KorEduBench

A Benchmarking Study of LMs for Korean Educational Text Classification

Jeongmin Park, Junghwi Kim, Minwoo Choi, Suhan Jang

Contents

Motivation & Problem Definition

Motivation

Why This Matters

- **LLMs** (Large Language Models) opened new paradigms in NLP
- Yet their **performance in Korean**, a **low-resource language**, and **specialized domains** remains unverified
- In Korean **EdTech**, classifying learning materials (textbooks, handouts, worksheets) enables:
  - **Automated curriculum alignment**
  - **Achievement-level tagging**
  - **Diagnostic feedback generation**
  - **Adaptive student evaluation**
- We focus on **2022 Achievement Standards** in Korea

Our Mission

text

어떤 용액을 만났을 때 그 용액의 성질에 따라 색깔이 변하는 물질을 **지시약**이라고 합니다. 용액에 지시약을 넣었을 때 지시약의 색깔 변화를 기준으로 용액을 분류할 수 있습니다. 색깔, 투명한 정도, 냄새 등 겉보기 성질이 비슷하여 구분하기 어려운 용액은 페놀프탈레인 용액이나 리트머스 종이와 같은 지시약을 이용하여 분류하기도 합니다.

part of textbook

[6과09-01] 여러 가지 용액에 지시약을 넣었을 때의 변화를 관찰하여 용액을 산성 용액과 염기성 용액으로 분류할 수 있다.
By observing the changes that occur when indicators are added to various solutions, we can classify them as acidic or basic solutions.

achievement standards

Text to achievement standard mapping!

Problem Definition

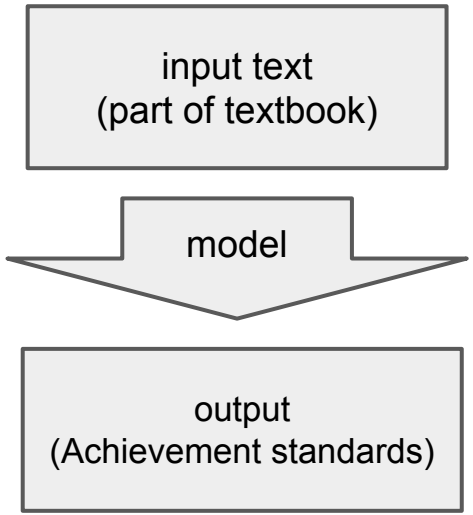
Goal:

Benchmark **zero-shot**, **few-shot**, **and fine-tuned** model performance for classifying Korean educational texts into the **2022 standards**

Key Questions

1. How well LLMs classify texts **without fine-tuning?**
2. Does **few-shot** prompting significantly improve performance?
3. Does increasing **model size** lead to better performance?
4. Can **smaller, fine-tuned models** achieve similar results efficiently?

```
[Sample #157]
Code: 10국01-06
Match Type: exact
Confidence: 1.00
Exact Match: YES
Input Text: 우리는 언어 공동체의 일원으로서 우리의 말과 행동이 어떻게 다른 사람에게 영향을 미치는지 성찰해야 합니다.
Content: 언어 공동체의 담화 관습을 성찰하고 바람직한 의사소통 문화 발전에 기여하는 태도를 지닌다.
True Content: 언어 공동체의 담화 관습을 성찰하고 바람직한 의사소통 문화 발전에 기여하는 태도를 지닌다.
Pred Content: 언어 공동체의 담화 관습을 성찰하고 바람직한 의사소통 문화 발전에 기여하는 태도를 지닌다.
LLM Response: 10국01-06
```



Prior Work and Our Contribution

- ***"Automated Curriculum Analysis Using Large Language Models and Knowledge Graphs" -*** Gacek & Adrian (2025)
  - Extract concepts & prerequisites from syllabi using LLM → Link to Wikidata → Build Knowledge Graph
  - Automated framework for curriculum consistency analysis
- **"Fine-Tuned 'Small' LLMs Still Significantly Outperform Zero-Shot Generative AI Models in Text Classification" -** Bucher & Martini (2024)
  - Compare fine-tuned BERT variants vs. GPT-4, Claude Opus
  - Small fine-tuned models outperform across 4 classification tasks

- **Our Contribution**

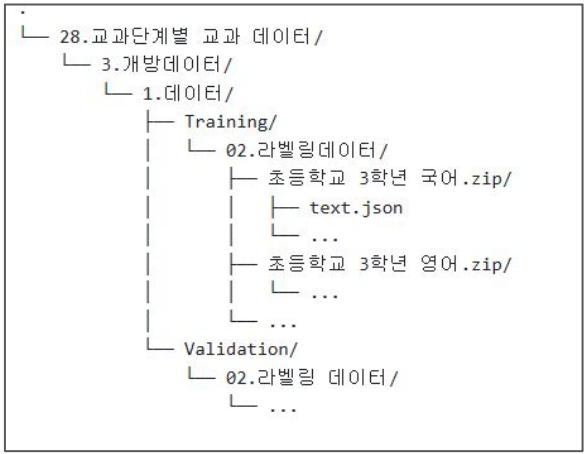
Aspect	Prior Work	KorEduBench
Language	English	**Korean (low-resource)**
Domain	University curricula / General text	**Korean textbook** → Achievement Standards
Scale	Concept extraction / 2-10 classes	**200+** standards (large-scale multi-class)
Pipeline	Single LLM	**Two-stage RAG** (Retriever → LLM)
Model Comparison	Limited	**7B FT vs. 80B** Zero-shot empirical study

Experiments Design

Dataset (Recap)

Raw Data Overview

- Data from *AI Hub*
- Korean educational texts across **elementary, middle, and high schools**
- Subjects include Math, Science, Korean, Social studies, etc.
- Each json has **info data, achievement, actual text, etc**



file structure

json file example

Baseline

1. Raw bi-encoder (ko-sroberta-multitask)
→ Multitask-trained model preserving inter-sentence semantic distance
→ Used as a strong zero-shot baseline for Korean semantic similarity

2. Fine-tuned cross encoder (albert-small-kor)
→ Lightweight yet efficient model; achieves BERT-level contextual understanding with fewer parameters, ideal for fine-tuned reranking.

