JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 2, pp. 31-42, February 2026
https://doi.org/10.9708/jksci.2026.31.02.031

An LLM-Based Automatic Selection of High-Difficulty Questions Using the Swiss Tournament Format

Jooeun Lee*, Minseob Song*, Namgyu Kim**

*Graduate Student, Graduate School of Business IT, Kookmin University, Seoul, Korea
**Professor, Graduate School of Business IT, Kookmin University, Seoul, Korea

[Abstract]

With the advancement of large language models(LLMs), in-context learning has become a key approach, driving research on prompting techniques. Among them, few-shot Chain-of-Thought(CoT) prompting, which induces explicit reasoning, shows strong performance but depends on example composition. While prior work focused on diversity or uncertainty in example selection, difficulty-based approaches remain underexplored. This study proposes a method to identify high-difficulty questions by combining pairwise difficulty comparisons conducted by an LLM with a Swiss tournament structure, constructing few-shot CoT exemplars with human reasoning annotations. Experiments on 1,319 GSM8K problems show that the proposed method outperforms random, uncertainty-based, and direct difficulty evaluation approaches by 2.12%p, 1.36%p, and 10.16%p, respectively.

▸ **Key words**: Large Language Model, CoT, Difficulty, Swiss Tournament, Inference

[요 약]

거대 언어 모델의 발전에 따라 문맥 내 학습은 언어 모델의 대표적인 활용법으로 주목받으며, 이에 다양한 프롬프트 기법이 연구되고 있다. 특히, 사고 과정의 명시를 유도한 few-shot CoT(Chain-of-Thought)는 소수의 예시 제공만으로 추론 성능을 극대화한 방법으로 알려져 있으나, 예시 구성에 따라 성능 편차가 발생한다는 한계가 존재한다. 기존 연구는 다양성이나 불확실성 등 일부 기준을 중심으로 예시를 구성할 질문을 선정해 왔으나, 난이도 기반 질문 선정에 관한 연구는 상대적으로 미진한 실정이다. 이에 본 연구는 언어 모델을 활용한 쌍대 난이도 비교와 스위스 토너먼트 구조를 결합하여, 고난도 질문을 체계적으로 선별하고 이를 기반으로 few-shot CoT 예시를 구축하는 새로운 방법론을 제안한다. 제안 방법론의 성능 평가를 위해 수학 서술형 벤치마크인 GSM8K 데이터셋 1,319개 문항을 대상으로 실험을 수행한 결과, 제안 방법론이 무작위 선정, 불확실성 기반, 그리고 난이도 직접 평가 방식 대비 정확도 측면에서 각각 2.12%p, 1.36%p, 그리고 10.16%p의 성능 향상을 보임을 확인하였다.

▸ **주제어**: 거대 언어 모델, CoT, 난이도 기반 예시 선정, 스위스 토너먼트, 추론

• First Author: Jooeun Lee, Corresponding Author: Namgyu Kim
*Jooeun Lee (wndms2047@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
*Minseob Song (magnet9805@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
**Namgyu Kim (ngkim@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
• Received: 2025. 12. 12, Revised: 2026. 01. 13, Accepted: 2026. 01. 28.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

딥러닝(Deep Learning)의 발전에 힘입어 등장한 거대 언어 모델(Large Language Model)은 자연어 처리 분야의 성장을 이끌며 핵심 연구 주제로 자리 잡았다. 거대 언어 모델은 수십억에서 수천억 개의 매개변수로 구성된 신경망 구조로, 방대한 말뭉치에 대한 사전학습(Pre-training)을 통해 언어 전반에 대한 지식을 내재한다. 대표적으로 GPT[1]와 LLaMA[2]는 인간에 근접한 수준의 텍스트 이해 및 생성을 구현하며, 기계 번역, 문서 요약, 질의응답, 그리고 코드 생성 등 다양한 작업에 활용되고 있다.

이러한 거대 언어 모델의 범용성에도 불구하고, 이를 특정 영역이나 작업에서 원하는 목적에 맞게 활용하기 위해서는 일반적으로 추가적인 조정이 필요하다. 대표적인 접근법으로 알려진 문맥 내 학습(In-Context Learning)[1, 3]은 입력 프롬프트(Prompt)에 몇 가지 지침과 예시를 제공함으로써, 별도의 가중치 갱신 과정 없이도 주어진 맥락을 토대로 모델이 새로운 작업을 해결하도록 한다. 하지만 본 방법은 단순 질의응답 작업에서는 성공적으로 작동하나, 다단계 수리 추론이나 복합 논리 추론과 같이 고차원적 인지를 요구하는 작업에서는 제한적인 성능을 보이는 경향이 있다[4].

이러한 한계를 극복하고 거대 언어 모델의 추론 성능을 향상하기 위한 다양한 연구가 이루어졌으며, 주목할 만한 방법론으로 Chain-of-Thought(CoT)[5]를 들 수 있다. CoT는 문제를 해결하는 과정에서 모델이 중간 추론 단계를 기술하도록 유도하는 방법이다. 이는 단순히 입력에 대한 최종 답을 제시하는 데 그치지 않고 답에 도달하기까지의 사고 과정(Rationale)을 단계적으로 나누어 서술함으로써, 복잡한 추론 작업에서 성능 향상을 이끌었다. CoT는 프롬프트에 제공되는 예시의 수에 따라 zero-shot[6], one-shot, 그리고 few-shot CoT[5]로 구분된다. 그 가운데 few-shot CoT는 추론 과정이 포함된 몇 개의 예시를 제공함으로써 모델이 이를 따라 단계별 사고를 전개하도록 한 기법으로, 가장 우수한 성능을 보인다. 그러나 프롬프트에 의존하는 방식의 특성상, few-shot CoT는 주어지는 예시에 따라 성능 편차가 나타난다는 근본적인 한계를 갖는다[3, 7, 8].

이와 같은 의존성은 적절한 예시 선정이 CoT의 성능에 결정적인 영향을 미친다는 점을 시사한다. 기존 연구에서는 연구자가 수동으로 적절한 예시를 선정하거나 단순히 무작위로 선별하는 방식을 채택하였으나, 이러한 방식은 다양한 작업에서 최적의 성능을 보장하기 어렵다는 점이 지적되었다[4]. 이에 연구자들은 어떤 질문을 주석(Annotation) 대상으로 선정할 것인가에 초점을 맞추었고, 이러한 배경 속에서 불확실성 기반의 예시 선정 방법론인 Active Prompt[4]가 고안되었다. Active Prompt는 모델의 일관적이지 않은 추론 결과를 나타내는 지표인 불확실성에 기반해 소수의 질문을 선정했으며, 이를 few-shot CoT 예시로 활용하여 추론 성능을 끌어올렸다. 이 연구는 예시 선정 방식이 CoT의 성능에 미치는 영향을 실증적으로 입증했다는 점에서 가치를 인정받았다.

본 연구는 동일한 문제의식에서 출발하여, Active Prompt의 추론 성능을 더욱 향상시키는 것을 목표로 한다. 이에 고난도 질문을 선별하고 이를 학습에 활용해 성능을 입증한 선행 연구[9]에서 아이디어를 차용하여, 난이도라는 새로운 접근에 기반한 예시 선정 방법론을 제안한다. 자세하게는 거대 언어 모델로 질문 간 난이도를 추정하고, 이 과정에서 높은 난도의 질문을 체계적으로 선별하기 위해 토너먼트(Tournament) 방법을 적용한다. 선정된 질문은 few-shot CoT로 제공할 예시 구축에 사용되며, 궁극적으로 추론 작업에서의 성능 향상을 목표로 한다.

