한국어 대규모 언어모델의 수학 문제 풀이 능력 향상을 위한 한국적 수학 말뭉치 합성

이동건[1,2,○] 백지수[2] 박원재[2] 유환조[1]
[1]포항공과대학교 [2]KT
{donggeonlee, hwanjoyu}@postech.ac.kr
{jisoo.100, wonjae.park}@kt.com

Synthesizing a Korean-centric Math Corpus to Enhance Math Problem-Solving Ability in Korean Large Language Models

DongGeon Lee[1,2,○] Jisoo Baik[2] Wonjae Park[2] Hwanjo Yu[1]
[1]Pohang University of Science and Technology (POSTECH) [2]KT

요 약

수학 문제 풀이는 대규모 언어모델(Large Language Model; LLM)의 연산 및 추론 능력을 판가름하는 핵심 역량으로 간주된다. 하지만 많은 기성 LLM에 대해 영어 문장으로 구성된 수학 문제보다 한국어 문장으로 구성된 수학 문제에서 문제 풀이 성능 저하가 나타난다. 본 연구는 이 문제의 원인을 한국의 문화적 맥락을 충분히 반영한 고품질 수학 말뭉치의 부족이라고 진단한다. 그 해결책으로 한국적 맥락(Korean-centric)을 반영하는 '합성-증강-검증' 3단계의 사전학습 데이터 구축 파이프라인을 제안한다. 이 파이프라인은 한국의 교육과정과 문화적 맥락을 반영한 교과서 본문과 연습문제를 생성(합성)하는 단계와 기존 문제에 대한 추가 풀이 및 역방향 문제를 생성(증강)하는 단계, 생성된 문제와 풀이의 수학적 정확성과 명확성을 평가(검증)하는 단계로 구성된다. 이 파이프라인을 이용해 최종적으로 약 2,300만 토큰 규모의 LLM 사전학습용 고품질 한국어 수학 말뭉치를 구축했다. 이 데이터로 연속 사전학습을 수행한 모델 성능은 한국어로 번역된 GSM8K 및 MATH 벤치마크에서 베이스라인 모델 성능 대비 평균 215% 향상되었다. 아울러 지도 미세조정 단계에서도 동일한 한국어 수학 데이터로 훈련할 때 미세조정 성능이 평균 39% 증가하였다.

주제어: Korean-centric Language Model, Mathematical Corpus Synthesis, LLM Pre-training, Ethnomathematics

1. 서론

대규모 언어모델(Large Language Model; LLM)의 수학적 추론 능력은 고차원적인 문제 해결을 위한 핵심 역량으로 주목받고 있다. 하지만 대다수 LLM은 영어 수학 문제에서는 높은 성능을 보이지만, 한국어 수학 문제에 대해서는 상대적으로 취약한 모습을 보인다 [1].

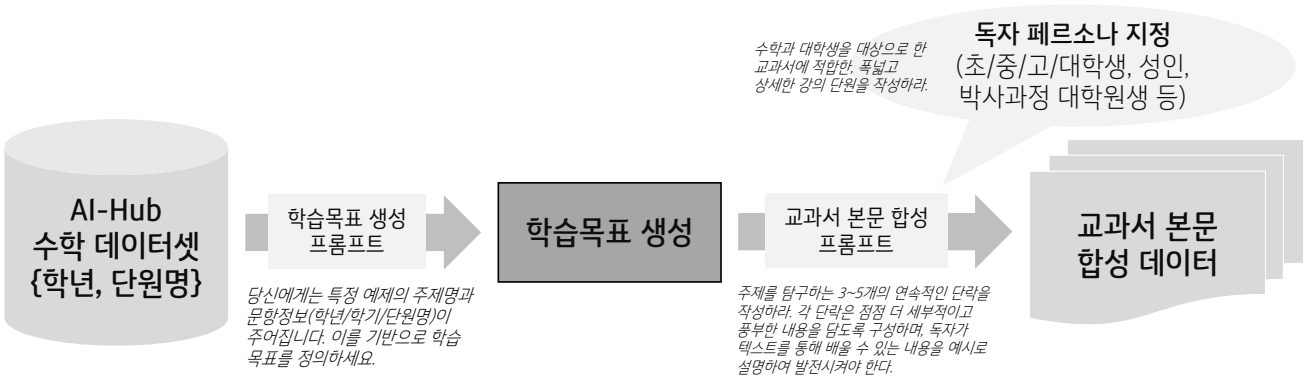
이러한 영어-한국어 간 성능 격차는 고품질 한국어 수학 데이터셋의 부족으로 인해 모델이 문제 해결 능력과 추론 능력을 충분히 학습하지 못한 데에서 비롯된다 [2]. 영어나 중국어와 같은 고자원(High-resource) 언어에서는 LLM의 수학적 능력을 체계적으로 강화하기 위한 수학 말뭉치 연구가 활발했다 [2]. 반면 한국어와 같은 저자원(Low-resource) 언어에 대해서는 상대적으로 이에 대한 연구가 부진했다 [3].