3. Fine-tuned bi-encoder (KLUE RoBERTa) w/ contrastive learning
→ Robust RoBERTa backbone fine-tuned for task-specific semantic alignment
→ Excels in large-scale retrieval and text–standard matching.

4. LLM (Qwen3-Next-80B)
→ Large-scale language model with strong generalization across domains
→ Capable of implicit semantic reasoning without task-specific training

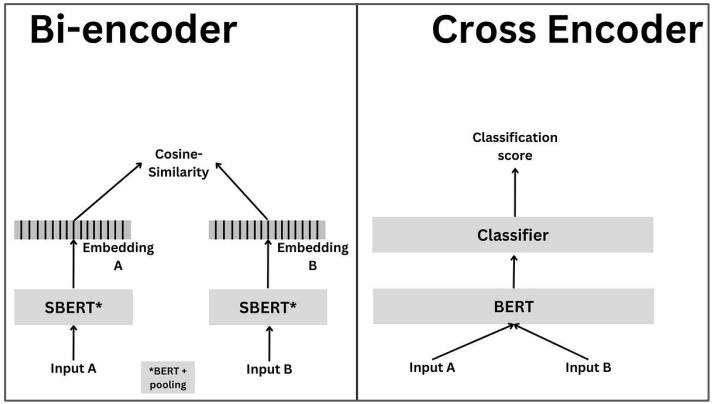
Baseline: bi-encoder / cross-encoder

Model predicts the most relevant *achievement standard* for a given educational text.

Model ① — Sentence-BERT (not fine-tuned)
Embeddings → cosine similarity for each text & achievement standards

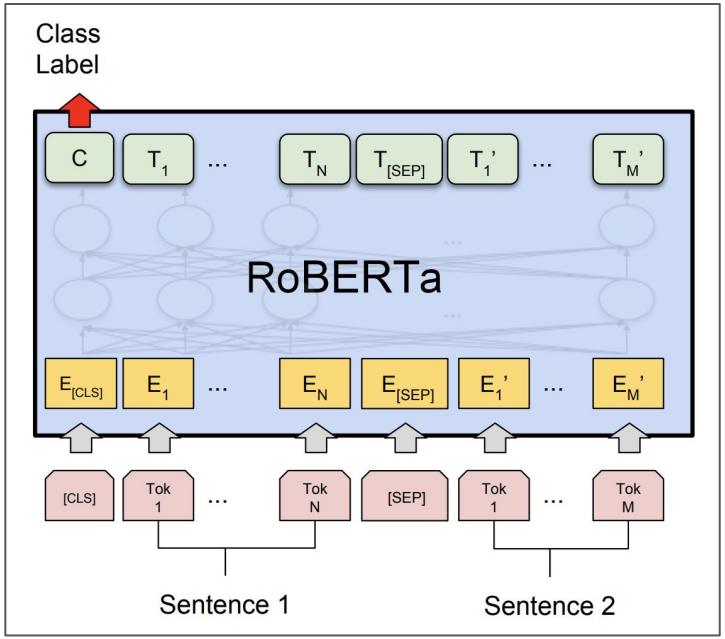
Model ② — Bi-Encoder + Cross-Encoder(fine-tuned) reranking:
Top-20 candidates from Model ① → re-rank using cross-encoder (refine)

Model ③ — Fine-tuned KLUE RoBERTa (contrastive learning):
fine-tune RoBerta on our task → predict by cosine similarity (w/ embeddings)