본 논문의 이후 구성은 다음과 같다. 2장에서는 본 연구와 밀접한 관련이 있는 선행 연구들을 고찰한다. 3장에서는 본 연구에서 제안하는 스위스 토너먼트(Swiss Tournament)를 적용한 난이도 기반 예시 선정 방법론을 구체적으로 설명하며, 4장에서는 제안 방법론의 실험 설계와 성능 평가 결과를 요약한다. 5장에서는 본 연구의 기여와 한계를 정리하고 향후 연구 방향을 논의한다.

II. Preliminaries

1. Language Model

언어 모델은 단어 시퀀스에 확률을 할당하는 모델[10]로, 오늘날 자연어 처리 전반을 이끄는 핵심 기술로 자리 잡고 있다. 초기에는 n-gram[11]과 같은 통계 기반의 단순 확률 모델이 주를 이루었으나[12], 신경망 기반 모델이 등장하면서 단어 간 의미적 관계를 보다 정교하게 포착하는 방향으로 연구가 발전해 왔다[13]. 이와 같은 언어 모델의 발전은 트랜스포머(Transformer)[14] 아키텍처와 함께 본격화되었다. 트랜스포머는 Self-attention 메커니즘을 통해 문장 내 모든 단어 간의 관계를 동시에 파악할 수 있게 하였으며, 기존의 장기 의존성 문제를 완화하고 병렬 처리 지원을 통해 학습 속도를 개선했다. 이를 근간으로

등장한 BERT[15]와 GPT[16] 모델은 대표적인 사전학습 언어 모델로, 방대한 텍스트 데이터를 통해 언어의 문법과 의미를 미리 학습함으로써 전이 학습만으로도 다양한 하위 작업에 적용할 수 있는 전환점을 마련했다.

구체적으로 BERT는 트랜스포머의 인코더 구조에 기반하여 마스킹된 단어를 맞히는 방식으로 양방향 문맥 정보를 학습함으로써, 문장 이해 작업에서 탁월한 성능을 달성했다. 이와 달리 GPT 모델은 디코더 구조를 통해 앞선 단어들을 보고 다음 단어를 예측하는 자기회귀적 언어 모델링 방식을 통해 자연스러운 텍스트 생성을 가능하게 했으며, 이는 대화형 인공지능 시스템의 발전을 촉진하였다.

나아가 연구자들은 언어 모델의 성능이 모델 크기, 학습 데이터의 양, 그리고 훈련에 투입되는 계산 자원의 확장과 직결된다는 스케일링 법칙(Scaling Law)[17]을 실험적으로 확인하였으며, 이 발견은 거대 언어 모델로 가는 연구의 가속화를 이끌었다. 거대 언어 모델은 수십억에서 수천억 개의 매개변수를 바탕으로 더욱 폭넓은 지식을 습득하며, 기존 사전학습 언어 모델의 성능을 훨씬 능가하고 있다. 최근에는 LLaMA, Gemini[18], 그리고 DeepSeek[19] 등 다양한 최신 언어 모델이 공개되면서, 일상과 산업 전반에서 활발히 활용되고 있다.

2. In-Context Learning

문맥 내 학습은 거대 언어 모델을 작업에 맞게 활용하기 위한 방식 중 하나로, 별도의 매개변수 업데이트 없이 입력 프롬프트에 포함된 지시문 및 예시를 통해 새로운 작업을 수행한다[3]. 이는 모델이 사전학습 과정에서 축적한 지식에 기반하여 작동하되, 사용자가 새로 제공한 작업 지침, 입출력 예시 및 원하는 응답 형식 등을 토대로 문맥에 맞는 적절한 결과를 출력한다.

문맥 내 학습은 제공되는 예시의 수에 따라 zero-shot, one-shot, 그리고 few-shot 학습으로 구분된다. zero-shot 학습은 단순한 지시문만으로 모델이 답변을 생성하는 방식이며, one-shot 학습은 단일 예시를, few-shot 학습은 둘 이상의 예시를 프롬프트에 포함하여 모델의 출력을 유도하는 방식이다[1]. 이들 방식은 모두 추가 학습 과정 없이 다양한 작업에 즉시 적용 가능하다는 특징을 가지며, 사용자의 목적에 따라 선택적으로 사용할 수 있다.

문맥 내 학습은 사전학습된 지식을 그대로 활용하면서 즉각적으로 다양한 문제 해결에 적용할 수 있다는 점에서 기존의 미세조정(Fine-Tuning) 방식과 차별화된다[20]. 특정 작업을 수행하기 위해 추가적인 학습을 필요로 하는 미세조정과 달리, 문맥 내 학습은 모델의 가중치를 조정하지 않기에 시간과 비용 절감이 가능하다[3]. 또한 대규모 주석 데이터가 필수적이지 않으며, 소수의 고품질 예시만으로도 성능을 확보할 수 있어 데이터 수집이 어려운 도메인에서 유용하게 활용된다[21, 22]. 이 외에도 단일 모델을 다양한 작업에 적용하기가 수월하여 새로운 도메인에 대한 적응이 빠르며[23], 단순한 프롬프트 설계만으로 문제 해결을 지원할 수 있어 비전문가도 손쉽게 모델을 활용할 수 있다는 이점을 지닌다[24]. 이에 최근에는 다양한 프롬프트 설계 기법들이 제안되는 방향으로 연구가 확장되고 있다.

3. Chain-of-Thought

전술한 바와 같이, 문맥 내 학습은 추가적인 학습 없이 프롬프트만으로 거대 언어 모델을 다양한 작업에 적용할 수 있는 기반을 마련하였다. 그러나 일반적으로 복잡한 추론 작업의 경우, 단순한 예시 제공만으로는 충분한 성능을 달성하기 어렵다는 한계를 보인다[4]. 이러한 한계는 보다 체계적이고 구조화된 프롬프트 설계의 필요성을 제기하였다.

이와 같은 맥락에서, 프롬프트를 기반으로 거대 언어 모델의 추론 성능을 개선하고자 한 CoT 기법이 제안되었다. CoT는 인간의 문제 해결 과정을 모방하여, 단순히 최종 답변만을 도출하는 것을 넘어 중간 추론 단계를 함께 서술하도록 유도한 방법이다[5]. 궁극적으로 CoT의 핵심은 모델이 답변을 도출하기까지 거치는 일련의 추론 과정 그 자체를 서술하는 것에 있으며, 이를 통해 추론 오류의 발생 가능성을 낮추고 전반적인 성능 향상을 이끌었다.

CoT의 대표적인 방식으로는 zero-shot CoT와 few-shot CoT가 있으며, one-shot CoT는 하나의 상세한 추론 예시를 사용하는 중간 형태이다. zero-shot CoT는 “Let’s think step by step”과 같은 단일 지시문을 통해 예시 제공 없이 단계적 사고 과정을 유도하는 반면[6], few-shot CoT는 추론 과정이 포함된 여러 예시를 제공함으로써 모델이 이를 바탕으로 답변을 출력하도록 한다[5]. few-shot CoT는 복잡한 추론 문제 해결에서 상대적으로 더 높은 성능을 보이는 것으로 알려져 있으나, 프롬프트로 제공되는 예시에 따라 성능이 크게 좌우된다는 한계를 갖고 있다[3, 7, 8, 25].

4. Example Selection for Few-shot CoT

few-shot CoT 학습에서 사용되는 예시는 모델 성능에 직접적인 영향을 미치는 핵심 요인이다. 이러한 예시를 구성하기 위해, 초기 CoT는 연구자가 예시로 사용할 질문을 선정하고 직접 주석을 작성하는 등 수동적인 방식에 의존

하였다[4]. 그러나 이는 상당한 시간과 비용을 수반할 뿐 아니라, 최적화된 예시를 보장하기 어렵다는 본질적인 한계를 지닌다.