현재 공개된 대부분의 한국어 수학 데이터셋은 기존 영어 수학 말뭉치나 문제를 단순히 번역한 것이다 [4, 5]. 그러나 이러한 단순 번역 데이터는 한국어의 표현 방식이나 교육 문맥을 충분히 반영하지 못하며 실제 한국 학습자의 사고 방식이나 문화적 배경과도 괴리를 보일 수 있다.

이러한 문화적 괴리는 LLM의 학습 효율성과 추론 성능에 직접적인 악영향을 미친다 [6]. 예를 들어, 영미권 수학 문제에 빈번하게 등장하는 "John", "Sarah" 같은 인명, "달러($)", "마일(mile)" 같은 단위, "M&M's", "Reese's" 같은 상품명은 한국어 말뭉치에서 출현 빈도가 현저히 낮아 LLM에는 분포 외(Out-of-Distribution) 데이터로 작용한다 [7]. 더 나아가 LLM이 사전학습 과정에서 습득한 한국어 기반 세계 지식과 번역된 데이터 간의 분포 차이(Distribution Shift)는 모델이 한국어 환경에서 수학적 개념을 자연스럽게 표현하고 추론하는 능력을 저해한다 [8]. 결과적으로, 번역 기반 데이터로만 학습된 LLM은 실제 한국어 수학 평가 환경에서 문제를 올바르게 해석하지 못하거나, 비자연스러운 풀이와 답변을 생성하는 등 성능 저하를 보이게 된다 [9].

이를 극복하기 위해, 본 논문에서는 한국적(Korean-Centric) 수학 말뭉치를 합성하기 위한 '합성-증강-검증' 세 단계로 구성된 데이터 구축 파이프라인을 제안한다. 본 파이프라인은 단순한 번역을 넘어 한국의 수학 교육과정, 단위 체계, 생활상, 문화적 요소 등을 말뭉치 합성 과정에 자연스럽게 반영되도록 설계되었다. 이를 통해 LLM이 사전학습 단계부터 한국어 환경에서의 수학 문제를 더 깊이 이해하고 정교하게 추론할 수 있도록 한다.

본 연구의 효과를 검증하기 위해, 구축된 한국적 수학 말뭉치로 다양한 크기의 LLM을 연속 사전학습(Continual Pre-training; CPT) 및 지도 미세조정(Supervised Fine-tuning; SFT)하여 한국어로 번역된 GSM8K 및 MATH [1, 10, 11] 수학 벤치마크에서 성능이 크게 향상됨을 실험적으로 입증한다.



합성 교과서 본문 데이터 예시: <중학교 1학년 - 소수와 합성수 >

∘ **독자: 중학교 1학년, 학습목표: 주어진 자연수가 소수인지 합성수인지 판별할 수 있다.**
(전략) 소수와 합성수에 대해 배워봅시다. 숫자 2, 3, 5와 같은 숫자들은 1과 자신 외에는 나누어 떨어지지 않으며, 이를 소수라고 합니다. 반면, 숫자 4, 6, 8과 같은 숫자들은 1과 자신 이외에도 다른 수로 나누어 떨어지며, 이를 합성수라고 합니다. (중략) 주어진 수가 소수인지 합성수인지 확인하려면 어떻게 해야 할까요? 1부터 주어진 수까지의 모든 수를 하나씩 나누어 볼 수 있습니다. 나누어 떨어지는 수가 1과 그 수 자신 이외에도 있다면 합성수, 없다면 소수입니다. (후략)

∘ **독자: 암호학 전공 박사과정 대학원생, 학습목표: 주어진 자연수가 소수인지 합성수인지 판별할 수 있다.**
(전략) 소수를 판별하는 과정에서 페르마의 소정리는 중요한 역할을 한다. 이 정리는 소수일 가능성이 있는 수를 먼저 선별하여 연산을 줄이는 데 유용하다. 그러나 페르마의 소정리는 합성수에 대해서도 참일 수 있는 '페르마의 거짓말쟁이'를 만들어 내므로, 이에 대한 해결책으로 밀러-라빈 소수성 테스트 같은 확률적 소수 판정법이 개발되었다. 이 알고리즘은 수가 소수일 가능성을 높이지만, 확률적으로 오차가 있기 때문에 실제 실용적인 소수 생성에 주로 활용된다. (후략)

