H. Davis, "Measuring individual differences in empathy: Evidence for a multidimensional approach," Journal of Personality and Social Psychology, Vol. 44, No. 1, pp. 113–126, 1983. https://doi.org/10.1037/0022-3514.44.1.113

[12] R. Plutchik, Emotion: A Psychoevolutionary Synthesis, Harper & Row, New York, 1980.

[13] H. Cai, X. Shen, Q. Xu, W. Shen, X. Wang, W. Ge, X. Zheng, and X. Xue, "Improving empathetic dialogue generation by dynamically infusing commonsense knowledge," in Findings of the Association for Computational Linguistics: ACL 2023, pp. 7858–7873, Toronto, Canada, 2023. https://doi.org/10.18653/v1/2023.findings-acl.498

[14] A. Sharma, A. S. Miner, D. C. Atkins, and T. Althoff, "A computational approach to understanding empathy expressed in text-based mental health support," in Proc. of EMNLP, pp. 5263-5276, Online, 2020. https://doi.org/10.18653/v1/2020.emnlp-main.425

[15] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in Proc. of ICLR, 2020. https://doi.org/10.48550/arXiv.1904.09675

[16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," Advances in Neural Information Processing Systems (NeurIPS), Vol. 35, pp.

24824-24837, 2022. https://doi.org/10.48550/arXiv.2201.11903

[17] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," Advances in Neural Information Processing Systems (NeurIPS), Vol. 36, 2023. https://doi.org/10.48550/arXiv.2306.05685

[18] National Information Society Agency (NIA), "AI Hub: Empathetic dialogue corpus," AI Hub Open Dataset, 2022. https://www.aihub.or.kr

Authors

So-Jeong Park received the B.S. degree in Computer Science from Sungkyunkwan University, Suwon, South Korea. She is currently pursuing the M.S. degree in the Graduate School of Information at Yonsei University, Seoul, South Korea.

Her research interests include Natural Language Processing, Large Language Models, AI Agents.

Beak-Cheol Jang received the Ph.D. degrees in Computer Science and Engineering from North Carolina State University, United States, in 2009. Dr. Jang joined the faculty of Graduate School of Information at Yonsei University, Korea, in 2021. He is currently a Professor. His research interests are Wireless Networking and Artificial Intelligence.