

국립국어원 2023-01-59

발간등록번호
11-1371028-000966-01

# 2023년 인공 지능의 한국어 처리 능력 평가 체계 운영 및 평가 과제 구축

연구책임자  
김 한 샘



## 제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 '2023년 인공 지능의 한국어 처리 능력 평가 체계 운영 및 평가 과제 구축'에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2023년 3월 24일 ~ 2023년 12월 20일

2023년 12월 20일

연구책임자: 김한샘(연세대학교)

연구 기관: 연세대학교 산학협력단, 고려대학교 산학협력단,  
(주)나라지식정보, (주)테디썸, (주)마이스앤드

연구책임자: 김한샘

공동연구원: 송상헌, 박승희, 송영숙,

유현조, 정유남, 홍승혜, 함영균, 임경태,

윤영민, 김현명, 여진영, 나승훈, FEI LI

보조연구원: 노강산, 황동진, 정용빈, 이이슬,

박재완, 오유진, 서현빈, 박서윤, 강예지,

이재원, 김유진, 강조은





## 2023년 인공 지능의 한국어 처리 능력 평가 체계 운영 및 평가 과제 구축

이 사업은 인공 지능의 한국어 처리 능력을 평가할 수 있는 체계를 마련하기 위해 기존 6종 말뭉치에 대한 정비를 진행하고 경진대회, 상시 과제를 운영하여 향후 평가 체계에 대한 개선점과 발전 방향을 제안하는 것을 목적으로 하였다. 이에 따른 주요 과업과 사업 성과는 다음과 같다.

**평가과제 설계 및 말뭉치 정비:** 본 과제에서는 국립국어원에서 구축된 감정 분석 말뭉치, 이야기 완성 말뭉치, 표 기반 문장 생성 말뭉치, 그림 기반 문장 생성 말뭉치, 그리고 부적절성 말뭉치와 함의 분석 말뭉치 6종을 분석하여 언어 능력 평가 과제를 설계하고, 과제에 맞게 말뭉치를 정비, 가공하여 평가용 말뭉치로 변환하였다. 정비 방법론 연구 시 각 말뭉치를 과제에 맞게 변환할 수 있는 지침(guideline)을 설계하고 실제 정비 시 검수 과정을 거침으로써 평가용 말뭉치에 대한 품질을 제고하였다.

**인공 지능(AI)말뭉 경진대회 운영:** 본 과제에서는 정비한 6종 말뭉치 중 감정 분석, 이야기 완성 말뭉치를 사용하여 자연어 이해와 생성을 다루는 경진대회 과제를 설계하였다. 경진대회 운영을 위해 운영 기준과 절차를 수립하였으며, 발주처와의 협의를 통해 원활한 경진대회를 개최할 수 있도록 준비하였다. 또한 민원 응대 체계를 마련하여 참가자들의 민원에 효율적으로 대응하였다. 경진대회 진행 시 인간 평가, 모델 기술서, 발표 평가 등 다각도의 평가를 활용하여 경진대회 결과에 대한 신뢰성과 타당성을 제고하였다.

**인공 지능 (AI) 말뭉 상시 과제 운영 절차 수립:** 해외 및 국내 인공 지능 언어 능력 평가 사례들을 조사하고 분석하여 상시 과제 운영 기준을 수립하고, 상시 과제 운영을 위한 절차 및 지침서를 마련하였다. 또한 상시 과제로는 정비된 함의 분석, 표 기반 문장 생성, 그림 기반 문장 생성, 그리고 부적절성 말뭉치를 사용하여 총 4종의 평가 과제(함의 분석, 표의 일부분에 대한 해석 생성, 부적절성 문장에 대한 태도 탐지, 문자가 포함된 이미지 기반 문장 생성)를 설계하고 상시 과제를 운영하였다.

**인공지능 언어 능력 평가 체계 전문가 위원회 구성 및 운영:** 산업계와 학계의 인공지능, 언어처리, 평가 분야에 경험이 풍부한 전문가로 구성된 과제 위원회를 구성하여 경진

대회 운영 계획과 절차, 그리고 상시 과제 계획과 실제 운영에 대한 자문을 받았다. 또한 평가 과제의 타당성, 평가 체계 방향성 전반에 대한 자문과 더불어 초거대 언어 모델 시대의 평가 체계 발전 방향에 대한 전문가별 자문을 받아 발전 방향 제안에 반영하였다.

**한국어 인공지능의 언어 능력 평가 발전 방향 제안:** 본 과제에서 수행한 인공지능(AI) 말평 경진대회, 상시 과제 운영 결과를 정리하여 향후 평가 체계 운영을 위한 자료로 활용하는 한편, 초거대 언어 모델 시대의 연구 동향 조사, 한국어 인공지능 연구 실태 조사, 그리고 한국어 인공지능 연구자 수요 조사를 실시함으로써 실제 상황에 부합하는 평가 체계 발전 방향을 제안하였다.

**주요어:** 인공지능 언어 능력 평가 체계, 벤치마크, 경진대회, 상시 과제, 감정 분석, 이야기 완성, 부적절성, 표 기반 유사 문장 생성, 그림 기반 유사 문장 생성, 함의 분석, 초거대 언어 모델(LLM)

<Abstract>

## Organization & Consolidation of the Korean Benchmarks and Evaluation Tasks For the Language Proficiency Evaluation on Korean Artificial Intelligence in 2023

This research aimed to align the six types of corpora from National Institute of Korean Language, operate competitions and leaderboards, and propose improvements directions for the future evaluation system in order to arrange the benchmark to evaluate Korean language processing capabilities of artificial intelligence. The main tasks and achievements of the research are as follows.

**Evaluation task design and corpus alignment:** In this research we analyzed six corpora from National Institute of Korean Language emotion analysis, story completion, table-based generation, image-based generation, inappropriateness and adversarial NLI corpus. This project designed evaluation task also constructed the datasets which fit the task. Through the process, this project made guidelines for converting each corpus to fit the task and improved the quality of the evaluation corpus.

**Organize an artificial intelligence (AI) contest:** In this research, we designed a contest task that deals with natural language understanding and generation using the emotion analysis and story completion corpora among the six corpora we aligned. To administrate the contest, we established standards and procedures, and prepared for a smooth contest through consultation with the client. We also established a complaint response system to efficiently respond to participants' complaints. When conducting the contest, we improved the reliability and validity of the contest results by utilizing various evaluations such as human evaluation, model technical documentation, and presentation evaluation.

**Establishment of AI Evaluation Leaderboard Procedures:** Established organization standards and guidelines for leaderboard by analyzing domestic and overseas AI language benchmark research. In addition, we designed and operated a total of four leaderboard tasks (implication analysis, interpretation generation for a table, attitude detection for an inappropriate sentence, and image-based sentence generation)

**Operation of AI Language Proficiency Evaluation System Expert Committee:** A task committee board consisting of experienced experts in the fields of artificial intelligence, language processing, and evaluation from industry and academia was formed to advise on the operation plan and procedures of the contest, as well as the planning and actual operation of the leaderboard. In addition, we received expert advice on the development direction of the evaluation system in the era of large language models, which was reflected in the proposal for the development direction.

**Proposing a direction for the development of language proficiency evaluation of Korean artificial intelligence:** The results of the artificial intelligence (AI) evaluation contest and leaderboard conducted in this research were summarized and used as materials for future evaluation system operation, and a research trend survey in the era of large language models (LLMs), a survey of Korean artificial intelligence research status, and a survey of Korean artificial intelligence researcher demand were conducted to propose a direction for the development of the evaluation system that meets the actual situation.

Key-words: benchmarks, contests, leaderboards, emotion analysis, story completion, inappropriateness, table-based sentence generation, image-based sentence generation, implication analysis, large language model (LLM)

<요약문>

## 1. 사업 개요

### ☐ 사업명

- 2023년 인공 지능의 한국어 처리 능력 평가 체계 운영 및 평가 과제 구축

### ☐ 사업 기간

- 2023년 3월 24일 ~ 2023년 12월 20일

## 2. 사업 목적 및 범위

### ☐ 인공 지능의 한국어 처리 능력 평가 체계 운영 및 개선

- 인공 지능의 한국어 처리 능력 상시 평가 체계 시범 운영
- 인공 지능의 한국어 처리 능력 경진 대회 한시 운영
- 평가 체계 운영 중 개선 사항을 도출하여 평가 과제에 반영

### ☐ 국립국어원 구축 말뭉치에 대한 정비·가공·변환을 통해 인공 지능의 한국어 처리 능력 평가 과제로 구축

- 총 6종 말뭉치(이야기 완성, 감정 분석, 함의 분석, 비유리 표현, 표 기반 문장 생성, 그림 기반 문장 생성)에 대한 정비·가공·변환

### ☐ 인공 지능의 언어 능력 평가 체계를 위한 신규 과제 구축

- 정비한 6종 말뭉치 중 1종 이상의 말뭉치를 사용하여 신규 과제로 구축

### ☐ 인공 지능의 언어 능력 평가 체계 홍보 계획 수립 및 홍보

### ☐ 2023년 인공 지능의 한국어 처리 능력 평가 체계 운영 결과 정리 및 발전 방향 제안

### ☐ 2023년 인공 지능의 언어 능력 평가 과제 개발 및 평가 체계 운영 및 평가 과제 구축 사업 수행을 위한 작업 도구 활용 및 인력 구성 계획, 사업 관리 계획, 보안·위험 관리 계획 마련 및 실행

### 3. 사업 수행 내용

#### □ 평가용 말뭉치 정비 방법론 마련

- 6종 말뭉치에 대한 평가별 세부 과제 상정 및 정비 방법론 연구
- 말뭉치 정비 계획 수립 및 평가 과제별 말뭉치 정비
- 말뭉치 검수 기준 및 재정비 방법론 설계, 문서화

#### □ 평가용 말뭉치 변환

- 정비 말뭉치를 바탕으로 평가용 말뭉치 변환
- 언어 모델 능력 평가가 가능한 데이터 세트 마련
- 평가 체계 설계를 통한 인공지능 언어능력평가 과제 수립

#### □ 과제 진행을 위한 전문가 위원회 운영

- 인공 지능, 언어 처리, 평가에 경험이 풍부한 전문가로 구성
- 언어 능력 평가 체계 운영과 언어 능력 평가 체계 발전 방향 제안을 위한 자문, 검토

#### □ 경진대회 과제 운영 절차 수립

- 2023년 국립국어원 경진대회 과제 운영 기준 설계 및 절차 수립
- 경진대회 과제 운영을 위한 지침서 마련
- 지침서 내 참여자 팀 구성, 제출물 형식, 제한 규정, 대회 진행 일정 및 방법 등 수록
- 민원 응대 체계를 마련하여 예상되는 위협 및 문의사항에 대응

#### □ 경진대회 진행

- 감정 분석 과제, 이야기 완성 과제를 중심으로 인공지능 언어 능력 평가 경진대회 진행
- 인간 평가, 모델 기술서, 발표 평가 등 다각도의 평가 방법을 사용하여 평가 신뢰성 및 타당성 제고

#### □ 상시 과제 운영 절차 수립

- 상시 과제 운영 기준 수립 및 상시 과제 운영을 위한 절차, 지침서 마련
- 지침서 내 참여자 팀 구성 방법, 제출물 형식, 제한 규정 등 제시

□ 한국어 인공 지능의 언어 능력 평가 발전 방향 제안

- 향후 인공지능(AI) 말평의 발전을 위해 LLM 시대의 연구 동향 조사, 한국어 인공지능 연구 실태 조사 그리고 한국어 인공지능 연구자 수요 조사를 실시하여 발전 방향 제시





# 차 례

## 제1장 서론

1. 서론 .....	23
1.1. 사업 개요 .....	23
1.2. 사업 목적 및 범위 .....	23
1.3. 사업 수행 내용 .....	24

## 제2장 평가 체계용 말뭉치 정비

2. 평가 체계용 말뭉치 정비 .....	29
2.1. 감정 분석 말뭉치 .....	29
2.2. 이야기 완성 말뭉치 .....	42
2.3. 그림 기반 유사 문장 말뭉치 .....	48
2.4. 표 기반 유사 문장 말뭉치 .....	54
2.5. 함의 분석 말뭉치 .....	59
2.6. 부적절성 말뭉치 .....	65

## 제3장 인공지능(AI) 말뭉치 경진대회 과제 개발 및 운영

3. 인공지능(AI) 말뭉치 경진대회 과제 개발 및 운영 .....	73
3.1. 감정 분석 과제 .....	73
3.2. 이야기 완성 과제 .....	78

## 제4장 인공지능(AI) 말뭉치 상시과제 개발 및 운영

4. 인공지능(AI) 말뭉치 상시과제 개발 및 운영 .....	87
4.1. 상시과제 과제 선정 과정 .....	87
4.2. 과제 정의 .....	89

# 차 례

## 제5장 평가 체계 발전 방향 논의

5. 평가 체계 발전 방향 논의 .....	113
5.1. LLM 시대 연구 동향 조사 ([부록1] 참조) .....	113
5.2. 한국어 인공지능 연구 실태 조사 ([부록 2] 참조) .....	114
5.3. 평가 체계 관련 설문 조사 .....	115

## 제6장 평가 체계 홍보 활동

6. 평가 체계 홍보 활동 .....	135
----------------------	-----

## 제7장 평가 체계 운영 지침

7. 평가 체계 운영 지침 .....	141
7.1. 경진대회 운영 지침 및 절차서 .....	141
7.2. 경진 대회 결과 .....	147
7.3. 이야기 완성 과제 인간평가 지침 .....	148
7.4. 상시 과제 운영 지침 및 절차서 .....	155

## 제8장 평가 체계 홍보물

8. 평가 체계 홍보물 .....	161
--------------------	-----

## 제9장 결론 및 기대 효과

9. 결론 및 기대 효과 .....	165
---------------------	-----

[참고 문헌] .....	166
[부록1] LLM 연구 동향 조사 .....	172
[부록2] 한국 자연어 처리 연구 동향 조사 .....	186
[부록3] 과제위원회 회의록 .....	202
[부록4] 인공지능 평가 체계 발전 방안 자문의견서 .....	212

## 표 차례

<표 1> 감정 분석 말뭉치 트위터 자료 수집 기준 .....	29
<표 2> 담화 선정 세부 기준 .....	30
<표 3> 감정 분석 표지 .....	31
<표 4> 감정 분석 말뭉치 대상 선정 규칙 .....	31
<표 5> 감정 분석 말뭉치 자료 통계 (이영희 외, 2022: 48) .....	33
<표 6> 고마움 주석 검수 예시 .....	34
<표 7> 감정 과소 주석 검수 예시 .....	34
<표 8> 기대감 주석 검수 예시 .....	35
<표 9> 불필요한 감정 주석 삭제 예시 .....	35
<표 10> 대상 수식어 주석 검수 예시 .....	35
<표 11> 수식 성분 길이 검수 예시 .....	36
<표 12> 감성 대상 우선 순위 검수 예시 .....	36
<표 13> 우선 등장 대상 선정 규칙 검수 예시 .....	36
<표 14> 다어절 고유명사 주석 검수 예시 .....	37
<표 15> 스펠 과소/과대 검수 예시 .....	37
<표 16> 숫자 대상 주석 검수 예시 .....	38
<표 17> 비식별화 표지 및 설명 .....	38
<표 18> 감정 분석 말뭉치 joy 주석 검수 사례 .....	39
<표 19> 감정 분석 말뭉치 감정 대상 주석 검수 사례 .....	40
<표 20> 감정 분석 JSONL 말뭉치 .....	40
<표 21> 이야기 완성 말뭉치 구성 .....	45
<표 22> 이야기 완성 말뭉치 검수 유형 및 건수 .....	47
<표 23> 이야기 완성 JSONL 말뭉치 .....	47
<표 24> 그림 기반 말뭉치 그림 ID별 데이터 개수 .....	48
<표 25> 수정 전 그림 기반 유사 문장 말뭉치 .....	50
<표 26> 수정 후 그림 기반 유사 문장 말뭉치 .....	50
<표 27> OCR 추가 사례 .....	50
<표 28> OCR 원문 내 현실 표기 인정 사례 .....	51
<표 29> 그림 기반 유사 문장 JSONL 말뭉치 .....	51
<표 30> 표 ID 별 데이터 개수 .....	54

## 표 차례

<표 31> 표 기반 유사 문장 말뭉치 내 문자 및 수치 오타 예시 .....	56
<표 32> 하이라이트 셀 참조 검수 사례 .....	57
<표 33> 하이라이트 셀 참조 검수 사례 문장 수정 .....	57
<표 34> 표 기반 유사 문장 JSONL 말뭉치 .....	57
<표 35> 함의 분석 말뭉치 수정 예시 1 .....	60
<표 36> 함의 분석 말뭉치 수정 예시 2 .....	60
<표 37> 가설 문장 생성 시 원문 자체 오류 예시 .....	60
<표 38> 함의 분석 말뭉치 수정 유형 및 건수 .....	61
<표 39> 파일별 수정 건수 .....	61
<표 40> 함의 분석 JSONL 말뭉치 .....	64
<표 41> 부적절성 말뭉치 맥락 예시 .....	66
<표 42> 부적절성 말뭉치 명시적/비명시적 예시 .....	66
<표 43> 부적절성 말뭉치 강도 예시 .....	67
<표 44> 부적절성 말뭉치 수정 건수 .....	68
<표 45> 부적절성 말뭉치 ‘맥락’ 정비 예시 .....	68
<표 46> 부적절성 말뭉치 ‘명시성’ 정비 예시 .....	69
<표 47> 부적절성 말뭉치 ‘강도’ 정비 예시 .....	69
<표 48> 부적절성 말뭉치 ‘부적절 표현’ 정비 예시 .....	70
<표 49> 부적절성 JSONL 말뭉치 .....	70
<표 50> 기준 모델 목록 .....	73
<표 51> 감정 분석 과제의 예시 .....	74
<표 52> 감정 분석 평가 지표 .....	74
<표 53> 감정 분석의 모델 입력과 출력의 예 .....	75
<표 54> 감정 분석 데이터 규모 .....	75
<표 55> 감정 분석 데이터 형식의 예 .....	75
<표 56> 이야기 완성 과제 예시 .....	78
<표 57> 이야기 완성 과제의 부적절한 예시 .....	79
<표 58> 이야기 완성의 모델 입력과 출력 예시 .....	81
<표 59> 이야기 완성 데이터 규모 .....	81
<표 60> 이야기 완성 데이터 형식의 예 .....	82

## 표 차례

<표 61> 상시과제 과제 개요 .....	87
<표 62> 상시과제 과제 데이터 형식 .....	88
<표 63> 기준 모델 목록 .....	88
<표 64> 함의 분석 과제 데이터 세트의 예시 .....	89
<표 65> 함의 분석 모델 입력과 출력의 예 .....	90
<표 66> 함의 분석 데이터 규모 .....	90
<표 67> 데이터 형식의 예 .....	91
<표 68> 표의 일부분에 대한 해석 생성 과제 모델 출력의 예 .....	92
<표 69> 표의 일부분에 대한 해석 생성 과제 데이터 규모 .....	94
<표 70> 표의 일부분에 대한 해석 생성 과제 데이터 형식의 예 .....	94
<표 71> 표의 일부분에 대한 해석 생성 과제 데이터 형식 변환 예시 .....	100
<표 72> 이미지 캡셔닝 과업의 예시 .....	100
<표 73> 문자가 포함된 이미지 기반 문장 생성 과제 모델 출력의 예 .....	101
<표 74> 문자가 포함된 이미지 기반 문장 생성 과제 데이터 규모 .....	102
<표 75> 문자가 포함된 이미지 기반 문장 생성 과제 데이터 형식의 예 .....	102
<표 76> 부적절성 문장에 대한 태도 탐지의 예시 .....	106
<표 77> 부적절성 문장에 대한 태도 탐지 모델 입력과 출력의 예 .....	107
<표 78> 부적절성 문장에 대한 태도 탐지 비식별화 태그 .....	107
<표 79> 부적절성 문장에 대한 태도 탐지 데이터 규모 .....	108
<표 80> 부적절성 문장에 대한 태도 탐지 데이터 형식의 예 .....	108
<표 81> 벤치마크 개발 필요성 과제 순위 .....	128
<표 82> 인공지능(AI)말뭉 관련 홍보 내역 .....	135
<표 83> 관련 분야 대학 기관·협회 TM .....	137
<표 84> 경진대회 운영 계획 및 상세 절차 .....	141
<표 85> 경진대회 운영 안내문 .....	142
<표 86> 참가자 시스템 검증을 위한 평가 항목 .....	145
<표 87> 발표 평가 시 심사 기준 및 배점 .....	146
<표 88> 경진대회 참가 현황 .....	147
<표 89> 감정분석 정량 평가 결과 .....	147
<표 90> 이야기완성 정량 평가 결과 .....	148

## 표 차례

<표 91> 인간평가 요약 .....	149
<표 92> 인간 평가 1단계 개요 .....	149
<표 93> 인간평가 1단계평가 준거 .....	150
<표 94> 인간 평가 2단계 개요 .....	153
<표 95> 인간 평가 워크 벤치 .....	153
<표 96> 표준 점수 계산식 .....	154
<표 97> 1단계, 2단계 반영식 .....	154
<표 98> 상시 과제 운영 계획 및 상세 절차 .....	155
<표 99> 개발된 평가 코드 .....	156
<표 100> 개발된 평가 코드 예시 .....	157
<표 101> 평가 체계홍보물 예시 .....	161
<표 102> 2023년 영어 리더보드/ 벤치마크 목록 .....	172
<표 103> 2023 중국어 리더보드/벤치마크 목록 .....	174
<표 104> 2023 리더보드/벤치마크 공개 데이터 세트 목록 .....	175
<표 105> 2023 리더보드/벤치마크 미공개 데이터 세트 목록 .....	176
<표 106> 2023 제안된 평가 방법 .....	178
<표 107> 시나리오 분류 체계 정의 .....	179
<표 108> 분류 체계에 따른 2023 공개 리더보드/ 벤치마크 분류 .....	179
<표 109> 분류 체계에 따른 국립국어원 언어자원 발전 방향 제안 .....	180
<표 110> 효율성 그래프 .....	183
<표 111> 2022년 토픽 추출 결과 by gensim .....	188
<표 112> 2022년 토픽 추출 결과 by tomotopy .....	189
<표 113> 2023년 토픽 추출 결과 by gensim .....	191
<표 114> 2023년 토픽 추출 결과 by tomotopy .....	193
<표 115> 2022년 HCLT 논문 키워드 빈도 .....	194
<표 116> 2023년 HCLT 논문 키워드 빈도 .....	196
<표 117> 2022년 HCLT 키워드 빈도(정규화) .....	197
<표 118> 2023년 HCLT 키워드 빈도(정규화) .....	198

## 그림 차례

[그림 1] 플루척의 감정 수레바퀴 .....	30
[그림 2] 깃허브 히스토리 .....	49
[그림 3] 테이블형 데이터 변환 예시 .....	49
[그림 4] OCR 원문 내 현실 표기 이미지 .....	51
[그림 5] 감정 분석 베이스라인 모델 개념도 .....	77
[그림 6] 이야기 완성 평가 metric. n = 정성 평가 대상 팀 수 .....	81
[그림 7] 이야기 완성 기준 모델(baseline model) 개념도 .....	83
[그림 8] 이미지 캡셔닝 예시 .....	100
[그림 9] 문자가 포함된 이미지 기반 문장 생성 과제 기준 모델 개념도 .....	105
[그림 10] 부적절한 문장에 대한 태도 탐지 기준 모델 개념도 .....	109
[그림 11] 성별 응답 .....	116
[그림 12] 연령대 응답 .....	116
[그림 13] 직업 응답 .....	116
[그림 14] 최종 학력 응답 .....	116
[그림 15] 전공 계열 응답 .....	117
[그림 16] 경력 응답 .....	117
[그림 17] 벤치마크 사용 빈도 .....	118
[그림 18] 벤치마크 대체 응답 .....	118
[그림 19] 국내 벤치마크 사용 빈도 순위 .....	118
[그림 20] 해외 벤치마크 사용 빈도 순위 .....	118
[그림 21] 국내 벤치마크 만족 요인 .....	119
[그림 22] 해외 벤치마크 만족 요인 .....	119
[그림 23] 국내 벤치마크 불만족 요인 .....	120
[그림 24] 해외 벤치마크 불만족 요인 .....	120
[그림 25] 진술1 응답 .....	123
[그림 26] 진술2 응답 .....	123
[그림 27] 진술3 응답 .....	123
[그림 28] 진술4 응답 .....	123
[그림 29] 진술1 응답 .....	124
[그림 30] 진술2 응답 .....	124

## 그림 차례

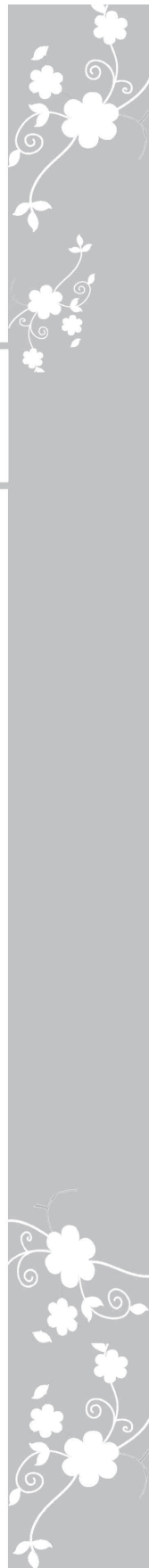
[그림 31] 진술3 응답 .....	124
[그림 32] 진술4 응답 .....	124
[그림 33] 진술1 응답 .....	125
[그림 34] 진술2 응답 .....	125
[그림 35] 진술3 응답 .....	125
[그림 36] 진술1 응답 .....	126
[그림 37] 진술2 응답 .....	126
[그림 38] 진술3 응답 .....	126
[그림 39] 리더보드 상금 필요성 응답 .....	129
[그림 40] 리더보드 상금 규모 응답 .....	129
[그림 41] 리더보드 운영 주체 응답 .....	129
[그림 42] 리더보드 홍보 채널 응답 .....	129
[그림 43] NLU 평가지표 응답 .....	130
[그림 44] NLG 평가지표 응답 .....	130
[그림 45] 정성적 평가 필요성 응답 .....	131
[그림 46] 종합적 평가 지표 개발 응답 .....	131
[그림 47] 인공지능(AI) 말평 홍보 사례 .....	135
[그림 48] 인공지능(AI) 말평 홍보글 업로드 사이트 목록1 .....	136
[그림 49] 인공지능(AI) 말평 홍보글 업로드 사이트 목록2 .....	136
[그림 50] 경진대회 심사 진행 단계 .....	144
[그림 51] 평가 지표에 따른 성능 레이더맵 .....	178
[그림 52] GAIA 성능 비교 그래프 .....	182
[그림 53] 다양한 규모의 학습용 LLM의 성능과 탄소 배출량 .....	183
[그림 54] 다양한 LLM의 에너지 소비량과 추론 속도 .....	183
[그림 55] 세 가지 프롬프트 전략을 사용한 GPT-4 성능 .....	184
[그림 56] 2022년 토픽 규모에 따른 perplexity, 일관성 변화 by gensim .....	187
[그림 57] 2022년 토픽 모델링 시각화 및 군집화 by gensim .....	188
[그림 58] 2022년 토픽 규모에 따른 .....	189
[그림 59] 2022년 토픽 모델링 시각화 및 군집화 by tomotopy .....	190
[그림 60] 2023년 토픽 규모에 따른 perplexity, 일관성 변화 by gensim .....	191
[그림 61] 2023년 토픽 모델링 시각화 및 군집화 by gensim .....	192
[그림 62] 2023년 토픽 규모에 따른 coherence 값 변화 by tomotopy .....	192
[그림 63] 2023년 토픽 모델링 시각화 및 군집화 by tomotopy .....	194





# 제 1 장

## 서론





## 1. 서론

### 1.1. 사업 개요

#### ☐ 사업명

- 2023년 인공 지능의 한국어 처리 능력 평가 체계 운영 및 평가 과제 구축

#### ☐ 사업 기간

- 2023년 3월 24일 ~ 2023년 12월 20일

### 1.2. 사업 목적 및 범위

#### ☐ 인공 지능의 한국어 처리 능력 평가 체계 운영 및 개선

- 인공 지능의 한국어 처리 능력 상시 평가 체계 시범 운영
- 인공 지능의 한국어 처리 능력 경진 대회 한시 운영
- 평가 체계 운영 중 개선 사항을 도출하여 평가 과제에 반영

#### ☐ 국립국어원 구축 말뭉치에 대한 정비·가공·변환을 통해 인공 지능의 한국어 처리 능력 평가 과제로 구축

- 총 6종 말뭉치(이야기 완성, 감정 분석, 함의 분석, 비유리 표현, 표 기반 문장 생성, 그림 기반 문장 생성)에 대한 정비·가공·변환

#### ☐ 인공 지능의 언어 능력 평가 체계를 위한 신규 과제 구축

- 정비한 6종 말뭉치 중 1종 이상의 말뭉치를 사용하여 신규 과제로 구축

#### ☐ 인공 지능의 언어 능력 평가 체계 홍보 계획 수립 및 홍보

#### ☐ 2023년 인공 지능의 한국어 처리 능력 평가 체계 운영 결과 정리 및 발전 방향 제안

#### ☐ 2023년 인공 지능의 언어 능력 평가 과제 개발 및 평가 체계 운영 및 평가 과제 구축 사업 수행을 위한 작업 도구 활용 및 인력 구성 계획, 사업 관리 계획, 보안·위험 관리 계획 마련 및 실행

#### □ 사업의 세부 범위

- 2023년 인공 지능 한국어 처리 능력 평가 체계를 운영하고, 이를 토대로 개선점을 도출하였다. 2023년 평가 체계 운영 기준 수립을 위해 2022년 평가 체계 운영에 대한 개선점을 찾은 후, 해당 결과를 기준 수립에 활용하였다. 마찬가지로 2023년 한시적 경진대회 및 상시 평가 체계 운영 결과를 분석하여 해당 결과를 향후 언어능력 평가 체계에 환류하였다.
- 국립국어원이 구축한 6종의 말뭉치를 인공 지능의 언어 능력 평가에 활용할 수 있도록 정비·가공하였다. 6종의 대상 말뭉치는 이야기 완성, 감정 분석, 함의 분석, 비유리 표현, 표 기반 문장 생성, 그림 기반 문장 생성이다.
- 정비·가공한 6종 말뭉치를 토대로 한국어 처리 능력 평가 신규 과제를 구축하였다.
- 인공 지능의 언어 능력 평가 체계에 대한 홍보 계획 수립 및 홍보를 수행하였다.
- 2023년 인공 지능의 한국어 처리 능력 평가 체계 운영 결과를 정리하고 발전 방향을 제안하였다.
- 2023년 인공 지능의 언어 능력 평가 과제 개발 및 평가 체계 운영, 평가 과제 구축 사업 수행을 위한 작업 도구 활용에 대한 계획을 마련하고 실행하였다. 또한 사업 수행 인력 구성 계획, 사업 관리 계획, 보안·위험 관리 계획을 수립하여 실행하였다.

### 1.3. 사업 수행 내용

#### □ 평가용 말뭉치 정비 방법론 마련

- 본 과제에서는 6종 말뭉치에 대한 평가별 세부 과제를 상정하고, 평가용 말뭉치로 사용하기 위한 정비 방법론 연구를 진행하였다.
- 이러한 연구를 통해 말뭉치 전반을 평가용 말뭉치로 정비하기 위한 계획을 수립하였으며, 인공 지능 언어 능력 평가 과제에 맞게 정비하였다.
- 말뭉치 목적에 타당한 검수 기준 및 말뭉치 재정비 방법론을 개발하므로 향후 기타 국립국어원 말뭉치에 적용할 수 있으며, 더 나아가 국내 동일, 혹은 유사 데이터 세트에 대한 평가 기준으로 자리 잡을 수 있도록 하였다.

#### □ 평가용 말뭉치 정비

- 본 과제에서 지향하는 인공 지능 언어 능력 평가 체계는 모델의 언어 능력을 평가하고 이를 개선하여 자연어 처리의 발전을 가능하게 하는 수단이다.

- 이에 따라 본 과제에서는 기존 말뭉치들을 평가용 말뭉치로 정비함으로써 언어 모델의 능력을 평가할 수 있는 데이터 세트를 마련하였다. 또한 평가 체계를 설계함으로써 누구나 언제든지 모델을 평가하고 성능을 향상시킬 수 있는 계기를 마련하였다.

#### □ 과제 진행을 위한 전문가 위원회 운영

- 인공 지능, 언어 처리, 평가에 경험이 풍부한 전문가들에게 인공 지능 언어 능력 평가 체계 운영과 언어 능력 평가 체계의 발전 방향 제안을 위한 자문을 받았다.
- 전문가 위원회는 국내외 인공 지능, 언어 처리, 평가 분야 전문가로 구성된 검토위원회와 자문위원회로써, 산업계, 학계, 정부기관 등 인공 지능 언어 처리 전문가 5인으로 각각 구성하였다. 사업 기간 중 원회 회의를 통해 발전 방향 제안과 인공 지능 언어 능력 평가 체계 수립·운영에 대한 자문과 검토를 받았다 (부록 3, 4 참조).

구분	검토위원 명단
산업계	김태윤(SKT)
	조원익(삼성전자)
	한지윤(업스테이지)
학계	김학수(건국대)
상시 검토위원	임준호(한국전자통신연구원 / 튜터러스 랩스)
구분	자문위원 명단
산업계	장두성(KT)
	박재현(NCsoft)
학계	최기선(카이스트)
	신효필(서울대)
	임희석(고려대)

#### □ 경진대회 과제 운영 절차 수립

- 본 과제에서는 2022년 국립국어원 인공 지능 언어 능력 평가 경진대회 및 국어원과의 유기적인 협의를 거쳐 2023년 경진대회 과제의 운영 기준과 절차를 상세하게 결정하고, 지침서를 마련하였다.
- 지침서에는 참여자 팀 구성 방법, 제출물 형식, 제한 규정 등에 대해 다루었으며, 예선 및 본선 진행 일정 및 방법을 제시하였다.
- 경진대회의 운영 기준 및 절차는 발주처와 수시로 협의하여 원활한 경진대회 진행이 이루어질 수 있도록 했으며, 민원 응대 체계를 마련하여 대회 참가자들의 문의 사항에 효율적으로 대응하였다.

□ **경진대회 진행 (23. 8. 21~23. 10. 20.)**

- 감정 분석 과제, 이야기 완성 과제를 중심으로 인공지능 언어 능력 평가 경진대회를 진행하였다.
- 경진대회 진행 시 인간 평가, 모델 기술서, 발표 평가 등 다각도의 평가 방법을 사용하여 경진대회 진행 결과 평가에 대한 신뢰성 및 타당성을 제고하였다.
- 자연어 이해 및 생성 분야에서의 경진대회 진행을 통해 국내 관련 분야 연구자 및 관심 있는 일반인들의 참여와 경쟁을 유도하였으며, 경쟁을 통해 국내 한국어 인공 지능의 학습 능력 향상 및 발전, 관련 기술 및 산업 분야의 발달을 꾀하였다.

□ **상시 과제 운영 절차 수립**

- 본 연구에서는 해외의 GLUE, SuperGLUE 등과 국내의 KLUE 같은 타 인공 지능 언어 능력 평가 사례의 경우를 참고하여 상시 과제 운영의 기준을 수립하고, 상시 과제 운영을 위한 절차 및 지침서를 마련하였다.
- 지침서에는 참여자 팀 구성 방법, 제출물 형식, 제한 규정 등을 제시하였으며, 해당 내용과 더불어 운영 기준 및 절차에 대해서는 발주처와 수시로 협의하여 운영 절차를 수립하였다.

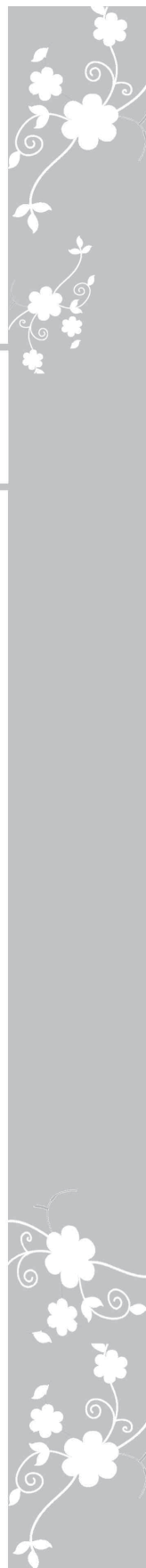
□ **한국어 인공 지능의 언어 능력 평가 발전 방향 제안**

- 본 과제에서는 국립국어원 인공지능 언어능력 평가 체계(인공지능(AI) 말평) 경진대회, 상시과제 운영과 더불어 향후 AI 말평의 발전을 위해 LLM 시대의 연구 동향 조사, 한국어 인공지능 연구 실태 조사 그리고 한국어 인공지능 연구자 수요 조사를 실시하여 발전 방향을 제시하였다.



## 제 2 장

# 평가 체계용 말뭉치 정비







## 2. 평가 체계용 말뭉치 정비

평가 체계용 말뭉치 정비는 경진대회, 상시 과제 진행을 위해 이루어졌으며, 과제용 말뭉치로 정비하기 전에 사전 검수를 통해 구축 당시 발생한 부정확한 주석이나 구축 지침과 모순되거나 어긋나는 데이터를 수정하는 단계를 거쳤다. 이후 과제 개발 사항에 따라 특정 주석 요소들을 중심으로 검수를 진행하여 평가용 말뭉치로 정비를 진행하였고, 형식 역시 평가 체계에 사용되는 JSONL 포맷으로 변환하였다. 일련의 과정을 통해 평가용 말뭉치에 대한 정확도와 신뢰도를 제고하였다.

### 2.1. 감정 분석 말뭉치

#### ○ 인수 말뭉치 분석

검수를 위해 인계받은 ‘2022 감정 분석 말뭉치’는 ‘국립국어원 2022년 말뭉치 감정 분석 및 연구 보고서(이영희 외, 2022)’에 따르면 트위터 자료 50,000건, 일상 대화 자료 10,000건으로 이루어져 있으며 트위터 자료는 문화 콘텐츠이면서 5어절 이상인 문서만을 대상으로 하였다. 문서에는 텍스트로 이루어진 데이터만이 포함되었으며, 텍스트가 아닌 자료 유형 혹은 외국어로만 구성된 자료 유형은 제외되었다.

또한, 문서는 2020년 1월 1일 이후에 작성된 게시글만 포함하였고 중복 자료, 상업 광고 자료 등은 포함하지 않았다. 모든 자료는 저작권 이용 허락 계약을 진행하여 향후 발생할 수 있는 법률적 분쟁을 최소화하고 민간 활용도를 제고하였다. 이상 트위터 자료 수집과 관련된 내용은 아래 표에 요약되어 있다.

<표 1> 감정 분석 말뭉치 트위터 자료 수집 기준

선별 대상	세부 기준
문화콘텐츠 관련 문서	영화/드라마/방송, 공연/전시/박람회, 도서/문학, 게임, 캐릭터, 음악/음반/콘서트, 연예인/유명인/팬덤/팬클럽 관련 게시글
포함 대상 기간	2020년 1월 1일 이후 작성 게시글
5어절 이상 문서	5어절 이상으로 구성된 게시글
비문서, 비국문 자료 제외	이미지, 스티커, 사진, 동영상, 파일 링크, 웹 주소, 해시태그로만 구성된 게시 자료, 전문 외국어로 구성된 게시 자료 삭제
중복 글, 펴 글, 홍보 글 제외	중복 게시 자료, 펴글(기사, 타인이 작성한 게시물 등)로만 구성된 게시 자료, 상업적 광고가 포함된 자료 삭제

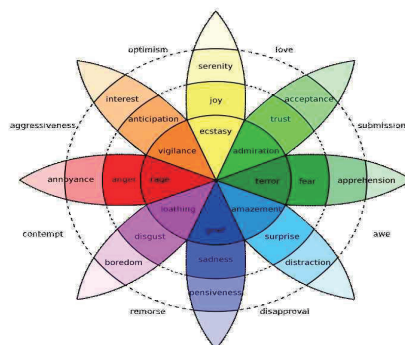
일상 대화 자료는 국립국어원 배포 말뭉치인 ‘2020 일상 대화 말뭉치’에서 감정 담화문만을 선별하여 구축하였다. 일상 대화 자료 또한 감정 분석 대상 문장과 그것의 선행 또는 후행 문장과의 맥락 유기성을 고려하여 5어절 이상 담화를 대상으로 하였다. 감정 담화문 내의 감정 개수에는 제한을 두지 않았고 연속하는 발화 중 특정 발화를 삭제하여 분석 대상 담화를 선별할 수 있도록 하였다. 정리된 담화 선정 세부 기준은 아래의 표와 같다.

<표 2> 담화 선정 세부 기준

idx	담화 선정 세부 기준
1	감정 담화문을 선별하기 위해서는 감정 문장을 찾는 작업이 가장 먼저 이루어져야 한다. 감정 문장이란 발화문에서 ‘발화자의 감정’이 드러나는 발화를 포함하는 문장이다. 감정 문장 선별 담당자는 일상 대화 말뭉치의 발화문을 직접 읽으며 감정 문장을 선별한다. 감정 문장은 색상으로 표시한 문장이다.
2	감정 문장을 찾은 후 담화 맥락을 파악할 수 있는 범위에서 선행 또는 후행 문장을 포함하여 감정 담화문을 선별한다. 이때 감정 담화문 내에서 발화자가 바뀌는 말차례 바꿈의 횟수나 전체 문장 수는 제한하지 않는다
3	감정 담화문은 하나 이상의 감정 문장을 포함할 수 있다. 또한, 하나의 문장만으로도 담화의 맥락을 이해할 수 있다면 단일 문장 담화 구성도 가능하다.
4	연속하는 발화 중간의 특정 발화를 삭제하여 분석 대상 담화를 선별할 수 없다.

구축된 감정 분석용 말뭉치는 감정 분석 방법론에 따라 주석되었다. 감정 분석 방법론은 로버트 플루치(Robert Plutchik)의 감정 체계를 차용하였다. SemEval 2018에서는 정서 심리학자 로버트 플루치(Robert Plutchik)이 주장한 기본 감정 및 복합 감정을 선별하여 감정 분석용 표지를 구성하였다. 아래의 그림은 로버트 플루치(Robert Plutchik)의 감정 수레바퀴이다. 플루치의 감정 수레바퀴(Plutchik’s wheel)는 8개의 핵심 감정의 조합으로 발생하는 감정의 상태를 표현한 것이다.

2022 감정 분석 말뭉치에서는 로버트 플루치(Robert Plutchik)의 기본 감정 ‘Joy, Anticipation, Trust, Surprise, Disgust, Fear, Anger, Sadness’를 차용하여 ‘기쁨, 기대, 신뢰, 놀람, 혐오, 공포, 분노, 슬픔’을 감정 분석의 표지로 활용하였다.



[그림 1] 플루치의 감정 수레바퀴

본 말뭉치에서는 8가지 기본 감정(기쁨, 기대, 신뢰, 놀람, 혐오, 공포, 분노, 슬픔)을 감정 분석의 표지로 활용하되, 기본 감정 스펙트럼에서 벗어나 어떠한 표지로도 정의될 수 없는 경우에는 ‘기타’ 표지를 활용하였다. 각 감정 분석 표지에 대한 세부 설명은 기술되어 있다.

<표 3> 감정 분석 표지

감정 분석표지	동일 스펙트럼 감정	설명
기쁨(Joy)	평온, 황홀	어떤 만족감에 의해 느끼는 즐겁고 흥겨운 감정.
기대(Anticipation)	경계, 관심	앞으로 있을 일이나 상황을 미리 짐작함. 또는 그런 내용.
신뢰(Trust)	수용, 감탄/존경	굳게 믿고 의지함.
놀람(Surprise)	놀라움, 부주의/방심	어떤 일이 뜻밖이거나 훌륭하거나 무서워서 신기해하거나 흥분하여 가슴이 뛰는 느낌.
혐오(Disgust)	지루함, 증오	싫어하고 미워함.
공포(Fear)	불안, 두려움	두렵고 무서움.
분노(Anger)	짜증, 격노	분개하여 크게 화를 냄.
슬픔(Sadness)	수심, 비탄	마음이 아프거나 괴로운 느낌.
기타	-	8가지 표지 외 기타 감정. (선후의 감정 정보 없는 궁금함, 의아함, 묘함 등)

감정 주석 시에는 인간의 감정에 나타나는 복합 감정, 연속 감정을 분석하기 위해 복수 주석을 허용하였다. 이와 더불어 감정의 대상(target)이 되는 표현을 선별하여 함께 주석하였다.

대상(target)은 지시하거나 의미하는 바가 무엇인지 명확하게 파악할 수 있는 구체 또는 추상 명사로 선정하였고, 이때 이를 수식하는 성분은 제외하였다. 문서 내에 명확한 구체 혹은 추상 명사가 없는 경우에는 수식 성분 + 의존 명사 혹은 대명사를 대상(target)으로 하였다. 여러 개의 어절로 구성된 고유명사가 대상(target)일 경우에는 전체 표현을 대상(target)으로 선정하였다. 마지막으로 감정에 대한 대상(target)이 문서 내에 없는 경우에는 대상(target)을 ‘null’로 처리하였다. 아래는 감정 말뭉치 구축 당시 세운 대상(target) 선정 규칙이다.

<표 4> 감정 분석 말뭉치 대상 선정 규칙

idx	감정 대상(target) 선정 세부 기준
1	대상(target)의 범위는 지시하거나 의미하는 바가 무엇인지 명확하게 파악할 수 있는 구체 명사 또는 추상 명사로 선정함. 이때 구체 명사나 추상 명사를 수식하는 성분은 대상(target) 선정에서 제외함.

	<p>[예시] 나온지 엄청 된 영화인데, 어제 잠이안와서 그냥 우연히 다시 봄. 여전히 다시 봐도 감동이고 진짜 대박 슬픔.. → target: 영화</p>
2	<p>문서 내에 대상을 가리키는 명확한 구체 명사 혹은 추상 명사가 없고 ‘수식 성분 + 의존 명사(-것, -거, -게, -분 등)’ 구조의 구문이 감정의 대상이 되는 경우 해당 구문을 대상(target)으로 선정함.</p> <p>[예시] 그~ 여행지에서 가장 인상 깊었던 건 사람들이 진짜 친절한 거였어요. → target: 사람들이 진짜 친절한 거</p>
3	<p>문서 내에 대상을 가리키는 명확한 구체 명사 혹은 추상 명사가 없고 대명사(이거, 저거, 그거 등)가 감정의 대상이 될 경우 해당 표현을 대상(target)으로 선정함.</p> <p>[예시] 저는 아무래도 이게 조금 안타깝긴 하더라고요. → target: 이게</p>
4	<p>동일한 것을 가리키는 추상 명사, 구체 명사, ‘수식 성분+ 의존 명사’ 구문, 대명사 등이 동일 문서 내에 존재할 경우 추상 명사·구체 명사 &gt; 수식 성분 + 의존 명사 &gt; 대명사 순으로 우선순위를 적용하여 대상(target)을 선정함.</p> <p>[예시] 영화 마지막에 나오는 장면 미쳤더라. 주인공 우는 거 보고 나도 울었음. → target: 장면</p>
5	<p>같은 대상을 가리키는 동일 수준 명사(또는 완전히 동일한 명사)가 복수로 등장하는 경우 문서 내에서 가장 먼저 등장하는 것을 대상(target)으로 선정함.</p> <p>[예시] 요즘 내가 보는 소설이 있거든. 근데 그 소설 진짜 읽다가 멈출 수가 없더라. → target: 소설</p>
6	<p>여러 개의 어절로 구성된 고유명사의 경우 전체 표현을 대상(target)으로 선정함. 단, 해당 표현이 하나가 아닌 여러 개의 발화(utterance)에 걸쳐 등장한다면 해당 고유명사의 일부가 포함된 마지막 발화(utterance) 내의 표현만 대상(target)으로 선정함.</p> <p>[예시] 어제 슬픔이여 이제 안녕 듣고 한참 울다가 밤에 잠을 못 잤지. → target: 안녕</p>
7	<p>감정에 대한 대상이 문서 내에 없는 경우 대상(target)은 ‘null’로 처리함.</p> <p>[예시] 너무 아름다워서,,, 화면에 잡힐때마다 와,,, 소리 절로나옴 → target: null</p>

‘국립국어원 2022년 말뭉치 감정 분석 및 연구 보고서(2022)’에 따르면 감정 분석 말뭉치는 주석자에 따른 주관성을 해결하기 위해 일치도 분석을 진행하였다. 일치도는 선발된 10명의 주석자가 동일 자료 1건에 대해 분석을 진행한 후 카파(Fleiss Kappa) 통계량을 도출하였다. 이때, 카파 통계량이 0.4를 초과할 경우 해당 문서를 적합 문서로 채택하였다. 또한, 감정 표지는 적합 문서 중 5명 이상의 분석 결과가 일치한 표지를 최종 채택하였다.

또한 동일 연구에서 트위터 자료 50,000건, 일상 대화 자료 10,000건에 대한 감정 분석 진행한 결과 전체 자료에서 가장 많이 나타난 표지는 ‘기쁨’이며, 가장 적게 나타난 표지는 ‘공포’였다. 감정 표지별 비중은 아래와 같다. 최종적으로 구축된 60,000건의 말뭉치는 JSON의 형태로 납품되었다.

<표 5> 감정 분석 말뭉치 자료 통계 (이영희 외, 2022: 48)

감정 분석표지	일상 대화		트위터		합계	
	분석표지 수 (건)	비중	분석표지 수 (건)	비중	분석표지 수 (건)	비중
기쁨(Joy)	5,710	57.1%	25,256	50.5%	30,966	51.6%
기대(Anticipation)	1,893	18.9%	10,773	21.5%	12,666	21.1%
신뢰(Trust)	414	4.1%	2,646	5.3%	3,060	5.1%
놀람(Surprise)	322	3.2%	3,229	6.5%	3,551	5.9%
혐오(Disgust)	1,296	13.0%	1,810	3.6%	3,106	5.2%
공포(Fear)	521	5.2%	1,015	2.0%	1,536	2.6%
분노(Anger)	144	1.4%	1,979	4.0%	2,123	3.5%
슬픔(Sadness)	1,097	11.0%	3,032	6.1%	4,129	6.9%
감정 없음	-	-	6,801	13.6%	6,801	11.3%

## ○ 감정 분석 말뭉치 검수 방법론

인공 지능의 한국어 처리 능력 평가를 위해 2022 감정 분석 말뭉치를 정비하였다. 정비를 위해 검수 지침을 구축하였으며, 2022 감정 분석 말뭉치의 구축 지침을 뼈대로 하여 주요 주석 요소인 감정, 감정 대상(target), 비식별화를 중심으로 검수 지침을 설계하였다.

### ▷ 감정 라벨 검수 방법

2022 감정 분석 말뭉치는 로버트 플루치(Robert Plutchik)의 감정 체계를 기반으로 8가지 기본 감정(기쁨, 기대, 신뢰, 놀람, 혐오, 공포, 분노, 슬픔)과 ‘기타’로 감정을 주석하였다. ‘기타’는 기본 감정 스펙트럼에서 벗어나 어떠한 표지로도 정의될 수 없는 경우에 사용하는 라벨이다. 감정 주석 시에는 인간의 감정에 나타나는 복합성, 연속성 등을 반영하기 위해 복수 태깅을 허가하였다.

감정 라벨 검수 시에는 기본적으로 구축 지침의 감정 주석 기준을 따르되, 새로 수립한 검수 지침에 따라 세 가지 사항을 중점적으로 정비하였다.

- 첫째, 문서에 드러난 감정이 알맞게 주석되었는지 확인하였다. 이때, ‘감사’, ‘고마움’이 문서에서 드러남에도 ‘joy’로 주석되지 않은 경우를 특히 중점적으로 검수하였다. ‘joy’ 즉 ‘기쁨’의 개념은 한국과 서양이 큰 차이를 보인다. 이에 감정 라벨 검수 시에는 인수한 데이터 내의 ‘감사’ 그리고 ‘고마움’에 대한 여러 주석 사례를 확인한 후 기존 대로 ‘감사’, ‘고마움’을 ‘joy’ 안에 포함하기로 결정하였다. 검수 시 감정이 잘못된 라벨로 주석되어있을 경우 적절한 라벨로 수정을 진행하였다. 이와 관련한 예시는 아래와 같다.

<표 6> 고마움 주석 검수 예시

sentence	emotion
당연함,, 솔직히 극만 좋으면 풀라고 뭐고 필요없음,,, 강 일해주시는거에 감사함...	[기존] joy: False → [수정] <u>joy: True</u> 문서 내 ‘감사’ 및 ‘고마움’이 드러남으로 joy: True로 주석

- 둘째, 문서에 주석된 감정이 하나일 경우, 실제 감정보다 과소하게 주석되었는지 확인하였다. 실제 감정보다 과소하게 주석되었을 때는 추가로 주석을 진행하였다. 아래는 감정이 과소하게 주석된 예시이다.

<표 7> 감정 과소 주석 검수 예시

sentence	emotion
체육대회 완전 짜증나지 않냐 경기 한번 땀 때마다 기력 짹짹빠지고 땀나서 개시름	[기존] anger: True → [수정] <u>disgust: True, anger: True</u> ‘개시름’에 대해 ‘disgust: True’, ‘짜증나지 않냐’에 대해 ‘anger: True’로 주석

- 이때, ‘기대감’을 나타내는 표현이 문서에 등장한 경우 다른 감정이 느껴지는지를 주로 검수하였다. 말뭉치 구축 시 ‘기대감’ 즉 ‘anticipation’은 표현 단위에 의해 주석되었다. 이는 ‘정말 기대돼요!’와 같은 표현이 등장할 때 ‘anticipation’이 주석됨을 의미한다. ‘anticipation’은 ‘감사’, ‘고마움’과 마찬가지로 구축 당시 의미와 관련해 논란이 있었다. 이에 따라 ‘기대감’ 검수 시에는 인수한 데이터의 표현 단위 키워드를 바탕으로 검수를 진행하되, 기타 감정이 등장하면 해당 라벨을 추가적으로 주석하였다.

<표 8> 기대감 주석 검수 예시

sentence	emotion
무척무척 좋아하는 노래도 예정에 있어서 넘 행복하고 기대중임!	[기존] anticipation: True → [수정] <u>joy: True, anticipation: True</u> 기대감과 행복감이 모두 드러나므로 'joy: True, anticipation: True'로 주석

- 셋째, 문서에 주석된 감정이 복수 개일 경우, 실제 감정보다 과도하게 주석되었는지 확인하였다. 실제 감정보다 과도하게 주석되었을 때는 불필요한 주석을 삭제하였다. 아래는 감정이 과대하게 주석된 예시이다.

<표 9> 불필요한 감정 주석 삭제 예시

sentence	emotion
아 영화 너무 재밌었어 진짜 존잼영화임 갓 갓	[기존] joy: True, trust: True → [수정] <u>joy: True</u> 문장 내에 재미만 드러나므로 'joy: True' 이외에 다른 주석을 달지 않음

2022 감성 분석 말뭉치는 Fleiss Kappa 일치도를 기반으로 감정 라벨을 채택하였다. 그러므로 감정 라벨 검수 시에는 최대한 기존 라벨을 존중하였다. 감정 라벨 변경은 검수자 간의 토론을 통해 변경에 대한 전원 동의를 얻었을 때만 이루어졌다.

#### ▷ 감정 대상(target) 검수 방법

2022 감정 분석 말뭉치는 감정과 더불어 감정 대상(target)을 주석 체계에 포함했다. 감정 대상(target)은 화자/작성자가 드러낸 감정의 대상이 되는 표현이다. 감정 대상(target) 검수 시에는 기본적으로 이가 구축 지침에 맞게 잘 주석되었는지 확인하였다. 이때, 주로 살펴본 사항은 다음과 같다.

- 첫째, '고유/구체/추상 명사'의 경우 수식어가 같이 주석되지 않았는지 확인하였다. 구축 지침에 명사의 수식 성분은 감정 대상(target) 주석 시 제외한다고 언급되어 있었기에 이를 중점적으로 검토하였다. 이와 관련한 예시는 아래와 같다.

<표 10> 대상 수식어 주석 검수 예시

sentence	target
영화 리틀포레스트 또 보는데.. 빛나는 김태 리 연기 너 무 좋아요~~	대상(target): '빛나는 김태리 연기' (X), '김태리 연기' (O)

- 둘째, ‘수식성분+의존 명사’의 경우 수식 성분의 길이가 과대/과소하지 않은지 검수하였다. 2022 감정 분석 말뭉치 구축 시, 문장 내에 명확한 고유/구체/추상 명사가 없을 때는 ‘수식성분+의존 명사’를 감정 대상(target)으로 주석하였다. 이때 주석자의 잘못된 판단으로 인해 감정 대상(target)이 너무 길게 주석되진 않았는지, 너무 짧게 주석되진 않았는지를 중심으로 문서를 살펴보았다. 관련 예시는 아래를 참조할 수 있다.

<표 11> 수식 성분 길이 검수 예시

sentence	target
좀비 나오는 드라마 난 너무 무서워서 못보겠는데..티타 사람들에게 인기있는거보면 신기..	대상(target): '티타 사람들에게 인기있는거' (X), '인기있는거' (O)

- 셋째, 주석 우선순위에 따라 감정 대상(target)이 잘 주석되었는지 검토하였다. 인계받은 말뭉치는 동일 지시 대상에 대해 ‘고유/구체/추상 명사’, ‘수식성분+의존 명사’, ‘대명사’가 함께 등장할 경우 ‘고유/구체/추상 명사’ > ‘수식성분+의존 명사’ > ‘대명사’ 순으로 주석을 시행하였다. 이에 따라 우선순위에 어긋나는 감정 대상(target)은 없는지 검수를 진행하였다. 다음은 감정 대상(target)의 우선 순위가 잘못된 예시이다.

<표 12> 감정 대상 우선 순위 검수 예시

sentence	target
남편의 데이오프를 맞이하여 해결을 떠올리며 스시 먹고 거기 경민 미술관 신나게 총 총 미술관이 내 안식처임	대상(target): '거기' (X), '경민 미술관' (O)

- 넷째, 동일 층위의 표현이 복수 개 등장하는 경우 먼저 등장한 표현이 감정 대상(target)으로 주석되었는지 확인하였다. 구축 지침을 살펴보면 한 대상을 가리키는 동일 수준 표현 혹은 같은 표현이 복수로 등장할 경우 가장 먼저 등장한 표현을 감정 대상(target)으로 선정한다는 규칙이 존재한다. 이를 기반으로 주석 시 우선 등장 대상(target) 선정 규칙이 지켜졌는지 검토하였다.

<표 13> 우선 등장 대상 선정 규칙 검수 예시

sentence	target
니노 목소리 조아해 ,, 니노 목소리 들으면 녹아벌임	대상(target): 가장 먼저 등장한 '니노 목소리' → 'begin'과 'end'를 잘 살펴보아야 함



- 다섯째, 다어절 고유명사를 규칙에 따라 전체/부분이 잘 주석되었는지 점검하였다. 구축 시 여러 개의 어절로 구성된 고유명사는 전체 표현을 감정 대상(target)으로 선정하였다. 단, 해당 표현이 여러 개의 발화에 걸쳐 등장할 경우 해당 고유명사 일부가 포함된 마지막 발화 내 표현만을 감정 대상(target)으로 지정하였기에 이를 주의하여 문서를 점검하였다. 아래는 다어절 고유명사의 주석 관련 예시이다.

<표 14> 다어절 고유명사 주석 검수 예시

sentence	target
집에서 혼자 보는 하지만 난 치어리더인 걸 대존잼	대상(target): '하지만 난 치어리더인 걸' (X), '치어리더인 걸' (O)

- 여섯째, 감정 대상(target)이 'None'일 때 대상으로 주석할 수 있는 표현이 문서 내에 있는지 다시 한번 살펴보았다. 구축 지침에 따르면 감정에 대한 대상이 문서 내에 없는 경우 감정 대상(target)을 'None'으로 처리해야 한다. 주석의 정확도를 높이기 위해 감정 대상(target)이 'None'으로 표기된 경우 주석할 대상이 정말로 없는지 다시 한번 확인하였다.

감정 대상(target) 검수는 감정 라벨과 마찬가지로 최대한 기존 라벨을 존중하였다. 추가로 감정 대상(target)이 구축 지침 기준에 맞게 잘 주석되었는지 외에 감정 대상(target)의 스패(span)과 숫자 포함 여부도 검수를 진행하였다. 스패(span)에 대해서는 감정 대상(target)의 스패(span)이 과소 혹은 과대하게 주석된 사례를 직접 수정하였다. 단, 이때 감정 대상(target)의 핵어(head)로 보이는 단어가 이미 포함되어있는 경우에는 수정을 진행하지 않았다.