그림 1. 동일 학습 목표로 독자 페르소나 및 학습과정에 맞는 한국적 교과서 본문 데이터를 합성하는 예시

2. 관련 연구

2.1 다국어 환경에서의 수학 추론 격차

LLM의 성능이 특정 언어(주로 영어 및 중국어)에 편중되는 현상은 여러 연구에서 지적되었다. 특히, [1]은 LLM이 한국어와 같은 저자원 언어 환경에서 수학 문제를 해결할 때 상당한 성능 저하를 겪는 다국어 수학 추론 격차 문제를 제기하였다. 이러한 연구들은 모델의 추론 능력 자체보다는 비영어권에 대한 맥락에 대한 학습이 부족한 것을 원인이라고 지적한다.

2.2 교과서 형식의 데이터 합성을 통한 추론 능력 강화

최근 LLM의 성능을 높이기 위해 양질의 데이터를 합성하려는 연구가 활발히 진행되고 있다. 그중 [12, 13]은 양질의 설명과 예제로 구성된 교과서 형식의 데이터를 학습한 LLM이 최종 학습 토큰 수가 더 적음에도 보다 나은 추론 능력을 나타냄을 보였다. 이는 무작위의 데이터보다 양질의 데이터를 학습하는 것이 모델 성능에 더 큰 영향을 미칠 수 있음을 시사한다. 또한, [14]는 합성 데이터의 다양성이 중요하다고 강조했다. 난이도, 복잡성, 서술 스타일 등 다양한 변주를 가진 데이터를 학습할 때 모델의 일반화 성능이 향상된다는 것이다. 본 연구는 이러한 선행 연구들의 아이디어를 적극적으로 채택하고 확장한다.

2.3 수학 문제 데이터 증강 기법

제한된 자원의 데이터셋을 효과적으로 확장하기 위한 데이터 증강 기법 또한 중요한 연구 주제이다. 특히 수학 문제 해결 분야에서는 문제와 풀이의 논리적 관계를 활용한 증강 방식이 주목받고 있다 [15]. [16]은 기존 문제에 대해 다양한 접근 방식의 풀이를 추가하거나, 문제의 일부를 미지수로 바꾸어 역으로 질문하는 방식을 통해 데이터를 증강한다. 또한 [17]은 전방-후방 추론(Forward-Backward Reasoning) 기법을 통해 주어진 답을 바탕으로 문제의 초기 조건을 추론하게 하는 역방향 문제를 생성하여 모델의 논리적 이해도를 높이는 방법을 제시했다.

표 1. 교육과정 분류 및 소단원에 따른 학습목표 합성 결과 예시

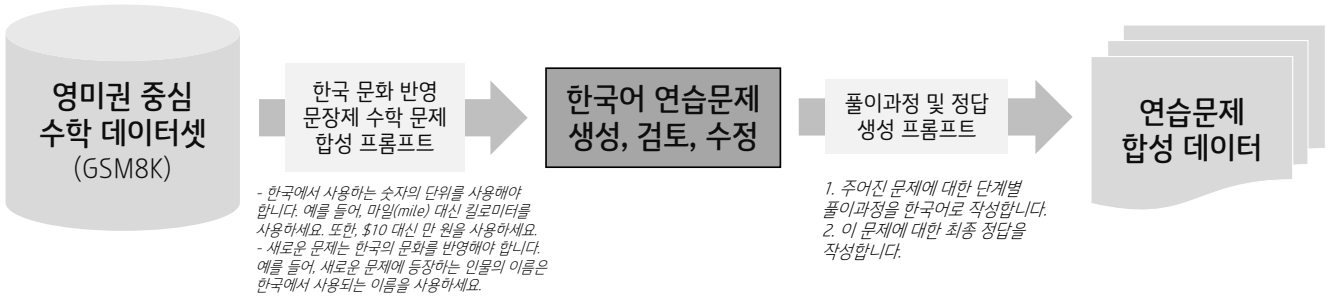
교육과정 분류	소단원	학습목표
중학교 1학년 1학기 1단원 자연수의 성질	소수와 합성수	"소수와 합성수의 정의를 이해하고 이를 구분할 수 있다.", "주어진 자연수가 소수인지 합성수인지 판별할 수 있다.", "소수를 찾기 위한 방법(예: 에라토스테네스의 체)을 이해하고 활용할 수 있다.", "소수와 합성수의 차이를 설명할 수 있다.", "자연수를 소인수분해하여 소수의 곱으로 표현할 수 있다."

