

# 전문 지식 및 대화 정책 예측이 결합된 프롬프트를 활용한 지식 기반 대화 생성

주어진<sup>o,1</sup>, 임채균<sup>1</sup>, 이도경<sup>2</sup>, 윤준영<sup>2</sup>, 성주원<sup>2</sup>, 최호진<sup>1</sup>  
한국과학기술원 전산학부<sup>1</sup>, KT AI2XL연구소<sup>2</sup>

{akwkek11, rayote, hojinc}@kaist.ac.kr, {dokyung.lee, junyoung.youn, jwsung}@kt.com

## Knowledge-Grounded Dialogue Generation Using Prompts Combined with Expertise and Dialog Policy Prediction

Eojin Joo<sup>o,1</sup>, Chae-Gyun Lim<sup>1</sup>, DoKyung Lee<sup>2</sup>, JunYoung Youn<sup>2</sup>, Joo-Won Sung<sup>2</sup>, Ho-Jin Choi<sup>1</sup>  
KAIST School of Computing<sup>1</sup>, KT AI2XL<sup>2</sup>

### 요약

최근 지식 기반 대화 생성에 많은 연구자가 초점을 맞추고 있다. 특히, 특정 도메인에서의 작업 지향형 대화 시스템을 구축하는 것은 다양한 도전 과제가 있으며, 이 중 하나는 거대 언어 모델이 입력과 관련된 지식을 활용하여 응답을 생성하는 데 있다. 하지만 현재 거대 언어 모델은 작업 지향형 대화에서 단순히 정보를 열거하는 방식으로 응답을 생성하는 경향이 있다. 이 논문에서는 전문 지식과 대화 정책 예측 모델을 결합한 프롬프트를 제시하고 작업 지향형 대화에서 사용자의 최근 입력에 대한 정보 제공 및 일상 대화를 지원하는 가능성을 탐구한다. 이러한 새로운 접근법은 모델 파인튜닝에 비해 비용 측면에서 효율적이며, 향후 대화 생성 분야에서 발전 가능성을 제시한다.

**주제어:** 거대 언어 모델, 작업 지향형 대화, 정보 검색, 지식 기반 문장 생성, 프롬프트 공학

### 1. 서론

최근 자연어 처리 분야에서 지식 기반 문장 생성에 많은 연구와 개발이 진행 중이다 [1]. 특히, 특정 도메인에서의 작업 지향형 대화 시스템을 구축하는 것에 있어 지식 검색 혹은 마치 사람과 대화하는 듯한 자연스러운 문장의 생성 등 많은 도전 과제를 동반하고 있다. 최근 작업 지향형 대화에서 지식을 포함한 문장을 생성할 때 거대 언어 모델(LLM)들[2, 3]을 활용하려는 연구들이 활발히 이루어지는 중이다 [4, 5]. 하지만 현재 주로 사용되는 거대 언어 모델은 작업 지향형 대화 내에서 답변을 생성할 때, 단순히 정보를 열거하는 방식으로 응답한다. 또한 다음의 그림 1을 보면 알 수 있듯이, 공감을 요구하는 상황에 있어 거대 언어 모델은 이러한 상황을 해결하기 위한 여러 방법을 열거식으로 제시하는 문장을 생성하는 것을 알 수 있다.

이러한 문제를 해결하기 위해 최근에는 프롬프트 공학(Prompt Engineering)을 활용하여 원하는 결과를 얻을 수 있도록 문장 생성을 제어한다 [6]. 프롬프트를 이용하면 파인튜닝(Fine-Tuning)하는 것보다 비용적인 측면에서 월등히 저렴하다. 최근에는 거대 언어 모델을 적은 비용과 컴퓨터 자원으로 파인튜닝을 할 방법들이 제시됐지만, 추론 시 시간이 굉장히 오래 걸리기 때문에 이를 활용하여 서비스를 운영하기엔 한계점이 존재한다 [7].

이 논문에서는 작업 지향형 대화 생성에 전문 지식과 대화 정책 예측 모델을 결합한 프롬프트를 바탕으로 자연스러운 대화 생성의 가능성을 제시한다. 이를 통해 작업 지향형 대화에서 사용자의 최근 입력에 대해 정보 제공 혹은 일상 대화를 최근 상담에 맞춰 번갈아 가면서 생성함을 보인다.