<표 15> 스패 과소/과대 검수 예시

sentence	target
후..... 이번 프리큐어..... 후..... 코코네랑 팼팼 넘귀여워 &others&	대상(target): '팼팼' (X), '코코네랑 팼팼' (O)
연경언니 팔 스트레칭 잠깐 따라했는데 어 깨에서 후두두두두 소리 누가 총 쏘는 줄	대상(target): '어깨에서 후두두두두 소리' (X), '후두두두 두 소리' (O)

숫자가 감정 대상(target)으로 주석된 경우에 이를 감정 대상(target)에서 제외하거나 수정하였다. 즉, 숫자를 감정 대상(target)으로 인정하지 않았다.

<표 16> 숫자 대상 주석 검수 예시

sentence	target
tvn 어쩌다사장2 봤는데.. 1은 재밌게 봤는데 2는 음 .. 1보다는 재미가 덜한것 같아요.. 뭔가 집중이 덜되는 듯한 느낌이랄까!?	대상(target): '1' (X), None(O)

## ▷ 비식별화 검수 방법

2022 감정 분석 말뭉치는 일부 감정 대상(target)에 대해 비식별화 처리를 하였다. 비식별화는 공격 대상(target)이 특정 인물 또는 집단이어서 개인정보 노출 위험이 있는 경우에만 진행하였다. 비식별화는 두 가지의 기준에 따라 검수하였다.

<표 17> 비식별화 표지 및 설명

비식별화 처리 유형	비식별화 표지	설명
이름	&name&	개인의 실명 (정치인, 연예인 등 공인·유명인 제외, 실존 인물이 아닌 캐릭터·극중 인물 제외)
온라인 계정 (아이디)	&account&	트위터 등 특정 사이트의 온라인 계정
고유 식별 번호 (주민등록번호)	&social-security-num &	개인의 주민등록번호
전화 번호	&tel-num&	휴대폰 번호, 사업장 번호 등
카드 번호	&card-num&	신용카드 번호 등
기타 번호	&num&	비밀 번호 등 기타 비식별화 대상 번호
주소	&address&	동 이하의 상세 주소
출신 및 소속	&affiliation&	개인의 출신 및 소속
기타 비식별화 필요 항목	&others&	위 항목 외 기타 비식별화 대상

- 첫째, 감정 대상(target)에 대한 비식별화가 잘 되어있는지 점검하였다. 만일 감정 대상(target)에 대해 비식별화가 잘 되어있을 경우, 해당 비식별화 표지가 문서(sentence)에도 잘 되어있는지 재확인하였다.
- 둘째, 비식별화 처리 유형에 적절한 비식별화 표지로 마스킹 되어있는지를 확인하였다. 만약 비식별화 처리 유형과 비식별화 표지가 불일치할 경우, 처리 유형에 맞는 비식별화 표지로 교체하였다.

## ○ 검수 결과

2022 감정 분석 말뭉치에 대해 수동 검수를 진행하였다. 그 결과 감정 라벨 변경이 필요한 경우가 15건, 감정 대상(target) 변경이 필요한 경우가 156건, 비식별화 변경이 필요한 경우가 1건이었다. 감정 라벨 변경은 검수자 전원 동의한 경우에만 이루어졌기에 그 개수가 많지 않았다. 감정 라벨 변경은 주로 (1) 감정이 알맞게 주석되지 않은 사례 중 ‘감사’, ‘고마움’이 문서에 드러남에도 ‘joy’가 주석되지 않은 경우, (2) 실제 감정보다 과소하게 주석된 사례 중 ‘기대감’과 그 외 감정이 문서에 드러남에도 ‘anticipation’만 주석된 경우, (3) 실제 감정보다 과도하게 주석된 사례에 대해서 이루어졌다. 실제 감정 라벨 변경 예시는 아래와 같다.

<표 18> 감정 분석 말뭉치 joy 주석 검수 사례

sentence	emotion
개인적으로 <name1> 유튜브 채널을 즐겨 보는데 영상 올라와서 볼 때마다 <name2> 님의 에너지를 전달 받는 거 같아서 너무 좋다ㅋㅋ 초기에 구독한 채널이라 쭈욱 봤는데 감량도 엄청 하셨고.. 넘 대단한 거 같다 앞으로도 계속 흥하셨으면..	[기존] joy: False, trust: True → [수정] joy: True, trust: True
내 인생에서 제일 힘들었을 때 너를 만나고, 너의 말들로 내가 일어서고, 너의 노래와 너의 무대로 힘을 얻어서 살아갈 수 있었고 .. 그래서 난 이 고마움을 오랜 시간에 걸쳐 너에게 사랑으로 갚을 거야	[기존] None → [수정] joy: True, anticipation: True
21. 그리고 정말로... 쉬어야지 하는 타이밍에 기작이 없어서 맘편히 쉬고 있었음.. 너무 좋았음 ..... 그래서 한 6,7월까지 쉬면 좋겠다 하는 찰나에 무슨 깜짝 차기작 다다음주에 개막이래 ..... 조금 슬퍼짐 나는 진짜 휴식이필요했걸랑 ... 그러니까 적당히 보겠습니다	[기존] sadness: True → [수정] sadness: True, surprise: True

감정 대상(target) 변경은 (1) 구축 지침 기준에 맞지 않는 사례, (2) 스펠(span)이 잘못 주석된 사례, (3) 숫자가 포함된 사례에 대해서 이루어졌다. 대부분의 변경은 스펠(span)의 과대 및 과소와 관련하여 이루어졌다. 구축 지침 기준에 맞지 않는 경우, 숫자가 포함된 경우는 그 수가 많지 않았다. 또한, 아예 감정 대상(target) 설정이 잘못된 경우가 있어 이를 수정하기도 하였다. 아래는 실제 감정 대상(target) 변경 예시이다.

<표 19> 감정 분석 말뭉치 감정 대상 주석 검수 사례

sentence	target
혁 그레타 거윅 <바비> 촬영 끝.. 열린 주세 요πππ &others&	[기존] 그레타 거윅 <바비 → [수정] <바비>
9,10위팀 그들만의 코시하고 앓았네ㅋㅋㅋ ㅋㅋㅋㅋ	[기존] 9 → [수정] 9,10위팀
짱 멋진 여자 선수들 인스타나 기사 찾아보 면서 결혼/애인 유무 찾아보면서 혼인 신고 올리려는 사람들 6486734명이라 너무 웃 김.... 귀여워 ㅋㅋㅋ	[기존] 6486734명 → [수정] 사람들

## ○ 정비 완료 말뭉치 예시

정비가 완료된 말뭉치는 기본적으로 json 형식으로 정비된다. 해당 말뭉치를 과제 수  
행을 위해 jsonl 형식으로 변환하여 평가체계 참가자들이 내려받을 수 있도록 하였다.  
아래는 jsonl 예시이다.

<표 20> 감정 분석 JSONL 말뭉치

```
{
  "id": "nikluge-2023-ea-train-000001",
  "input": {
    "form": "하... 근데 준프사 너무 고소각임...",
    "target": {
      "form": "준프사",
      "begin": 8,
      "end": 11
    }
  },
  "output": {
    "joy": "True",
    "anticipation": "False",
    "trust": "False",
    "surprise": "False",
    "disgust": "False",
    "fear": "False",
    "anger": "False",
    "sadness": "False"
  }
},
{
  "id": "nikluge-2023-ea-train-000002",
  "input": {
    "form": "2기였나 지은북이랑 4기 메거진은 지금도 읽는데",
    "target": {
      "form": "4기 메거진",
      "begin": 11,
      "end": 17
    }
  },
  "output": {
    "joy": "True",
    "anticipation": "False",
    "trust": "False",
    "surprise": "False",
    "disgust": "False",
    "fear": "False",
    "anger": "False",
    "sadness": "False"
  }
},
{
  "id": "nikluge-2023-ea-train-000003",
  "input": {
    "form": "흐아아아아악 흐아아아아악악악악 뒤흔 손차이가 절케  
난다니 알고는 있었지만 놀랍다 | 아아악악",
    "target": {
      "form": "뒤흔 손차이",
      "begin": 17,
      "end": 23
    }
  },
  "output": {
    "joy": "False",
    "anticipation": "False",
    "trust": "False",
    "surprise": "True",
    "disgust": "False",
    "fear": "False",
    "anger": "False",
    "sadness": "False"
  }
},
{
  "id": "nikluge-2023-ea-train-000004",
  "input": {
    "form": "도브가 반반을 가고 프린스가 안밀린다면 하는 상상",
    "target": {
      "form": null,
      "begin": null,
      "end": null
    }
  },
  "output": {
    "joy": "False",
    "anticipation": "True",
    "trust": "False",
    "surprise": "False",
    "disgust": "False",
    "fear": "False",
    "anger": "False",
    "sadness": "False"
  }
},
{
  "id": "nikluge-2023-ea-train-000005",
  "input": {
    "form": "담주에 티켓팅 공지 뜨고 다담주에 티켓팅할 느낌",
    "target": {
      "form": "티켓팅 공지",
      "begin": 4,
      "end": 10
    }
  },
  "output": {
    "joy": "False",
    "anticipation": "True",
    "trust": "False",
    "surprise": "False",
    "disgust": "False",
    "fear": "False",
    "anger": "False",
    "sadness": "False"
  }
},
{
  "id": "nikluge-2023-ea-train-000006",
  "input": {
    "form": "천러 춤 존나존나존나 늘은거 볼때마다 강 하염없이  
눈물 남 ... 포지션 명확하니까 굳이 춤에 욕심 내지 않아도 됐을텐데 욕심 갖고 꾸준히 연습해온 결과잖아 아 존  
나 기특해 ㅁㅁ 천러의 욕심이랑 독기는 너무 건강하고 언제나 결과를 동반함 그게 너무 조음",
    "target": {
      "form": "천러 춤",
      "begin": 0,
      "end": 4
    }
  },
  "output": {
    "joy": "True",
    "anticipation": "False",
    "trust": "True",
    "surprise": "False",
    "disgust": "False",
    "fear": "False",
    "anger": "False",
    "sadness": "False"
  }
},
{
  "id": "nikluge-2023-ea-train-000007",
  "input": {
    "form": "도일이 개눈깔 뜰 때마다 나는 설레...",
    "target": {
      "form": "도일이 개눈깔",
      "begin": 0,
      "end": 7
    }
  },
  "output": {
    "joy": "True",
    "anticipation": "False",
    "trust": "False",
    "surprise": "False",
    "disgust": "False",
    "fear": "False",
    "anger": "False",
    "sadness": "False"
  }
}
```

```
{
  "id": "nikluge-2023-ea-train-000008",
  "input": {
    "form": "나 지금 너무 감동적이라 눈물나는데 &others&",
    "target": {
      "form": null,
      "begin": null,
      "end": null
    },
    "output": {
      "joy": "True",
      "anticipation": "False",
      "trust": "False",
      "surprise": "False",
      "disgust": "False",
      "fear": "False",
      "anger": "False",
      "sadness": "False"
    }
  },
  "id": "nikluge-2023-ea-train-000009",
  "input": {
    "form": "역시 브로커... 1년 기다린 그 기대감을 저버리지 않는  
구나□",
    "target": {
      "form": "브로커",
      "begin": 3,
      "end": 6
    },
    "output": {
      "joy": "True",
      "anticipation": "True",
      "trust": "False",
      "surprise": "False",
      "disgust": "False",
      "fear": "False",
      "anger": "False",
      "sadness": "False"
    }
  }
}
```

## 2.2. 이야기 완성 말뭉치

### ○ 인수 말뭉치 분석

평가 체계용 이야기 완성 말뭉치는 <2022년 이야기 완성 평가 말뭉치 연구 분석> 사업을 통해 구축된 말뭉치를 기반으로 하였다. 이야기 완성 말뭉치는 첫 번째 문장과 세 번째 문장 사이에 들어가기에 적절한 문장을 선택하는 ‘이야기 완성 추론 말뭉치’와 적절한 문장을 생성하는 ‘이야기 완성 생성 말뭉치’ 두 가지 형태로 구성하여 정비하였다. 다만 이번 평가 체계용 이야기 완성 말뭉치는 생성에 초점을 맞추고 있으므로 정비 과정에서 ‘이야기 완성 생성 말뭉치’, 특히 첫 번째 문장과 세 번째 문장을 중정적으로 검수를 진행하였다. 결과물 평가 과정과 추후 활용 가능성을 고려하여, 추론과 생성 말뭉치의 나머지 문장들에 대해서도 검수를 진행하였다.

‘2022년 국립국어원 이야기 완성 평가 말뭉치 연구 분석 보고서(송상헌 외, 2022)’에 따르면 2022년 국립국어원 이야기 완성 평가 말뭉치는 이야기 완성 평가 방법을 최초로 제안했던 LSDSem 2017 Shared Task의 Story Cloze Test와는 형식에서 두 가지 차이를 보인다. 먼저 기존 Story Cloze Test에서는 총 네 문장의 이야기가 주어지면 두 개의 가설 문장 중 주어진 문장에 적절하게 이어질 수 있는 한 문장을 고르는 방식으로 과업이 구성되었다.

반면 이번 2023년 평가 체계에 활용될 2022년 국립국어원 이야기 완성 평가 말뭉치는 총 세 문장이 하나의 이야기를 이룬다. 즉, 가설 문장 두 개를 제외하고 맥락을 제시하는 두 개의 문장이 주어지며, 그 주어진 맥락에 알맞은 하나의 문장을 두 개의 가설 문장 중에서 고르는 것이 과업의 기본적인 구조이다. 두 번째 차이는, 가설 문장의 위치이다. Story Cloze Test의 경우, 주어진 네 개의 문장 이후에 이어지기에 적절한 문장을 고르는 형태였다면, 2022년 이야기 완성 평가 말뭉치에서는 가장 마지막에 올 문장이 아닌, 첫 번째 문장과 마지막 문장 사이에 들어가기에 적절한 형태의 문장을 고르게 하는 방식으로 구성되어 있다.

이 데이터는 '문장 생성 > 가설 부착 > 담화 평정 > 모델 검증'으로 이루어진 평가용 데이터이다. 결과물은 이야기와 가설이 각기 json의 형식으로 되어 있다. 데이터의 갯수는 이야기와 가설이 각 150,176개 그리고 150,276개이다. 이야기는 시간의 흐름에 따라 기술된 세 문장으로 구성되어 있다(S1, S2, S3). 이러한 일련의 문장을 생성하기 위해 복수의 방법을 사용하였는데 각기, 문장생성 방식 3종: 자유창작(우리말샘 활용), 키워드(한국어 교육용 자료에서 발췌 및 확장), 그림기반(2021년 그림기반 유사문장 활용) 방식이다.

또한 한 문장은 최소 4어절에서 최대 12어절로 구성하며, 모두 과거시제로 기술한다는 기본 원칙을 두었다. 가설 부착은 위 세 문장에서 가운데 S2를 감추고, 두 문장 S1, S3의 가운데 등장하기에 알맞은 문장과 그렇지 않은 문장 2개를 적절 가설(H1)과 부적절 가설(H2)로 추가 생성하는 과정이다. 이때 가운데에 새롭게 들어가는 H1과 H2가 너무 동떨어져 있어서는 안된다. 즉, 구조와 어조를 크게 변형하지 않게 체하여 인공 지능이 가설을 구별하기 어렵게 만드는 것을 목적으로 하였다.

### ○ 이야기 완성 말뭉치 검수 방법론

이야기 완성 말뭉치의 경우 효율적인 검수 절차를 위해 단발어를 추출하여 오타자, 띄어쓰기 및 기타 비표준적인 표현을 수정하였다. 명백한 오타자인 경우 단발어 목록에서 바로 교정을 진행하였으며, 명백하지 않은 명사 및 표현형에 대해서는 검수자가 직접 맥락을 살펴 수정 필요성을 판단하였다. 메타 데이터(metadata) 중 ‘title’, ‘type’, ‘clue1’, ‘clue2’의 경우 말뭉치 구축 과정에서 작업자의 문장 생성을 돕기 위해 사용된 것들로, 목표로 하는 과업의 수행에 활용되지 않는 요소이기 때문에 검수 대상에 포함시키지 않았다.

검수 과정에서 오타자임이 확실하지 않지만 익숙하지 않은 전문용어로 추론되는 경우 맥락을 통해 확인한 뒤 검수를 진행하였다. 일례로, 제라늄과 같이 해당 표현이 학명 및 전문용어임을 확인할 수 있는 경우 수정을 진행하지 않았다.

외래어의 경우, 주어진 맥락에서 문장의 이해 여부를 해치는지 여부를 우선적인 기준으로 두고 검수를 진행하였다. 일례로, ‘알러지’라는 표현 규범 표기는 ‘알레르기’이다. ‘알러지’라는 표현이 <표준국어대사전>에는 등재되어 있지 않지만, <우리말샘>에는 등재되어 있다. 또 일반적으로 두 가지 표현 모두 언중에게 수용 가능하고, ‘알레르기’가 아닌 ‘알러지’를 사용해도 문장의 의미나 맥락을 이해하는 데 문제가 되지 않으므로 이는 수정 대상에 포함되지 않는다. 단, 문장의 이해 여부에 영향을 미치는 외래어 오타자를 수정해야 하는 경우에는 외래어표기법에 맞추어 수정하였다.

띄어쓰기의 경우, 해당 말뭉치가 언어 모델의 학습과 평가에 활용되는 데 있어 그 적형성이 핵심적으로 작용한다고 보기 어렵다. 일례로 본용언과 보조 용언의 경우 붙여 쓰기와 띄어 쓰기가 모두 허용되기도 한다. 따라서 띄어쓰기는 문장의 이해 여부를 해치는지를 가장 우선적인 기준으로 삼아 검수를 진행하였고, 문장의 중의적인 해석을 발생시키지 않는 경우 원본의 형태를 유지하였다. 이와 관련된 사례는 아래와 같다.

1) 보조 용언

그는 담벼락에 금이 가있는 것을 발견했다.

2) 부정어

나는 바로 전화를 걸어서 전화를 계속 안받은 것에 대해 사죄했다.

3) 관형어

그래서 그는 피부에 선크림을 한겹 더 덧칠했다.

4) 복합명사

일연이 땀을 뻘뻘 흘리며 회의장소에 들어왔다.

5) 기타

가족들은 오히려 이때까지 고생 많았다며 안아주었다.

다만 그 밖에 비표준적인 표현의 경우, 문장의 의미나 글의 맥락을 이해하는데 문제를 발생시키지 않더라도 수정을 진행하였다. 예를 들어, “승희는 엄마에게 학원을 안갈려고 거짓말을 쳤다.”라는 표현에서 ‘안갈려고’는 ‘안 가려고’의 잘못된 표현이다. 사람이라면 이를 ‘가다’로 바꾸어 이해할 수 있지만, 인공지능 모델의 경우 형태소의 원형을 ‘가다’가 아닌 ‘갈다’로 복원할 수 있으므로 비표준적인 표현들은 모두 수정 대상에 포함되었다. 비표준어의 경우 우리말샘 사전에서 인덱싱하는 수준을 기준으로 검수하는 것을 목표로 하였다.

더불어 이야기 완성 말뭉치의 경우, 하나의 일관된 이야기를 구성하는 세 개의 문장이 하나의 세트로 구성되어 있다는 측면에서 문장간 연결성 및 논리성을 검토할 필요가 있다. 따라서 오탈자를 수정하는 과정에 세 문장의 논리적 일관성에 명백한 오류가 존재하는 경우 수정을 진행하였다. 이때 문장들의 의미나 맥락에 변화를 가져올 수 있는 정도의 수정은 최대한 지양하는 방향으로 검수를 진행하였다.

## ○ 검수 결과 및 통계

이야기완성 말뭉치는 세 개의 문장이 논리적으로 일관된 하나의 이야기를 전달할 수 있는 형태로 구성된 말뭉치이다. 이는 첫 번째 문장과 세 번째 문장만을 활용하여 가추적인 추론을 통해 그 사이에 들어가기에 적절한 문장을 생성하는 과업을 평가하기 위해 고안된 형태이다. 각각의 이야기 세트에서 적절한 문장을 중심으로 세 문장의 논리적 인과에 명백한 오류가 존재하는 경우에 한하여 문장을 수정하였고, 그 외에는 단발어를 중심으로 오탈자, 띄어쓰기 및 기타 비표준적인 표현을 수정하였다. 이야기완성 말뭉치의 구성은 다음과 같다.



<표 21> 이야기 완성 말뭉치 구성

```
"document": [  
  {  
    "id":  
    "metadata": {  
      "title":  
      "type":  
      "clue1":  
      "clue2":  
    },  
    "sentences": {  
      "sentence1":  
      "sentence2":  
      "sentence3":  
    }  
  }  
]
```

metadata 중 title, type, clue1, clue2는 과업에 활용되지 않는 데이터로 검수 대상에 포함하지 않았으나, 데이터의 무결성을 위해 값의 존재 여부와 형식을 확인하였다. 이야기 완성 말뭉치의 예시는 다음과 같다.

- (1) sentence 1: 민서는 친구네 집에 초대를 받아 방문했다.  
sentence 2: 친구가 직접 만든 음식을 한 상 가득 차려서 대접했다.  
sentence 3: 민서는 음식을 먹고 친구에게 엄지손가락을 들어 보였다.

(1)의 사례의 경우 세 번째 문장에 민서가 음식을 먹고 친구에게 엄지손가락을 들어 보이는 방식으로 칭찬을 표하는 상황이 제시되어 있다. 다만 첫 번째 문장에는 이와 관련된 내용이 언급되어 있지 않으므로 세 번째 문장에 앞서 민서가 음식을 대접받은 상황이 제시되지 않으면 이야기의 연결이 어색해진다. 이러한 점을 고려하여 첫 번째 문장과 세 번째 문장 사이에 들어가기 적절한 문장으로 친구가 민서에게 음식을 만들어 대접한 내용이 포함된 문장이 제시되어 있다.

- (2) sentence 1: 민수는 요즘 보드 타는 것에 취미를 들였다.  
sentence 2: 그는 맨날 보드를 타다가 다쳐서 나타났다.  
sentence 3: 하지만 보드 타는 것을 멈추지 않았다.

(2)의 사례의 경우 세 번째 문장에 제시된 접속 부사 ‘하지만’의 의미를 고려하여 문장이 생성되어야 한다. 첫 번째 문장에서 민수가 보드에 취미를 들였다는 점이 진술된 상태이므로, 민수가 보드 타는 것을 멈출만한 적절한 이유가 제시되지 않는다면 세 번째 문장과의 내용 연결이 어색해진다. 이러한 점을 고려하여 첫 번째 문장과 세 번째 문장 사이에 들어가기 적절한 문장으로 민수가 보드를 타다가 다쳐서 나타난 상황이 포함된 문장이 제시되어 있다.

- (3) sentence 1: 나는 인공 지능 스피커에게 음악을 재생하라는 명령을 내렸다.  
sentence 2: 인공지능 스피커는 내 명령을 이해하지 못했다고 말했다.  
sentence 3: 나는 이런 모습을 보며 인공 지능 스피커를 산 것을 후회했다.

(3)의 사례의 경우 세 번째 문장에 제시된 ‘후회하다’라는 어휘의 의미를 고려하여 두 번째 문장이 생성되어야 한다. 첫 번째 문장에서 인공 지능 스피커에서 구체적인 요청을 한 상태이므로, 인공 지능 스피커를 산 것을 후회할 만한 상황이 제시되지 않는다면 이야기의 연결이 어색해진다. 이러한 점을 고려하여 첫 번째 문장과 세 번째 문장 사이에 들어가기 적절한 문장으로 인공 지능 스피커가 명령을 이해하지 못해 제 기능을 다하지 못한 상황이 포함된 문장이 제시되어 있다.

전체 정비 결과 말뭉치 수정 건수는 1,597건으로, 4개 오류에 대해 이를 교정하는 정비를 진행하였다. 최종적으로 정비를 진행한 결과 이야기 완성 문장 세트는 150,175건, 전체 단발어는 110,740건으로 집계되었다. 아래는 세부 오류 유형과 해당 오류 수정 건수를 나 타낸 것이다.

<표 22> 이야기 완성 말뭉치 검수 유형 및 건수

수정 유형	수정 건수
명백한 띄어쓰기 오류	978건
명백한 고유명사 오류	10건
확인이 필요한 고유명사	122건
기타 확인이 필요한 경우	487건
합계	1,597건

## ○ 정비 완료 말뭉치 예시

<표 23> 이야기 완성 JSONL 말뭉치

```
{
  "id": "nikluge-2023-sc-train-000001",
  "input": {
    "sentence1": "시은이는 다음 주의 여름 휴가 이전에 기분을 전환하고 싶었다.",
    "sentence3": "예쁘게 꾸민 손톱을 보며 여행을 갈 생각에 한층 더 들떴다.",
    "output": "그래서 네일샵에 가서 예쁘게 손톱을 칠했다."
  }
},
{
  "id": "nikluge-2023-sc-train-000002",
  "input": {
    "sentence1": "우리는 의자를 큰 책상 주위에 빙 둘러놓았다.",
    "sentence3": "학생들이 모두 착석한 뒤 회의가 시작되었다.",
    "output": "그러자 의자에 학생들이 하나 둘 앉기 시작했다."
  }
},
{
  "id": "nikluge-2023-sc-train-000003",
  "input": {
    "sentence1": "우진이는 폭우로 독서실이 침수되었다는 소식을 들었다.",
    "sentence3": "다행히 우진이가 다니는 독서실은 무사해서 우진이는 안도했다.",
    "output": "그래서 우진이가 다니는 독서실도 침수되었을까 봐 걱정됐다."
  }
},
{
  "id": "nikluge-2023-sc-train-000004",
  "input": {
    "sentence1": "나는 줄을 서서 배식을 기다렸다.",
    "sentence3": "나는 밥을 국에 말아 먹고, 반찬은 모두 버렸다.",
    "output": "배식을 받아보니 내가 싫어하는 반찬만 있었다."
  }
},
{
  "id": "nikluge-2023-sc-train-000005",
  "input": {
    "sentence1": "그는 손빨래해야 하는 옷을 건조기에 넣어버렸다.",
    "sentence3": "그래서 결국 그는 다시 옷을 구매해야만 했다.",
    "output": "그러자 그가 산 옷이 하루아침에 줄어들었다."
  }
},
{
  "id": "nikluge-2023-sc-train-000006",
  "input": {
    "sentence1": "나는 외국의 한 공항에 도착했다.",
    "sentence3": "나는 바로 달려가 내가 한국어를 읽을 수 있다고 말했다.",
    "output": "그때 공항에서 한국어를 읽을 수 있는 사람을 찾았다."
  }
},
{
  "id": "nikluge-2023-sc-train-000007",
  "input": {
    "sentence1": "어린 아이가 자신의 방 안에서 턱을 꺾은 채 공부하고 있었다.",
    "sentence3": "아이는 문제집을 들고 언니의 방에 들어가 도움을 요청했다.",
    "output": "그 아이는 어려운 문제를 만나서 한참을 고민했다."
  }
},
{
  "id": "nikluge-2023-sc-train-000008",
  "input": {
    "sentence1": "애인과 데이트를 하는 날이라 하이힐을 신고 외출했다.",
    "sentence3": "하이힐을 버리고 근처 신발 가게에서 아무 신발이나 샀다.",
    "output": "약속 장소에 가다가 하이힐 굽이 하수구에 끼었다."
  }
},
{
  "id": "nikluge-2023-sc-train-000009",
  "input": {
    "sentence1": "선배가 나에게 보험 사기를 쳤다는 사실을 알았다.",
    "sentence3": "그러나 선배는 전화를 받지 않았다.",
    "output": "나는 배신감에 떨며 선배에게 전화했다."
  }
}
```

## 2.3. 그림 기반 유사 문장 말뭉치

### ○ 인수 말뭉치 분석

그림 기반 유사 문장 말뭉치는 이미지에서 OCR로 텍스트를 추출한 결과로 말뭉치에 OCR 정보 포함된다. 텍스트를 추출한 후에는 OCR 텍스트를 바탕으로 키워드 선정하고 키워드를 이용하여 기본 문장 생성한 후 기본 문장을 바탕으로 유사 문장 생성하는 방식으로 구축된다.

2023년 그림 기반 유사 문장 생성 말뭉치에서는 ‘2022 표 기반 유사 문장 말뭉치’의 1만건 데이터를 모두 활용하였다. 또한 2023년 데이터 세트는 2022년 평가 세트와 마찬가지로 학습(train), 개발(dev), 시험(test) 데이터 세트를 8 : 1 : 1의 비율로 분할하여 구성하였다. 데이터 세트는 무작위로 분할하되, 통계적으로 데이터 세트 간에 편향 없이 유사성을 가지도록 분할하였다. 출력(output) 값의 분포와 ‘문맥(context)’ 및 ‘프롬프트(prompt)’의 문자 개수가 유사할 수 있도록 분할하는 것이 목적이며, 이를 위해 원 데이터의 통계적 특성을 검토한다.

2021 그림 기반 유사 문장 말뭉치의 경우 기준이 되는 문장 1은 평균 40글자, 문장 2~5는 평균 30, 28, 26, 33글자로, 참조가 되는 문장 1이 상대적으로 긴 것으로 확인하였다. 구체적으로 국어원에서 인수한 데이터 명세는 아래와 같다.

#### ■ 국어원에서 수령한 데이터 건수(5/10)

- 말뭉치 JSON 파일: 1건 (14MB, id: GIPS2202305020)
- 그림 JPG 파일: 9168건 (34GB)
- 국어원 230611 수정본 수령: OCR 정보를 수정한 JSON 말뭉치 파일 수령
- 그림 ID별 데이터 개수

<표 24> 그림 기반 말뭉치 그림 ID별 데이터 개수

id	개수
P0xxxx	136
P1xxxx	2411
P2xxxx	2895
P3xxxx	2683
P4xxxx	455
PKxxxx	588

## ○ 그림 기반 유사 말뭉치 검수 방법론

### ▷ 전체 검수 방법론

한글 파일을 이용하여 전체 문장의 맞춤법을 검수, 수정한 후 깃허브(Github) 히스토리 를 참조하여 교차 검수하였다.

```

P00001 worker3 하늘색 테두리 안내판에는 2인승이라고 적혀 있는데, 그 옆에 띄워져 있는 것은 다양한 색깔의 우산들이다.
P00001 worker4 안내판은 하늘색 테두리에 2인승이라고 적혀 있으며, 그 옆에 띄워져 있는 것은 다양한 색의 우산들이다.
P00002 ref 나무 옆에 있는 갈색 껍탈은 대왕참과 소왕참으로 가는 방향을 알려주고 있다.
- P00002 worker1 대왕참과 소왕참으로 가는 방향은 나무 옆에 있는 갈색 껍탈이 알려준다.
+ P00002 worker1 대왕참과 소왕참으로 가는 방향은 나무 옆에 있는 갈색 껍탈이 알려준다.
P00002 worker2 갈색 껍탈이 알려주는 것은 대왕참과 소왕참으로 가는 방향이고, 껍탈은 나무 옆에 있다.
P00002 worker3 대왕참과 소왕참으로 가는 방향을 알려주고 있는 것은 나무 옆에 있는 갈색 껍탈이다.
P00002 worker4 나무 옆 껍탈은 갈색이며, 대왕참과 소왕참으로 가는 방향을 알려준다.
P00006 worker3 사람 한 명이 안에 있는 한옥 모양의 건물에 붙어 있는 갈색 간판에는 고한정이라고 적혀 있다.
P00006 worker4 안에 한 명의 사람이 있는 한옥 모양의 건물에는 간판이 부착되어 있는데, 갈색이며 고한정이라고 적혀 있다.
P00007 ref 유리창에는 파란색 장애인 스티커와 고정문이니 열문을 이용하라고 적혀 있는 흰색 안내문, 감시용 카메라 녹화 중임을 알리는 흰색 껍탈 등이 붙어 있다.
- P00007 worker1 파란색 장애인 스티커와 고정문이니 열문을 이용하라고 안내하는 흰색 안내문, 감시용 카메라가 녹화 중임을 안내하는 흰색 껍탈 등이 유리창에 붙어 있다.
+ P00007 worker1 파란색 장애인 스티커와 고정문이니 열문을 이용하라고 안내하는 흰색 안내문, 감시용 카메라가 녹화 중임을 안내하는 흰색 껍탈 등이 유리창에 붙어 있다.

```

[그림 2] 깃허브 히스토리

최종 검토는 작업자 3인이 오류 문장에 대해 논의 및 검토 후 검토한 내용을 반영하였다. 수정 후 수정 파일(Diff to HTML)을 별도로 제출하였다.

### ▷ 유사 문장 검수 방법론

그림 아이디, 작업자 구분 표시, 유사 문장의 3컬럼으로 구성된 테이블형 데이터로 변환하여 진행하였다.

	A	B	C
1	P11601	ref	엘레베이터 버튼 위에 붙어 있는 안내문은 유모차와 장애인 우선 탑승을 안내하고 있다.
2	P11601	worker1	엘레베이터 오른쪽에 붙어 있는 안내문에는 유모차와 장애인이 우선 탑승할 수 있도록 안내하고 있다.
3	P11601	worker2	엘레베이터 오른쪽에 부착된 안내문은 유모차, 장애인 우선 탑승을 안내하고 있다.
4	P11601	worker3	2호기 표지판 아래에 부착된 안내문은 유모차, 장애인 우선 탑승을 지시하고 있다.
5	P11601	worker4	엘레베이터 층수 버튼 상단에 유모차 장애인 우선이라고 적힌 안내문이 부착되어 있다.
6	P11601	ocr	유모차 장애인 우선
7	P11602	ref	통로 옆에 설치된 표지판에 따르면 진열상품은 구매 시 교환과 환불이 불가능하다.
8	P11602	worker1	통로 옆 표지판에는 진열되어 있는 상품 구매 시 교환이나 환불이 불가능하다고 적혀있다.
9	P11602	worker2	통로 옆의 안내판에 따르면 진열상품 구매 시 교환 또는 환불이 불가능하다.
10	P11602	worker3	통로 옆에 설치된 표지판에는 진열상품 구매 시 교환과 환불이 불가능하다고 적혀있다.
11	P11602	worker4	통로 옆에 설치된 표지판은 진열상품 구매 시 교환과 환불이 불가능하다고 안내하고 있다.
12	P11602	ocr	진열상품 구매 시 교환/환불 불가합니다
13	P11603	ref	주차요금 정산기 왼쪽에 위치한 표지판을 보면 주차장 이용 방법을 알 수 있다.
14	P11603	worker1	우측 주차요금정산기의 옆에 붙인 표지판을 보면 주차장 이용 관련 안내가 적혀 있다.
15	P11603	worker2	주차장 이용에 관한 내용은 주차요금 정산기 왼쪽에 위치하고 있다.
16	P11603	worker3	주차장 이용 방법이 적힌 표지판 오른쪽에 주차요금 정산기가 있다.
17	P11603	worker4	주차요금 정산기 왼쪽에 주차장 이용을 설명하는 배너가 설치되어 있다.
18	P11603	ocr	주차장 이용 \$ 주차요금 정산기
19	P11604	ref	이 두 대의 무인회수기 사용 방법을 모를 때에는 그 사이에 부착된 빈 용기 무인회수기 사용안내문을 보면 된다.
20	P11604	worker1	빈 용기 무인회수기 두 대가 설치되어 있는데, 두 무인회수기 사이에 빈 용기 무인회수기 사용안내문이 부착되어 있다.
21	P11604	worker2	두 대의 무인회수기 사용안내문을 읽은 후 무인회수기를 사용하면 된다.
22	P11604	worker3	두 대의 빈 용기 무인회수기 사이에 있는 벽에 빈 용기 무인회수기 사용안내문이 붙어있다.
23	P11604	worker4	빈 용기 무인회수기 기계가 두 대 설치되어 있고, 그 사이에 빈용기 무인회수기 사용 안내문이 부착되어 있다.
24	P11604	ocr	빈 용기 무인회수기 사용안내
25	P11605	ref	진열된 화분 상단에 부착된 안내문에 따르면 화분이 크고 무거워 주의하여 옮겨야 한다.
26	P11605	worker1	화분 여러 개가 설치된 진열대 상단에 화분이 크고 무거우니 주의하여 옮겨야 한다는 안내문이 부착되어 있다.
27	P11605	worker2	화분 진열대 상단에 화분이 무거우니 주의하여 옮겨야 한다는 당부의 말이 적혀있다.
28	P11605	worker3	화분이 배치된 선반에 화분이 크고 무거우니 주의하여 옮겨야 한다는 안내문이 붙어있다.
29	P11605	worker4	화분이 진열되어 있고, 그 상단에 화분이 크고 무거우니 주의하여 옮겨달라는 안내문이 부착되어 있다.
30	P11605	ocr	화분이 크고 무겁습니다 주의하여 옮겨주세요
31	P11606	ref	사람이 걸터 있는 두 자판기의 하단에는 지폐투입구가 있다.

[그림 3] 테이블형 데이터 변환 예시

## <수정 전>

<표 25> 수정 전 그림 기반 유사 문장 말뭉치

P11601	worker1	엘레베이터 오른쪽에 붙어 있는 안내문에는 유모차와 장애인이 우선 탑승할 수 있도록 안내하고 있다.
P11601	worker2	엘레베이터 오른쪽에 부착된 안내문은 유모차, 장애인 우선 탑승을 안내하고 있다.
P11601	worker3	2호기 표지판 아래에 부착된 안내문은 유모차, 장애인 우선 탑승을 지시하고 있다.
P11601	worker4	엘레베이터 층수 버튼 상단에 유모차 장애인 우선이라고 적힌 안내문이 부착되어 있다.

## <수정 후>

<표 26> 수정 후 그림 기반 유사 문장 말뭉치

P11601	worker1	<b>엘리베이터</b> 오른쪽에 붙어 있는 안내문에는 유모차와 장애인이 우선 탑승할 수 있도록 안내하고 있다.
P11601	worker2	<b>엘리베이터</b> 오른쪽에 부착된 안내문은 유모차, 장애인 우선 탑승을 안내하고 있다.
P11601	worker3	<b>엘리베이터</b> 의 우측에 붙은 안내문에는 유모차, 장애인 우선 탑승이라고 안내하고 있다.
P11601	worker4	<b>엘리베이터</b> 의 우측에 부착된 안내문에 따르면 유모차, 그리고 장애인이 우선 탑승할 수 있다.

## ▷ OCR 검수 방법론

- OCR에서 누락되었던 8개 항목의 정보를 검수 과정 중에 새롭게 추가하였다.

<표 27> OCR 추가 사례

P11940	OCR	해나온식자재마트 \$ 과일 \$ 채소 \$ 정육
P11942	OCR	무단횡단금지 \$ 교량 하부 하천 산책로를 이용하시기 바랍니다.
P11944	OCR	즉석인화기 \$ 정말 간단한 사용방법
P11946	OCR	도구해수욕장 \$ 청룡회관 \$ 화장실 \$ 호미반도해안둘레길
P11948	OCR	도시숲사랑 \$ 달팽이마라톤 \$ 2019
P11950	OCR	하윤이네 \$ 반찬생각 \$ 즉석반찬 \$ 주문예약 받습니다
P11952	OCR	고양평화통일교육전시관 \$ 통일에 \$ 배우고 체험하는 \$ 평화통일교육전시관입니다
P11954	OCR	재래김 \$ 5,480 \$ 조미김



- 또한 OCR에 포함되어 있는 상호명 등에서 사용되는 다양한 현실 표기를 인정하여 정비하였다.

예) 띄어 쓰기와 맞춤법을 수정하여 ‘맛있는족보’로 표기한 경우는 이미지 표기에 있는 그대로 ‘맛있는족보’로 수정



[그림 4] OCR 원문 내 현실 표기 이미지

<표 28> OCR 원문 내 현실 표기 인정 사례

수정전	수정후
P11225 ref 빨갇고 하얀빛이 나는 간판이 있는 맛있는족보는 족발, 보쌈 전문점이다.	P11225 ref 빨갇고 하얀빛이 나는 간판이 있는 맛있는족보는 족발, 보쌈 전문점이다.

- 명백한 철자 오류가 있는 경우 수정

(예) 파란 천장에 부착되어 있는 표지판에, 빨간 글씨로 주차금지라 써져 있다. -> 철장

도서 검색대 컴퓨터 위로 독서는 정신적으로 충실한 사람을 만든다는 벤자민 크랭클린의 말이 적혀 있다. -> 프랭클린

## ○ 정비 말뭉치 예시

<표 29> 그림 기반 유사 문장 JSONL 말뭉치

{       "id": "nikluge-gips-2023-train-000000",       "input": {         "id": "P00001",         "image_width": 6000,         "image_height": 4000,         "ocr_info": [           {             "words": "2인승",             "type": "rect",             "bbox": {               "x": 486,               "y": 1091,               "width": 891,               "height": 193             }           }         ]       },       "output": ["2인승이라고 적혀 있는 하늘색 테두리 안내판 옆에는 다양한 색깔의 우산들이 띄워져 있다."]     }
---

다양한 색깔의 우산들이 띄워져 있는 곳 옆에는 2인승이라고 쓰인 하늘색 테두리 안내판이 있다.", "다양한 색깔의 우산이 띄워진 곳 옆에 있는 하늘색 테두리 안내판에는 2인승이라고 적혀 있다.", "하늘색 테두리 안내판에는 2인승이라고 적혀 있는데, 그 옆에 띄워져 있는 것은 다양한 색깔의 우산들이다.", "안내판은 하늘색 테두리에 2인승이라고 적혀 있으며, 그 옆에 띄워져 있는 것은 다양한 색의 우산들이다."}}

{"id": "nikluge-gips-2023-train-000001", "input": {"id": "P00002", "image\_width": 5328, "image\_height": 4000, "ocr\_info": [{"words": "대왕릉", "type": "rect", "bbox": {"x": 2883, "y": 1890, "width": 538, "height": 201}}, {"words": "소왕릉", "type": "rect", "bbox": {"x": 1536, "y": 2209, "width": 520, "height": 230}}], "output": ["나무 옆에 있는 갈색 팻말은 대왕릉과 소왕릉으로 가는 방향을 알려주고 있다.", "대왕릉과 소왕릉으로 가는 방향은 나무 옆에 있는 갈색 팻말이 알려준다.", "갈색 팻말이 알려주는 것은 대왕릉과 소왕릉으로 가는 방향이고, 팻말은 나무 옆에 있다.", "대왕릉과 소왕릉으로 가는 방향을 알려주고 있는 것은 나무 옆에 있는 갈색 팻말이다.", "나무 옆 팻말은 갈색이며, 대왕릉과 소왕릉으로 가는 방향을 알려준다."]}

{"id": "nikluge-gips-2023-train-000002", "input": {"id": "P00004", "image\_width": 3168, "image\_height": 2112, "ocr\_info": [{"words": "입장권을 구입해 주십시오", "type": "rect", "bbox": {"x": 578, "y": 1594, "width": 335, "height": 42}}], "output": ["사람들이 올라가고 있는 계단에는 입장권을 구입해 달라고 적혀 있는 흰색 안내문이 붙어 있다.", "입장권을 구입해 달라고 쓰여 있는 흰색 안내문이 사람들이 올라가고 있는 계단에 붙어 있다.", "사람들이 올라가고 있는 계단에 부착된 흰색 안내문에는 입장권을 구입해 달라고 적혀 있다.", "입장권을 구입해 달라고 적혀 있는 흰색 안내문이 부착되어 있는 곳은 사람들이 올라가고 있는 계단이다.", "계단에는 사람들이 올라가고 있으며 흰색 안내문이 부착되어 있는데, 입장권을 구입해 달라고 적혀 있다."]}

{"id": "nikluge-gips-2023-train-000003", "input": {"id": "P00005", "image\_width": 1080, "image\_height": 1440, "ocr\_info": [{"words": "청정원", "type": "rect", "bbox": {"x": 574, "y": 292, "width": 172, "height": 78}}, {"words": "고구마쥬", "type": "rect", "bbox": {"x": 56, "y": 797, "width": 278, "height": 131}}, {"words": "100%", "type": "polygon", "bbox": {"all\_points\_x": [755, 629, 664, 825], "all\_points\_y": [680, 753, 806, 724]}}, {"words": "고구마", "type": "polygon", "bbox": {"all\_points\_x": [804, 786, 664, 685, 834], "all\_points\_y": [756, 748, 825, 874, 791]}}, {"words": "그대로", "type": "polygon", "bbox": {"all\_points\_x": [802, 721, 753, 832], "all\_points\_y": [818, 860, 902, 856]}}, {"words": "100% 고구마 그대로라는 문구가 적힌 고구마쥬의 봉투는 하얀색이고, 청정원에서 만들었다.", "words": "100% 고구마 그대로라는 문구가 적힌 고구마쥬의 봉투는 하얀색이고, 청정원에서 만들었다.", "words": "100% 고구마 그대로라는 문구가 적힌 고구마쥬의 봉투는 하얀색이고, 청정원에서 만들었다.", "words": "100% 고구마 그대로라는 문구가 적힌 고구마쥬의 봉투는 하얀색이고, 청정원에서 만들었다."}], "output": ["청정원에서 만든 고구마쥬의 봉투는 흰색이고 100% 고구마 그대로라는 문구가 적혀 있다.", "청정원은 고구마쥬의 봉투는 흰색이고 100% 고구마 그대로라는 문구가 적혀있도록 만든다.", "100% 고구마 그대로라는 문구가 적힌 고구마쥬의 봉투는 하얀색이고, 청정원에서 만들었다.", "흰색으로 된 고구마쥬 봉투에 적혀 있는 것은 100% 고구마 그대로라는 문구인데, 청정원에서 만든 것이다.", "고구마쥬는 청정원에서 만들었으며, 100% 고구마 그대로라는 문구가 적혀있는 봉투는 흰색이다."]}

{"id": "nikluge-gips-2023-train-000004", "input": {"id": "P00007", "image\_width": 4032, "image\_height": 3024, "ocr\_info": [{"words": "감시용 카메라", "type": "rect", "bbox": {"x": 2222, "y": 45, "width": 1037, "height": 197}}, {"words": "녹화중", "type": "rect", "bbox": {"x": 2235, "y": 237, "width": 974, "height": 322}}, {"words": "고정문", "type": "rect", "bbox": {"x": 1985, "y": 2123, "width": 1319, "height": 434}}, {"words": "옆문 이용", "type": "rect", "bbox": {"x": 2302, "y": 2771, "width": 845, "height": 224}}], "output": ["유리창에는 파란색 장애인 스티커와 고정문이니 옆문을 이용하라고 적혀 있는 흰색 안내문, 감시용 카메라 녹화 중임을 알리는 흰색 팻말 등이 붙어 있다.", "파란색 장애인 스티커와 고정문이니 옆문을 이용하라고 안내하는 흰색 안내문, 감시용 카메라가 녹화 중임을 안내하는 흰색 팻말 등이 유리창에 붙어 있다.", "장애인 스티커는 파란색이고, 고정문이니 옆문을 이용하라고 적힌 안내문과 감시용 카메라 녹화 중임을 알리는 팻말은 하얀색이며, 모든 것은 유리창에 붙어 있다.", "유리창에 부착되어 있는 것은 파란색 장애인 스티커와 고정문이니 옆문을 이용하라고 적혀 있는 흰색 안내문, 감시용 카메라 녹화 중임을 알리는 흰색 팻말 등이다.", "유리창에 부착되어 있는 것은 파란색 장애인 스티커와 흰색 안내문, 감시용 카메라 녹화 중임을 알리는 흰색 팻말 등이며, 안내문에는 고정문이니 옆문을 이용하라고 적혀 있다."]}

{"id": "nikluge-gips-2023-train-000005", "input": {"id": "P00008", "image\_width": 5328, "image\_height": 4000, "ocr\_info": [{"words": "중명전 일원", "type": "rect", "bbox": {"x": 3207, "y": 2540, "width": 750, "height": 219}}], "output": ["중명전 일원에 관한 설명이 적힌 갈색 팻말 뒤에는 빨간색 건물이 있다.", "갈색 팻말



에는 증명전 일원에 대한 설명이 적혀 있고, 그 뒤에는 빨간색 건물이 있다.", "빨간 건물 앞으로 증명전 일원에 관한 설명이 적힌 팻말이 있고, 팻말은 갈색이다.", "빨간색 건물의 앞에 있는 것은 증명전 일원에 관한 설명이 적힌 갈색 팻말이다.", "팻말은 갈색에 증명전 일원에 관한 설명이 적혀 있으며 그 뒤에는 빨간색 건물이 있다."}}

{"id": "nikluge-gips-2023-train-000006", "input": {"id": "P00009", "image\_width": 3168, "image\_height": 2112, "ocr\_info": [{"words": "목사 김효성의 비", "type": "rect", "bbox": {"x": 1401, "y": 1647, "width": 249, "height": 98}}], "output": ["검은색 비석과 돌비석 앞에는 목사 김효성의 비라고 적힌 회색 팻말이 있다.", "목사 김효성의 비라고 쓰인 회색 팻말 뒤에 검은색 비석과 돌비석이 있다.", "회색 팻말에는 목사 김효성의 비라고 적혀 있고, 그 뒤로 있는 것은 돌비석과 검은색 비석이다.", "목사 김효성의 비라고 적혀 있는 회색 팻말은 검은색 비석과 돌비석 앞에 있다.", "검은색 비석과 돌비석 앞 팻말은 회색이며 목사 김효성의 비라고 적혀 있다."]}

{"id": "nikluge-gips-2023-train-000007", "input": {"id": "P00010", "image\_width": 3168, "image\_height": 2112, "ocr\_info": [{"words": "관광안내소", "type": "rect", "bbox": {"x": 1170, "y": 1057, "width": 263, "height": 119}}], "output": ["한옥의 갈색 문 위에는 관광 안내소라고 적힌 갈색 안내판이 붙어 있다.", "한옥의 갈색 문 위에 붙어 있는 갈색 안내판에는 관광 안내소라고 쓰여 있다.", "안내판의 색은 갈색이며 관광 안내소라고 적혀 있고 한옥의 갈색 문 위에는 달려 있다.", "관광안내소라고 적혀 있는 갈색 안내판이 붙어 있는 곳은 한옥의 갈색 문 위이다.", "한옥의 갈색 문 위 부착된 안내판은 갈색이며 관광 안내소라고 적혀 있다."]}

{"id": "nikluge-gips-2023-train-000008", "input": {"id": "P00011", "image\_width": 3168, "image\_height": 2112, "ocr\_info": [{"words": "세계유산 백제역사 유적지구", "type": "rect", "bbox": {"x": 572, "y": 980, "width": 1451, "height": 400}}, {"words": "공주 무령왕릉과 왕릉원", "type": "rect", "bbox": {"x": 632, "y": 1465, "width": 787, "height": 351}}], "output": ["회색 돌에는 세계유산 백제 역사 유적 지구, 공주 무령왕릉과 왕릉원이라는 문구가 적혀 있고, 뒤에는 나무들이 있다.", "나무들 앞의 회색 돌에는 세계유산 백제 역사 유적 지구, 공주 무령왕릉과 왕릉원이라는 안내가 쓰여 있다.", "회색 돌에 적힌 문구는 세계유산 백제 역사 유적지구, 공주 무령왕릉과 왕릉원이고, 그 뒤로 서 있는 것들은 나무이다.", "나무들 앞을 보면 세계유산 백제 역사 유적 지구, 공주 무령왕릉과 왕릉원이라는 문구가 회색 돌에 쓰여 있다.", "회색 돌에는 문구가 적혀 있는데 세계유산 백제 역사 유적 지구, 공주 무령왕릉과 왕릉원이라는 문구이며, 돌 뒤에 있는 것은 나무들이 있다."]}

{"id": "nikluge-gips-2023-train-000009", "input": {"id": "P00012", "image\_width": 3168, "image\_height": 2112, "ocr\_info": [{"words": "구둔역", "type": "rect", "bbox": {"x": 1978, "y": 700, "width": 437, "height": 86}}], "output": ["구둔역 간판이 붙어 있는 흰색 건물 앞에는 나무들과 노란색 팻말이 있다.", "흰색 건물에는 구둔역 간판이 붙어 있고, 그 앞에는 나무들과 노란색 팻말이 있다.", "노란색 팻말과 나무 뒤로 있는 흰 건물에는 구둔역 간판이 붙어 있다.", "나무들과 노란색 팻말이 있는 곳은 구둔역 간판이 부착되어 있는 흰색 건물 앞이다.", "건물은 흰색에 구둔역 간판이 붙어 있으며, 그 앞에 있는 것은 나무들과 노란색 팻말이 있다."]}

## 2.4. 표 기반 유사 문장 말뭉치

### ○ 인수 말뭉치 분석

표 기반 문장 생성 말뭉치는 주어진 표의 특정 부분을 설명하는 문장을 생성하는 과제로써, 수행하는 과제의 성격이 명확하다. 이에 따라 표 기반 문장 생성 말뭉치는 전반적인 말뭉치 형식에 대한 검수, 표에 대한 형식적 오류 검수, 그리고 생성 문장에 대한 오류 검수를 중심으로 이루어진다. 2021년 표 기반 유사 문장 말뭉치는 JSON 형식으로 제공되었으며 이에 따라 표의 내용 역시 JSON 형식의 2차원 배열로 제시되었다. 이때 표를 JSON으로 변환하는 과정에서 표 형식 생성과 관련한 일부 오류가 발생하였다.

2022년 말뭉치에서도 유사한 오류의 사례가 발생할 가능성이 있으므로 우선적으로 표의 형식을 검토하였다. 2021 표 기반 유사 문장 말뭉치의 표는 한글 확장자(.hwp) 파일로 원본 문서의 내용을 정리한 뒤, 해당 파일을 전산적으로 불러들일 수 있는 형식으로 변환한 후 자동으로 JSON 형식으로 구축한 형태이다. 때문에 JSON의 배열을 이용한 테이블 표현은 HTML 테이블 형식과 일대일로 대응하며, HWP의 경우 XML 형식의 HWPML 형식을 이용하여 대응되는 결과를 얻을 수 있다는 장점이 있었다. 다만 2021년 말뭉치에서 표 중 일부에서 내용에 반영되는 칸이 한 행 어긋나는 오류를 발견하여 수정하였던 사례가 있었다. 해당 사례는 한글 파일의 표에서 실제로 열의 이름을 표시하는 헤더의 위에 범례 정보 등을 표현하거나, 문서에서 적절한 간격을 유지하기 위해 모든 칸이 병합된 한, 두 개 정도의 행이 있는 경우 JSON 변환에 오류가 발생한 것으로 분석되었다. 이에 따라 이번 사업의 대상인 2022 표 기반 유사 문장 말뭉치를 정비하는 과정에서도 표의 형식적 오류 존재 여부를 우선적으로 살폈다.

#### ■ 국어원에서 수령한 데이터 건수(5/10)

- 말뭉치 JSON 파일: 1건 (112MB, id: GTPS220202305060)
- 표 데이터 HTML 파일 : 10062건 (95MB)
- 표 데이터 HWP 파일: 10062건 (589MB)
- 표 ID별 데이터 개수

<표 30> 표 ID 별 데이터 개수

id	개수
TBxxxxxx-xxx	2498
TPxxxxxx	2048
TSxxxxxx-xxx	5516

## ○ 표 기반 유사 말뭉치 검수 방법론

유사 문장에 대한 내용 검수는 한국어 어문 규범에 부합하는지 여부를 중심으로 진행한다. 주로 띄어쓰기 오류, 외래어 및 외국어 표기 오류, 오타자 등이 수정 대상이 된다. 한편 2022년 표 기반 유사 문장 말뭉치는 공적 문서를 비롯한 전문 분야의 텍스트를 포함하고 있어 전문 용어들이 다수 등장한다. 기계 부품명, 화학물질명 등 특수한 용어를 오타자로 오인하여 수정할 가능성이 있으므로 검수 시 전문 용어를 고려하여 검수를 진행한다.

말뭉치의 문장 전체에 대해서도 형태소 분석과 맞춤법 검사를 이용한 검수를 진행한다. 다만 전체 문장에 대해 형태소 분석부터 진행하는 것은 시간과 인력의 제약이 있기 때문에, 유니그램(1-gram), 바이그램(2-gram), 트라이그램(3-gram)을 기반으로 단발어(hapax legomenon)를 추출하여 단발어가 등장한 사례를 집중 검수 대상으로 선정하여 검수한다. 단발어란 빈도가 1인 토큰으로, 유니그램에서 트라이그램까지를 기반으로 단발어를 설정할 경우 비정규적으로 발생하는 오자 및 탈자를 간접적으로 찾아낼 수 있는 장점이 있다.

### ▷ 기본 검수 원칙: 유사 문장 원문 검수, 서로 다른 작업자 2인에 의한 검수

- 한글 파일을 이용하여 전체 문장의 맞춤법, 표의 내용을 충실히 반영하지 않은 경우(숫치 오타 등) 등을 검수, 수정한 후 깃허브(Github) 히스토리를 참조하여 교차 검수

### ▷ 주요 검수 내용

- 표 기반 문장 생성에서 비표준어의 허용 수준은 표준, 우리말샘 사전에서 인덱싱하는 수준을 목표로 하고 하이라이트 셀이 누락되었거나 셀 인덱스 정보가 없는 데이터는 최종 제출 데이터에서 제외한다.

### ▷ 주요 오류 수정 예시

- 문장 정보 불일치 오류 유형으로 문자 및 수치에 오타가 있는 경우

<표 31> 표 기반 유사 문장 말뭉치 내 문자 및 수치 오타 예시

구분	전체	바디프랜드	휴테크산업	LG전자	SK매직
알고 있는 활동 없음	197(21.9)	93(31.0)	81(27.0)	8(5.3)	15(10.0)
에너지절약 기술 사용	175(19.4)	52(17.3)	54(18.0)	36(24.0)	33(22.0)
친환경 소재 사용	172(19.1)	45(15.0)	53(17.7)	35(23.3)	39(26.0)
고객만족경영	108(12.0)	41(13.7)	33(11.0)	22(14.7)	12(8.0)
코로나19 의료진/자원봉사자/군부대/소 방서 등에 안마의자 기부	69(7.7)	22(7.3)	23(7.7)	10(6.7)	14(9.3)
이메일 영수증, 온라인 안내서 사용 (종이문서 사용 자제)	65(7.2)	14(4.7)	23(7.7)	14(9.3)	14(9.3)
소외·취약계층 지원, 농촌 일손돕기 참여	44(4.9)	15(5.0)	11(3.7)	8(5.3)	10(6.7)
청년고용 지원	26(2.9)	4(1.3)	9(3.0)	9(6.0)	4(2.7)
집중호우 등 자연재해 피해고객 안마의자 무상 교체	21(2.3)	8(2.7)	7(2.3)	3(2.0)	3(2.0)
회사직원복지 증대활동	12(1.3)	3(1.0)	4(1.3)	3(2.0)	2(1.3)
지배구조 개선	11(1.2)	3(1.0)	2(0.7)	2(1.3)	4(2.7)
총합	900(100.0)	300(100.0)	300(100.0)	150(100.0)	150(100.0)

TP10208 worker3 ESG 경영 주요 실천 활동의 인식에 대해 조사한 결과, 전체의 19.1%(172명)이 친환경 소재 사용에 응답하였으며, 전체의 19.4%(174명)이 에너지절약 기술 사용에 대한 인식에 응답하여, 두 항목 간 차이는 미미한 것으로 나타났다. -> 175명으로 수정

■ 하이라이트 셀을 참조하지 않고 다른 정보를 사용하여 문장을 생성한 경우

- 하이라이트 셀의 정보를 사용하여 수치를 올바르게 수정하고 문장 중 문제가 되는 부분 삭제하였다.

<표 32> 하이라이트 셀 참조 검수 사례

		2021. 5			2022. 5		
		졸업(중퇴)후 취업유경험자	고졸이하	대졸이상	졸업(중퇴)후 취업유경험자	고졸이하	대졸이상
< 전 체 >		4,056	1,495	2,562	4,117	1,504	2,613
		(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)
취 업 경 로	가족, 친지 소개(추천)	(16.7)	(25.2)	(11.8)	(16.0)	(22.5)	(12.2)
	그 직장 근무자 소개(추천)	(9.1)	(9.9)	(8.6)	(9.1)	(12.0)	(7.5)
	학교(학원) 선생님 추천	(6.7)	(7.9)	(6.0)	(6.7)	(6.2)	(6.9)
	신문, 잡지, 인터넷 등 응모	(31.8)	(34.9)	(29.9)	(31.7)	(36.2)	(29.1)
	공개채용시험	(20.6)	(6.0)	(29.1)	(21.2)	(7.0)	(29.4)
	특별채용	(2.6)	(2.0)	(3.0)	(2.5)	(2.1)	(2.8)
	그 외	(12.5)	(14.1)	(11.6)	(12.8)	(13.9)	(12.2)

<표 33> 하이라이트 셀 참조 검수 사례 문장 수정

수정 전	수정 후
TP11807 worker2 고졸 이하 취업 유경험자의 취업경로를 조사하였더니, 전체의 31.7%의 인원이 신문, 잡지, 인터넷응모를 통해 취업했다고 응답했으며, 이보다 약 1/2에 해당하는 16.0%의 인원이 가족, 친지 소개(추천)를 통해 취업한 것으로 나타났다.	고졸 이하 취업 유경험자의 주된 취업경로는 신문,잡지,인터넷등 응모(31.7%), 가족,친지 소개(추천)(16.0%) 등이 있었다.