Auto-CoT[26]는 인간의 개입에서 비롯된 한계를 극복하고자 제안된 방법으로, 기존의 수동 작업을 자동화된 절차로 대체하였다. 해당 연구는 질문을 선정하는 단계에서 다양성을 핵심 기준으로 삼으며, 이를 자동화하는 과정에서 클러스터링 기법을 적용하였다. 자세하게는 데이터셋의 질문들을 클러스터링한 후, 각 클러스터에서 대표 질문을 하나씩 선정하는 방식을 통해 질문의 다양성을 확보했다. 이후 선정된 질문에 zero-shot CoT를 적용하여 주석을 생성함으로써 수동 설계의 부담을 제거하고, 결과적으로 우수한 성능을 달성하였다.

이후 등장한 Active Prompt는 모델의 불확실성(Disagreement)을 감소시키는 것이 성능 향상에 기여한다는 선행 연구[27] 결과를 토대로, 다양성이 아닌 불확실성에 기반한 질문 선별 방안을 제안하였다. 이에 동일한 질문에 대해 여러 차례 추론을 수행하고, 답변의 일관성을 분석함으로써 불확실성을 정량적으로 측정하였다[4]. 해당 과정에서 일관성이 낮게 나타나는 질문일수록 높은 불확실성을 지닌 것으로 간주하며, 불확실성이 높은 질문들을 주석의 대상 후보로 선별했다. 저자들은 작업 성능 향상을 위해 비교적 적은 수로 선별된 질문에 대해서만 인간 전문가가 주석을 부여하게 했으며, 이를 few-shot CoT 학습의 새로운 예시로 활용하여 추론을 수행하였다. 그 결과, 유의미한 성능 향상을 보임을 입증했다.

5. Difficulty-based Learning Approaches

난이도 기반의 학습과 추론 성능 간의 관계는 오랜 기간에 걸쳐 지속적으로 연구되어 왔다. 특히 커리큘럼 학습(Curriculum Learning)[28]은 쉬운 데이터로 학습을 시작하여 점차 높은 난도의 데이터로 확장함으로써, 점진적 난이도를 반영한 대표적인 학습 방식으로 알려져 있다. 한편, 이러한 접근과는 달리 높은 난도만을 고려한 학습 방식이 우수한 성능을 나타내는 연구 결과도 존재한다.

Less is More for Reasoning[9] 연구는 고난도 질문을 활용한 학습이 모델의 성능 향상으로 이어질 수 있음을 실증적으로 입증하였다. 구체적으로, 모델이 비교적 쉽게 해결한 문제를 제외하는 필터링 과정을 통해 높은 난도의 문제만을 남기고, 이를 학습 자원으로 활용하여 성능을 개선하였다. 이러한 결과는 모델이 상대적으로 어려움을 보이는 문제가 중요한 학습 자원으로 활용될 수 있음을 시사한다.

이렇듯 few-shot CoT 학습은 예시 선정에 있어 다양성, 불확실성 등 다양한 지표를 고려해왔다[4, 26, 29]. 하지만 비지도 학습 환경에서 난이도에 기반한 접근은 여전히 부족한 실정이다. 이에 본 연구는 고난도 문제를 학습에 활용한 선행 연구[9]의 관점을 차용하여 난이도 기반의 예시 선정 방법을 구현하고자 하며, 이를 통해 결과적으로 모델의 추론 성능 향상을 도모하고자 한다.

III. The Proposed Method

1. Research Process

본 장에서는 고난도 질문을 자동으로 선별하고, 이를 few-shot CoT 예시로 구축하여 추론을 수행하는 방법론을 소개한다. 제안 방법론의 전체 과정은 Fig. 1과 같다.

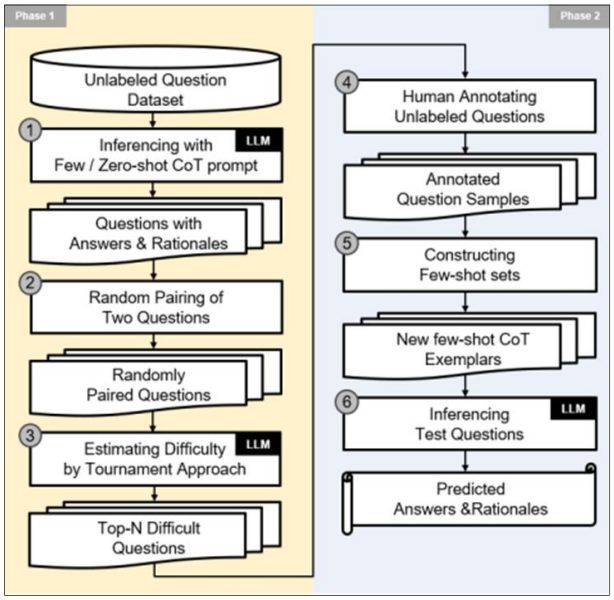

Fig. 1. Overall Research Process

Phase 1은 난이도에 기반해 질문을 선정하는 단계로, 먼저 질문 데이터에 few-shot CoT 또는 zero-shot CoT를 적용하여 질문, 사고 과정, 그리고 답변 구조의 데이터셋을 구축한다(①). 이후 두 개씩 무작위로 짝지어 쌍(Pair)을 구성하고(②), 토너먼트 알고리즘에 기반하여 여러 라운드에 걸친 난이도 평가를 수행한다(③). 이러한 절차를 통해 최종적으로 상위 N개의 고난도 질문을 선별한다.

Phase 2에서는 앞에서 선별된 질문에 대해, 인간 주석자가 직접 질문에 대한 사고 과정과 답변을 작성한다(④). 주석이 완료된 질문들은 새로운 few-shot CoT의 예시로 활용되며(⑤), 테스트 질문 앞에 추가되어 추론을 수행한다(⑥).

각 단계에 대한 세부적인 프로세스는 본 장의 이후 절에서 상세히 설명하며, 실제 데이터를 적용한 제안 방법론의 성능 평가 결과는 4장에서 제시한다.

2. Data Construction with Few/zero-shot CoT

본 절에서는 Fig. 1의 Phase 1에서 이루어지는 절차 가운데, CoT 프롬프팅 기법을 통한 데이터셋 재구성 과정을 소개한다(①).

제안 방법론은 정답 레이블 없이 질문만으로 구성된 데이터셋을 사용하며, 실제 정답이 주어지지 않은 문제의 난이도를 평가하기 위해 언어 모델을 활용한다. 그러나 해당 과정에서 모델이 활용할 수 있는 정보는 제한적이며, 특히 질문의 표면적 텍스트에만 의존하여 문제의 난이도를 충분히 포착하기에는 어려움이 존재한다[30, 31, 32]. 본 연구에서는 모델이 생성한 주석이 학습 및 추론 개선에 유효한 신호로 작용할 수 있음[33]에 착안하여, 주어진 질문을 사고 과정과 답변을 포함하는 확장된 형태로 재구성하는 작업을 수행한다.

이를 위해, 제안 방법론은 주어진 질문 데이터에 단순 지시문을 주는 zero-shot CoT, 또는 소수의 예시를 제공하는 few-shot CoT 프롬프팅을 사용한다. 구체적으로, 각 질문에 대해 zero-shot 또는 few-shot 프롬프팅 방식을 적용하고, 모델은 이를 기반으로 질문에 대한 사고 과정과 답변을 생성한다(Fig. 2).