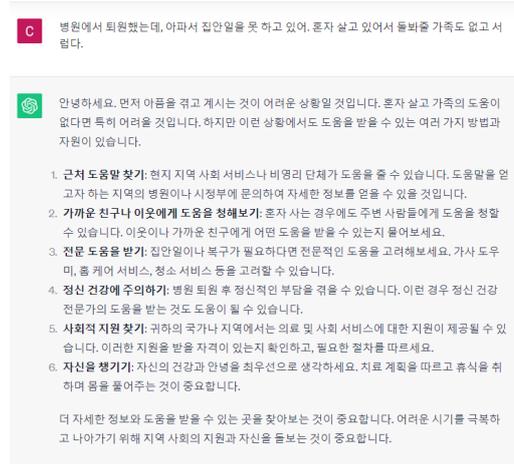


그림 1. 공감 요구 상황에 대한 ChatGPT [2]의 응답 예시

### 2. 관련 연구

#### 2.1 지식 기반 문장 생성

지식 기반 문장 생성 모델의 예시로는 Retrieval Augmented Generation (RAG) [8]와 Fusion-in-Decoder (FiD) [9]등이 있다. 최근에는 거대 언어 모델을 추가적인 파인튜닝 없이 활용하려는 시도들[4, 5]이 있으나 특정 도메인에서의 지엽적인 정보에 대해선 환각 현상을 보인다 [10]. 이를 해결하기 위해 생성 모델과 별개로 정보 검색 모델을 따로 학습한다 [11]. 전통적인 정보 검색 모델은 입력으로 들어온 질문과 가장 유사한 문서 목록을 추출하는 TF-IDF, BM25 [12]와 같은 단어 기반 유사도를 사용했으나 최근에는 DPR (Dense Passage Retrieval) [13]과 같은 언어 모델 기반의 검색 모델을 많이 활용한다.

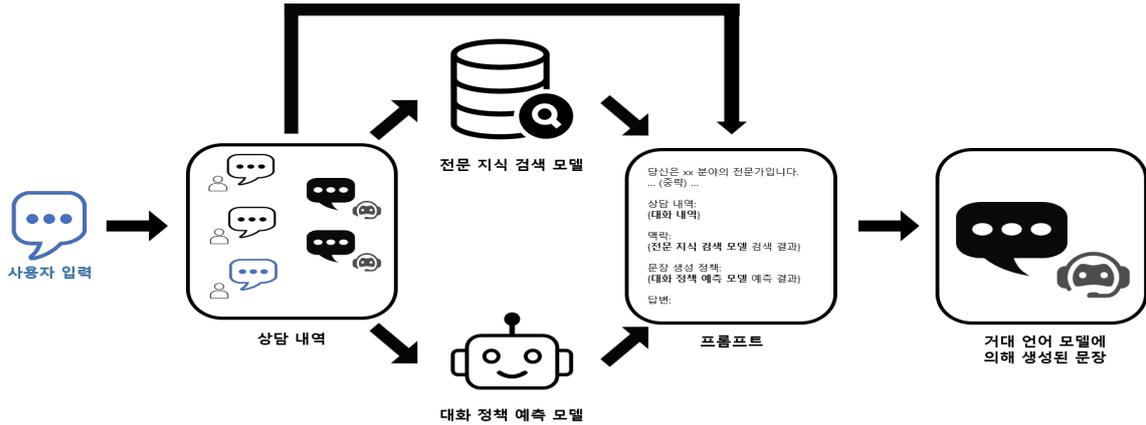


그림 2. 전문 지식과 대화 정책 예측 모델을 결합한 제안 기법

## 2.2 프롬프트 공학

프롬프트 공학은 생성형 모델을 제어하기 위한 지시사항을 모델이 잘 이해할 수 있도록 가공하는 분야이며, 최근 거대 언어 모델의 발달과 함께 부각되고 있다 [6]. 적절한 프롬프트를 사용하고 이를 조정하여 모델이 원하는 방식으로 작동하도록 도울 수 있는데, 이는 만족하는 결과를 얻기까지 프롬프트 작성과 수정의 반복을 통해 이루어진다. 최근에는 거대 언어 모델을 활용해 자동으로 프롬프트를 생성하는 사례가 있으며, 이는 인간이 직접 프롬프트를 만드는 것과 비슷한 성능을 보인다고 알려져 있다 [14].

## 3. 제안 기법

이 논문에서 제안한 방법은 그림 2와 같다. 사용자의 현재 입력은 과거의 사용자 입력과 모델이 생성한 문장으로 이루어진 상담 내역에 저장되며, 상담 내역의 가장 최근 문장 3개를 입력으로 받는 전문 지식 검색 모델(DPR) [13]의 검색 결과와 대화 정책 예측 모델의 예측 결과를 프롬프트로 결합해서 거대 언어 모델에게 입력으로 제공한다.

### 3.1 전문 지식 검색 모델

전문 지식 검색 모델은 사용자의 질문에 대해 가장 관련 있는 구절을 찾아주는 모델이다. 이 논문에서는 DPR [13]를 활용했다. DPR은 밀집 인코더를 질문 인코더와 구절 인코더 총 2개를 사용하며, 이 둘을 각각  $E_P$ 와  $E_Q$ 라고 칭하겠다. DPR은 입력으로 받은 질문  $q$ 를  $d$ 차원의 벡터( $E_Q(q)$ )로 변환한 후, 질문 벡터와 가장 유사한  $k$ 개의 구절 벡터( $E_P(p)$ )를 검색하는 작업을 수행한다. 여기서 질문과 구절의 유사도를 벡터의 내적과 유사한 행렬 곱셈으로 다음과 같이 표현할 수 있다.