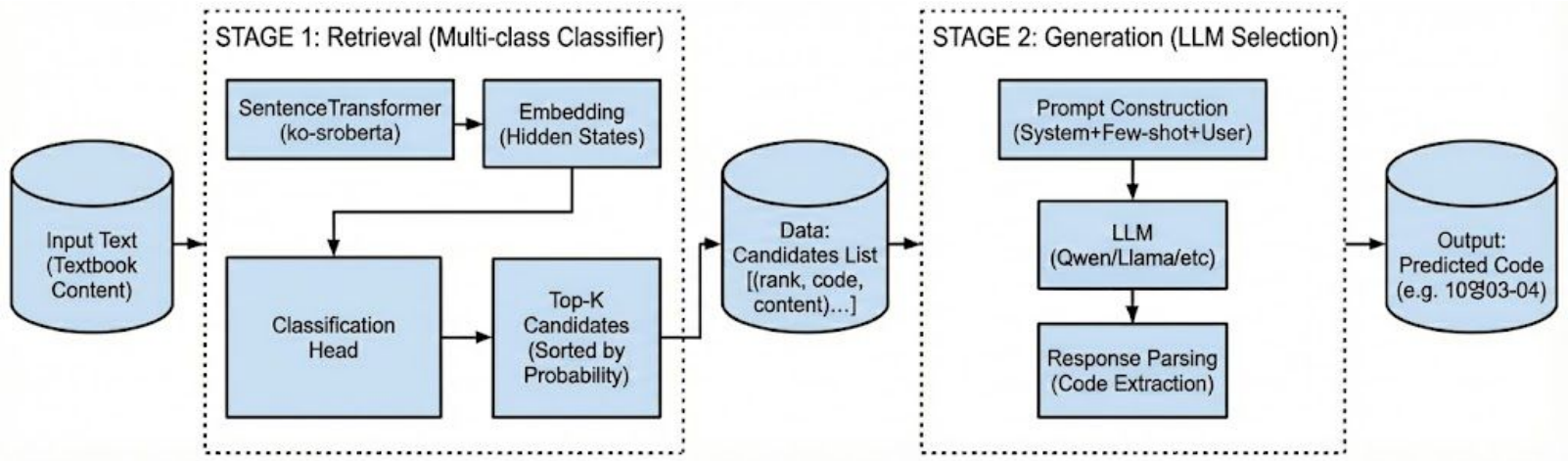
Classifier(Retrieval) Specification

- Fine-tuned to predict relevant achievement standard when text is given
- RoBERTa-large encoder with classification head
- Input: Single sentence (sample text)
- [CLS] token used as pooled representation
- Classification Head: Linear layer + softmax
- Cross-entropy loss for multi-class classification



RAG: Pipeline

- **Two-Stage RAG Pipeline -** Retrieval + Generation separation for reduced prompt length and improved accuracy
- **Stage 1: Learned Retrieval -** Fine-tuned ko-sroberta classifier outputs probability distribution, retrieves top-k candidates (default: 20)
- **Stage 2: LLM Selection -** Selects best match from candidates, supports both local and API-based models, few-shot available(optional).



RAG: Prompt Format

- **Achievement Standard Definition** — Clear explanation of what standards represent (knowledge, skills, expected performance)
- **Matching Guidelines** — Step-by-step instructions for aligning textbook content to standards
- **Task Emphasis & Output Format** — Explicit task description, code-only output constraint, no explanations allowed

input prompt example

top-k candidates

few-shot (optional)

```
# === SYSTEM MESSAGE ===
"""You are an educational curriculum expert. Your task is to select the most appropriate achievement standard from a list of candidates.

WHAT ARE ACHIEVEMENT STANDARDS:
Achievement standards are specific learning objectives that each standard describes:
- The specific knowledge or skills students need to acquire
- The level of understanding or performance expected
- The context or situation where learning should be applied

HOW TO MATCH TEXTBOOK CONTENT TO STANDARDS:
1. Read the textbook text carefully and determine the subject area
2. Identify the primary educational purpose of the text
3. Select the standard that most directly aligns with the main learning goal

# Achievement Standards List
The infer-top-k tool has returned the following candidates (ranked by similarity):
1: 10영03-05; 글을 읽고 주제 및 요지를 파악할 수 있다
2: 10영03-04; 글을 읽고 필자의 의도나 글의 목적을 파악할 수 있다
3: 10영03-01; 글을 읽고 세부 정보를 파악할 수 있다
... (top k=20개까지)

# Few-Shot Examples
Example 1:
Text: 다음 글을 읽고 빈칸에 들어갈 말로 가장 적절한 것을 고르시오. The most important thing in communication is...
Achievement Standard: 글을 읽고 세부 정보를 파악할 수 있다
Answer code: 10영03-01
... (num examples=5개까지)"""

# === USER MESSAGE ===
"""# Textbook Text
Managers of any organization must consider how various employees learn in order to
.... (교과서 지문 전체)

# Task
Analyze the textbook text and select the ONE achievement standard that best matches
its primary educational objective.

# Instructions
Select ONLY ONE achievement standard that best describes the textbook text above.

IMPORTANT: Output ONLY the achievement standard code. Do NOT add any explanations,
reasoning, or additional text.

Correct format:
10영03-04

# Answer"""


# === ASSISTANT MESSAGE (학습 시에만, inference 시 제외) ===
"""10영03-05"""
```

output analysis

```
[Sample #23]
Code: 10영01-03
Match Type: exact
Confidence: 1.00
Exact Match: YES
Input Text: Jake: Yeah, I think it's about using resources in a way that they won't run out.
Content: 친숙한 일반적 주제에 관한 말이나 대화를 듣고 내용의 논리적 관계를 파악할 수 있다.
True Content: 친숙한 일반적 주제에 관한 말이나 대화를 듣고 내용의 논리적 관계를 파악할 수 있다.
Pred Content: 친숙한 일반적 주제에 관한 말이나 대화를 듣고 내용의 논리적 관계를 파악할 수 있다.
LLM Response: 10영01-03
------------------------------------------------------------
```