## ○ 정비 말뭉치 예시

<표 34> 표 기반 유사 문장 JSONL 말뭉치

<pre>{   "id": "nikluge-gtps-2023-train-000000",   "input": {     "metadata": {       "title": "4차 산업혁명에 따른 조세환경 변화와 정책 과제",       "table_title": "4차 산업혁명 관련 조세 부문 주요 의원발의 법률안",       "date": "2020-06-09",       "publisher": "국회예산정책처",       "url": "https://www.nabo.go.kr/Sub/01Report/01_01_Board.jsp",       "highlighted_cells": [[0, 13], [1, 14], [3, 14]]     },     "table": [       {         "value": "조세특례제한법",         "is_header": true,         "col": 0,         "colspan": 4,         "row": 0,         "rowspan": 1       },       {         "value": "2009580",         "is_header": false,         "col": 0,         "colspan": 1,         "row": 1,         "rowspan": 1       },       {         "value": "특허 출원·등록비용 소득세·법인세 공제",         "is_header": false,         "col": 1,         "colspan": 1,         "row": 1,         "rowspan": 1       },       {         "value": "정병국의원",         "is_header": false,         "col": 2,         "colspan": 1,         "row": 1,         "rowspan": 1       },       {         "value": "2017.9.25.",         "is_header": false,         "col": 3,         "colspan": 1,         "row": 1,         "rowspan": 1       },       {         "value": "2011799",         "is_header": false,         "col": 0,         "colspan": 1,         "row": 2,         "rowspan": 1       },       {         "value": "지능정보 기술 분야에 대한 R&amp;D비용 세액 공제",         "is_header": false,         "col": 1,         "colspan": 1,         "row": 2,         "rowspan": 1       },       {         "value": "박광온 의원",         "is_header": false,         "col": 2,         "colspan": 1,         "row": 2,         "rowspan": 1       },       {         "value": "2018.2.6.",         "is_header": false,         "col": 3,         "colspan": 1,         "row": 2,         "rowspan": 1       },       {         "value": "2013345",         "is_header": false,         "col": 0,         "colspan": 1,         "row": 3,         "rowspan": 1       },       {         "value": "창업자, 신기술사업자, 벤처기업에 출자한 내국법인의 주식양도차익 비과세 및 배당소득 비과세",         "is_header": false,         "col": 1,         "colspan": 1,         "row": 3,         "rowspan": 1       },       {         "value": "...",         "is_header": false,         "col": 1,         "colspan": 1,         "row": 15,         "rowspan": 1       },       {         "value": "박성중의원",         "is_header": false,         "col": 2,         "colspan": 1,         "row": 15,         "rowspan": 1       },       {         "value": "2018.11.30.",         "is_header": false,         "col": 3,         "colspan": 1,         "row": 15,         "rowspan": 1       }     ]   },   "output": ["2018년 11월 6일 발의된 부가가치세법의 내용은 부가가치세 과세대상 전자적 용역 범위를 인터넷광고, 클라우드컴퓨팅서비스, 공유경제서비스까지 확대하는 것이다."] }</pre>
--

가치세 과세대상 전자적 용어범위는 인터넷, 클라우드컴퓨팅서비스, 공유경제서비스까지를 포함하는 것으로 2018년 11월 6일 발의된 부가가치세법 내용에서 정의하고 있다.". "부가가치세법의 경우 인터넷광고, 클라우드컴퓨팅서비스, 공유경제서비스 등 부가가치세 과세대상 전자적 용역 범위를 확대한다는 내용이 2018년 11월 6일 발의되었다.". "부가가치세 과세대상 전자적 용역 범위를 인터넷광고, 클라우드컴퓨팅서비스, 공유경제서비스까지 확대하는 것이 2018년 11월 6일 발의된 부가가치세법의 내용이다.". "2018년 11월 6일, 부가가치세법은 인터넷광고, 클라우드컴퓨팅서비스, 공유경제서비스까지 부가가치세 과세대상 전자적 용역 범위를 확대하는 것으로 결의되었다."}}

## 2.5. 함의 분석 말뭉치

### ○ 인수 말뭉치 분석

이번 과제의 “함의 분석” 과제는 협의의 자연어 추론, 즉 한 문장 쌍 내부에서 두 문장 사이의 논리적 추론 관계를 파악하는 과제를 의미한다. 일반적으로 함의 분석 과업에서 추론에 사용되는 문장들은 선행 문장과 후행 문장의 두 문장으로 이루어진다. 선행 문장은 전제(premise) 혹은 문맥(context) 불리며, 후행 문장은 가설(hypothesis)로 불린다. 선행 문장은 후행 문장이 이해될 수 있는 맥락을 제공하며, 후행 문장은 선행 문장을 기반으로 만들어진다. 함의 분석 과제는 언어 모델로 하여금 위와 같은 형식으로 주어진 선행 문장과 후행 문장의 쌍에 대하여 문장들 사이의 관계가 함의(entailment), 중립(neutral), 모순(contradiction)의 세 가지 라벨 중 어떤 것에 해당하는지를 추론한다. 본 평가 체계에 서 정비한 2022년 국립국어원 한국어 적대적 함의 분석 말뭉치의 경우, 앞서 언급된 함의 분석 과업의 기본적인 틀을 가져오되 난도를 높여 만들어진 말뭉치이다.

이 데이터는 인공 지능 평가를 목적으로 적대적 함의 관계 2만 건 이상을 포함한다. 구축의 흐름은 (1) 선행 및 대상 문장 추출, (2) 가설 문장 생성, (3) 작업자 간 검수, (4) 모델 fooling의 네 단계로 구성된다. 선행 및 대상 문장은 “국립국어원 신문 말뭉치 2021”에서 무작위 추출되었다. 국립국어원 신문 말뭉치 2021 버전은 총 729,280 건의 기사로 구성되어 있으며 매체 수는 모두 35개이다. 이들 가운데 제목을 제외한 본문에서 이어지는 문장을 연달아 추출하여 각 작업자에게 매회마다 5천건의 기사 조각을 할당하였다. 가설들은 적절하게 생성될 수 있는 것만 골라 수백 건씩을 매회(round) 생성되었다.

### ○ 함의 분석 말뭉치 검수 방법론

본 사업단은 <2022년 말뭉치 함의 분석 및 연구> 과제를 통해 구축된 함의 분석 말뭉치 데이터에 대한 정비 작업을 수행하였다. 정비 작업은 단발어 기준 검사와 전수 검사의 두 단계로 나뉘어 진행되었다.

첫 번째 단계인 단발어 기준 검사에서는 함의 분석 말뭉치 데이터에서 137,164개의 단발어(Hapax legomenon)를 추출하여 수정 작업을 수행하였다. 이어서 두 번째 단계에서는 구축된 함의 분석 말뭉치에 대한 전수 검사를 수행하였다. 특히 전수 검사 단계의 경우, 말뭉치 정비에 앞서 검수 작업을 수행할 두 명의 작업자들을 위한 전수 소자 작업 지침을 구축 및 배포하여 일관적이고 안정적인 검사가 이루어지도록 하였다. 구체적으로 <2022년 말뭉치 함의 분석 및 연구>의 전수 검사 과정에서는 아래와 같은 세부 지침이 사용되었다.

- ▷ **대원칙:** 생성된 가설문장에 맞춤법 및 띄어쓰기 오류로 인한 비문이 존재하는 경우 적절히 수정하도록 한다. 단, 수정의 범위는 형식적인 오류의 수정에 국한되며 가설문장의 내용(함의, 중립, 모순 중 어느 쪽에 해당하는지 여부)에 대한 변경이 이루어지지 않도록 한다.

- 위의 대원칙을 벗어나지 않는 선에서 아래 [1]과 [2]의 수정 사항을 준수하는 것을 원칙으로 한다.

[1] 가설문장에서 의존 명사 등의 띄어쓰기 오류가 있을 경우 수정하도록 한다.

[2] 가설문장에서 잘못된 조사의 사용 등 맞춤법 오류가 있을 경우 수정하도록 한다.

#### ■ 수정 예시 1

<표 35> 함의 분석 말뭉치 수정 예시 1

수정 이전: 임미숙씨의 성별은 남자가 아니다.
수정 이후: 임미숙 씨의 성별은 남자가 아니다.

#### ■ 수정 예시 2

<표 36> 함의 분석 말뭉치 수정 예시 2

수정 이전: 한화이글스의 노태형은 6회말에 교체 선수로 투입되어 한화의 승리에 기여하였다.
수정 이후: 한화이글스의 노태형은 6회 말에 교체 선수로 투입되어 한화의 승리에 기여하였다.

- 일반적으로 수정된 오류의 경우, 가설 문장을 생성한 작업자의 오류에 기인하기 보다는 가설문장 생성에 사용된 원문 자체에 오류가 존재하는 경우에 해당하였다. 이러한 경우에는 원문 데이터를 존중하여 원문에는 별도의 수정을 하지 않고 원문에 기반한 가설 문장만을 수정하는 것을 원칙으로 하였다.

<표 37> 가설 문장 생성 시 원문 자체 오류 예시

원문: 그러나 대형마트에는 QR코드를 체크하거나 방문지 기록 카드를 작성하지 않아도 아무런 제제 없이 입장이 가능하며 체온 체크도 제대로 이루어지지 않아 방문 고객들이 불안해하고 있다. 지난 추석 연휴동안 북새통을 이룬 대형마트는 차량 진입부터 어려운 상황에서 쇼핑 결제까지 10여 분을 기다리는 촌극에 이르기까지 군민들이 그동안 경험해 보지 못한 상황이 연출됐다.
수정 이전: 추석 연휴동안 대형마트에는 사람들이 몰려들었다.
수정 이후: 추석 연휴 동안 대형마트에는 사람들이 몰려들었다.



## ○ 검수 결과 및 통계

우선 단발어 기준 검사의 전체적인 결과는 아래와 같다.

<표 38> 함의 분석 말뭉치 수정 유형 및 건수

수정 유형	수정 건수
명백한 오타자	164
명백한 띄어쓰기 오류	520
명백한 고유명사 오류	12
확인이 필요한 고유명사	4
기타 확인이 필요한 경우	40
합계	740

함의 분석 말뭉치의 경우 신문 말뭉치를 기반으로 구축되었기 때문에 인명, 지명, 전문 용어 등의 단발어가 고빈도로 분포하고 있다. 이러한 이유로 인해 전체 단발어의 발생 빈도는 함의 분석 말뭉치가 이야기 분석 말뭉치에 비해서 더욱 높음에도 불구하고 수정 건수는 더 적은 것으로 드러났다. 전수 검사 작업의 전체적인 결과는 아래와 같다.

<표 39> 파일별 수정 건수

파일	수정 건수
1번 파일	190
2번 파일	116
3번 파일	449
4번 파일	404
합계	1159

전수 검사의 경우, 전수 검사를 맡은 작업자 2인이 1차적으로 전체 데이터를 두 부분(작업자 1: 1번 파일 & 2번 파일, 작업자 2: 3번 파일 & 4번 파일)으로 나누어 각각 검수 및 수정하였고 그다음 2차적으로 각자가 작업했던 1차 결과물을 상호 검증(cross-check)하는 방식으로 이루어졌다. 함의 분석 말뭉치는 두 개의 문장으로 구성된 “전제(이하 premise)”와 해당 “premise”에 상응하는 “명제(이하 proposition)” 문장으로 구성된다. 이 중에서 “premise”는 작업자에게 원래부터 주어지는 원문 데이터에 해당하며 이 “premise”를 근거로 작업자는 함의/중립/모순 중 어느 하나에 해당하는 “proposition”을 생성하였다. 이 과정에서 6가지 방법(“method”)이 사용되었는데 각각의 “proposition” 생성에 있어 서로 다른 방법이 다양하게 사용되었다.

함의 분석 말뭉치의 구성은 다음과 같다. id, metadata 의 경우 본 말뭉치가 목표하는 과업의 수행에 활용되지 않는 요소이기 때문에 검수 대상에 포함시키지 않고, 값의 존재 여부와 형식만 확인하였다.

```

"document": [
  {
    "id":
    "metadata": {
      "source":
      "topic":
    },
    "method": {
      "method_num":
      "method_ref":
      "method_std":
      "method_lex":
      "method_tricky":
      "method_reas":
    },
    "sentences": {
      "premise":
      "proposition":
      "label":
      "explanation":
    },
    "prediction": {
      "model_prediction1":
      "model_prediction2":
    }
  }
]

```

이 중 실질적으로 수정이 이루어진 항목은 “proposition”에 국한되며 이는 함의 분석 말뭉치 작업자에 의해 생성된 “proposition”과 달리 원래부터 주어진 “premise”에 대한 수정은 하지 않음을 검수 과정에서 원칙으로 했기 때문이다. 함의 분석 말뭉치의 예시는 다음과 같다.

(1) premise: 마지막으로 관리부실이다. 설계상 통행량이 하루 평균 8만대에 하중은 32t였던 다리에 40t이 넘는 과적 차량을 방치했고 통행량은 설계 대비 2배 이상인 16만대를 웃돌았다.

proposition: 하루 평균 실제 통행량은 설계상 통행량보다 8만 대 이상 많다.

(1)에서는 두 개의 문장으로 구성된 premise가 주어져 있고 해당 premise를 근거로 작업자가 생성한 proposition ‘하루 평균 실제 통행량은 설계상 통행량보다 8만 대 이상 많다.’가 주어져 있다. 실제로 설계상 통행량이 하루 평균 8만 대이고, 실제 통행량은 16만대 이므로 하루 평균 실제 통행량은 설계상 통행량보다 8만 대 이상 많다. 그러므로 (1)의 proposition은 premise에 대하여 함의(entailment)에 해당함을 알 수 있다.

(2) premise: 27일에는 고2, 중3, 초1,2학년, 유치원이 개학을 하며, 6월 3일에는 고1, 중2, 초 3,4학년, 8일에는 중1, 초 5,6학년이 순차적 등교 수업을 진행할 예정이다. 코로나19 감염 우려를 배제할 수 없는 상황에서 수시 모집, 맞벌이, 한부모 가정의 자녀 돌봄, 교육 관계를 통한 인성교육, 기초학력 부진 등의 우려가 등교 개학을 추진한 배경이 됐다.

proposition: 총 3일에 걸쳐 전 학년이 순차적 개학을 한다.

(2)에서도 마찬가지로 두 개의 문장으로 구성된 premise가 주어져 있고 해당 premise를 근거로 작업자가 생성한 proposition ‘총 3일에 걸쳐 전 학년이 순차적 개학을 한다.’가 주어져 있다. 이 경우에는 모든 학년이 아닌 일부 학년만이 순차적 개학을 하므로 (2)의 proposition은 premise에 대하여 모순(contradiction)에 해당함을 알 수 있다.

(3) premise: 이 전 총리가 직전 총리였던 만큼 유족들의 기대감도 컸던 것이 사실이다. 다만, 장 의원의 페이스북 글처럼 이 전 총리와 유족간 대화는 짧은 문답 형식은 아니었고 이에 대한 평가도 다소 엇갈리고 있다.

proposition: 이 전 총리와 유족 간 대화는 긴 문답 형식이었다.

(3)의 경우에도 마찬가지로 두 개의 문장으로 구성된 premise가 주어져 있고 해당 premise를 근거로 작업자가 생성한 proposition ‘이 전 총리와 유족 간 대화는 긴 문답 형식이었다.’가 주어져 있다. 이 경우에는 짧은 문답 형식이 아니었다는 정보만으로는 이 전 총리와 유족 간 대화가 긴 문답 형식인지의 여부를 확정할 수 없으므로 (3)의 proposition은 premise에 대하여 중립(neutral)에 해당함을 알 수 있다.

## ○ 정비 말뭉치 예시

<표 40> 함의 분석 JSONL 말뭉치

```
{
  "id": "nikluge-2023-te-train-000001",
  "input": {
    "premise": "경찰은 현재 과학수사 수준이 사건 발생 당시보다 비약적으로 발전한 점에 착안해 지난해 7월 15일 피해자 유류품을 국립과학수사연구원으로 보내 DNA 검출·분석을 의뢰했다. 지난해 8월 9일 9차 사건 유류품에서 이춘재의 DNA가 처음 검출됐고 그의 자백이 더해져 1년에 걸친 재수사는 마무리됐다.",
    "proposition": "국립과학수사연구원은 경찰의 의뢰를 받아들여 피해자 유류품을 분석했을 것이다.",
    "output": "entailment"
  },
  "id": "nikluge-2023-te-train-000002",
  "input": {
    "premise": "다만 미국의 행동에 따라 조치를 취하겠다는 논리를 편 것은 여전히 트럼프 대통령의 재선 가능성을 고려해 대화의 창은 완전히 닫지 않았다는 방증이라는 관측도 나온다. 김정 북한대학원대 교수는 “한미의 관심을 끌고자 저장도 도발은 할 수 있으나 ‘레드라인’인 핵실험이나 대륙간탄도미사일(ICBM) 시험발사는 리스크가 크다고 생각할 것”이라고 말했다.",
    "proposition": "현시점에서는 아직 다음 미국 대선의 결과가 나오지 않았다.",
    "output": "entailment"
  },
  "id": "nikluge-2023-te-train-000003",
  "input": {
    "premise": "우리나라 상반기 총 수출규모는 2406억 달러로 전년 동기대비 11.3% 감소했는데, 부산의 수출 감소폭은 전국의 2배에 육박한다. 이처럼 부진이 두드러지면서 부산의 지자체 수출 순위는 11위, 수출증감률은 14위로 최하위권을 기록했다.",
    "proposition": "부산의 수출 감소 폭이 전국의 두 배에 달함에도 불구하고 그보다 수출규모가 작은 지자체가 존재한다.",
    "output": "entailment"
  },
  "id": "nikluge-2023-te-train-000004",
  "input": {
    "premise": "이 흐름이 팬데믹 이후 심화했고 유럽에서 사회 안 전망이 가장 잘 돼있다고 여겨지는 국가에서도 기아와 빈곤 상황이 심각해졌다. 빈곤 퇴치 단체들은 위기를 근본적으로 해결하려면 식량 생산과 공급, 분배 시스템 전반을 재점검해야 한다고 주장한다.",
    "proposition": "코로나 19 이후로 빈곤 위기를 크게 느끼지 못했던 국가에서도 기아와 빈곤 문제가 심각해지기 시작했다.",
    "output": "entailment"
  },
  "id": "nikluge-2023-te-train-000005",
  "input": {
    "premise": "화이자는 독일 바이오엔테크와 백신을 공동개발했으며 미 식품의약국(FDA) 긴급승인을 거쳐 14일부터 백신 접종이 시작됐다. 미국이 전시법까지 동원하며 백신을 확보하면서 전 세계 코로나 19 양극화 현상은 심화할 것으로 예상된다.",
    "proposition": "화이자는 독일 바이오엔테크와 같이 백신을 개발했는데 최종 승인에는 이르지 못했다.",
    "output": "contradiction"
  },
  "id": "nikluge-2023-te-train-000006",
  "input": {
    "premise": "미국 음악 전문 매체 빌보드가 3일(현지시간) 발표한 최신 차트(11월 7일 자)에 따르면, 방탄소년단의 디지털 싱글 ‘다이너마이트(Dynamite)’는 ‘팝 송’ 차트에서 전주 대비 2계단 상승한 9위를 기록했다. ‘팝 송’ 차트는 빌보드가 발표하는 라디오 차트 중 하나로, 팝 장르의 상위 40개 곡을 대상으로 미국 내 약 160개 주요 라디오 방송국의 주간 방송 횟수를 집계해 순위를 매긴다.",
    "proposition": "다이너마이트는 방탄소년단 멤버 중 한 명이 따로 낸 음원이다.",
    "output": "contradiction"
  }
}
```

## 2.6. 부적절성 말뭉치

### ○ 인수 말뭉치 분석

2022년 국립국어원 말뭉치 비윤리성 분석 및 연구 보고서(조태린 외, 2022)에 따르면 부적절성 말뭉치는 비윤리성과 더불어 엄격한 의미에서 비윤리성으로 간주하기 어려운 부정적 특성까지 폭넓은 개념을 채택한 말뭉치로써, 공격성, 편향성, 비하성, 편견 등을 포함한 텍스트로 구성되어 있다. 해당 말뭉치는 문장을 주석 단위로 채택하고 있으며 명시적, 비명시적 표현을 주석한 명시성, 맥락(감성), 영역, 강도가 주석되었다.

명시성 측면에서 비명시적 부적절성 문장이 명시적인 사례보다 2배 더 많이 발생하였으며 강도는 강-약 비슷한 수준으로 관찰되었다. 맥락의 경우 부정이 긍정보다 3배 이상 많이 주석되었다. 영역은 문화, 신체, 성 등 여러 개로 나누어져 있으나 본 사업에서는 영역은 평가 대상으로 다루지 않아 정비에서는 제외되었다. 말뭉치 내 문장 수는 총 16,240건으로 본 사업에서는 16,240건 전수에 대해 평가용 말뭉치로 정비를 진행하였다. 개인정보의 경우 비식별화 되었다.

### ○ 부적절성 말뭉치 검수 방법론

상시 과제 평가용 말뭉치로 부적절 말뭉치를 정비하였으며 구체적인 정비 기준은 부적절성 말뭉치 분석 작업 지침 지침을 따라 맥락(context), 명시성(is\_explicit), 강도(intensity), 부적절 표현(inappropriate\_expression)에서 지침 내용에 벗어난 오류를 찾아 수정하였다. 이때, ‘부적절성’ 개념에 대한 주관성으로 작업자 간의 완벽한 일치를 이루기 어려움을 인정하였으며, 구축 말뭉치에서 지침과 다른 부분을 발견하여 정비하고자 하였다.

비윤리, 혐오 표현 등과 같은 말뭉치의 주석 작업의 난점은 판단 과정에서 개입되는 주관성과 강도 높은 작업 피로도이다. 해당 문제점을 부적절성 말뭉치 정비 과정에서 해결하기 위해 3명 이상의 검수자들이 문장을 검수하며 주관성 개입의 여지를 줄였으며, 표본(샘플링) 선검수와 지침 반영 여부 확인 과정을 통해 주관성을 제한하였다. 또한 정비 작업 이전에 검수자들에게 부적절성 말뭉치에 대한 데이터 특성을 알리고, 피로감 호소 시 작업 휴식 혹은 작업자 교체 등 적극적인 조치를 적용하였다.

부적절성 말뭉치의 정비 대상이 되는 라벨은 ‘맥락(context), 명시성(is\_explicit), 강도(intensity), 부적절 표현(inappropriate\_expression)’으로 4가지이다. 라벨의 정비 기준은 부적절성 말뭉치 분석 작업 지침 지침의 정의를 따랐으며, 아래는 각 라벨의 정비 기준이 되는 정의와 그 내용을 후술한 것이다.

맥락(context) 검수는 구축 지침의 내용을 반영한다. 지침에서의 맥락 정의는 문장의 부적절성이 화자의 태도(의도)나 맥락 내용 측면에서 부정적인지, 긍정적인지를 판정하는 것이다. 화자의 태도(의도)나 맥락 내용이 모두 부정적이거나, 긍정성과 부정성을 판단할 수 없는 무표적인 문장을 재검토하며 ‘부정적’ 태깅으로 변경하였다. 마찬가지로 화자의 태도(의도)와 맥락 내용이 모두 긍정적으로 판단되며, 부정적인지 않은 문장을 긍정적으로 수정하였다. 아래는 지침에서 분류하고 있는 부정적/긍정적 맥락 문장이다.

<표 41> 부적절성 말뭉치 맥락 예시

sentence	맥락
대학생이나 되는데, 생각은 참 초딩 같아.	부정적
오늘도 구라치다 하루가 다 갔어~	
제발 애 싸지려면 기본상식 시험 치고 통과해서 싸질렀으면 좋겠다.	
완벽한 S라인을 자랑하는 아름다운 몸매가 그 첫 번째.	
나랑 딱칠래?	긍정적
기장도 살짝 다듬어주시고, 거지존 극복 할 수 있게	
미소빌런	
진짜 저 사진 찍어놓고 너무 멍충해보여서 한동안 웃었다.ㅋㅋ	
시발 존멋/존예	
개웃기다시발 나 왜 속음	

명시성(is\_explicit)의 개념은 문장에서 부적절성이 구체적인 어휘나 표현을 통해 명시적으로 나타나는지, 문장의 맥락에서 드러나는지(비명시)를 판정하는 것이다. 명시성을 판단하기 위한 구체적인 표현 범위는 어절 단위의 욕설, 비속어 등이 있으며, 띄어쓰기 없이 하나의 어절로 제시되기도 한다. 대부분 어절이나 구 단위로 나타나며, <표준국어대사전>과 <고려대한국어대사전> 등의 대사전류에서 욕설, 비어, 비하성 속어 등으로 기술되는 어휘나 표현이 나타나는 모든 문장을 대상으로 한다. 이러한 지침 내용에 준하여 명시적인 어휘 및 표현이 있는 문장을 검토하고, 주석 여부를 검토하였다. 아래는 지침에 따라 명시적 표현과 비명시적 표현을 구분한 예이다.

<표 42> 부적절성 말뭉치 명시적/비명시적 예시

sentence	명시성
여기 <b>틀딱</b> 들이 대거 나온 듯.	명시
시발진짜내가쪽팔림을무릅쓰고학원 쌤한테잠깐만폰해도되냐고물어봤는데	
하지만 전문가 말만큼 믿을 게 못 되는 것도 없다.	비명시
외국 아가씨들이랑 놀기 위해..	

강도(intensity)는 문장의 부적절성의 심각성이 어느 정도(강/약)인지를 판정하는 것으로, 맥락 및 명시성과 긴밀한 연관성을 지닌다. 예를 들어, 명시적 부적절성이 부정적 맥락에서 나타날 때는 강도가 강(strong)하고, 명시적 부적절성이 긍정적 맥락일 때는 약(weak)하게 주석된다. 문장이 비명시적일 경우에는 성적 폭력성, 선정성 등과 같은 부정적 맥락 및 내용에 따라 강도를 주석하게 된다. 아래는 강도에 대한 지침 예시이다.

<표 43> 부적절성 말뭉치 강도 예시

sentence	강도
말본새를 보면 애미도 없는 것들이지	강
넌 조선족 사는데 같이 살고 싶어?	
저 중에 누가 제일 따먹고 싶게 생김?	
왜 키우냐 묻거든 강 웃지요 근데 꼴에 머리는 겁나 좋다 그냥 똥똥한 사람같다ㅋㅋ 오래 살아라.	약
그나저나 오늘 날씨 왠케 좋아!!! 이런 X!! 욕 나오게 하네.	
한약 먹는 보람도 없이 이게 무슨 짓이람..	

부적절 표현(inappropriate\_expression)은 명시성의 표현 범위를 나타낸 것으로 문장에서의 시작-종료 위치와 표현이 주석되어 있다. 명시성의 표현 범위는 어절 단위로 나타나 주로 체언+조사, 용언+어미 등의 형태로 나타난다. 이때 띄어쓰기가 이루어지지 않아 하나의 어절로 제시되는 경우에는 전체 문장을 표현 범위로 표시한다. 또한 명시적 부적절성을 발생시키는 관용구는 하나의 어절로 간주하여 구 단위를 표현 범위로 주석한다. 이러한 지침 내용에 따라 부적절 표현을 검토하여 정비 과정을 거쳤다.

## ○ 검수 결과

부적절성 말뭉치는 총 16,238건 문장 단위로 구축되어 있으며, 정비 과정에서 문장 및 주석 라벨을 정성적으로 검토한 결과 186건이 수정되었다. 정비 기준은 부적절성 말뭉치 분석 작업 지침 지침으로 정하고, 지침 기준 및 예시에 맞추어 검토하였다. 검수 대상이 되는 세부 주석 라벨은 1) 맥락(context), 2) 명시성(is\_explicit), 3) 강도(intensity), 4) 부적절 표현(inappropriate\_expression)이다. 각 라벨별 세부 정비 건수는 다음과 같다.

<표 44> 부적절성 말뭉치 수정 건수

검수 대상	수정(개수:건)
1) 맥락(context),	64건
2) 명시성(is_explicit),	38건
3) 강도(intensity),	111건
4) 부적절 표현 (inappropriate_expression)	30건

부적절성 말뭉치에서 ‘맥락’ 주석 정비는 분석 지침 지침의 정의에 따라 부적절성이 화자의 태도(의도)나 맥락 내용 측면에서 부정적인지, 긍정적인지를 판정하고, 실제 주석된 내용을 정성적으로 살펴보며 주석 내용을 변경하였다. 아래 정비 예시 결과를 살펴보면, 검수 대상이 되는 문장은 주로 무표적인(부정적이지 않은) 문장에 부정적으로 주석이 되어 있는 오류와 부적절성이 명시적으로 나타나는 어휘를 사용하여 부정적인 태도(의도)를 표현하는 문장에 대해 긍정적으로 주석된 경우로 한정하였다. 즉 긍정적인 맥락에 대해 부정적으로 주석하고, 명시적인 어휘를 사용하고 있음에도 긍정적으로 오류 주석한 경우를 발견하여 분석 지침 지침의 기준에 따라 수정하는 작업을 거쳤다. 아래 예시는 무표적인(부정적이지 않은) 문장에 긍정적으로 주석을 수정하고, 명시적 어휘를 사용하여 부정적인 태도(의도)를 나타내는 문장에 대해 부정적으로 수정한 정비 결과이다.

<표 45> 부적절성 말뭉치 ‘맥락’ 정비 예시

sentence-id	sentence_form	context (정비 이전)	context (정비 이후)
EBRW1908000303 .114.19.1	졸려죽겠지만 &name& 보고 이겨냈다.	Negative	Positive
ESRW190700198 0.104.1.1	존나 파우치 시발 ㅋㅋㅋㅋㅋㅋ	Positive	Negative



분석 지침 지침에서 정의하고 있는 명시성은 대사전류에서 정의되는 욕설, 비하성 속어, 비어 등의 어휘 분류이다. 따라서 명시성 정비 과정은 대사전류 검색을 통해 정성적으로 검수하였으며, 아래 표는 명시성 주석을 정비한 예시이다. 대사전류 검색을 통해 욕설, 비하성 속어, 비어 등으로 분류되지만, 주석 내용이 비명시로 되어 있는 오류를 발견하여 주석 내용을 명시로 수정하는 과정을 진행하였다. 예로 <표준국어대사전>에서 ‘지랄’은 ‘마구 법석을 떨며 분별없이 하는 행동을 속되게 이르는 말.’이라 풀이한다. 이는 ‘속어’에 해당하기 때문에 명시적인 표현을 사용한 경우로 주석을 수정하였다. 또한 ‘노예’는 ‘남의 소유물로 되어 부림을 당하는 사람. 모든 권리와 생산 수단을 빼앗기고, 물건처럼 사고팔리던 노예제 사회의 피지배 계급이다.’을 뜻하며 일반 명사에 해당하기 때문에 명시적이지 않은 것으로 주석을 변경하였다.

<표 46> 부적절성 말뭉치 ‘명시성’ 정비 예시

sentence-id	sentence_form	is_explicit (정비 이전)	is_explicit (정비 이후)
ESRW190700172 3.287.1.1	참., 지랄도 대풍년이다	False	True
ESRW190500083 8.442.3.1	#숯불갈비#갈비#화요일#점심#일의노예	True	False

부적절성 말뭉치에서 강도 주석은 명시성과 맥락을 고려하여 정비하였다. 분석 지침 지침에서 명시적 부적절성이 부정적 맥락에서 나타날 때는 강도를 ‘강’으로 주석하고, 명시적 부적절성이 긍정적 맥락일 때는 ‘약’으로 나타내야 하지만 아래 정비 예시와 같이 맥락을 고려하지 않은 강도 주석이 발견되어 정비 대상으로 삼아 수정하였다. 또한 비명시적인 문장의 경우 성적 폭력성, 선정성 등과 같은 부정적 맥락 및 내용에 따라 강도가 결정되는 것도 검토하였다. 아래 정비 예시는 대사전류에서 비어, 속어 등으로 분류되는 어휘가 명시적으로 드러나 명시적 문장이 부정적 맥락에서 나타나는 경우로 강도 주석을 강으로 변경하였다. 또한 명시적인 표현이 나타나지 않은 무표적인(부정적이지 않은) 문장은 강도 주석을 약으로 수정하였다.

<표 47> 부적절성 말뭉치 ‘강도’ 정비 예시

sentence-id	sentence_form	intensity (정비 이전)	intensity (정비 이후)
ESRW190700198 0.104.1.1	존나 파우치 시발 ㅋㅋㅋㅋㅋㅋ	Weak	Strong
EPRW1909002732 .206.2.1	시집가야겠다	Strong	Weak

부적절성 말뭉치 분석 지침 지침에서 부적절 표현의 정의는 명시성의 표현 범위를 의미한다. 따라서 명시적 어휘에 대해 시작-종료 위치와 표현이 주석되어 있으며, 비명시적인 문장의 경우 전체 문장이 시작-종료 위치와 함께 나타나 있다. 이때, 명시적 표현이 문장 안에 존재함에도 그 시작-종료 위치가 전체 문장으로 되어 있는 경우나 부적절 표현의 범위가 올바르게 표현되지 않은 경우를 발견하여 정비 대상으로 삼았다. 부적절 표현의 범위 주석을 정비한 예시는 아래와 같다. 어절 분리 및 띄어쓰기가 온전히 이루어진 문장에서 명시적인 표현이 나타날 때, 부적절 표현의 범위를 별도로 표시하며 주석을 수정하였다.

<표 48> 부적절성 말뭉치 ‘부적절 표현’ 정비 예시

sentence-id	sentence_form	inappropriate_exp ression (정비 이전)	inappropriate_exp ression (정비 이후)
ESRW190700195 3.2416.1.1	아 너무 웃겨 지랄하지 마ㅏ	[{'begin': 0, 'end': 14, 'form': '아 너무 웃겨 지랄하지마ㅏ'}]	[{'begin': 8, 'end': 14, 'form': '지랄하지마ㅏ'}]

## ○ 정비 말뭉치 예시

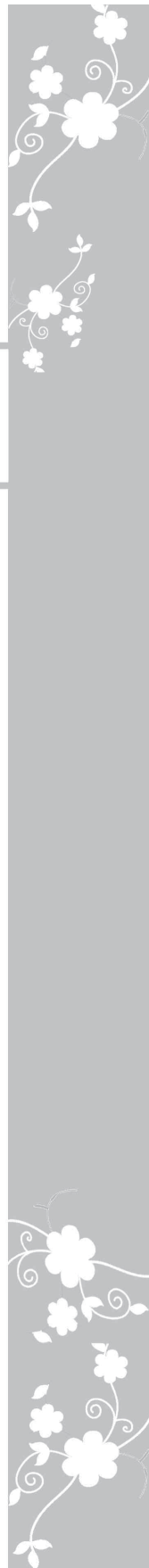
<표 49> 부적절성 JSONL 말뭉치

```
{
  "id": "nikluge-2023-iau-train-000001",
  "input": "존나웃기다ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ",
  "output": "POSITIVE"
}
{
  "id": "nikluge-2023-iau-train-000002",
  "input": "마간호사 존나멋있고 존나웃겨",
  "output": "POSITIVE"
}
{
  "id": "nikluge-2023-iau-train-000003",
  "input": "가던말던니 좇대로해~~",
  "output": "NEGATIVE"
}
{
  "id": "nikluge-2023-iau-train-000004",
  "input": "진짜 존나 무기력하다 큰일남",
  "output": "NEGATIVE"
}
{
  "id": "nikluge-2023-iau-train-000005",
  "input": "미친 &name&",
  "output": "NEGATIVE"
}
{
  "id": "nikluge-2023-iau-train-000006",
  "input": "b조식은 좇같았는데 ㅋㅋㅋㄱㅋㅋ",
  "output": "NEGATIVE"
}
{
  "id": "nikluge-2023-iau-train-000007",
  "input": "개 휘돌린다 ..",
  "output": "NEGATIVE"
}
{
  "id": "nikluge-2023-iau-train-000008",
  "input": "아 시파 ㅋㅋㅋㅋㅋㅋㅋㅋ",
  "output": "NEGATIVE"
}
{
  "id": "nikluge-2023-iau-train-000009",
  "input": "#&company& 은 뭐가 그리 무서워서 노조 하나 못 만들게  
하는지",
  "output": "NEGATIVE"
}
{
  "id": "nikluge-2023-iau-train-000010",
  "input": "치명적인 뒤통스..",
  "output": "POSITIVE"
}
{
  "id": "nikluge-2023-iau-train-000011",
  "input": "양심이 없어?",
  "output": "NEGATIVE"
}
{
  "id": "nikluge-2023-iau-train-000012",
  "input": "게을러서 건방져졌네여...",
  "output": "NEGATIVE"
}
{
  "id": "nikluge-2023-iau-train-000013",
  "input": "&art& 좇집남매얼마나 좇같길래",
  "output": "NEGATIVE"
}
{
  "id": "nikluge-2023-iau-train-000014",
  "input": "생일 축하해 시발럼",
  "output": "POSITIVE"
}
{
  "id": "nikluge-2023-iau-train-000015",
  "input": "우 차단존나 많이 당했누",
  "output": "NEGATIVE"
}
{
  "id": "nikluge-2023-iau-train-000016",
  "input": "나는 어떤 녀석직원이 해줬는데 굶고나서 느낌 이상해서 보니  
까 큐티클 있는데랑 그 손톱 옆라인 손톱 밑에까지 발라놔서 같이 구워짐ㅋㅋㅋㅋㅋㅋ말했더니 또 한숨  
조오 오오오나 폭 쉬면서 손탁 잡아서 다시함ㅋㅋㅋㅋ 그나마도 또 큐티클라인쪽에 고였어",
  "output": "NEGATIVE"
}
```



## 제 3 장

# 인공지능(AI) 말평 경진대회 과제 개발 및 운영





### 3. 인공지능(AI) 말평 경진대회 과제 개발 및 운영

- 경진대회 진행을 위해 과제로 ‘감정 분석’ 과제와 ‘이야기 완성’을 선정하였다. 두 과제는 자연어 이해와 생성을 대표하는 과제로서, 기본적으로 정비 대상 말뭉치 중 데이터 규모, 개발 과제 난이도, 그리고 경진대회 과제로써의 개발 적합성, 자연어 처리에서의 과업 중요성 등을 고려하여 선정되었다. 감정 분석은 2022년 국립국어원 경진대회 과제였던 ‘속성 기반 감성 분석’에 이어 자연어 이해의 가장 기본적인 과제이면서도 대화 시스템 개발에서부터 서비스에 이르기까지 광범위하게 활용되는 과제이기에 선정되었고 ‘이야기 완성’은 국립국어원 경진대회의 첫 자연어 생성 과제로써, 향후 인공지능의 다양한 분야에서 활용될 수 있는 중요한 과제이기에 선정되었다.
- 기준 모델(베이스라인 모델)을 개발하여 가공된 데이터에 대한 적합성을 검증하고 깃허브(github) 등을 통해 모델 및 소스코드를 참가팀에게 공개하여 참가자들이 데이터 활용이나 모델 개발에 참고할 수 있는 기준을 제공하였다.
- 기준 모델 개발 과정에서 기계학습을 수행하고, 학습에 필요한 충분한 양의 데이터가 확보되었음을 보였다.
- 이하 과제 설명은 2023년 인공지능(AI) 말평 경진대회 내 과제 기술서들의 내용으로, 해당 과제 기술서는 참가자들의 이해를 돕기 위해 작성되었다.

<표 50> 기준 모델 목록

과제	기준 모델	기준 성능(2023년)
이야기완성	<a href="https://github.com/teddysum/Korean_SC_2023">https://github.com/teddysum/Korean_SC_2023</a>	0.323 (ROUGE-1) 0.386 (BLEURT) 0.762 (BERT score)
감정분석	<a href="https://github.com/teddysum/Korean_EA_2023">https://github.com/teddysum/Korean_EA_2023</a>	0.850 (F1 micro)

#### 3.1. 감정 분석 과제

##### ▷ 과제 개요

감정 분석은 주어진 텍스트에 대한 화자의 감정 상태를 파악하는 과제이다. 이 과제는 텍스트에 드러나는 8가지 감정 유형을 분류하는 것을 목표로 한다. 감정 분석은 고객 서비스, 사회 네트워크 분석, 피드백 시스템, 인공지능 대화 시스템 등에 널리 활용된다.

<표 51> 감정 분석 과제의 예시

항목	내용
텍스트	"아 뉴스레터에서 뮤지컬 킹아더 관람 신청받는데... 가고싶은데 약속이네 어어아앙악 짜증"
대상	'약속'
감정	"joy": "False", "anticipation": "False", "trust": "False", "surprise": "False", "disgust": "False", "fear": "False", "anger": "True", "sadness": "False"

## ▷ 과제 정의

감정 분석이란 긍정, 부정, 중립으로만 판단하는 감성 분석과는 다르게 기쁨, 신뢰, 놀람, 공포 등 사람이 느끼는 감정들을 분석하는 것을 의미한다. 이 과제는 국립국어원이 제공하는 '감정 분석 말뭉치'를 활용하여 제시된 텍스트에서 특정 대상(target)에 대한 화자의 감정 상태를 파악하고 이를 "joy(기쁨)", "anticipation(기대)", "trust(신뢰)", "surprise(놀람)", "disgust(혐오)", "fear(공포)", "anger(분노)", "sadness(슬픔)"의 8가지 감정으로 분류한다.

감정 분석 과제는 텍스트와 특정 대상이 주어졌을 때, 대상에 대한 화자의 감정을 분류하는 과제로 정의할 수 있다. 각 대상에 대한 감정 레이블은 8가지이며, 이에 따라 기본적으로 다중 분류(multi-class classification)를 수행하여 특정 레이블에 해당하는 감정이 드러날 경우 'True', 아닐 경우 'False' 값으로 표시한다. 한편 한 텍스트 내에 여러 개의 감정이 드러날 수도 있으므로 하나의 텍스트가 여러 개 레이블에 대해 'True' 값을 가지는 다중 레이블 분류(Multi-label classification)로 수행할 수도 있다. 이 과제에서의 평가는 각 레이블의 'True/False' 값에 대한 'F1-score(micro)'를 기준으로 순위를 결정한다.

<표 52> 감정 분석 평가 지표

$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}), \text{Recall} = \text{TP} / (\text{TP} + \text{FN}), \text{F1 Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$	
- True Positives (TP): 모델이 'True'로 예측했고, 실제 측정값도 'True'인 경우.	즉, 모델이 해당 레이블을 제대로 예측한 경우
- False Positives (FP): 모델이 'True'로 예측했지만, 실제 측정값은 'False'인 경우.	즉, 모델이 레이블을 잘못 예측한 경우
- False Negatives (FN): 모델이 'False'로 예측했지만, 실제 측정값은 'True'인 경우.	즉, 모델이 레이블을 놓친 경우
- True Negatives (TN): 모델이 'False'로 예측했고, 실제 측정값도 'False'인 경우.	즉, 해당 레이블이 없다고 예측하였고, 실제로도 그 레이블이 없는 경우

<표 53> 감정 분석의 모델 입력과 출력의 예

분류	내용		예시	자료형
입력	텍스트		"아 뉴스레터에서 뮤지컬 킹아더 관람 신청받는데... 가고 싶은데 약속이네 ㅇ어으아앙악 짜증"	문자열
	대상	form	"약속"	문자열
		begin, end	35, 37	정수
출력	8가지 감정에 대한 분석(True, False)		{ "joy": "False", "anticipation": "False", "trust": "False", "surprise": "False", "disgust": "False", "fear": "False", "anger": "True", "sadness": "False" }	딕셔너리 (dictionary )
평가	F1 점수			

## ▷ 자료 형식

데이터 세트는 'JSONL(jsonlines)' 형식으로 제공되며, 각 'JSON'은 텍스트, 텍스트 내 감정 분석 대상, 대상에 대한 화자의 감정 정보를 제공한다. 아래 표는 데이터의 규모이다. 훈련, 검증, 시험 데이터는 무작위로 분할되었다.

<표 54> 감정 분석 데이터 규모

	훈련	검증	시험
텍스트 수	37,932	4,751	4,748

<표 31>은 데이터의 예시이다. 주어진 훈련 데이터와 시험 데이터는 동일한 'JSONL' 형식으로 제공되며, 시험 데이터의 경우에는 각 텍스트에 대한 출력 항목이 빈 목록으로 제공된다. 참가팀은 해당 목록에 대해 모델의 출력 결과를 추가하여 제출한다. 훈련 데이터와 제출용 데이터의 형태 및 구성 요소는 동일하다.

<표 55> 감정 분석 데이터 형식의 예

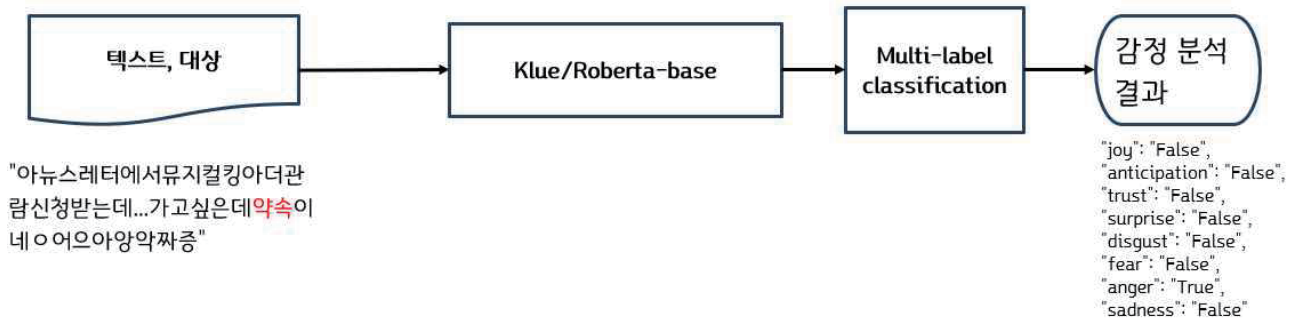
항목	내용
훈련용 데이터의 예	{ "id": "데이터id1", "input": { "form": "아뉴스레터에서뮤지컬킹아더관람신청받는데...가고싶은데약속"

	<pre> 이네ㅇ어으아앙악짜증", "target": {   "form": "약속",   "begin": 35,   "end": 37 } }, "output": {   "joy": "False",   "anticipation": "False",   "trust": "False",   "surprise": "False",   "disgust": "False",   "fear": "False",   "anger": "True",   "sadness": "False" } } </pre>
	<p>- 아이디(id)와 입력(input), 그리고 출력(output)으로 구성</p>
평가용 데이터의 예 (제출 전)	<pre> {   "id": "데이터id1",   "input": {     "form": "아뉴스레터에서뮤지컬킹아더관람신청받는데...가고싶은데약속 이네ㅇ어으아앙악짜증",     "target": {       "form": "약속",       "begin": 35,       "end": 37     }   } } </pre>
	<p>- 학습용 데이터와 동일한 형태</p> <p>- “output” 키와 값을 제거한 데이터</p>
제출 데이터	<pre> {   "id": "데이터id1",   "input": {     "form": "아뉴스레터에서뮤지컬킹아더관람신청받는데...가고싶은데약속 이네ㅇ어으아앙악짜증",     "target": {       "form": "약속",       "begin": 35,       "end": 37     }   },   "output": {     "joy": "False",     "anticipation": "False",     "trust": "False",     "surprise": "False",     "disgust": "False",     "fear": "False",     "anger": "True",     "sadness": "False"   } } </pre>
	<p>- 평가용 데이터에 “output”을 생성. 감정 category는 고정</p>



## ▷ 베이스라인 모델

본 대회에 베이스라인 모델은 깃허브(github)<sup>1)</sup>를 통해 공개되어 있다. 해당 모델은 Klue/Roberta-base<sup>2)</sup>를 기반으로 다중 레이블 분류(multi-label classification)로 각 라벨은 이진 분류로 진행한 모델이다.



[그림 5] 감정 분석 베이스라인 모델 개념도

1) [https://github.com/teddysum/Korean\\_EA\\_2023](https://github.com/teddysum/Korean_EA_2023)

2) <https://huggingface.co/klue/roberta-base>

## 3.2. 이야기 완성 과제

### ▷ 과제 개요

이야기 완성 과제는 제공된 문장들을 논리적으로 연결하는 문장을 생성하는 과제이다. 이 과제를 통해 문장들의 맥락을 파악하고 연결고리를 찾는 과정을 통해 기계의 언어 이해 능력을 향상시키는 데 기여할 수 있으며, 이어지는 문장을 생성하게 함으로써 언어 생성 능력을 측정할 수 있다. 이야기 완성 과제는 인공지능 챗봇, 자동 번역, 문서 요약 등 다양한 분야에서 활용될 수 있다. 아래는 이야기 완성 과제의 적절한 예시와 부적절한 예시들이다.

<표 56> 이야기 완성 과제 예시

항목	입출력		내용	비고
적절한 예시 1	입력	문장 1	나는 입사하고 나서 몇 달 동안은 조심스럽게 행동했다.	주어진 예시의 1번 문장에서는 입사 후 초기 화자의 태도를, 3번 문장에서는 달라진 화자의 태도에 대한 동료들의 반응을 말해주고 있습니다. 특히, 3번 문장은 ‘그랬더니’로 시작하는데, 이는 앞선 내용이 원인이 되고 그 결과가 뒤에 이어짐을 나타냅니다. 2번 문장은 입사 이후 몇 달이 지나고 더 나아가 회사 분위기를 다 파악할 정도의 시간이 지나면서 화자의 태도가 어떤 식으로 변화했는지를 설명하고 있습니다. 2번 문장의 앞부분은 논리 흐름상 1번 문장 뒤에 자연스럽게 이어지며, 2번 문장의 뒷부분 ‘눈치껏 내 성격을 드러냈다’는 3번 문장의 ‘그랬더니’의 지시 대상을 보여 줍니다. 그러므로 주어진 예시의 2번 문장은 1번과 3번 문장 사이에 들어가기에 적절하다고 할 수 있습니다.
		문장 3	그랬더니 동료들은 첫인상과 다른 나의 모습에 놀랐습니다.	
	출력	문장 2	회사 분위기를 파악하고 나서는 눈치껏 내 성격을 드러냈다.	
적절한 예시 2	입력	문장 1	민수는 요즘 보드 타는 것에 취미를 들였다.	1번 문장은 ‘민수’가 최근에 새로운 취미를 가지게 되었음을 제시하고 있습니다. 3번 문장에서는 ‘하지만’이라는 접속사와 함께, 민수가 계속해서 보드를 타고 있음을 말해 주고 있습니다. 1번에서 민수가 보드에 취미를 들였다고 이야기를 한 상황이기 때문에 적절한 이유가 제시되지 않는다면, 3번 문장에서 ‘하지만’이라는 접속사와 함께 그럼에도 불구하고 보드 타는 것을 계속했다는 내용을 전달하는 것은 어색합니다. 따라서, 1번과 3번 문장의 흐름을 자연스럽게 연결하기 위해서는 보통 그 정도의 일이 발생하면 보드 타는 것을 멈춘만한 이유가 2번 문장에 제시되어야 합니다. 위의 예시에서는 보드를 타다가 다쳤다는 내용이 2번 문장에 제시되어 있기 때문에 세 문장이 적절하게 연결된다고 할 수 있습니다.
		문장 3	하지만 보드 타는 것을 멈추지 않았다.	
	출력	문장 2	그는 맨날 보드를 타다가 다쳐서 나타났다.	
적절한 예시 3	입력	문장 1	최근에 잠을 못 자서 피로가 쌓인 기분이 들었다.	1번 문장에 화자가 현재 잠을 제대로 자지 못해 피로가 쌓였다는 상황이 제시되어 있습니다. 3번 문장에는 ‘덕분에’라는 접속사와 함께, 잠을 잘 자서 피로가 풀렸다는 이야기가 제시되어 있습니다. 그러므로 2번 문장에 잠을 잘 잘 수 있었던 이유가 제시되어야 내용이 적절하게 이어진다고 평가할 수 있습니다. 2번 문장에서 자기 전에 따뜻한 우유를 마신 것은 잠을 잘 잘 수 있었
		문장 3	덕분에 잠을 잘 자서 피로가 풀렸다.	

항목	입출력		내용	비고
	출력	문장 2	잠을 잘 자기 위해 자기 전에 따뜻한 우유를 마셨다.	던 이유로 타당성이 인정되기 때문에, 세 문장이 자연스럽게 연결된다고 할 수 있습니다.

<표 57> 이야기 완성 과제의 부적절한 예시

항목	입출력		내용	비고
부적절한 예시 1	입력	문장 1	나는 목표했던 일을 이루지 못할까봐 떨렸다.	1번 문장은 현 상황에 대한 화자의 감정을, 3번 문장은 앞의 상황으로 인한 화자의 감정을 드러냅니다. 1번 문장과 3번 문장 사이에서 화자의 심경에 변화('떨림'에서 '뿌듯함, 시원한 감정'으로 변화)가 생겼습니다. 그러므로 이 두 문장을 자연스럽게 이어주기 위해서는 이러한 심경 변화의 원인이나 심경 변화를 일으킬 수 있는 상황이 제시되어야 합니다. 그러나 2번 문장에는 노력에 대한 화자의 생각이 드러나 있을 뿐, 화자의 심경 변화에 대한 원인을 추론할 수 있는 정보는 없습니다. 그 결과, 주어진 맥락에서 2번 문장은 부적절하다고 할 수 있습니다.
		문장 3	나는 뿌듯함과 시원한 감정이 동시에 들었다.	
	출력	문장 2	나는 노력이 중요하다고 생각한다.	
부적절한 예시 2	입력	문장 1	나는 학교 축제에서 사회자를 맡기 위해 오디션에 서류 지원을 했다.	맥락을 이해하는 데 문장들 사이의 시간적 흐름이 중요합니다. 1번 문장에서 화자가 학교 축제 사회자를 위한 오디션에 지원했다는 언급이 등장하고, 3번 문장에서 최종 면접을 보러 갔다는 이야기를 하고 있습니다. 3번 문장이 되어서야 최종 면접을 보러 갔다는 언급이 등장하기 때문에, 그 전 맥락에서는 최종 면접을 보러 가기 전 상황(예컨대 '서류 전형 합격 통지를 받는 일' 등)에 대한 이야기만 등장해야 합니다. 그러나 현재 예시의 2번 문장에서는 최종 면접 합격이라는 정보가 제시되어 있습니다. 이는 최종 면접 이후에만 파악할 수 있는 정보이므로 3번 문장에 앞서 등장하기에 부적절합니다.
		문장 3	나는 문자로 통보받은 날짜에 최종 면접을 보러 학교 방송실로 갔다.	
	출력	문장 2	최종 면접에 합격했다는 문자를 받았다.	
부적절한 예시 3	입력	문장 1	나는 숙제를 하지 않고 휴대전화로 게임을 하면서 놀았다.	주어진 세 문장 중, 1번 문장에서는 화자가 숙제를 미루고 있는 상황을, 3번 문장에서는 화자가 숙제를 하러 들어가는 상황을 보여 주고 있습니다. 3번 문장은 '그래서'라는 접속사가 문장 첫머리에 등장합니다. '그래서'는 앞의 내용이 뒤의 내용의 원인, 근거, 조건 등이 될 때 쓰는 말이기 때문에, 그 앞의 내용에 해당하는 2번 문장에는 화자가 숙제하러 방에 들어간 적절한 원인이 제시되어야 합니다. 가령 "숙제도 하지 않고 논다고 엄마한테 꾸중을 들었다." 정도의 문장이 2번 문장 자리에 오면 논리상 문장 3으로 순조롭게 연결될 수 있습니다. 그러나 현재 2번 문장의 경우, 화자의 엄마가 화자의 현 상태를 보고 칭찬을 하는 것이기 때문에 1번과 3번 문장에서 발생하고 있는 상황 변화에 대한 원인을 적절하게 설명하지 못합니다.
		문장 3	그래서 나는 숙제를 하러 방에 들어갔다.	
	출력	문장 2	그러자 엄마가 나에게 좋은 습관을 가지고 있다며 칭찬했다.	
부적절한 예시	입력	문장 1	나는 환경을 보호하기 위해 쓰레기들을 주웠다.	주어진 예시의 1번 문장과 3번 문장은 화자가 환경보호를 위해 길 위의 쓰레기를 주었으며, 이러한 행동을 계기로 동생과 함께 동아리를 만들었다고 이야기하고 있습니다. 1번 문장에는 화자의 동생에 대한 언급이 등장하지 않지만, 3번 문장에는 화자와 화자의 동생이 동

항목	입출력		내용	비고
4		문장 3	그러다 동생과 나는 같이 쓰레기를 줍는 동아리를 만들었다.	아리를 만들었다는 이야기가 제시되어 있기 때문에, 두 문장을 조금 더 자연스럽게 이어주기 위해서는 화자의 행동에 대한 동생의 태도를 설명하는 문장이 필요합니다. 화자와 화자의 동생이 함께 동아리를 만들었다는 점에서, 쓰레기를 줍는 화자의 행동에 동생 또한 동조했음을 추론할 수 있습니다. 주어진 예시의 2번 문장에서는 이러한 추론과는 반대로 쓰레기를 줍는 내 행동을 동생이 보고도 모른 체했다는 언급이 나오기 때문에 부적절하다고 할 수 있습니다.
	출력	문장 2	내가 쓰레기를 줍는 모습을 동생은 본체만체했다.	
부적절한 예시 5	입력	문장 1	그는 뮤지컬 공연 오케스트라를 지휘하여 공연을 성공적으로 마쳤다.	1번 문장에서는 '그'가 공연을 성공적으로 마쳤음을 제시하고 있으며, 3번 문장에서는 공연을 성공적으로 마친 '그'가 기쁜 마음으로 다음 공연을 준비하러 갔다는 내용을 제시하고 있습니다. 그러므로 두 문장 사이에는 성공적인 공연에 대한 반응으로 적절하면서, 동시에 '그'가 다음 공연을 기쁜 마음으로 준비할 수 있게 하는 동기가 될 수 있을 만한 상황이 제시되어야 합니다. 그러나 주어진 예시의 2번 문장의 경우, 관객들이 '그'에게 야유를 보냈다는 내용이 등장합니다. '관객의 야유'는 공연을 성공적으로 마친 지휘자에 대한 적절한 반응으로 보기 어렵고, 그 지휘자로 하여금 기쁜 마음으로 다음 공연을 준비하러 나가게 하는 적절한 동기로 보기도 어렵습니다. 따라서 예시된 2번 문장은 부적절한 문장이라고 할 수 있습니다.
		문장 3	그리고 그는 기쁜 마음으로 바로 다음 공연을 준비하러 나갔다.	
	출력	문장 2	그러나 관객들은 그에게 야유를 보냈다.	

## ▷ 과제 정의

이야기 완성 과제는 주어진 '문장 1번'과 '문장 3번'을 논리적으로 연결하는 '문장 2번'을 생성하는 것이 목표이다. 생성된 문장은 문맥적으로 일관성 있고, 문법적으로 정확하며, 논리적으로 문장 1번과 문장 3번을 연결할 수 있어야 한다. 이에 '이야기 완성' 과제의 학습용 데이터 세트는 '문장 1 - 문장 2 - 문장 3'의 형태로 구성되며, 생성된 문장 2의 문맥적/문법적 정확성 및 논리적 일관성이 평가 기준이 된다.

생성된 문장 2의 품질은 정량적 평가와 정성적 평가를 종합하여 판단한다. 정량적 평가는 ROUGE, BERTScore, BLEURT의 평균으로 평가되며, 정성적 평가는 생성된 문장에 대해 다수의 인간 평가자가 순위를 매긴다. 정량적 평가를 먼저 진행하고, 정량적 평가에서 높은 점수를 받은 상위 팀들을 대상으로 정성적 평가를 진행한다. 최종 점수는 정량적 평가 50%, 정성적 평가 50%를 반영한 종합점수로 평가한다. 정성적 평가의 점수는 아래와 같은 수식을 통해 정량적 평가 결과와 같은 가중치를 부여한다. 예를 들어 10개 팀에 대해 정성 평가를 진행한다고 했을 때, 정량 평가 1등 팀 점수가 100점, 10등 팀 점수가 90점일 때, 정성 평가 1등 팀은 10점 $[(100-90)/10 * (10-1+1)]$ , 정성 평가 10등 팀은 1점 $[(100-90)/10 * (10-10+1)]$ 이 된다.