$$sim(q, p) = E_Q(q)^T E_P(p) \quad (1)$$

이 논문에서는 질문 인코더와 구절 인코더 모두 KLUE 데이터로 사전학습한 RoBERTa를 사용했다 [15, 16]. 질문과 구절의

임베딩 벡터는 BertPooler에서 반환하는 마지막 은닉 상태의 첫번째 토큰인 [CLS]의 임베딩을 가져와서 사용했다. 여기에 그림 3과 같이 질문에 들어갈 입력에 가장 최근의 사용자 입력뿐만이 아닌 이전의 상담 내역까지 입력 질문에 포함시키는 방식으로 (질문, 구절) 쌍을 늘려 DPR을 학습할 때의 데이터를 증강했다. 총 4가지 경우에 대해 각각 학습을 진행했으며 각각 (1) 가장 최근 사용자의 입력만을 고려하는 1-Turn, (2) 가장 최근의 상담 내역 세 발화를 고려하는 3-Turn, (3) 가장 최근 사용자의 입력과 상담 내역 두 발화를 전부 고려하는 1, 2-Turn, 마지막으로 (2)와 (3)을 합한 경우에 대해 각각 학습을 진행했다.

### 3.2 대화 정책 예측 모델

대화 정책 예측 모델은 입력으로 상담 내역의 일부가 들어가고, 출력의 경우 다음에 생성할 문장의 생성 정책이 일상적인 대화(공감)인지 사실을 제공(사실제공)해야 하는지를 예측한 결과를 반환한다. 그림 4는 모델이 상담 내역의 세 발화를 입력으로 받아 다음 발화의 대화 정책을 예측하는 것을 시각화한 것이다. 이 모델 역시 전문 지식 검색 모델과 동일한 RoBERTa를 사용했다 [15, 16].

### 3.3 모델들의 출력을 결합하는 프롬프트

이 논문에서 제안한 기법에 따라 상담 도메인에서의 문장 생성 결과 비교 실험을 진행하기 위해 총 2개의 프롬프트를 만들었다.

1. 상담 내역 예시 중 하나의 전체 발화를 제공하는 프롬프트: 프롬프트로 상담 작업내에서의 역할을 부여하며, 상담 내역 예시 중 하나의 전체 발화를 제공한다.
2. 대화 정책에 따른 예시를 자세히 서술한 프롬프트: 생성할 대화의 정책 키워드당 각각 하나의 예시에 더해 대화 정책 예측 모델의 예측 결과가 제공된다.

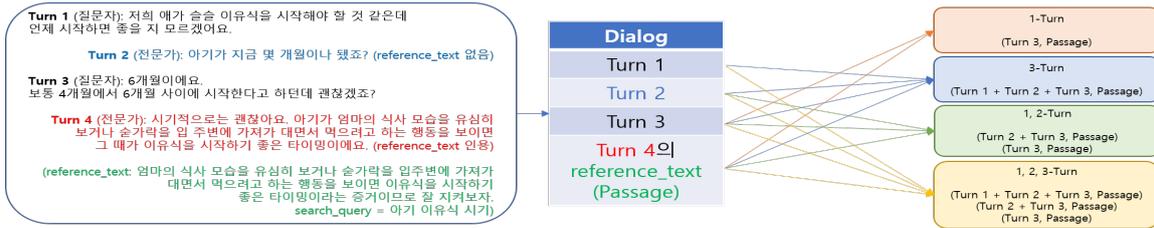


그림 3. DPR [13] 학습 데이터 증강 방법

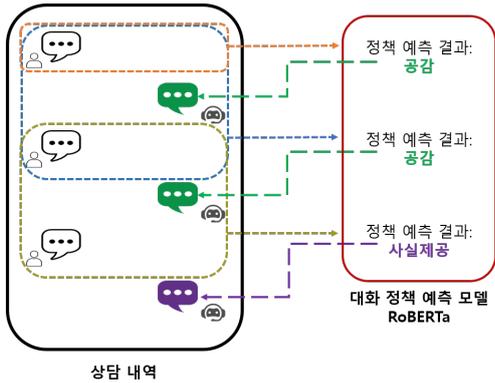


그림 4. 대화 정책 예측 모델의 시각화

프롬프트의 맥락에는 상담 내역의 최근 세 발화를 입력으로 받아 검색한 결과 중 상위 유사도 5개의 구절이 들어간다. 단, 입력 토큰 수가 과하게 많아지는 경우 가장 오래된 발화를 지운다. 각각의 자세한 예시는 표 3에 기재되어 있다.

#### 4. 실험 및 결과

실험을 위해 AIHub의 지식 검색 대화의 샘플 데이터를 이용했다 [17]. 이 실험 데이터의 대화 총수는 18,278개이며 주제는 총 10개로 분류되어 있다. 대화 세션당 발화 수는 평균 20.61, 최솟값은 20이며 최댓값은 28이다. 이 논문에서는 특정 도메인에서의 지식 기반 문장 생성 작업을 위해 실험 데이터 중 '가족과 주거' 주제만을 활용했다. 가족과 주거 주제의 대화 총수는 2,111개이고 이 중 200개를 무작위로 추출해 테스트 데이터로, 나머지 1,911개는 전문 지식 검색 모델과 대화 정책 예측 모델의 학습 데이터로 사용했다. 지식 검색 대화의 데이터 구조는 그림 5와 같다.