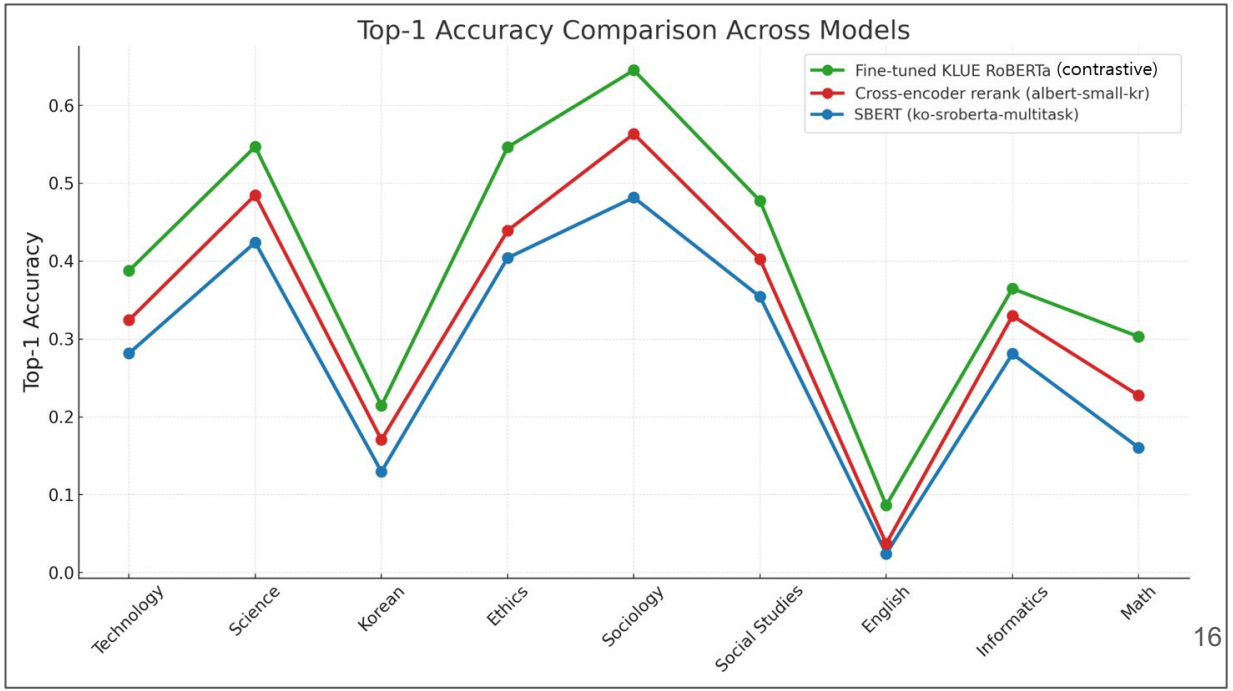
Experiments Result

Baseline: bi-encoder / cross-encoder

Bi-Encoder + Cross-Encoder > Plain Bi-Encoder

Contrastive Fine-Tuning boosts performance on core subjects:

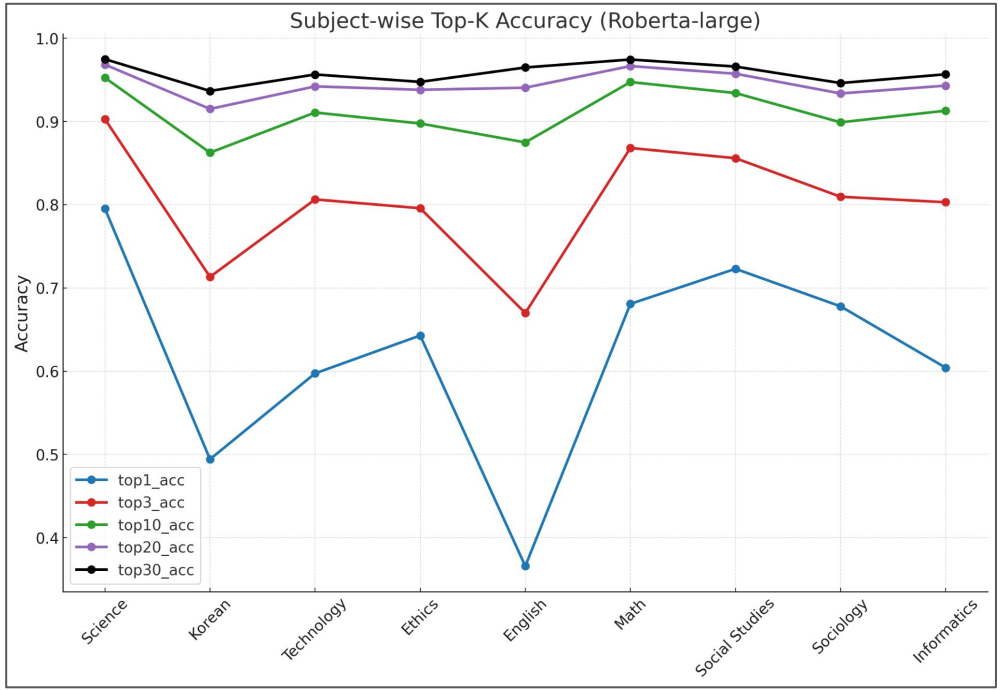
- Science ↑ to 54.7%
- Social studies ↑ to 64.5%
- Math ↑ to 39.0%
- Korean ↑ to 21.4%