3. 한국어 수학 말뭉치 합성-증강-검증 파이프라인

본 연구에서는 '합성-증강-검증' 3단계로 구성된 한국적 수학 말뭉치 구축 파이프라인을 제안한다. 교과서 형식 데이터의 학습 효율을 입증한 이전 연구 [12, 13]을 참고하여 실제 국내 수학 교과서와 같은 산출물이 구축되도록 했다.

3.1 교과서 데이터 합성

합성 단계에서는 실제 수학 교과서와 유사한 형태로 국내 교육과정에 맞는 수학적 개념과 예시를 한국어로 설명하는 본문과 교육과정 상에 등장하는 각 개념과 한국의 문화적 요소를 동시에 학습하도록 하는 연습문제를 합성한다.



∘ 기존 영미권 중심 수학 데이터셋 예시
It's Ava's birthday party. Her parents bought a unicorn piñata for $13 and filled it with all of her favorite treats. They bought 4 bags of Reese's for $9 per bag, 3 bags of Snickers for $5 per bag, and 5 bags of Skittles for $7 per bag. How much did the unicorn piñata and the treats cost altogether?

∘ 합성한 한국어 연습문제 데이터 예시
미진이의 생일 파티를 위해 부모님이 바구니를 1만 3천원에 사고, 그녀가 좋아하는 간식을 채워 넣었습니다. 그들은 '몽쉘'을 4갑 사는데 한 갑당 9천 원, '자유시간'을 3봉지 사는데 한 봉지당 5천 원, '새콤달콤'을 5박스 사는데 한 박스당 7천 원의 가격으로 샀습니다. 바구니와 간식 비용을 다 합하면 얼마일까요?

그림 2. 영미권 중심의 수학 문제를 바탕으로 한국의 문화적 맥락에 맞는 연습문제를 합성하는 과정 및 예시

3.1.1 한국적 교과서 본문 데이터 합성

교과서 본문 데이터 합성 과정은 단순히 개념을 나열하는 것을 넘어 내용의 깊이와 서술 스타일의 다양성을 확보하는 데 중점을 둔다. 그림 1은 수학적 개념을 한국 교육과정에 맞게 서술하는 교과서 본문 글을 합성하기 위한 과정을 나타낸다.

교육과정별 학습 목표 합성 교과서 본문을 합성하기 전 실제 국내 수학 교육과정에서 다루는 단원과 개념 정보를 추출한다. 추출된 개념 정보는 교과서 본문을 생성하기 위한 토대 역할을 한다. 구체적으로 AI-Hub에 공개된 수학 데이터셋 [18, 19]를 바탕으로 교육과정(학년, 학기, 단원명, 소단원) 정보를 추출한다. 다음으로 LLM을 이용하여 학습 목표를 소단원별로 합성한다. 예를 들어 중학교 1학년 1학기 1단원의 소단원 '소수와 합성수'가 추출되면 이 개념에 대해 알아야 할 세부 학습 목표를 표 1과 같이 합성한다.

페르소나 기반 본문 합성 다음으로 LLM을 이용하여 소단원 및 학습목표에 맞는 본문 글을 생성한다. 이때 실제 교육 수요자를 상정한 독자 페르소나를 지정한다. 각 페르소나의 눈높이에 맞춰 본문을 생성하라는 지시를 통해 개념 설명의 깊이와 복잡성을 조절한다. 예를 들어 동일한 '소수와 합성수' 소단원에 대해서 독자 페르소나에 따라 다른 데이터를 얻을 수 있다. 중학생 페르소나의 경우, 해당 개념에 대한 기본적인 정의와 판별법을 서술한 본문 데이터를 얻는다. 반면 암호학 전공 대학원생 페르소나의 경우, '페르마 소정리'나 '밀러-라빈 소수판정법' 같은 심화 개념을 서술한 데이터를 얻는다.