##### 4.1 학습 방법

전문 지식 검색 모델의 경우 전체 상담 내역 중에 전문가의 발화에서 reference\_text가 비어있지 않은 경우 이 키에 들어 있는 값들을 구절로 활용하고 그 이전의 상담 내역의 일부를 질문으로 활용해 DPR의 학습을 진행했다. 그림 3의 대화 예시가 학습 데이터 중 하나이며 전문가의 답변 Turn 4와 그 이전의 Turn 1부터 3까지를 활용하는 것을 시각화한 것이다. 테스트는

```
{
  "reference_date": "2021-10-07",
  "reference_text": [
    {
      "value": "엄마의 식사 모습을 유심히 보거나 손가락을 입 주변에 가져가 대면서 먹으려고 하는 행동을 보이면",
      "value_time": "2022-05-25",
      "value_type": "variable"
    }
  ],
  "role": "전문가",
  "search_URL": "https://kizmom.hankyung.com/news/view.html?id=202110072105o",
  "search_query": "아기 이유식 시기",
  "terminate": false,
  "text": "시기적으로는 괜찮아요. 아기가 엄마의 식사 모습을 유심히 보거나 손가락을 입",
  "turn_id": "BK22000057-2",
  "utterance_id": "BK22000057.4"
}
```

그림 5. AIHub의 지식 검색 대화[17] JSON 파일의 내용 일부

reference\_text가 있는 문장 직전의 가장 최근 세 발화를 결합해서 질문으로 사용했다. 대화 정책 예측 모델의 경우 질문자의 마지막 발화 이후 다음 전문가의 발화에서의 reference\_text가 있다면 사실제공, 없다면 공감으로 분류하게 학습을 진행했다. 전문가의 발화로부터 바로 이전 문장(1-Turn)만 반영하는 방식부터 4개의 문장(4-Turn)을 반영하는 방식까지 각각 학습 후 테스트 데이터에 대해 예측 정확도를 측정했다.

##### 4.2 정량적 실험 결과 및 분석

전문 지식 검색 모델의 실험 결과는 표 1에 제시되어 있다. 총 테스트 케이스는 1,637개로, 테스트 데이터 200개 내 모든 전문가의 발화 중 reference\_text가 있는 발화 이전의 상담 내역에서 가장 최근의 발화 3개가 입력으로 들어가는 방식이다. 제시된 수치는 기존 논문에서 제안한 Top-k retrieval accuracy이다. 배치 크기는 32이며 부정 샘플링은 단순한 무작위 추출로 학습을 수행했다. 기존 DPR 논문에서 다른 영문 데이터로 실험한 결과보다 전체적으로 성능이 더 좋게 나왔는데, 오픈 도메인에서의 질의응답이 아닌 특정 도메인에서의 맥락을 고려한 질의응답이고 탐색해야 하는 전체 구절의 수가 3,390개로 많지 않았기 때문에 성능이 높게 나온 것으로 추측된다. 생성형 모델을 활용한 상호작용 결과 생성에는 1,2-Turn 모델을 사용했다. 대화 정책 예측 모델의 예측 성능은 표 2에 기재되어 있으며 거대 언어 모델에서의 대화 정책을 결정하는 모델은 성능이 가장 좋은 3-Turn을 사용했다. 다만 테스트 데이터와 실제 상호작용에서 모델이 생성하는 데이터에는 명확한 차이가 있기에 테스트 데이터에서의 예측 성능과 실제 상호작용에서의 성능과는 상이할 수 있다.

표 1. 전문 지식 검색 모델의 성능 실험 결과

	1-Turn	3-Turn	1,2-Turn	1,2,3-Turn
Top-5	0.6048	0.8064	<b>0.8259</b>	0.8240
Top-10	0.7728	0.8943	0.9040	<b>0.9083</b>
Top-20	0.8717	0.9450	0.9468	<b>0.9499</b>
Top-100	0.9707	<b>0.9902</b>	0.9877	0.9853