Classifier(Retrieval) Results

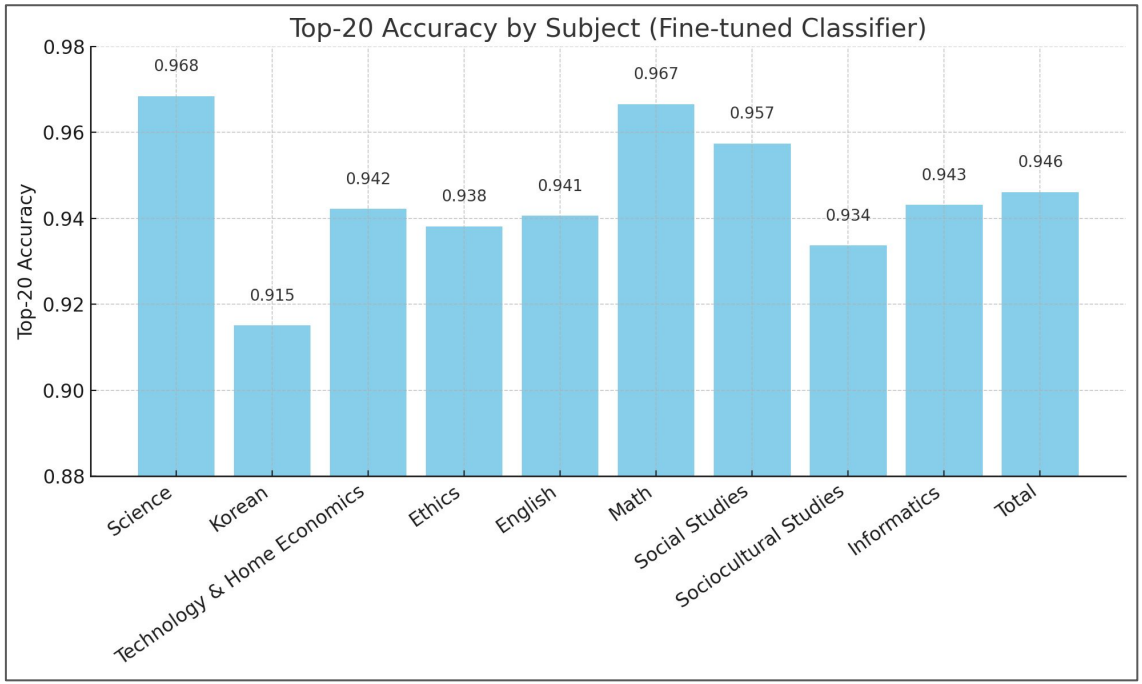
ko-sroberta classifier

- Significant accuracy gains from Top-1 → Top-10 → Top-20
- Accuracy gain sharply decreases beyond Top-20
- Top-30 only ~1–2% higher than Top-20
- Correct standard is usually within the Top-20



Classifier(Retrieval) Results

- RoBERTa-large
- For Top 20, accuracy > 90%
- most subjects around 94~96%
- High generalizability of the fine-tuned classifier for candidate selection



Exp 1: Subjects

1. **How accurate are LLMs in classifying texts by subject?**
   - Strong Performance in **Math** and **Science**
   - Strong in **Social Culture** (except Qwen)

Table 1: Few-shot (k=5) accuracy of four models (best bold, second underline).

Subject	Qwen3-80B	Llama3.2-70B	GPT-4.1-mini	Gemini-2.5-Flash
Science	0.630	**0.630**	**0.695**	**0.710**
Korean Language	0.370	0.305	0.365	0.330
Technology & Home	0.390	0.395	0.445	0.420
Moral	0.555	0.495	0.615	0.560
Social Studies	0.505	0.495	0.615	0.605
Social Culture	0.505	0.575	0.705	0.680
Mathematics	**0.650**	0.530	0.605	0.640
English	0.225	0.190	0.255	0.265
Information	0.510	0.450	0.520	0.530
Average	0.482	0.452	0.536	0.527

table for experiment 1

Exp 1: Subjects

1. How accurate are LLMs in classifying texts by subject?

Strong for Math and Science

- **Standardized text patterns**: Learning objectives highly structured (phrases like “calculates ~”, “understands ~”, or “explains ~”)
- **Clear conceptual boundaries:** Sentence variation low, meaning distinctions sharp, and objectives easier to separate.
- **Low abstraction**: Content is largely concrete and fact-based rather than interpretive.
- **Model fit**: These subjects align well with LLMs' strengths in surface-level semantic matching.

→ As a result, Math and Science are the **most accurately classified subjects** across all models.

Exp 1: Subjects

1. How accurate are LLMs in classifying texts by subject?

Weak in English

- **High Semantic Overlap:** Similar phrasings across objectives (e.g., “identify topic” vs. “infer mood”) make distinctions difficult.
- **Abstract Nuances:** English standards involve subtle cognitive differences, unlike the fact-based clarity of Science.

True: "친숙한 일반적 주제에 관한 말이나 대화를 듣고 **주제 및 요지**를 파악할 수 있다."
Predicted: "친숙한 일반적 주제에 관한 말이나 대화를 듣고 **세부 정보**를 파악할 수 있다."

True: "말이나 글의 **주제나 요지**를 파악한다."
Predicted: "말이나 글의 **분위기나 화자나 인물의 심정** 및 의도 등을 **추론**한다."

examples of wrongs in english

Exp 1: Subjects (Error Analysis)

- **Error Analysis with Cosine Similarity**
  - Calculated Cosine Similarity of wrong predict context and true context
    - Invalid-code errors reflect structural failures.
    - High-similarity errors indicate semantically aligned, near-miss predictions.

Invalid code

```
True Content: 빛이 나아가는 현상을 관찰하여 빛이 직진, 반사, 굴절하는 성질이 있음을 말할 수 있다.
Pred Content: INVALID
True Code: 6과02-02
LLM Response: 11과06-03
```

cos-sim:0.45

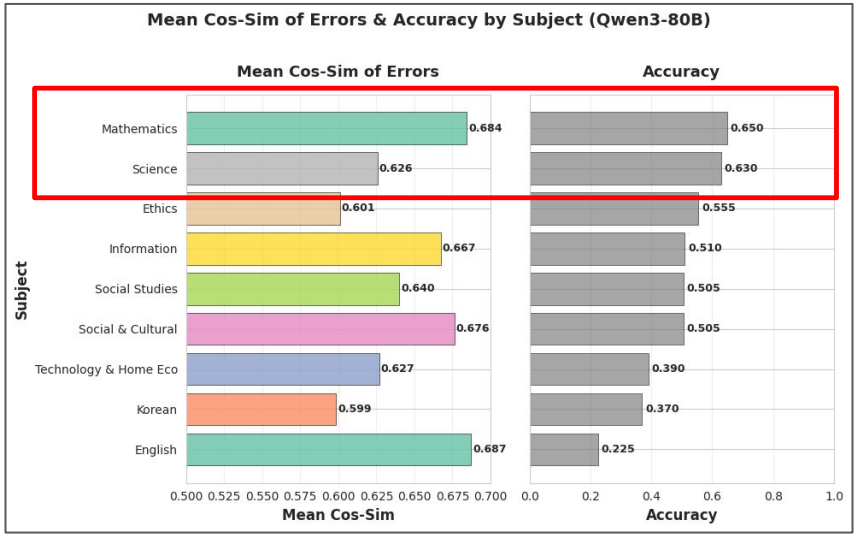
```
True Content: 단어, 어구, 문장의 함축적 의미를 추론한다.
Pred Content: 친숙한 주제에 관해 경험이나 계획을 설명한다.
True Code: 9영01-07
LLM Response: 9영02-04
```

