Tulu3 학습 데이터

Tulu3는 다양한 데이터셋을 활용하여 학습된 대규모 언어 모델입니다. 이 문서에서는 Tulu3의 학습 데이터 수집, 처리, 그리고 학습 과정에 대해 상세히 설명합니다.

학습 데이터 수집 및 구성

오픈 소스 데이터 (57%)

다양한 사용자 요청을 일반화하여 처리할 수 있는 능력 향상을 목적으로 합니다. 공개적으로 이용 가능한 데이터셋 광범위 조사하고, 전문가 주석 데이터, 실제 사용자 제공 데이터, 합성 데이터를 포함하며, 사람이 직접 리뷰하여 엄격한 선별 과정을 진행합니다.

합성 데이터 (43%)

기존 공개 데이터로 부족한 특정 기술 영역 보완을 목적으로 합니다. 페르소나(Persona) 기반 데이터 합성 방법론을 활용하여 특정 관점에서 데이터를 합성하도록 모델을 유도합니다. 예: "신경망을 연구하는 머신러닝 연구자" 페르소나로 코딩 문제 생성합니다.

오픈 소스 데이터

1

일반 (General)

OpenAssistant Guanaco는 7,132개 프롬프트를 포함하며, SFT와 DPO 모두에 사용됩니다. OpenAssistant 프로젝트의 일환으로 수집된 대화형 데이터로, 다양한 언어와 주제를 포함하여 AI 모델의 대화 능력을 향상시키는 데 사용됩니다. 다양한 언어 지원, 고품질 대화 데이터, 대화 및 지시 수행 능력 향상에 기여합니다.

2

지식 회상 (Knowledge Recall)

No Robots는 9,500개 프롬프트를 포함하며, SFT와 DPO 모두에 사용됩니다. 숙련된 인간 주석자들이 작성한 10,000개의 지시문과 데모로 구성된 고품질 데이터셋입니다. 다양한 작업 범주를 포함하고, 인간 주석자가 작성하여 데이터의 다양성과 품질을 보장하며, 언어 모델의 지시 이해 및 수행 능력 향상을 위한 지도 미세 조정에 사용됩니다.

3

수학 및 추론 (Math & Reasoning)

TÜLU 3 Persona MATH, GSM, Algebra 등 다양한 수학 관련 데이터셋이 포함됩니다. 수학적 추론과 문제 해결 능력 향상을 위해 특별히 설계된 데이터셋들로 구성되어 있습니다.

합성 데이터

Tulu 3 Persona MATH

수학 문제 149,960개를 포함하는 합성 데이터셋입니다.

NuminaMath-TIR

수학적 추론 관련 34,439개의 데이터를 포함합니다.

Tulu 3 Persona GSM

GSM 형태 합성 데이터 49,980개를 포함합니다.

Tulu 3 Persona Algebra

대수학 분야 20,000개의 프롬프트를 포함합니다.

OpenMathInstruct

수학 교육 목적 50,000개 데이터셋입니다.

합성 데이터는 기존 공개 데이터로 부족한 특정 기술 영역을 보완하는 목적으로 생성되었습니다. 페르소나 기반 데이터 합성 방법론을 활용하여 특정 관점에서 데이터를 합성하도록 모델을 유도했습니다. 예를 들어 "신경망을 연구하는 머신러닝 연구자" 페르소나로 코딩 문제를 생성하는 방식입니다.

프롬프트 큐레이션 (Prompt Curation)

프롬프트 큐레이션의 개념

기존 데이터셋에서 품질이 높은 프롬프트를 선별하여 모델이 효율적으로 학습하도록 합니다. 불필요한 데이터(단순하거나 반복적인 데이터)를 제거하여 학습 효율성을 향상시킵니다.