표 2. 대화 정책 예측 모델의 성능 실험 결과

	1-Turn	2-Turn	3-Turn	4-Turn
Accuracy	0.9500	0.9464	<b>0.9574</b>	0.9569

표 3. 모델들의 출력을 결합하는 두 프롬프트의 예시

<p>1. 상담 내역 예시 전체를 하나 제공하는 프롬프트</p> <p>당신은 가족과 주거에 대한 전문가입니다. 당신은 상담에서 전문가의 역할을 담당하고 있습니다. 당신은 질문자의 가장 최근 응답에 대해 상담 내역과 맥락을 참고하여 전문가의 입장에서 최대한 적절하게 답변합니다. 생성할 응답의 길이는 40단어 이하입니다.</p> <p>상담의 예시는 다음과 같습니다.</p> <p>상담 예시</p> <p>질문자: 전셋값이 올라서 이사할 집을 찾아봐야 해. 할 일도 많은데 이사라니 겁 없는 사람은 서럽다.</p> <p>전문가: 요즘 전세 사기행위도 많아서 부동산 거래하는 게 무섭겠다.</p> <p>질문자: 맞아, 혹시 나도 사기당하면 어쩌나 걱정이랑이까. 서울에서 집 계약할 때 1인가구 전월세 안심도움서비스를 신청하면 된다고 하던데 그게 뭔지 알고 있어?</p> <p>전문가: 부동산 계약에 어려움이 있는 사회초년생이나 고령층 등 1인 가구의 전세나 월세 계약에 도움을 주는 서비스야. (중략)</p>	
<p>상담 내역</p> <p>{상담 내역}</p>	
<p>맥락</p> <p>{맥락}</p>	
<p>답변 (40단어 이내):</p> <p>2. 대화 정책에 따른 예시를 자세히 서술한 프롬프트</p> <p>당신은 가족과 주거에 대한 전문가입니다. 당신은 상담에서 전문가의 역할을 담당하고 있습니다. 문장 생성 정책은 공감, 사실제공 둘 중 하나입니다. 문장 생성 정책에 따른 답변 생성은 다음과 같습니다.</p> <p>1. 공감: 질문자의 가장 최근 응답에 호응하거나 공감하는 문장을 생성합니다. 공감의 경우 맥락의 정보를 참조하지 않으며 상담 내역에 최대한 적합한 문장이어야 합니다.</p> <p>공감 예시</p> <p>질문자: 전셋값이 올라서 이사할 집을 찾아봐야 해. 할 일도 많은데 이사라니 겁 없는 사람은 서럽다.</p> <p>전문가: 요즘 전세 사기행위도 많아서 부동산 거래하는 게 무섭겠다.</p> <p>2. 사실제공: 질문자의 가장 최근 응답에 대해 제공되는 맥락에서 가장 관련있는 키워드나 정보들을 인용하여 답변합니다.</p> <p>사실제공 예시</p> <p>질문자: 맞아, 혹시 나도 사기당하면 어쩌나 걱정이랑이까. 서울에서 집 계약할 때 1인가구 전월세 안심도움서비스를 신청하면 된다고 하던데 그게 뭔지 알고 있어?</p> <p>전문가: 부동산 계약에 어려움이 있는 사회초년생이나 고령층 등 1인 가구의 전세나 월세 계약에 도움을 주는 서비스야.</p> <p>당신은 질문자의 가장 최근 응답에 대해 상담 내역과 맥락을 참고하여 전문가의 입장에서 최대한 적절하게 답변합니다. 답변은 문장 생성 정책을 반드시 지켜야 하며 맥락과 상담 내역에 대해 모순이 없어야 합니다. 생성할 응답의 길이는 40단어 이하입니다.</p>	
<p>상담 내역</p> <p>{상담 내역}</p>	
<p>맥락</p> <p>{맥락}</p>	
<p>문장 생성 정책: {대화 정책}</p> <p>답변 (40단어 이내):</p>	

표 4. 두 프롬프트 간 SSA [18]측정 비교 결과

	감수성	구체성	평균
상담 내역 예시 전체를 하나 제공하는 프롬프트	0.7527	0.7018	0.7273
대화 정책에 따른 예시를 자세히 서술한 프롬프트	<b>0.8036</b>	<b>0.7291</b>	<b>0.7664</b>

### 4.3 문장 생성 결과의 정성적 비교와 분석

200개의 테스트 데이터 중, 100개를 무작위로 추출하여 GPT-3.5-turbo-0301 모델을 활용해 2개의 프롬프트 각각에서 상호 작용한 결과를 저장해서 비교해 보았다. 일단, 표 3에서의 2번 프롬프트를 활용하여 문장을 생성하는 예시는 표 5과 같다. 표 6에서 제공된 생성된 대화의 예시는 동일한 테스트 데이터에서의 1번 프롬프트와 2번 프롬프트의 문장 생성 결과다. 2번 프롬프트의 예시에는 전문가 발화 끝의 괄호에 발화 생성시의 문장 생성 정책을 남겨놨다. 1번 프롬프트에 비해 2번 프롬프트에서 상황에 맞는 답변을 생성하여 더 자연스러운 대화가 진행됐음을 확인할 수 있다. 또한 1번 프롬프트에서는 상황에 맞지 않은 정보를 제공하고 있으나 2번 프롬프트는 대화의 흐름이 자연스러워졌다. 또 다른 예시로는 표 7를 보면 알 수 있듯이 동일한 대화의 시작에서 1번 프롬프트와 2번 프롬프트의 응답이 상이한 것을 알 수 있다. 2번 프롬프트는 의문문으로 끝나면서 대화를 더 구체적으로 끌어 나가는 대화 정책에 적합한 문장을 생성했지만 1번의 경우 결혼 일정과 관련된 추가적인 정보를 활용해 문장을 생성했다는 것을 확인할 수 있다. 어떤 프롬프트가 더 자연스럽고 구체적인지는 사람마다 엇갈릴 수 있기 때문에 테스트 데이터 50개를 활용해 각각의 프롬프트로 생성한 상담 대화 전문을 가지고 구글에서 제안한 Sensibleness and Specificity Average (SSA) 평가를 수행했다 [18].