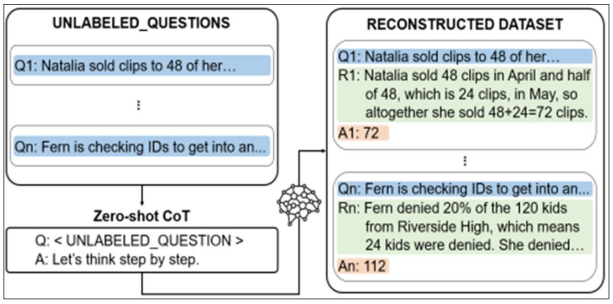

Fig. 2. Dataset Reconstruction via zero-shot CoT

Fig. 2는 주어진 질문 "Natalia sold clips to 48 of her … ?"에 대해 zero-shot CoT 프롬프트인 "Let's think step by step"을 적용하여 질문(Q), 사고 과정(R), 그리고 답변(A)으로 구성된 트리플(Triple)을 생성한 예를 보인다. 이렇게 재구성된 데이터셋은 질문과 답변뿐 아니라 문제 해결의 중간 추론 단계까지 명시적으로 포함하고 있으며, 이러한 구성 요소들은 추후 모델이 난이도를 평가하는 과정에서 추가적인 정보로 활용된다.

3. Pairwise Difficulty Comparison via Swiss Tournament

본 절에서는 앞서 구축된 데이터셋을 무작위로 두 개씩 매칭하고(②), 토너먼트 기반 난이도 추정을 통해(③) 고난도 질문을 선별하는 과정을 소개한다. 일반적으로 질문의 난이도를 추정하기 위해서는 정답률 또는 전문가의 난이도 판단과 같은 명시적인 난이도 레이블이 필요하다[33]. 그러나 실제 환경에서는 이러한 사전 정보가 부재한 경우가 대부분이기 때문에, 이와 같은 레이블 구축에 막대한 시간과 비용이 소요된다[34]. 따라서 본 연구에서는 별도의 레이블이 없는 비지도 환경에서의 난이도 추정을 목표로, 쌍대 비교 기반의 난이도 평가를 수행한다. 해당 절차의 핵심은 언어 모델에 두 문제를 한 쌍으로 제시하고 어느 쪽이 더 어려운가에 대한 판단을 반복적으로 요청함으로써, 전체 문제의 난이도 순위를 점진적으로 파악하는 것이다.

본 논문에서는 모델이 동시에 비교할 수 있는 데이터의 개수를 두 개로 한정하였으며, 이를 위해 앞서 구축된 데이터셋에서 질문을 무작위로 두 개씩 짝지어 쌍을 구성한다(Fig. 3). 해당 과정에서 각 쌍은 질문, 사고 과정, 그리고 답변의 동일한 구조를 유지한다.

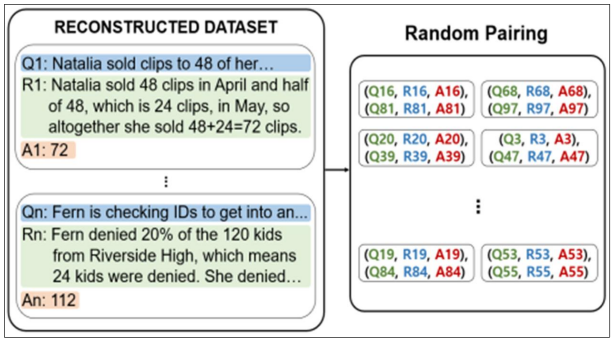

Fig. 3. Random Pairing

이렇게 매칭된 각 쌍은 난이도 비교를 요청하는 프롬프트와 함께, 언어 모델의 입력으로 제공된다(Fig. 4).

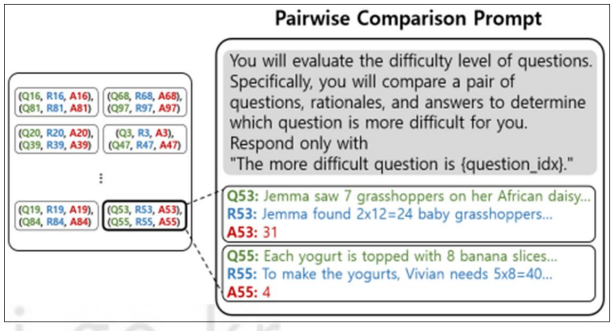

Fig. 4. Pairwise Comparison

쌍대 비교 방식은 전체 문제를 동시에 평가하여 난이도 점수를 산정하는 방식과 비교하여 몇 가지 이점을 지닌다. 우선 한 번에 비교할 대안의 수를 최소화하여 직관적인 판단이 가능하고, 순간적인 인지 부하를 완화할 수 있다[35, 36]. 아울러, 명확한 기준을 사전에 정의하기 어려운 상황에서도 두 문제 간 상대적 우위 판단만으로 난이도 비교를 수행할 수 있다[37]. 언어 모델은 이와 같은 원리를 바탕으로 주어진 두 문제 중 난도가 높은 문제의 인덱스를 선택하는 과정을 수행한다.

이러한 쌍대 비교는 토너먼트 구조로 확장되어 여러 라운드를 거치며 진행된다. 토너먼트 알고리즘은 다수의 대안을 한 번에 두 개씩만 단계적으로 비교하며, 라운드를 거듭해 최종 우승자 또는 전체 순위를 결정하는 방식이다[38]. 각 라운드에서 상대적으로 더 어렵다고 판단된 문제는 다음 라운드로 진출한다(Fig. 5).

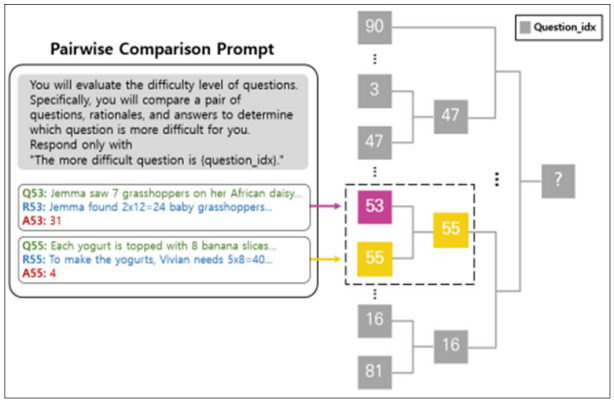

Fig. 5. Tournament based Pairwise Comparison

이를 통해 최종 상위 N개의 고난도 문제를 선별할 수 있지만, 이와 같은 일반적인 토너먼트 방식은 초기 대진 운이 결과에 크게 영향을 미친다는 한계를 갖는다. 예를 들어, 초기 무작위 매칭에 의해 우연히 가장 강한 상대와 만나게 된 참가자는 설령 전체 중 두 번째로 강한 실력을 갖고 있더라도 첫 라운드에 탈락하게 된다[39, 40]. 이러한 한계는 본 제안 방법론에서도 동일하게 적용되어, 문제의 난이도를 제대로 평가하지 못하는 경우가 발생할 수 있다. 이에 본 연구는 이러한 한계를 보완하고자 무작위 매칭의 영향을 완화할 수 있는 스위스 토너먼트 시스템을 채택한다.

스위스 토너먼트 방식은 체스, 바둑, 그리고 e스포츠와 같은 대회에서 널리 사용되는 방식으로, 모든 질문이 동일한 횟수의 비교에 참가하되 라운드마다 유사한 전적의 질문끼리 맞붙는 구조로 설계된다[38, 41]. 먼저 수행할 라운드를 지정한 후, 모든 문제가 0승 0패의 동일한 상태에서 토너먼트를 시작한다. 매 라운드 종료 시, 해당 시점의 누적 승점 기록을 기준으로 동일 점수군 내에서 새로운 매칭이 이루어진다. 해당 과정에서 특정 전적 그룹의 질문 수가 홀수인 경우 부전승으로 처리할 수 있으며, 일정 횟수 이상 패배 시 경기에서 제외되는 규칙을 설정할 수 있다. Fig. 6은 스위스 토너먼트의 예를 보이며, 각 라운드에서의 승자는 흰색, 패자는 검정색, 그리고 부전승은 회색으로 표시되어 있다. 본 그림에서는 참가자가 세 번 패배하는 경우 이후 경기에서 제외되는 규칙을 예시적으로 적용한다.