데이터 필터링 및 정제

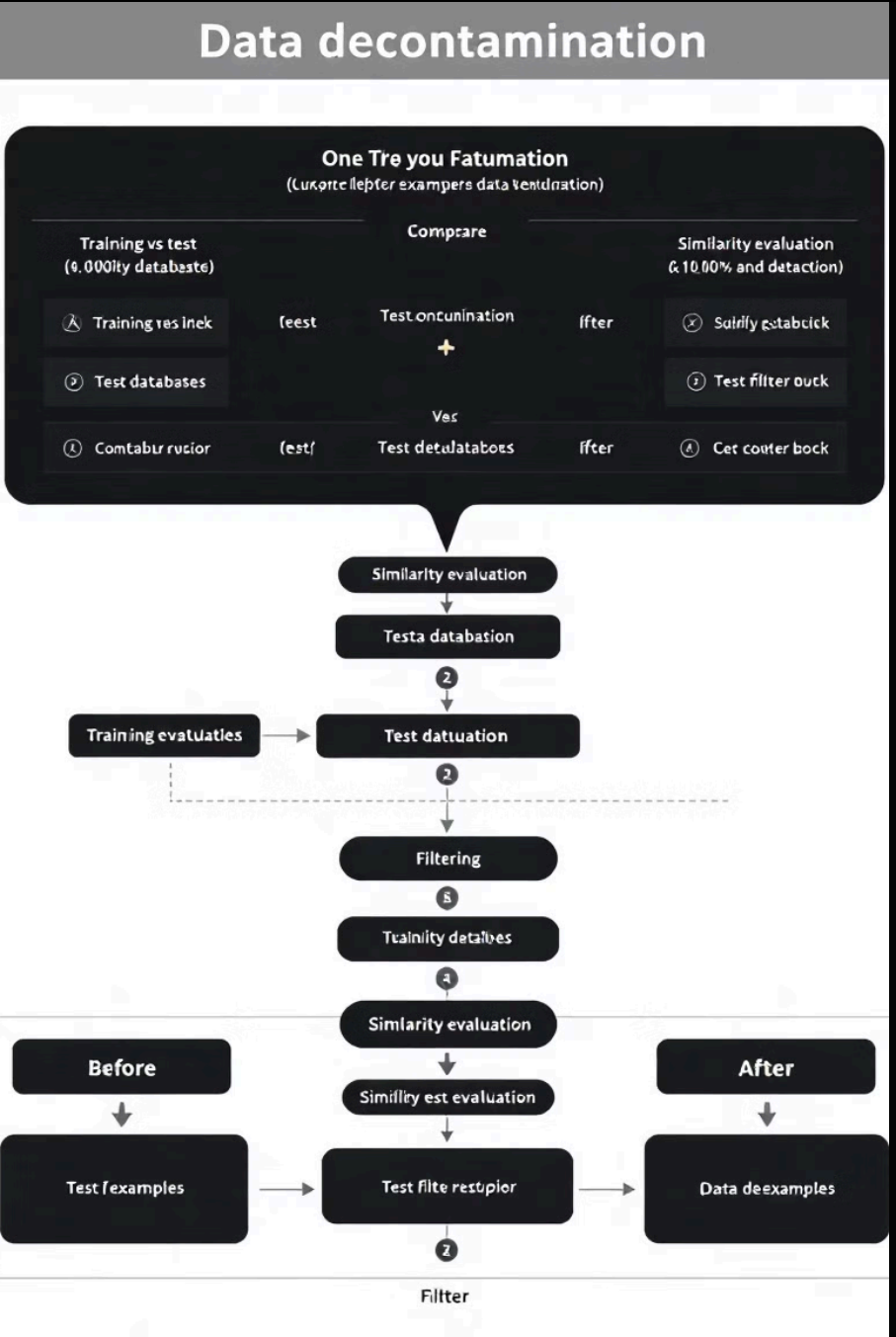
단순하거나 불필요한 프롬프트 (예: "2+2는?") 삭제하고, 중복 데이터 및 비효율적 질문을 제거합니다.

프롬프트 개선 및 보강

단순한 질문을 추론적이고 논리적인 사고가 필요한 형태로 변경합니다. 예시로, 변경 전: "2+2는 얼마인가요?" (단순 계산형 질문)에서 변경 후: "사과를 2개 가지고 있는데, 2개를 더 샀다면 사과는 총 몇 개인가요? 답을 설명해 주세요." (추론과 설명이 요구되는 질문)으로 바꿉니다.

프롬프트 합성 (Prompt Synthesis)

프롬프트 합성은 새로운 프롬프트를 인공적으로 생성하는 과정입니다. 기존 데이터에서 좋은 질문을 고르는 큐레이션과 달리, 완전히 새로운 질문을 만드는 작업입니다. 페르소나 기반 방식은 특정 역할이나 관점(페르소나)을 설정하여, 해당 관점에서의 질문과 응답 데이터를 생성합니다. 예를 들어 고등학교 수학 교사, 파이썬 개발자, 금융 전문가 등의 페르소나를 활용합니다. Self-Play & Bootstrapping은 AI가 스스로 질문을 만들고 답을 생성하여 데이터를 확장하는 방식입니다.