cos-sim:0.89

```
True Content: 실생활 자료를 그림그래프로 나타내고, 이를 활용할 수 있다.
Pred Content: 실생활 자료를 수집하여 간단한 그림그래프나 막대그래프로 나타낼 수 있다.
True Code: 6수05-02
LLM Response: 4수05-01
```

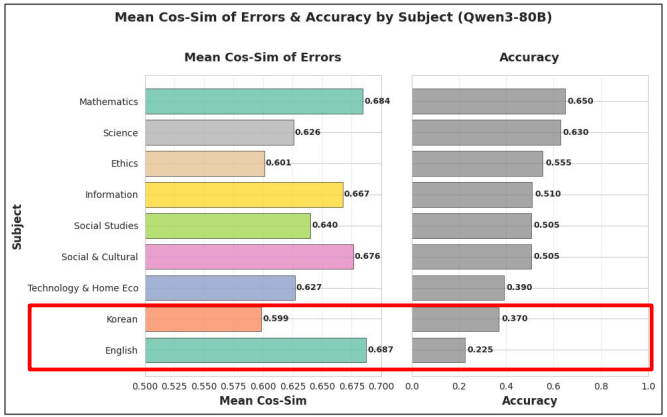
Exp 1: Subjects (Error Analysis)

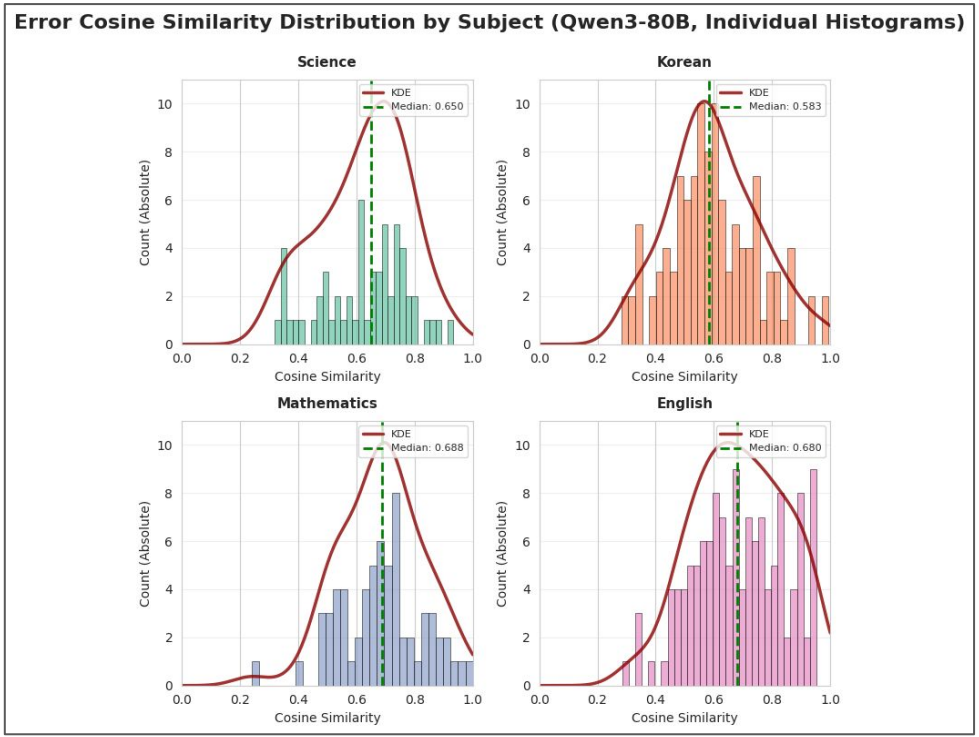
- **Error Analysis with Cosine Similarity**
  - **Math**: high accuracy + high similarity
    → stable and well-aligned understanding
  - **Science**: high accuracy + moderate similarity
    → correct answers but less consistent alignment



Exp 1: Subjects (Error Analysis)

- **Error Analysis with Cosine Similarity**
  - **English**: low accuracy + high similarity → structure is correct; errors arise from fine-grained semantic distinctions
  - **Korean**: low accuracy + low similarity → difficulty capturing both structure and meaning of objectives



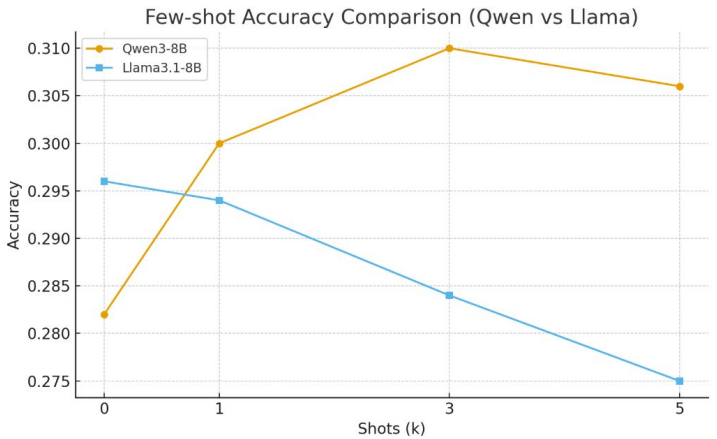


Exp 2: Few-shot Results

2. **Does few-shot learning improve accuracy?**
   - Data using Qwen3-8B and Llama3.1-8B
   - Did **NOT** have significant effects on accuracy
   - showed weak trend but the trend were contrary between Qwen and Llama

Table 2: Few-shot performance comparison (Qwen and Llama).

