Embedding-based Option Quality Control for Choice Questions in AI-Driven Teaching

Zhenghang Ren
Beijing Zhongguan Academy
s-rzh25@bjzqca.edu.cn

Xin Zou
Beijing Zhongguan Academy
zouxin@bjzqca.edu.cn

Jian Li
Beijing Zhongguan Academy
lijian@bjzqca.edu.cn

Xin Liu
Beijing Zhongguan Academy
v-lx@zgci.ac.cn

Abstract

In AI-driven interactive teaching scenarios, automatically generating high-quality, difficulty-controllable test questions is crucial for enhancing teaching effectiveness. However, traditional Large Language Model (LLM)-based question generation methods face issues such as uncontrollable quality, unreasonable semantic distribution, excessive similarity to correct answers, or insufficient distractiveness in generating distractors, lacking effective quantitative control mechanisms. This study aims to evaluate the application value of cosine similarity of embedding vectors in distractor quality control and difficulty grading for multiple-choice questions. By constructing an evaluation dataset containing 9 core semantic relation types with a total of 1800 samples, and utilizing the Qwen3-Embedding-8B model, we performed cosine similarity calculations and statistical analysis in multi-dimensional spaces from 32 to 512 dimensions. The results indicate that five relation types—conceptual confusion, synonyms, partially correct, antonyms, and overgeneralization—are suitable as distractors; while double negation, syntactic variation, and synonymous different expressions should be filtered out. Dimensionality analysis shows that 128–256 dimensions are sufficient for option-level text representation, with limited gains from higher dimensions. This study provides a quantifiable semantic basis for automated distractor screening and difficulty control in test question generation.

Keywords: Semantic similarity, distractor quality, interactive teaching, question generation, embedding dimensionality

1 Introduction

In the current landscape of smart education, transforming one-way knowledge dissemination into immersive, quantifiable two-way interaction is a core challenge. Platforms such as PQ adopt a "speaker–audience" bidirectional interaction model, using AI to automatically generate quiz questions based on course materials and quantify teaching effectiveness in real time. The practical effectiveness of such platforms heavily depends on the quality of generated questions. However, traditional LLM-based multiple-choice question generation leads to low-quality distractors and uncontrollable difficulty, failing to meet the fine-grained requirements for question difficulty and discrimination in teaching processes [1].

To address this issue, this paper proposes using semantic vector space analysis to construct a quantifiable and controllable distractor generation and screening mechanism. In contrast to previous

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

semantic network-based methods [2], our approach emphasizes quantitative control in semantic vector spaces. Through large-scale semantic similarity experiments, we systematically evaluate the similarity distribution characteristics of 9 typical and common option semantic relations across different embedding dimensions. We then propose a similarity-threshold-based distractor filtering strategy and dimensionality selection recommendations, laying a foundation for building more reliable automated question generation systems.

The main contributions of this paper are as follows:

- Construction of a large-scale evaluation dataset containing 9 semantic relation types and 1800 sentence pairs.
- Systematic analysis of semantic similarity distribution patterns across different dimensions using the Qwen3-Embedding-8B model, with recommendations for dimensions balancing quality and efficiency.
- Validation of the effectiveness of semantic vector space analysis for distractor quality control, providing a reproducible semantic evaluation benchmark for question generation in intelligent education.

2 Methodology

2.1 Experimental Data Construction

We constructed a benchmark dataset for evaluating the semantic quality of distractors. The dataset contains 1800 sentence pairs, evenly distributed across 9 common and representative core semantic relation types, with 200 samples per type. Table 1 provides the definitions and examples for each semantic relation type.

Table 1: Definition and examples of semantic relation types in the dataset.

Relation Type	**Definition**	**Example (Sentence Pair)**
Synonymous different expression	Semantically identical, only differing in expression.	("Integration of classical and Renaissance styles", "Renaissance and classical styles integration")
Double negation	Expresses affirmation through double negation.	("Gothic architecture is not inapplicable to the classical category", "Gothic architecture is applicable to the classical category")
Syntactic variation	Sentence structure changes while core information remains unchanged.	("Gothic architecture, due to its unique structural system, creates a transcendental atmosphere visually", "The unique structural system enables Gothic architecture to create a transcendental atmosphere visually")
Conceptual confusion	Intentionally replaces or confuses two similar but distinct core concepts.	("Gothic architecture is similar to Byzantine architecture", "Gothic architecture is similar to Renaissance architecture")
Synonyms	Uses words with similar but not identical meanings.	("Gothic architecture is elusive", "Gothic architecture is not easy to understand")
Partially correct	Statements contain partially correct information but are overall inaccurate.	("Gothic architecture has only structural significance", "Gothic architecture has both structural and symbolic significance")
Antonyms	Expresses meanings opposite to the original.	("The Renaissance admired Gothic architecture", "The Renaissance disliked Gothic architecture")
Overgeneralization	Inappropriately generalizes specific cases to general situations.	("All medieval architecture is Gothic", "Some medieval architecture is Gothic")
Irrelevant item	Content is completely unrelated to the question stem topic.	("The flying buttress structure of Gothic architecture", "The weather is nice today")

2.2 Embedding Model and Parameters

This study employs the Transformer-based embedding model Qwen3-Embedding-8B, which has demonstrated strong semantic representation capabilities in sentence representation tasks [3]. Qwen3-Embedding-8B performs excellently in multiple embedding tasks [4], making it suitable for semantic similarity computation in this study.