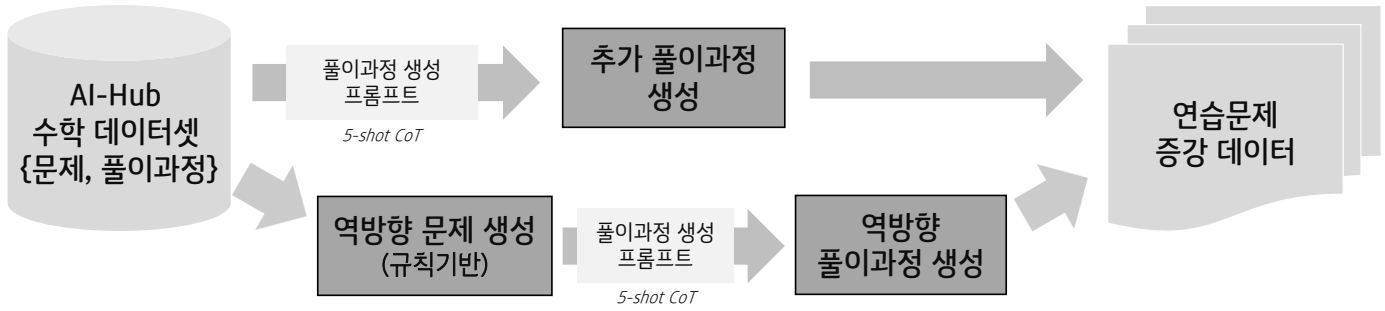
3.1.2 한국적 연습문제 합성

수학적 개념을 응용한 문제 풀이 능력 학습을 위해 각 개념과 연관된 연습문제를 합성한다. 연습문제 합성은 페르소나 기반 본문 합성 시 진행한다. 이에 따라 각 단원과 개념에 대해 각 페르소나 수준에 맞춘 연습문제를 구축한다. 더불어 한국의 사회문화적 맥락을 반영한 연습문제를 생성하기 위해 GSM8K [10] 같은 기존 영미권 중심의 수학 데이터셋을 한국의 실정에 맞게 변환하는 과정을 포함한다. 그림 2는 한국적 연습문제 합성 과정과 영미권 원본 수학 문제가 한국 중심의 문제로 변환된 예시이다. 해당 과정은 프롬프트를 통해 원본 영어 문제에 사용되는 단위를 한국식(mile → 킬로미터, $ → 원)으로 변환하고, 등장인물의 이름을 한국 이름으로 변경하며, "Reese's" 같은 외국 상품명을 "몽쉘", "자유시간" 등 한국인에게 친숙한 상품명으로 대체한다.

3.2 교과서 데이터 증강

증강 단계에서는 학습용 데이터의 양과 다양성을 확보하기 위해, 공개된 수학 데이터셋의 문제 유형과 풀이과정을 증강하는 것을 목표로 한다. 본 논문에서는 공개된 수학 데이터셋 [18, 19]에 두 가지 전략을 활용한다. 그림 3은 증강하는 과정을 그림으로 나타낸다.

추가 풀이과정 생성 먼저 수학 데이터셋의 각 문제에 대해 단계적인 사고 과정이 반영된 풀이과정을 생성한다. 이는 LLM이 하나의 문제에 대해 다각적인 문제 접근 방식과 보다 구조화된 풀이를 학습하도록 유도하려는 목적이다 [16]. 아울러 이 단계는 생성한 풀이의 정확성이 보장되도록 원본 문제의 정답과 생성된 최종 정답이 일치하는지 확인하는 과정을 포함한다.



∘ 기존 연습문제 예시
민수는 천원 권 3장과 만원 권 5장을 갖고 있습니다. 민수는 얼마를 갖고 있나요? *정답: 53000*

∘ 역방향 문제 생성 예시
민수는 천원 권 X장과 만원 권 5장을 갖고 있습니다. 민수가 5만 3천원을 갖고 있다면, X는 몇 인가요? *정답: 3*

그림 3. 연습문제 데이터 증강 과정

역방향 문제 생성 다음으로, 모델의 논리적 유연성을 강화하기 위해 역방향(Forward-Backward) 문제를 합성한다. 역방향 문제란 기존 문제의 숫자 일부를 미지수 X로 바꾸고, 기존 문제의 정답을 문제의 조건으로 제시하는 것을 말한다 [16, 17]. 예를 들어 "천원권 3장과 만원권 5장을 합하면 얼마인가?"라는 문제를 "천원권 X장과 만원권 5장을 합해 5만3천원이라면 X는 몇인가?"와 같은 형태의 문제로 변환한다.