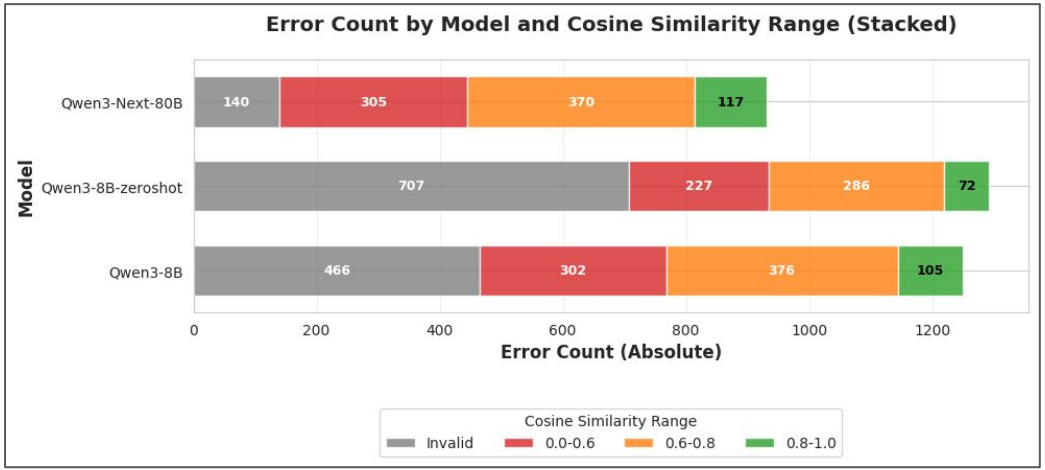
Model	Shots (k)	Accuracy	Δ vs Zero-shot
Qwen3-8B	0	0.282	–
Qwen3-8B	1	0.300	+0.018
Qwen3-8B	3	**0.310**	+0.028
Qwen3-8B	5	0.306	+0.024
Llama3.1-8B	0	**0.296**	–
Llama3.1-8B	1	0.294	−0.002
Llama3.1-8B	3	0.284	−0.012
Llama3.1-8B	5	0.275	−0.021

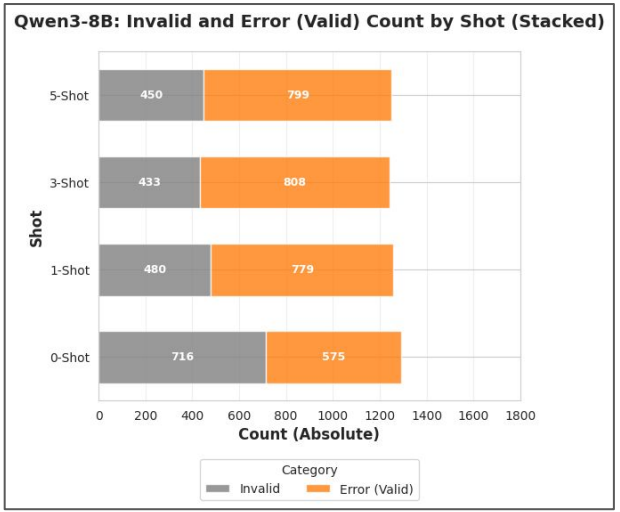


Exp 2: Few-shot Results

2. Does few-shot learning improve accuracy?

- The Few-shot strategy **reduces the Invalid output count (Grey Bar)** compared to zero-shot model.
- The overall shape of the valid error Cosine Similarity distribution is **nearly identical** across models, even with Few-shot prompting.





Exp 3: Model Size Comparison

3. **Accuracy change with increasing model size**
   - **clear upward trend** as model size increases
     - **larger models improve performance** under few-shot conditions.

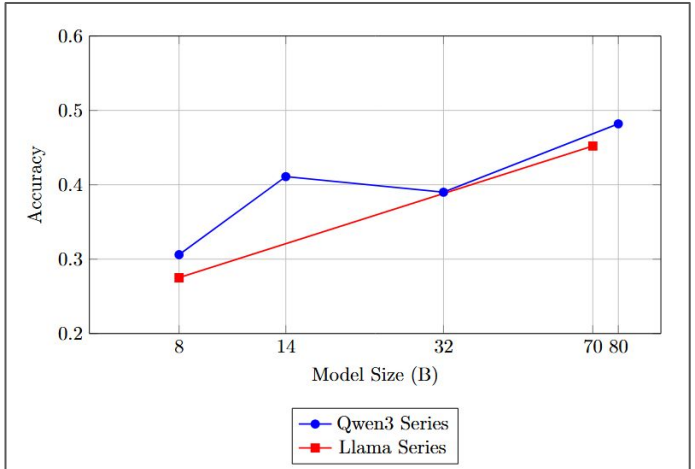


Figure 1: Model size vs. Few-shot (k=5) accuracy (log scale).

Table 3: Effect of model size (Qwen & Llama).

