

KEDI LLM Similarity Evaluation Methodology

November 2025

1. Introduction to LLM Similarity Evaluation

A. **Necessity**

Rapid industrial shifts make it difficult for traditional sector classifications to reflect modern themes in index design. This creates a need for new portfolio construction methods beyond conventional industry standards.

B. **Use of Large Language Models (LLM)**

Traditional TF-IDF methods rely on keyword frequency, which lacks context and requires a complex list of synonyms. LLM technology, however, understands the context and evaluates semantic similarity, allowing for more accurate analysis.

C. **Limitations**

LLM scores do not always correlate with actual financial performance. To address this, the KEDI Index Committee reviews quarterly earnings reports and excludes constituents with low relevance. These are replaced by the next highest-ranking stocks.

D. **KEDI LLM Model**

KEDI uses LLM and embedding methods to measure similarity between companies and industry categories (KICS) or specific keywords.

2. Similarity Evaluation Method using KEDI LLM

A. **Keyword Selection**

KEDI selects keywords based on credible reports from government agencies or official public disclosures. These keywords may be updated following a review by the Index Committee to reflect changes in the industry or market.

B. **Data Collection**

Input data for the evaluation is based on official corporate disclosures.

- **US Companies:** SEC filings such as 10-K, 10-Q, or F20.
- **Korean Companies:** Business reports from the DART (Electronic Disclosure System) operated by the Financial Supervisory Service.

Earnings transcripts may be used to supplement these disclosures. For Korean companies where transcripts are difficult to obtain, refined earnings news data from 'hankyung.com' may be used instead.

C. Corporate Report Generation

Relevant document pools are collected from SEC filings or business reports based on specific themes or keywords. KEDI uses a RAG (Retrieval Augmented Generation) process to ensure the model only references the provided data. This process maintains the reliability of the input data.

D. Similarity Evaluation

i. Primary Evaluation

For companies in the investment universe, the LLM evaluates similarity to keywords based on business reports and categorizes them into five levels:

$$L_{abs}(K, P_i) = \begin{cases} 1.00, & \text{Keyword is a core business with strong competitiveness} \\ 0.75, & \text{Keyword is an active business with competitiveness} \\ 0.50, & \text{Developing or expanding the keyword as part of the business} \\ 0.25, & \text{Not a direct business, but key partners or customers are involved} \\ 0.00, & \text{No relation to the keyword business} \end{cases}$$

L_{abs} : LLM-based absolute similarity score between an individual company and a keyword.

K : Individual keyword.

P_i : Corporate information (Business reports and earnings materials for company i).

ii. Secondary Evaluation: Embedding

A refined secondary evaluation is conducted for companies that scored 0.5 or higher in the primary stage. This stage calculates a final score by measuring the embedding-based similarity between the keyword and the company report.

Embedding-based similarity converts reports and keywords into numerical vectors to measure the distance between them using cosine similarity. The final score ranges from 0.0 to 1.0 for each company.

$$L_{final}(K, R_i) = \begin{cases} 1.0, & \text{High similarity between company } i\text{'s report and keyword } K \\ 0.0, & \text{Low similarity between company } i\text{'s report and keyword } K \end{cases}$$

L_{final} : Final LLM score (Final similarity score between the report and the keyword).

K : Individual keyword.

R_i : Individual company report (Report generated for company i based on the keyword)

※ The embedding model ensures reproducibility by delivering consistent results under the same input data conditions.

iii. **Secondary Evaluation: Categorization**

For companies with a primary score of 0.5 or higher, a more detailed analysis may be performed using sub-categories. In this case, companies are classified into three types (High, Mid, or Low) based on their relevance to specific sub-keywords. This evaluation also uses the corporate reports generated in the first stage.

$$L_{final}(K, R_i) = \begin{cases} High & \text{High similarity to sub – keyword K} \\ Mid & \text{Medium similarity to sub – keyword K} \\ Low & \text{Low similarity to sub – keyword K} \end{cases}$$

L_{final} : Final LLM score

K : Individual keyword.

R_i : Individual company report (Report generated for company i based on the keyword)

※ Individual index methodologies supersede this document in case of any discrepancies

<Disclaimer>

The index content of KEDI (Korea Economic Daily Index), including this document, may not be reproduced, transmitted, or distributed without the written consent of Korea Economic Daily Co., Ltd. The purpose of KEDI's index content is to provide information, and it does not guarantee the accuracy or completeness of the calculation and publication. Korea Economic Daily Co., Ltd. does not express any investment opinions regarding third-party investment products based on KEDI indices, and it has no legal obligation to be involved in disputes arising between users of the indices and third parties through the use of services. Additionally, it does not bear any responsibility for losses incurred due to investment activities or similar actions.