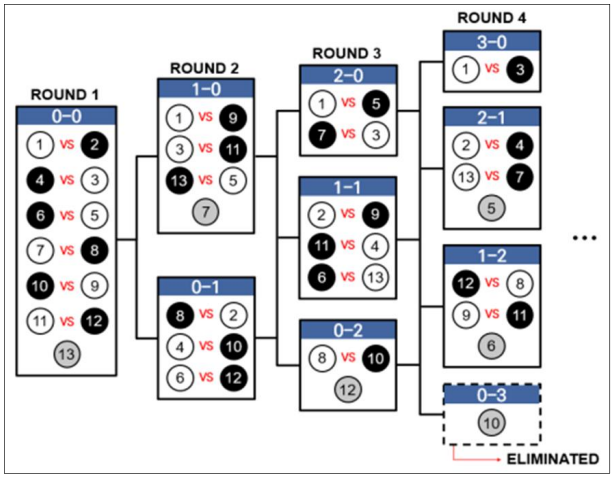

Fig. 6. Swiss Tournament

정해진 라운드만큼 반복을 거치며, 언어 모델은 단계마다 두 질문 중 상대적으로 더 난도가 높은 질문 항목을 선택한다. 이로써 모든 질문은 동일한 횟수의 비교 기회를 부여받으며, 이를 통해 대진 운의 편향에서 벗어난 공정한 난이도 평가 수행이 가능하다. 지정한 라운드가 최종적으로 종료되면, 누적 승률 기준 상위 구간에 포함된 N개의 질문들이 고난도 질문으로 선정된다.

4. Inference using New Few-shot CoT Exemplars

본 절에서는 Phase 2에서 이루어지는 과정 중, 선별된 질문에 인간 전문가가 주석을 달고(④), 이를 새로운 few-shot CoT 예시로 구축한 후(⑤) 테스트 질문의 추론에 사용하는 과정을 소개한다(⑥).

통상적으로 높은 품질의 주석 라벨을 확보하기 위해서는 인간 전문가의 개입이 필요하다[42, 43]. 이에 본 연구는 선별된 상위 N개의 고난도 질문에 한정해 기존 모델이 생성한 주석을 사용하는 대신, 인간 전문가가 작성하고 검증한 주석을 활용한다. 주석자는 각 질문에 대해 사고 과정을 단계적으로 서술하고 최종 답변을 작성하며, 이는 이

후 few-shot CoT 프롬프팅의 새로운 예시로 사용되어 테스트 질문과 함께 입력 프롬프트로 언어 모델에 제공된다. 해당 예시는 모델이 참조해야 할 사고 전개의 시연(Demonstration)으로 작용하며, 모델은 이를 모방하여 새로운 질문에 대한 추론을 단계적으로 수행한다(Fig. 7).

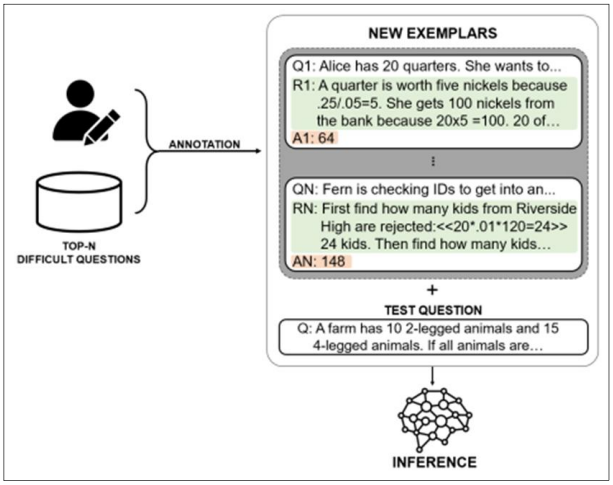

Fig. 7. Pipeline of Human Annotation and Inference

IV. Experiment

1. Experiment Overview

본 장에서는 앞서 소개한 제안 방법론인 난이도 기반 예시 선별 기법을 실제 데이터에 적용한 실험 과정 및 결과를 소개한다.

실험에는 수학 서술형 벤치마크인 GSM8K[44] 데이터를 사용하였다. 본 데이터셋은 여러 단계의 산술 추론을 평가하기 위해 설계된 CoT 추론 기법 평가 데이터셋이며, 약 7,500개의 학습 데이터와 1,300개의 평가 데이터로 구성되어 있다. 본 실험은 Python 3.12 환경에서 수행되었으며, 구체적인 실험 환경은 Table 1과 같다. 또한 실험의 전체적인 프로세스는 Fig. 8과 같다.

Table 1. System Environment

HW	CPU	24 core 3.4GHz
	GPU	NVIDIA GeForce GTX 1660 SUPER
	Memory	32GB
SW	Python	3.12.7
	Pytorch	2.5.1

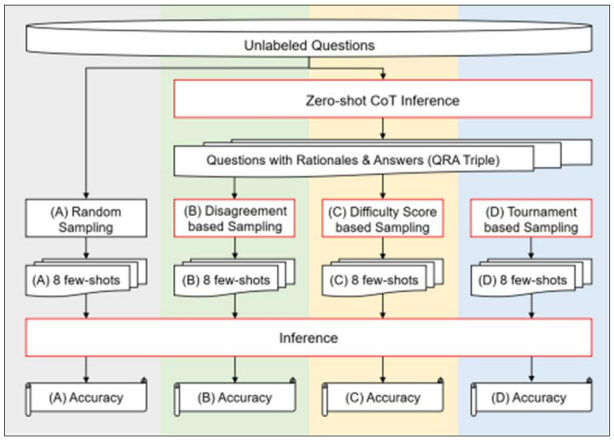

Fig. 8. Overall Process of Performance Evaluation

Fig. 8은 제안 방법론(D) 및 비교 방법론의 상대적인 성능 비교를 위해 설계한 전체 평가 파이프라인을 나타낸다. 모든 방법론에 대해 Few-shot CoT 예시는 8개로 고정하였으며, 비교를 위해 예시를 구성하는 방식만 달리하고 그 밖의 추론 절차와 평가 방식은 동일하게 유지하였다. 도식 내의 빨간색 테두리는 언어 모델이 관여하는 처리 단계를 의미하며, 본 실험에서는 Llama-3.2-1B-Instruct 모델을 사용하였다.

방법론 (A)는 데이터셋에서 질문을 무작위로 선택해 예시를 구성하는 방식이다. (B)는 Active Prompt 방식으로, 동일 질문에 대해 10번 추론한 뒤 응답 불일치를 기반으로 불확실성이 높은 질문을 예시로 선별한다. (C)는 전체 문제를 대상으로 모델이 도출한 난이도 점수에 따라 상위 항목을 예시로 구성한다. 이 방식은 각 문제를 동시에 평가하여 난도를 부여한다는 점에서 (D)와 구별된다. 제안 방법론(D)는 토너먼트 알고리즘 기반의 쌍대 비교 결과를 라운드별로 누적해, 상대 난도가 높은 질문을 점진적으로 선별한다.

이렇게 선정한 상위 8개 질문으로 few-shot CoT 예시를 구성한 뒤, 동일한 추론 프롬프트로 평가 데이터셋에 대해 추론을 수행하고, 도출된 정확도(Accuracy)를 바탕으로 방식별 성능을 비교한다.

2. Data Construction and Example Selection

본 절에서는 데이터셋 구축, 그리고 질문 선별 절차와 결과를 소개한다. 본 연구는 정답 레이블 없이 질문만으로 구성된 데이터셋을 사용하므로, 모델이 활용할 수 있는 정보를 확장하기 위해 각 문항에 대해 zero-shot CoT를 적용하여 질문, 사고 과정, 그리고 최종 답변으로 이루어진 트리플을 생성하였다. 이러한 과정은 예시를 임의로 추출

하는 (A) Random Sampling에는 적용하지 않고, (B)-(D)의 세 방식에만 적용된다.