3.3 합성 데이터 검증

검증 과정에서는 추론 모델을 이용하여 풀이 검증을 통해 앞서 확보한 합성 데이터의 품질을 보장한다. 모델에 합성된 본문과 연습문제에 대해 수학적 정확성과 명확성을 기준으로 통과 여부를 참/거짓으로 판단하도록 지시한다. 단, 증강 과정에서 생성된 연습 문제는 기존 정답 정보를 활용한 검증이 가능하여 휴리스틱 검증으로 대체하였다.

4. 실험

4.1 말뭉치 합성-증강-검증

표 2. 검증 전후별 구축 말뭉치의 샘플 수 및 토큰 수 비교

	검증 전		검증 후	
	샘플 수	토큰 수	샘플 수	토큰 수
① 교과서 본문 합성	18,739	8,688,815	17,050 (▽1,689)	7,891,094 (▽797,721)
② 교과서 연습문제 합성	7,473	1,936,305	7,047 (▽426)	1,806,016 (▽130,289)
③ 연습문제 증강	48,337	13,324,999	48,337 (–)	13,324,999 (–)
전체 구축 말뭉치	74,549	23,950,119	72,434 (▽2,115)	23,022,109 (▽928,010)

표 2는 제안한 파이프라인을 활용하여 GPT-4o로 한국어 수학 말뭉치를 합성 · 증강한 뒤, DeepSeek-R1-Distill-Qwen-32B로 검증한 결과를 보여준다. 합성 과정을 통해 약 26,000 건의 데이터(약 1,000만 토큰)와 증강 과정을 통해 약 48,000 건의 데이터(약 1,300만 토큰)를 확보하였다.

그러나 합성 데이터에 대한 검증 결과 데이터 2,115 건이 부적합 판정을 받았다. 최종적으로 사전학습 실험에는 AI-Hub 수학 데이터 원본을 포함하여 134,186 건의 말뭉치(약 3,500만 토큰)를 말뭉치 데이터로 사용한다.

4.2 실험 설계

본 논문에서는 합성한 말뭉치의 효과를 사전학습과 수학 문제 풀이 과제에 대한 SFT 관점에서 검증한다. 각 실험에 대한 평가는 한국어로 번역된 GSM8K 및 MATH 벤치마크 [1, 10, 11]에 대한 Exact Match 점수로 측정한다.

4.2.1 CPT의 영향

CPT는 이미 사전학습된 모델에 추가로 사전학습을 수행하는 것을 말한다. 먼저 구축한 말뭉치가 사전학습 단계에서 수학 문제 풀이 능력 향상에 얼마나 효과가 있는지 검증한다. 이를 위해 이미 대규모 말뭉치로 사전학습된 모델에 합성 말뭉치를 이용하여 CPT를 수행한 모델의 성능을 확인한다. 특히 사전학습 중 말뭉치의 영향도를 분석하기 위해 파이프라인 중 합성 증강 단계까지만 거친 데이터와 검증 단계까지 거친 데이터로 각각 CPT를 진행한다. 두 경우의 성능을 비교하여 본 연구가 제시하는 전체 파이프라인의 효용성을 확인한다. 베이스라인 모델로는 Llama-3.2 1B, 3B 및 Llama-3.1 8B를 사용한다.

표 3. 합성 말뭉치 기반 CPT 및 CPT-SFT 모델의 한국어 수학 벤치마크 Exact Match 성능 비교 (단위: %)

	한국어 GSM8K	한국어 MATH	평균	증감률
Continual Pre-trained (CPT) Model				
Llama-3.2-1B-Base	1.29	2.70	1.99	
Llama-3.2-1B-CPT	**3.03**	**4.40**	**3.71**	+86.21%
Llama-3.2-3B-Base	3.94	6.72	5.33	
Llama-3.2-3B-CPT	**14.48**	**9.71**	**12.09**	+126.74%
Llama-3.1-8B-Base	1.90	4.02	2.96	
Llama-3.1-8B-CPT	**12.28**	**19.27**	**15.78**	+432.94%
Supervised Fine-tuned (SFT) Model				
Llama-3.2-1B-Base-SFT	6.22	7.07	6.64	
Llama-3.2-1B-CPT-SFT	**10.31**	**8.18**	**9.25**	+39.16%
Llama-3.2-3B-Base-SFT	19.71	12.10	15.90	
Llama-3.2-3B-CPT-SFT	**29.64**	**14.80**	**22.22**	+39.72%

4.2.2 SFT의 영향

베이스라인 모델과 CPT 모델을 각각 동일한 데이터로 SFT 한 후 성능을 비교함으로써 SFT 시에 합성 말뭉치의 영향을 확인한다. SFT 시에는 NuminaMath-CoT [20]을 한국어로 단순 번역한 공개 데이터 [4]를 40,000 건 샘플링하여 사용한다.