$quantscore_k$  = 정량 평가 k등 팀의 정량 평가 점수

$qualscore_k$  = 정성 평가 k등 팀의 정성 평가 점수

$$qualscore_k = \frac{(quantscore_1 - quantscore_n)}{n} * (n - k + 1)$$

정량 평가 i등, 정성 평가 j등의 최종 score =  $0.5 * quantscore_i + 0.5 * qualscore_j$

[그림 6] 이야기 완성 평가 metric. n = 정성 평가 대상 팀 수

<표 58> 이야기 완성의 모델 입력과 출력 예시

분류	내용	예시	자료형
입력	문장 1(앞 문장)	"나는 할아버지 댁에 건너가기 전에 어머니께 연락을 드렸다."	문자열
	문장 3(뒤 문장)	"나는 어머니께 정확히 언제 돌아올지 모르겠다고 말했다."	문자열
출력	문장 2(중간 문장)	"어머니는 나에게 언제 돌아올 것인지 물으셨다."	문자열
평가	ROUGE-1, BERTScore, BLEURT의 평균		

## ▷ 자료 형식

데이터 세트는 'JSONL(jsonlines)' 형식으로 제공되며, 각 'JSON'의 입력에는 문장 1과 문장 3, 출력에는 문장 2를 제공한다. <표 35>는 이야기 완성 데이터의 규모이다. 훈련, 검증, 시험 데이터는 무작위로 분할되었다.

<표 59> 이야기 완성 데이터 규모

구분	훈련	검증	시험
발화 수	120,140	15,017	15,018

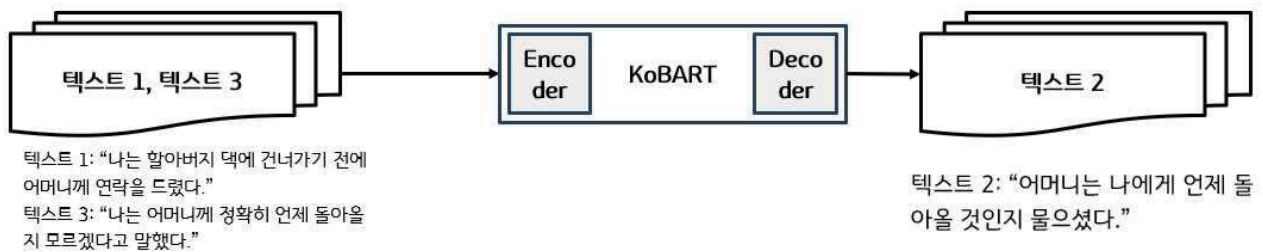
<표 36>은 데이터의 예시이다. 주어진 훈련 데이터와 시험 데이터는 동일한 'JSON-L' 형식으로 제공되며, 시험 데이터의 경우에는 각 발화에 대한 출력 항목이 빈 목록으로 제공된다. 참가팀은 해당 목록에 대해 모델의 출력 결과를 추가하여 제출한다. 훈련 데이터와 제출용 데이터의 형식 및 형태는 동일하다.

<표 60> 이야기 완성 데이터 형식의 예

항목	내용
훈련용 데이터의 예	<pre>{   "id": "데이터id1",   "input": {     "sentence1": "나는 할아버지 댁에 건너가기 전에 어머니께 연락을 드렸다.",     "sentence3": "나는 어머니께 정확히 언제 돌아올지 모르겠다고 말했다."   },   "output": "어머니는 나에게 언제 돌아올 것인지 물으셨다." }</pre> <ul style="list-style-type: none"> <li>- 아이디(id)와 입력(input), 그리고 출력(output)으로 구성</li> </ul>
평가용 데이터의 예 (제출 전)	<pre>{   "id": "데이터id1",   "input": {     "sentence1": "나는 할아버지 댁에 건너가기 전에 어머니께 연락을 드렸다.",     "sentence3": "나는 어머니께 정확히 언제 돌아올지 모르겠다고 말했다."   } }</pre> <ul style="list-style-type: none"> <li>- 학습용 데이터와 동일한 형태</li> <li>- “output” 키와 값을 제거한 데이터</li> </ul>
제출 데이터	<pre>{   "id": "데이터id1",   "input": {     "sentence1": "나는 할아버지 댁에 건너가기 전에 어머니께 연락을 드렸다.",     "sentence3": "나는 어머니께 정확히 언제 돌아올지 모르겠다고 말했다."   },   "output": "어머니는 나에게 언제 돌아올 것인지 물으셨다." }</pre> <ul style="list-style-type: none"> <li>- 평가용 데이터에 “output”에 생성된 문장 입력</li> </ul>

▷ 기준 모델(baseline model)

이 대회<sup>3)</sup>의 기준 모델은 ‘깃허브(github)<sup>3)</sup>’를 통해 공개되어 있다. 해당 모델은 koBART를 기반으로 학습된 모델이다.



[그림 7] 이야기 완성 기준 모델(baseline model) 개념도

3) [https://github.com/teddysum/Korean\\_SC\\_2023](https://github.com/teddysum/Korean_SC_2023)

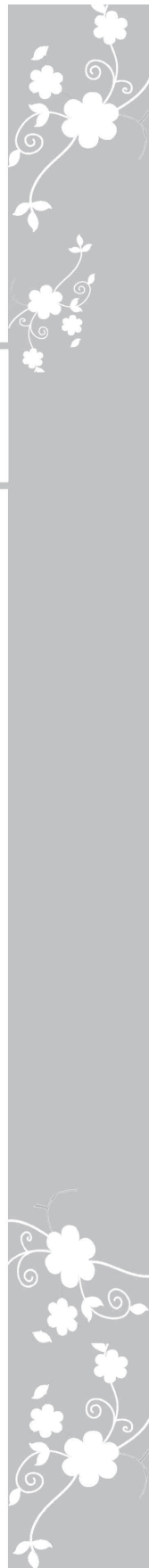






## 제 4 장

# 인공지능(AI) 말평 상시과제 개발 및 운영





## 4. 인공지능(AI) 말평 상시과제 개발 및 운영

### 4.1. 상시과제 과제 선정 과정

- 상시과제 언어 능력 평가 문제(이하 과제)는 표의 일부분에 대한 해석 생성, 문자가 포함된 이미지 기반 문장 생성, 합의 분석, 부적절성 문장에 대한 태도 탐지 4가지로 구성되어 있다.
- 생성 문제의 경우 두 문제 모두 정답이 5개로 많은 편이고, 정답이 될 수 있는 문장의 다양성이 그리 크지 않아 rouge1, rouge-L, bleu 와 같은 지표로 평가될 수 있다. 분류 문제의 경우 micro-f1 score와 macro-f1 score를 함께 사용하여 데이터 수가 적은 class에 대해서도 잘 분류하는지 평가할 수 있도록 하였다.
- 이하 과제 설명은 2023년 인공지능(AI) 말평 상시과제 내 과제 기술서들의 내용으로, 해당 과제 기술서는 참가자들의 이해를 돕기 위해 작성되었다.

<표 61> 상시과제 과제 개요

과제	평가 지표	과제 정의
표의 일부분에 대한 해석 생성	rouge1, rouge-l, bleu	표와 가리키는 부분(highlight)을 주어졌을 때, 가리키는 부분에 대한 해석 문장을 생성하는 과제
문자가 포함된 이미지 기반 문장 생성	rouge1, rouge-l, bleu	이미지와 이미지에 포함된 문자에 대한 OCR 데이터가 주어졌을 때, 이를 설명하는 문장을 생성하는 과제
합의 분석	micro-f1, macro-f1	상식 및 추론을 바탕으로 하여 전제와 명제 문장들이 함의하는지 모순되는지를 판단하는 과제
부적절성 문장에 대한 태도 탐지	micro-f1, macro-f1	부적절하게 표현된 문장 표현의 문맥상 긍정성 또는 부정성을 판단하는 과제

○ 자동 평가를 위하여 json 데이터가 한 줄마다 들어간 jsonl 데이터 형식을 사용한다.

<표 62> 상시과제 과제 데이터 형식

json 하나의 형식	<pre>{   "id": 데이터id,   "input": 입력 데이터,   "output": 출력 데이터 }</pre>
jsonl 형식	<pre>{ "id": 데이터id1, "input": 입력 데이터1, "output": 출력 데이터1 } { "id": 데이터id2, "input": 입력 데이터2, "output": 출력 데이터2 } { "id": 데이터id3, "input": 입력 데이터3, "output": 출력 데이터3 } { "id": 데이터id4, "input": 입력 데이터4, "output": 출력 데이터4 } { "id": 데이터id5, "input": 입력 데이터5, "output": 출력 데이터5 } { "id": 데이터id6, "input": 입력 데이터6, "output": 출력 데이터6 }</pre>
jsonl 데이터 예시	<pre>{ "id": "nikluge-2023-te-dev-000001",   "input": {     "premise": "시리즈 1차분에는 김언희의 첫 시집 ‘트렁크’를 필두로 김사인 ‘밤에 쓰는 편지’, 이수명 ‘새로운 오독이 거리를 메웠다’, 성석제 ‘낮선 길에 묻다’, 정미정 ‘대머리와 사랑’, 함민복 ‘우울씨의 일일’, 진수미 ‘달의 코르크 마개가 열릴 때까지’, 박정대 ‘단편들’, 유형진 ‘피터래빗 저격사건’, 박상수 ‘후르츠 캔디 버스’가 들어 있다.",     "proposition": "시리즈 1차분에는 10가지 시가 들어있다."   },   "output": "contradiction" }</pre>

○ 또한 기준 모델을 개발하여 가공된 데이터에 대한 적합성을 검증하고, 깃허브(github) 등을 통해 모델 및 소스코드를 참가팀에게 공개하여 참가자들이 데이터 활용이나 모델 개발에 참고할 수 있는 기준을 제공하였다. 기준 모델 개발 과정에서 기계학습을 수행하고, 학습에 필요한 충분한 양의 데이터가 확보되었음을 보였다.

<표 63> 기준 모델 목록

과제	기준 모델	기준 성능(2023년)
표의 일부분에 대한 해석 생성	<a href="https://github.com/teddysum/korean_T2_T_2023">https://github.com/teddysum/korean_T2_T_2023</a>	0.3945 (ROUGE-1) 0.4585 (BLEU)
문자가 포함된 이미지 기반 문장 생성	<a href="https://github.com/teddysum/korean_IC_2023">https://github.com/teddysum/korean_IC_2023</a>	0.3191 (ROUGE-1) 0.4061 (BLEU)
함의 분석	<a href="https://github.com/teddysum/Korean_TE_2023">https://github.com/teddysum/Korean_TE_2023</a>	0.68 (F1-micro) 0.52 (F1-macro)
부적절성 문장에 대한 태도 탐지	<a href="https://github.com/teddysum/Korean_IAU_2023">https://github.com/teddysum/Korean_IAU_2023</a>	0.885 (F1-micro) 0.825 (F1-macro)

## 4.2. 과제 정의

### ○ 4.2.1. 함의 분석 과제

#### ▷ 과제 개요

‘자연어 추론(Natural Language Inference)’ 또는 ‘함의 분석(Textual Entailment)’은 대표적인 언어 능력 평가체계(벤치마크)인 GLUE, KLUE 등에서도 제공하는 과제로, 주로 두 문장의 관계를 ‘함의/중립/모순’ 중 하나로 분류하는 것을 평가한다.

함의 분석(Textual Entailment)은 상식 및 추론을 바탕으로 하여 전제와 명제 문장들이 함의하는지 모순되는지를 판단하는 과제인데, 이러한 과제를 해결하기 위해서는 인공지능이 유의어를 이해하고 논리적·산술적으로 추론하는 능력이 필요하다. 그래서 함의 분석 과제는 최근 거대언어모델(LLM)의 자연어 추론 능력을 평가하거나(NLI, SNLI) 생성 능력(요약 등)을 평가하는 데에 활용되고 있다.

국립국어원에서는 2022년 ‘말뭉치 함의 분석 및 연구’ 사업을 통해 함의 분석 말뭉치를 구축하였다. 이 말뭉치는 주어진 문장(전제, premise)에 대해 명제(proposition)가 내용을 함의하는지를 주석한 말뭉치이며, 데이터 세트의 예시는 <표 40>과 같다.

<표 64> 함의 분석 과제 데이터 세트의 예시

항목	내용
전제 Premise	납입 한도도 월 50만원으로 한도가 높은 편이다. 적금 개설 시 영업점 창구에 스마트폰 등을 통해 반려견과 함께 찍은 사진을 제시하면 별도의 까다로운 과정 없이 가입할 수 있다.
명제 Proposition	반려견과 함께 찍은 사진을 제시할 수 있는 적금은 한달 동안 최소 50만원 이상을 납입해야 한다.
레이블 Label	contradiction

## ▷ 과제 정의

함의 분석 과제는 국립국어원에서 2022년 ‘말뭉치 함의 분석 및 연구’ 사업을 통해 구축한 “함의 분석 말뭉치”를 활용한다. 해당 데이터 세트의 함의 주석은 함의(entailment), 모순(contradiction), 중립(neutral) 세 가지로 주석되어 있다. 함의 분석 과제는 입력 전제(Premise)와 명제(Proposition)에 대해 레이블을 예측하는 방식으로 정의되어 있다. 참가자들은 평가 데이터 세트의 입력(Premise, Proposition)을 기반으로 함의 정보(entailment, contradiction, neutral)를 예측한다. 본 과제에서는 예측 결과에 대한 F1-score(micro, macro) 점수를 평가 점수로 제공한다.

<표 65> 함의 분석 모델 입력과 출력의 예

분류	내용	예시	비고
입력	Premise	납입 한도도 월 50만원으로 한도가 높은 편이다. 적금 개설 시 영업점 창구에 스마트폰 등을 통해 반려견과 함께 찍은 사진을 제시하면 별도의 까다로운 과정 없이 가입할 수 있다.	문자열
	Proposition	반려견과 함께 찍은 사진을 제시할 수 있는 적금은 한 달 동안 최소 50만원 이상을 납입해야 한다.	문자열
출력	Label	contradiction	문자열
평가	Micro F1-score, Macro F1-score		

## ▷ 데이터 형식

데이터 세트는 JSON-L 형식으로 제공되며, 각 입력(Premise, Proposition) 문장들에 대한 함의 레이블이 문자열(string)로 주석되어 있다. <표 42>는 데이터 규모를 보여준다. 출력(output)값의 분포가 유사하도록 무작위로 분할되었다.

<표 66> 함의 분석 데이터 규모

구분	훈련	검증	시험
문장 수	12,019	1,502	1,503

<표 43>은 데이터 예시이다. 주어진 훈련 데이터와 시험 데이터는 동일한 JSONL 형식으로 제공되며, 시험 데이터의 경우에는 각 문장에 대한 출력(output) 항목이 빈 목록으로 제공된다. 참가팀은 해당 목록에 대해 모델의 출력 결과를 추가하여 제출한다.

<표 67> 데이터 형식의 예

항목	내용
훈련용 데이터의 예	<pre> {   "id": "nikluge-2023-te-dev-000001",   "input": {     "premise": "시리즈 1차분에는 김언희의 첫 시집 ‘트렁크’를 필두로 김 사인 ‘밤에 쓰는 편지’, 이수명 ‘새로운 오독이 거리를 매웠다’, 성석제 ‘낮 선 길에 묻다’, 성미정 ‘대머리와 사랑’, 함민복 ‘우울씨의 일일’, 진수미 ‘달의 코르크 마개가 열릴 때까지’, 박정대 ‘단편들’, 유형진 ‘피터래빗 저격 사건’, 박상수 ‘후르츠 캔디 버스’가 들어 있다.”,     "proposition": "시리즈 1차분에는 10가지 시가 들어있다."   },   "output": "contradiction" } </pre> <p>- 아이디(id)와 입력 문장(input), 그리고 출력으로 구성</p>
평가용 데이터의 예 (제출 전)	<pre> {   "id": "nikluge-2023-te-dev-000001",   "input": {     "premise": "시리즈 1차분에는 김언희의 첫 시집 ‘트렁크’를 필두로 김 사인 ‘밤에 쓰는 편지’, 이수명 ‘새로운 오독이 거리를 매웠다’, 성석제 ‘낮 선 길에 묻다’, 성미정 ‘대머리와 사랑’, 함민복 ‘우울씨의 일일’, 진수미 ‘달의 코르크 마개가 열릴 때까지’, 박정대 ‘단편들’, 유형진 ‘피터래빗 저격 사건’, 박상수 ‘후르츠 캔디 버스’가 들어 있다.”,     "proposition": "시리즈 1차분에는 10가지 시가 들어있다."   },   "output": "contradiction" } </pre> <p>- 학습용 데이터와 동일한 형태 - “output” 키와 값을 제거한 데이터</p>
제출 데이터	<pre> {   "id": "nikluge-2023-te-dev-000001",   "input": {     "premise": "시리즈 1차분에는 김언희의 첫 시집 ‘트렁크’를 필두로 김 사인 ‘밤에 쓰는 편지’, 이수명 ‘새로운 오독이 거리를 매웠다’, 성석제 ‘낮 선 길에 묻다’, 성미정 ‘대머리와 사랑’, 함민복 ‘우울씨의 일일’, 진수미 ‘달의 코르크 마개가 열릴 때까지’, 박정대 ‘단편들’, 유형진 ‘피터래빗 저격 사건’, 박상수 ‘후르츠 캔디 버스’가 들어 있다.”,     "proposition": "시리즈 1차분에는 10가지 시가 들어있다."   },   "output": "contradiction" } </pre> <p>- 평가용 데이터에 “output”과 label 추가하여 제출</p>

#### ▷ 기준 모델(베이스라인 모델)

이 대회 기준 모델(베이스라인 모델)은 깃허브(github)<sup>4)</sup>를 통해 공개되어 있다. 해당 모델은 klue/RobERTa 모델을 사용하여 학습되었으며, 모델 구조는 klue/roberta-base 모델의 <s> 토큰 output에 SimpleClassifier인 FFNN을 붙인 형태의 모델이다.

4) [https://github.com/teddysum/korean\\_TE\\_baseline](https://github.com/teddysum/korean_TE_baseline)

## ○ 4.2.2. 표의 일부분에 대한 해석 생성

### ▷ 과제 개요

‘표의 일부분에 대한 해석 생성’은 자료로부터 텍스트를 생성하는 과제 중 하나로, 주어진 표의 특정 부분을 설명하는 문장을 만드는 과제이다. 위키피디아 등 다양한 웹 문서 내에서 핵심적인 정보는 표 형식으로 기술되어 있는 경우가 많다. 이러한 데이터를 인공지능이 잘 이해하기 위해서는 인공지능 언어 처리 기술을 통해 표의 내용을 잘 요약하고 설명할 수 있는지 평가할 필요가 있다.

### ▷ 과제 정의

‘표의 일부분에 대한 해석 생성’ 과제는 국립국어원의 ‘2022년 유사 문장 생성 말뭉치 연구 및 구축’ 사업을 통해 구축한 자료 중 표 기반 문장 생성 결과물을 활용하여 개발되었다. 이 자료는 해외 표 기반 문장 생성의 대표적인 데이터인 구글의 ‘ToTTo’ 데이터 세트를 참조하였다. 데이터 세트는 HTML로 작성된 표의 형식을 유지하여 JSON 형식으로 변환하고, 해당 표에 음영으로 표시한 부분을 설명하는 문장 5개로 구성된다.

이 과제의 목표는 표에 음영으로 표시한 부분을 설명하는 문장 한 개를 생성하는 것이다. 주어진 표에 대하여 모델이 생성한 문장과 정답 문장 5개 각각을 비교하여 산출한 ROUGE-1, ROUGE-L, BLEU 점수의 평균값을 평가 성능 지표로 사용한다.

<표 68> 표의 일부분에 대한 해석 생성 과제 모델 출력의 예

분류	내용	예시	비고
입력	표	<pre> <input": "2020-06-09",="" "4차="" "date":="" "highlighted_cells":="" "https:="" "metadata":="" "publisher":="" "table":="" "table_title":="" "title":="" "url":="" "국제조세="" "국회예산정책처",="" 0,="" 01_01_board.js="" 01report="" 1="" 1,="" 2,="" <="" [="" ]="" ],="" p",="" pre="" sub="" www.nabo.go.kr="" {="" },="" 개념",="" 과세원칙="" 과제",="" 따른="" 변화와="" 산업혁명에="" 일반="" 정책="" 조세환경=""> </input":></pre>	JSON



		<pre> {   "value": "과세원칙",   "is_header": true,   "col": 0,   "colspan": 1,   "row": 0,   "rowspan": 1 }, {   "value": "특징",   "is_header": true,   "col": 1,   "colspan": 1,   "row": 0,   "rowspan": 1 }, {   "value": "이중과세 조정",   "is_header": true,   "col": 2,   "colspan": 1,   "row": 0,   "rowspan": 1 }, {   "value": "원천지국 과세",   "is_header": false,   "col": 0,   "colspan": 1,   "row": 1,   "rowspan": 1 }, {   "value": "소득이 발생한 국가(원천지국)에서 과세관할권 보유",   "is_header": false,   "col": 1,   "colspan": 1,   "row": 1,   "rowspan": 1 }, {   "value": "국외소득면제",   "is_header": false,   "col": 2,   "colspan": 1,   "row": 1,   "rowspan": 1 }, {   "value": "거주지국 과세",   "is_header": false,   "col": 0,   "colspan": 1,   "row": 2,   "rowspan": 1 }, {   "value": "거주자의 전세계 소득에 대해 거주지국에서 과세관할권 보유",   "is_header": false,   "col": 1,   "colspan": 1,   "row": 2,   "rowspan": 1 }, {   "value": "외국납부세액공제",   "is_header": false,   "col": 2,   "colspan": 1,   "row": 2, </pre>	
--	--	---	--

		<pre> "rowspan": 1     }   ] } </pre>	
출력	설명 문장	"output": "국제조세 과세원칙의 개념을 살펴보면 원천지국 과세는 소득 원천 국가에서 과세관할권을 보유하기 때문에 국외소득면제를 조정해야 한다."	문자열
평가	ROUGE 1, ROUGE L, BLEU		

## ▷ 자료 형식

데이터 세트는 JSONL 형식으로 제공되며, 입력은 JSON으로 작성된 표 및 하이라이트 셀이고, 출력은 이를 설명하는 문장 5개이다. <표 45>는 데이터 규모를 보여준다. 훈련, 검증, 시험 데이터는 무작위로 분할되었다. 아래 표는 데이터의 예시이다.

<표 69> 표의 일부분에 대한 해석 생성 과제 데이터 규모

구분	훈련	검증	시험
문장 수	7170	896	896

주어진 훈련 데이터와 시험 데이터는 동일한 JSONL 형식으로 제공되며, 시험 데이터의 경우에는 각 문장에 대한 출력(output) 항목이 빈 목록으로 제공된다. 참가자(팀)은 해당 목록에 대해 모델의 출력 결과를 추가하여 제출한다. 훈련 데이터에서는 “출력(output)”에 5개의 문장을 목록으로 제공하지만, 제출용 데이터에서는 “출력(output)”에 생성된 문장 1개를 문자열로 제출한다.

<표 70> 표의 일부분에 대한 해석 생성 과제 데이터 형식의 예

항목	내용
훈련용 데이터의 예	<pre> {   "id": "nikluge-gtps-2023-train-000003",   "input": {     "metadata": {       "title": "4차 산업혁명에 따른 조세환경 변화와 정책 과제",       "table_title": "국제조세 과세원칙 일반 개념",       "date": "2020-06-09",       "publisher": "국회예산정책처",       "url": "https://www.nabo.go.kr/Sub/01Report/01_01_Board.jsp",       "highlighted_cells": [         [           0,           1         ],         [           1,           1         ],         [           2,           1         ]       ]     }   } } </pre>

항목	내용
	<pre> ] } }, "table": [ { "value": "과세원칙", "is_header": true, "col": 0, "colspan": 1, "row": 0, "rowspan": 1 }, { "value": "특징", "is_header": true, "col": 1, "colspan": 1, "row": 0, "rowspan": 1 }, { "value": "이중과세 조정", "is_header": true, "col": 2, "colspan": 1, "row": 0, "rowspan": 1 }, { "value": "원천지국 과세", "is_header": false, "col": 0, "colspan": 1, "row": 1, "rowspan": 1 }, { "value": "소득이 발생한 국가(원천지국)에서 과세관할권 보유", "is_header": false, "col": 1, "colspan": 1, "row": 1, "rowspan": 1 }, { "value": "국외소득면제", "is_header": false, "col": 2, "colspan": 1, "row": 1, "rowspan": 1 }, { "value": "거주지국 과세", "is_header": false, "col": 0, "colspan": 1, "row": 2, "rowspan": 1 }, { "value": "거주자의 전세계 소득에 대해 거주지국에서 과세관할권 보유", "is_header": false, "col": 1, "colspan": 1, "row": 2, "rowspan": 1 } }, } </pre>

항목	내용
	<pre> "value": "외국납부세액공제", "is_header": false, "col": 2, "colspan": 1, "row": 2, "rowspan": 1 } ] }, "output": [ "국제조세 과세원칙의 개념을 살펴보면 원천지국 과세는 소득 원천 국 가에서 과세관할권을 보유하기 때문에 국외소득면제를 조정해야 한다.", "원천지국 과세가 소득 원천 국가에서 과세관할권을 보유하기 때문에 국외소득면제를 조정해야 한다는 근거는 국제조세 과세원칙의 개념에서 발 견할 수 있다.", "국제조세 과세원칙의 개념에 따르면 원천지국 과세는 국외소득면제를 조정해야 하는데, 이는 소득 원천 국가에서 과세관할권을 보유하기 때문이 다.", "원천지국 과세는 소득 원천 국가에서 과세관할권을 보유하기 때문에 국외소득면제를 조정해야 함을 국제조세 과세원칙의 개념을 통해 확인할 수 있다.", "국외소득면제를 조정해야 하는 이유는 국제조세 과세원칙의 개념에서 원천지국 과세가 소득 원천 국가에서 과세관할권을 보유하기 때문이다." ] } </pre> <p>- 아이디(id)와 입력 표(input), 그리고 출력(5개의 설명 문장)으로 구성</p>
평가용 데이터의 예 (제출 전)	<pre> {   "id": "nikluge-gtps-2023-train-000003",   "input": {     "metadata": {       "title": "4차 산업혁명에 따른 조세환경 변화와 정책 과제",       "table_title": "국제조세 과세원칙 일반 개념",       "date": "2020-06-09",       "publisher": "국회예산정책처",       "url": "https://www.nabo.go.kr/Sub/01Report/01_01_Board.jsp",       "highlighted_cells": [         [           0,           1         ],         [           1,           1         ],         [           2,           1         ]       ]     },     "table": [       {         "value": "과세원칙",         "is_header": true,         "col": 0,         "colspan": 1,         "row": 0,         "rowspan": 1       },       {         "value": "특징",         "is_header": true,         "col": 1,         "colspan": 1,         "row": 0,         "rowspan": 1       },       {         "value": "이중과세 조정", </pre>

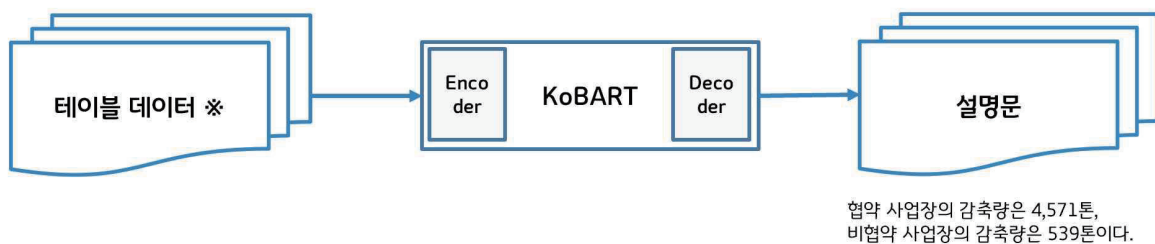
항목	내용
	<pre> "is_header": true, "col": 2, "colspan": 1, "row": 0, "rowspan": 1 }, {   "value": "원천지국 과세",   "is_header": false,   "col": 0,   "colspan": 1,   "row": 1,   "rowspan": 1 }, {   "value": "소득이 발생한 국가(원천지국)에서 과세관할권 보유",   "is_header": false,   "col": 1,   "colspan": 1,   "row": 1,   "rowspan": 1 }, {   "value": "국외소득면제",   "is_header": false,   "col": 2,   "colspan": 1,   "row": 1,   "rowspan": 1 }, {   "value": "거주지국 과세",   "is_header": false,   "col": 0,   "colspan": 1,   "row": 2,   "rowspan": 1 }, {   "value": "거주자의 전세계 소득에 대해 거주지국에서 과세관할권 보유",   "is_header": false,   "col": 1,   "colspan": 1,   "row": 2,   "rowspan": 1 }, {   "value": "외국납부세액공제",   "is_header": false,   "col": 2,   "colspan": 1,   "row": 2,   "rowspan": 1 } } ] } </pre>
	<ul style="list-style-type: none"> <li>- 학습용 데이터와 동일한 형태</li> <li>- “출력(output)” 키와 값을 제거한 데이터</li> </ul>
제출 데이터	<pre> {   "id": "nikluge-gtps-2023-train-000003",   "input": {     "metadata": {       "title": "4차 산업혁명에 따른 조세환경 변화와 정책 과제",       "table_title": "국제조세 과세원칙 일반 개념",       "date": "2020-06-09",       "publisher": "국회예산정책처",       "url": </pre>

항목	내용
	<pre> "https://www.nabo.go.kr/Sub/01Report/01_01_Board.jsp", "highlighted_cells": [   [     0,     1   ],   [     1,     1   ],   [     2,     1   ] ] }, "table": [   {     "value": "과세원칙",     "is_header": true,     "col": 0,     "colspan": 1,     "row": 0,     "rowspan": 1   },   {     "value": "특징",     "is_header": true,     "col": 1,     "colspan": 1,     "row": 0,     "rowspan": 1   },   {     "value": "이중과세 조정",     "is_header": true,     "col": 2,     "colspan": 1,     "row": 0,     "rowspan": 1   },   {     "value": "원천지국 과세",     "is_header": false,     "col": 0,     "colspan": 1,     "row": 1,     "rowspan": 1   },   {     "value": "소득이 발생한 국가(원천지국)에서 과세관할권 보유",     "is_header": false,     "col": 1,     "colspan": 1,     "row": 1,     "rowspan": 1   },   {     "value": "국외소득면제",     "is_header": false,     "col": 2,     "colspan": 1,     "row": 1,     "rowspan": 1   },   {     "value": "거주지국 과세",     "is_header": false,     "col": 0,     "colspan": 1, </pre>

항목	내용
	<pre> "row": 2, "rowspan": 1 }, { "value": "거주자의 전세계 소득에 대해 거주지국에서 과세관할권 보유", "is_header": false, "col": 1, "colspan": 1, "row": 2, "rowspan": 1 }, { "value": "외국납부세액공제", "is_header": false, "col": 2, "colspan": 1, "row": 2, "rowspan": 1 } ] }, "output": "국제조세 과세원칙의 개념을 살펴보면 원천지국 과세는 소득 원천 국가에서 과세관할권을 보유하기 때문에 국외소득면제를 조정해야 한 다." } </pre> <p>- 평가용 데이터에 “출력(output)”과 생성된 문장을 제공한다.</p>

#### ▷ 기준 모델(베이스라인 모델)

이 대회 기준 모델은 깃허브(github)<sup>5)</sup>를 통해 공개되어 있다. 해당 모델은 KoBART 모델을 사용하여 학습되었으며, JSON 형태의 표 데이터를 문자열로 변환하는 전처리 모듈을 포함하고 있다. 테이블 데이터는 JSON 데이터를 텍스트 형태의 입력으로 변환하였다.



5) [https://github.com/teddysum/korean\\_T2T\\_2023](https://github.com/teddysum/korean_T2T_2023)

<표 71> 표의 일부분에 대한 해석 생성 과제 데이터 형식 변환 예시

```
"table": [
  "구 분[TAB]협약 사업장(톤, %)[TAB]비협약 사업장(톤, %)[NL]'19.12[TAB]...58[TAB]58[TAB]0",
  "구 분[TAB]협약 사업장(톤, %)[TAB]비협약 사업장(톤, %)[NL]'19.12[TAB]...58[TAB]58[TAB]0",
  "구 분[TAB]협약 사업장(톤, %)[TAB]비협약 사업장(톤, %)[NL]'19.12[TAB]...58[TAB]58[TAB]0",
]
```

### ○ 4.2.3. 문자가 포함된 이미지 기반 문장 생성

#### ▷ 과제 개요

‘문자가 포함된 이미지 기반 문장 생성’은 자료로부터 텍스트를 생성하는 과제이다. 즉, 주어진 그림이나 사진을 설명하는 문장을 생성하는 과제이다. 이 과제는 이미지 캡셔닝(Image Captioning)으로도 알려져 있으며, 의료 분야의 이미지 캡셔닝이나 산업계 내 시각 정보를 사용한 품질 관리, 교통 관리 등을 수행하는 인공지능 챗봇 등 다양한 분야에 접목될 수 있다.

<표 72> 이미지 캡셔닝 과업의 예시

항목	내용
그림(사진)	 <p>[그림 8] 이미지 캡셔닝 예시</p>



캡션	붉은 벽돌 벽 앞에 비상 버튼과 안내 버튼이 있는 서울교통공사의 비상전화가 설치되어 있다.
----	--

## ▷ 과제 정의

‘문자가 포함된 이미지 기반 문장 생성’ 과제는 국립국어원에서 ‘2022년 유사 문장 생성 말뭉치 연구 및 구축’ 사업을 통해 구축한 데이터 세트 내 그림(사진) 기반 문장 생성 결과물을 활용하여 개발되었다. 이 과제는 그림(사진)이 주어졌을 때 이를 설명하는 문장 1개를 생성하는 것으로 정의할 수 있다. 학습용 데이터 세트는 하나의 그림(사진) 당 이를 설명하는 정답 문장 5개로 구성되어 있으며, 해당 정답 문장 5개에 대한 ROUGE-1, ROUGE-L, BLEU 점수의 평균 점수를 평가 점수로 사용한다.

<표 73> 문자가 포함된 이미지 기반 문장 생성 과제 모델 출력의 예

분류	내용	예시	비고
입력	그림(사진)	<pre> "input": {   "id": "P10974",   "image_width": 4032,   "image_height": 3024,   "ocr_info": [     {       "words": "비상버튼",       "type": "rect",       "bbox": {         "x": 1026,         "y": 1700,         "width": 205,         "height": 135       }     },     {       "words": "안내버튼",       "type": "rect",       "bbox": {         "x": 1259,         "y": 1705,         "width": 278,         "height": 122       }     },     {       "words": "비상전화",       "type": "rect",       "bbox": {         "x": 960,         "y": 1129,         "width": 608,         "height": 216       }     }   ] } </pre>	JSON

		<pre> {   "words": "서울교통공사",   "type": "rect",   "bbox": {     "x": 1172,     "y": 3680,     "width": 463,     "height": 123   } } </pre>	
출력	설명 문장	"붉은 벽돌 벽 앞에 비상버튼과 안내버튼이 있는 서울교통공사의 비상전화가 설치되어 있다."	문자열
평가		ROUGE 1, ROUGE L, BLEU	

## ▷ 자료 형식

데이터 세트는 JSONL 형식으로 제공되며, 입력(그림/사진 파일)에 대한 설명 문장 5개가 목록으로 제공된다. 입력에 제공된 그림 파일명에 대응하는 실제 그림(사진) 파일도 함께 포함되어 있다. <표 50>은 데이터의 규모이다. 훈련, 검증, 시험 데이터는 출력(output)값의 분포가 유사하도록 무작위로 분할되었다.

<표 74> 문자가 포함된 이미지 기반 문장 생성 과제 데이터 규모

구분	훈련	검증	시험
문장 수	7,334	917	917

<표 51>은 데이터의 예시이다. 주어진 훈련 데이터와 시험 데이터는 동일한 JSONL 형식으로 제공되며, 시험 데이터의 경우 각 문장에 대한 출력(output) 항목이 빈 목록으로 제공된다. 참가자(팀)은 해당 목록에 대해 모델의 출력 결과를 추가하여 제출한다. 훈련 데이터에서는 “출력(output)”에 5개의 문장을 목록으로 제공하지만, 제출용 데이터에서는 “출력(output)”에 생성된 문장 1개를 제출한다.

<표 75> 문자가 포함된 이미지 기반 문장 생성 과제 데이터 형식의 예

항목	내용
훈련용 데이터의 예	<pre> {   "id": "nikluge-2022-image-dev-000110",   "input": {     "id": "P10974",     "image_width": 4032,     "image_height": 3024,     "ocr_info": [       {         "words": "비상버튼",         "type": "rect",         "bbox": {           "x": 1026,           "y": 1700,           "width": 205, </pre>

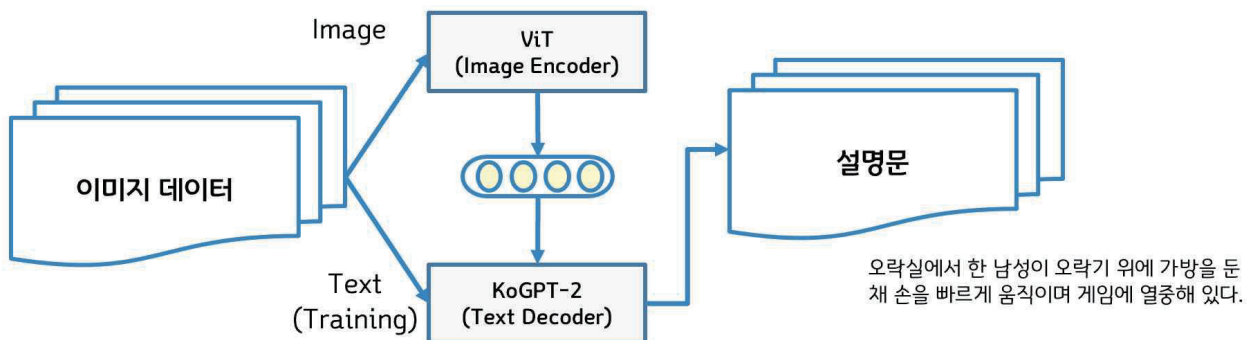
항목	내용
	<pre> "height": 135     }   },   {     "words": "안내버튼",     "type": "rect",     "bbox": {       "x": 1259,       "y": 1705,       "width": 278,       "height": 122     }   } }, {   "words": "비상전화",   "type": "rect",   "bbox": {     "x": 960,     "y": 1129,     "width": 608,     "height": 216   } }, {   "words": "서울교통공사",   "type": "rect",   "bbox": {     "x": 1172,     "y": 3680,     "width": 463,     "height": 123   } } ] }, "output": [   "붉은 벽돌 벽 앞에 비상버튼과 안내버튼이 있는 서울교통공사의 비상전화가 설치되어 있다.",   "비상버튼과 안내버튼이 있는 서울교통공사의 비상전화가 붉은 벽돌 벽 앞에 설치되어 있다.",   "붉은 벽돌 벽 앞에 설치되어 있는 서울교통공사의 비상전화에 비상버튼과 안내버튼이 있다.",   "앞에 비상버튼과 안내버튼이 있는 서울교통공사의 비상전화가 설치되어 있는 벽은 붉은 벽돌로 되어 있다.",   "비상버튼과 안내버튼이 있는 서울교통공사의 비상전화가 설치된 곳은 붉은 벽돌 벽 앞이다." ] } </pre> <p>- 아이디(id)와 입력(input), 그리고 출력(5개의 설명 문장)으로 구성</p>
평가용 데이터의 예 (제출 전)	<pre> {   "id": "nikluge-2022-image-dev-000110",   "input": {     "id": "P10974",     "image_width": 4032,     "image_height": 3024,     "ocr_info": [       {         "words": "비상버튼",         "type": "rect",         "bbox": {           "x": 1026,           "y": 1700,           "width": 205,           "height": 135         }       }     ]   },   {     "words": "안내버튼",     "type": "rect",     "bbox": { </pre>

항목	내용
	<pre> "x": 1259, "y": 1705, "width": 278, "height": 122 } }, { "words": "비상전화", "type": "rect", "bbox": { "x": 960, "y": 1129, "width": 608, "height": 216 } }, { "words": "서울교통공사", "type": "rect", "bbox": { "x": 1172, "y": 3680, "width": 463, "height": 123 } } ] } } </pre>
	<ul style="list-style-type: none"> <li>- 학습용 데이터와 동일한 형태</li> <li>- "output" 키와 값을 제거한 데이터</li> </ul>
제출 데이터	<pre> {   "id": "nikluge-2022-image-dev-000110",   "input": {     "id": "P10974",     "image_width": 4032,     "image_height": 3024,     "ocr_info": [       {         "words": "비상버튼",         "type": "rect",         "bbox": {           "x": 1026,           "y": 1700,           "width": 205,           "height": 135         }       },       {         "words": "안내버튼",         "type": "rect",         "bbox": {           "x": 1259,           "y": 1705,           "width": 278,           "height": 122         }       },       {         "words": "비상전화",         "type": "rect",         "bbox": {           "x": 960,           "y": 1129,           "width": 608,           "height": 216         }       }     ]   } } </pre>

항목	내용
	<pre> "words": "서울교통공사", "type": "rect", "bbox": {   "x": 1172,   "y": 3680,   "width": 463,   "height": 123 } } ], "output": "붉은 벽돌 벽 앞에 비상버튼과 안내버튼이 있는 서울교통공사 의 비상전화가 설치되어 있다." } </pre> <p>- 평가용 데이터에 “output”과 생성된 문장을 제공한다.</p>

#### ▷ 기준 모델(베이스라인 모델)

이 대회 기준 모델은 깃허브(github)<sup>6)</sup>를 통해 공개되어 있다. 해당 모델은 이미지 인코더 ViT 모델<sup>7)</sup>과 한국어 텍스트 디코더 KoGPT-2<sup>8)</sup>를 사용하여 입력 이미지 인코더의 결과에 대해 한국어 텍스트를 생성할 수 있도록 설계되었다.



[그림 9] 문자가 포함된 이미지 기반 문장 생성 과제 기준 모델 개념도

6) [https://github.com/teddysum/korean\\_IC\\_2023](https://github.com/teddysum/korean_IC_2023)

7) [google/vit-base-patch16-224-in21k](https://github.com/google/vit-base-patch16-224-in21k)

8) [skt/kogpt2-base-v2](https://github.com/skt/kogpt2-base-v2)

#### ○ 4.2.4. 부적절성 문장에 대한 태도 탐지

##### ▷ 과제 개요

부적절성 문장에 대한 태도 탐지 과제는 부적절하게 표현된 문장 표현의 문맥상 긍정적 또는 부정적 태도를 판단하는 작업이다. 자연어 처리에서는 상식 추론(Commonsense reasoning), 자연어 추론(NLI), 위노그라드 스키마 챌린지(Winograd schema challenge) 등에서 텍스트의 숨겨진 의미를 파악하는 능력을 검증한 사례가 있었으나, 부적절 표현을 대상으로 한 과제는 이전에 존재하지 않았다. 따라서 이 과제는 학술적으로도 의미가 크며, 기계의 추론 능력 확장에 기여하는 바 역시 클 것으로 예상된다. 부적절 표현 맥락에서의 태도 판단은 향후 온라인 플랫폼의 콘텐츠를 감시하거나 댓글 필터링 등 다양한 분야에서 응용될 수 있다는 점에서 의의가 있다.

<표 76> 부적절성 문장에 대한 태도 탐지의 예시

항목	내용
입력 문장	"쥐엔장 ~ 믿고있었다구~"
분류	"POSITIVE"

##### ▷ 과제 정의

부적절성 문장에 대한 태도 탐지 과제는 문장 내의 부적절한 표현이 주어졌을 때, 해당 표현의 문맥상 긍정적 태도 또는 부정적 태도를 탐지하는 것을 목표로 한다. 즉 과제의 주요 목표는 문장의 전체적인 맥락을 고려하여 부적절한 표현에서 드러난 태도를 올바르게 판단하는 것이며, 입력 문장의 부적절성은 문맥에 명시적으로 표현된 경우와 함의된 경우를 모두 포함한다. 학습용 데이터 세트는 부적절한 표현이 포함된 문장과 해당 문장의 '문맥(context)' 라벨로 구성된다.

이 과제의 성능 기준은 제공된 부적절한 표현의 문맥상 긍정성 또는 부정성을 정확히 판단하는 것으로, 평가는 F1-score(micro, macro)를 기준으로 이루어진다. 최종적인 성능 평가는 모델이 예측한 라벨과 실제 라벨 간의 일치 정도를 바탕으로 전체 결과에 대한 Micro F1-score와 Macro F1-score의 평균을 계산하여 진행된다. 따라서 평가 데이터의 각 입력 문장에 대해 ‘부정적(NEGATIVE)’ 혹은 ‘긍정적(POSITIVE)’ 라벨로 분류하는 것을 과제로 정의하며, F1-score를 평가 점수로 제공한다. 평가는 정답 데이터 세트와 예측 데이터 세트의 주석(annotation)을 문장 단위로 비교하여 F1-score 점수로 측정한다.

<표 77> 부적절성 문장에 대한 태도 탐지 모델 입력과 출력의 예

분류	내용	예시	비고
입력	문장	“마간호사 존나멋있고 존나웃겨” "진짜 존나 무기력하다 큰일남"	문자열
출력	분류 결과	“NEGATIVE”: 문맥상 부정적 문장, “POSITIVE”: 문맥상 긍정적 문장	문자열
평가	Micro F1-score, Macro F1-score		

데이터 세트 구축 과정에서 개인정보는 비식별화하였다. 이름, 출신/소속, 번호, 온라인 계정, 주소, 상호명, 상표명은 비윤리적 표현의 대상 여부의 관계없이 모두 비식별화되어 있으며 그 외 장소 이름, 창작물 이름 등은 비윤리적 표현의 대상일 경우만 비식별화되어 있다. 데이터 세트에서 비식별화 태그와 항목은 아래와 같다.

<표 78> 부적절성 문장에 대한 태도 탐지 비식별화 태그

분류		태그	항목
이름		&name&	실명, 특수 애칭, 별명, 대화명, 필명 가수 그룹명도 포함
출신 소속	출신 학교, 지역	&affiliation&	출신 학교, 지역 (출신 지역이 아닌 지역명은 장소로 주석)
	온라인 커뮤니티		
	정당		변형된 형태이지만 맥락에서 어떤 정당인지 유추가 가능하면 주석
	팬클럽		
	기타		
번호	고유 식별 번호	&social-security-num&	주민등록번호
	전화번호	&tel-num&	
	카드번호	&card-num&	
	계좌번호	&bank-account&	
	기타 번호	&num&	일련번호, (구매자) 식별 번호, 사업자 등록 번호, 비밀 번호
온라인 계정		&online-account&	아이디, 전자우편 주소
주소		&address&	상세 주소, 아파트 및 거주 건물명
상호명		&company&	기업/회사/상점 이름
장소명		&location&	나라, 도시 이름

상표명	&brand&	제품명, 상품명(브랜드명)
창작물명	&art&	소설, 영화, 드라마, 만화 등의 작품명
기타	&other&	위에서 언급하지 않은 항목

## ▷ 데이터 형식

데이터 세트는 'JSONL' 형식으로 제공되며, 각 문장이 문맥상 긍정적인지 부정적인지에 대한(NEGATIVE 또는 POSITIVE) 주석이 포함되어 있다. <표 55>는 데이터 규모를 보여준다.

<표 79> 부적절성 문장에 대한 태도 탐지 데이터 규모

구분	훈련	검증	시험
문장 수	12,990	1,624	1,624

<표 56>은 데이터 형식의 예시이다. 주어진 훈련 데이터와 시험 데이터는 동일한 JSONL 형식으로 제공되며, 시험 데이터의 경우에는 각 문장에 대한 출력(output) 항목이 빈 목록으로 제공된다. 참가팀은 해당 목록에 대해 모델의 출력 결과를 추가하여 제출한다.

<표 80> 부적절성 문장에 대한 태도 탐지 데이터 형식의 예

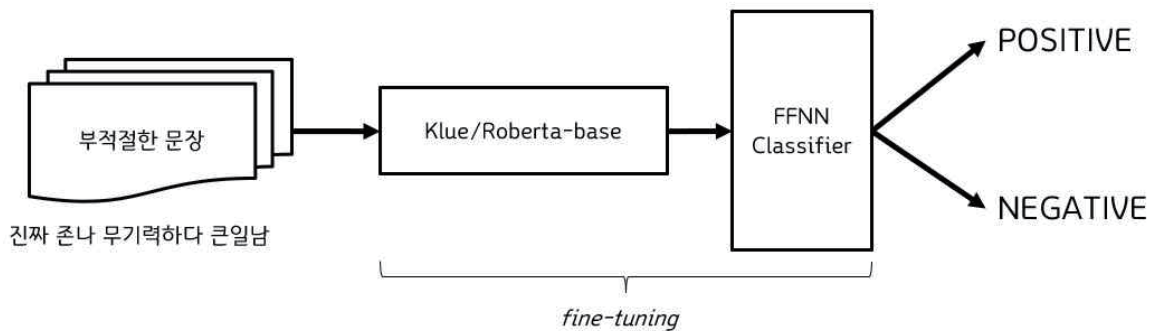
항목	내용
훈련용 데이터의 예	<pre>{   id: "nikluge-2023-iau-train-000031",   input: "이거 진짜 졸맛탱!!!",   output: "POSITIVE"} ...</pre>
	- 아이디(id)와 입력 문장(input), 그리고 출력("NEGATIVE", "POSITIVE")로 구성
평가용 데이터의 예 (제출 전)	<pre>{   id: "nikluge-2023-iau-train-000031",   input: "이거 진짜 졸맛탱!!!",   ... }</pre>
	<ul style="list-style-type: none"> <li>- 학습용 데이터와 동일한 형태</li> <li>- "출력(output)" 키와 값을 제거한 데이터</li> </ul>
제출 데이터	<pre>{   id: "nikluge-2023-iau-train-000031",   input: "이거 진짜 졸맛탱!!!",   output: "POSITIVE"} ...</pre>
	- 평가용 데이터에 "출력(output)"과 클래스를 추가하여 제출



### ▷ 기준 모델(베이스라인 모델)

이 대회 기준 모델은 ‘깃허브(github)<sup>9)</sup>’를 통해 공개되어 있다. 해당 모델은 klue/RobERTa 모델을 사용하여 학습되었으며, 모델 구조는 klue/roberta-base 모델의 <s> 토큰 출력(output)에 단순 분류기(SimpleClassifier)인 FFNN을 붙인 형태의 모델이다.

- 모델 입력 예시: <s>이거 진짜 졸맛탱!!!</s>
- 모델 출력 예시: 1 (POSITIVE)



[그림 10] 부적절한 문장에 대한 태도 탐지 기준 모델 개념도

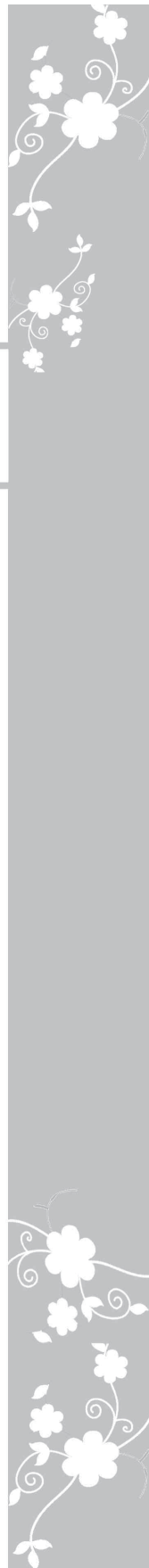
9) [https://github.com/teddysum/Korean\\_IAU\\_2023](https://github.com/teddysum/Korean_IAU_2023)





## 제 5 장

# 평가 체계 발전 방향 논의





## 5. 평가 체계 발전 방향 논의

본 과제에서는 인공지능(AI)말평 내 경진대회, 상시과제 운영과 더불어 향후 인공지능(AI) 말평의 발전을 위해 LLM 시대의 연구 동향 조사, 한국어 인공지능 연구 실태 조사 그리고 한국어 인공지능 연구자 수요 조사를 실시하여 발전 방향을 제시하였다. 본 절에서는 LLM 연구 동향과 한국어 인공지능 연구 동향 조사 내용에 대한 요약과 한국어 인공지능 연구자 수요 조사 결과를 제시한다.

### 5.1. LLM 시대 연구 동향 조사 ([부록1] 참조)

기술의 발전에 따라 해외에서는 LLM의 연구가 폭발적으로 이루어지고 있다. 특히 모델의 규모가 커지면서 연구 데이터 및 도메인, 그리고 응용 분야 등 다양한 방면에서 연구가 활발하게 진행되고 있는 추세이며, 개별 과제(task) 중심에서 실생활 시나리오 중심으로 연구가 변경되는 흐름을 보이고 있다. 또한 특기할 만한 점은 평가 체계(Benchmark) 연구가 기존 주류 언어였던 영어, 중국어를 넘어 다양한 언어에서 동시다발적으로 진행되고 있는 점이다. 주류 언어들의 평가 체계 연구는 다른 언어들에 대한 평가체계의 기준이 된다는 점에서 의의가 있다. 영어에서는 Open-llm-leaderboard, JEEBench, Hallucination Leaderboard등이 연구되었으며, 중국어에서도 C-EVAL, CLEVA등이 있다. 한국어로도 역시 평가체계 연구가 진행되고 있는데, Open-Ko-LLM LeaderBoard, Hae-rae 등이 대표적인 예시이다.

LLM 시대의 평가 체계 연구는 또한 평가 지표에 대해서도 새로운 시각으로 연구가 진행되고 있다. 2023년 국내외 LLM 평가 체계 내 모델 평가에 사용된 평가 지표의 흐름을 살펴보면 새로운 정량적 평가 방법이 다수 등장하였고, 정량적 평가에 대한 연구도 지속적으로 진행되고 있는 것으로 보인다. 또한 평가하려는 모델 및 시나리오, 평가 대상 등의 특성에 따라 정량적 평가는 물론 정성적 평가를 함께 진행하는 경향을 관찰하였다. 특히 다양성, 창의성과 같은 정성 평가가 필요한 평가의 자동화를 위한 평가 지표에 대한 연구들이 필요하며, 향후 더 활발한 연구들이 필요한 분야이기도 하다. 향후 국립국어원의 인공지능 언어능력 평가체계는 기존의 평가 체계/리더보드의 데이터 세트를 참고하여 국립국어원에서 기구축한 언어 자원을 사용해 시나리오를 정의할 필요성이 있다.

더불어 향후 국립국어원의 인공지능 언어능력 평가체계에는 기존의 자원을 발전시키는 것뿐만 아니라 새로운 평가 지표의 구축도 당연히 고려되어야 한다. 이에 본 보고서에서는 평가 시 생각의 사슬(Chain of Thoughts, CoT)을 사용하거나 논리적 사고, 배경지식, 문제 해결 등 다양한 능력에 대해 세분화된 평가를 제안하는 등의 사례를

분석하였으며, LLM 평가 시 정량적 지표와 정성적 지표 간의 관계, 인간 평가 필요성, 그리고 효율성 측면에서의 새로운 평가 지표 필요성을 제안한다. 효율성의 경우 LLM의 크기가 커질수록 늘어나는 탄소 배출량의 증가에 대한 사회적인 우려와, 효율적으로 GPU를 사용하는 것에 대한 논의가 해외에서는 이미 활발하게 이루어지고 있다는 점에서 특히 논의가 필요하다.

## 5.2. 한국어 인공지능 연구 실태 조사 ([부록 2] 참조)

한국어 인공지능 연구 실태는 지난 2022년, 2023년 HCLT에서 발표된 연구 성과물들을 대상으로 진행하였다. 성과물들의 제목과 요약(abstract)을 사용하였으며 LLM이 등장한 전(2022년), 후(2023년) 연도들에 이루어진 연구를 조사함으로써 LLM이 한국어 인공지능 연구에 어떤 영향을 미쳤는지를 살펴보고자 하였다. 이를 위해 토픽 모델링과 키워드 빈도 분석, 그리고 평가 지표 분석을 수행하였다.

토픽 모델링은 문서 집합에서 주제를 찾아내는 통계적 모델링 방법으로, 본 보고서에서 gensim과 tomotopy 라이브러리를 사용했다. 이를 통해 2022년과 2023년의 인공지능 연구 논문에서 토픽을 분석했으며, 주요 발견은 ChatGPT의 등장 이후 연구 트렌드에 상당한 변화가 있었다는 점이다. 2023년에는 특히 초거대 언어 모델 관련 연구가 증가했으며, 이는 인공지능 분야에서 새로운 연구 방향의 출현을 나타낸다.

키워드 빈도 분석에서는 2022년 논문은 '기계학습', '예측 모델', '데이터 분석'과 같은 개별적 과제에 초점을 맞춘 키워드가 주를 이루었다. 반면, 2023년에는 '대규모 언어 모델', '자연어 처리', '딥러닝' 등의 키워드가 강조되었는데, 이는 연구의 초점이 개별 과제 해결에서 언어 모델의 크기 증대와 복합적 활용으로 옮겨가고 있음을 나타낸다. 이러한 경향은 인공지능 연구의 새로운 방향성을 보여주는 중요한 지표이다.

평가 지표 분석에 따르면, 2022년에는 전통적인 성능 지표들이 사용되었다. 이에 반해, 2023년에는 언어 모델의 성능을 더 깊이 분석하는 지표들이 새롭게 사용되었으며, 생성된 텍스트의 다양성과 창의성을 평가하는 데 중요한 역할을 하는 지표들도 등장했다. 2022년 대비 2023년의 평가 방법에서는 언어 모델의 성능을 평가하는 새로운 지표들이 두드러졌으며, 이는 생성 모델과 관련된 연구가 증가하고 있으며 이러한 모델들의 성능을 보다 정밀하게 평가하려는 연구자들의 노력을 반영한다. 이는 연구의 방향성과 평가 방법의 다변화를 나타내며, 연구자들이 보다 구체적이고 세밀한 평가 기준을 모색하고 있음을 보여준다.

위 사실을 통해 실제로 초거대 언어 모델 전후로 연구 동향이 변화한 것을 확인할 수 있었다. 또한 언어 모델 크기 증가와 성능 향상에 따라 개별 과제를 통한 성능 측정정보다는 종합적인 과제 수행 결과를 통해 성능을 측정하고자 하는 시도들이 증가하고 있음을 확인하였다. 평가 지표 측면에서는 특히 생성 과제 측면에서 전통적인 지표

들 외에 신규 평가 지표들이 등장한 것을 확인함으로써 생성 결과에 대한 성능을 정확하게 측정하고자 하는 시도들이 증가하고 있음을 확인하였다. 또한 언어 모델 크기 증가 및 성능의 비약적인 향상에 따라 한국어 역시 데이터의 양과 질을 증대시키고, 한국어의 특성에 맞는 평가 지표 개발, 연구 생태계 조성이 필요함을 확인하였다. 이러한 노력은 데이터 부족 문제를 해결하고, 대규모 언어 모델의 성능을 향상시키는 데 기여할 것이다. 또한, 한국어 언어 모델의 성능을 보다 정확하게 평가하고, 연구 생태계를 통해 한국어 언어 모델 관련 연구를 활성화시킬 수 있을 것으로 기대된다.

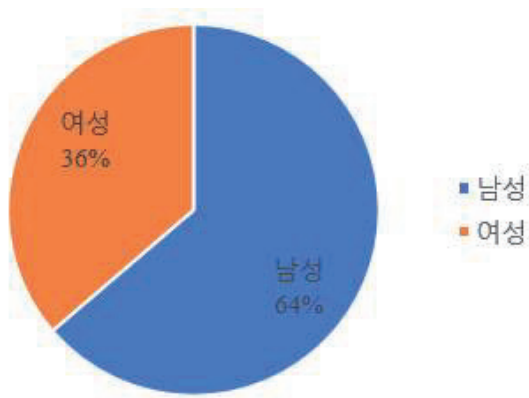
### 5.3. 평가 체계 관련 설문 조사

본 과제에서는 인공 지능 언어 능력 평가 체계에 대한 학계, 산업계 등의 의견과 동향을 조사하고자 ‘평가 체계 설문’을 진행하였다. 설문은 ‘2023 HCLT’ 등록자들을 대상으로 진행했는데, 이는 학술대회 특성상 자연어 처리 관련 학과, 연구기관 혹은 업계 종사자 등 다양한 사람들이 가장 많이 모일 수 있는 기회이기 때문이다. 이에 따라 실제 설문 참여 대상자는 크게 학교, 기업, 공공기관에 소속된 사람으로 조사되었다. 학교 안에는 대학생, 대학원생, 소속 교원, 그리고 교내 연구소 연구원들이 포함되며, 기업에는 회사원, 임원, 연구원, 공공기관은 공공기관 종사자 및 정부 출연 연구소 연구원 등이 포함되었다.

