FAQ

생성일	@2025년 7월 1일 오후 9:10
최종 편집 일시	@2025년 7월 10일 오후 4:47
태그	대회 공고

▷자주 찾는 질문(공통)

Q: 대회 참여는 개인 또는 팀 모두 가능한가요?

A: 네, 한국어 인공지능에 관심 있는 개인 또는 단체라면 누구나 참가할 수 있습니다.

Q. 팀으로 참여 시 팀원 수에 제한이 있나요?

A: 본 대회에서는 팀원 수를 제한하고 있지 않습니다.

Q. 한 사람이 여러 팀에 소속되어 참가할 수 있나요?

A: 답안 제출 시에는 한 사람이 여러 팀에 소속되어도 되지만 수상 시에는 소속 팀 중 반드시 하나의 팀을 선택해야 합니다.

Q: 대회 종료 후 모범 답안을 공개하나요?

A: 모범 답안은 별도로 공개하지 않습니다.

Q: 모델 학습 시 RTX 4090 24GB보다 상위 스펙의 GPU를 사용할 수 있나요?

A: 학습은 어떠한 환경에서 진행돼도 괜찮습니다. 단, 추론 시 VRAM 24GB 이하 환경에서 구동할 수 있어야 합니다.

Q: 데이터 세트 모두를 학습에 사용해도 되나요?

A: 참가 신청한 과제의 데이터(train, validation)는 모두 학습에 자유롭게 사용할 수 있습니다. 단, 시험(test) 데이터는 학습에 이용할 수 없습니다.

Q: Chat GPT API를 사용하여 데이터 증강을 해도 되나요?

A: 네 가능합니다. (나) 유형에서는 평가 데이터를 사용하는 것을 제외한 모든 방식의 데이터 증강이 허용됩니다.※ 증강된 데이터 사용 시 저작권 등의 데이터에 대한 책임 문제는 참가팀에게 있습니다.

Q: (가) 유형(데이터 변형/증강 불가)과 (나) 유형(데이터 변형/증강 허용)은 무슨 차이인가요?

A: (가) 유형은 (증강 등의 방법 없이) 주어진 데이터만 활용해야 하고, (나) 유형은 평가 데이터를 제외한 모든 데이터를 (증강 등의 기법에) 활용할 수 있습니다.

Q: (가) 유형(데이터 변형/증강 불가)에서 "모델 입력을 위한 형식 및 형태 변형만이 허용된다."의 허용 범위가 어떻게 되나요?

A: (가) 유형의 대회 규정에서 정의하고 있는 데이터 형식 및 형태 변형은 학습에 사용될 데이터 구조를 변경하는 것을 의미합니다.- (가) 유형에 참가하시는 경우 모델 학습에 활용할 입 력 데이터의 부분적 제거(제거 전처리, 불용어 처리)는 가능하지만, 입력 데이터의 내용을 수 정하거나 내용을 추가하는 것은 불가능합니다.- (나) 유형에 참가하시면 데이터에 대한 변형 및 증강을 자유롭게 하실 수 있습니다.

※ "입력 데이터의 형식 및 형태 변형"의 허용 범위

허용- 데이터의 제거 전처리(특수문자, 한글 자음모음 제거, 특정 단어 제거)- 불용어 처리

불가- 데이터의 수정 전처리(예: name1 → 화자1)- 입력하는 텍스트 데이터 내용 변경- 불용어를 다른 토큰으로 변경- 학습 데이터의 추가적인 태깅- 프롬프트에 입력 외 다른 정보 포함- 학습 데이터를 제외한 외부 데이터를 사용하여 샘플 추출 후 데이터 변형하여 사용

Q: (가) 유형(데이터 변형/증강 불가)에서 LLM을 이용한 증강이 아닌 셔플, 믹스, 리버스 등 알고리즘으로 학습 데이터를 수정해도 되나요?