Decontamination (데이터 비오염화)

1

Decontamination의 개념

모델이 평가 데이터(Test Set)를 미리 학습하는 문제를 방지하기 위한 과정입니다. 학습 데이터(Training Set)에서 평가 데이터와 겹치는 부분을 제거합니다.

2

평가 데이터 전처리

평가 데이터(test set)를 문장 단위로 분리하여 분석 가능한 형태로 준비합니다. 데이터 정규화 및 표준화 작업을 수행합니다.

3

학습 데이터와 평가 데이터 간 유사성 검사

N-gram 매칭, Cosine Similarity, Jaccard Index, Dense Embeddings 등의 방법을 통해 유사도를 평가합니다.

4

중복 데이터 필터링 및 수정

학습 데이터가 평가 데이터와 매우 유사(90% 이상)하면 완전히 삭제합니다. 부분적으로 유사(60~90%)하다면, 사람이 직접 확인하고 수정합니다.

Tulu-3 SFT 데이터셋 구조 상세 설명

기본 데이터 구조

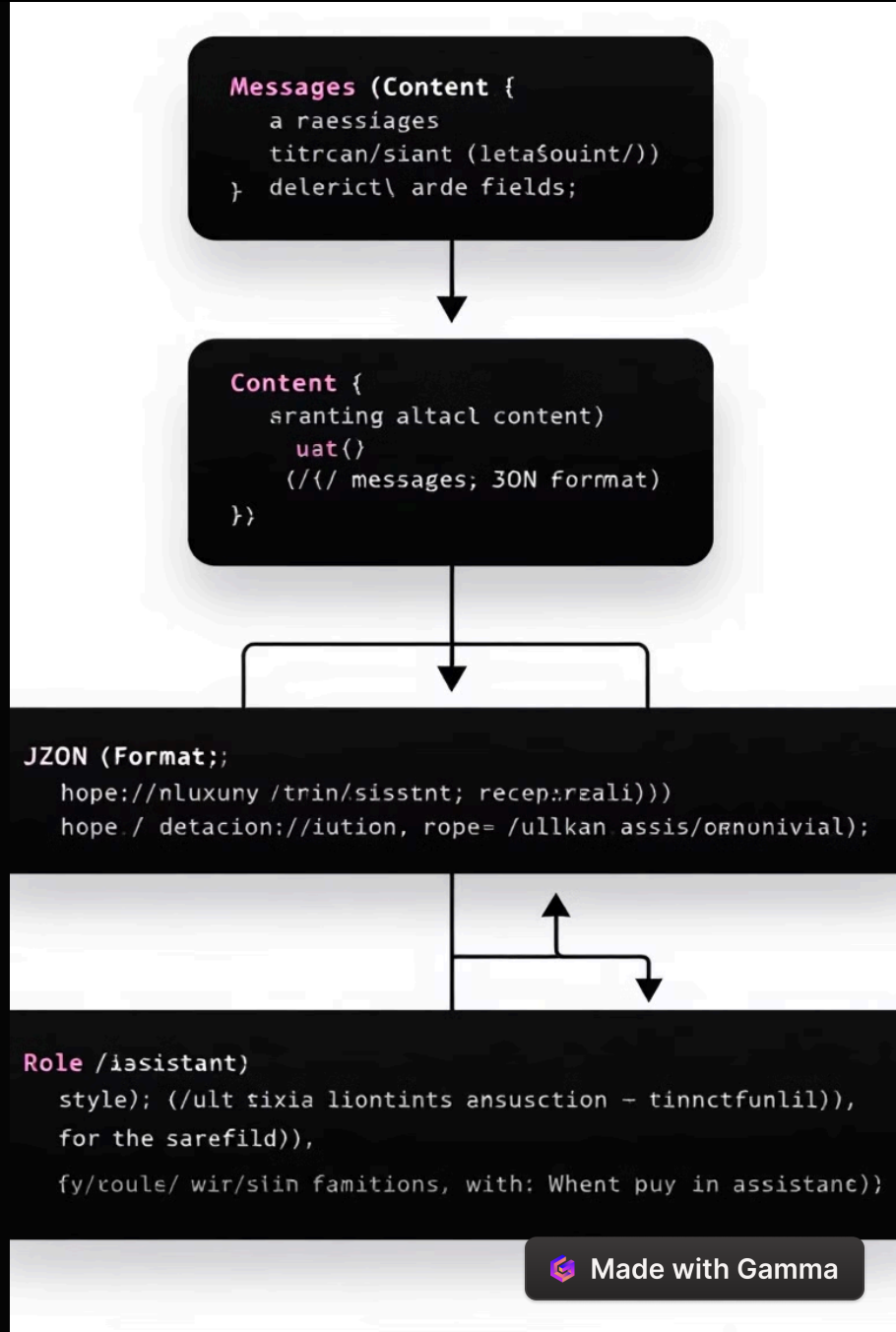
Hugging Face에 공개된 allenai/tulu-3-sft-mixture 데이터셋은 messages (대화 형식의 리스트로, 모델 학습에 직접 사용되는 핵심 데이터), id (각 데이터 샘플의 고유 식별자), source (데이터의 출처 데이터셋)의 구조로 되어 있습니다.