4.3 실험 결과

4.3.1 CPT 및 SFT의 효과

표 3은 베이스라인 모델(Base), CPT 모델, Base 모델을 SFT 한 모델(Base-SFT), 그리고 CPT 후 SFT를 거친 모델(CPT-SFT)의 한국어 수학 벤치마크 성능 평가 결과를 나타낸다.

CPT 모델은 베이스라인 모델보다 평균 215.30% 향상된 성능을 보였다. 특히 모델의 파라미터 크기가 클수록 합성 말뭉치를 통한 성능 개선 효과가 더 컸다. 이는 모델 규모가 클수록 합성 말뭉치에 담긴 복잡한 개념과 문화적 맥락을 더 효과적으로 학습하고 일반화하는 능력이 뛰어남을 시사한다.

또한 CPT-SFT 모델의 성능은 Base-SFT 모델의 성능 대비 평균 39.44% 향상되어 합성 말뭉치가 SFT 단계에서의 학습 능력을 비롯한 모델의 잠재력을 효과적으로 끌어올림을 확인했다. 이 결과는 CPT에서 문화적 배경을 고려한 수학적 사고를 학습시키는 방법이 단순히 특정 유형의 문제 풀이 능력만을 주입하는 것을 넘어 SFT 과정에서 학습 효율을 높이는 견고한 기초를 마련해 주었음을 의미한다. 즉, 한국적 맥락에 맞춰 사전 학습을 거친 모델이 후속 미세조정 단계에서도 새로운 지식을 더 잘 흡수하고 응용할 수 있게 된 것이다.

4.3.2 말뭉치 검증의 효과

검증을 거치며 2,115개 샘플(전체의 2.84%)이 제거되었지만 (표 2), 검증된 말뭉치로만 학습한 모델은 검증되지 않은 데이터를 포함한 말뭉치로 학습된 모델 대비 평균 +36.67%의 성능 향상을 보였다(표 4). 이러한 결과는 수학과 같이 정밀한 논리적 추론이 요구되는 분야에서는 데이터의 양보다 수학적 오류가 없고 명확하게 서술된 고품질 데이터가 모델 성능에 결정적인 영향을 미친다는 것을 보여준다. 즉, 검증을 거치지 않은 데이터에 포함될 수 있는 미세한 논리적 오류나 부정확한 풀이 과정이 모델의 학습을 방해하고 오히려 성능 저하의 원인이 될 수 있음을 시사한다. 이러한 결과로써 본 논문에서 제안한 파이프라인에서 합성, 증강뿐 아니라 검증 과정까지 모두 중요하게 작동하였음을 입증했다.

표 4. CPT 과정에서 말뭉치 검증 여부에 따른 한국어 수학 벤치마크 Exact Match 성능 변화 (단위: %)

	한국어 GSM8K	한국어 MATH	평균	증감률
Llama-3.2-1B-Base	1.29	2.70	1.99	
Llama-3.2-1B-CPT (미검증)	1.59	3.85	2.72	
Llama-3.2-1B-CPT (검증)	**3.03**	**4.40**	**3.71**	+36.67%

5. 결론 및 논의

본 연구는 LLM의 한국어 수학 추론 격차 문제를 해결하기 위해 '합성-증강-검증' 3단계 파이프라인을 제안하고, 이를 통해 한국의 교육적·문화적 맥락을 반영한 고품질 수학 말뭉치를 구축하였다. 실험 결과, 본 데이터로 CPT를 수행한 모델은 한국어 GSM8K 및 MATH 벤치마크에서 베이스라인 대비 월등한 성능 향상을 보였으며, 데이터 검증 과정 또한 필수적임을 확인하였다. 결론적으로 본 연구는 단순 번역을 넘어 목표 언어의 고유한 맥락을 반영한 고품질 데이터 합성이 LLM의 전문 분야 추론 능력을 강화하는 데 매우 효과적임을 입증하였다. 제안한 파이프라인은 향후 다른 저자원 언어나 수학 외 전문 분야에서도 고품질 데이터를 확보하고 성능 격차를 해소하기 위한 효과적인 방법론으로 확장될 것으로 기대된다.