Model	Accuracy
Qwen3-8B	0.306
Qwen3-14B	0.411
Qwen3-32B	0.390
Qwen3-80B	**0.482**
Llama3.1-8B	0.275
Llama3.3-70B	**0.452**

graph/table for experiment 3

Exp 4: Fine-tuning Trade-off

4. **Can small fine-tuned models match large few-shot LLMs?**
   a. Fine-tuning **drastically increased performance**
   b. The fine-tuned 7B model **outperformed** Qwen3-80B few-shot.
   c. It **surpassed closed-source** models (GPT-4.1-mini, Gemini-2.5-Flash)
   d. The fine-tuned 7B model achieved top scores in most subjects

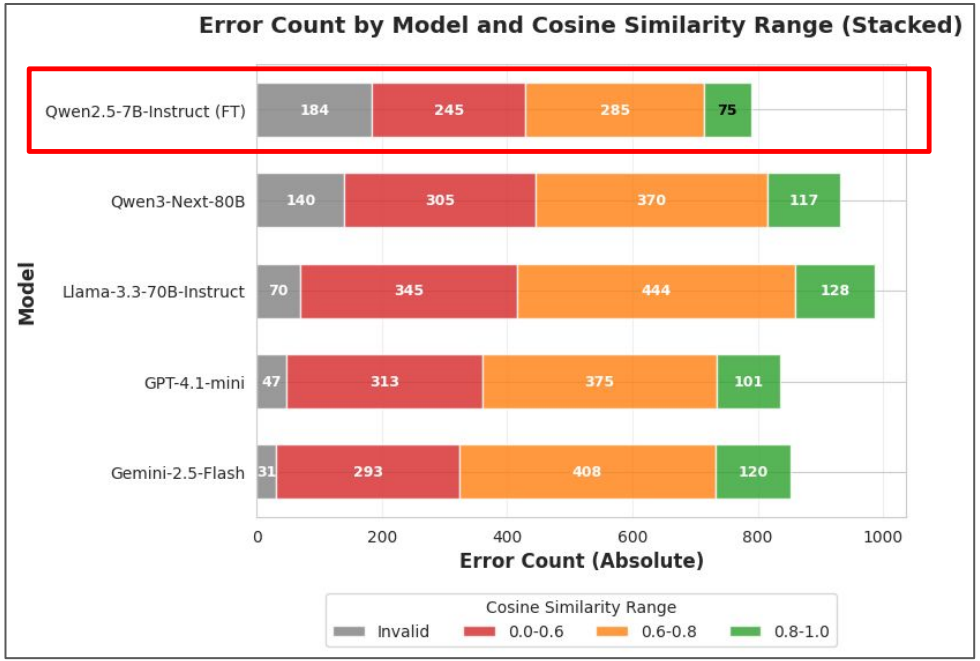
Table 5: Accuracy of five models (best bold, second underline for each subject).

Subject	Qwen2.5-7B-FT	Qwen3-80B	Llama3.2-70B	GPT-4.1-mini	Gemini-2.5-Flash
Science	**0.730**	0.630	0.630	0.695	<u>0.710</u>
Korean Language	**0.375**	0.370	0.305	<u>0.365</u>	0.330
Technology & Home	**0.630**	0.390	0.395	<u>0.445</u>	0.420
Moral	**0.700**	0.555	0.495	<u>0.615</u>	0.560
Social Studies	**0.665**	0.505	0.495	<u>0.615</u>	0.605
Social Culture	**0.730**	0.505	0.575	<u>0.705</u>	0.680
Mathematics	0.515	**0.650**	0.530	<u>0.605</u>	0.640
English	**0.340**	0.225	0.190	0.255	<u>0.265</u>
Information	0.370	0.510	0.450	<u>0.520</u>	**0.530**
Average	**0.562**	0.482	0.452	<u>0.536</u>	0.528

Exp 4: Fine-tuning Trade-off

4. **Can small fine-tuned models match large few-shot LLMs?**
   e. **large number of invalid outputs** (184)
   f. **trade-off**: higher accuracy but weaker robustness
   g. Still, valid predictions retain competitive semantic quality.

- **Practical implication**: a small fine-tuned model can deliver **strong performance** at much **lower cost**, but requires careful validation or post-processing to handle invalid outputs.



Limitation & Conclusion

Limitation

```
[Sample #28]
True Code: 9국04-06
Pred Code: 10국04-04
Match Type: exact
Confidence: 1.00
Exact Match: YES
Input Text: 한글 맞춤법은 '표준어를 소리대로 적되, 어법에 맞도록 함'을 원칙으로 삼습니다.
True Content: 한글 맞춤법의 기본 원리와 내용을 이해하고 국어생활에 적용한다.
Pred Content: 한글 맞춤법의 기본 원리와 내용을 이해한다.
LLM Response: 10국04-04
```

figure: Overlapping Standards example

Data Quality

- **Short Text Samples** — AI Hub dataset contains brief excerpts, some nearly indistinguishable without broader context
- **Overlapping Standards** — Near-duplicate achievement standards exist across subjects, making precise classification inherently ambiguous

Experimental Constraints

- **Limited Experiment Density** — Time/budget constraints prevented thorough hyperparameter tuning, larger model comparisons, and fine-grained few-shot ablation studies

Conclusion

Q1. Can AI be used in education alignment?

→ Yes. LLMs show promising performance, especially in structured domains like Math and Science.
They're not perfect, but not unusable either.

Q2. What needs to improve for reliable classroom use?

→ *Better handling of ambiguous or abstract inputs (e.g., Korean/English standards)*
reduced invalid output, and robust fine-tuning techniques.

Q3. How can we evaluate future models for educational use?

→ *Use our benchmark! It offers a realistic, domain-specific testbed*
with diverse subjects and task types across the Korean curriculum.

Thank you