참여 인원은 총 11명이며, 테스트 데이터 50개 내 모든 사용자의 입력에 대해 거대 언어 모델이 두 프롬프트로 생성한 상담 대화 전문을 가지고 대화의 흐름이 자연스러운지(감수성) 그리고 구체적인지(구체성)를 그렇다/아니다로 평가를 진행했다. 단, 감수성에서 '아니다'를 체크하면 구체성에서 '그렇다'를 체크해도 '아니다'로 간주했다. 평가의 결과는 표 4에서 확인할 수 있는데, 1번 프롬프트보다 2번 프롬프트에서 모든 항목에 대해 더 우수한 성능을 보임을 확인할 수 있었다. 1번 프롬프트로 생성한 대화 전문에 감수성 항목에서 '그렇다'라고 선택된 경우는 총 399회, 그 중 구체성에도 '그렇다'가 선택된 경우는 총 386회이며, 2번 프롬프트는 각각 442회, 401회다. 하지만 두 경우 모두 구글의 Meena보다는 SSA 지수가 낮았는데, 이는 GPT-3.5-turbo-0301를 파인튜닝하지 않고 In-context learning [19]을 사용했기에 성능이 더 낮게 나오는 것이라 추측할 수 있다.

하지만 두 프롬프트 모두 문제점이 있는데, 테스트 데이터의 평균 발화 수가 20을 넘으며 모든 상담 내역을 거대 언어 모

표 5. 대화 정책에 따른 예시를 자세히 서술한 프롬프트의 다음 발화 생성 예시

<p><b>상담 내역</b></p> <p>질문자: 와이프가 이제 임신 15주가 됐어요. 그동안 많은 변화가 있었는데, 여러 가지 궁금한 점이 생겼어요.</p> <p>전문가: 임신 중 궁금증이 많다면 당연한 일이에요. 부족한 부분을 채워드릴 수 있도록 최선을 다할게요.</p> <p>질문자: 음, 먼저 임신 15주면 태아의 크기는 어느 정도 인가요?</p> <p>전문가: 임신 15주 태아의 크기는 9.3cm이며, 몸무게는 50g 정도입니다. 이 단계에서는 양수를 들이마실 뿐입니다.</p> <p>질문자: 아, 아직 많이 작네요. 태어는 어느 정도 발달이 되었나요?</p>
<p><b>맥락</b></p> <p>맥락0: 임신 15주 태아 특징: 태어는 양수를 들이마시고 내뿜기를 반복하면서 폐의 공기주머니를 발달시켜요.</p> <p>맥락1: 임신 15주 태아 특징: 눈을 뜨지 못하지만, 빛을 감지할 수 있게 되며, 불완전하지만 뇌가 발달하며 외부 자극에 대해 쾌감과 불쾌감, 불안, 조조 등의 기본적인 감정을 느끼기도 합니다.</p> <p>맥락2: 임신 15주 태아 특징: 특징: 조음파로 남아와 여아 구별이 가능해집니다.</p> <p>맥락3: 임신 17주 태아 특징: 특징: 몸 전체에 태반이 생기기 시작하고, 귓속에는 작은 뼈가 단단해지면서 청각 기관이 크게 발달해 조용한 음악을 들으면 안정된 모습을 보여요.</p> <p>맥락4: 임신 26주 태아 특징: 폐 속에서 폐포가 발달하기 시작하여 호흡을 위한 연습을 합니다.</p>
<p><b>문장 생성 정책: 사실제공</b></p> <p>답변 (40단어 이내): 전문가: 임신 15주 태어는 눈을 뜨지는 못하지만, 뇌가 발달해 감정을 느낄 수 있으며, 빛을 감지할 수 있게 됩니다.</p>

표 6. 프롬프트 간의 생성 대화 비교 (1)