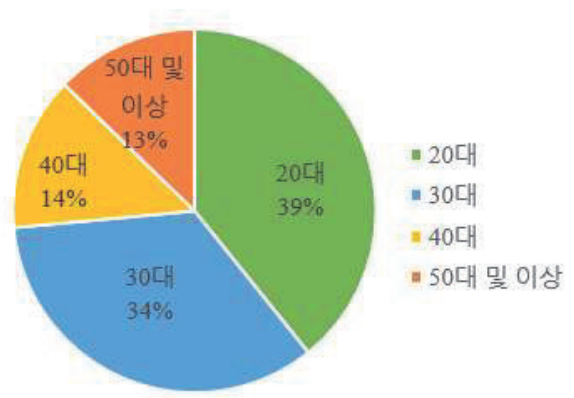
설문 내용으로는 크게 참여자 인적 사항, 벤치마크 사용 경험, 벤치마크 설계, 벤치마크 과제 구성 인식, 리더보드 및 평가 지표 인식을 다루었으며, 약 60개의 세부 문항으로 구성되어 있다.

#### 1. 참여자 인적 사항

참여자 인적사항으로는 성별, 연령대, 소속, 최종 학위, 전공 계열, 연구/업무 경력을 조사하였다. 조사 결과 가장 대다수의 응답자는 2-30대의 학교 소속으로 전공은 공학 계열이 가장 많았다. 또한 연구, 업무 경험은 6년 이하가 가장 많았는데, 이는 학술대회 주요 참여층이 대학원생 혹은 저년차 직장인임을 나타낸다. 성별 측면에서는 여성 37명, 남성 65명으로 남성이 여성보다 1.8배정도 많았다. 연령별로는 20대 (39%)가 가장 많이 응답하였고, 30대 (34%), 40대(14%), 그리고 50대(13%) 순으로 응답하였다.

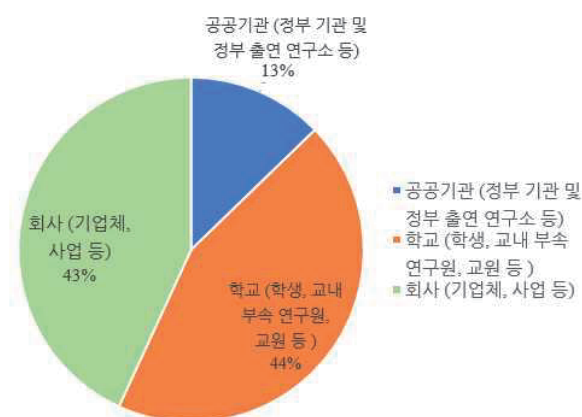


[그림 11] 성별 응답

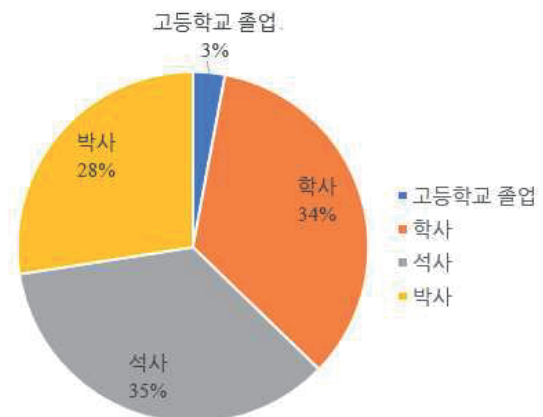


[그림 12] 연령대 응답

연구기관 외에도 산업계에서 참여가 활발한 자연어처리 학술대회 특성상 소속이 학교인 사람들(45명, 44%)과 회사를 소속으로 둔 사람들(44명, 43%)이 비슷한 숫자로 설문  
에 응답하였으며, 공공기관 소속의 경우 매우 적은 수(13%, 13명)만이 응답하였다. 마찬  
가지로 학술대회 특성상 대학원 이상의 사람들이 많이 참석하기에 최종 학력 측면에서  
학부 재학은 가장 적은 숫자(3명, 3%)로 나타났으며, 학사 졸업(35명, 34%), 석사 졸업  
(36명, 35%)이 비슷한 숫자로 나타났다. 박사 졸업은 28명(28%)로 나타났다.



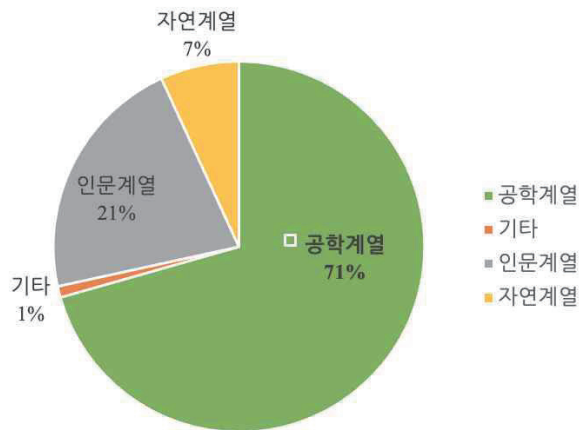
[그림 13] 직업 응답



[그림 14] 최종 학력 응답

응답자들의 전공은 공학 계열이 가장 많았고(72명, 71%), 이어서 인문 계열(22명, 21%), 자연 계열(7명, 7%) 그리고 기타(1명, 1%) 순으로 설문에 응답하였다. 연구 혹은  
업무 경력은 3년 단위로 설문을 진행하였는데, 조사 결과 3년 미만의 경력을 가진 응답  
자(34명, 33%)와 3년에서 6년 사이의 경력을 가진 응답자(32명, 31%)가 비슷한 숫자로  
나타났다. 이외 응답자 분포는 10년 이상 경력자 (24명, 24%), 7-10년 사이 경력자(12  
명, 12%)로 나타났다.





[그림 15] 전공 계열 응답



[그림 16] 경력 응답

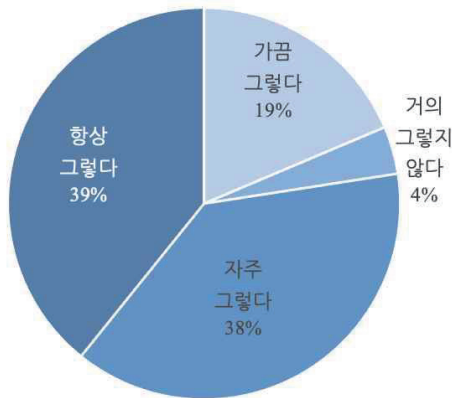
## 2. 벤치마크 사용 경험 관련 문항

벤치마크 사용 경험 항목은 세부 5개 문항으로 구성되었으며, 각 문항들은 사용자들의 벤치마크 사용 빈도, 벤치마크 부재 상황, 국내/외 벤치마크 사용 경험 및 만족 요인, 불만족 요인에 대해 다루었다.

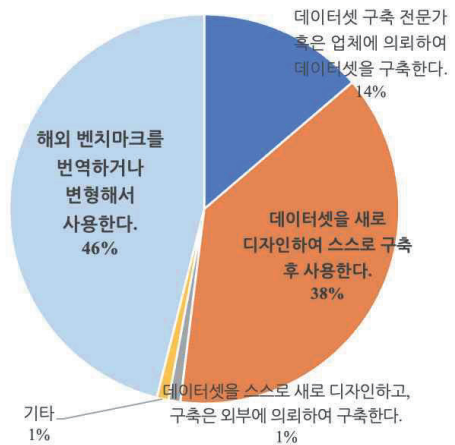
### 2.1. 벤치마크 사용 빈도 및 부재 관련 설문

사용자 벤치마크 사용 빈도 문항에서 대다수의 응답자들이 벤치마크가 있는지 항상 확인하고 실험을 진행하거나(39%) 자주 먼저 확인하고 실험을 진행하는 편(38%)이라고 답하여 자연어 처리 연구에서 벤치마크는 실험 성능의 기준이자 필수적인 요소로 자리잡았음을 확인할 수 있었다. 적합한 국내 벤치마크가 부재한 상황에서 사용자들은 주로 해외 벤치마크를 번역하거나 변형해서 실험에 사용하거나 (46%), 데이터 세트를 목적에 맞게 새롭게 만들어 사용하는 (38%) 경향을 보였다.

이외에 데이터 세트 구축 전문가 혹은 업체에 의뢰하여 데이터 세트를 새롭게 만드는 방법(14%)도 자주는 사용되고 있으나 위의 2가지 방법에 비해서는 대중적인 편은 아니다. 제일 적은 비율을 차지한 방법들에는 ‘데이터 세트를 디자인하되 외부에 구축 요청’ 등이 있었다. 이러한 설문 응답으로 미루어보았을 때 국내에는 외부에 데이터 세트 구축을 의뢰할 수 있는 주체들보다는 개인 혹은 소규모 단위를 중심으로 연구가 이루어지는 것을 확인할 수 있었으며 해외 벤치마크에 대한 의존도가 높은 것 역시 확인할 수 있었다. 또한 해외 벤치마크들의 경우 국내 연구자들의 수요를 어느 정도는 충족할 수 있는 다양성이 있음 역시 확인할 수 있었다. 특히 해외 벤치마크의 다양성에 대한 연구자들의 인식 및 사용 양상은 사용자들의 해외 벤치마크 사용 경험과 더불어 해외 벤치마크 만족 요인 설문 문항에 의해서도 방증된다.



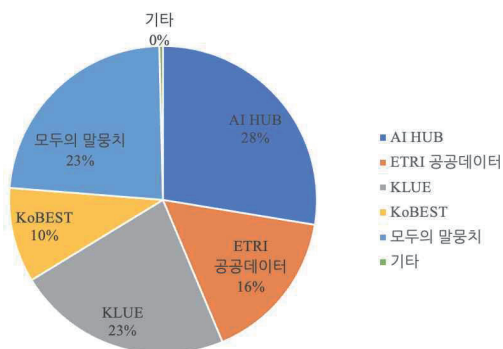
[그림 17] 벤치마크 사용 빈도



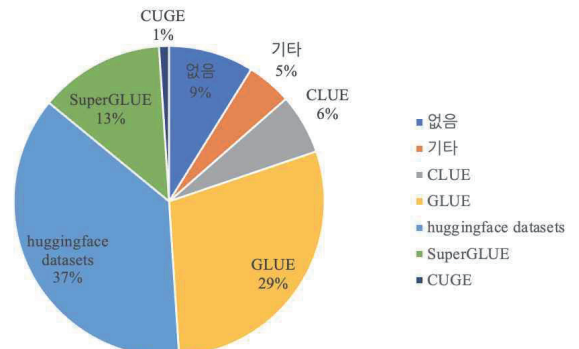
[그림 18] 벤치마크 대체 응답

## 2.2. 국내/외 벤치마크 사용 경험 관련 설문

본 설문에서는 ‘벤치마크’를 언어 모델의 언어 이해 능력을 평가할 때 사용하는 평가체계와 더불어 성능 평가를 위해 사용하는 데이터 세트’까지 개념을 확장하였다. 사용자들이 사용해본 국내 벤치마크는 AIHub(28%), 모두의 말뭉치, KLUE(23%)순으로 나타났으며, 이외에 ETRI 공공 데이터, KLUE의 심화 벤치마크인 KoBEST, 기타 순이었다. 한편 사용자들이 경험한 해외 벤치마크는 huggingface dataset(37%), GLUE(29%), 그리고 GLUE의 심화버전인 SuperGLUE(13%)순으로 나타났다. 이외에 중국 벤치마크인 CLUE(6%)와 CUGE(1%)도 나타났으며, 기타 벤치마크로는 TED, ACE2003, IWSLT, OntoNotes5, GUGE, HANS, DocRED, Financial phrasebank, CoNLL2003 등 매우 다양한 벤치마크와 데이터 세트들이 사용되었다.



[그림 19] 국내 벤치마크 사용 빈도 순위

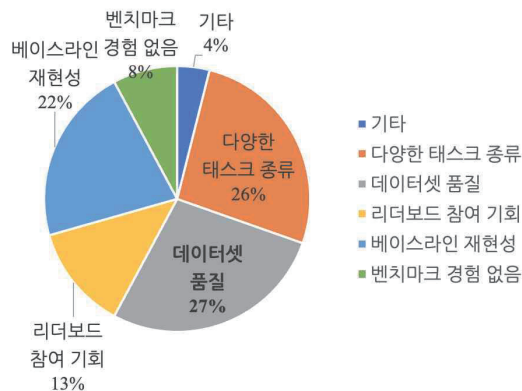


[그림 20] 국외 벤치마크 사용 빈도 순위

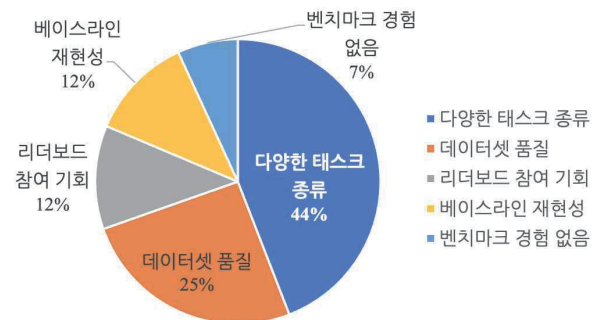
### 2.3. 국내/외 벤치마크 만족 요인 설문

이와 같이 자연어 처리 연구자들은 국내/외 다양한 벤치마크들을 사용하고 있음을 설문을 통해 확인할 수 있었다. 이에 따라 각 벤치마크별 만족 및 불만족 요인에 대해서도 설문을 진행하였으며 그 결과 국내 벤치마크와 해외 벤치마크의 상보적인 특성을 확인할 수 있었다.

국내 벤치마크의 만족 요인은 ‘데이터 세트 품질(27%)’로, 주로 한국어/한국 지식이 포함된 데이터 세트란 점에서 만족 요인이 높았다. 또한 해외 벤치마크를 번역할 때 감수해야 하는 번역 품질에 대해 구애받지 않는 점 역시 만족 요인이 된 것으로 보인다. 또한 AIHUB와 모두의 말뭉치 등 국내 데이터 세트들의 종류가 다양한 편이기에 ‘다양한 과제 종류(26%)’ 역시 만족 요인으로 꼽혔다. 한편 해외 벤치마크의 절대적인 만족 요소로는 ‘다양한 과제 종류(44%)’가 꼽혔다. 해외 벤치마크의 과제 종류는 위에서 언급된 OntoNote5, GUGE 등 데이터셋 뿐만 아니라 실질적인 벤치마크의 다양성을 모두 포함한다. 또한 각 벤치마크들은 개별 데이터셋을 파생하는 경우가 많아 해외는 국내보다 더욱 다양한 데이터 세트들이 사용될 수 있는 환경을 갖추고 있다. 이후의 만족 요인은 데이터 세트 품질(25%)로, 영어권 데이터 세트들의 경우 기존부터 자주 사용되어온 데이터 세트이기에 품질이 어느 정도 보증되어 있거나 개선되고 있는 편이고 주석자간 일치도 정보 등 말뭉치 품질을 확인할 수 있는 정보들이 있기 때문인 것으로 보인다.



[그림 21] 국내 벤치마크 만족 요인



[그림 22] 해외 벤치마크 만족 요인

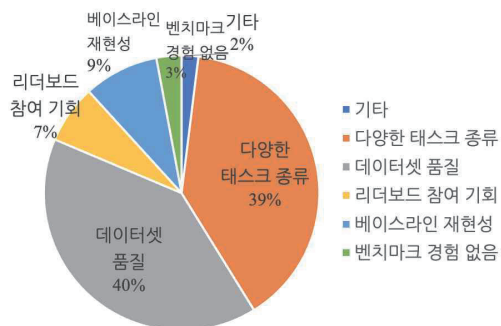
### 2.4. 국내/외 불만족 요인 설문

국내 벤치마크의 불만족 요인으로서는 ‘데이터 세트 품질(40%)’과 ‘다양한 과제 종류(39%)’가 꼽혔다. 전자의 경우 해외에 비해 말뭉치 품질 개선이나 주석자간 일치도 등 품질에 대한 정보를 접하기 어렵고, 연구자들의 선형적인 측면에 비추어보았을 때 품질이 낮다고 판단한 것으로 추측된다. 과제 종류의 경우에도 해외와 비교했을 때 벤치마크 수가 적고, 벤치마크들이 파생하는 데이터 세트 숫자가 비교적 적기 때문에

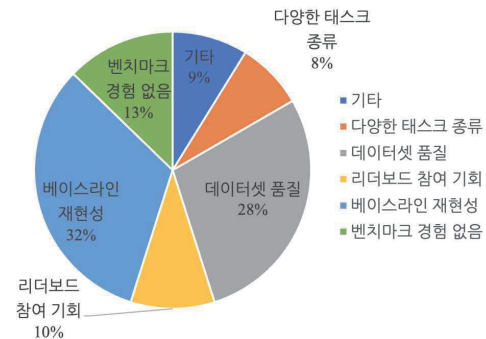
이러한 결과가 나왔음을 유추할 수 있다. 한편 국외 벤치마크의 불만족 요인으로는 ‘베이스라인 재현성(32%)’과 ‘데이터 세트 품질(28%)’이 꼽혔는데, 두 요인 모두 해외 벤치마크를 한국어로 학습된 모델에 적용하면서 나타나는 문제인 것으로 보인다. 즉 영어 벤치마크를 한국어로 학습된 모델에 적용하면서 베이스라인 재현성에 문제가 생기는 것으로 추측되며, 한국어 번역 품질 등으로 인해 데이터 세트 질 저하가 발생할 가능성 역시 높을 수밖에 없다. 국내 만족 요인 중 ‘베이스라인 재현성’은 이러한 현상과 관련된 것으로, 영어 벤치마크에 비해 한국어에 특화된 벤치마크가 보다 안정적인 베이스라인 재현이 가능함을 알 수 있다.

이처럼 국내/외 벤치마크 만족/불만족 요인은 상보적인 분포를 보이는 것을 확인할 수 있었다. 이를 정리하면 아래와 같이 정리할 수 있다.

- 국내 만족 요인: 데이터 세트 품질, 베이스라인 재현성
  - ↔ 해외 불만족 요인: 데이터 세트 품질, 베이스라인 재현성
  - ⇒ 한국어, 한국 지식 등을 반영한 데이터 세트 수요 높음
- 해외 만족 요인: 다양한 과제 종류
  - ↔ 국내 불만족 요인: 다양한 과제 종류
  - ⇒ 해외와 같이 다양한 벤치마크 및 데이터 세트 발굴 필요



[그림 23] 국내 벤치마크 불만족 요인



[그림 24] 해외 벤치마크 불만족 요인

요약하자면 피설문자들은 자연어 처리 실험 시 벤치마크를 매우 자주 활용하며, 국내에 적절한 벤치마크가 없을 경우 해외 벤치마크를 번역하거나 자급하는 경향을 보인다. 국내 벤치마크 중 가장 자주 활용되고 있는 벤치마크로 AIHUB와 모두의 말뭉치가 꼽혔으며, 해외 벤치마크의 경우 huggingface dataset과 GLUE이다. 국내 벤치마크의 만족 요인은 데이터 세트 품질과 베이스라인 재현성, 그리고 다양한 과제 종류로, AIHUB와 모두의 말뭉치의 데이터 세트의 종류가 다양한 점, 그리고 한국어 데이터 세트인 점 때문인 것으로 보인다. 해외 벤치마크의 만족 요인은 다양한 과제 종류와 데이터 세트 품질로, 항목은 같으나 국내 벤치마크 만족 요인과는 결을 조금 달

리한다. 다양한 과제의 종류는 데이터 세트는 물론 벤치마크에도 적용이 되며, 데이터 세트 품질의 경우 국내에 비해 품질 개선 이력 및 주석자간 일치도 등 데이터 세트 품질에 대한 정보가 많기 때문이다.

한편 국내 벤치마크의 불만족 요인은 다양한 과제 종류와 데이터 세트 품질로, 이는 해외 만족 요인과 상보적인 경향을 보인다. 곧 해외 벤치마크에 비해 실질적으로 벤치마크 및 데이터 세트의 종류가 부족한 점, 그리고 데이터 세트 품질에 대한 정보가 명확하지 않은 점이 단점으로 꼽혔다. 해외 벤치마크 단점으로는 데이터 세트 품질과 베이스라인 재현성이 꼽히는데, 이는 한국어로 학습된 모델에 대해 영어 벤치마크를 적용하면서 발생하는 문제 때문이다.

이에 따라 향후 한국어 벤치마크 구축 시에는 한국어, 한국 지식, 맥락 등이 반영된 다양한 벤치마크를 구축하는 것이 중요하며, 데이터 세트 품질 제고에 대한 노력이 필요하다.

### 3. 벤치마크 설계 관련 문항

벤치마크 설계와 관련된 세부 문항으로는 벤치마크 (데이터 세트) 설계 관련, 벤치마크 발전 방향 관련, 그리고 한국어 LLM 모델 설계 및 학습, LLM 벤치마크 설계 방향 관련 문항으로 총 4문항이다. 이번 문항에서는 화제에 대한 설문자의 동의 정도를 중심으로 문항을 설계하였으며, 문항에 따라 4점 리커트 척도 및 순위 매기기가 사용되었다.

#### 3.1. 벤치마크 (데이터 세트) 설계 설문

해당 설문은 설문자들에게 벤치마크 (데이터 세트) 설계 관련 4가지 진술별로 동의하는 정도에 4점 리커트 척도로 응답하도록 설계하였다. 진술 1과 2는 데이터 세트 품질, 진술 3과 4는 데이터 세트와 관련된 내/외적 문제점을 다루었다.

**진술 1) 향후 벤치마크 데이터 세트 구축 인력은 ‘클라우드 워커’가 아닌 ‘주석 전문 인력’이 되어야 한다.**

위 설문에 대해서는 약간 그렇다(54%) > 매우 그렇다(35%) > 약간 그렇지 않다(9%) > 전혀 그렇지 않다(2%) 순으로 결과가 나타났다. 이에 따라 설문자들은 벤치마크 데이터 세트에 대해 주석 전문 인력을 통해 얻은 고품질 주석을 기대하고 있음을 알 수 있다.

진술 2) 앞으로 한국어 벤치마크 데이터 세트 개발에서 ‘양’보다는 ‘질’을 더욱 중요시해야한다.

위 진술에 대해서는 매우 그렇다(59%) > 약간 그렇다 (33%) > 약간 그렇지 않다 (8%) 순으로 결과가 집계되었으며, ‘전혀 그렇지 않다’ 응답은 나타나지 않았다. 1)과 종합하였을 때, 자연어 처리 연구자들의 ‘고품질’ 데이터 세트에 대한 수요가 매우 높은 것을 알 수 있다.

진술 3) 앞으로의 한국어 벤치마크 데이터 세트 내 사회적/윤리적 편향성 문제 해결에서 ‘실제성’보다는 ‘공정성’을 중시해야 한다<sup>10)</sup>.

해당 진술에 대해서는 약간 그렇다(45%) > 약간 그렇지 않다 (29%) > 매우 그렇다 (19%) > 전혀 그렇지 않다 (7%)의 비율로 응답이 나타났다. 과반 이상의 응답자들이 ‘공정성’을 중시해야 한다고 답해 전반적으로는 데이터 세트 내 윤리적 편향성 문제를 해결하는 것에 동의하는 경향을 관찰하였다. 다만 이외의 응답 비율(36%)을 살펴보았을 때 ‘실제성’을 중시하는 경향 역시 드러나고 있어 모델이 실제 현실 세계(real-world)에 반영되어 있는 사회적/윤리적 편향성을 학습할 수 있는 데이터 세트, 혹은 학습한 결과를 평가할 수 있는 벤치마크에 대한 잠재적인 수요도 있음을 확인하였다.

진술 4) 앞으로의 벤치마크 개발 시 원천 데이터 측면에서 ‘다양성’보다는 ‘합법성’을 중시해야 한다<sup>11)</sup>.

진술 4에 대해서는 약간 그렇다 (40%) > 약간 그렇지 않다 (30%) > 매우 그렇다 (21%) > 전혀 그렇지 않다 (9%) 순으로 응답이 나타났다. 진술 4 역시 진술 3과 마찬가지로 전반적으로는 저작권 등 데이터 세트 사용 시 제약으로 작용하는 요인들이 해결된 데이터 세트에 대한 수요가 높으나, 저작권 준수 여부와 상관없이 현실 세계가 반영된 원천 데이터의 다양성 추구에도 수요가 있음을 확인할 수 있었다.

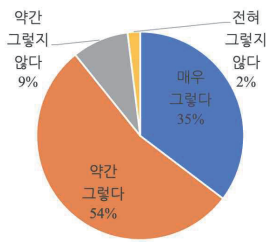
위 4가지 진술을 살펴보았을 때 자연어 처리에 있어 전문 주석자들이 주석해 데이터 품질이 높으면서도 사회적, 윤리적 편향이나 저작권 등의 문제들이 해결된 데이터 세트에 대한 수요가 높은 것으로 보인다. 한편으로는 제약 사항들을 해결하면서도 동시에 현실 세계를 반영할 수 있는 데이터 세트 역시 원하고 있음을 알 수 있었다.

---

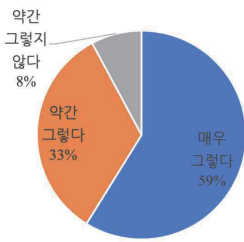
10) 실제성: 현실(Real-life) 반영을 위해 데이터 세트 내의 사회적/윤리적 편향성 문제를 있는 그대로 사용  
공정성: 데이터 세트에 나타나는 사회적/윤리적 편향성 문제를 없애야 함

11) 다양성: 원천 데이터의 다양성 추구(저작권 해결 여부와 상관없이)  
합법성: 원천 데이터의 법적 이슈 해결 여부를 중요시 (저작권 문제가 없는 데이터만 사용)

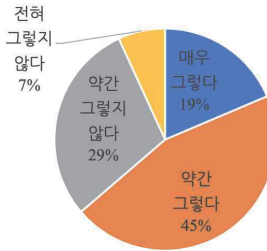




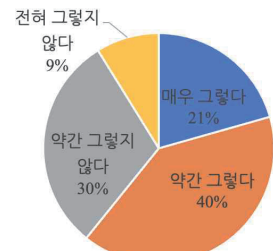
[그림 25] 진술1 응답



[그림 26] 진술2 응답



[그림 27] 진술3 응답



[그림 28] 진술4 응답

### 3.2. 벤치마크 발전 방향 관련 설문

해당 항목 역시 위와 마찬가지로 향후 벤치마크 발전 방향에 대한 4가지 진술들에 대해 응답자들이 동의하는 정도를 4점 리커트 척도로 측정하였다.

**진술 1) 앞으로의 한국어 벤치마크 발전 방향은 ‘해외 유명 벤치마크 모방’보다는 ‘한국어 특징 혹은 사회적 맥락 반영’이 되어야 한다.**

위 진술에 대해서는 매우 그렇다 (52%) > 약간 그렇다 (42%) > 약간 그렇지 않다 (6%)로, 한국어 특징 혹은 맥락을 반영한 벤치마크에 대해 매우 높은 수요가 있음을 확인할 수 있었다. 이는 앞서 설문한 국내/외 벤치마크 만족/불만족 요인에서도 확인한 바와 일치한다.

**진술 2) 앞으로의 한국어 벤치마크 발전 혹은 개발 주체는 ‘민간’이 아닌 ‘정부’가 되어야 한다.**

해당 진술은 벤치마크 개발에 있어 연구소 혹은 기업이 주도해야 할지, 아니면 정부 기관이 주도해야 할지를 묻는 진술로써, 약간 그렇지 않다 (38%) > 약간 그렇다 (29%) > 매우 그렇다 (20%) > 전혀 그렇지 않다 (13%)로 결과가 집계되었다. 이에 따라 해외 벤치마크 사례들과 같이 민간 주도의 발전에 대한 수요와 더불어 현재 AIHUB, 모두의 말뭉치 등 국가 기관이 주도하는 벤치마크에 대한 수요가 동시에 있음을 확인할 수 있었다. 전자의 경우 연구 경향에 따라 벤치마크 개발 및 운영 등이 유동적인 장점이 있으나 주체의 규모가 작을수록 개발이 어렵게 되며, 후자의 경우 예산 지원 등을 통한 개발 지원 등이 가능하나, 유연한 개발이 어려운 단점이 존재한다.

**진술 3) 앞으로의 한국어 벤치마크 발전을 위해 ‘학계’보다는 ‘업계’의 의견을 더 많이 수렴해야 한다.**

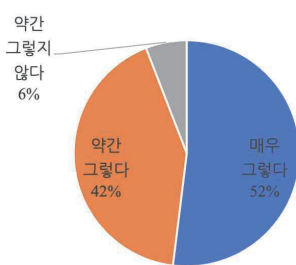
위 진술에 대해서는 약간 그렇다 (47%) > 약간 그렇지 않다 (24%) > 매우 그렇다 (21%) > 전혀 그렇지 않다 (8%)로 나타났다. 업계 의견 수렴에 대한 수요가 높은 것

을 알 수 있으며, 기존 벤치마크 과제가 아닌 시의성 있는, 현실 세계의 문제를 반영한 다양한 벤치마크에 대한 수요로 해석 가능하다.

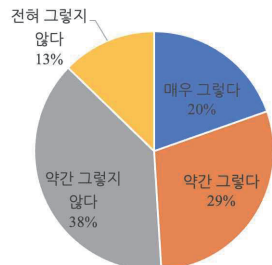
**진술 4) 앞으로의 한국어 벤치마크 개발 플랫폼의 발전 방향은 ‘세분화된, 다수의 특화 플랫폼’보다는 ‘하나의 종합적 플랫폼’으로 가야한다.**

해당 진술에 대해서는 약간 그렇다 (39%) > 매우 그렇다 (32%) > 약간 그렇지 않다 (23%) > 전혀 그렇지 않다 (6%) 순으로 결과를 확인할 수 있었다. 동의 비율이 매우 높은 것으로 보아 KLUUE와 같이 여러개의 과제 수행을 통해 인공 지능의 언어 능력을 한번에 종합적으로 판단할 수 있는 벤치마크 개발이 필요하며, 더 나아가 Papers with Code 등 해외 사례와 같이 벤치마크와 논문 및 코드 공유가 가능한 자연어 처리 커뮤니티 역시 발전 방향으로 고려해볼 수 있다.

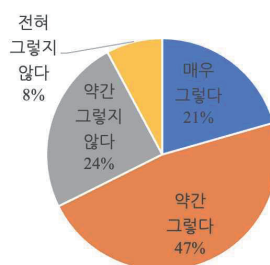
위 네 가지 진술에 대한 설문자들의 응답을 종합하여 보면, 벤치마크 발전 방향은 한국어에 특화된 벤치마크이며 시의성 있고 다양한 과제들을 포함할 수 있어야 한다. 이를 위해서는 정부보다는 민간 주도의 유연한 발전이 적절하다. 또한 벤치마크는 개별 과제를 사용하여 진행하는 특화 플랫폼보다는 하나의 종합적인 플랫폼을 지향하는 것이 바람직하다.



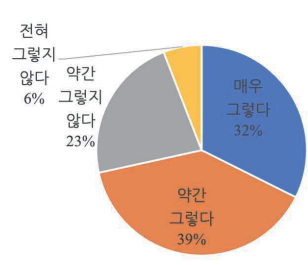
[그림 29] 진술1 응답



[그림 30] 진술2 응답



[그림 31] 진술3 응답



[그림 32] 진술4 응답

### 3.3. 한국어 LLM 설계 및 학습 관련 설문

이번 설문은 한국어 거대 언어 모델 설계 및 훈련과 관련된 진술을 사용해 진행하였으며, 이번 항목은 리커트 척도가 아닌 순위 매기기를 통해 설문을 진행하였다. 이를 통해 설문자들이 어떤 사안을 시급하게 생각하는지 확인할 수 있었다.

**진술 1) 한국어 LLM 설계 및 학습은 영어로 학습된 LLM을 토대로 이루어져야 한다.**

3위 (57%) > 2위 (26%) > 1위 (17%)



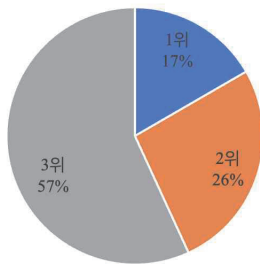
진술 2) 한국어 LLM 설계 및 학습은 한국어 데이터 세트를 직접 구축하여 밑바닥 (scratch)부터 학습해야 한다.

1위(44%) > 2위 (41%) > 3위 (15%)

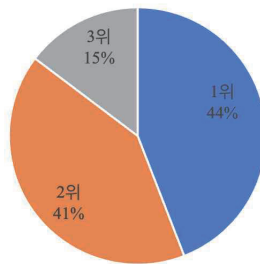
진술 3) 한국어 LLM 설계 및 학습은 기존 다국어 데이터 세트 중 한국어 텍스트 비중을 높여 강화학습을 진행해야 한다.

1위 (49%) > 2위 (40%) > 3위 (11%)

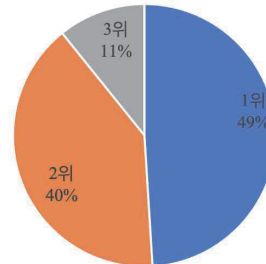
진술1의 경우 가장 선호되지 않은 진술로써, 설문 응답자들은 한국어 LLM 학습 시 영어로 이미 학습된 LLM 토대로 하는 것보다는 ‘한국어’로 학습하는 것을 선호하는 것을 알 수 있다. 진술2와 진술3은 비슷한 비율로 선호도가 나타났는데, 진술3이 근소하게 높아 현재 자연어 처리 연구자들이 가장 현실적으로 생각하는 LLM 학습 방법이 기존 다국어 데이터 세트 내 한국어 텍스트 비중을 높여 강화 학습을 진행하는 것임을 알 수 있다. 일련의 진술에 대한 선호도 응답을 통해 현재 한국어 LLM 설계 및 학습에 있어서도 한국어 특성 및 지식을 반영하는 것이 중요함 역시 확인할 수 있었다.



[그림 33] 진술1 응답



[그림 34] 진술2 응답



[그림 35] 진술3 응답

### 3.4. 한국어 LLM 벤치마크 설계 관련 설문

이번 문항은 LLM 벤치마크 설계와 관련된 진술들으로써, 3가지 진술로 구성되어 있다.

진술1) 한국어 LLM 벤치마크 설계는 영어권 LLM 벤치마크를 벤치마킹하여 설계해야 한다.

2위 (47%) > 3위 (31%) > 1위(22%)

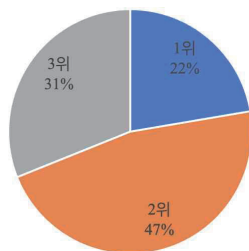
진술2) 한국어 LLM 벤치마크 데이터 세트는 클라우드 소싱을 통해 기존 한국어 데이터 세트를 변형하여 만들어야 한다.

2위(44%) > 3위 (31%) > 1위 (25%)

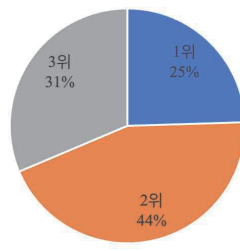
진술3) 한국어 LLM 벤치마크 데이터 세트는 주석 전문가를 통해 양질의 소규모 한국어 데이터 세트로 구축해야 한다.

1위(55%) > 2위(27%) > 3위(18%)

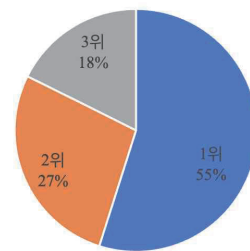
진술1과 진술2는 비슷한 비율로 응답되어, 설문자들이 영어권 LLM 벤치마킹과 클라우드 소싱에 대해 비슷하게 선호함을 알 수 있었다. 가장 선호되는 벤치마크 설계 방법은 ‘주석 전문가를 통해 양질의 소규모 한국어 데이터 세트’를 구축하는 것으로, 이 역시도 LLM 벤치마크에 한국어, 한국 지식 반영 수요가 반영된 것이라 볼 수 있다.



[그림 36] 진술1 응답



[그림 37] 진술2 응답



[그림 38] 진술3 응답

### 3.5. 벤치마크 과제 인식 관련 문항

해당 문항에서는 자연어 이해, 자연어 생성, 그리고 멀티 모달 과제에 대한 설문자들의 인식을 조사하였다. 설문은 먼저 위 세 가지 과제 중 어떤 과제에 가장 관심이 있는지부터 설문하였다. 설문 결과 가장 관심이 많은 과제는 NLG (46%), 이후에는 NLU(38%), 멀티모달(16%) 과제 순으로 나타났다. 이후 각 과제별로 세부 과제들에 대한 설문자들의 인식을 조사하였다. 과제에 대한 인식은 인간 성능에 도달한 과제, 시일 내 벤치마크로 개발해야 하는 과제, 윤리적/사회적 논란이 잠재된 과제, 미래 활용 가치가 높은 과제로 나누어 조사하였다.

#### 3.5.1. 자연어 이해 과제 인식 관련

설문에 포함된 자연어 이해 세부 과제는 언어학 기반 과제(예: 형태소 분석 등), 감정 분석, 혐오 표현 탐지, 정보/관계 추출, 어휘 중의성, 상호 참조 해결, 속성 기반 감성분석이다. 해당 세부 과제에 대해 응답자들이 1위로 선정한 과제는 아래와 같다.

1) 인간 성능(human performance)에 도달했거나, 거의 접근한 과제: 언어학 기반 과제 (36%)

- 2) 현실적인 수요가 높아 벤치마크 개발이 이루어져야 함: 정보/관계 추출(25%)
- 3) 윤리적 혹은 사회적 논란을 일으킬 우려가 있는 과제: 혐오 표현 탐지 (60%)
- 4) 현재에는 접근이 쉽지 않지만 미래 활용 가치가 높은 과제: 상호 참조 해결 (20%)

자연어 이해 과제 인식을 살펴보았을 때 가장 기초적 층위에서 이루어지는 언어학 기반 과제들은 응답자들이 모델이 사람만큼 수행이 가능하다고 판단하였다. 현실적 수요 측면에서는 정보/관계 추출 과제가 1위로 꼽혔는데, 이는 LLM 상용화에 따라 검색 기능 및 성능에 대한 기준이 높아졌기 때문이다. 마찬가지로 윤리적 혹은 사회적 논란을 일으킬 우려가 있는 과제로 꼽힌 혐오 표현 탐지와 높은 활용 가치가 있다고 간주된 상호 참조 해결 역시 LLM에 대한 활발한 연구와 상용화로 인해 선정된 것으로 보인다. 두 과제는 LLM과 사용자 간 대화 시 필요한 모델의 능력을 다룬다.

### 3.5.2. 자연어 생성 과제 인식 관련

자연어 생성 항목에서는 이야기 생성 과제, 일상 대화 시스템, 목적 지향 대화 시스템, 문서 요약, 산업 리포트 생성 과제에 대한 인식을 조사하였다. 결과는 아래와 같다.

- 1) 인간 성능(human performance)에 도달했거나, 거의 접근한 과제: 문서 요약 (48%)
- 2) 현실적인 수요가 높아 벤치마크 개발이 이루어져야 함: 일상 대화 시스템 (30%)
- 3) 윤리적 혹은 사회적 논란을 일으킬 우려가 있는 과제: 일상 대화 시스템 (37%)
- 4) 현재에는 접근이 쉽지 않지만 미래 활용 가치가 높은 과제: 산업 리포트 생성(31%)

자연어 이해 설문 흐름에서 드러난 것과 마찬가지로 LLM 동향과 관련된 인식들이 드러나고 있다. 먼저 인간 성능에 도달한 과제로 인식되는 것은 문서 요약이다. 요약 과제는 전통적인 자연어 생성 과제로써, 연구된 기간이 비교적 길어 이러한 결과가 나타난 것으로 보인다. 2), 3)은 LLM 사용과 관련된 인식으로써, 사용자와의 상호 작용 시 윤리적, 사회적 논란이 발생할 가능성과 더불어 가장 벤치마크가 필요한 과제라는 인식이 반영되어 있다. 3)의 산업 리포트 생성 과제의 경우 의료 리포트, 금융 리포트 생성 등 산업계와 연계될 경우 실제로 수익 등을 창출할 수 있는 과제라는 인식이 반영되어 있다.

### 3.5.3. 멀티 모달 과제 인식 관련

멀티 모달 과제는 주로 이미지와 관련된 과제에 대해 설문을 진행하였다. 여기에는 비주얼 설명 생성, 비주얼 추론(reasoning), 비주얼 질의응답, 비주얼 스토리 텔링이 포함되었다.

- 1) 인간 성능(human performance)에 도달했거나, 거의 접근한 과제: 비주얼 설명 생성(50%)
- 2) 현실적인 수요가 높아 벤치마크 개발이 이루어져야 함: 비주얼 질의응답(32%)
- 3) 윤리적 혹은 사회적 논란을 일으킬 우려가 있는 과제: 비주얼 스토리텔링(43%)
- 4) 현재에는 접근이 쉽지 않지만 미래 활용 가치가 높은 과제: 비주얼 추론(41%)

설문 조사 결과 인간 성능에 도달한 멀티 모달 과제는 비주얼 설명 생성, 즉 캡셔닝이다. 캡셔닝은 자연어 생성 과제의 일환으로, 상대적으로 연구된 시간이 다른 과제들에 비해 길다. 현실적인 수요가 높아 벤치마크로 개발되어야 하는 과제로는 LLM 트렌드와 관련된 비주얼 질의응답 과제가 꼽혔으며 설문자들은 비주얼 스토리텔링, 즉 이야기 생성 과제에 윤리적, 사회적 논란의 소지가 발생할 수도 있다고 인식하였다. 비주얼 추론의 경우 추론 과제 특성상 난도가 다른 과제들에 비해 높으며, 실제로 설문자들도 이를 인식하고 응답한 것으로 보인다.

위의 결과를 표로 정리하면 아래와 같으며, 응답자들이 벤치마크로 개발이 필요하다고 응답한 2)의 과제들부터 벤치마크로 개편하는 것이 필요하다.

<표 81> 벤치마크 개발 필요성 과제 순위

	NLU	NLG	멀티 모달
1)	언어학 기반 과제	문서 요약	비주얼 설명 생성
2)	<b>정보/관계 추출</b>	<b>일상 대화 시스템</b>	<b>비주얼 질의응답</b>
3)	협오 표현 탐지 과제	일상 대화 시스템	비주얼 스토리텔링
4)	상호 참조 해결	산업 리포트 생성	비주얼 추론

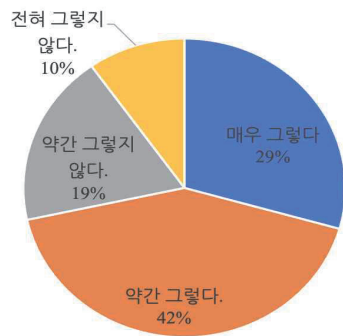
## 4. 리더보드 및 평가 지표 인식 문항

### 4.1. 리더보드 인식 관련

리더보드에 대한 인식 조사는 상금 및 상금 규모, 리더보드 운영 주체, 리더보드 홍보 채널로 나누어 설문을 진행하였다.

#### 4.1.1. 상금 및 상금 규모

상금 필요성에 대해 조사를 진행한 결과 약간 그렇다 (42%), 매우 그렇다 (29%)로, 대다수의 응답자들이 리더보드 운영에 상금이 필요하다고 인식하였다. 이때 적절한 리더보드 우승 상금 규모는 1백-5백만원(48%)이라고 답하였다. 이에 따라 리더보드 활성화 및 참여 동기 부여를 위해서는 상금을 제시할 필요가 있으며, 이때 적절한 상금 규모는 ~500만 원까지이다.



[그림 39] 리더보드 상금 필요성  
응답



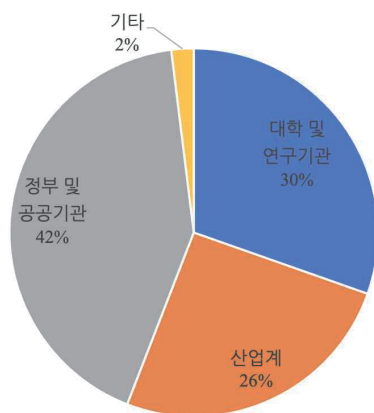
[그림 40] 리더보드 상금 규모  
응답

#### 4.1.2. 리더보드 운영 주체

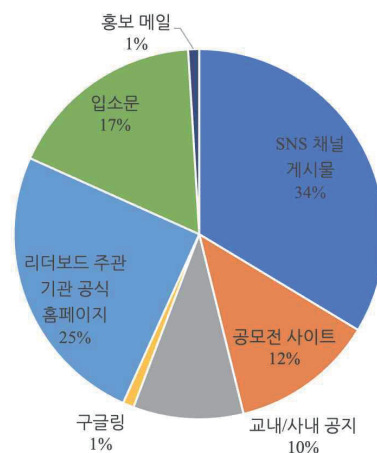
설문 응답자들이 응답한 리더보드 운영 주체로는 정부 및 공공기관 (42%) > 대학 및 연구기관 (30%) > 산업계(26%) 순으로 나타났다. 정부 및 공공기관이 1위를 차지 함으로써 설문 응답자들이 안정적인 리더보드 운영을 원하고 있음을 확인할 수 있었다.

#### 4.1.3. 리더보드 홍보 채널

설문 응답자들이 리더보드 소식을 접하는 곳은 크게 SNS 게시물(34%) > 리더보드 주관 기관 누리집(25%) > 입소문 (17%) > 공모전 사이트 (12%) > 교내/사내 공지 (10%) 순이었다. 이에 따라 SNS 홍보를 중심으로 리더보드 홍보를 진행하되, 공식 홈페이지에 상세한 정보를 올려 리더보드에 대한 정확한 정보를 전달할 필요가 있으며 주 타겟층이 유관 전공 대학생 혹은 대학원생임을 감안하여 학교 및 공모전 사이트에서도 활발한 홍보를 진행하는 것이 필요하다.



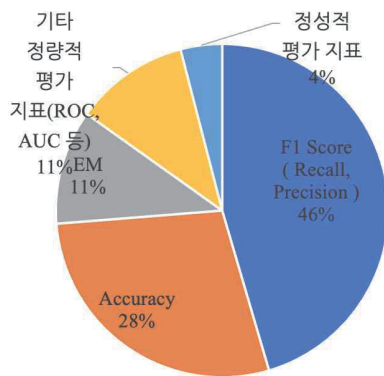
[그림 41] 리더보드 운영 주체 응답



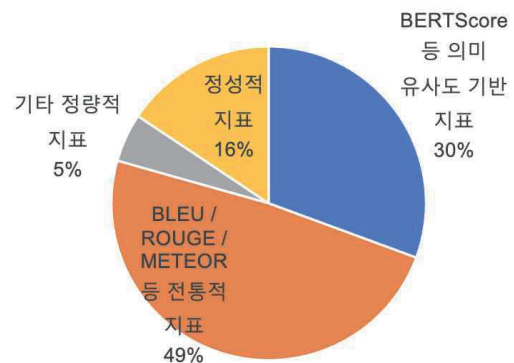
[그림 42] 리더보드 홍보 채널 응답

## 4.2. 평가지표 인식 관련

평가지표는 자연어 이해, 자연어 생성으로 나누어 응답자들의 인식을 조사하였다. 먼저 자연어 이해 평가 지표의 경우 여전히 분류 성능지표(46%) 및 정확도(28%)가 많이 쓰이며, 자연어 이해 과제 수행 시 인간의 품이 들어가는 정성적 평가 지표는 거의 이루어지지 않았다(4%). 자연어 생성의 경우에도 전통적으로 많이 쓰이는 n-gram 기반 지표나 BERTScore 등 의미 유사도 기반의 평가 지표들이 사용되었다. 단, 자연어 이해에 비해 약 4배정도 더 많이 정성적 평가 (16%)가 시행되었다.

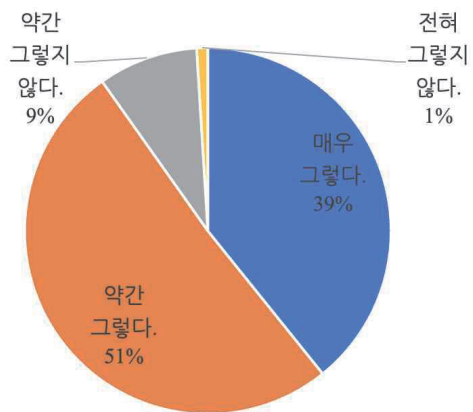


[그림 43] NLU 평가지표 응답

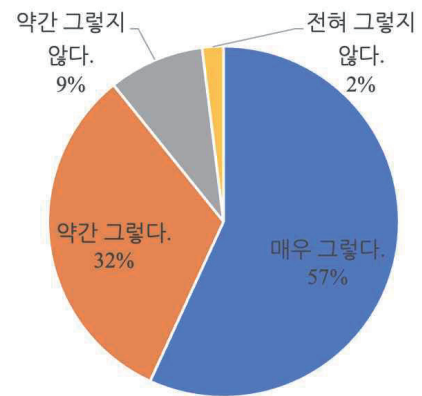


[그림 44] NLG 평가지표 응답

향후 모델 성능 평가 시 인간 평가 결과(human score)를 기준(gold standard)로 제시해야 한다는 말에는 대부분의 응답자들이 긍정적으로 답하였다 (매우 그렇다 39%, 약간 그렇다 51%). 이에 따라 향후 설계하는 벤치마크에 인간 평가 결과를 적극적으로 도입할 필요성이 있다. 아울러 과제별 지표와 더불어 인공 지능에 대한 종합적인 평가 지표를 벤치마크에 포함하는 것 역시도 대부분의 응답자들이 긍정적으로 답해, 벤치마크 개발 시 인공 지능을 총체적으로 평가할 수 있는 지표가 반드시 필요함을 확인할 수 있었다.



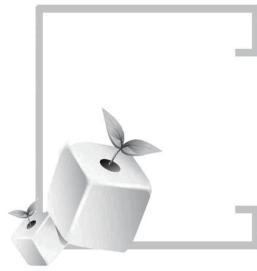
[그림 45] 정성적 평가 필요성 응답



[그림 46] 종합적 평가 지표 개발 응답







## 제 6 장

평가 체계  
홍보 활동





## 6. 평가 체계 홍보 활동

국립국어원의 인공 지능 언어능력 평가체계인 ‘인공지능(AI)말평’에 대한 홍보를 진행하고 자 국립국어원 누리소통망, 공신력 있는 공모전 사이트에 배너, 포스터 등을 게재하는 한편 유관기관 DB를 활용하여 홍보 진행 및 인공 지능 언어 처리 관련 학회·학과를 홍보 대상으로 하여 홍보를 진행하였다. 구체적인 홍보 내용은 아래와 같다.

<표 82> 인공지능(AI)말평 관련 홍보 내역

국립국어원 누리소통망	온라인 배너 게시
관련 학회·학과 TM	관련 분야(학과) TM 진행 후 메일 발송
카카오톡 오픈채팅방	공모전, 대외활동 채널 내 오픈채팅방 활용
공모전 유료 배너	유료 배너 진행
공모전 사이트 홍보글 업로드	공모전 사이트(55)



[그림 47] 인공지능(AI) 말평 홍보 사례

### ○ 공모전 사이트 홍보글 업로드

경진대회 홍보 시에는 전년도 평가 체계 홍보 시 즉각적인 반응이 높았던 공모전 사이트를 적극적으로 활용하여 홍보를 진행하였다. 또한 경진대회 참가 대상층을 고려하여 대학생, 대학원생, 관심 있는 대상자들이 많이 방문하는 공모전 사이트 활용하여 홍보를 진행하였다.

NO	구분	홍보 채널
1	공모전 및 경진대회	Contest index
2	공모전 및 경진대회	Contest Korea
3	공모전 및 경진대회	대터존
4	공모전 및 경진대회	다콘테스트
5	공모전 및 경진대회	데이콘
6	공모전 및 경진대회	독취사
7	공모전 및 경진대회	디자인정글
8	공모전 및 경진대회	링크리어
9	공모전 및 경진대회	링크리어
10	공모전 및 경진대회	배틀콘테스트
11	공모전 및 경진대회	슈퍼루키
12	공모전 및 경진대회	스펙업
13	공모전 및 경진대회	스펙토리
14	공모전 및 경진대회	속삭
15	공모전 및 경진대회	섬국
16	공모전 및 경진대회	섬유
17	공모전 및 경진대회	아이러브 콘테스트
18	공모전 및 경진대회	아이캠핑
19	공모전 및 경진대회	온오프믹스
20	공모전 및 경진대회	올콘
21	공모전 및 경진대회	요즘것들
22	공모전 및 경진대회	위비디
23	공모전 및 경진대회	이벤터스
24	공모전 및 경진대회	인공지능 펍토리
25	공모전 및 경진대회	정글
26	공모전 및 경진대회	자콘테스트
27	공모전 및 경진대회	캠퍼스온
28	공모전 및 경진대회	크라우드픽

[그림 48] 인공지능(AI) 말뚱 홍보글 업로드 사이트 목록1

NO	구분	홍보 채널
29	대학 및 대중매체	딕치고 취업
30	대학 및 대중매체	대학매일
31	대학 및 대중매체	아웃 캠퍼스
32	대학 및 대중매체	앵드루페이퍼
33	대학 및 대중매체	데브리타임
34	대학 및 대중매체	전대모
35	대학 및 대중매체	취업 대학교
36	대학 및 대중매체	취업 뽀개기
37	대학 및 대중매체	취업의 달인
38	대학 및 대중매체	캠퍼스픽
39	스타트업 대상 홍보	넥스트 유니콘
40	스타트업 대상 홍보	데모데이
41	스타트업 대상 홍보	로켓 펀치
42	스타트업 대상 홍보	바이러인 네트워크
43	스타트업 대상 홍보	양재시 허브
44	스타트업 대상 홍보	판교 창업존
45	스타트업 대상 홍보	플래텀
46	언어정보 처리 관련 커뮤니티	AI NLP KOREA
47	언어정보 처리 관련 커뮤니티	IAAE 국제 인공 지능 & 윤리 협회
48	언어정보 처리 관련 커뮤니티	okky
49	언어정보 처리 관련 커뮤니티	언어 공학 연구회
50	언어정보 처리 관련 커뮤니티	짧간 자연어처리
51	언어정보 처리 관련 커뮤니티	정보 과학회
52	언어정보 처리 관련 커뮤니티	캐글 코리아
53	언어정보 처리 관련 커뮤니티	케라스 코리아
54	언어정보 처리 관련 커뮤니티	텐서플로우 코리아
55	언어정보 처리 관련 커뮤니티	파이썬 코리아

[그림 49] 인공지능(AI) 말뚱 홍보글 업로드 사이트 목록2

## ○ 관련 학회·학과 TM

국립국어원의 협조를 통해 인공 지능 언어 처리 관련 학회·학과를 대상으로 하여 TM을 진행하였다. 이에 관련 분야 대학 기관·협회 TM 진행 후 홍보 방안을 확인하였으며, 홍보 이미지 파일 전달, 홍보물 게시 요청 등의 방법을 통해 게재될 수 있도록 하였다. 또한 적극적인 홍보 및 정보 전달을 위해 각 단체 담당자 메일로 행사 자료 전달하였다.

<표 83> 관련 분야 대학 기관·협회 TM

[2023 국립국어원 인공 지능 언어 능력 평가 대회 TM 스크립트]

안녕하세요, 혹시 \*\*\* (학과 또는 단체 이름) 맞으실까요?

2023 국립국어원 인공 지능 언어 능력 평가 대회 홍보를 맡고 있는 마이스앤드라고합니다.

대회와 관련된 정보 및 포스터, 공문을 이메일로 발송해드리면 학교(기관 또는 단체) 온라인으로 게시해주실 수 있으실까요~?

*\*가능하다고 하는 경우 메일 연락처 확인*

안내 주신 메일 주소로 관련 정보와 포스터 보내드리도록 하겠습니다.

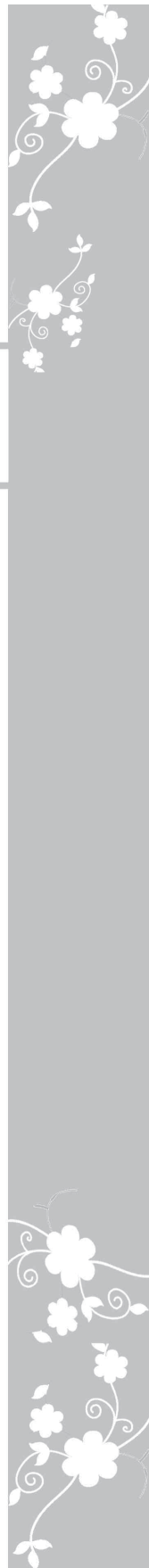
감사합니다.





## 제 7 장

# 평가 체계 운영 지침 및 절차서







## 7. 평가 체계 운영 지침

### 7.1. 경진대회 운영 지침 및 절차서

#### ○ 경진대회 운영 및 절차

<표 84> 경진대회 운영 계획 및 상세 절차

단계	항목	내용	비고
1	경진대회 과제 정의	- 경진대회 과제 정의(초안)	
2	데이터 준비 (1차 검수)	- 데이터 세트 1차 검토 (개인정보 등 비식별화 진행, 오류 데이터 교정)	데이터 세트 검토 진행 중 비식별화 1차 검수 완료 데이터 세트 형식 검토
3	외부 전문가 검토	- 검토 위원회 검토 - 자문 위원회 자문	1, 2차 검토 위원회 진행 1차 자문 위원회 진행
4	경진대회 과제 확정	- 외부 전문가 자문을 바탕으로 경진대회 과제 정의 보완 및 확정 - 평가 지표, 데이터 세트 형식 등 확정	
5	데이터 준비 및 형식 변환 (2차 검수)	- 데이터 세트 2차 검토 (개인정보 등 비식별화 진행, 오류 데이터 교정) - 데이터 세트 형식에 맞추어 변환 - 학습 데이터/개발 데이터/평가 데이터 구분 및 확정	데이터 세트 검토 진행 중 비식별화 2차 검수 완료 데이터 세트 형식 검토 데이터 구분 진행
6	평가 코드 준비	- 각 과제의 목적에 부합하는 평가 코드 준비 - 공개 라이브러리 등을 활용하여 일관성/안정성 있는 평가 코드 개발	상시 과제 리더보드 시스템 적용 중
7	베이스라인 모델 개발 및 공개	- 각 과제에 대한 베이스라인 모델 개발 - 베이스라인 모델에 대한 상세 설명 - 베이스라인 모델은 개발 데이터에 대한 평가 코드를 포함 - 베이스라인 모델은 github를 통해 공개함	

8	경진대회 개설	<ul style="list-style-type: none"> <li>- 과제 기술서 공개 (과제 개요, 데이터 설명, 베이스라인 모델 설명, 평가 지표 및 해석, 안내문, 주요 일정 등)</li> <li>경진대회 리더보드 개설 (참가 방법 및 제출 방법 공지)</li> </ul>	필요 시 외부 전문가 검토 진행
9	홍보	경진대회에 대한 언론 및 관련 커뮤니티 홍보 진행	
10	경진대회 진행	<ul style="list-style-type: none"> <li>- 참가팀 정보 수집</li> <li>- 참가팀 문의사항 답변 (영업일 3일 내)</li> <li>- 참가팀 점수 및 순위 모니터링</li> </ul>	
11	경진대회 평가	<ul style="list-style-type: none"> <li>- (특정 종료 시점이 정해진 경우) 참가팀 결과에 대한 정량적 평가 진행(리더보드 결과)</li> <li>- 참가팀 모델(모델 기술서)에 대한 정성적 평가 진행</li> <li>- 참가팀 결과에 대한 인간평가 진행</li> <li>- 상위 참가팀에 대한 심사위원회 개최(발표평가)</li> </ul>	
12	시상	<ul style="list-style-type: none"> <li>- 시상 안 준비 (과제별 시상 팀 수, 수상자 수 확정)</li> <li>- 시상식 확정</li> </ul>	

경진대회는 2023년 8월 21일부터 2023년 10월 20일까지 진행되었다. 감정분석 및 이야기 완성 두 가지 과제로 진행되었고, 많은 참여를 유도하기 위해 과제 참가 개수는 크게 제한하지 않았다. 과제 참가자들을 위하여 과제 기술서를 작성함으로써 경진대회와 과제에 대한 이해도를 높였으며, 문의의 경우 전화, 전자우편 등 소통 창구를 통하여 대처하였다.