- **Embedding model**: Qwen3-Embedding-8B
- **Similarity metric**: Cosine Similarity
- **Test dimensions (D)**: 32, 64, 128, 256, 512

2.3 Experimental Procedure

1. Compute embedding vectors for each sentence pair across different dimensions.
2. Calculate cosine similarity and summarize statistics.
3. Analyze similarity distribution characteristics of different relation types.
4. Propose thresholds for distractor screening and dimensionality recommendations.

3 Experiments and Results

3.1 Semantic Relation Similarity Overall Distribution

Table 2 presents comprehensive statistics of relation type similarity across all dimensions.

Table 2: Comprehensive statistics of relation type similarity across all dimensions.

Relation Type	Total Samples	Mean	Median	SD	5%–95% Interval
Double negation	200	0.7554	0.8446	0.1942	[0.402, 0.961]
Syntactic variation	200	0.7559	0.8293	0.1947	[0.433, 0.974]
Synonymous different expression	200	0.7388	0.7996	0.1872	[0.425, 0.961]
Conceptual confusion	200	0.7238	0.7873	0.1723	[0.422, 0.932]
Synonyms	200	0.7132	0.7763	0.1885	[0.375, 0.939]
Partially correct	200	0.6817	0.7062	0.1737	[0.377, 0.924]
Antonyms	200	0.6761	0.6956	0.1637	[0.395, 0.908]
Overgeneralization	200	0.6228	0.6205	0.1693	[0.345, 0.890]
Irrelevant item	200	0.2351	0.2290	0.1174	[0.051, 0.432]

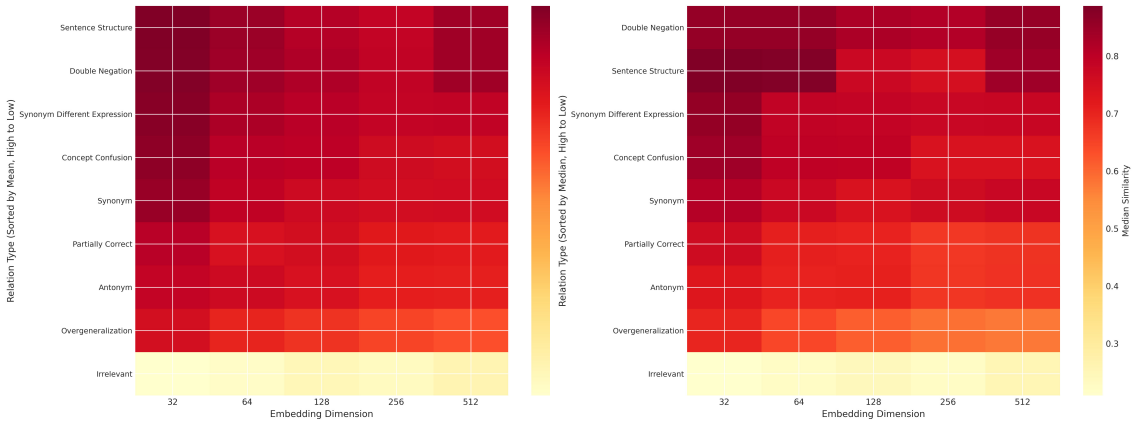
3.2 Key Findings

3.2.1 Semantic Relation Types and Similarity Exhibit Strong Correlation and Interpretable Ranking.

The similarity ranking of semantic relation types demonstrates strong consistency across different measures, as shown in Figure 1. Both the mean similarity heatmap (Figure 1a) and the median similarity heatmap (Figure 1b) reveal identical ordering patterns:

- From a semantic perspective, "synonymous different expression," "double negation," and "syntactic variation" essentially introduce no new information or errors, merely representing the same fact differently.
- From a similarity analysis perspective, these three also rank as the top three in similarity.

Such options are unsuitable as distractors because they do not test whether students grasp the knowledge point but merely rephrase the same content, likely variants of the correct answer. They cannot effectively differentiate whether students truly understand the point of knowledge and should be filtered out. In contrast, types like "conceptual confusion," "synonyms," "partially correct," "antonyms," and "overgeneralization" introduce varying degrees of errors or deviations, with similarity falling within a more discriminative range, making them ideal sources for distractors.