<p><b>1. 상담 내역 예시 전체를 하나 제공하는 프롬프트</b></p> <p>질문자: 어제 아기 낳고 처음으로 치과에 다녀왔거든요. 1년 만에 가서 그런지 충치도 생기고 치석도 많아서 스케일링을 하는데 한참 걸렸어요.</p> <p>전문가: 첫니 충치치료는 영구치 건강을 위해 중요합니다. 충치가 생기면 치아가 일찍 빠질 수 있고, 유치 충치는 영구치 발달에 영향을 줍니다. 조기 진단과 치료가 필요합니다.</p> <p>질문자: 맞아요. 이제는 6개월마다 꼭 가려고요. 제가 이가 약해서 아기도 닳을까 봐 걱정인데 몇 살부터 치과에 데려가면 될까요?</p> <p>전문가: 생후 12개월 이전이나 첫 치아가 나오는 시기에 치과에 방문하는 것을 추천합니다. 또한, 생후 18~29개월에 영유아 구강검진을 받는것이 좋습니다. (후략)</p>
<p><b>2. 대화 정책에 따른 예시를 자세히 서술한 프롬프트</b></p> <p>질문자: 어제 아기 낳고 처음으로 치과에 다녀왔거든요. 1년 만에 가서 그런지 충치도 생기고 치석도 많아서 스케일링을 하는데 한참 걸렸어요.</p> <p>전문가: 정말 힘들었겠네요. 치아 건강은 생각보다 중요한데 출산으로 인해 놓치는 경우가 많아요. 앞으로 좀 더 신경 쓰시면 좋을 것 같아요. (공감)</p> <p>질문자: 맞아요. 이제는 6개월마다 꼭 가려고요. 제가 이가 약해서 아기도 닳을까 봐 걱정인데 몇 살부터 치과에 데려가면 될까요?</p> <p>전문가: 아기의 치아가 나오면 생후 12개월 이전이나 치아가 나온 시기에 치과에 방문하는 것이 좋아요. 국내에서는 영유아 구강검진을 하는데 18~29개월에 1차 구강검진이 이루어지는데 이때 치아의 상태나 구강검진을 받을 수 있어요. (사실제공) (후략)</p>

델의 입력으로 제공하기에는 토큰의 한계가 존재하기 때문에 과거의 맥락이 잊혀지는 문제가 발생했다. 특히 상담 내역이 길어질수록 맥락과 전혀 관련없는 엉뚱한 대화를 생성하거나 같은 문장을 계속 반복하는 경우가 있었다. 특히나 1번 프롬프트는 같은 문장을 반복하는 경우가 2번 프롬프트에 비해 더 많이 발생했다. 1번 프롬프트의 토큰 수가 2번 프롬프트보다 훨씬 많기 때문에 거대 언어 모델의 입력에 더 적은 상담 내역이 들어갈 수 밖에 없는 것이 원인으로 보인다. 하지만 2번 프롬프트는 1번 프롬프트에 비해 비문을 더 많이 생성하는 경향이 있었다. 그 중에서도 공감에 해당하는 문장을 생성할 때 영어를 한국어로 번역한듯한 문장을 생성하는 경우가 많았는데, 이를 바탕으로 GPT-3.5-turbo-0301의 학습 당시 한국어 학습 데이터가 적었음을 예상할 수 있다. 표 8를 보면 두 프롬프트로 생성했을 때의 문제점들을 하나씩 확인할 수 있다.

표 7. 프롬프트 간의 생성 대화 비교 (2)

<p><b>1. 상담 내역 예시 전체를 하나 제공하는 프롬프트</b></p> <p>질문자: 나 드디어 결혼 날짜 잡았어.</p> <p>전문가: 축하드려요! 결혼에 필요한 준비를 최대한 미리 하시는 것이 좋겠어요.</p> <p>코로나 바이러스의 여파로 인해 예식 준비와 관련된 일정 조정도 필요할 수 있어요. (후략)</p>
<p><b>2. 대화 정책에 따른 예시를 자세히 서술한 프롬프트</b></p> <p>질문자: 나 드디어 결혼 날짜 잡았어.</p> <p>전문가: 축하해요! 결혼은 정말 기쁜 일이지요. 어디서, 어떤 스타일의 결혼식을 하시려고 하세요? (공감) (후략)</p>

표 8. 프롬프트로 생성한 대화의 문제점 예시

<p><b>1. 상담 내역 예시 전체를 하나 제공하는 프롬프트에서 더 많이 발생한 문제점 - 반복 (전략)</b></p> <p>질문자: 와, 엄청 무서운 병이었네요. 아내에게 꼭 얘기해 줘야겠어요.</p> <p>전문가: <b>정기검진, 휴식, 혈압 관리 등을 통해 자간전증을 조기에 진단</b>하고 건강한 출산을 위해 노력하세요.</p> <p>질문자: 네, 그럴게요. 오늘 좋은 정보 친절하게 알려주셔서 정말 감사드려요.</p> <p>덕분에 궁금했던 것들이 많이 해소되었어요.</p> <p>전문가: <b>정기검진, 휴식, 영양균형 중요. 자간전증 조기 진단과 예방이 필요.</b> 혈압 관리도 중요한데, 의료진과 상의하시길 권해드립니다.</p>
<p><b>2. 대화 정책에 따른 예시를 자세히 서술한 프롬프트에서 더 많이 발생한 문제점 - 비문 (전략)</b></p> <p>질문자: 만약 병설 유치원 급식이 맵지 않게 개선된다고 하더라도 시간이 걸릴 테니까. 그래도 아이가 맵은 음식에 조금씩 적응할 수 있게 도와줘야겠어.</p> <p>전문가: 아이가 맵은 음식을 못 먹어 걱정이시군요. 부실급식 문제도 있어 고민이 많을 것 같아요. <b>함께 움켜잡아서 좋은 선택을 하시길 바랍니다.</b></p>