예시 후보 풀은 학습 데이터 7,473개 질문 중 무작위로 1,000개를 추출하여 구성하였다. 모든 항목에 동일한 프롬프트 포맷과 디코딩 설정을 적용하고, 최대 생성 길이 256 토큰으로 추론을 진행하였다. 본 과정의 결과에 대한 예시는 Table 2와 같다.

Table 2. Example of QRA Triple

Idx	Question	Rationale	Answer
1156	A beadshop earns a third of its profit on Tuesday and...	To find out the profit made on Wednesday, let's follow the...	$1,200
2965	Bertha plays tennis. Every ten games, one of her tennis...	Let's break down the problem step by step: 1. Bertha starts...	8 balls
⋮	⋮	⋮	⋮
7137	John bakes 12 coconut macaroons, each weighing...	To find the total weight of the remaining coconut...	12

이러한 과정을 통해 구축한 트리플을 바탕으로 (B)-(D)의 질문 선별을 진행하였다. Active Prompt 방식 (B)는 각 후보 질문에 대해 10회 반복 추론을 수행하고, 응답 불일치를 불확실성 점수로 환산하여 랭킹을 도출하였다. 그 결과, 10회 모두 불일치를 보인 문항 12개를 도출했으며, 이 중 무작위로 8개를 선택하여 few-shot CoT 예시의 질문으로 사용하였다.

LLM 난이도 점수 방식 (C)는 각 문항에 대해 0-100 범위의 난이도 점수를 도출하도록 간단한 규칙 프롬프트를 적용하였다. 모델은 한 문장으로 난이도를 선언하도록 설계하였으며, 예컨대 “The difficulty of this question is 90.”과 같은 형태로 결과가 반환된다. 형식을 위반한 출력은 제외하였고, 각 문항에 대해 Self-Consistency 기법을 적용하여 반복적인 추론을 통해 난이도 점수를 산출하였다. 이후 도출된 점수를 기준으로 상위 8개 문항을 예시 질문으로 채택하였다.

제안 방법론 (D)는 스위스 토너먼트 기반 쌍대 비교를 통해, 상대 난도가 높은 질문을 점진적으로 선별한다. 총 10라운드를 수행하며, 3패에 도달한 문항은 탈락시킨다. 각 라운드 종료 후에는 유사 전적끼리 재매칭하고, 후보 수가 홀수일 때는 부전승을 부여하였다. 최종적으로 승수 상위 8개 질문을 few-shot CoT 예시로 구축하였다.

이러한 절차를 통해 프로세스별 예시 질문 선정을 모두 완료하였으며, 다음 절에서 동일한 추론 프롬프트와 평가 체계로 성능을 비교한다.

3. Performance Evaluation

본 절에서는 제안 방법론의 성능을 비교 방식들과 동일한 조건에서 평가한 결과를 소개한다.

방식별로 선별된 질문의 사고 과정과 최종 답변은 전문가가 작성 및 검증한 주석을 활용하며, 이를 위해 GSM8K의 기 주석된 데이터[4]를 사용하였다. 해당 과정에서 질문과 주석 간의 정합성은 인덱스 기반 매칭을 통해 확보했으며, 이를 통해 구성된 few-shot CoT 예시는 태스크 질문 추론 시 시연으로 활용하였다. 평가는 GSM8K 테스트 데이터셋 1,319개 문항을 대상으로 수행하였고, 추론 프롬프트는 모든 방식에 동일하게 적용하였다(Fig. 9).

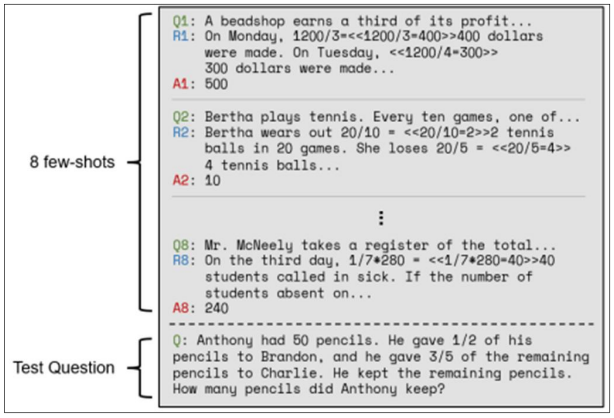

Fig. 9. Inference Prompt Template

생성된 출력에 대해서는 품질 일관성을 확보하기 위해 후처리 규칙을 적용하였다. 수치형 답변은 실수 변환 후 반올림하여 표준화를 수행하였으며, 포맷 오류나 파싱 실패는 빈 문자열로 대치하여 후속 단계에서 제외하였다. 제안 방법론과 비교 방법론의 few-shot CoT 추론 성능은 Table 3과 Fig. 10에 요약되어 있다.

Table 3. Results of Few-shot CoT Inference

Method	Accuracy	Correct
(A) Random Sampling	0.3518	464
(B) Active prompt	0.3594	474
(C) Difficulty based Sampling	0.2403	317
(D) Proposed Method	**0.3730**	492

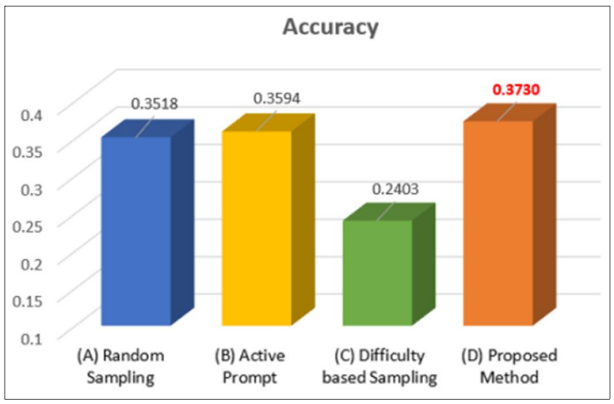

Fig. 10. Method-wise Accuracy

실험 결과, 제안 방법론(D), Active prompt(B), Random Sampling(A), 그리고 LLM 난이도 점수 방식(C) 순으로 높은 정확도를 보이는 것으로 나타났다. 제안 방법론의 정확도는 0.3730으로, Active prompt 방법론 대비 1.36%p의 개선을 보였다. 반면, 난이도 점수를 직접 부여해 상위 항목을 택한 방식(C)은 정확도 0.2403으로 가장 낮은 성능을 보였다. 종합하면, 제안 방법론은 무작위, 불확실성, 그리고 난이도 직접 평가 방식 대비 정확도 측면에서 가장 우수한 성능을 보임을 확인하였다.

4. Comparative Analysis

본 절에서는 제안 방법론의 성능을 보다 면밀히 검증하기 위해 추가적인 비교 실험을 수행한 결과를 소개한다. 모든 실험은 제안 방법론과 동일한 예시 후보 풀 및 실험 환경을 유지한 상태에서 진행하였으며 토너먼트 구조 변화, 그리고 계산 비용 측면에서의 차이를 분석하였다.

먼저, 스위스 토너먼트 방식의 적용으로 인한 성능 향상 여부를 확인하기 위해, 스위스 토너먼트가 아닌 일반 토너먼트 방식을 적용한 추가 실험을 수행하였다. 이때 응답 생성 과정과 평가 기준은 동일하게 유지하여 토너먼트 구조에만 차이를 두어 성능을 비교하였다. 실험 결과는 Table 4와 같다.