<표 85> 경진대회 운영 안내문

<p><b>1) 리더보드 운영</b></p> <p>참가자들이 주어진 데이터에 대해 형식에 맞춰서 모델이 추론한 결과를 제출하면 서버 내부의 정답지와 자동 비교하여 점수가 매겨지도록 하였다.</p> <p><b>2) 대회 참가 규정 및 제출 방법</b></p> <p><b>(1) 팀 구성</b></p> <ul style="list-style-type: none"> <li>○ ‘모두의 말뭉치’ 회원만 참가할 수 있다.</li> </ul>
--

- 참가자는 접수 기간 중 여러 개의 팀에 참여할 수 있으며, 팀 구성 시 인원 제한은 없다.
- 다만, 최종 심사 대상 중 참가자가 속한 팀이 여러 개의 팀이면 참가자는 하나의 팀을 소속 팀으로 결정하여야 한다.
- 중복 참여는 가능하나 중복 수상은 할 수 없다.

## (2) 답안 제출

- 참가자(팀)는 대회에서 제시한 과제를 해결한 결과를 표본(샘플) 파일과 동일한 형식으로 작성하여 제출한다.
- 한 참가자(팀)가 복수의 결과물을 제출할 경우 각 팀의 제출 모델 및 결과 중 가장 높은 성적만을 순위표(리더보드)에 게시한다.

## (3) 모델 사용 및 제출

- 라이선스에 문제가 없는 모델 사용 가능(라이선스의 검토 책임은 참가팀에게 있음)
  - 외부 데이터 추가 사용 불가
    - 외부 데이터로 학습된 사전학습 언어모델 중 8월 10일 이전에 공개된 언어모델 사용 가능(BERT, polyglot-ko, Llama-2 등)
    - 사용된 기본모델과 함께 제출(제출된 기본모델도 평가될 수 있음, 공개된 모델과 차이가 있는 경우 재현성 없음으로 판단될 수 있음)
  - 경진대회 진행 시 외부 API(예: chatGPT api) 이용 불가
  - 로컬 환경에서 동작하는 모든 모델 사용 가능(여러 LLM 포함)
    - 조건 1: 외부 API 이용 불가
- > 참가자(팀)의 서버에서 구동한 생성 AI의 경우에 한정함(프롬프트 엔지니어링 가능)
- 조건 2: RTX 4090 24GB 1개에서 구동 가능한 모델만 사용 가능
    - > 이 과제에서는 최근의 연구 동향을 반영하여 모델의 경량화에 초점을 맞추어 제한된 하드웨어에서 동작하도록 제한함
    - > 기업 및 연구실 외에도 많은 개인 참가자가 참가하여 경쟁할 수 있도록 함

## (4) 시스템 사용 안내

- 참가 신청(팀 구성)
  - 과제 목록 선택 후 참가 신청을 눌러 신청서를 작성한다.
    - 신청서는 '모두의 말뭉치' 회원만 작성할 수 있고, '모두의 말뭉치' 회원의 전자 우편 주소를 입력하여 팀을 구성할 수 있다(회원이 아닌 경우 팀원이 될 수 없음에 유의할 것).
    - 과제 참가 신청서에 작성한 팀원에게 과제 참가 신청 동의서가 발송된다.
    - 신청자가 구성한 팀원 모두가 전자 우편을 확인하고 과제 참가에 동의하면 신청서 접수가 완료된다.
- 제출 관리
  - 참가 신청을 완료한 후 '제출 관리'를 통해 예측 결과를 제출한다.
  - 모델명 및 모델 설명을 작성하고, 모델 예측 결과(JSON-L 파일)를 등록한다.
  - 하루에 5회까지 제출할 수 있으며, 제출 결과는 모두 순위표(리더보드)에 반영된다.
- 순위표(리더보드)
  - '제출 관리'에서 등록한 예측 결과 중 일정 비율(예 70%)을 무작위 추출하여 평가한 후 순위표

(리더보드)에 평가 점수 및 순위를 제공한다.

- 제출한 결과 중 가장 높은 평가 점수가 순위표에 제공된다(좌측 화살표 버튼을 누르면 다른 결과물의 점수 및 순위 확인 가능).

○ 참여자 게시판: 참가자들이 과제별 정보 공유를 위해 활용 가능하다.

(5) 기타: 시험 데이터의 정답 공개 계획은 없다.

## ○ 경진대회 운영 시 질의 응답

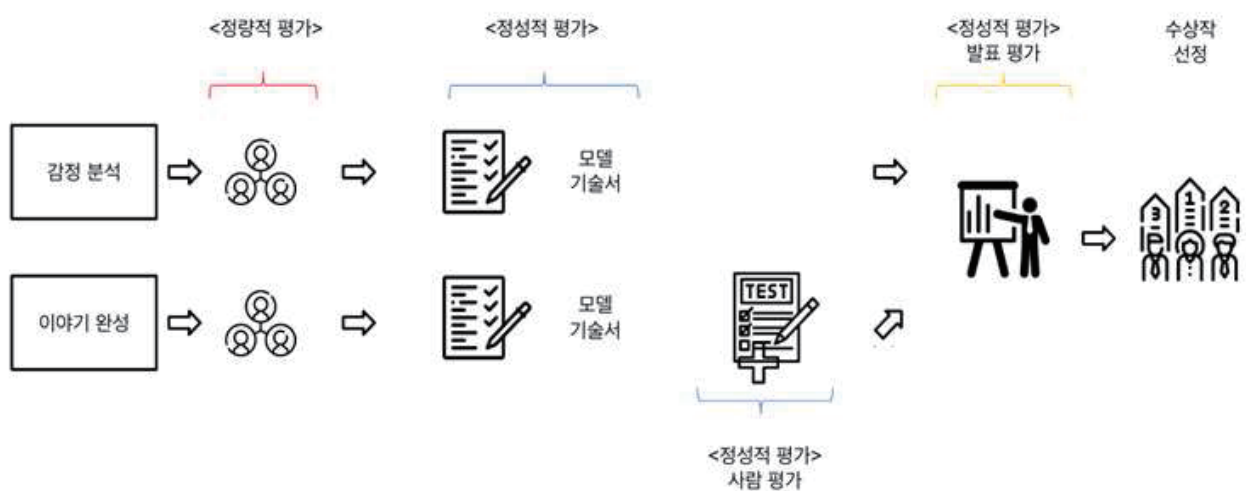
경진대회 진행 과정에서 참가자들의 질문에 대해 응답할 수 있도록 임시 연락처를 오픈하였고 질의응답 게시판 및 메일을 활용하였다.

- 전화로 질문에 대해 응답 - 5건
- 게시판 또는 전자우편으로 응답 - 이야기완성 17건, 감정분석 49건

## ○ 경진대회 심사 진행 단계

정량적 평가, 정성적 평가(모델 기술서, 사람 평가(인간평가), 발표 평가)를 종합하여 수상작 선정

- 감정분석의 경우 정량적 평가 50, 모델기술서 50, 발표평가 100으로 진행
- 이야기완성의 경우 정량적 평가 50, 모델기술서 25, 인간평가25, 발표평가 100으로 진행



[그림 50] 경진대회 심사 진행 단계

## ▷ 정량적 평가 수행

- 최종 모델의 inference 결과에 대한 AI 말평 리더보드의 정량적 점수 → 과제별

## 상위 10팀 선정

### ▷ 정성적 평가 수행

#### (1) 모델기술서 평가

- 모델기술서와 코드를 바탕으로 재현성 및 모델의 우수성 정성적 평가
- 모델 재현성 검정은 ‘모델 재현성’, ‘데이터의 재현성’, ‘편의성’, ‘성능 재현’으로 평가하며, 모델 우수성 검증은 ‘문제 인식 및 방법론’, ‘알고리즘의 우수성’, ‘프로그램의 우수성’으로 평가

<표 86> 참가자 시스템 검증을 위한 평가 항목

검증	기준 모델	항목
모델 재현성 검증	참가팀의 소스코드는 학습 코드를 포함하여 재현이 가능한가?	모델 재현성
	참가팀의 모델은 라이선스에 문제가 없고 공개된 데이터를 사용하였는가? (데이터 활용의 재현성)	데이터 재현성
	참가팀의 소스코드가 공개되어 다른 사람이 바로 사용할 수 있는 수준인가?	편의성
	참가팀의 모델은 제출된 경진대회의 결과물을 재현하는가?	성능 재현
모델 우수성 검증	참가팀은 본 과제의 성격을 잘 이해하고 있는가?	문제 인식 및 방법론
	참가팀이 풀고자 하는 문제는 얼마나 중요한가?	
	참가팀이 제시한 방법론은 참가팀이 풀고자 하는 문제에 적당한가?	
	참가팀이 제안한 알고리즘은 참가팀의 방법론에 적합한 방법인가?	알고리즘의 우수성
	참가팀이 제안한 알고리즘은 최신의 연구 동향을 반영한 우수한 방법인가?	
	참가팀이 제안한 알고리즘은 독창성이 있는가?	
	참가팀이 제출한 소스코드는 방법론 및 알고리즘을 충실히 구현하였는가?	프로그램의 우수성
	참가팀이 제출한 소스코드는 가독성이 높고 간결하게 작성되어 활용 가능성이 높은가?	

- 결과 재현 평가를 통과한 참가팀의 코드 리뷰 수행(과제 인력으로 코드에 대한 전수검사 실시)
- 학습 코드에 대한 정성평가 및 재현
- 사용 모델의 최신 인공지능 연구 동향과의 정합성 평가
- 코드 내 부정행위가 포함되어 있는지를 전수검사 수행
- 필요 시 코드 품질 평가(코드 모듈화, 재사용성, 코드 수행 시간 등)

(2) 인간평가(이야기 완성)

- 이야기 완성 과제의 경우 문장 1, 문장 3이 주어졌을 때, 적절한 문장 2를 생성해야 하므로 창의성에 따라 다양한 응답이 정답으로 여겨질 수 있음. 따라서 일반적인 rouge와 같은 지표만으로는 정확한 평가를 하기 어렵기 때문에 인간평가 수행

(3) 발표 평가(심사위원회 평가)

- 심사위원은 정량적 평가(리더보드), 모델기술서 평가, 인간 평가 결과를 참고하여 아래 표를 기반으로 평가

<표 87> 발표 평가 시 심사 기준 및 배점

평가 분야	배점
1. 과제 이해도	10점
2. 방법론의 적합성	10점
3. 모델의 우수성	10점
4. 모델의 독창성	10점
5. 연구 결과 분석	10점
총점	50점

## 7.2. 경진 대회 결과

### ○ 참가 현황

<표 88> 경진대회 참가 현황

과제	참가 팀	제출 모델 수	참가 인원	최종 후보 팀*
감정 분석	146팀 참가	1,974개(1916개)	203명	6팀
이야기 완성	87팀 참가	696개(680개)	122명	4팀
(합계)	233팀	2,670개	325명	10팀

\* 최종 후보 팀의 경우 각 리더보드 상위 10위 팀 중 모델 제출 동의 및 최종 평가 발표회에 참가 의사를 밝힌 팀으로, 총 10팀(자연어 이해 6팀, 자연어 생성 4팀)이 최종 후보로 선정

### ○ 감정분석 상위 10개 팀(정량 평가 지표 기준)

<표 89> 감정분석 정량 평가 결과

순위	팀명	F1 점수	팀 인원	모델명
1	LIA	90.2520061	1	tmp_18
2	사과는맛있어맛있으면바나나	90.1989049	1	또 다른 내 친구는 내 어쩔 두드리며 잊어버리라 했지만 잊지 못할 것 같아
3	로도스 아일랜드	90.1693419	2	0.5
4	동물의 왕국	90.1516253	6	최종5
5	오늘은맑음	90.1462726	1	맑음3
6	가천킹	90.1170947	1	고래밥왕
7	방생된노예	89.9049112	1	노예의28번째모임
8	야부영외부영	89.6571584	7	힐볼
9	자허블	89.6059217	1	only
10	라이언	89.5996691	7	polyglot-ko-all2

## ○ 이야기완성 상위 10개 팀(정량 평가 지표 기준)

<표 90> 이야기완성 정량 평가 결과

순위	팀명	F1 점수	팀 인원	모델명
1	엘리	60.2040711	3	test_v31_3
2	R.content	60.1837255	4	sheep-dock-mhs-v3
3	MLP Lab	59.889418	9	봉준호BTS임경태Les'sGo
4	개쩌는 팀	59.7559598	1	꼬시19
5	국내산 자연어	59.7559545	6	최종꼬시
6	603310	59.7315135	2	한번만 더
7	도레미파	59.6596753	1	북23
8	ロシゝヒョウ	59.6395971	1	〇スネ
9	최고의공격	59.5941489	1	콤보
10	아무거나	59.5934229	1	24

## 7.3. 이야기 완성 과제 인간평가 지침

- 2023년 국립국어원 ‘AI 말평’ 내 이야기 완성 과제에 대해 정량적 지표만으로는 생성 AI 모델의 성능을 온전히 평가할 수 없다고 판단하여 생성 텍스트에 대한 인간 평가를 진행하였다. 이를 위해 평가 데이터(test data) 15,018건 중 1,000건에 대해 각 참가팀이 생성한 결과물에 대해 두 단계로 인간 평가를 진행하였다.
- 인간 평가는 1) 평가 준거에 대한 적합/부적합(P/NP) 판단, 그리고 2) 답변에 대한 선호도 평가 토너먼트로 구성하였으며 1단계 평가의 경우 대학 교양 글쓰기 수업을 이수한 학부생 25명, 2단계 평가는 생성 AI 텍스트에 익숙한 전산언어학 전공 대학원생 7명이 평가자로 참여하였다. 2단계 평가 결과에 대해서는 대학 글쓰기 수업을 담당하는 전문가 2인의 자문을 받아 평가의 신뢰성과 타당성을 제고하였다. 인간 평가는 1, 2단계 모두 워크벤치를 사용하여 진행되었으며, 이를 통해 효율적인 평가 작업 관리 및 결과 분석이 가능하였다. 평가에 대한 요약은 아래 표와 같다.



<표 91> 인간평가 요약

항목	내용
평가 대상	이야기 완성 과제에 대해 생성AI 모델이 생성한 텍스트 <ul style="list-style-type: none"> <li>1단계: 평가 데이터 문제 1,000건에 대해 정량적 평가 상위 10위팀에서 제출한 답변으로, 총 10,000개 텍스트</li> <li>2단계: 1단계에서 평가한 문제 1,000건별 평균 점수의 사분위수별 50건으로 총 200건</li> </ul>
평가 방법	2단계로 진행, 최종 점수 산출 시 각 단계 순위를 50%씩 반영 1, 2단계 모두 워크벤치를 활용하여 평가 진행 <ul style="list-style-type: none"> <li>1단계: 평가 준거 기준 적합/부적합(P/NP) 판단</li> <li>2단계: 답변 선호도 평가 기반 토너먼트 + 전문가 감수</li> </ul>
평가 주체	<ul style="list-style-type: none"> <li>1단계: 대학 교양 글쓰기 수업을 이수한 학부생</li> <li>2단계: 전산언어학 전공 대학원생, 글쓰기 전문가</li> </ul>

### ○ 1단계) 평가 준거 기준 적합/부적합(P/NP) 판단

1단계 평가 대상은 평가 데이터 문제 1,000건에 대해 정량 평가 상위 10팀이 제출한 답변으로, 총 10,000개 텍스트이다. 평가 준거에 비추어 보았을 때 생성된 텍스트가 해당 준거에 적합한지, 부적합한지를 평가자가 평가하였으며, 적합일 때는 1점, 부적합일 때는 0점을 부여하였다. 평가자는 학부생 25명으로, 인당 400개 텍스트를 평가하였다.

<표 92> 인간 평가 1단계 개요

구분	주요 내용
적합/부적합 (P/NP) 판단	<b>대상:</b> 평가 데이터(test data) 문제 1,000건에 대해 프로그램 자동 평가 상위 10위 팀에서 제출한 답변 총 10,000개 <b>방법:</b> 평가 준거에 비추어 답변에 대한 적합/부적합(P/NP) 판단 <ul style="list-style-type: none"> <li>적합(P)의 경우는 1점, 부적합(NP)의 경우 0점 부여</li> <li>평가자 25명(학부생)이 1인당 하루 80건, 총 400건 평가</li> </ul>

### ▷ 1단계 평가 준거

1단계 평가 준거는 평가 준거 자문 위원들의 자문 내용을 토대로 수립되었다. 평가 준거는 크게 내용, 표현, 표기 준거 3가지로 나누어지며, 각 준거들에 대해 세부 판단 기준을 둬으로써 일관된 인간 평가가 가능하도록 하였다. 이후 국립국어원과의 논의를 통해 실제 경진대회 진행 시 적용할 수 있는 수준으로 판단 기준에 대한 조정이 이루어졌으며, 본격적인 평가에 들어가기 앞서 조정된 내용을 바탕으로 평가자들에 대한 교육

을 진행하였다. 평가자들에게는 평가 시 주어진 대상 문장을 세부 판단 기준에 비추어 보았을 때 크게 위배하지 않은 경우는 ‘적합’, 그렇지 않은 경우 ‘부적합’으로 판단하도록 대원칙을 안내하였고, 지나치게 어렵게 생각하지 않고 ‘첫 인상’에 든 느낌을 평가에 활용하도록 하여 인간 평가 목적에 어긋나지 않도록 주의 사항을 전달하였다.

또한 교육 시 평가자들의 일치도 제고 및 인간 평가 대상 문장에 대한 적응 차원에서 오류가 포함된 연습용 생성 AI 텍스트에 대해 시범 평가를 수행하였다. 시범 평가 시 나온 질의응답이나 전체 논의 사항은 준거에 반영함으로써 보다 정확한 평가가 이루어질 수 있도록 하였으며 일련의 작업을 통해 최종적으로 아래와 같은 준거를 마련하였다.

<표 93> 인간평가 1단계평가 준거

평가 대원칙: 주어진 평가 대상 문장을 세부 판단 기준에 비추어 보았을 때 판단 기준을 크게 위배하지 않은 경우는 ‘적합’, 그렇지 않은 경우는 ‘부적합’ 판정		
준거	판단 기준	예시
내용	<ul style="list-style-type: none"> <li>● 문제 상황에 대한 이해 : 문장1에서 제시된 문제 상황에 대해 충분히 이해하고 작성되었는가?</li> </ul>	<p>&lt;적절 예시&gt;</p> <p>문장1: 민수는 요즘 보드 타는 것에 취미를 들었다.</p> <p>문장2: 그는 <u>맨날 보드를 타다가 다쳐서 나타났다.</u></p> <p>(▶ ‘보드타는 것’이라는 상황에 적절하게 문장 생성)</p> <p>문장3: 하지만 보드 타는 것을 멈추지 않았다.</p> <p>&lt;부적절 예시&gt;</p> <p>문장1: 나는 목표했던 일을 이루지 못할까봐 떨렸다.</p> <p>문장2: 나는 <u>노력이 중요하다고 생각한다.</u></p> <p>(▶ ‘목표 달성’이라는 상황에 적절하지 않게 문장 생성)</p> <p>문장3: 나는 뿌듯함과 시원한 감정이 동시에 들었다.</p>
	<ul style="list-style-type: none"> <li>● 논리적 연결성 : 생성한 문장이 앞.뒤 문장과 문맥적으로 자연스럽게 연결되도록 사건을 논리적으로 명료하게 설명하는가?</li> </ul>	<p>&lt;적절 예시&gt;</p> <p>문장1: 나는 입사하고 나서 몇 달 동안은 조심스럽게 행동했다.</p> <p>문장2: 회사 분위기를 파악하고 나서는 <u>눈치껏 내 성격을 드러냈다.</u></p> <p>(▶ 문장1의 ‘조심스럽게 행동한 것’과 문장3의 동료들이 놀란 이유에 논리적으로 적절하게 문장2 생성)</p> <p>문장3: 그랬더니 동료들은 첫인상과 다른 나의 모습에 놀랐다.</p> <p>&lt;부적절 예시&gt;</p> <p>문장1: 그는 뮤지컬 공연 오케스트라를 지휘하여 공연을 성공적으로 마쳤다.</p> <p>문장2: <u>그러나 관객들은 그에게 야유를 보냈다.</u></p> <p>(▶ 문장1의 ‘공연을 성공적으로’ 마친 것과 문장3의 ‘기쁜 마음으로’ 공연을 준비하려 나가는 내용을 논리적으로 연결하지 못하는 문장 생성)</p> <p>문장3: 그리고 그는 기쁜 마음으로 바로 다음 공연을 준비하러 나갔다.</p>
표현	<ul style="list-style-type: none"> <li>● 문장 표현의 자연스러움 : 문장이 자연스럽게</li> </ul>	<ul style="list-style-type: none"> <li>● 앞뒤 문장 혹은 문맥에 비추어 보았을 때 문장 길이가 지나치게 짧거나 길게 생성되었는지에 따라 적절/부적절</li> </ul>

	<p>생성되었으며 길이가 적절한가?</p> <p>→ “문장”에 대한 이슈로 어순, 문장 길이 등을 판단</p>	<p>판단</p> <p>● 어순 혹은 문장 내 어휘, 동어 반복 등으로 인해 문장이 부자연스럽게 느껴지는 경우 부적절 판단</p> <p>&lt;적절 예시&gt;</p> <p>문장1: 나는 오늘 날씨가 정말 마음에 들었다.</p> <p>문장2: <u>그래서 나는 친구들에게 학교 광장에 나가서 수다를 떨자고 말했다.</u></p> <p>문장3: 친구들도 날씨가 좋다며 흔쾌히 그렇게 하자고 답했다.</p> <p>&lt;부적절 예시&gt;</p> <p>문장1: 나는 오늘 날씨가 정말 마음에 들었다.</p> <p>문장2: <u>날씨가 정말 마음에 들었기 때문에 친구들에게 학교 광장에 나가서 앉아서 좋은 날씨를 즐기면서 오랜만에 서로 이야기하는 시간을 가지자고 이야기했다.</u></p> <p>(▶ 지나치게 장황한 문장 생성)</p> <p>문장3: 친구들도 날씨가 좋다며 흔쾌히 그렇게 하자고 답했다.</p> <p>&lt;부적절 예시2&gt;</p> <p>문장2: 전화기라고 생각했는데, 귀에 대는 것은 다리미였다.</p> <p>▶ 모어가 한국어인 화자 직관으로 보았을 때 부자연스러운 문장이므로 ‘부적절’ 판단</p>
	<p>● 어휘 사용의 적절성: 문맥에 적절한 어휘를 사용하였는가?</p> <p>→ “어휘”의 의미의 적절성, 높임법 등에 대한 이슈</p> <p>→ 높임법의 경우 조사, 어미 사용의 엄밀성이 아닌 문장 전체의 뉘앙스로 판단</p>	<p>&lt;적절 예시&gt;</p> <p>문장2: 할머니가 침대에 누워 계셨다.</p> <p>(▶ ‘할머니’에 ‘께서’라는 체언 높임 조사가 쓰이지 않았으나, 문장 전반적인 뉘앙스가 높임법이므로 ‘적절’ 판단)</p> <p>&lt;부적절 예시&gt;</p> <p>문장1: 나는 할아버지 댁에 건너가기 전에 어머니께 연락을 드렸다.</p> <p>문장2: <u>어머니는 나에게 복귀 시간을 물어보셨다.</u></p> <p>(▶ ‘복귀 시간’은 일반적인 상황에서 잘 쓰이지 않는 어휘이므로 부적절 판단)</p> <p>문장3: 나는 어머니께 정확히 언제 돌아올 지 모르겠다고 말했다.</p>
표기	<p>● 표기의 수용성: 의미 전달에 방해되는 표기 오류가 포함되어 있지 않은가?</p> <p>→ 단, 일상 생활에서 자주 틀리는 띄어쓰기나 어문 규범 등 용납 가능 범위의 표기는 ‘적절’판단</p>	<p>&lt;적절 예시1&gt;</p> <p>문장2: 잠을 잘 자기 위해 자기 전에 따뜻한 우유를 마셨다.</p> <p>&lt;적절 예시2&gt;</p> <p>문장2: 그와 못만난지가 한참 되어서 매우 반가웠다.</p> <p>▶ 못V만난V지이나, 일상적인 띄어쓰기 오류이므로 ‘적절’</p> <p>&lt;부적절 예시&gt;</p> <p>문장2: <u>잘 잠을 자기 위해 자기 전 땀한 우유를 마셨다.</u></p> <p>▶ 오타 등은 없으나 부사 위치 및 비표준어로 인해 표기의 수용성에서 오류 발생</p>

		<부적절 예시2> 문장2: 그 일이 있는 지가 벌써 몇일 지났다. ▶ ‘몇일’은 ‘며칠’의 명백한 오타이므로 부적절 판단
--	--	---

평가 결과 전체 문제 1,000건에 대한 상위 10위팀의 답안들은 최고 4.75점, 최저 4.31점의 평균 점수를 보였다. 세부 판단 기준별로는 어휘 사용의 적절성 기준을 통과한 텍스트가 가장 많았으며, 그 뒤로는 표기의 수용성, 문장 표현의 자연스러움, 문제 상황에 대한 이해, 논리적 연결성 순으로 각 기준을 통과한 텍스트 수가 감소하였다. 이를 통해 인공 지능이 생성한 텍스트가 가장 만족하기 쉬운 기준은 ‘어휘 사용의 적절성’, 가장 만족하기 어려운 기준은 ‘논리적 연결성’임을 확인할 수 있었다.

## ○ 2단계) 답변 선호도 평가 토너먼트

2단계 평가 대상은 1단계에서 사용한 평가 데이터 문제 1,000건에 대한 문제별 답변들의 점수 평균을 사분위수로 구하여 골라낸 문제 200건이다. 이후 선정된 문제 200건에 대해 10팀이 생성했었던 텍스트들을 토너먼트 형식으로 비교하는 선호도 평가를 진행하였다. 해당 평가 방법은 인간 피드백 기반 강화 학습(RLHF; Reinforcement Learning with Human Feedback) 시의 데이터 평가 방식을 참고하여 두 답변 간 선호도를 인간 평가 과정에 반영한 것이다. 선호도 평가 시 보편적으로  $n$ 인 이상의 다수결 혹은 과반 이상 등의 방법을 사용하며, 본 2단계 평가에서도 이러한 선례들을 고려하여 3인 평가 및 과반 찬성의 방법을 채택하였다.

이에 따라 평가 시 한 문제당 대학원생 3인이 평가하였으며, 주어진 두 팀의 답변 중 어느 한 답변에 대해 2인 이상 선호 시 해당 답변은 다음 라운드로 진출하는 것을 기본 원칙으로 하였다. 다만 실제로 각 문제들에 대해 팀별로 생성한 텍스트를 살펴보았을 때, 문제에 대해 동일한 답변을 생성한 복수의 팀들이 발생하는 것을 확인하였다. 이에 대해서는 불필요한 토너먼트 진행 방지 및 보다 타당성 있는 평가를 위해 해당하는 팀들을 그룹으로 묶어 토너먼트를 진행하였다. 즉, A, B, C 팀끼리 같은 문장 생성 시, {A, B, C}를 하나의 팀으로 간주하여 나머지 팀과 토너먼트를 진행하며, 평가자가 A(혹은 B, C)의 답변이 나머지 팀 답변보다 좋다고 응답 시 A, B, C가 동시에 점수를 얻는 방식으로 평가하였다.

평가 준거 및 세부 판단 기준이 있었던 1단계 평가와는 달리 2단계 평가에서는 총체적인(holistic) 평가로 진행하였으며, 이에 따라 판단 기준은 평가자의 선호에 입각하게 된다. 2단계 평가는 평가자들에게 이야기 완성 과제의 목적에 비추어 주어진 두 문장 중 선행, 후행 문장에 대해 보다 논리적인 흐름을 보이는 문장을 개인의 선호에 따라 선택하도록 안내

하였다. 2단계 인간 평가에서 대학원생들이 인간 평가를 진행한 결과에 대해서는 글쓰기 전문가 2인의 자문을 받음으로써 평가에 대한 신뢰성 및 타당성을 제고하였다.

<표 94> 인간 평가 2단계 개요

구분	주요 내용
선호도 평가 (preference ranking)	<p><b>대상:</b> 1단계(적합/부적합 판단)에서 평정한 문제 1,000건의 평균 점수값 사분위수에 따라 선정한 200건</p> <p><b>방법:</b> 선정된 200건에 대해 문제별로 10개 팀 답변을 토너먼트 형식으로 선호도 평가두 팀의 답변 중 선호하는 답변 선택</p> <ul style="list-style-type: none"> <li>선호 답변 개수가 과반수면 승리</li> <li>한 문제 당 3인 평가(대학원생)</li> </ul>

인간 평가 진행 시에는 워크벤치를 사용하여 체계적이고 효율적으로 인간 평가를 진행하였으며, 평가 결과 누락 등을 미연에 방지하였다.

<표 95> 인간 평가 워크 벤치

teddysum   데이터 구축 플랫폼				
<div> <div>로그인</div> <div>회원가입</div> </div>				
teddysum   2023인간평가 데이터 구축 플랫폼				
관리 yonsei_m11				
	전체 문서수	작업현황		
		전 기간	작업 가능	작업 완료
선호도 평가 작업가능	0	yonsei_26	0	408
		yonsei_27	0	409
		yonsei_28	0	409
		yonsei_29	0	409
		yonsei_30	0	409
		yonsei_31	0	409
		yonsei_32	0	409
선호도 평가 작업중가	0			
선호도 평가 작업완료	2862			

## ○ 경진대회 최종 순위 산정

경진대회 최종 순위는 기본적으로 정량적 평가 지표 순위와 모델 혹은 결과물에 대한 정성적 평가 결과를 토대로 산정하였다. 이에따라 감정 분석 과제는 리더보드 내 정량적 점수, 모델 기술서 평가 점수가 순위 결정에 반영되었으며, 이야기 완성 과제의 경우 감정 분석과 동일한 요소들과 더불어 인간 평가를 순위 결정 시 반영하였다. 이러한 흐름에 따라 감정 분석, 이야기 완성 과제 모두를 통틀어 상위 10위팀이 결정되었다. 평가 프로세스에 대한 대략적인 흐름은 아래와 같다.

- ▷ 감정 분석, 이야기 완성: 정량적 평가 지표 순위 확인 & 과제별 정량적 평가 상위 10위팀 선정
- ▷ 이야기 완성: 상위 10팀 과제 결과에 대한 인간 평가
- ▷ 감정 분석, 이야기 완성: 상위 팀 과제 모델에 대한 모델 기술서 평가
- ▷ 감정 분석, 이야기 완성: 심사위원회 개최 및 팀별 발표
- ▷ 감정 분석, 이야기 완성: 최종 수상자 선정

결론적으로 최종 순위 산정 시에는 정량적 점수, 인간 평가(이야기 완성 과제), 모델 기술서, 심사위원회 점수가 모두 고려되었다. 정량적 점수의 경우 점수에 대한 정규화를 수행하였다. 정규화는 두 과제의 정량적 점수를 동일 선상에서 비교하기 위해 수행되었으며, 상위 10팀의 원점수에 대해서는 아래의 식과 같이 계산하였다. 이때 최저 점수는 25점이 되도록 조정하였다.

<표 96> 표준 점수 계산식

$$\left( \frac{\text{원점수} - \text{최저 점수}}{\text{최고점수} - \text{최저점수}} \times 25 \right) + 25$$

인간 평가 점수는 총점을 25점으로 하여 계산하였다. 점수 산정 시 1단계, 2단계는 50:50 비율로 반영되었으며 원점수는 전체 10위에 대해 각 팀이 받은 1, 2단계 등수를 평균한 것의 차를 사용하였다. 모델 기술서 원점수는 5점이며, 순위 산정 시에는 25점을 만점으로 하여 계산하였다.

<표 97> 1단계, 2단계 반영식

$$\text{원점수} = (10 - \text{인간평가1, 2단계 순위 평균})$$

## 7.4. 상시 과제 운영 지침 및 절차서

### ○ 상시 과제 운영 및 절차

<표 98> 상시 과제 운영 계획 및 상세 절차

단계	항목	내용	비고
1	상시 과제 정의	- 상시 과제 정의(초안)	
2	데이터 준비 (1차 검수)	- 데이터 세트 1차 검토 (개인정보 등 비식별화 진행, 오류 데이터 교정)	데이터 세트 검토 진행 중 비식별화 1차 검수 완료 데이터 세트 형식 검토
3	외부 전문가 검토	- 검토 위원회 검토 - 자문 위원회 자문	1, 2차 검토 위원회 진행 1차 자문 위원회 진행
4	상시 과제 확정	- 외부 전문가 자문을 바탕으로 상시 과제 정의 보완 및 확정 - 평가 지표, 데이터 세트 형식 등 확정	
5	데이터 준비 및 형식 변환 (2차 검수)	- 데이터 세트 2차 검토 (개인정보 등 비식별화 진행, 오류 데이터 교정) - 데이터 세트 형식에 맞추어 변환 - 학습 데이터/개발 데이터/평가 데이터 구분 및 확정	데이터 세트 검토 진행 중 비식별화 2차 검수 완료 데이터 세트 형식 검토 데이터 구분 진행
6	평가 코드 준비	- 각 과제의 목적에 부합하는 평가 코드 준비 - 공개 라이브러리 등을 활용하여 일관성/안정성 있는 평가 코드 개발	상시 과제 리더보드 시스템 적용 중
7	베이스라인 모델 개발 및 공개	- 각 과제에 대한 베이스라인 모델 개발 - 베이스라인 모델에 대한 상세 설명 - 베이스라인 모델은 개발 데이터에 대한 평가 코드를 포함 - 베이스라인 모델은 github를 통해 공개함	
8	상시 과제 개설	- 과제 기술서 공개 (과제 개요, 데이터 설명, 베이스라인 모델 설명, 평가 지표 및 해석, 안내문, 주요 일정 등) - 상시 과제 리더보드 개설 (참가 방법 및 제출 방법 공지)	필요 시 외부 전문가 검토 진행
9	홍보	- 상시 과제에 대한 언론 및 관련 커뮤니티 홍보 진행	

10	상시 과제 진행	<ul style="list-style-type: none"> <li>- 참가팀 정보 수집</li> <li>- 참가팀 문의사항 답변 (영업일 3일 내)</li> <li>- 참가팀 점수 및 순위 모니터링</li> </ul>	
11	상시 과제 평가	<ul style="list-style-type: none"> <li>- (특정 종료 시점이 정해진 경우) 참가팀 결과에 대한 평가 진행</li> <li>- (필요 시) 참가팀 모델에 대한 정성적 평가 진행, 참가팀 모델 및 결과에 대한 심사위원회 개최</li> </ul>	(필요 시 진행)
12	시상	<ul style="list-style-type: none"> <li>- 시상 안 준비 (과제별 시상 팀 수, 수상자 수 확정)</li> <li>- 시상식 확정</li> </ul>	(필요 시 진행)

#### □ 상시 과제 및 AI 말평 체계 시스템 운영 지원

##### ○ 상시 과제 자동 평가를 위한 평가 코드 제공

<표 99> 개발된 평가 코드

항목	기능	내용
분류 평가	분류(classification)을 위한 F1 평가 코드	<ul style="list-style-type: none"> <li>- 평가 지표: Micro F1, Macro F1, Weighted F1, multi-label-classification micro F1</li> <li>- 활용: 감정분석 및 함의분석, 부적절성 문장에 대한 태도 탐지 등 분류 문제에 활용</li> </ul>
수치 평가	수치적 예측에 대한 MSE 평가 코드	<ul style="list-style-type: none"> <li>- 평가 지표: MSE</li> <li>- 활용: 추론 확신성 과제(2022)에 적용</li> </ul>
생성 평가	생성 모델에 대한 평가 코드	<ul style="list-style-type: none"> <li>- 평가 지표: ROUGE-1, ROUGE-L BLEU, BLEURT, BERTScore</li> <li>- 활용: 표의 일부분에 대한 해석 생성, 문자가 포함된 이미지 기반 문장 생성과 같은 생성 과제에 적용</li> </ul>



평가 코드의 사용 예시는 아래와 같다.

<표 100> 개발된 평가 코드 예시

```
from evaluation import evaluation
```

```
result = evaluation(submit_data, test_data, evaluation_metrics=['ROUGE-1', 'ROUGE-L',  
'BLEU'], ratio=0.5, iteration=10)
```

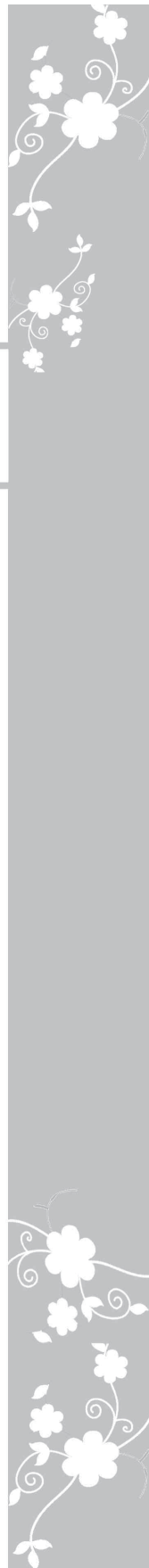
- submit\_data: 참가팀이 제출한 데이터 (jsonl을 list화)
- test\_data: 평가 데이터 (jsonl을 list화)
- evaluation\_metrics: 평가하고자 하는 평가 지표. 복수 선택 가능.  
선택 가능한 목록: ['classification\_micro\_F1', 'classification\_macro\_F1',  
'classification\_weighted\_F1', 'MSE', 'ROUGE-1', 'BLEU', 'bleurt', 'bertscore',  
'multi\_label\_classification\_micro\_F1', 'ROUGE-L']
- ratio: 평가 데이터 중 임의로 사용하고자 하는 비중 (예: 1일 경우 100% 사용)
- iteration: 평가 데이터에 대해 평가를 수행하는 횟수 (예: 10일 경우 10회 수행 후 평균을 계산)





## 제 8 장

# 평가 체계 홍보물





## 8. 평가 체계 홍보물

- 포스터의 경우 문치 캐릭터를 활용하여 친근감 있는 디자인으로 제작하였다. 또한 행사 정보를 쉽게 습득 할 수 있도록 우측 상단에 QR 삽입하였다. 한편 사이트에 맞는 온라인 배너 제작을 통해 홍보 효과를 노리는 한편 많은 사람들이 AI 말뚝에 쉽게 접근 가능하도록 하였다.

>

<표 101> 평가 체계홍보물 예시

## 인공 지능, 인간의 감정을 이해하고 이야기를 완성하다

# 2023 국립국어원 인공 지능 언어 능력 평가

# AI 말풍



접수 바로가기



접수 기간
**2023. 8. 21. (월) ~ 2023. 10. 20. (금)**

접수 방법
국립국어원 모두의 말풍치 <https://corpus.korean.go.kr> 에 접속

참가 자격
국어 정보 처리 또는 국어 인공 지능 관련된 개인 및 단체  
\*공제 제출은 가능하나, 공제 수산성 접수 없습니다.

평가 과제
**이마지 완성**

<p style="text-align: center;"><b>감정 분석</b></p> <p>인공 지능이 문장을 이해해서 주어진 대상에 대한 대화의 감정을 분석하는 과제</p>	<p style="text-align: center;"><b>이마지 완성</b></p> <p>인공 지능이 두 문장 사이에 논리적으로 연결되는 문장을 생성하는 과제</p>
---	--

주요 일정

2023. 8. 11. (금)	과제용 말풍치 (개별 · 시합용) 및 기출 모델 공개
2023. 8. 21. (월) ~ 10. 20. (금)	과제 선정자 접수 및 답안 제출(리얼타임으로 반영)
	<b>참가자 출품작 접수 및 수상작 결정</b>
2023. 11. 30. (목)	시상식

시상/해택

대상(1팀) 문화체육관광부 장관상 500만원	금상(2팀) 국립국어원장상 200만원	은상(2팀) 국립국어원장상 100만원	특별상 관련 기업 상장
--------------------------------	----------------------------	----------------------------	-----------------

\*제1차 평가 지원: 인공지능 기제 제공, 제1차 평가 및 제2차 평가 진행 시 대회 진행에서 부가적인 예산지출을 위해 내내 내용을 주목 하십시오

\*제2차 평가 지원: 제2차 평가 지원, 제2차 평가 진행 시 대회 진행에서 부가적인 예산지출을 위해 내내 내용을 주목 하십시오

문의
국립국어원 모두의 말풍치(<https://corpus.korean.go.kr>) > AI 말풍, 인공 지능 언어 능력 평가 > 인공 지능 과제 > 문의 사항 게시판 이용

한국어교육원  
국립국어원

인공 지능, 인터넷을 이해하고 아이기를 완성하다

## 2023 국립국어원 인공 지능 언어 능력 평가

참가 기간 2023. 8. 21. ~ 2023. 10. 20.

# AI 말평



한국어 교육원  
국립국어원

인공 지능, 인터넷을 이해하고 아이기를 완성하다

### 2023 국립국어원 인공 지능 언어 능력 평가

## AI 말평

참가 기간 2023. 8. 21. ~ 2023. 10. 20.

국립국어원 홈페이지 <https://koripi.kor.ac.kr>에서 접수

한국어 교육원  
국립국어원

인공 지능, 인터넷을 이해하고 아이기를 완성하다

### 2023 국립국어원 인공 지능 언어 능력 평가

## AI 말평

참가 기간 2023. 8. 21. ~ 2023. 10. 20.

국립국어원 홈페이지 <https://koripi.kor.ac.kr>에서 접수

한국어 교육원  
국립국어원

인공 지능, 인터넷을 이해하고 아이기를 완성하다

### 2023 국립국어원 인공 지능 언어 능력 평가

## AI 말평

참가 기간 2023. 8. 21. ~ 2023. 10. 20.

국립국어원 홈페이지 <https://koripi.kor.ac.kr>에서 접수

한국어 교육원  
국립국어원

인공 지능, 인터넷을 이해하고 아이기를 완성하다

### 2023 국립국어원 인공 지능 언어 능력 평가

## AI 말평

참가 기간 2023. 8. 21. ~ 2023. 10. 20.

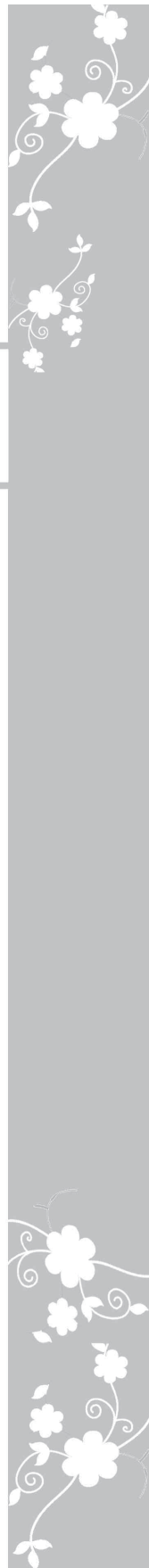
국립국어원 홈페이지 <https://koripi.kor.ac.kr>에서 접수





## 제 9 장

### 결론 및 기대 효과







## 9. 결론 및 기대 효과

본 과제는 인공지능의 한국어 처리 능력 평가가 가능한 평가 체계 운영을 목표로 6종 말뭉치 정비 및 과제 설계, 인공지능 언어능력 평가체계 운영과 전문가 자문, 그리고 개선점 제안 및 평가 체계 발전 방향을 진행하였다.

먼저 국립국어원 기구축 6종 말뭉치(감정 분석, 이야기 완성, 표 기반 문장 생성, 그림 기반 문장 생성, 함의 분석, 부적절성 말뭉치)를 바탕으로 평가 과제를 설계하고, 평가 과제용 말뭉치로 정비, 가공하여 기구축 말뭉치를 평가 말뭉치로 변환하였다. 이 중 감정 분석, 이야기 완성 과제는 경진대회 과제로 설계되었다. 경진대회 운영을 위해 운영 기준과 절차를 수립하고 민원 응대 체계를 마련하여 원활한 경진대회 운영이 가능하도록 하였다. 경진대회 진행 시에는 인간 평가, 모델 기술서, 발표 평가 등 다양한 평가를 통해 심사 결과의 신뢰도와 타당도를 높였다.

경진대회 외에 상시 과제 운영 절차도 수립하여 상시 과제 운영을 위한 절차 및 지침서를 마련하였다. 또한 상시 과제로는 함의 분석, 표의 일부분에 대한 해석 생성, 부적절성 문장에 대한 태도 탐지, 문자가 포함된 이미지 기반 문장 생성을 설계하여 상시 과제를 운영하였다.

평가 체계 운영 시에는 산업계와 학계의 인공지능, 언어처리, 평가 분야 전문가로 구성된 과제 위원회를 구성하여 경진대회와 상시 과제 운영 계획과 절차에 대한 자문을 받았다. 또한 평가 과제의 타당성, 평가 체계 방향성 전반에 대한 자문과 평가 체계 발전 방향에 대한 전문가별 자문을 받아 발전 방향 제안에 반영하였다.

한국어 인공 지능의 언어 능력 평가 발전 방향 제안을 위해서는 경진대회, 상시 과제 운영 결과를 정리하였으며, 초거대 언어 모델 시대의 연구 동향 분석, 한국어 인공지능 연구 동향, 한국어 인공지능 연구 수요 설문을 수행하여 평가 체계 발전 방향을 제안하였다.

본 과제는 인공 지능의 한국어 처리 능력 평가를 위한 평가 체계를 마련하고, 이를 위한 절차서와 지침서를 개발하여 신규 인공지능 언어능력 평가 체계 설계 시에도 활용할 수 있도록 한 것, 그리고 과제 평가용 말뭉치 정비를 위한 지침과 절차를 문서화하여 이후의 신규 과제 개발 및 말뭉치 정비에 참고할 수 있도록 한 것에 의의가 있다. 또한 실제로 평가 체계를 운영하고, 그 결과를 정리하여 발전 방향으로 제시함으로써 향후 인공지능의 한국어 능력 평가 체계 수립에 기여할 수 있도록 하였다.

향후에는 급속도로 발전한 LLM의 성능을 충분히 측정하기 위해서는 더욱 긴 길이의, 다양한 구조를 가진 텍스트를 평가 대상으로 삼아야 한다. 또한 단일 모달 외 멀티 모달(multi-modal)로서의 LLM을 측정할 수 있는 방법 역시 고민해야 하며, 한국어에 특화된 과제들을 개발하여 LLM의 한국어 이해 능력을 넓고 심도 있게 평가할 수 있도록 연구를 진행할 필요성이 있다.

## [참고 문헌]

- 이영희 외 (2022), 2022년 말뭉치 감정 분석 및 연구, 국립국어원 연구보고서, 국립국어원
- 조태린 외 (2022), 2022년 말뭉치 비윤리성 분석 및 연구, 국립국어원 연구보고서, 국립국어원
- Akhtar, S., Basile, V., & Patti, V. (2019). A new measure of polarization in the annotation of hate speech. In *AI\* IA 2019-Advances in Artificial Intelligence: XVIIIth International Conference of the Italian Association for Artificial Intelligence*, Rende, Italy, November 19-22, 2019, Proceedings 18 (pp. 588-603). Springer International Publishing.
- Alkomah, Fatimah, and Xiaogang Ma. "A Literature Review of Textual Hate Speech Detection Methods and Datasets." *Information* 13.6 (2022): 273.
- Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, & Dipanjan Das. 2020. ToTTo: A Controlled Table-To-Text Generation Dataset.
- Barnes, J., Oberländer, L., Troiano, E., Kutuzov, A., Buchmann, J., Agerri, R., ... & Velldal, E. (2022, July). SemEval 2022 Task 10: Structured Sentiment Analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 1280-1295).
- Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
- Chandra Bhagavatula et al. (2020). Abductive Commonsense Reasoning. Paper presented at International Conference for Learning Representation (ICLR).
- Charles Sanders Peirce. *Collected papers of Charles Sanders Peirce*, volume 5. Harvard University Press, 1965.
- Chatterjee, A., Narahari, K. N., Joshi, M., & Agrawal, P. (2019, June). SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 39-48).
- Crowdfunder(2016) / Mohammad, S. M., & Bravo-Marquez, F. (2017). Emotion intensities in tweets. *arXiv preprint arXiv:1708.03696*.

- De Bruyne, L., & De Clercq, O. (2022). Prospects for Dutch Emotion Detection: Insights from the new EmotionNL Dataset. *Computational Linguistics in the Netherlands Journal*, 11, 231-255.
- Del Vigna<sup>12</sup>, F., Cimino<sup>23</sup>, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017, January). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)* (pp. 86-95).
- Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*
- dos Santos, A., Júnior, J. D. B., & de Arruda Camargo, H. (2018). Annotation of a corpus of tweets for sentiment analysis. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24-26, 2018, Proceedings 13* (pp. 294-302). Springer International Publishing.
- Gholipour Shahraki, A. (2015). *Emotion Mining from Text.*, Master's thesis, University of Alberta, USA
- Gref, M., Matthiesen, N., Venugopala, S. H., Satheesh, S., Vijayananth, A., Ha, D. B., ... & Köhler, J. (2022). A study on the ambiguity in human annotation of german oral history interviews for perceived emotion recognition and sentiment analysis. *arXiv preprint arXiv:2201.06868*.
- Huang, C., Trabelsi, A., Qin, X., Farruque, N., & Zaïane, O. R. (2019). Seq2emo for multi-label emotion classification based on latent variable chains transformation. *arXiv preprint arXiv:1911.02147*.
- Jason Obeid, Enamul Hoque. 2020. Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model.
- Jiawen Zhu, Jinye Ran, Roy Ka-wei Lee, Kenny Choo, Zhi Li. 2021. AutoChart: A Dataset for Chart-to-Text Generation Task.
- Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2020, May). Evaluating aggression identification in social media. In *Proceedings of the second workshop on trolling, aggression and cyberbullying* (pp. 1-5).
- Lee, Jean, et al. "K-MHaS: A Multi-label Hate Speech Detection Dataset in Korean Online News Comment." *arXiv preprint arXiv:2208.10684* (2022).
- Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). Dailydialog: A

- manually labelled multi-turn dialogue dataset. arXiv preprint arXiv:1710.03957.
- Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Koreeda, Y. (2022). Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.
- Liu, C., Osama, M., & De Andrade, A. (2019). DENS: A dataset for multi-class emotion analysis. arXiv preprint arXiv:1910.11769.
- Liu, V., Banea, C., & Mihalcea, R. (2017, October). Grounded emotions. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 477-483). IEEE.
- Miestamo, M., Karlsson, F., & Sinnemäki, K. (2008). Language complexity. *Language Complexity*, 1-374.
- Mohammad, S. M., & Bravo-Marquez, F. (2017). WASSA-2017 shared task on emotion intensity. arXiv preprint arXiv:1708.03700.
- Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4), 480-499.
- Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018, June). Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 1-17).
- Mollas, Ioannis, et al. "ETHOS: a multi-label hate speech detection dataset." *Complex & Intelligent Systems* 8.6 (2022): 4663-4678.
- Mossie, Z., & Wang, J. H. (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3), 102087.
- Mulki, H., Haddad, H., Ali, C. B., & Alshabani, H. (2019, August). L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the third workshop on abusive language online* (pp. 111-118)
- Nasrin Mostafazadeh et al. (2016). A Corpus and Evaluation Framework for Deeper Understanding of Commonsense Stories. arXiv preprint arXiv:1604.01696.
- Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2019). Adversarial NLI: A new benchmark for natural language understanding.

- arXiv preprint arXiv:1910.14599.
- Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, Amanpreet Singh. 2020. TextCaps: a Dataset for Image Captioning with Reading Comprehension: Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.
- Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., ... & Cho, K. (2021). Klue: Korean language understanding evaluation. arXiv preprint arXiv:2105.09680.
- Patwa, P., Aguilar, G., Kar, S., Pandey, S., Srinivas, P. Y. K. L., Gambäck, B., ... & Das, A. (2020). SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets. SemEval@ COLING, 774-790.
- Pérez, Juan Manuel, et al. "Assessing the impact of contextual information in hate speech detection." arXiv preprint arXiv:2210.00465 (2022).
- Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. Language Resources and Evaluation, 55, 477-523
- Qian, Jing, et al. "Learning to decipher hate symbols." arXiv preprint arXiv:1904.02418 (2019).
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1), 5485-5551.
- Sabat, Benet Oriol, Cristian Canton Ferrer, and Xavier Giro-i-Nieto. "Hate speech in pixels: Detection of offensive memes towards automatic moderation." arXiv preprint arXiv:1910.02334 (2019).
- Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., ... & Rush, A. M. (2021). Multitask prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207.
- Schuff, H., Barnes, J., Mohme, J., Padó, S., & Klinger, R. (2017, September). Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 13-23).

- Sharma, C., Bhageria, D., Scott, W., Pykl, S., Das, A., Chakraborty, T., ... & Gamback, B. (2020). SemEval-2020 Task 8: Memotion Analysis--The Visuo-Lingual Metaphor!. arXiv preprint arXiv:2008.03781.
- Sharma, H. K., & Kshitiz, K. (2018, June). Nlp and machine learning techniques for detecting insulting comments on social networking platforms. In 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE) (pp. 265-272). IEEE.
- Sharma, R., Allen, J., Bakhshandeh, O., & Mostafazadeh, N. (2018). Tackling the Story Ending. Biases in The Story Cloze Test. ACL.
- Tafreshi, S., & Diab, M. (2018, May). Sentence and clause level emotion annotation, detection, and classification in a multi-genre corpus. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
- Toshniwal, S., Shi, H., Shi, B., Gao, L., Livescu, K., & Gimpel, K. (2020). A cross-task analysis of text span representations. arXiv preprint arXiv:2006.03866.
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. Advances in neural information processing systems, 32.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.
- Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., ... & Shen, X. (2022, December). Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing(pp. 5085-5109).
- Yang, Kichang, Wonjun Jang, and Won Ik Cho. "APEACH: Attacking Pejorative Expressions with Analysis on Crowd-Generated Hate Speech Evaluation Datasets." arXiv preprint arXiv:2202.12459 (2022).
- Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). SWAG: A Large-Scale Adversarial Dataset. for Grounded Commonsense Inference. EMNLP.
- Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J.

(2022). Large language models are human-level prompt engineers. arXiv preprint arXiv:2211.01910.



## [부록1] LLM 연구 동향 조사

ChatGPT 등장 이후로 자연어 처리에 대한 연구나 산업계 판도가 완전히 초거대 언어 모델(이하 LLM)로 전환되었으며, 본 보고서에서는 초거대 언어 모델의 연구 동향을 분석함으로써 향후 평가 체계가 나아갈 방향에 대해 고찰하였다.

LLM은 기존 언어 모델에 비해 훨씬 큰 크기의 모델들이라 할 수 있으며, 모델 크기와 성능이 비례하는 언어 모델의 특성상 개별 과제들에 대한 LLM의 성능은 비약적으로 증가하였다. 이에 따라 LLM의 성능을 종합적으로 측정할 수 있는 벤치마크 필요성이 대두되었으며, 실생활에서 일어나는 상황을 다루는 ‘시나리오 중심’의 전환이 일어나고 있다. 때문에 이를 뒷받침할 수 있는 다양한 데이터 세트를 마련할 필요성 역시 커지고 있다. 평가 측면에서도 실생활의 여러 상황들을 평가하게 되었기 때문에 다양한 정량적 평가 지표에 대한 고찰이 필요하며, 실제 사용자의 수요를 만족시킬 수 있는지, 그리고 인간과의 일치(alignment)를 보이는지 등을 측정할 수 있는 정성적 지표를 마련해야 한다. 더 나아가 초거대 언어 모델의 효율적인 학습 혹은 추론(inference)을 평가할 수 있는 ‘효율성’ 지표도 눈여겨보아야 할 지표이다.

한편 기존 영어, 중국어와 같은 주류 언어들로 학습된 LLM 특성상 기존의 벤치마크들 역시 주류 언어로 개발되는 특성들이 있었다. 이에 기존 언어들로 개발된 벤치마크들의 시나리오들을 분석하여 분류 체계(taxonomy)로 제시하였으며, 국립국어원이 보유한 언어 자원들을 해당 분류 체계(taxonomy)로 분류하여 향후 LLM 학습 및 성능 평가를 위한 데이터 세트 개발 기초 자료를 제시하였다.

결론적으로 향후 국립국어원의 인공지능 언어능력 평가 체계는 실생활 시나리오를 포함한 **종합적 벤치마크**로 나아가야 하며, 이를 위해 다양한 데이터 세트 개발과 여러 정량적, 정성적 지표들을 마련할 필요성이 있다 이때 개발의 기준으로서는 주류 언어 벤치마크로부터 개발한 **분류 체계**를 활용할 수 있으며, **한국어 벤치마크**를 위한 **지속적인 과제 발굴 및 평가 지표 마련, 데이터 세트 개발 노력**이 필요하다.

### 1. 평가 체계(벤치마크, Benchmark) 연구 현황

영어로 훈련된 LLM의 영향으로 다수의 벤치마크가 영어를 중심으로 연구되고 있으며 종합적인 LLM의 성능을 보기 위한 벤치마크/리더보드와 특정 도메인에 대한 성능 평가를 진행하기 위한 벤치마크/리더보드 두 가지로 양분되어 연구가 진행되고 있다. 2023년 상반기보다 하반기에 다수 공개가 되었으며, 이를 통해 벤치마크의 개발이 빠르게 이루어지고 있음을 알 수 있다.



<표 102> 2023년 영어 리더보드/ 벤치마크 목록

공개 시기	리더보드/벤치마크	비고
2023 상반기	Open-llm-leaderboard	오픈소스 커뮤니티에서 진정한 진전이 이루어지고 있는지, 어떤 모델이 최신 기술인지 용이하게 가려내기 위한 리더보드
	JEEBench	LLM의 문제 해결 능력을 평가
	HIPPO (High-level Interlingual Performance Proximity Optimized)	번역의 문법적 정확성과 의미적 친밀성 측면에서 LLM의 효능을 평가
2023 하반기	instruction_following_eval(IFEval)	모델이 지침을 따르는 능력 평가
	Hallucination Leaderboard	문서 요약 시 모델이 허위 정보를 얼마나 만들어내는지 평가
	PsyBench	심리학 영역에서 모델의 성능 종합적 평가
	LLM data Contamination	오픈 LLM, 독점 LLM 모두에 적용될 수 있는 평가 데이터 세트 오염 탐지
	LLMEval	영어 LLM평가 벤치마크 개발
	JudgeLM	고품질 데이터세트를 활용하여 LLM을 파인튜닝하여 구축
	PandaLM	LLM의 hyperparameter를 평가하고 최적화하기 위해 구축
	MT-bench	LLM-as-a-judge' VS 'human preference' 일치 여부를 확인하기 위함
	Chatbot Arena	클라우드소싱 방식으로 익명의 무작위 대전을 제공하는 대규모 언어 모델(LLM)을 위한 벤치마크

특히 2023년 하반기에 공개된 Zheng, Lianmin, et al(2023)의 MT-bench는 LLM을 종합적으로 평가하는 대표적인 벤치마크이다. 이 연구에서는 LLaMA-13B와 Vicuna-13B를 파인튜닝하여 기존 벤치마크만을 사용하여 평가할 때 인간 평가와 불일치하다는 것을 입증하였다. 고품질 다중 라운드 질문으로 구성된 벤치마크로 글쓰기, 역할 놀이(role-play), 추출, 추론, 수락, 수학, 코딩, 지식(STEM), 지식2(인문/사회과학) 카테고리를 대상으로 하여 종합적으로 모델을 평가한다.

영어에 대한 LLM 뿐만 아니라 중국어에 대한 벤치마크도 빠른 속도로 개발이 되고 있다. 중국어는 통합적으로 모델을 평가하는 벤치마크가 다수 있다. 특히 C-eval의 경우 'C-eval hard'를 별도로 구축하여 고급 추론 능력을 평가하고자 더 어려운 벤치마크를 함께 구축한 점이 눈에 띄는 부분이다. 이 외에도 C-eval은 중국 환경에서 사용하기 위한 한정적인 도메인의 성격을 띄고 있는데, 중국 사용자의 관심사에 대한 지식, 중국 문

화, 역사, 법률과 같은 중국 사회 고유한 역량에 대한 평가를 목적으로 한다. 데이터는 52개 카테고리 및 4개의 난이도에 걸쳐 13,948개의 객관식 문항으로 구성되어 있다.