메시지 구조

messages 필드는 대화 형식의 리스트로, content (실제 텍스트 내용)와 role ("user" 또는 "assistant")을 포함합니다. 모델은 "user" 역할의 content를 입력으로 받고, "assistant" 역할의 content를 생성하도록 학습됩니다.

다양한 대화 형태

Tulu-3 데이터셋은 단순한 1:1 질문-응답뿐만 아니라, 여러 턴의 대화도 포함할 수 있습니다. 이런 형태의 데이터를 통해 모델은 대화의 맥락을 이해하고 이전 대화를 기억하며 응답하는 능력을 학습합니다.



데이터셋 관련 프로젝트 구조 분석

1 데이터셋 준비 도구

scripts/data/sft/ 디렉토리에는 다양한 데이터셋을 처리하는 스크립트가 있습니다. prepare_all.sh는 모든 SFT(Supervised Fine-Tuning) 데이터셋을 준비하는 스크립트이며, 각 데이터셋별 처리 스크립트(aya.py, coconot.py, lima.py, sharegpt.py 등)는 데이터셋을 다운로드하고, 필터링하고, 메시지 형식으로 변환하는 작업을 수행합니다.

2 합성 선호도 파이프라인

scripts/synth_pref/ 디렉토리에는 합성 선호도 데이터 생성을 위한 파이프라인이 있습니다. generate_responses.py는 모델 응답 생성을, create_annotation_mix.py는 응답 혼합 생성을, annotate_preferences.py는 선호도 주석 생성을, parse_preferences.py는 선호도 데이터 파싱을 담당합니다.

3 페르소나 기반 데이터 생성

scripts/persona_driven_data_gen/ 디렉토리에는 페르소나 기반 데이터 생성 도구가 있습니다. prompt_templates.py는 다양한 페르소나 기반 프롬프트 템플릿을 제공하고, persona_driven_generate_ifdata.py는 명령어 따르기 데이터를 생성하며, persona_driven_generate_math_code.py는 수학 및 코드 문제/해결책을 생성합니다.

4 Decontamination

decontamination/ 디렉토리에는 훈련-평가 데이터 중복 측정 도구가 있습니다. index.py는 훈련 데이터셋을 Elasticsearch 인덱스로 변환하고, search.py는 테스트 데이터로 인덱스를 검색하여 중복을 측정합니다. 텍스트 매칭과 벡터 기반 매칭을 모두 지원합니다.

결론

939,344

총 프롬프트 수

다양한 출처에서 수집된 총 프롬프트 수

57%

오픈 소스 비율

전체 데이터 중 오픈 소스 데이터의 비율

43%

합성 데이터 비율

전체 데이터 중 합성 데이터의 비율

Tulu3의 학습 데이터 준비 과정은 다양한 출처의 데이터를 수집하고, 철저한 큐레이션과 합성, 그리고 Decontamination을 통해 고품질의 학습 데이터를 구축하는 과정입니다. 이러한 과정을 통해 Tulu3는 다양한 작업에서 높은 성능을 발휘할 수 있게 되었습니다.

데이터 준비 과정에서 특히 주목할 만한 것은 프롬프트 큐레이션과 합성 전략입니다. 프롬프트 큐레이션을 통해 기존 데이터에서 양질의 질문들을 선별했을 뿐만 아니라, 프롬프트 합성을 통해 완전히 새로운 질문들을 추가로 생성하여 학습 데이터셋을 한층 더 풍부하게 만들었습니다.

또한 페르소나 기반 데이터 합성 방법론은 특정 분야의 전문적 지식과 기술을 모델에 효과적으로 학습시키는 데 큰 역할을 했으며, 철저한 Decontamination 과정은 모델의 평가 결과가 실제 성능을 정확히 반영하도록 보장했습니다. 이러한 종합적인 데이터 준비 과정을 통해 Tulu3는 보다 풍부하고 질 높은 학습 데이터를 확보할 수 있었습니다.