Table 4. Performance Comparison Across Tournament Structure

Method	Accuracy	Correct
General Tournament	0.3397	448
Swiss Tournament (Proposed)	0.3730	492

Table 4에서 확인할 수 있듯이, 스위스 토너먼트 구조인 제안 방법론이 일반 토너먼트 방식을 적용한 경우보다 우수한 성능을 보였다. 이를 통해 스위스 토너먼트 구조가 반복적인 쌍대 비교를 통해 초기 대진 편향을 완화하여, 성능 향상에 기여하는 것을 확인하였다.

추가로, 방법론 간 성능 비교의 공정성을 확보하기 위해 각 방법론별 LLM 호출 횟수를 분석하였다. 본 분석은 예시 후보 풀 1,000개를 기준으로 수행되었으며, 평가를 위한 추론 과정은 제외하였다. 분석 결과, 방법론 (B)와 (C)는 각각 총 11,000회의 LLM 호출이 요구된 반면, 제안 방법론 (D)는 약 4,000회의 호출로 수행되었다. 이러한 결과는 제안 방법론이 비교 방법론 대비 상대적으로 적은 호출 횟수 환경에서도 우수한 성능을 보이며, 계산 비용 측면에서도 효율적임을 시사한다.

V. Conclusions

거대 언어 모델의 우수성이 입증됨에 따라, 모델의 내재적 지식을 효과적으로 활용하기 위한 문맥 내 학습에 대한 연구가 이루어지고 있다. 대표적으로 사고 과정을 포함하도록 유도하는 few-shot CoT 프롬프팅은 단 몇 개의 예시 제공만으로 추론 능력을 극대화할 수 있는 것으로 알려져 있다. 그러나 예시 구성이 수작업으로 이루어져 시간과 비용 측면에서 비효율적이며, 선정된 예시를 모든 작업에 고정적으로 사용하는 구조적 제약으로 인해 작업별 최적 성능 확보에 한계가 존재하였다. 이러한 한계를 극복하기 위해 예시 구성의 자동화 연구가 활발히 진행되었으며, 불확실성과 다양성 등 특정 기준을 예시 선정에 반영하기 위한 여러 시도가 이루어졌다.

이러한 흐름 속에서 본 연구는 난이도를 기준으로 질문을 선별하는 방법론을 제안하였다. 제안 방법론은 비지도 환경에서 난이도를 추정하기 위해 zero-shot CoT를 이용하여 트리플 구조를 구축하고, 언어 모델에 쌍대 난이도 비교를 요청하였다. 이를 스위스 토너먼트 기반으로 반복 수행함으로써 누적 승률이 높은 고난도 질문을 선별하였다. 선별된 질문에는 인간 주석을 결합하여 새로운 few-shot CoT 예시를 구성했으며 이를 활용한 추론 실험 결과, 타 지표 대비 향상된 추론 정확도를 확인하였다.

본 연구는 스위스 토너먼트 기반의 쌍대 비교 구조를 활용하여 자동으로 난이도를 추정할 수 있는 절차를 새롭게 제안했다는 점에서 학술적 기여를 인정받을 수 있다. 실험을 통해 난이도 기반 예시 선정이 모델의 추론 정확도를 향상시킬 수 있음을 실증함으로써, few-shot CoT 예시 구성 연구의 방법론적 확장 가능성을 제시하였다. 또한 비지도 기반의 난이도 추정을 통해, 라벨이 존재하지 않거나

데이터 라벨링이 어려운 상황에서도 질문 선정을 효과적으로 수행할 수 있다는 점에서 실무적 기여를 찾을 수 있다. 나아가 전체 질문 중 선별된 일부 질문에만 인간의 주석을 결합하는 방식을 통해, 실제 환경에서 인적 자원을 효율적으로 활용할 수 있을 것으로 기대한다.

다만, 본 연구는 영어 기반의 수학 추론 데이터셋과 특정 언어 모델 환경에 국한된 실험을 수행하였다. 이에 확장 연구를 통해 다양한 태스크 전반으로의 일반화 가능성을 입증할 필요가 있으며, 향후 모델 버전 및 규모에 따른 성능 변동성을 포괄하는 체계적인 분석이 이루어져야 한다. 또한, 스위스 토너먼트 기반의 쌍대 비교를 반복적으로 수행하는 구조적 특성상 예시 집합의 규모가 확장될수록 선정 과정의 계산 시간 및 비용이 증가하므로, 계산 효율성을 고려한 후속 연구가 요구된다.

한편, 본 연구의 실험 결과에서 보고된 성능 차이는 단일 실행을 기반으로 산출되어 통계적 유의성 검증 측면에서 한계가 존재한다. 따라서 반복 실험 및 신뢰구간 분석과 같은 통계적 검정 절차를 적용해 신뢰성 및 재현성을 확보할 필요가 있으며, 이와 함께 정답률 외의 보조 지표를 도입하여 보완적 평가 체계를 구축하는 것이 바람직하다. 나아가 난이도뿐만 아니라 불확실성, 다양성 등의 다중 기준을 통합하여 예시를 선정하는 방안을 모색하는 것이 향후 연구 과제로 남아있다.

REFERENCES

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, "Language Models are Few-Shot Learners," Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS '20), pp. 1877-1901, December 2020.

[2] A. Grattafiori et al., "The Llama 3 Herd of Models," arXiv:2407.21783, Jul 2024. DOI: 10.48550/arXiv.2407.21783

[3] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, B. Chang, X. Sun, L. Li, Z. Sui, "A Survey on In-context Learning," Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), pp. 1107-1128, Miami, Florida, USA, November 2024. DOI: 10.18653/v1/2024.emnlp-main.64

[4] S. Diao, P. Wang, Y. Lin, R. Pan, X. Liu, T. Zhang, "Active Prompting with Chain-of-Thought for Large Language Models," Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1330-1350, Bangkok, Thailand, August 2024. DOI: 10.18653/v1/2024.acl-long.73

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022), pp. 24824-24837, November 2022.

[6] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, "Large Language Models are Zero-shot Reasoners," Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022), pp. 22199-22213, November 2022.

[7] X. Wang, W. Zhu, M. Saxon, M. Steyvers, W. Y. Wang, "Large Language Models Are Implicitly Topic Models: Explaining and Finding Good Demonstrations for In-Context Learning," ES-FoMO 2023 Poster, July 2023.

[8] C. Qin, A. Zhang, C. Chen, A. Dagar, W. Ye, "In-Context Learning with Iterative Demonstration Selection," arXiv:2310.09881, October 2023. DOI: 10.48550/arXiv.2310.09881

[9] Y. Ye, Z. Huang, Y. Xiao, E. Chern, S. Xia, P. Liu, "LIMO: Less is More for Reasoning," arXiv:2502.03387, February 2025. DOI: 10.48550/arXiv.2502.03387

[10] Y. Bengio, "A Neural Probabilistic Language Model," The Journal of Machine Learning Research, pp. 1137-1155, March 2003.

[11] C. E. Shannon, "Prediction and Entropy of Printed English," The Bell System Technical Journal, Vol. 30, No. 1, pp. 50-64, January 1951. DOI: 10.1002/j.1538-7305.1951.tb01366.x

[12] S. F. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," Proceedings of the 34th annual meeting on Association for Computational Linguistics (ACL '96), pp. 310-318, June 1996. DOI: 10.3115/981863.981904

[13] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, "Large Language Models: A Survey," arXiv:2402.06196, February 2024. DOI: 10.48550/arXiv.2402.06196

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention is All You Need," Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017), pp. 6000-6010, December 2017.

[15] J. Devlin, M. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171-4186, Minneapolis, Minnesota, June 2019. DOI: 10.18653/v1/N19-1423

[16] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," 2018.