<표 103> 2023 중국어 리더보드/벤치마크 목록

공개시기	리더보드/벤치마크	비고
2023 상반기	C-EVAL	중국어 기초 모델의 고급 기능을 평가. 중국 특화 데이터 세트 구축.
2023 하반기	ZhuJiu	종합적인 중국 벤치마크 제안. 영어로도 동등한 평가 가능
	CLEVA	중국 LLM을 총체적으로 평가. prompt를 표준화하여 LLM 평가의 비교 가능성을 높임

영어, 중국어뿐만 아니라 한국어로 구축된 벤치마크 및 리더보드도 발표되고 있다. 가장 최근에 개발된 대표적인 한국어 LLM 리더보드로는 Open-Ko-LLM 리더보드(Leader Board)가 있다. 이는 한국어와 한국 문화 특성이 반영된 모델 평가를 위해 운영되는 리더보드이다. 리더보드 순위를 높이기 위해 훈련 과정에 평가 세트를 포함하는 등 부정 사용을 잡아내는 절차를 포함하고 있다. 리더보드 과제 유형은 Ko-HellaSwag (업스테이지 제공, 기계 번역), Ko-MMLU(업스테이지 제공, 사람 번역 및 변형), Ko-Arc(업스테이지 제공, 사람 번역 및 변형), Ko-Truthful QA(업스테이지 제공, 사람 번역 및 변형), Ko-CommonGen V2 (고려대학교 NLP&AI 연구실에서 제공, 자체 제작)이 있다. 이 외에도 한국어 맥락에서 언어 모델의 숙련도를 평가하기 위해 개발된 ‘해레(Hae-rae) 벤치마크가 있다. 어휘, 역사, 일반지식 등을 포함하여 6개의 포괄적인 과제로 구성되어 한국어 말뭉치에만 있는 정보를 이해하고 기억하는 능력을 평가하는 방식으로 진행된다.

이 외에도 소수 언어 벤치마크로 스칸디나비아 언어에 대한 벤치마크 ScandEval이 있다. 본 벤치마크는 언어적 수용성과 QA 데이터는 새로이 구축하였고 스칸디나비아 본토 언어 간의 언어 전이, 스칸디나비아 본토언어와 스칸디나비아 섬의 언어에 대한 연구를 함께 진행하였다. 이처럼 소수 언어에서도 언어학이라는 한정적인 도메인에 대해 벤치마크가 개발되고 있다.

## 2. LLM 연구의 흐름 변화

LLM에 대한 기술의 발전과 함께 전체적인 연구의 흐름도 변화하고 있다. 기존에는 단일 과제(task)에 대해 연구가 이루어 졌다면 2023년에는 시나리오 중심으로 변화하고 있다. 시나리오는 과제의 상위 개념으로써 2023년 공개된 다수의 벤치마크에서 사용되고

있다. 예를 들어 Percy Liang et al (2022)의 HELM 리더보드는 116개의 시나리오를 제안한다. 질문답변/ 정보검색/ 요약/ 감성분석 /독성감지/ 텍스트분류 / 언어 /지식/ 추리/ 능력/ 피해/ 객관식 전략 을 비롯한 116개의 시나리오를 적용하고 있다. 중국어로 구축된 FlagEval은 자연어 처리(NLP), 컴퓨터 비전(CV), 오디오, 멀티 모달 4가지의 시나리오를 적용하고 있다.

또한 기존의 언어 모델(이하 LM)에서 LLM으로 언어 모델의 크기가 확대되고 있다. 이와 함께 여러가지 요소가 변화되었다. LM은 모델 트레이닝 아키텍처도 인코더(encoder) 중심으로 생성이 불가능하여 이해 중심으로 활용되었다. 모델 크기도 하이퍼파라미터 백만(million) 수준으로 한정적이기에 데이터 셋의 규모도 작고 도메인(domain)의 다양성도 부족한 편이었다. 평가 방식도 훈련(training) 이후 정량적인 평가가 주로 이루어졌으며 벤치마크도 과제 중심 평가로 이루어졌었다. 전체적으로 LM은 규모가 작고 한정적이어서 task를 이해하는데 국한되는 점이 있었다. 하지만 LLM으로 변화되면서 가장 큰 변화는 생성이 가능해졌다는 점과 모델의 크기도 하이퍼파라미터 조(billion) 단위로 대규모로 변화된 부분이다. 이에 맞추어 적용 가능한 데이터 셋의 규모도 대규모로 훈련이 가능하며 다양한 도메인을 적용할 수 있게 되었다. 특히 평가와 관련하여 LM은 훈련 이후 정량적 평가가 가능했으나 LLM은 훈련 도중 정성적 평가가 가능하며 벤치마크도 과제 중심에서 시나리오 중심으로 이동하였다. 정성적 평가가 이루어지는 LLM으로의 이동은 곧 평가에 인간의 개입이 필요함을 의미한다. 이에 뒤에서 평가의 흐름을 함께 살펴볼 예정이며 본 보고서에서도 시나리오 중심으로 벤치마크를 분석하고 발전 방향을 제안한다.

2023년에 공개된 벤치마크를 살펴보면 LLM을 대상으로 평가하는 것이기 때문에 데이터셋의 규모가 다양하다. 상반기보다 하반기에 공개된 데이터의 수가 더 많으며 데이터의 규모는 최소 40개의 질문부터 30만 개 이상까지 다양하다.

<표 104> 2023 리더보드/벤치마크 공개 데이터 세트 목록

공개시기	평가체계 이름	데이터 세트 규모
2023 상반기	c-eval	13,948 개의 객관식 문제로 구성
	JEEBench	515개의 문제
	The Rakuda Ranking of Japanese AI	일본어질문 40개
2023 하반기	instruction_following_eval (IFEval)	500건의 프롬프트
	hallucination Leaderboard	source-summary 구조 데이터 세트, 약 9,000건
	CLEVA	84개 데이터 세트에서 37만 개의 테스트 인스턴스 + 증강 후 9백만 개의 쿼리 수행
	safety benchmark	11,435개 객관식 문항
	judgeLM	모델 학습용 dataset

		JudgeLM-100K(train:100,000건, vali: 5,000건)
	pandaLM	모델 학습용 dataset 300,000개
	LLMEval	2,553 samples

대부분의 벤치마크에서 데이터 세트는 공개하고 있으나 일부 데이터 세트는 선택적 공개 혹은 미공개이다. 공개되지 않은 데이터 세트 목록은 다음과 같다.

<표 105> 2023 리더보드/벤치마크 미공개 데이터 세트 목록

공개시기	평가체계 이름	데이터 세트 미공개 여부
2023 상반기	Chatbot Arena	미공개
	Open LLM Leaderboard (Humans and GPT4 evaluations)	미공개
	Hae-rae	선택적 공개(메일로 신청)
2023 하반기	PsyBench	논문만 공개된 상태
	Open-Ko-LLM LeaderBoard	독점(비공개) -> 평가 과정에서만 사용 가능
	ZhuJiu	미공개
	LLM data Contamination	논문만 공개된 상태

미공개 데이터 세트도 존재하고 있으나, 대체로 데이터의 규모가 다양하고 커졌고 데이터의 형식도 다양해지고 있으며 대부분의 데이터 세트가 공개를 함께 하는 흐름으로 진행되고 있다.

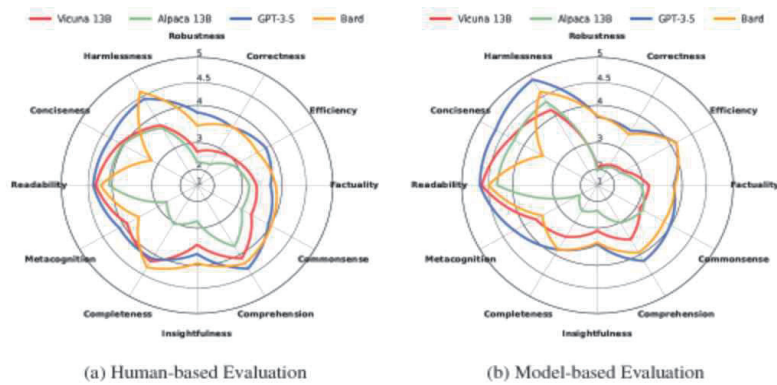
### 3. 평가 지표

2023년 모델 평가에 사용된 평가 지표의 흐름은 새로운 정량적 평가 방법이 등장한다는 점이다. 기존처럼 정량적 평가만 진행하거나 정량적 평가와 정성적 평가를 함께 진행하는 연구도 있다. 정량적 평가만으로는 생성 결과물의 품질을 명확히 평가하기 어려워 정성적 평가를 추가로 진행하는 편이다. 특히 정량적 평가와 정성적 평가를 함께 사용하는 연구의 경우 정성적 평가의 비율이 높은 편이다. Changrong Xiao et al. (2023)은 LLM의 생성 능력을 평가하고자 정량적 평가로는 NLL, SMOG, Flesch, TTR, Rep를 사용하였고 인간의 정성 평가는 다차원 품질 점수(가독성, 정확성, 일관성, 참여도, 전체적인 품질)와 쌍별 비교(모델 생성 점수, 주제 일관성 점수, 적합성 점수)를 진행하였다. 수치만으로 확인하기 어려운 평가 항목에 대해서는 인간 평가를 진행하여 평가의 신뢰성을 뒷받침한다.

이에 정량적 평가 지표도 계속해서 연구되고 있는 것으로 보인다. Yang, Liu et al.(2023)의 G-EVAL은 자연어 생성시스템으로 생성된 텍스트의 품질 평가(창의성, 다양

성)에 대한 기존 metric 적용의 어려움 극복을 위해 생각의 사슬(Chain-of-Thought, CoT)를 적용한 평가 프레임 워크이다. G-EVAL은 텍스트 요약과 대화 생성이라는 두 가지 NLG 작업에 대한 광범위한 실험을 수행한 결과 타 평가 지표(ROUGE-2, BERTScore, MoverScore etc) 대비 인간과 더 높은 일치도를 달성하였다. 또 다른 정량 평가지표로 Ye, Seonghyeon, et al.(2023)의 FLASK 인간 및 모델 기반 평가 프로토콜을 통해 세분화된 평가를 진행하는 연구가 있다. 논리적 사고(Logical Thinking:: Logical Correctness, Logical Robustness, Logical Efficiency), 배경 지식(Background Knowledge: Factuality, Commonsense Understanding), 문제 해결(Problem Handling: Comprehension, Insightfulness, Completeness, Metacognition), 사용자 일치(User Alignment: Conciseness, Readability, Harmlessness) 총 12가지 평가 체계를 활용하여 세분화된 평가를 제안하며 레이더 맵을 함께 제공하고 있다.

다만 다양한 LLM 과제에 가장 효과적인 평가 지표가 무엇인지에 대해 추가적인 연구를 수행할 필요가 있어 보인다. 일례로, Nimah et al. (2023)에서는 자동 평가 결과와 인간 평가 결과의 correlation 점수가 평가 도구로서의 평가 지표의 효과성을 의미하는 것은 아니라고 여겨 자동 평가 지표의 선호도 체크리스트를 진행하였다. [1] Transfer Experiment : 도메인 내/외의 사용 사례에서 자동지표와 인간 일치 평가 지표(human aligned metric)의 상관관계가 일관되게 유지되는지, [2] System-level Evaluation : human aligned metric이 LLM의 성능 평가에 효과적인지?, [3] System-level Preference : human aligned metric과 자동 평가 지표가 언어 모델의 순위를 비슷하게 매기는지, [4] Aspect-level Evaluation : human aligned metric이 human-like quality의 품질을 식별해내는데 더 효과적인지, [5] Aspect-level Preference : human aligned metric과 자동 평가 매트릭이 결과물의 품질을 비슷하게 평가해내는지 여부에 대해 살펴보았다. 그 결과 human aligned metric이 반드시 자동 평가 지표보다 더 낫다고 할 수 없고 특히 llm system 평가에서는 자동 평가 지표가 더 효과적이라고 결론을 내렸다.



[그림 51] Ye, Seonghyeon, et al. (2023)의 "Flask: Fine-grained language model evaluation based on alignment skill sets. 평가 지표에 따른 성능 레이더맵

이처럼 대부분의 연구에서 인간의 평가를 대체하기 위한 정량적 지표의 필요성을 언급하며 이에 맞는 지표를 개발하는 흐름이다. 다만 태스크에 따라 적용 가능한 평가 방법이 다르다는 것을 염두할 필요가 있어 보이며 이에 인간 평가가 적합한 과제가 있다는 것도 고려하여 인간평가의 발전도 함께 추가적으로 연구가 진행되어야 한다. 2023년 제안된 평가 지표를 정리하면 다음과 같다.

<표 106> 2023 제안된 평가 방법

평가 METRIC	설명
G-EVAL	자연어 생성 시스템으로 생성된 결과물의 품질 평가에 CoT를 적용하는 프레임 워크.
FLASK	instruction에 메타데이터 주석을 하고, skill, instance별 rubric을 기반으로 평가를 수행하여 skill, domain, level에 따라 LLM 기능에 대한 포괄적인 분석 제공
Exact Match	레벨1~3까지의 질문을 통해 LLM의 성능 평가
JudgeLM	agreement, precision, recall, F1-score 를 활용하여 성능 평가
PandaLM,	LLM의 hyperparameter를 평가하고 최적화하기 위해 accuracy, precision, recall, F1-score 사용.
ChatEval	인건비와 평가 시간의 효율을 도모하기 위해 인간의 대안으로 LLM의 잠재력 탐구하고자 인간 평가자의 주석을 사용한 평가 지표. natural, coherence, engagingness, groundedness, understandable.
LLMEval	영어 LLM 벤치마크로 summarization, Logical Reasoning, Semantic Understanding, Knowledge QA, Multilingual, Harmlessness, Dialogue, Text Composition 사용
LLM data Contamination	평가 데이터 세트 오염 탐지를 위한 평가 지표. BM25 score, SacreBLEU, BLEURT, semantic similarity

#### 4. 발전 가능한 언어 자원

대다수의 벤치마크에서 질문을 활용한 데이터, 프롬프트, ChatGPT를 활용하여 데이터 세트를 구축하거나 공공 데이터 세트를 사용하였다. 기 구축된 말뭉치를 사용하는 것이 아니라 새로 데이터 세트를 구축하는 방식으로 진행된 벤치마크가 많은 편이나, 현재 국어원에서 구축되어 있는 자원 중 참고가 가능한 데이터 세트는 Hallucination Leaderboard와 SafetyBench가 있다. Hallucination Leaderboard는 원문-요약 구조로 된 데이터 세트를 사용하였고 이는 국립국어원 말뭉치 중 문서요약말뭉치와 유사한 성격을 가져 이를 참고하여 발전시킬 수 있을 것으로 생각된다. SafetyBench는 객관식 문항으로 구성되어 있지만 안전 문제와 관련한 문항으로 이는 비윤리적 표현 말뭉치 지침 개선을 응용하여 적용이 가능할 것으로 보인다. 다만 벤치마크에서 데이터 세트를 자체 구축 혹은 증강의 방법을 사용하고 있는데 이는 한편으로 기구축된 언어 자원을 활용하여 새로운 평가 데이터 세트를 만들 수 있는 것을 의미한다. 본 장에서는 국립국어원에서 현재 구축되어 있는 언어 자원을 시나리오에 따라 발전시켜 적용할 수 있는 방향을 제안, 정리하였다.

이를 위해 시나리오를 우선 본 보고서에서 설정하였다. 자연어 처리 분야의 특성을 고려하여 언어(Language), 지식(Knowledge), 인간 일치(Human value), 멀티 모달(Multimodal)로 구성하였다. 자세한 시나리오 분류 체계는 다음과 같다.

<표 107> 시나리오 분류 체계 정의

대분류	소분류	비고
Language	NLU	비정형 텍스트 데이터 이해, 정형 텍스트 데이터 이해
	NLG	비정형/정형 텍스트 데이터 생성, 창의적 글쓰기(이야기 짓기 등)
	linguistic 층위의 Tasks	pos, parsing, SRL, minimum pairs ..
Knowledge	general knowledge	일반적인 사실, 상식에 대한 hallucinations, commonsense reasoning
	specific knowledge	math, code 등 STEM 지식, 전문분야(의료, 법률 등) 지식을 다루는 경우
Human value		harmfulness, toxicity, ethics, bias ..
Multimodal		텍스트가 아닌 음성, 이미지, 비디오 등을 다루는 데이터 세트

우선, 2023년 시나리오에 따라 공개된 리더보드/벤치마크의 현황을 파악하고 연구가 집중된 분야와 추가적으로 개발이 필요한 부분을 살펴보았다. 시나리오 분류 체계를 활용하여 2023년 공개된 벤치마크를 분류하면 다음과 같다.



<표 108> 분류 체계에 따른 2023 공개 리더보드/ 벤치마크 분류

시나리오	Benchmark/Leaderboard
Human value	Leaderboard) Open LLM Leaderboard (Humans and GPT4 evaluations) Benchmark) ZhuJiu
Language-Task	Leaderboard) CLEVA, Open LLM Leader board, Benchmark) ScandEval, LLMEval
Language-NLG	Leaderboard) Hallucination Leaderboard, The Rakuda Ranking of Japanese AI, CLEVA, Chatbot Arena, Open LLM Leader board, Benchmark) Instruction_following_eval(IFEval), HIPPO (High-level Interlingual Performance Proximity Optimized)
Language-NLU	Leaderboard) CLEVA, Open LLM Leader board, Benchmark) Instruction_following_eval(IFEval), ZhuJiu, LLM data Contamination, JEEBench,
Knowledge-General	Leaderboard) Open-Ko-LLM LeaderBoard, Open LLM Leader board, C-EVAL, CLEVA Benchmark) Hae-rae, ZhuJiu, ScandEval
Knowledge-Specific	Leaderboard) C-EVAL(hard), CLEVA Benchmark) ZhuJiu, PsyBench, SafetyBench

LLM의 성능 평가에 사용된 시나리오를 살펴보면 human value에 대한 평가가 저조한 편이다. 이는 정성적으로 인간 평가가 필요한 부분으로 시간, 비용의 측면을 고려하여 쉽게 적용이 어려운 평가 지표이기 때문인 것으로 보인다. 앞서 정의한 시나리오를 기준으로, 기구축된 국립국어원의 언어 자원을 활용하여 평가체계에 적용하는 방안을 함께 제안한다.

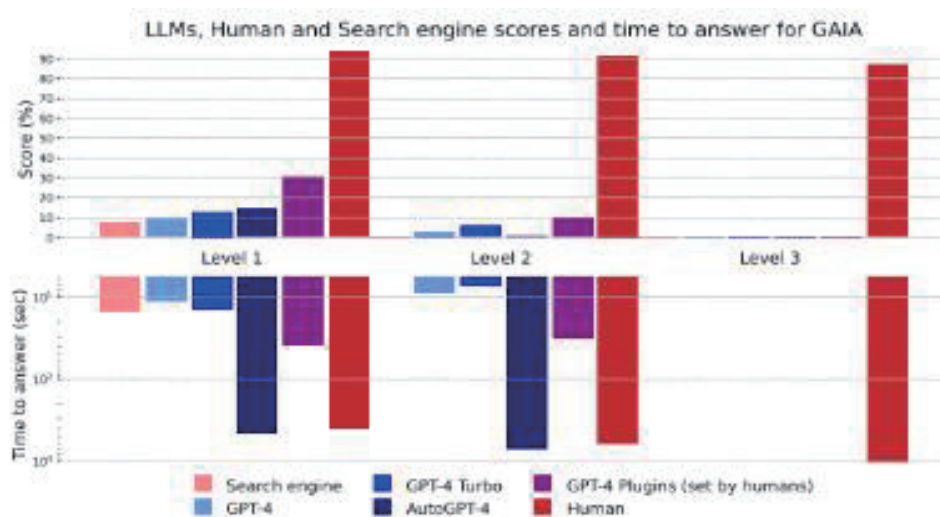
<표 109> 분류 체계에 따른 국립국어원 언어자원 발전 방향 제안

번호	시나리오	언어자원명	해외 사례 여부	활용 가능성 여부	활용방안
1	Task	형태분석말뭉치		○	한국어 형태소 분석 평가 활용 가능
2	Task	구문분석말뭉치		○	한국어 구문 분석 평가 활용 가능
3	Task	상호참조해결말뭉치2019		○	한국어 상호참조 해결 평가 활용 가능, 문서 혹은 문장 내부 동일 대상 지시 표현 관련 평가 가능
4	NLG	문서요약말뭉치	Hallucination Leaderboard	○	주제문에 따른 요약문 생성 평가 활용 가능
5	NLG	무형대용어복원말뭉치2020		○	적절한 단어 생성 여부 평가 활용 가능
6	NLU	문법성판단말뭉치		○	문법성 평가 활용 가능
7	NLU	맞춤법교정말뭉치2022		○	맞춤법, 교정 평가 가능



8	NLU	맞춤법교정말뭉치2021		○	맞춤법, 교정 평가 가능
9	Knowledge	개체명사전2022		○	개체명 관련 지식 평가 가능
10	Knowledge	개체명사전2021		○	
11	Knowledge	개체명분석말뭉치개체연결2022		○	
12	Knowledge	개체명분석말뭉치개체연결2021		○	
13	Knowledge	개체명분석말뭉치2022		○	
14	Task	2022 인공지능언어능력평가말뭉치: ABSA		○	모델 감성 이해 능력 평가
15	Task	한국어-힌디어병렬말뭉치2021		○	기계 번역 능력 평가 활용 가능
16	Task	한국어-필리핀타갈로그어 병렬말뭉치2021			
17	Task	한국어-태국어병렬말뭉치2021			
18	Task	한국어-캄보디아크메르어 병렬말뭉치2021			
19	Task	한국어-인도네시아어병렬말뭉치2021			
20	Task	한국어-우즈베크어병렬말뭉치2021			
21	Task	한국어-베트남어병렬말뭉치2021			
22	Task	한국어-러시아어병렬말뭉치2021			
23	Task	2023년한국어-외국어병렬말뭉치구축			
24	NLG	추론-확신성분석말뭉치2021		○	의미가 유사한 문장 생성 및 평가 관련 활용 가능
25	NLG	추론-확신성분석말뭉치2020		○	
26	NLG	유사문장말뭉치		○	유사한 문장 생성 및 평가 관련 활용 가능
27	Task	의미역분석말뭉치		○	의미역 관계 QA 평가 관련 활용 가능
28	NLG	어휘관계자료: NIKLex		○	어휘 관계 자료를 이용하여 문장 생성 평가 활용 가능
29	Human Value	2022년비윤리적표현말뭉치연구분석및구축	SafetyBench	○	비윤리적 표현에 대한 인간 평가 활용 가능
30	NLG	2022년이야기완성평가말뭉치연구분석		○	이야기 생성 능력 평가 활용 가능
31	NLU	2024년도국어생활자료 정비및온라인상답		○	한국어 문법 및 맞춤법 오류 평가 활용 가능
32	Multimodal	2023년 한국수어 말뭉치 구축		○	한국어-수어 이미지와 일치 평가 활용 가능
33	Multimodal	2023년 한국어-한국수어 병렬 말뭉치 구축		○	

위의 구축된 자원을 활용하여 인간 평가를 강조하는 평가 체계를 구축할 수 있을 것으로 생각된다. 또한 Human value를 다루는 벤치마크/리더보드가 적은 것을 통해 이를 발전 시키면 경쟁력을 가질 수 있을 것이다. Jones, C., & Bergen, B. (2023)에서는 튜링테스트로 인간과 모델을 비교한다. 튜링 테스트를 통해 인간이 인공지능이라 판단한 근거가 자연스러운 표현, 문장 구조, 이상한 말투에서 비롯함을 확인하였다. 이를 통해 인간과 인공지능을 비교할 때 분명한 차이가 있음을 알 수 있다. 인간 평가를 통해 인공지능의 부족함을 구별할 수 있을 것이다. 한국어 문법에 대하여 자연스럽게 사용을 하였는지, 자연스러운 문장을 만들기 위해 비문을 찾아낼 수 있는지, 종결 어미가 올바르게 쓰였는지 등을 한국어 특화 평가체계에 적용하는 방법으로 응용이 가능할 것으로 생각된다. 뿐만 아니라 Mialon, G et al.(2023)에서 Exact match는 3단계의 평가 과정을 가진다. 단계가 높아질수록 인간 수준에 가까운 능력이 필요한데, LLM 모델은 단계가 높아질수록 성능 저하를 보였다.



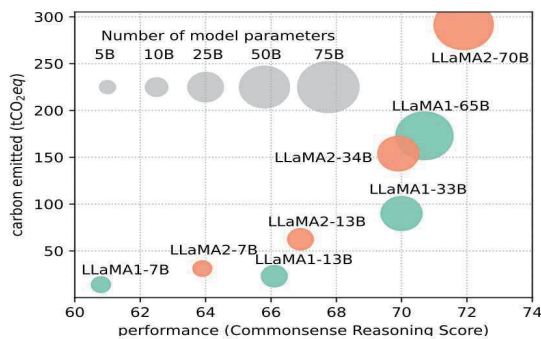
[그림 52] GAIA: a benchmark for General AI Assistants의 성능 비교 그래프

레벨 1에 해당하는 질문은 일반적으로 도구가 필요하지 않거나, 최대 하나의 도구가 필요하지만 5단계를 넘지 않는 수준, 레벨 2 질문은 일반적으로 5~10개의 더 많은 단계를 포함하며 문제 해결을 위해 다양한 도구를 결합해야 하는 수준이며, 레벨 3은 거의 완벽한 인간 수준의 능력이 있어야 하는 수준으로, 긴 작업을 수행하기 위해 다양한 도구를 사용하는 것은 물론 외부로부터 정보를 습득해야 하는 능력이 필요하다. 본 평가에서 가장 높은 정확도를 기록한 것은 GPT4이지만 인간의 점수와 높은 차이가 있어 아직 인간만이 해결할 수 있는 문제가 있다는 것을 의미한다. 정량적인 평가 방법도 중요하나 인간 평가가 LLM의 성능 평가에 유의미한 평가임을 확인할 수 있다.

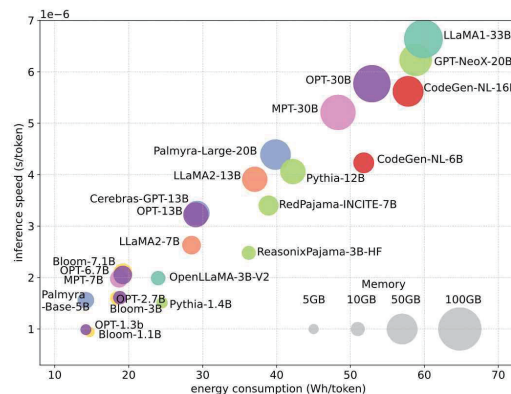
## 5. 효율성

최근 LLM의 프레임 워크는 적은 GPU의 사용을 통해 메모리 처리의 효율성과 최적화를 이루어 내는 것이다. 단순 성능의 향상을 위한 효율성, 최적화에 국한된 것이 아니다. LLM의 규모가 커질수록 추론 과정에서 많은 메모리 사용량, 에너지 소비량이 발생하게 된다. 이에 따라 탄소배출량도 증가하기 때문에 환경적 측면에서도 적은 GPU를 사용하는 것에 대한 목소리가 꾸준히 나오고 있다.

<표 110> 효율성 그래프



[그림 53] 다양한 규모의 학습용 LLM의 성능과 탄소 배출량

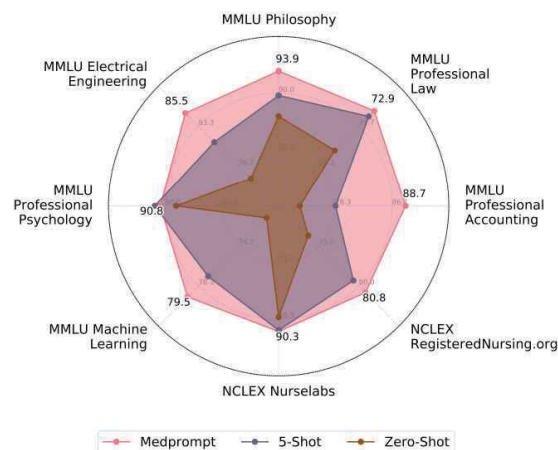


[그림 54] 다양한 LLM의 에너지 소비량과 추론 속도를 비교

LLM의 학습, 추론 등을 효율적으로 처리하기 위한 프레임 워크의 개발은 다음과 같이 진행되어지고 있다. Zhongwei Wan et al.(2023)에서는 LLM의 효율성을 위한 다양한 방법을 언급하고 있는데, Microsoft의 DeepSpeed는 LLM의 훈련과 배포를 위한 통합 프레임 워크이다. DeepSpeed는 대규모 모델의 GPU 메모리 제약을 해결하기 위해 만들어진 ZeRO-Infernece를 핵심으로 한다. ZeRO-Infernece는 모델을 여러 GPU와 CPU로 나누어 개별 장치의 메모리 제약을 관리하는 접근 방식을 제공한다. DeepSpeed의 딥퓨전 메커니즘을 기반으로 하는 DeepSpeedMII 모듈은 성능, 유연성, 비용 효율성을 강조하는 기술로 CPU와 GPU 모두에서 훈련할 수 있다. Megatron은 LLM의 훈련과 배포를 위해 구성되어 GPU모델의 병렬 처리의 효과적인 배포에 중점을 두고 있다. Megatron은 처리 속도와 메모리 활용의 최적화를 위해 여러 GPU에 분산된 모델의 텐서 연산을 전략적으로 분해하여 잠재적으로 훈련 처리량을 향상시키는 방식이다. Nvidia의 TensorRT-LLM은 처리량 향상을 위해 LLM에 맞춤형 고급 도구와 최적화를 제공한다. 특히 FasterTransformer의 최적화된 커널을 통합하고 텐서 병렬 처리를 채택하여 여러 GPU와 서버에서 대규모 효율적 추론이 가능하도록 하고 있다. 뿐만 아니라 효율적

인 추론만 제공하는 프레임 워크도 등장하고 있다. vLLM, Paralleformers, OpenLLM, RayLLM은 모델의 성능 향상, 활용 최적화, 운용 시간 및 비용 감축을 위한 프레임 워크이다.

이 외에 효율적으로 LLM을 맞춤화하고 응답의 정확도를 높이기 위해 프롬프트 엔지니어링도 중요하게 연구되어오고 있다. few-shot프롬프팅은 추가 훈련이나 파인튜닝 없이도 다양한 작업 수행이 가능하도록 하며 이는 LLM을 효율적으로 사용할 수 있도록 한다. 좋은 예시를 제공하거나, 예시의 순서를 고려하여 제공하는 방법을 사용할 수 있다. 프롬프트는 예시 제공 외에도 프롬프트 템플릿의 영향을 받는데, 인스트럭션 생성과 다단계 추론이 있다. 인스트럭션 생성의 경우 LLM이 생성도 가능하며, LLM과 상호작용을 통해 인스트럭션 생성 혹은 최적화된 인스트럭션의 생성이 가능함이 입증되기도 했다. 다단계 추론은 LLM이 출력 전 일련의 중간 단계를 거치도록 하여 답변의 품질을 향상시키는 방법(CoT)이다. LLM의 단계별 CoT를 제안하는 Auto-CoT, 각 단계의 생성 질문을 CoT에 통합하는Self-ASK, 복잡한 질문을 세분화하여 이전 질문의 답변 맥락을 반영하여 출력하는 ReAct, “self-consistency”를 통해 일관된 답을 결정하는 CoT-SC, 유효한/유효하지 않은 추론 모두를 제공하여 대조 사고를 제안하는 Constrastive CoT 등이 연구되어 발전된 답변을 출력하도록 한다. 이러한 프롬프트 엔지니어링 방식인 CoT와 GPT-4를 활용하여 작성한 프롬프트를 기반으로 하는 Nori, H et al.(2023)의 Medprompt를 활용하여 데이터 세트의 정확도를 향상시켰다.



[그림 55] Nori, H. et al.(2023) 도메인 외 데이터 세트에 대한 세 가지 프롬프트 전략을 사용한 GPT-4 성능

그뿐만 아니라 Medprompt는 의료 도메인(MedQA)에 국한되지 않고 일반적인 도메인으로 확장 가능한 방법임을 확인하였다. 이처럼 효율성, 지속 가능성에 대한 요구와 필요성 등을 고려하여 앞으로의 평가 지표의 하나로 효율성에 대한 평가 지표가 추가될 필요성이 있다.

## [부록2] 한국 자연어 처리 연구 동향 조사

ChatGPT의 등장 이후에 변화한 학계의 연구 흐름을 파악하기 위해서 한국어와 관련된 인공지능 연구가 발표되는 ‘한글 및 한국어정보처리 학술대회(HCLT)’에서 ChatGPT 이전의 2022년 학술대회, ChatGPT 이후 시점인 2023년 학술대회에서 발표된 논문의 제목, 초록, 키워드, 평가 지표의 데이터를 수집하여 비교하였다. ChatGPT를 필두로 한 LLM(Large Language Model)들은 단숨에 학계에서도 많은 관심을 받았다. LLM의 가장 큰 특징이라고 하면 모델의 크기가 이전에 비해서 급격하게 증가했다는 점과 생성 기반 모델이라는 점이기에 때문에 LLM 연구가 많아졌다면 1년 새 연구에서 사용된 인공지능의 종류와 크기에 변화가 있을 것으로 기대할 수 있다. 또한 LLM이 대부분 생성형 모델을 기반으로 하기 때문에 이전까지 연구의 대상이 되었던 언어 이해 모델을 평가하기 위한 방식이 변화했을 것을 기대할 수 있고, 특정 과제를 해결하기 위한 언어 모델이 아니라 많은 과제에 모두 일정 이상의 성능을 보이는 일반 인공지능(General AI)의 모습을 보이기 때문에 해결하고자 하는 연구 목표에도 변화가 있을 것을 기대할 수 있다. 본 보고서에서는 이에 기반하여 아래의 세 가지 가설을 제시하고 데이터 분석을 통해서 가설을 증명하였다.

- ▷ 가설 1: 2022년과 2023년의 연구에서 인공지능을 평가하는 방식에 변화가 있을 것이다.
  - ☞ 참: 정량적 지표 종류와 더불어 생성 모델 평가 지표가 증가, 생성 모델 평가를 위한 정성적 지표 연구 등장
- ▷ 가설 2: 2022년과 2023년의 연구에서 해결하고자 하는 연구 목표에 변화가 있을 것이다.
  - ☞ 참: 다양한 과제 해결에서 LLM 활용을 위한 연구로 전환
- ▷ 가설 3: 2022년과 2023년의 연구에서 사용된 인공지능의 종류와 크기에 변화가 있을 것이다.
  - ☞ 참: 초거대 언어 모델이 압도적인 빈도로 등장, 모델 크기 증가 흐름 관찰

### 1. 토픽 모델링(Topic Modeling)

토픽 모델링(Topic Modeling)을 위해서 가장 먼저 2022년과 2023년에 발표된 논문들의 초록을 전처리하는 과정을 거쳤다. 전처리에는 Kiwipiepy<sup>12)</sup>를 사용하여, 토큰나이징(tokenizing)을 진행했다. 토픽 모델링을 통해서 확인하고자 했던 것이 논문의 동향 변화이기 때문에 명사 태그(NNP, NNG)만 추출했다. 또한 한글자 명사는 맥락 없이는 의미를 파악하기 힘든 노이즈로 작용하기 때문에 두 글자의 이상의 명사에 대해서만 토픽 모델링을 진행했다. 토픽 모델링은 모두 LDA 기반의 gensim<sup>13)</sup>과 tomatopy<sup>14)</sup>를 활

12) <https://github.com/bab2min/kiwipiepy>

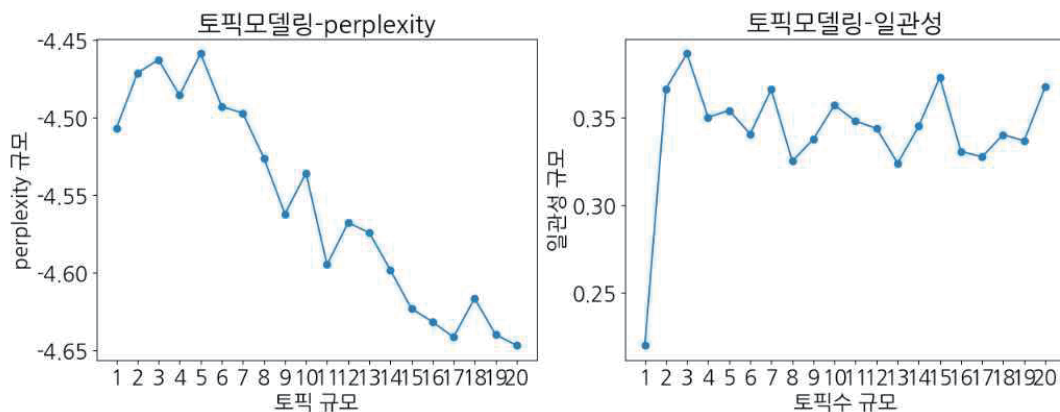
13) <https://radimrehurek.com/gensim/>

용했다.

토픽 모델링에서 기계가 아니라 사람이 직접 설정해야하는 하이퍼파라미터(hyperparameter) 중 하나로 토픽의 수가 있다. 토픽의 수를 결정하기 위해서는 토픽의 수에 따라서 변화하는 지표를 확인하여 인간의 판단으로 지표의 적절한 부분을 선택한다. gensim에서는 토픽의 수를 선택하기 위한 지표로 perplexity와 일관성을 확인할 수 있으며, tomotopy에서는 coherence(응집성)을 지표로 제공한다. 위 지표를 통해서 연도별, 모델별로 토픽의 수를 지나치게 늘리지 않으면서도 최대한 효과적으로 토픽을 뽑아낼 수 있는 숫자를 선정했다. 또 다른 하이퍼파라미터인 min\_df(단어가 나타난 문서의 수의 최솟값)에는 모든 모델과 년도에 공통적으로 5로 적용했다.

### 1.1. 2022년

2022년의 결과는 다음과 같다. 아래 그림은 gensim을 사용해서 확인한 토픽의 수에 따른 perplexity와 응집성 수치의 변화를 나타낸 그래프이다. 그래프를 토대로 가장 적절한 토픽의 수를 10개로 정했다.



[그림 56] 2022년 토픽 규모에 따른 perplexity, 일관성 변화 by gensim

gensim을 이용해서 토픽 모델링을 통해서 10개의 토픽을 뽑은 결과는 아래의 ‘표 N’과 같다. gensim에서는 각 토픽마다 총 10개까지의 단어를 묶어서 보여주고 있다.

14) <https://bab2min.github.io/tomotopy/v0.5.1/kr/>



<표 111> 2022년 토픽 추출 결과 by gensim

토픽
언어 방법 평가 사전 결과 수행 한국어 제시 존재 부족
구축 분석 방법 말뭉치 언어 생성 활용 사전 특성 자동
해결 한국어 문제 제공 표현 문장 구축 결과 실험 분야
생성 유형 텍스트 포함 언어 특징 반영 활용 방식 작업
표현 생성 사용 결과 평가 지능 실험 분류 문제 인공
대화 생성 시스템 도메인 목적 가능 입력 사용자 문장 정보
인식 한국어 평가 분류 처리 벡터 텍스트 비교 일반 유형
정보 문장 방법 추출 분류 관계 기존 텍스트 의미 사용
한국어 적용 사전 사용 처리 방법 과제 공개 언어 자동
검색 생성 응답 지식 문서 향상 대화 활용 오픈 방법

이렇게 추출한 10개의 토픽은 모두 동일한 정도의 대표성을 지니는 것은 아니다. 10개의 토픽들 중에서도 서로 상관관계가 높은 토픽들은 묶어서 더 효율적으로 동향을 파악하기 위해서 10개의 토픽을 시각화하여 서로 상관관계가 높다고 판단되는 토픽들을 군집화(clustering)하였다. 시각화의 결과는 아래 그림을 통해서 확인할 수 있다. 상관관계가 높아서 묶을 수 있다고 판단되는 주제들은 빨간색으로 경계를 표시하였다.



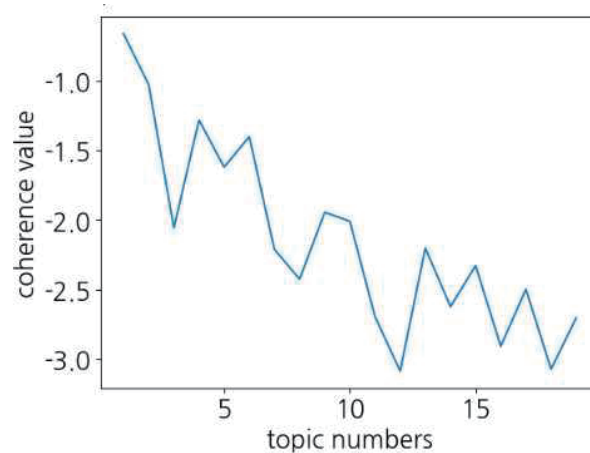
[그림 57] 2022년 토픽 모델링 시각화 및 군집화 by gensim

토픽 1,2,3,6,7,9가 묶이는 군집에서 가장 핵심이 되는 단어는 ‘한국어’로 보인다. 각 토픽에서 거의 빠짐없이 등장하며 ‘한국어 평가’, ‘한국어 말뭉치’, ‘한국어 과제/문제’ 등의 세부적인 주제로 나누어 진 것으로 보인다. 그 다음은 토픽 4,5가 묶인 군집은 ‘생



성'을 핵심으로 하는 것으로 보인다. 인간의 언어의 특징을 반영한 생성 혹은 생성된 표현을 평가하는 것을 연구주제로 하는 토픽 등이 포함되었다. 토픽 8은 '정보'를 핵심으로 문장의 핵심이 되는 정보를 어떻게 추출할 것인지에 대한 주제로 인식되고, 토픽 10은 '검색'을 핵심으로 문서에서 필요한 지식을 어떻게 검색할 것인지에 대한 연구가 있었던 것으로 볼 수 있을 것이다.

tomotopy는 gensim과 달리 coherence 하나의 지표로 토픽의 수를 결정해야한다. tomatopy로 계산한 토픽의 수에 따른 coherence 값은 다음 그림과 같다. 그래프를 통해서 도출해 낼 수 있는 2022년의 최적의 토픽의 수는 12개이다.



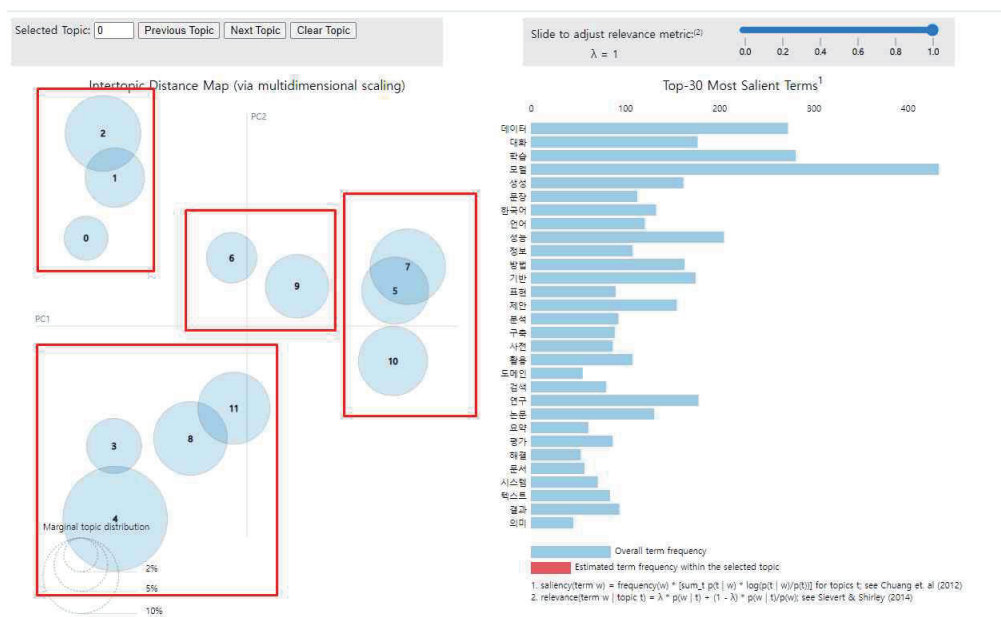
[그림 58] 2022년 토픽 규모에 따른 coherence 값 변화 by tomatopy

아래 표에서 tomatopy에서 추출된 12개의 토픽을 확인할 수 있다. gensim과의 차이점은 gensim은 각 토픽에 포함된 단어의 수가 10개인 것에 반해 tomatopy에서는 5개라는 점이 있다.

<표 112> 2022년 토픽 추출 결과 by tomatopy

토픽
도메인 해결 상호 참조 오픈
표현 분석 말뭉치 교정 연구
데이터 구축 자동 경우 연구
대화 관계 개체 시스템 제안
방법 성능 제안 논문 결과
생성 모델 연구 유형 입력
문장 학습 의미 문맥 오류
학습 모델 언어 사전 과제
한국어 텍스트 평가 유사 활용
요약 문서 질의 구조 모델
모델 정보 추출 일반 작업
기반 검색 활용 성능 지식

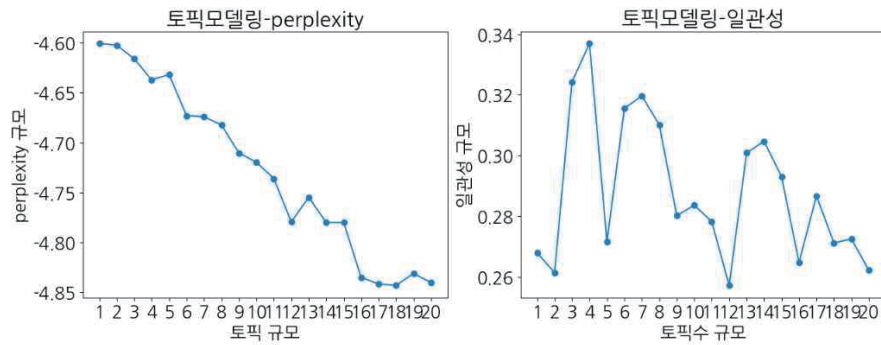
위 토픽들을 시각화하여 좌표평면 상에서 가까운 것들끼리 군집화하면 다음 그림과 같이 4개의 군집으로 묶는 것이 가능하다. 첫 번째 군집은 토픽 0,1,2의 묶음으로 ‘상호 참조 해결’, ‘표현 교정’, ‘자동 데이터 구축’과 같은 ‘과제’를 핵심으로 하는 토픽들이 모인 것으로 확인 할 수 있다. 토픽 3,4,8,11이 묶인 군집에서는 ‘대화 시스템’, ‘한국어 텍스트 평가’, ‘지식 기반 검색’등을 해결하는 모델들의 ‘성능’을 핵심으로 묶인 것으로 보인다. ‘문맥 오류’, ‘문서 요약’ 등이 포함된 토픽 6,9가 포함된 군집은 요약을 위해서 혹은 오류를 찾기 위해 ‘의미’를 이해하고자 한다는 공통점이 있으며 ‘생성 모델’, ‘사전학습 언어모델’, ‘일반 정보 모델’등 ‘모델’에 대한 토픽이 공통점이 있는 5,7,10도 하나의 군집을 이룬다.



[그림 59] 2022년 토픽 모델링 시각화 및 군집화 by tomatopy

## 1.2. 2023년

ChatGPT가 등장하고 난 이후에 이루어진 2023년 HCLT에서 발표된 논문의 초록으로 토픽 모델링을 2022년과 마찬가지로 gensim, tomatopy의 두 가지 모델로 진행했다. 먼저 gensim의 경우 [그림 60]의 perplexity, 일관성 지표를 확인하여 최적의 토픽의 개수는 16개로 확정하여 토픽 모델링을 진행하였다. 그렇게 추출된 토픽들은 <표 113>의 16가지이다.

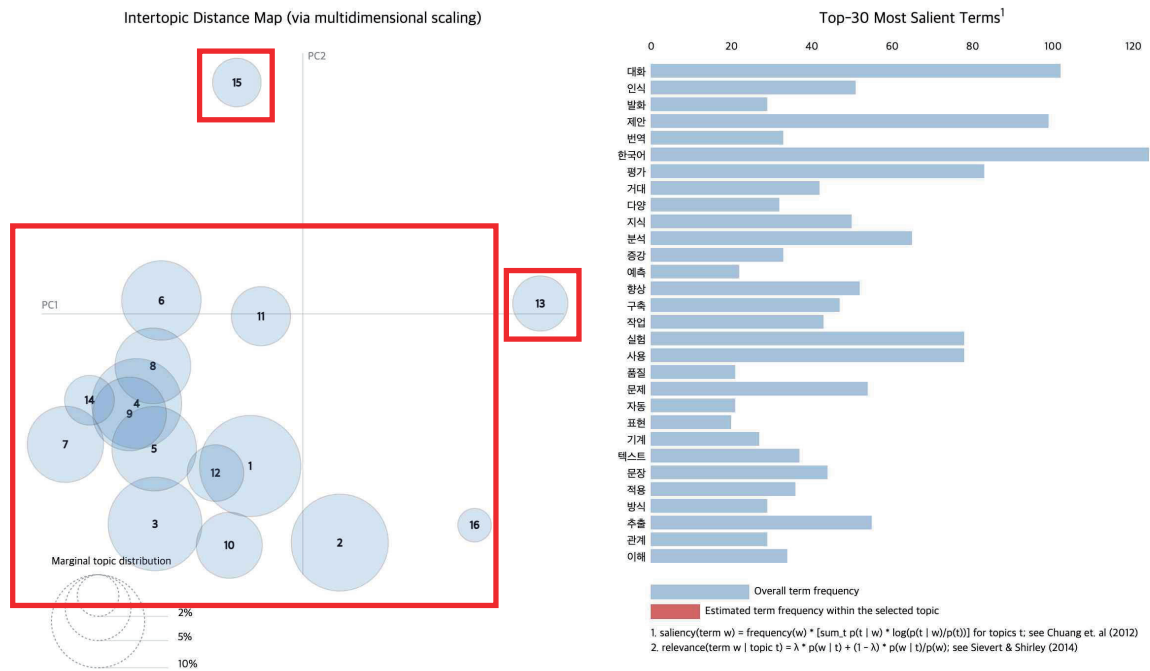


[그림 60] 2023년 토픽 규모에 따른 perplexity, 일관성 변화 by gensim

<표 113> 2023년 토픽 추출 결과 by gensim

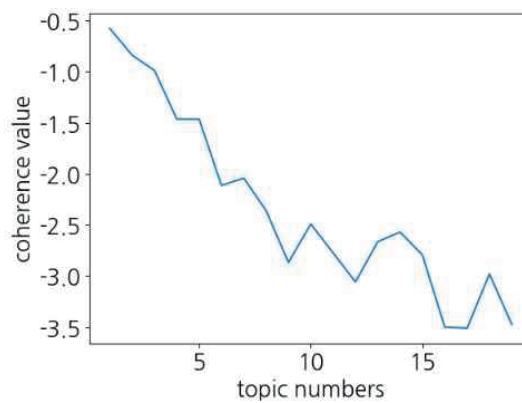
토픽
발화 제안 예측 한국어 구축 텍스트 사용 다양 분석 조정
대화 지식 기반 시스템 사용 제안 한국어 구축 발화 사용자
표현 분석 문장 해당 결과 실험 텍스트 한국어 문제 기반
한국어 관계 추론 거대 추출 평가 기법 자연어 제안 반영
한국어 인식 사전 처리 기반 결과 단위 문서 분석 구축
제안 기존 결과 문서 분류 기반 검색 분석 시스템 정답
문제 제안 분류 기존 정보 검색 작업 실험 진행 경우
평가 추출 결과 과제 정보 문장 구조 수행 규모 사용
이해 실험 처리 정보 적용 수행 문맥 기계 자연어 과제
인식 번역 기계 자동 구축 관련 기반 러닝 다양 의미
사용 사전 분야 조정 방식 최근 미세 질문 효율 지식
다양 향상 품질 자동 평가 제안 적용 이용 진행 방식
대화 평가 요약 기반 방식 문서 수행 분석 조정 결과
거대 실험 증강 작업 분석 부족 입력 상황 사용 영향
번역 기계 가능 작업 사용 추출 제시 추가 발생 품질
규모 기반 향상 증강 제안 능력 기법 도메인 다양 한국어

2023년의 논문 초록을 토대로 토픽 모델링을 gensim을 이용해서 진행해서 나온 16개의 토픽들을 시각화 해보면 [그림 61]과 같이 토픽 간의 거리를 확인할 수 있다. 추출된 토픽의 수는 16개로 많지만 군집화하자면 3개 정도로 묶을 수 있으며 그 중에서 14개를 하나의 군집으로 묶을 수 있을 만큼 토픽들이 2022년에 비해 균질화되었다고 추측할 수 있다. 14개의 토픽들은 보통 각자의 ‘발화 예측’, ‘대화 시스템’, ‘문장 표현 분석’ 등의 세부 주제로 이루어져 있지만 2022년에 비해서 특징적으로 나타나는 단어들은 ‘거대’, ‘규모’, ‘기존’이라고 볼 수 있다. 이는 ‘거대 언어 모델(LLM)’이 확실히 큰 규모의 모델로 많이 논의되기 시작했다는 것의 증거로 볼 수 있다. 군집에 포함되지 못한 두 토픽은 좌표 평면 상에서는 서로 떨어져 있지만 모두 ‘증강’ 관련으로 LLM을 이용한 과제의 해결이 아니라 LLM의 생성 능력에 집중하여 한국어에 부족한 라벨링 데이터의 양을 늘리는 것의 가능성을 모색했다는 점에서 상관관계가 적게 나타난 것으로 보인다.



[그림 61] 2023년 토픽 모델링 시각화 및 군집화 by gensim

tomotopy에 따르면 2023년 HCLT 초록의 토픽 수에 따른 coherence 값의 변동은 아래 [그림 62]와 같다. 가장 최적의 토픽의 수를 14개로 선정하여 이후 토픽 모델링을 진행하여 결과로 <표 114>의 토픽들을 얻었다.

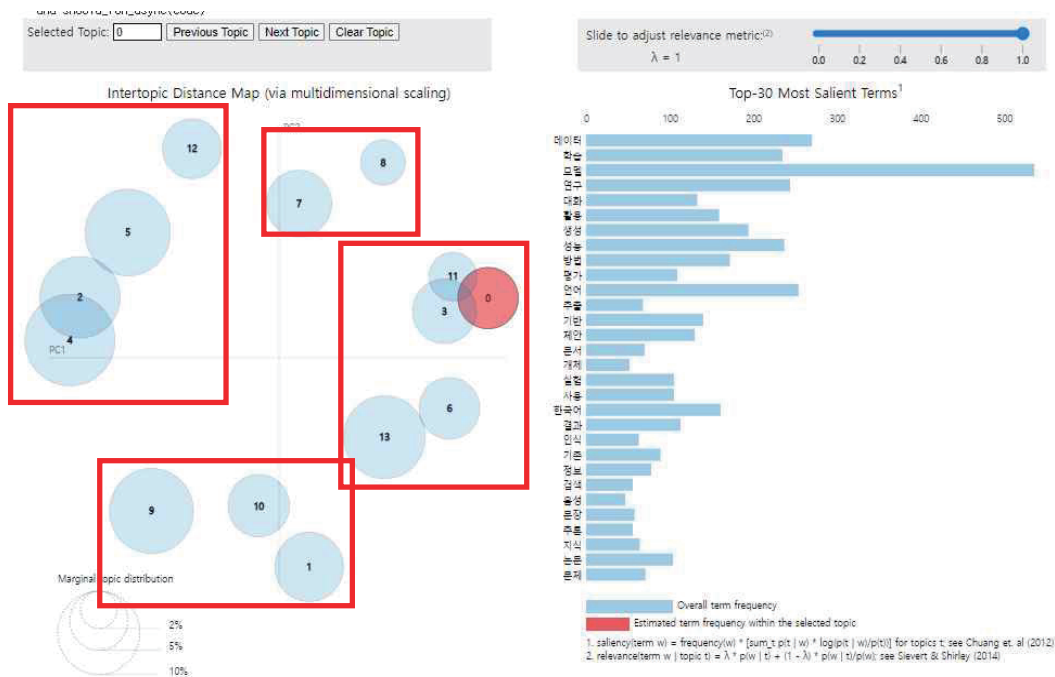


[그림 62] 2023년 토픽 규모에 따른 coherence 값 변화 by tomatopy

<표 114> 2023년 토픽 추출 결과 by tomotopy

토픽
대화 문서 검색 시스템 발화
연구 결과 적용 제시 기계
데이터 모델 성능 제안 텍스트
분류 의도 조정 관련 미세
모델 언어 학습 성능 한국어
생성 모델 제안 문제 정보
평가 처리 요약 기반 편향
문장 수행 이해 과정 방식
추출 개체 관계 인식 구조
모델 실험 분석 언어 확인
성능 추론 연구 기존 추가
음성 인식 번역 지식 기반
학습 사용 진행 사용자 라벨
활용 방법 언어 기존 구축

추출된 14개의 토픽들을 시각화하여 상관 관계가 높다고 나타나는 토픽들끼리 군집화 하면 [그림 63]처럼 묶을 수 있다. 토픽 1,9,10은 ‘연구 결과 적용’, ‘모델 실험 분석’, ‘추론 성능 연구’등의 모델의 능력을 확인하고 ‘기존’ 키워드처럼 그전까지와 차별점을 확인하려는 점에서 공통점을 확인할 수 있다. 토픽 0, 3, 6, 11, 13은 ‘대화 시스템’, ‘문서 검색’, ‘의도 분류’, ‘요약’, ‘음성 인식’, ‘번역’ 등의 LLM으로 이전에 비해 더 잘 해결이 가능해진 여러 과제들이라는 공통점이 있으며 ‘기존’이라는 키워드와 ‘활용/구축 방법’이라는 키워드가 기존과 차별화된 모델을 어떻게 더 잘 활용할 것인지에 대한 고민하는 연구들이라는 점에서 상관관계가 높았다고 보인다. 토픽 2, 4, 5, 12는 ‘모델’이라는 말이 가장 공통적으로 나타나며, ‘생성 모델’, ‘언어 모델’, ‘모델 데이터’ 등의 키워드로 모델 관련한 연구라는 관련성으로 묶인 것으로 추측된다. 마지막으로 토픽 7,8은 ‘문장’, ‘개체명 인식’, ‘관계 추출’ 등의 키워드로 보아 언어학적인 접근을 시도한 연구들도 2023년에도 꾸준히 있었다는 것을 말해준다.



[그림 63] 2023년 토픽 모델링 시각화 및 군집화 by tomotopy

본 보고서에서는 gensim과 tomotopy의 두 가지 도구를 이용하여 2022년과 2023년에 HCLT에서 발표된 논문들의 초록 데이터를 기반으로 각 년도별 특징적으로 나타난 토픽들을 분석하고자 시도했다. 2022년에 비해 2023년에는 ‘기존’이라는 키워드가 등장하면서 LLM의 등장 이후 연구의 패러다임이 변화했고 기존 연구들과의 차별점을 보이려는 것이 더 두드러지게 나타났다는 것을 볼 수 있었고, 2023년에 등장하는 ‘거대/생성 모델’이라는 키워드도 LLM 트렌드로의 변화를 뒷받침하는 근거라고 볼 수 있다.

## 2. 키워드 빈도 분석

토픽 모델링은 기계가 정량적으로 분석한 각 논문의 주제를 기반으로 한 분석이고, HCLT에 제출된 논문들은 저자가 정성적인 판단에 근거하여 핵심 단어로 선정한 키워드들이 있기 때문에 이를 바탕으로 2022년과 2023년에 키워드의 변화를 빈도를 통해서 살펴보았다.

<표 115> 2022년 HCLT 논문 키워드 빈도

Keyword	Frequency (2022)
자연어처리	8
자연어 생성	8

딥러닝	7
데이터 증강	6
대조 학습	4
생성 모델	4
언어 모델	4
BERT	4
오픈 도메인 질의응답	3
질의응답	3
KoBERT	3
언어모델	3
챗봇	3
상호참조해결	3
멀티 모달	3
자연어 추론	3
트랜스포머	2
대화 이해	2
관계추출	2
평가 지표	2
감성분석	2
검색 모델	2
한국어 페르소나 대화	2
한국어 언어 모델	2
기계번역	2
회의록 요약	2
자연어 처리	2
지식 그래프	2
관계 추출	2
사전학습 모델	2
대화 생성	2
병렬말뭉치	2
유사대화 검색	2
임베딩 모델	2
학습데이터 구축	2
대화 시스템	2
전이학습	2
멀티턴 대화	2

Dense Retrieval	2
감정 분석	2
대화시스템	2

<표 116> 2023년 HCLT 논문 키워드 빈도

Keyword	Frequency (2023)
데이터 증강	8
대규모 언어 모델	7
프롬프트	6
Large Language Model	5
LLM	5
멀티모달	4
자연어 생성	4
언어 모델	4
사전학습 언어모델	3
대규모 언어모델	3
ChatGPT	3
언어모델	3
자연어처리	3
자연어 처리	3
대규모 언어 모델(LLM)	3
개체명 인식	2
요약	2
질의 응답	2
거대 언어 모델	2
GPT	2
초거대 언어모델	2
휴지 예측	2
거대 언어모델	2
통계생산 방법론	2
Instruction Tuning	2
BERT	2
페르소나	2
음성인식	2
한국어	2



기계 번역	2
대조 학습	2
에세이 자동 평가	2
생성 모델	2
기계번역	2
기계독해	2

위의 <표 115>, <표 116>은 논문의 저자가 직접 선정한 키워드를 최소 2개 이상의 빈도로 나타난 키워드에 대해서만 빈도수를 계산한 표이다. 다만 위의 데이터는 학계에서 아직 용어의 통일이 안된 개념 혹은 단순 띄어쓰기 차이 등으로 빈도수에 대한 적절한 계산이라고 보기 어렵기 때문에 본 보고서에서는 수작업으로 동일한 개념에 대한 키워드는 빈도수를 합쳐 정규화하는 작업을 진행했다.