감사의 글

이 논문은 2025년 정부(과학기술정보통신부) 및 지자체(대구광역시)의 재원으로 (재)대구디지털혁신진흥원에서 주관하는 지역 디지털 혁신거점 조성지원 사업의 지원을 받아 수행된 연구입니다.(No.25DIH-11, 모델 컨텍스트 프로토콜(MCP) 기반 언어모델(LLM) 멀티 에이전트 협업 시스템 개발).

이 연구는 과학기술정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 구축된 "수학 과목 자동 풀이 데이터" 및 "수학 과목 문제 생성 데이터"를 활용하여 수행된 연구입니다. 본 연구에 활용된 데이터는 AI 허브에서 다운로드 받으실 수 있습니다.

참고문헌

[1] H. Ko, G. Son, and D. Choi, "Understand, solve and translate: Bridging the multilingual mathematical reasoning gap," *arXiv preprint arXiv:2501.02448*, 2025.

[2] Z. Azerbayev, H. Schoelkopf, K. Paster, M. D. Santos, S. M. McAleer, A. Q. Jiang, J. Deng, S. Biderman, and S. Welleck, "Llemma: An open language model for mathematics," *The Twelfth International Conference on Learning Representations*, 2024.

[3] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, 2020.

[4] G. Chu, "AI-MO-NuminaMath-CoT-Ko," https://huggingface.co/datasets/ChuGyouk/AI-MO-NuminaMath-CoT-Ko, 2024.

[5] Allganize Inc. LLM TEAM and S. Ryu, "AIME2025-ko," https://huggingface.co/datasets/allganize/AIME2025-ko, 2025.

[6] A. Karim, A. Karim, B. Lohana, M. Keon, J. Singh, and A. Sattar, "Lost in cultural translation: Do LLMs struggle with math across cultural contexts?" *arXiv preprint arXiv:2503.18018*, 2025.

[7] S. Tu, K. Zhu, Y. Bai, Z. Yao, L. Hou, and J. Li, "Dice: Detecting in-distribution contamination in LLM's fine-tuning phase for math reasoning," *arXiv preprint arXiv:2406.04197*, 2024.

[8] S. Jiang, Y. Liao, Y. Zhang, Y. Wang, and Y. Wang, "Taia: Large language models are out-of-distribution data learners," *Advances in Neural Information Processing Systems*, Vol. 37, pp. 105 200–105 235, 2024.

[9] A. Tomar, N. R. Sahoo, A. Mittal, R. Murthy, and P. Bhattacharyya, "Mathematics isn't culture-free: Probing cultural gaps via entity and scenario perturbations," *arXiv preprint arXiv:2507.00883*, 2025.

[10] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168v2*, 2021.

[11] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the MATH dataset," *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*, 2021.

[12] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. D. Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, H. S. Behl, X. Wang, S. Bubeck, R. Eldan, A. T. Kalai, Y. T. Lee, and Y. Li, "Textbooks are all you need," *arXiv preprint arXiv:2306.11644v2*, 2023.

[13] Y. Li, S. Bubeck, R. Eldan, A. D. Giorno, S. Gunasekar, and Y. T. Lee, "Textbooks are all you need II: Phi-1.5 technical report," *arXiv preprint arXiv:2309.05463*, 2023.

[14] H. Chen, A. Waheed, X. Li, Y. Wang, J. Wang, B. Raj, and M. I. Abdin, "On the diversity of synthetic data and its impact on training large language models," *arXiv preprint arXiv:2410.15226v2*, 2024.

[15] J. Jung and S. Jung, "Generating mathematical curriculum types and problems using large language models," *Proceedings of the 36th Annual Conference on Human & Cognitive Language Technology*, pp. 298–303, 2024.

[16] L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. T. Kwok, Z. Li, A. Weller, and W. Liu, "MetaMath: Bootstrap your own mathematical questions for large language models," *The Twelfth International Conference on Learning Representations*, 2024.

[17] W. Jiang, H. Shi, L. Yu, Z. Liu, Y. Zhang, Z. Li, and J. T. Kwok, "Forward-backward reasoning in large language models for mathematical verification," *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6647–6661, 2024.

[18] AI 허브, "수학 과목 자동 풀이 데이터," https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71716, 2023.

[19] AI 허브, "수학 과목 문제 생성 데이터," https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71718, 2023.

[20] J. LI, E. Beeching, L. Tunstall, B. Lipkin, R. Soletskyi, S. C. Huang, K. Rasul, L. Yu, A. Jiang, Z. Shen, Z. Qin, B. Dong, L. Zhou, Y. Fleureau, G. Lample, and S. Polu, "NuminaMath," https://huggingface.co/AI-MO/NuminaMath-CoT, 2024.