## 5. 결론

이 논문에서는 전문 지식을 검색하는 모델과 대화 생성 정책을 예측하는 모델을 프롬프트로 결합해 거대 언어 모델에게 작업 지향형 대화 작업 내에서 지식 기반 문장 생성과 일상 대화를 상황에 맞춰 선택해 생성해 나갈 수 있는 것을 보였다. 단순히 과거의 상담 내역과 예시를 제공하는 것보다 대화 생성 정책을 추가로 제공하고 각각의 정책에 대한 예시를 제공하는 것이 더 자연스러운 문장을 생성할 수 있음을 보였다. 이 논문에서 제안한 기법은 프롬프트 공학을 통해 여러 도메인에 적은 비용으로 적용할 수 있으며 각각의 모듈이 독립적으로 학습 가능하기 때문에 각 모델 간의 종속성을 고려하지 않아도 되는 장점이 있다. 하지만 한국어 말뭉치 비율이 굉장히 낮은 모델을 사용했기에 생성 모델에서의 과거의 맥락 손실과 대화가 길어짐에 따른 적절하지 않은 문장 생성 등 근본적인 문제가 해결되지 않았다. 향후 연구에서는 더 많은 한국어 말뭉치로 학습한 거대 언어 모델을 활용하여 더 품질이 좋은 문장을 생성할 수 있음을 보일 것이다. 또한 생성된 대화의 정성적 평가가 부족한 점을 보완하기 위해 이를 GPT4Eval와 같은 프롬프트 기반 거대 언어 모델의 정성적 평가를 추가로 수행할 계획이다 [20].

## 감사의 글

이 논문은 2023년도 정부(산업통상자원부)의 재원으로 한국산업기술평가관리원의 지원을 받아 수행된 연구임. (No. 1415184727, 전문개인투자자 맞춤형 투자 정보 제공을 위한 실시간 금융 텍스트 심층 이해 및 투자 정보 지원 서비스 개발)

## 참고문헌

- [1] W. Yu, C. Zhu, Z. Li, Z. Hu, Q. Wang, H. Ji, and M. Jiang, “A survey of knowledge-enhanced text generation,” *ACM Computing Surveys*, Vol. 54, No. 11s, pp. 1–38, 2022.
- [2] OpenAI. (2023) Chatgpt august 3 version. [Online]. Available: <https://chat.openai.com/chat>
- [3] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and finetuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [4] S. Swamy, N. Tabari, C. Chen, and R. Gangadharaiyah, “Contextual dynamic prompting for response generation in task-oriented dialog systems,” *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3102–3111, May 2023. [Online]. Available: <https://aclanthology.org/2023.eacl-main.226>
- [5] V. Hudeček and O. Dušek, “Are llms all you need for task-oriented dialogue?” *arXiv preprint arXiv:2304.06556*, 2023.
- [6] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, “A prompt pattern catalog to enhance prompt engineering with chatgpt,” *arXiv preprint arXiv:2302.11382*, 2023.
- [7] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “Qlora: Efficient finetuning of quantized llms,” *arXiv preprint arXiv:2305.14314*, 2023.
- [8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in Neural Information Processing Systems*, Vol. 33, pp. 9459–9474, 2020.
- [9] G. Izacard and E. Grave, “Leveraging passage retrieval with generative models for open domain question answering,” *arXiv preprint arXiv:2007.01282*, 2020.
- [10] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, Vol. 55, No. 12, pp. 1–38, 2023.
- [11] G. Lee, V. Hartmann, J. Park, D. Papailiopoulos, and K. Lee, “Prompted LLMs as chatbot modules for long open-domain conversation,” *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4536–4554, Jul. 2023. [Online]. Available: <https://aclanthology.org/2023.findings-acl.277>
- [12] S. Robertson, H. Zaragoza *et al.*, “The probabilistic relevance framework: Bm25 and beyond,” *Foundations and Trends® in Information Retrieval*, Vol. 3, No. 4, pp. 333–389, 2009.
- [13] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, “Dense passage retrieval for open-domain question answering,” *arXiv preprint arXiv:2004.04906*, 2020.
- [14] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, “Large language models are human-level prompt engineers,” *arXiv preprint arXiv:2211.01910*, 2022.
- [15] 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,” *arXiv preprint arXiv:1907.11692*, 2019.
- [16] 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.
- [17] AIHub. (2023) 지식검색 대화. [Online]. Available: <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realM&dataSetSn=71304>
- [18] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu *et al.*, “Towards a human-like open-domain chatbot,” *arXiv preprint arXiv:2001.09977*, 2020.
- [19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.
- [20] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, “GpTeval: Nlg evaluation using gpt-4 with better human alignment,” *arXiv preprint arXiv:2303.16634*, 2023.