A: 불가능합니다. LLM을 이용한 증강뿐만 아니라 어떠한 방법을 통한 증강이 불가능합니다. (나) 유형에서 자유롭게 증강 기법을 적용하시길 권장드립니다.

Q: 베이스라인 코드에서 제시한 프롬프트와 다른 프롬프트를 사용할 수 있나요?

A: 가능합니다. 자유롭게 프롬프트를 구성할 수 있습니다. 다만 프롬프트를 구성할 때 주어진 입력에 다른 정보가 들어가면 안 되니 주의 부탁드립니다.

Q: 말뭉치 데이터를 이용하여 Foundation LLM 모델을 파인튜닝해도 될까요?

A: Foundation 모델에 참가 신청한 과제의 데이터 세트(train, dev)를 활용하여 Fine-tuning할 수 있습니다.

Q: (나) 유형의 "외부 API를 통해 호출하는 모델(OpenAI API 등)은 제출할 수 없다."가 어떤 의미일까요?

A: 데이터 증강을 위해 외부 API를 이용하는 것은 허용하지만, 시스템이 테스트 데이터에 대해 추론을 수행하는 과정에서 외부 API를 이용하는 것은 불가합니다.즉, 모든 추론 과정은 로컬 환경에서 완결되어야 합니다.

Q: 주어진 베이스라인 이외의 다른 모델을 사용해도 되나요?

A: 제공한 코드는 베이스라인으로 제공한 코드이며 추론 방법은 어떠한 방식을 사용하셔도 좋습니다. 베이스라인의 방법은 하나의 예시일 뿐이며, 더 좋은 방법을 사용하여 좋은 모델을 만드시길 기대합니다. 다만 (가) 유형은 데이터 증강이 허용되지 않습니다. 따라서 제한된 데이터 이외의 데이터를 이용하여 학습한 후 공개하여 베이스 모델로 활용하는 경우는 규정에 어긋난다고 판단합니다. 즉 대회가 시작되기 전 공개된 모델은 모두 이용 가능 하지만, 대회가 시작되고 난 후 공개된 모델(25.6.11 이후) 중 본 과제를 위해 특별히 학습된 사전 학습 모델은 허용되지 않습니다. (주어진 학습 데이터 외의 다른 학습데이터를 활용했다 고 판단합니다) 이에 대해서는 수상 대상자를 선정하기 위한 정성 평가 단계에서 내부 전문가가 정성적으로 검증할 예정입니다.

Q: 모델을 개발할 때 라이선스에 문제가 없어야 한다고 하는데 무슨 뜻일까요?

A: 참가자가 제출한 모델의 저작권은 참가자에게 있으며, 국립국어원은 참가자가 제출한 모델을 상업적으로 활용하지 않습니다(참가자가 사용한 모델 및 이용 기술 등은 홍보용으로 공개할 수 있음).

또한, 접수 기간 종료 후 평가 상위 팀에 대하여 개발하신 모델을 공개하도록 요청할 계획입니다.

이를 고려하여 라이선스에 문제가 없는 모델을 선정하여 과제에 참가하여 주시기를 바랍니다.

Q: 여러 모델을 이용하여 모델 결과를 앙상블하는 앙상블 모델 추론이 가능한가요?

A: 이번 대회에서는 앙상블 모델 제출을 제한합니다. 안내해 드린 바와 같이 이번 대회의 제출 모델은 RTX4090 24GB 1개에서 구동(추론) 가능하여야 하며, 접수 기간 종료 후 정성평가 신청서를 제출한 상위 10개 팀에 대하여 모델의 재현성 및 우수성 평가, 발표 평가 등을 거쳐 수상자를 선정합니다.

제한된 조건 내에서 앙상블 모델을 활용하는 경우 여러 개의 모델로 순차적으로 추론한 후 최적의 답변을 선택하게 되어, 추론 시간이 단일 모델에 비해 현저히 길어지게 되므로, 모델 정성 평가 시 매우 낮은 점수를 받을 가능성이 높습니다.