<표 117> 2022년 HCLT 키워드  
빈도(정규화)

Keyword	정규화 빈도
자연어처리	10
언어 모델	9
자연어 생성	8
딥러닝	7
데이터 증강	6
질의응답	6
대조 학습	4
생성 모델	4
BERT	4
대화 시스템	4
KoBERT	3
챗봇	3
상호참조해결	3
멀티 모달	3
자연어 추론	3
트랜스포머	2
대화 이해	2
관계추출	2
평가 지표	2
감성분석	2

검색 모델	2
한국어 페르소나 대화	2
기계번역	2
회의록 요약	2
지식 그래프	2
관계 추출	2
사전학습 모델	2
대화 생성	2
병렬말뭉치	2
유사대화 검색	2
임베딩 모델	2
학습데이터 구축	2
전이학습	2
멀티턴 대화	2
Dense Retrieval	2
감정 분석	2

<표 118> 2023년 HCLT 키워드 빈도(정규화)

Keyword	정규화빈도
대규모 언어 모델	24
데이터 증강	8
언어 모델	7
프롬프트	6
자연어처리	6
멀티모달	4
자연어 생성	4
기계 번역	4
사전학습 언어모델	3
ChatGPT	3
개체명 인식	2
요약	2
질의 응답	2
GPT	2
휴지 예측	2
통계생산 방법론	2
Instruction Tuning	2

BERT	2
페르소나	2
음성인식	2
한국어	2
대조 학습	2
에세이 자동 평가	2
생성 모델	2
기계독해	2

정규화 결과 2022년에는 없는 개념인 LLM(대규모 언어 모델)이 2023년에는 다른 모든 키워드들에 비해서도 압도적으로 많이 나타났다는 사실이 가장 눈에 띄게 나타난다. 또한 2번 이상한 등장한 키워드의 개수는 2022년이 더 많다는 사실에서 2022년에 다양한 주제에 대해서 논의를 하다가 2023년에는 LLM 관련 연구에 집중되었다고도 해석이 가능하다. 2023년에 비해 2022년에 더 다양한 세부적인 과제가 키워드로 등장했다는 사실도 의미가 있다.

### 3. 평가 지표(Metric) 분석

2022년 논문의 자동 평가지표에서는 'F1', 'Recall', 'Precision' 같은 전통적인 평가 지표들이 주로 사용됐다. 이들 지표는 정밀도, 재현율, F1 점수 등을 통해 논문의 성능을 평가하는 데 중요한 역할을 했다. 또한, 'Accuracy' 같은 일반적인 성능 지표도 중요하게 다뤄졌으며, 연구의 명확성과 정확성 평가에 기여했다. 각 연구의 특성에 맞춰 다양한 평가 방법들이 적용된 것도 눈에 띈다.

2023년 자동평가에서는 새로운 평가 지표들이 등장했다. 예를 들어, 'Semantic Score', 'SentenceBERT(SBERT)', 같은 언어 모델 관련 지표들이 새롭게 사용되었다. 이러한 지표들은 언어 처리 모델의 성능을 더 깊이 분석하는 데 중점을 두고 있다. 'Diversity', 'Distinct-2', 'FID' 같은 지표들은 생성된 텍스트의 다양성과 창의성을 평가하는 데 중요한 역할을 하며, 언어 생성 모델의 발전을 반영한다.

2022년과 2023년 자동평가 방법을 비교하면, 2023년에는 다양하고 새로운 평가 지표들이 등장한 것이 두드러진다. 특히, 언어 모델의 성능을 평가하는 지표들이 눈에 띄는데, 'Semantic Score', 'BERTScore' 같은 지표들은 모델의 정교한 성능 평가에 초점을 맞춘다. 2022년과 2023년 데이터를 비교해보면, 'Semantic Score'와 'BERTScore' 지표의 사용 빈도가 2023년에 증가한 것을 볼 수 있다. 'Semantic Score'는 2022년에는 사용되지 않았으나, 2023년에 1회 등장했고, 'BERTScore'의 경우, 2022년에 3회 사용된 반면, 2023년에는 6회 사용되어 두 배 가량 증가했다. 더불어, 'Diversity',

'Distinct-2' 같은 생성 관련 지표들이 새롭게 등장함으로써, 생성 모델의 다양성과 창의성을 평가하는 데 중점을 두는 연구가 많아졌음을 보여준다.

2022년의 정성평가 데이터에서 'X'는 논문에 구체적인 정성평가 지표가 제시되지 않았음을 나타내며, 이는 94회에 달했다. 이외에 나타난 평가 지표들은 각각 단 한 번씩만 사용되었다. 이러한 지표들은 '적절성', '매력도', '유창성', '표현', '구성', '내용', '탐지 오류', '해석 오류' 등과 같은 다양한 측면을 반영하며, 인공지능의 생성한 텍스트의 내용과 스타일, 기술적 오류에 대한 세부적인 평가를 측정하고자 한 것으로 보인다.

2023년에는 'X'의 빈도가 108회로 증가했다. 이는 더 많은 논문들에서 구체적인 정성평가 지표가 나타나지 않았음을 의미한다. 그러나 새롭게 등장한 평가 지표들로는 '오류 유형 분석', 'coherence', 'consistency', 'fluency', 'relevance' 등이 있으며, 이들은 텍스트의 일관성, 유창성, 관련성 등을 평가하는 데 중점을 두고 있다. 이러한 지표들은 특히 생성 모델의 성능을 세밀하게 평가하는 데 유용하다.

2022년 대비 2023년의 정성평가에서 두드러진 차이점은 논문에 정성평가 지표가 덜 나타났다는 것과 함께, 새로운 평가 지표들의 도입이다. 특히 2023년에 등장한 'coherence', 'consistency', 'fluency', 'relevance'와 같은 지표들은 생성 모델의 다양한 특성을 세밀하게 평가하는 데 중요하다. 이는 생성 모델과 관련된 연구가 증가하고 있으며, 이러한 모델들의 성능을 보다 정밀하게 평가하려는 연구자들의 노력을 반영한다. 이러한 변화는 연구의 방향성과 평가 방법의 다변화를 나타내며, 연구자들이 보다 구체적이고 세밀한 평가 기준을 모색하고 있음을 보여준다.

#### 4. 가설 검증

본 보고서에서는 이전에 제시한 세 가지 가설을 검증하기 위해서 2022, 2023년도에 HCLT에서 발표된 논문들의 초록, 키워드, 평가 지표의 데이터를 가지고 분석을 진행하였다. 분석은 토픽 모델링과 키워드 빈도 분석의 방법으로 진행되었다.

첫 번째 가설, 2022년과 2023년의 연구에서 인공지능을 평가하는 방식에 변화가 있을 것이라는 점은 2022년과 2023년의 평가 지표를 추출해 낸 것을 빈도 분석한 결과를 통해서 검증이 가능했다. 실제로 자동 평가 지표에서는 semantic score, BERTscore 같은 지표들이 늘거나 새로 등장하여 생성 모델을 평가하기 위한 지표가 늘어난 것을 확인할 수 있었으며, 정성 평가도 정성 평가 자체가 늘어나지는 않았지만 정성 평가에 사용한 지표가 조금 더 생성 모델을 평가하는 것에 용이하도록 변화하고 있다는 것을 확인했다.

두 번째 가설은 2022년과 2023년의 연구에서 해결하고자 하는 연구 목표에 변화가 있을 것인데 이는 키워드 분석, 토픽 모델링의 결과를 보았을 때, 제시된 과제들이 변화한

것을 근거로 참이라고 할 수 있을 것으로 보인다. 2022년에는 인공지능을 더 다양한 과제들을 해결하는 것에 초점이 맞춰져 있었다면 2023년에는 해결하고자 하는 과제의 종류는 줄고, LLM을 잘 활용할 수 있는 방안을 찾는 것을 목표로 하는 키워드와 토픽들이 더 두드러지게 나타난 것을 확인했다.

마지막 가설은 2022년과 2023년의 연구에서 사용된 인공지능의 종류와 크기에 변화가 있을 것이라는 것으로 키워드 분석에서 ‘거대 언어 모델(LLM)’이 압도적으로 나타난 2023년의 분석 결과나, 2023년에 나타난 새로운 토픽 ‘거대’ 등에서 2022년에 비해서 크기가 큰 모델이 연구의 주류가 되었다는 점을 확인할 수 있었다.

이런 트렌드의 변화에 발맞추기 위해서는 다음과 같은 방향이 필요할 것으로 보인다. 첫째, 방대한 양의 질 좋은 한국어 데이터 구축이 필요하다. 키워드 분석과 토픽 모델링 결과에서 2023년에 더 많이 등장한 내용 중 하나가 ‘증강’이라는 키워드이다. 이는 영미권, 중국어권에 비해서 상대적으로 적은 데이터의 양을 해결하기 위한 노력의 시도가 LLM의 시대에 더 늘어난 것으로 보인다. 따라서, 한국어 LLM을 위한 방대한 양의 질 좋은 데이터를 구축하여 연구자들이 사용할 수 있도록 해야 한다. 둘째, 한국어 특성에 맞는 평가 지표 개발이 필요하다. 한국어는 영어와 같은 서양 언어와 비교하여 문법, 어휘, 발음 등에서 차이점이 있다. 따라서, 한국어 LLM의 성능을 정확하게 평가하기 위해서는 한국어 특성에 맞는 평가 지표를 개발해야 한다. 셋째, 연구 인력, 연구 장비, 연구 자금 등 연구 생태계 조성이 필요하다. 한국어 LLM 연구는 아직 초기 단계에 있으며, 연구 인력과 연구 장비, 연구 자금이 부족한 실정이다. 따라서, 정부와 기업, 대학 등 다양한 주체의 협력을 통해 연구 인력 양성과 연구 장비 구축, 연구 자금 지원 등 연구 생태계를 조성해야 한다. 이러한 노력을 통해 한국어 LLM의 성능을 향상시키고, 한국어 LLM을 활용한 다양한 응용 연구를 활성화시킬 수 있을 것으로 기대할 수 있다.

## [부록3] 과제위원회 회의록

### 제1회 국어원 인공지능 언어능력 평가체계 검토위원회 회의록

일시: 2023.5.30.

장소: 비대면 화상회의

#### 1. 감정 분석 task 관련 검토

- 1) 감정 분석 성능의 기본 상한선(upperbound) 설정 관련
  - 상한 설정에 대한 기준: 사람의 성능 vs 베이스라인의 성능으로 상정 가능
  - 이를 위해서는 데이터 난도 측정에 따른 과제 난도 산정 필요
  - 현재 연구진 내에서 모의 실험 수행
    - ▷ 감정 분석 데이터에 대해 멀티 레이블 분류만 시행: 준수한 성능
    - ▷ 감정 분류 과제는 문장 분류 과제로써 난도는 높지 않을 것으로 예상
    - ▷ 다만 2개 과제 수행 시 f1-score는  $\pm 60$ 으로 예측 중
- 참고) ABSA: 베이스라인 f1-score 50, 1등팀 f1-score 68
- 2) 서브 과제 설정 관련
  - 데이터를 소수로 주는 과제, 전체 데이터(5만 건)를 제공하는 과제로 조절 가능

#### 2. 생성 AI 시대 리더보드 운영 관련 검토

- 이미 성능이 좋은 생성 AI들이 많이 개발되었기에 리더보드 운영 시 고려 필요
- 프롬프팅 허용 시 모델의 재현성을 우선으로 해야 함
  - ▷ 상업적 API 사용 제한, 오픈 소스 모델로 수행 및 재현할 수 있는 방향으로 설계 필요
- 예) 제출 시 query를 제외한 prompt 제출
- 예) 재현성을 평가할 수 있도록 개발한 모델 & 데모를 위한 서버 주소 제출 등
- 프롬프트 제출 등의 방법 생각 가능: 발주처와 논의 필요
- 생성 AI를 고려한 리더보드 운영 역시 국립국어원을 비롯한 유관기관의 논의 반영 필요
- 차후 검토위원회 논의사항으로 재상정 예정

#### 3. 이야기 완성 과제 관련 검토

- 분류보다는 인간 평가를 수행하더라도 사람이 평가하는 생성 과제 방향성이 시의적절
- 과제를 통해 생성 결과에 대한 이해 측정 가능
- 이에 따라 이야기 완성 말뭉치는 생성 과제로 조직하는 것이 바람직함
- 1) 이야기 '생성' 과제 시 인간 평가 관련
  - 정량적인 것을 가미하되, 재현 가능한 수작업 평가가 중심이 되어야 함
- (1) 다수의 평가자의 결과를 voting하는 방법 개발 필요성 존재
  - 정답에 대한 인간 평가 의견 수합 시 결과 취합 이상의 방향성을 가져야 함
  - 인간의 선호도 랭킹을 매기는 RLHF 참고 가능
  - 또한 nDCG(normalized Discounted Cumulative Gain) 등 기존 랭킹 관련 지표를 참고 가능

- ▷ 랭킹 매기기의 경우 쌍별(pairwise) 평가 시 산출물 후보(output candidate) 간 퀄리티 비교가 가능하나, 평가가 어렵다는 단점 존재
- ▷ round 1 / 2 로 나누어 hybrid로 하는 방법, 기계적 지표와 병행하여 hybrid로 하는 방법 등도 사용 가능

(2) 정확한 과제 정의 및 scoring 공지 요구

- 정성적 평가의 경우 평가 과정 중 불만 발생 가능성을 고려해야 함

(3) 인간 평가 시 명확한 평가 지침, 체크리스트 및 IAA 담보 필요

- 지침, 체크리스트의 경우 이야기 말뭉치 구축 방식을 근거로 활용 가능
- evaluator간 IAA 및 컨센서스, 과제 이해도 증명 필요
- 평가 예제를 주어 평가의 신뢰도를 올리는 방법도 사용 가능

2) 경진대회 상황에서의 인간평가

- 인간평가 방법이 이상적이나, 시간과 노력 등의 비용이 많이 소요될 것으로 예상

(1) 참가자들에게 정량 & 정성평가 중 어디에 방점이 있는지 안내 필요

예) human eval.은 메인, 다른 정량평가는 사람 평가와 관련된 간접적 평가 수단(proxy measure)로 판단

- 정량적 스코어를 리더보드에 게재하거나, 스크립트를 주어 참가자들이 평가하는 것도 좋음

▷ 참가자들이 자기 결과를 눈으로 보아가면서 proxy measure과 자연스러움에 대한 스스로의 판단 가능

▷ 이때 사용된 정량적 평가 지표는 보조적 수단으로만 활용

(2) 전체 참가팀에 대한 상위 10팀 선정 방법 관련

- 인간평가 시나리오 안에서 상위 10위 팀을 선정하는 것은 까다로운 문제
- 초기에는 좋다/아니다로 평가하여 이 중 상위 10팀을 선정

▷ 이후 상위 10팀에 대해서는 세밀한 인간 평가(예: 랭킹 매기기)를 수행하는 것도 방법

3) 정량평가 추가 지표

- 정량평가 기준 중 NLI score 도입 가능

▷ 정답과의 일치를 산정하며 비슷한 의미에 있는 것들에 대해 점수화 가능

- 지식 추론 과제(Reasoning task) 관점의 ROSCOE 등 정량적 평가 지표 사용 가능

▷ 문장 레벨 evaluation scheme으로 연속적인 문장에 특화

- 펄플렉서티(perplexity) 역시 자연어 생성 과제에서 쓰이는 지표

▷ 문장 & 전체에 대한 펄플렉서티로 구분

▷ 펄플렉서티가 낮을수록 생성이 잘 되고 있음을 의미

- 또한 정량 평가 과정에서 생성 AI의 적극적 활용도 가능할 것으로 전망

예) 이야기 완성 말뭉치 1, 3번 문장을 토대로 생성한 2번 문장을 입력 후 생성 score 연산 → 자연스럽게 연결되는 문장들일수록 score가 높게 산출

## 제2회 국어원 인공지능 언어능력 평가체계 검토위원회 회의록

일시: 2023.7.24.  
장소: 비대면 화상회의

### 1. 경진대회 모델 사용 관련

#### 1) 외부 데이터에 대한 정의 필요

- 현재 경진대회 세팅: LLM 사용을 하지 않고서는 참여 어려움
- 이에 따라 체크포인트만 활용 가능한 것인지 등을 명확히 해야 함  
예) 프롬프트 인컨텍스트 러닝 시 새로운 데이터 주입/입력 여부 등

#### 2) 향후 API 사용 가능 여부

- 현재 경향: 잘 학습된 API를 활용하는 방향으로 나아가고 있음  
예) ACL 2023: LLM 경량화 연구 트렌드
- 또한 작은 LLM일 경우 성능이 원활하게 나오지 않을 가능성도 존재  
ChatGPT: GPU 메모리 40GB에서 구동
- 미래를 보았을 때 LLM 기업/기관 등과 협업하여 API를 사용 검토 필요  
→ 참여자 확대 가능: 공학적 배경을 가지지 않은 참가자들도 참여 가능

#### 3) 경진대회 중 '모델의 독창성 평가' 관련

- 질의응답을 포함한 제출팀의 모델 독창성 발표 시간 필요
- LLM 사용 시 독창성: 인스트럭션 튜닝 을 얼마나 잘 하느냐로 평가 가능
- 이에 따라 독창성 평가 시 '발표' 형식을 채택할 것인지 '문서' 형식을 채택할 것인지 확인 필요
- LLM으로 인해 진입장벽이 내려가 참가자 수는 크게 감소하지 않을 것으로 예상
- 기본 LLM 모델들에 대해 프롬프트 엔지니어링을 넘어 '인스트럭션 디자인/튜닝 및 모델 변형 등' 필요  
→ 위 내용에 따라 독창성 평가 및 리더보드 활성화 가능

### 2. 이야기 완성 task 인간 평가 적절성 관련

#### 1) 현재 인간 평가 표본 수: 한 팀당 정량 점수 상위 15개 문장

- 표본 수가 적으며, 15개 문장이 모두 동일한 문장이어야 함
- 기존 관례: 데이터 및 연구에 따라 다르나 기본적으로 수 백건 ~ 천 건을 평가셋으로 사용  
→ 다만 현재 세팅에 비추었을 때 분량이 많으므로 문제에 대한 난도 낮추고 평가 수량을 늘려 인간 평가를 진행하는 것이 바람직
- 또한 상위 15개 문장만 다룰 경우 '다양성, 독창성'이 드러나는 하위 순위 표본이 묻힐 수도 있음  
→ 전체 문장에 대해 랜덤 샘플링을 하여 평가 대상을 선정하는 것도 방법

#### 2) 리커트 점수 외 점수 산정 방법 고려 필요

- 리커트 점수: 시간, 비용이 많이 드는 평가 방법



→ 랭킹 매기기, yes/no 이진 평가 등 고려 가능

### 3. 최종 순위 결정 관련

#### 1) 정량적 평가의 한계 및 보완

- 정량적-정성적 지표 간 어떤 비율을 써도 정량적 지표는 평가 근거가 불충분할 수 있음
- 따라서 최종 순위는 정성적 평가에 따라 결정해야 함
- GPT-4 점수: 정량 평가 항목 중 하나로 고려 가능
- 예) 고려대 KULLM: 평가 방식 참고 가능

#### 2) 과제 가중치 관련

- 과제에 대해 가중치를 둘 수도 있으나 평가위원회에서 최종 평가 시 고려하는 것도 방법

#### 3) 대상 결정 관련

- 현재 기준: '두 과제 참여 시 가산점 부여', 두 과제 참여 시 대상 심사 후보 등록
- 다만 두 과제 참여 시 참가자 부담 증가
  - 대상에 조건으로 제약을 거는 것보다 없애는 것을 권장
- 시상의 구체적인 사안(과제별 시상 비율, 대상 선정 방식 등)은 평가위원회에서 최종 결정
  - 탑 티어 컨퍼런스 예에 따라 모델 독창성을 대상 결정의 요인으로도 활용 가능

### 4. 상시 과제 전환 시 인간 평가 운용 관련

#### 1) 2023 상시 과제 중 인간평가 방법

- 과도기적 상황이므로 인간평가에 대해 바로 자동평가로 전환하는 것은 불가
- 주기별로 인간 평가 진행 혹은 human 데이터 세트를 구축하여 기준으로 활용 가능
- 혹은 GPT-4와 같이 기준이 되는 golden score나 기준 평가 프롬프트를 두는 것도 고려
  - 단 GPT-4 활용 시 데이터 유출 등에 대한 고민은 필요

#### 2) 인간 평가의 자동화 관련

- 최종 목표가 되어야 하며, 이에 대한 연구 반드시 필요
- 예) 현재 개발된 루브릭의 자동 평가화 등...
- 자동 전환 시 자동 평가와 인간 성능 간 비교가 가능하게 해야 신뢰성 제고 가능

### 5. 기타

#### 1) 경진대회 과제 기술서 내 리더보드 조건

- '등록한 예측 결과 중 일정 비율(예 70%)을 무작위 추출하여 평가한 후 순위표(리더보드)에 평가 점수 및 순위를 제공한다.'
- 결과 중 70%를 랜덤 샘플링: 상위권 중 점수 간 차이가 근소할 때 랜덤 샘플링으로 인한 데이터 세트 차이 우려
  - 이에 따라 전수 평가를 하되, 1일 제출 횟수를 10회에서 더 줄이는 것이 필요

2) 경진대회 참가 독려 타게팅 대상

- 채용 특전 등으로 인해 전년도와 같이 학부생이 주일 것으로 예상
- 다만 향후 기업에 대한 제도적 특전 마련 시 다른 주체들에 대한 참여 독려가 가능할 것으로 보임

## 제1회 인공지능 언어능력 평가체계 자문위원회 회의록

일시: 2023. 7. 26.

장소: 비대면 화상회의

### 1. 생성 AI 사용 및 데이터 제약 관련

- 이번 경진대회의 경우 생성 AI 평가 방법 및 리더보드를 마련하는 것에 의의
- 1) 다만 최근 트렌드인 '초거대 규모'와 차이가 있는 접근 방법
  - 최근 생성 AI 트렌드와 반대가 되는 진행 내용 혹은 평가가 이루어질 우려 존재
- 2) 모델 용량 제한 관련
  - RTX 4080 24GB 1장으로 가능한 모델: 소규모일 것으로 예상
  - 올해는 과도기적 단계로, 경량화 조건 내 성능 향상 기술 사용 등을 측정하고자 함
    - 용량 조건 내에 인스트럭션 튜닝, 파인 튜닝 사용 가능
  - 또한 참가자 간 대규모 모델 운용 차이를 줄이고자 하는 목적
  - 향후 대규모 모델 사용 가능성 검토 예정

### 2. 평가 방식 관련

- 1) 정량적 평가 지표 운용 관련
  - 생성AI 결과물에 대해 전통적 지표 가중합으로 평가할 경우 다양성 누락 위험 존재
  - 상위 평가팀을 선발 용도로 사용 가능, 다만 다양성 측정을 위해서는 정성 평가 필요
  - 또한 최종 결과의 객관성 담보를 위해 상위 평가팀에 대해서도 정량적+정성적 평가를 함께 진행하는 것도 고려 가능
- 2) 평가 방법 고려: 리커트 척도, 랭킹 방식 등
  - 현재 주요 평가자: 학부생
    - 문장 수용성 차원에서 평가 가능, 평가 전문성은 검수자로 보강 예정
  - 단, 평가 결과의 신뢰성, 타당성 보장을 위해서는 평가 방식에 대한 고민 필요
  - 리커트 척도: 주요 평가자인 학부생들에게 평가 난이도가 높은 방법일 가능성 존재  
예) 비밀관적인 점수 매기기, 매긴 점수에 대한 타당성 문제 등 ...
  - 이에 따라 리커트 척도와 binary 랭킹 방식을 동시에 사용하는 방법 고려 가능
    - binary의 경우 두 대상에 대해 우위를 가리면 되므로 평가 난이도가 비교적 낮음
    - 수용성 측면에서 평가하되 균질한 평가, 평가 타당성을 위해 평가 방식 재고 필요
- 3) 장기적으로는 인간 평가를 자동화할 수 있는 방법 필요
  - 이를 위해 인간 평가 데이터 세트 수집 및 운용 로드맵 수립 필요
  - 개발된 평가지표를 통해 평가된 인간 평가 데이터 세트가 누적되면 향후 LLM 자동 평가 기준으로

활용 가능

- 최근 LLM 평가 트렌드 참고 가능: fastchat
  - 평가 아레나로, 서로 다른 LLM의 결과에 대해 스코어 비교 우위를 가릴 수 있음

### 3. 향후 생성 AI 활용 방향성 관련

#### 1) 외부 데이터 허용

- 최근 초거대 AI 기반 관점에서 보았을 때 외부 데이터도 허용하는 것이 필요
- 이에 따라 향후 별도의 트랙(비허용/허용)으로 운영하는 것 고려 가능

#### 2) 평가 자체에 대한 평가 방법론 경진대회

- 새로운 평가 방법론들이 많이 연구되고 있는 추세
- 평가 데이터 축적 및 트렌드에 따라 평가 방법론 자체에 대한 경진대회 진행 가능
- 경진대회 참여 주체의 전문성 (예: 연구자 등)을 높일 수 있을 것으로 기대

#### 3) LLM API 허용

- 향후 LLM들을 허용하는 것 역시 검토 가능
- 참고: ACL 챌린지 상위팀의 경우 대규모 LLM api를 사용
- 2)와 비슷하게 2가지 트랙(비허용/허용)으로 경진대회 운영 가능
- 최근 국내 LLM들의 개발이 이루어지고 있어 국내 LLM을 활용하는 방향도 고려 필요

### 4. 향후 진행 사항

- 상시과제 오픈, 평가체계 발전 방향 등에 대해 자문 예정
- 현재까지는 생성 AI 보안 이슈 해결에 한계점 존재
- 상시 이후에 3.에서 제안된 방법들에 대해 검토 및 운용 방향 구상 예정
- 자료 내 '경진대회 타게팅'에 대해서는 추후 자문 예정

## 국립국어원 인공지능 언어능력 평가체계 과제위원회 회의록

일시: 2023.9.7.

장소: 비대면 화상회의

### 1. 인간 평가 관련 논의

- 전반적으로 기존 피드백 잘 반영
- 절대평가, 상대 평가 상관관계에 대한 고려 필요

#### 1) 평가 기준 관련

- 문장 루브릭에 대한 이해도 저하 우려
- 이에 판단 기준 자체가 여러 가지로 해석되지 않도록 유의하는 것이 필요
- 평가 대상이 '기계가 생성한 것'임을 평정자들에게 알리는 것 권장

#### 2) 평정자 일치도 관련

- 평가자들을 대상으로 평가에 대한 워크숍 진행 필요
- 일치도 평가를 통해 일치도를 제고할 수 있도록 해야 함

#### 3) 실제 평가 전 시범 평가 관련

- 시뮬레이션을 통해 평가 과정 점검 가능

##### (1) 선호도 평가 방식

- 적합/부적합으로만 평가 시 적합/부적합 기준에 대해 공통적인 의견 수렴이 어려움
  - 즉 평정자 별로 평가 편차가 심해짐
  - 바이너리 평가가 아닌 스코어링 도입 검토 필요: 시범 평가 수행 必

##### (2) 토너먼트 방식

- 대진운이 따르는 방법, 퀄리티에 따른 분배가 중요
- 토너먼트 대상을 선정하기 위한 4분위수 진행 시 편차가 심하게 나타날 가능성도 존재
- 모든 200건을 한 팀과 대진하는 것이 아닌 문제별로 다른 팀과 경쟁하는 것도 좋은 방법  
예) a팀 1번 문제-b팀 1번 문제, a팀 2번 문제-c팀 2번 문제 등...

### 2. 상시 과제 개발 관련

#### 1) 부적합성 말뭉치 과제

- 부적절성을 탐지하는 과제는 중요할 것으로 예상

##### (1) 주관성 이슈 관련

- '부적절성' 자체의 명칭은 좋으나 사람들의 관점과 주관성에 따라 논란 가능
  - 정도성에 따라 공격성(offensive), 선정성 등 만을 다루는 것도 방법이나 완벽하게 논란 차단은 어려움
  - '부적절성'에 대한 기준을 마련하고 해당 기준에 준하여 과제 개발하는 것이 필요
- ex) 언어 사용자별로 부적절성에 대한 정도 차이, 맥락에 따른 차이 등...  
(중고등 학생의 언어생활, 즐거운 상황 속에서의 비속어 ...)

(2) 비명시적 표현 관련

- 비명시적 주석의 경우 주관적으로 느껴질 수 있음
- ‘큐’: 감성/정 분석 시 명시적 표현이 명확한 큐가 되는 경우가 많으므로 중요한 부분  
→ 또한 큐가 없는 케이스기에 일관성을 해칠 우려

(3) 데이터 세트 규모 점검 관련

- 명시성의 경우 기존 사전, 보고서를 기준으로 함  
→ 명시적인 것만 대상으로 하여 과제를 진행 시 너무 단순한 과제가 될 수 있음  
→ 또한 데이터 세트 규모를 살펴보았을 때 과제 진행이 가능한지도 확인 필요
- 데이터 세트 사이즈에 따른 튜트랙 운영도 고려 가능

(4) 부적절성 과제 개발 관련 사항

- 현재 데이터에 ‘맥락’이 없는 것이 문제가 될 가능성  
문맥을 고려하지 않은 부적절성 판단 사례가 존재할 수도 있음
- 개발 방향: chatGPT가 부적절성을 가릴 수 없는, 직관을 반영한 부적절성 과제  
현재 데이터 세트는 모든 경우를 아우르는 것처럼 보임
- 과제 난도는 상승하는 것이 바람직하며, 처음 시작임을 고려했을 때 과제 범위를 축소하는 것이 좋아보임

2) 추론 확신성 말뭉치 과제 개발 관련

- 실제 현실에서는 자연어 추론이 구조화되어있지 않은 경우가 많음
- 분류 과제는 소위 ‘라벨 찍기’가 가능하므로 과제로 운영하기에는 위험할 수 있음

3) 표/그림 기반 문장 생성 과제 개발 관련

- 정답과 다른 문장 생성 가능  
ex) 정답은 ‘SOS 신호가 쓰여져 있는 소화전과 빨간 벽돌이 보인다’이나 참가자들이 ‘빨간 벽돌’에 관심을 가지지 않을 수 있음  
→ 주변 정보에 대한 과제 확장 필요
- 현재 표 그림 설명(캡션) 생성 과제는 영어권에서는 옛날부터 수행했던 과제
- 지금은 더 복잡한 과제로 진화, 단 한국어로 된 과제의 시작으로 간주 가능  
→ 나중에는 더 어려운 과제를 조직하는 것도 중요

### 3. 상시 평가 운영 관련

#### 1) 리더보드 운영 관련

- 리더보드 결과 자체만 보면 API를 사용한 것인지 테스트 데이터 치팅을 한 것인지 확인 불가  
→ 리더보드 작성 시 라이브 평가 도입 가능
- 평가 기간 공고, 평가 서버 내 동시 평가 진행 등...

#### 2) 모델 평가 관련

- 로컬 서버에서 돌리는 모델에 대한 평가 방법론 필요
- 이번에는 모델 크기를 제한하지 않기에 평가에 어려움이 따를 것으로 예상
- 모델 평가 수행 시 재현성 관련 방법론 모색 필요

#### 3) 테스트 데이터 세트 관리 관련

- 테스트 데이터 유출로 인해 성능이 부정한 방법으로 상승할 가능성 존재
- PaperswithCode 사례처럼 테스트 데이터 공개와 더불어 해당 과제를 수행한 모델, 논문 등을 공유하는 플랫폼 도입 가능  
→ 사회적 가치 측면에서 바람직하나 참여도가 낮아질 수 있음  
→ 미도입 시 서약서 등의 보안 장치 마련 필요

#### 4) 참가 조건 제약 관련

- 상시 과제에 참여하려면 어쨌거나 모델과 일정 성능 이상의 서버가 필요  
→ 자원이 없는 참가자들의 참가 제한 가능  
→ 크레딧을 지원할 수 있는 회사와 협업하는 것도 좋은 방법

#### 5) 전반적 방향 관련

- 평가셋 공개, 모델 제출 등에 대해 보완 사항들이 존재
- 모델 크기에 따른 다중 트랙 등을 도입해볼 수 있음
- API 허용 가능성 재고 필요  
→ 부정행위 및 재현성 평가를 위해 API 사용 이력, prompt, 모델 등을 제출하는 것이 바람직  
자기 모델만 쓰는 것에 비해 허들이 낮아질수도 있을 듯함

## [부록4] 인공지능 평가 체계 발전 방안 자문의견서

1. 국립국어원은 2023년 인공지능 언어능력 평가체계 사업을 통해 사용자들이 모델의 언어 능력을 수시로 평가할 수 있는 인공지능(AI)말평의 상시 과제를 개설할 계획입니다. 현재 상시 과제 목록은 아래와 같습니다.

※ ‘인공지능(AI)말평’은 국립국어원의 인공지능 한국어 처리 능력 평가 체계(벤치마크)의 홍보 명칭(브랜드)임.

2022 상시과제 시범 운영 과제	2023 상시과제 운영 과제
<ul style="list-style-type: none"> <li>• 확산성 추론</li> <li>• 그림(사진) 기반 문장 생성</li> <li>• 속성기반 감성 분석</li> <li>• 혐오 발언 탐지</li> <li>• 표 기반 문장 생성</li> </ul>	<ul style="list-style-type: none"> <li>• 함의 분석</li> <li>• 문자가 포함된 이미지(OCR) 기반 문장 생성</li> <li>• 표의 일부분에 대한 해석 생성</li> <li>• 부적절성 문장에 대한 태도 탐지</li> </ul>

- 1) 위 과제 외에도 향후 국립국어원이 상시 과제 개설 시 생각해볼 수 있는 과제들에는 어떤 것들이 있을지 여쭙습니다.
- 2) 사용자들의 많은 참여가 이루어졌던 국립국어원 인공지능 언어능력 경진대회에 비해 상시 과제 참여율은 상대적으로 저조한 편입니다. 상시과제 참여율을 제고할 수 있는 방법에 대해 여쭙고자 합니다. 아울러 참고할 수 있는 벤치마크나 리더보드 등을 말씀해주주시면 감사하겠습니다.

자문  
위원1

○ AI말평에서 개발이 필요한 상시 과제 제안

새롭게 릴리즈 된 OpenAI의 GPT 모델의 Max Token이 32k 이상으로 늘어나고, 이로 인해 기존의 짧은 텍스트만을 처리할 수 있는 것이 아니라, 길고 다양한 형태의 텍스트들을 처리할 수 있는 상황입니다. Xiao et al., 2023의 연구에서 볼 수 있는 것처럼 늘어난 token에 따라서 기존의 attention 메커니즘 단순히 사용하는 것보다 나은 성능을 보이는 새로운 형태의 attention 메커니즘의 연구 결과도 나오고 있습니다.

이에 발맞추어 국립국어원의 과제에서도 현재 짧은 길이의 텍스트로 이루어진 과제뿐만 아니라, A4 용지로 몇 장에 이르는 길이의 텍스트와 같은 긴 텍스트에 대한 과제를 만들어 낼 수 있다면 관련 연구에 큰 도움이 될 수 있을 것으로 생각합니다.

늘어난 token 개수에 맞추어 새롭게 정의하는 과제는 단순히 길이에만 초점을 맞추는 것 이외에 현재는 테이블 형태의 데이터 등과 같은 텍스트의 구조를 대상으로 하는 것을 넘어서서 다양한 구조들의 텍스트를 정의하고 이를 과제화할 수도 있을 것으로 생각합니다. 예를 들어서 특허 문서, 정부 공문, 재판 기록들을 분석하고 이를 기반으로 한 결과물을 생성하기 위한 과제를 만들어 볼 수 있을 것입니다. 또한, 이와 반대로 짧은 텍스트에서 구조화 된 문서를 생성하기 위한 과제도 생각해 볼 수 있을 것입니다. 짧은 텍스트로부터 길고 구조화 된



	<p>텍스트를 생성하기 위한 과제에서는 Task의 정의뿐만 아니라, 이에 따른 새로운 형태의 평가 지표 역시 새롭게 만들어 볼 수 있는 것이 많은 거로 생각합니다.</p> <p>○ AI말평의 상시 과제 활성화 방안 제안</p> <p>참여자 입장에서 생각해 보았을 때, 상시 과제의 경우 관심을 받을 가능성이 낮아 참여 동기가 약하다고 생각합니다. 이미 운영 중에 있는 Leader board의 운영으로 경진대회와 같은 기회가 없이도 상시적으로 참여자들의 모델의 평가 결과가 대중적으로 할 수 있도록 해줌으로써 참여자의 관심을 끌 수 있는 가장 좋은 방법이라고 생각합니다. 이에 더해 현재의 news letter 등과 같은 좀 더 상시 과제의 적극 참여를 유도하기 위한 운영 방식의 개선 정도가 가능하지 않을까요?</p> <p>○ AI말평 운영에 참고할 수 있는 벤치마크나 리더보드</p> <p>구글링으로 찾을 수 있는 leaderboard 이외의 독특한 leaderboard는 알고 있는 것이 없습니다.</p> <p>[1] Xiao, Guangxuan, et al. "Efficient streaming language models with attention sinks." arXiv preprint arXiv:2309.17453 (2023).</p>
<p>자문 위원 2</p>	<p>○ AI말평에서 개발이 필요한 상시 과제 제안</p> <p>인공 지능의 한국어 처리 능력 평가 체계를 주제로 가져간다면 국립국어원의 설립 취지에 맞게 “국어의 발전과 국민의 언어생활을 향상하는 연구 사업”에 활용할 수 있는 인공 지능 개발과 개발이라는 목표를 설정하고, 이를 달성하기 위하여 어떠한 능력을 실험할 것인지에 집중하는 것이 좋을 수 있을 것 같습니다.</p> <p>현재는 범용인공지능(AGI)의 개발에 전체적인 인공지능 개발의 방향이 맞춰져 가는 것이 흐름이라고 생각하는데요. 인공 지능의 능력을 평가하는 것은 앞으로 언어 능력에 국한되지 않고 멀티 모달로 진화할 것이라고 생각합니다. 현재도 LLM을 넘어서서 LMM으로 나아가고 있고, 구글의 gemeni같은 경우도 멀티모달 벤치마크에 대한 성능을 리포팅하였습니다. 이런 흐름을 잘 반영하여 말평의 과제의 영역을 발전시켜 왔다고 생각하는데요.</p> <p>그럼에도 불구하고, 아직 이렇게 평가되는 인공지능들이 어떻게 한국어 연구에 활용될 수 있는 지에 대해서 국립국어원의 언어정보나눔터 이용자들이 직관적으로 이해하기는 어렵다는 생각이 듭니다.</p> <p>따라서 세부 과제로 근대국어 이해능력 평가나, 음성(방언) 이해나 처리 벤치마크, 아시아 언어 번역 평가 등 기존 국어원에서 진행해온 연구 사업 중 말뭉치 구축 사업의 영역을 넘어서는 새로운 아이디어를 공모받아도 좋을 것 같습니다.</p> <p>○ AI말평의 상시 과제 활성화 방안 제안</p> <p>앞서 이야기 한대로, 기존의 인공지능 모델 개발을 중점으로 하는 연구자 또는 학생 층만을 대상으로 하지 않고, 국어 연구에 뜻이 있는 연구자들이 관심을 가질 수 있는 연구 주제를</p>

	<p>설정하고, 국어 연구에 활용할 수 있는 과제를 설정하면 상시 과제에 대한 관심을 제고할 수 있을 것으로 생각합니다.</p> <p>국어정보화 방법론 자체보다 국어정보화를 통해 획득된 인공지능 기술을 어떻게 활용할 수 있을지에 초점을 맞추어 과제를 새로 생성하는 것이 더 유익할 것으로 생각합니다.</p> <p>○ AI말평 운영에 참고할 수 있는 벤치마크나 리더보드</p> <p>실질적인 운영 측면에서 접근성이나 대중성 모두를 갖춘 리더보드는 허깅페이스에서 운영하는 리더보드(<a href="https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard">https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard</a>) 일 것입니다. 한국어 관련해서는 저희 회사에서 허깅페이스 플랫폼과 협업한 한국어 리더보드가 운영되고 있습니다.(<a href="https://huggingface.co/spaces/upstage/open-ko-llm-leaderboard">https://huggingface.co/spaces/upstage/open-ko-llm-leaderboard</a>) 벤치마크 데이터 자체는 위 리더보드에서 메인으로 삼고 있는 데이터에 더하여 구글의 gemini의 성능 보고에 활용된 데이터들도 함께 검토할 수 있을 것입니다.(<a href="https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf">https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf</a>) 추후 ACL 계열 컨퍼런스의 대안으로 올해부터 개최될 COLM에서 공유될 데이터와 논의들도 검토된다면 도움이 될 것으로 생각합니다. 최예진 교수님 등이 구성한 신규 컨퍼런스입니다.(<a href="https://colmweb.org/cfp.html">https://colmweb.org/cfp.html</a>)</p>
자문 의견3	<p>○ AI말평에서 개발이 필요한 상시 과제 제안</p> <ul style="list-style-type: none"> <li>- 각 과제의 연구 진보 성과를 집성하여 결론을 어느 정도 내리지 못한다면 계속 진행하도록 함. 리더보드를 만드는 이유가 그 과제 혹은 분야의 연구 방향 집성을 하려고 하는 것임.</li> </ul> <p>○ AI말평의 상시 과제 활성화 방안 제안</p> <ul style="list-style-type: none"> <li>- 과제 분야별로 가장 연구발표가 활발한 연구회/컨퍼런스에서 해당 과제에 대한 워크숍이 이루어질수 있도록 한다.</li> <li>- 국제적 워크숍을 각 과제 분야별로 개최할수 있도록 한다. 이렇게 하기 위해서는 국내외 협력 체계가 필요하다.</li> <li>- 문제를 영어로도 공지하여, 외국에서도 접근할 수 있도록 한다.</li> </ul> <p>○ AI말평 운영에 참고할 수 있는 벤치마크나 리더보드</p> <ul style="list-style-type: none"> <li>- NTCIR (최근 과제 NTCIR-17): <a href="https://research.nii.ac.jp/ntcir/ntcir-17/index.html">https://research.nii.ac.jp/ntcir/ntcir-17/index.html</a></li> <li>- 각 과제관리자 그룹을 별도로 두어, 리더보드에 올린 각 과제팀들이 발표를 할수있도록 하고, 그 과제관리 그룹이 1회성이 아닌 2-3년 정도의 과제 관리를 하여 연구진보의 한 집성을 할 수 있도록 한다.</li> </ul>
자문 의견4	<p>○ AI말평에서 개발이 필요한 상시 과제 제안</p> <ul style="list-style-type: none"> <li>- 국어원의 기관 역할과 관련된 과제를 운영하는 게 당위성이 있을 것으로 생각합니다.</li> <li>- 예를 들어, (단락 단위) 맞춤법 교정 글 생성 과제, 온라인 게시글 말뭉치에서 신조어 탐색 및 정의문 자동 생성 과제 등이 타 벤치마크 과제 대비 차별점이 있을 것으로 생각합니다.</li> </ul> <p>○ AI말평의 상시 과제 활성화 방안 제안</p> <ul style="list-style-type: none"> <li>- 단순한 챌린지 개최 이상의 구체적인 타겟팅이 필요해 보입니다. 주 참여 대상으로 AI전공 대학생들을 목표로 할지, AI스타트업을 목표로 할지, 네이버/카카오 등 AI 대기업을 목표로</li> </ul>

	<p>할지를 먼저 구체화하고, 각 타겟층에 맞는 보상 방안 설계가 필요해 보입니다.</p> <ul style="list-style-type: none"> <li>- 예를 들어, AI전공 대학생을 주 참여 대상으로 설정한다면, 취업 지원 또는 학과 과정(수업 텀프로젝트 대체 등)과 연계한 보상 방안 설계가 적절해 보입니다.</li> <li>o AI말평 운영에 참고할 수 있는 벤치마크나 리더보드</li> <li>- Huggingface LLM Leaderboard 등 해외 유명 벤치마크는 이미 충분히 파악하셨을 것으로 생각합니다.</li> </ul>
--	--

2. 2023년 이후 국립국어원에서는 인공지능 평가를 위한 평가체계 말뭉치를 아래와 같이 계획 중에 있습니다. 이에 말뭉치 구축 관련 사안에 대해 자문을 여쭙고자 합니다.

구분	주요 평가 항목
한국어 능력 종합	문법 정확도 및 복잡도
	주어진 맥락에 대한 문장 의미 파악 등 한국어 의미 종합 추론 능력
한국어 대화 능력	한국어 특성을 반영한 대화 이해 능력 평가
	다층적 대화(멀티턴) 기반 대화 이해 및 생성 능력
한국언어문화 이해 능력	한국언어문화 이해 질의응답 능력
	지역별, 연령별 다양한 배경의 언어 이해·생성 능력
부적절 표현 탐지 능력	부적절 표현 탐지 및 분류 능력(필수 과제로 매년 개발)

<표> 평가 과제 구축 관련 중장기 계획 기본 방향

1) 위 인공지능 벤치마크 데이터셋을 마련하는 데 있어 기존과 같이 클라우드 소싱을 통해 구축해야 할지, 아니면 소수의 주석 전문가를 사용한 양질의 소규모 한국어 데이터셋을 구축해야 하는 것이 좋을지 여쭙습니다.

2) 아울러 적절한 인공지능 벤치마크 데이터셋 규모에 대해서도 여쭙고자 합니다. (현재 계획: 데이터셋별 instance 1000건 내외)

3) 국립국어원 인공지능 벤치마크 데이터셋 개발 시 해외의 다양한 벤치마크 데이터셋, 혹은 기구축 instruction 말뭉치들을 어떤 방식으로 활용할 수 있을지, 그리고 자문위원회에서 국립국어원에 기대하는 한국어/한국 문화 특화 데이터셋이 있는지도 여쭙습니다.

자문  
의견1

○ 벤치마크 데이터셋 구축 시 구축 방법 (클라우드 소싱 vs 소수 전문가)  
앞선 항목에서 제시한 더 긴 길이의 task 혹은 다양한 구조의 task에 대한 평가 데이터의 경우 더 고도의 전문적인 지식을 요하므로 소수 전문가를 통한 데이터 구축이 적합하다고 생각합니다. 다만, 제한된 자원의 문제를 극복하기 위해서, 클라우드 소싱으로 초기 데이터셋의 구축후 소수 전문가의 검토를 받는 방식도 가능할 것으로 생각됩니다.

○ 적절한 벤치마크 데이터셋 규모  
제시된 규모가 적절하다고 생각합니다.

○ 기존 해외 벤치마크 데이터셋, instruction 말뭉치 활용 가능성  
기존의 해외 벤치마크 데이터셋을 번역하여 활용하는 방법의 경우 parameter 볼륨이 작은 모델에서는 문제가 없지만, 모델의 크기가 커지면 번역 오류의 전파로 인한 문제가 무시할 수 없을 정도가 될 수 있다고 생각합니다. 따라서, 해외 벤치마크를 번역하고 이를 다시 한번

	<p>정교하게 검수한 데이터가 있다면 연구하시는 분들에게 큰 도움이 될 수 있다고 생각합니다.</p> <p>번역 결과물에 대한 검수에 더해서, 자동 생성된 chain-of-thought에 대한 검수 데이터가 있다면 이 역시 큰 도움이 될 것이라 생각합니다.</p> <p>○ 국립국어원에 기대하는 한국어/한국 문화 특화 데이터셋</p> <p>지난 회의에 나온 것처럼 방언, 연령대별 표현에 대한 병렬 데이터가 있다면 유용할 것이라는 점은 자명하다고 생각합니다.</p>
<p>자문 의견 2</p>	<p>○ 벤치마크 데이터셋 구축 시 구축 방법 (클라우드 소싱 vs 소수 전문가)</p> <p>이제 대규모의 데이터셋 구축은 생성형 인공지능을 활용한 합성데이터로 대체될 것이라고 생각합니다. 이에 따라 전문 지식을 갖춘 양질의 데이터셋 구축에 집중하는 것이 차별화 방식이 될 것입니다.</p> <p>○ 적절한 벤치마크 데이터셋 규모</p> <p>과제에 따라 다를 것이라고 생각합니다만, 실질적으로 연내에 실현가능한 규모로 제시된 규모가 적합하다고 생각합니다.</p> <p>○ 기존 해외 벤치마크 데이터셋, instruction 말뭉치 활용 가능성</p> <p>요새는 기존에 고전적인 과제들 이외에 다양한 활용 목적에 맞춰서 새로운 과제를 제안하는 경우가 늘어나고 있습니다. 따라서 해외 벤치마크나 기존 존재하는 말뭉치를 단순히 번역하는 것보다는 국립국어원 특화 과제를 만드는 것이 더 유용할 것이라고 생각합니다.</p> <p>○ 국립국어원에 기대하는 한국어/한국 문화 특화 데이터셋</p> <p>디지털 인문학 분야의 수요를 조사하여 고전 국어 및 근대 국어 등 현대 한국어와 문자셋(유니코드)가 다른 자료들이 구축된다면 연구의 지평을 넓히는데 도움이 되리라 생각합니다.</p> <p>이외에도 CJK(중국어, 일본어, 한국어)를 넘어서서 동남아시아권의 언어까지 영역을 넓힌 병렬 말뭉치들이 구축된다면, 한국어 교육 이외에도 다양한 어플리케이션이 나오는데 기여할 수 있을 것이라고 생각합니다.</p>
<p>자문 의견3</p>	<p>○ 벤치마크 데이터셋 구축 시 구축 방법 (클라우드 소싱 vs 소수 전문가)</p> <ul style="list-style-type: none"> <li>- 클라우드소싱을 하되 소수 전문가의 검수를 받는 형식</li> <li>- 과제에 따라 자격 조건으로서 전문성을 평가하도록 함.</li> </ul> <p>○ 적절한 벤치마크 데이터셋 규모</p> <ul style="list-style-type: none"> <li>- 분포가 편중되지 않도록</li> </ul> <p>○ 기존 해외 벤치마크 데이터셋, instruction 말뭉치 활용 가능성</p> <ul style="list-style-type: none"> <li>- (의견 없음)</li> </ul> <p>○ 국립국어원에 기대하는 한국어/한국 문화 특화 데이터셋</p> <ul style="list-style-type: none"> <li>- 한국어 시조 짓기</li> </ul>

<p>자문 의견 4</p>	<ul style="list-style-type: none"> <li>○ 벤치마크 데이터셋 구축 시 구축 방법 (클라우드 소싱 vs 소수 전문가)</li> <li>- 클라우드 소싱 vs 전문가 구축 데이터는 양 vs 질의 이슈로 생각되고, 최근 우수한 LLM 학습에 질이 중요하다는 연구 결과를 고려하면, 질이 우수한 데이터 구축이 필요하다고 생각합니다.</li> <li>○ 적절한 벤치마크 데이터셋 규모</li> <li>- 데이터셋의 규모는 과제의 난이도에 따라 다를 수 있음을 전제로 이야기 드리고,</li> <li>- 쉬운 과제라면 1천개도 충분하지만, 어려운 과제라면 1만개 이상의 데이터가 필요할 수 있습니다. (예: 에세이 자동 채점)</li> <li>○ 기존 해외 벤치마크 데이터셋, instruction 말뭉치 활용 가능성</li> <li>- 역시 과제 별로 차이가 있을 것 같습니다.</li> <li>- 공개 해외 데이터의 활용이 가능하다면, 활용하여 해당 데이터셋의 한국어/영어 성능을 비교하는 것도 의미가 있다고 생각합니다.</li> <li>○ 국립국어원에 기대하는 한국어/한국 문화 특화 데이터셋</li> <li>- 한국어 내에 존재하는 다양성을 보존하는 말뭉치 및 평가 체계가 있으면 좋겠습니다.</li> <li>- 예를 들어, 지역 별 방언에 대한 이해 및 생성 능력 평가, 연령 별 대화 데이터셋, 한국 문화에 대한 질의응답 데이터셋 등이 구축, 활용되면 좋겠습니다.</li> </ul>
--------------------	---

3. 국립국어원은 인공지능(AI)말평의 2023년 인공지능 언어능력 평가 경진대회 운영 당시 참가자들의 생성 AI 모델 학습 및 구동 사양을 'RTX 4090 24GB 1개'로 제한하여 대회를 개최한 바 있습니다. 향후 경진대회 운영에 대해 현재 3가지 방향으로 논의를 진행 중에 있습니다.

<방향>

- 1) 2023년과 같이 참가자들이 생성 AI 모델로 참가 시 스펙 제한 트랙으로 운영
- 2) 모델 훈련 및 구동에 대한 스펙 무제한 트랙으로 운영
- 3) 스펙 제한 / 스펙 무제한 두 트랙 운영

이에 위 세 가지 방향에 대해 자문을 여쭙고자 합니다.

자문 의견1	단일 트랙이든 혹은 두 트랙 운영이든, 스펙 제한 트랙의 운영은 공정한 경쟁이라는 관점에서 필요하다고 생각합니다. 다만 현재의 RTX 4090 24GB 라는 제약 조건을 올리는 것은 필요하다고 생각합니다. A100과 같이 단순히 메모리량이 더 큰 하드웨어를 사용할 수 있게 해주는 방향 뿐만 아니라, GPU 여러 장을 사용한 Model Parallel을 하여 참여하는 것도 허용을 하면서 평가 척도에 TPS와 같은 속도에 관한 평가척도를 추가하여 LLM을 위한 인프라를 배우고 연구하는 분들의 관심도 유도를 할 수 있으면 좋을 것으로 생각합니다.
자문 의견2	기업에서 별도의 지원을 받아서 인프라를 제공할 수 없다면, 누구나 참여 가능한 수준에서 스펙을 제한 하는 것이 바람직하다고 생각합니다. 또한, 최근 리더보드 순위를 올리는 것이 과열되었다는 지적이 이어지면서 불필요하게 많은 탄소를 배출하는 것에 대한 자성의 목소리도 나오는 상황입니다. 요 근래 허깅페이스 llm 리더보드 상위권에 랭킹된 모델에 stop carbon이나 CarbonVillain 같은 네이밍이 등장하기도 하였고요. 따라서 자원을 절제하여 사용하면서 경량화, 효율화에 신경쓴 모델이 높은 점수를 받아갈 수 있도록 하는 구조를 만드는 것이 더 좋다는 의견을 남깁니다.
자문 의견3	좋음.
자문 의견4	<ul style="list-style-type: none"> <li>○ 가능하다면 2번과 같이 H/W 환경의 제약 없이 운영하는 게 참가자 입장에서는 가장 좋습니다.</li> <li>○ 하지만, 개최자 입장에서는 예상 비용 산정도 어렵고, 행사 개최에 많은 어려움이 따를 것입니다. 그리고, 모델이 너무 크다면 추후 실서비스 적용 시에도 고비용을 감당해야 하는 어려움이 있습니다.</li> <li>○ 국립국어원에서 컴퓨팅 리소스 감당이 가능하다면 2번과 같이 무제한 트랙이 좋으나,</li> <li>○ 현실적으로 어려움이 있다면, 1번과 같이 운영하여도 충분히 의미가 있을 것으로 생각합니다.</li> </ul>

<p>4. 본격적으로 인공지능 벤치마크를 운영하게 될 경우 모델 성능에 대한 ‘평가 방법 및 지표’ 측면 역시 반드시 고려해야 할 것 같습니다. 이에 아래 사안에 대해 자문을 구하고자 합니다.</p> <p>1) 인공지능 벤치마크의 경우 ‘텍스트 생성’ 과제 결과에 대한 정성적 평가 수요가 높아지고 있습니다. 이에 대해서는 ① 전문가 수동(manual) 평가와 ② G-Eval을 검토하려 합니다. 두 가지를 인공지능 벤치마크에 효율적으로 사용할 수 있는 방법에 대해 고견을 여쭙습니다.</p> <p>2) 앞서 여쭙 바와 같이 국립국어원의 경우 크게 한국어(언어), 한국 문화 두 가지 방향으로 인공지능 벤치마크 과제를 구성 중에 있습니다. 이에 인공지능의 성능을 ① 과제별 성능 지표로 제시해야 할지, 아니면 ② 통합 성능 지표로 제시해야 할지 여쭙고자 합니다.</p>	
<p>자문 의견1</p>	<p>○자연어 생성 결과에 대한 정성적 평가 방법 활용 방안</p> <p>이 부분에 대해서는 현재 정성적인 평가의 중요성에 대해서는 공감하지만, 어떤 식으로 평가의 방향이 정해져야 할지에 대해서는 아이디어가 없습니다.</p> <p>○인공지능의 과제 성능 표기 방법 (과제별 성능 지표 vs 통합 성능 지표)</p> <p>한국 문화를 기준으로 한 성능 지표는 여러 과제에 공통으로 사용될 수 있는 통합 성능 지표를 정의하고 만드는 것은 힘든 목표이지 않을까 생각합니다.</p>
<p>자문 의견2</p>	<p>○자연어 생성 결과에 대한 정성적 평가 방법 활용 방안</p> <p>최대한 공정과 채점 방식을 단순화하고, 여러 사람이 직관적으로 평가한 내용에 대해서 IAA를 체크하는 방식으로 진행할 수 있는 구조를 만들어야 운영 및 관리의 효율성을 높일 수 있다고 생각합니다.</p> <p>○인공지능의 과제 성능 표기 방법 (과제별 성능 지표 vs 통합 성능 지표)</p> <p>통합 성능을 유의미하게 사용하기 어려울 것으로 판단됩니다. 과제별 성능 지표를 나타내는 것으로 충분할 것입니다. 해석하는 쪽에서 유의미한 결과를 도출할 수 있도록 상세한 지침이 제시되면 좋을 것입니다.</p>
<p>자문 의견3</p>	<p>○자연어 생성 결과에 대한 정성적 평가 방법 활용 방안</p> <p>- (의견없음)</p> <p>○인공지능의 과제 성능 표기 방법 (과제별 성능 지표 vs 통합 성능 지표)</p> <p>- 과제별 성능 지표</p>



<p><b>자문 의견4</b></p>	<p>○자연어 생성 결과에 대한 정성적 평가 방법 활용 방안</p> <ul style="list-style-type: none"> <li>- 최근 해외 벤치마크와 유사하게 G-Eval 평가 활용 방안에 찬성합니다.</li> <li>- 전문가 수동 평가는 ‘수시’ 과제에는 가능하지만, ‘상시’ 과제에는 가능하지 않을 것으로 예상되고,</li> <li>- G-Eval 결과와 전문가 수동 평가 결과의 경향성을 근거 자료로 제시하면 어떨지 의견 드립니다.</li> </ul> <p>○인공지능의 과제 성능 표기 방법 (과제별 성능 지표 vs 통합 성능 지표)</p> <ul style="list-style-type: none"> <li>- 각 과제 자체의 가치를 생각하면, 과제별 성능 지표 제시는 반드시 필요해 보입니다.</li> <li>- 다만, 순위화 목적으로 통합 성능 지표가 필요하다면, 평균 등의 방법을 사용하여 통합 성능 지표도 같이 산정하는 게 어떨지 의견 드립니다.</li> </ul>
--------------------------	---



<기획·연구>

국립국어원 강미영 언어정보과장

국립국어원 이보라미 학예연구관

국립국어원 장연지 연구원

국립국어원 한송이 연구원

국립국어원 김예영 연구원

<연구 참여자>

연구책임자 김한샘(연세대학교)

공동연구원 송상현, 홍승혜(고려대학교),

박승희, 송영숙, 유현조, 정유남(나라지식정보)

함영균, 임경태(테디썸),

윤영민, 여진영, FEI LI(연세대학교),

나승훈(전북대학교), 김현명(마이스앤드)

연구보조원 노강산, 황동진(고려대학교),

정용빈, 이이슬, 박재완, 오유진, 서현빈(테디썸),,

박서윤, 강예지, 이재원, 김유진, 강조은(연세대학교)

---

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2023년 12월 20일

발행일: 2023년 12월 20일

인 쇄: 연이프린텍

---

※ 이 보고서는 국립국어원의 용역비로 수행한 ‘2023 인공 지능의 한국어 처리 능력 평가 체계 운영 및 평가 과제 구축’ 사업의 결과물을 발간한 것입니다.