[17] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, "Scaling Laws for Neural Language Models," arXiv:2001.08361, January 2020. DOI: 10.48550/arXiv.2001.08361

[18] G. Comanici et al., "Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities," arXiv:2507.06261, July 2025. DOI: 10.48550/arXiv.2507.06261

[19] D. Guo et al., "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," arXiv:2501.12948, January 2025. DOI: 10.48550/arXiv.2501.12948

[20] A. K. Lampinen, A. Chaudhry, S. C. Y. Chan, C. Wild, D. Wan, A. Ku, J. Bornschein, R. Pascanu, M. Shanahan, J. L. McClelland, "On the Generalization of Language Models from In-context Learning and Finetuning: A Controlled Study," arXiv:2505.00661, May 2025. DOI: 10.48550/arXiv.2505.00661

[21] L. Wang, N. Yang, and F. Wei, "Learning to Retrieve In-Context Examples for Large Language Models," Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, pp. 1752–1767, St. Julian's, Malta, March 2024. DOI: 10.18653/v1/2024.eacl-long.105

[22] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," ACM Computing Surveys, Vol. 55, No. 9, pp. 1–35, January 2023. DOI: 10.1145/3560815

[23] B. Lester, R. Al-Rfou, and N. Constant, "The Power of Scale for Parameter-Efficient Prompt Tuning," Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. DOI: 10.18653/v1/2021.emnlp-main.243

[24] M. Mosbach, T. Pimentel, S. Ravfogel, D. Klakow, Y. Elazar, "Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation," Findings of the Association for Computational Linguistics: ACL 2023, pp. 12284–12314, Toronto, Canada, July 2023. DOI: 10.18653/v1/2023.findings-acl.779

[25] C. Qin, A. Zhang, C. Chen, A. Dagar, W. Ye, "In-Context Learning with Iterative Demonstration Selection," Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 7441–7455, Miami, Florida, USA, November 2024. DOI: 10.18653/v1/2024.findings-emnlp.438

[26] Z. Zhang, A. Zhang, M. Li, A. Smola, "Automatic Chain of Thought Prompting in Large Language Models," arXiv:2210.03493, Oct 2022. DOI: 10.48550/arXiv.2210.03493

[27] C. Gentile, Z. Wang, and T. Zhang, "Fast Rates in Pool-Based Batch Active Learning," The Journal of Machine Learning Research, Vol. 25, No. 1, pp. 12671–12712, January 2024.

[28] Y. Bengio, J. Louradour, R. Collobert, J. Weston, "Curriculum Learning," Proceedings of the 26th Annual International Conference on Machine Learning, pp. 41–48, June 2009. DOI: 10.1145/1553374.1553380

[29] K. Shum, S. Diao, and T. Zhang, "Automatic Prompt Augmentation and Selection with Chain-of-Thought from Labeled Data," Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 12113–12139, Singapore, December 2023. DOI: 10.18653/v1/2023.findings-emnlp.811

[30] L. Benedetto, P. Cremonesi, A. Caines, P. Buttery, A. Cappelli, A. Giussani, R. Turrin, "A Survey on Recent Approaches to Question Difficulty Estimation from Text," ACM Computing Surveys, Vol. 55, No. 9, pp. 1–37, January 2023. DOI: 10.1145/3556538

[31] Y. H. El Masri, S. Ferrara, P. W. Foltz, J. Baird, "Predicting Item Difficulty of Science National Curriculum Tests: The Case of Key Stage 2 Assessments," Curriculum Journal, Vol. 28, No. 1, pp. 59–82, November 2016. DOI: 10.1080/09585176.2016.1232201

[32] L. Benedetto, A. Cappelli, R. Turrin, P. Cremonesi, "R2DE: A NLP Approach to Estimating IRT Parameters of Newly Generated Questions," Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, pp. 412–421, March 2020. DOI: 10.1145/3375462.3375517

[33] E. Zelikman, Y. Wu, J. Mu, N. Goodman, "STaR: Bootstrapping Reasoning With Reasoning," Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022), pp. 15476–15488, December 2022.

[34] A. Thuy, E. Loginova, and D. F. Benoit, "Active Learning to Guide Labeling Efforts for Question Difficulty Estimation," arXiv:2409.09258, Sep 2024. DOI: 10.48550/arXiv.2409.09258

[35] E. Loginova, L. Benedetto, D. Benoit, P. Cremonesi, "Towards the Application of Calibrated Transformers to the Unsupervised Estimation of Question Difficulty from Text," Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pp. 846–855, Held Online, September 2021.

[36] M. Pérez-Ortiz, A. Mikhailiuk, E. Zerman, V. Hulusic, G. Valenzise, R. K. Mantiuk, "From Pairwise Comparisons and Rating to a Unified Quality Scale," IEEE Transactions on Image Processing, Vol. 29, No. 1, pp. 1139–1151 , August 2019. DOI: 10.1109/TIP.2019.2936103

[37] J. Ramík, "Ranking Alternatives by Pairwise Comparisons Matrix and Priority Vector," Scientific Annals of Economics and Business, Vol. 64, No. SI, pp. 85–95, December 2017. DOI: 10.1515/saeb-2017-0040

[38] R. W. Saaty, "The Analytic Hierarchy Process—What It Is and How It Is Used," Mathematical Modelling, Vol. 9, No. 3-5, pp. 161-176, 1987. DOI: 10.1016/0270-0255(87)90473-8

[39] B. R. Sziklai, P. Biró, and L. Csató, "The Efficacy of Tournament Designs," Computers & Operations Research, Vol. 144, p. 105821, August 2022. DOI: 10.1016/j.cor.2022.105821

[40] C. Groh, B. Moldovanu, A. Sela, U. Sunde, "Optimal Seedings in Elimination Tournaments," Economic Theory, Vol. 49, pp. 59-80, January 2012. DOI: 10.1007/s00199-008-0356-6

[41] A. J. Schwenk, "What is the Correct Way to Seed a Knockout Tournament?," The American Mathematical Monthly, Vol. 107, No. 2, pp. 140-150, February 2000. DOI: 10.2307/2589435

[42] K. Devriesere, L. Csató, and D. Goossens, "Tournament Design: A Review from an Operational Research Perspective," European Journal of Operational Research, Vol. 324, No. 1, pp. 1-21, July 2025. DOI: 10.1016/j.ejor.2024.10.044

[43] J. Ma, Y. Ushiku, and M. Sagara, "The Effect of Improving Annotation Quality on Object Detection Datasets: A Preliminary Study," Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4849-4858, New Orleans, LA, USA, June 2022. DOI: 10.1109/CVPRW56347.2022.00532

[44] Human Annotators in AI: Adding Context & Meaning to Raw Data, https://sigma.ai/human-annotators-in-ai/

[45] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, J. Schulman, "Training Verifiers to Solve Math Word Problems," arXiv:2110.14168, Oct 2021. DOI: 10.48550/arXiv.2110.14168

Authors

Jooeun Lee received the B.S. degree in Statistics from Sungshin Women's University and is currently enrolled in the Graduate School of Business IT, Kookmin University.

She is interested in large language models, natural language processing, and deep learning.

Minseob Song received the B.S. degree in Computer Information Engineering from Inha Technical College and is currently enrolled in the Graduate School of Business IT, Kookmin University.

He received the Best Paper Award at a conference hosted by the Korea Intelligent Information Systems Society. He is interested in natural language processing, deep learning, and time series analysis.

Namgyu Kim received the B.S. in Computer Engineering from Seoul National University in 1998, M.S. and Ph.D. degrees in Management Engineering from KAIST, Korea, in 2000 and 2007, respectively.

Dr. Kim joined the faculty of the School of Management Information Systems at Kookmin University, Seoul, Korea, in 2007. He served as the Dean of the Graduate School of Business IT at Kookmin University and is currently a professor at the Business IT. He is interested in LLM, text mining, and deep learning.