특히 '한국어 어문 규범 기반 생성'과제의 경우, 국어 전문가가 모델을 직접 사용해 보면서 평가를 진행하게 되므로(챗봇 아레나 방식) 추론 속도는 매우 중요합니다. 그 밖에도 정성 평가 시 단일 모델과 앙상블 모델을 같은 선상에 놓고 비교하기 어려운 문제 등이 발생할 수 있습니다. 이와 같은 이유로 인하여 올해 대회에서는 앙상블 모델 제출을 제한하게 되었습니다.

Q: 한국어 어문 규범 기반 생성(RAG) 과제에서 추론 모델(LLM) 외 임베딩 모델도 VRAM 24GB 제한에 포함되나요?

A: 임베딩 모델의 경우 24GB 제한에 포함되지 않으며 임베딩 모델의 실행 환경 및 방식은 참가자 재량에 따라 자유롭게 선택 가능합니다.

다만 주 추론 모델은 VRAM 24GB에서 단일 모델로 구동(추론) 가능하여야 하는 점 유의해 주시기 바랍니다.

▷한국어 어문 규범 기반 생성(RAG)(가 유형)

Q: 학습 데이터와 주어진 참조 문서를 결합해 튜닝하는 방식(프롬프트 정보에 더하는 방식)은 가능할까요?

A: 가능합니다. 규정상 (가) 유형(데이터 변형/증강 불가) 과제의 경우에는 “모델 입력을 위한 형 식 및 형태 변형만이 허용된다.”라고 명시되어 있습니다. 따라서 새로운 프롬프트 구조를 만 들어서 모델을 튜닝하는 것은 창의적인 학습 전략으로 간주하여 허용됩니다. 다만 학습 데이터의 내용을 사람이나 GPT가 변형하거나 증강하는 것은 허용되지 않습니다.

Q: 말뭉치외에 "한국어 어문 규범 기반 생성(RAG) 참조 문서"는 데이터 증강처리해도 되나요?, 로컬 LLM을 이용(외부 통신 X, 추론 모델과 동일)하여 내부적으로 온톨로지를 구축하고 RAG에 이용하는건 문제가 없을까요?

A: 참조문서를 증강하는 것은 허용되지 않습니다. 규정상 가 유형(데이터 변형/증강 불가) 과제의 경우에는 “과제 데이터(말뭉치)를 거대언어모델(LLM) 등을 활용해 변형 및 증강하여 사용할 수 없다(㉮ 유형)” 라고 명시되어 있습니다. 참조 문서 역시 여기에 포함됩니다. 다만 참조문서를 지식 구조(온톨로지, 지식 그래프)로 변환하는 것은 허용됩니다. 그러나 정보 의 출처는 오로지 대회에서 제공된 데이터만을 기반으로만 하여야 합니다.

Q: 말뭉치의 train, dev, json 의 질문과 답변 내용을 입력의 few shot 예제로 이용해도 괜찮을까요?

A: train, dev 내에 포함된 데이터라면 이용 가능합니다. 다만 이 과정에서 입력 데이터의 변형은 일어나서는 안 됩니다.

Q: Retrieval도 질문 변형 없이 사용해야 하나요? 문제 생성하여 학습할 때 추가 문제를 증강하면 안 되나요?

A: RAG 사용 시, DB 검색을 위한 쿼리(질문)를 변형하여 사용하는 것은 허용됩니다. 이는 실시 간으로 입력된 질문을 처리하는 모델의 내부 로직으로 판단할 수 있습니다. 문제 생성하여 학습할 때 추가 문제를 증강하는 것은 허용되지 않습니다. (가) 유형(데이터 변 형/증강 불가) 과제는 데이터가 동일할 때, 모델의 튜닝 방법이나 알고리즘 등에 의해 성능을 향상시키는 것에 초점을 맞춘 과제입니다. 따라서 질문의 변형이나 증강은 불가합니다.