(a) qwen8B Mean Similarity Heatmap

(b) qwen8B Median Similarity Heatmap

Figure 1: Similarity ranking of semantic relation types across different dimensions.

3.2.2 Embedding Dimensionality Has Limited Impact on Similarity Judgment; 128/256 Dimensions Offer the Best Cost-Effectiveness.

The impact of embedding dimensionality on similarity assessment is summarized in Table 3. As dimensionality increases, the average similarity for all relation types exhibits a slight declining trend, indicating that higher-dimensional spaces provide finer semantic differentiation. Apart from a clearer separation between 32 dimensions and others, differences between 64 and 512 dimensions are relatively small, as low-dimensional spaces (e.g., 32 dimensions) have fewer feature dimensions, leading to insufficient discrimination and generally higher similarity, which is unfavorable for fine-grained discrimination. The average coefficient of variation across all relation types is only 0.0446 (far less than 0.1), indicating that dimensionality variation has a relatively limited overall impact on similarity judgment. This pattern is visually confirmed in Figure 2, which shows consistent ranking trends across dimensions. Considering the balance between computational efficiency and performance, for short texts like options, 128 or 256 dimensions already provide sufficiently stable and precise semantic representation, without blindly pursuing higher dimensions.

Table 3: Impact of embedding dimensions on similarity.

Relation Type	Avg. Sim.	Std. across Dims	CoV	Max Diff.	Trend
Double negation	0.7554	0.0255	0.0338	0.0780	Decreasing
Syntactic variation	0.7559	0.0291	0.0385	0.0855	Decreasing
Synonymous different expression	0.7388	0.0267	0.0362	0.0730	Decreasing
Conceptual confusion	0.7238	0.0328	0.0453	0.0927	Decreasing
Synonyms	0.7132	0.0297	0.0416	0.0797	Decreasing
Partially correct	0.6817	0.0278	0.0408	0.0762	Decreasing
Antonyms	0.6761	0.0264	0.0391	0.0669	Decreasing
Overgeneralization	0.6228	0.0372	0.0598	0.1059	Decreasing
Irrelevant item	0.2351	0.0155	0.0658	0.0436	Increasing

Note: Avg. Sim. = Average Similarity; Std. across Dims = Standard Deviation across Dimensions; CoV = Coefficient of Variation; Max Diff. = Maximum Difference. The average coefficient of variation is 0.0446 (CoV < 0.10 indicates low variability). Trends indicate slight changes with increasing dimension.

3.2.3 Distractor Screening Threshold Recommendations

Average similarity gradually decreases with increasing dimensionality, stabilizing from 256 dimensions onward. Using 256 dimensions for further analysis of semantic relations:

- Top three rankings: Synonymous different expression, double negation, syntactic variation.

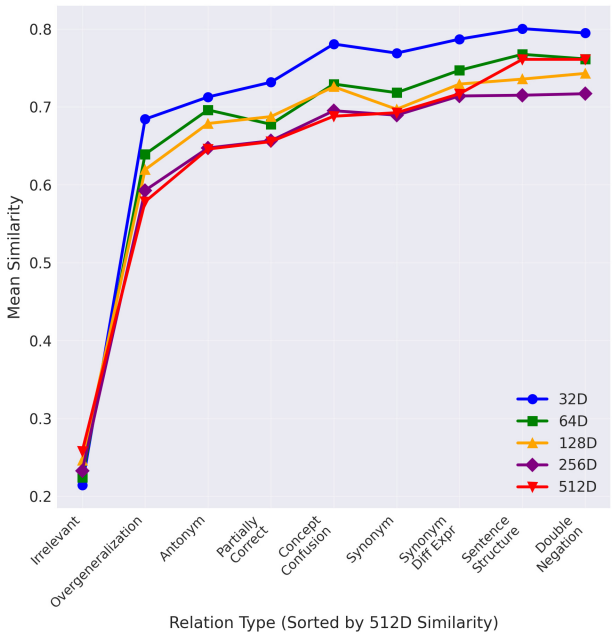


Figure 2: qwen8B Impact of embedding dimensions on similarity.

- Conceptual confusion and synonyms tend to create "higher-difficulty" options.
- Partially correct and antonyms are medium-difficulty options.
- Overgeneralization are low-difficulty options.

Figure 3 shows that options with similarity in the 0.62–0.73 range are most suitable as high-quality distractors. Options in this range ensure moderate semantic differences: "conceptual confusion" and "synonyms" tend to produce higher-difficulty distractors (mean around 0.72–0.71), "partially correct" and "antonyms" belong to medium difficulty (mean around 0.68–0.67), while "overgeneralization" is relatively easier (mean 0.62).

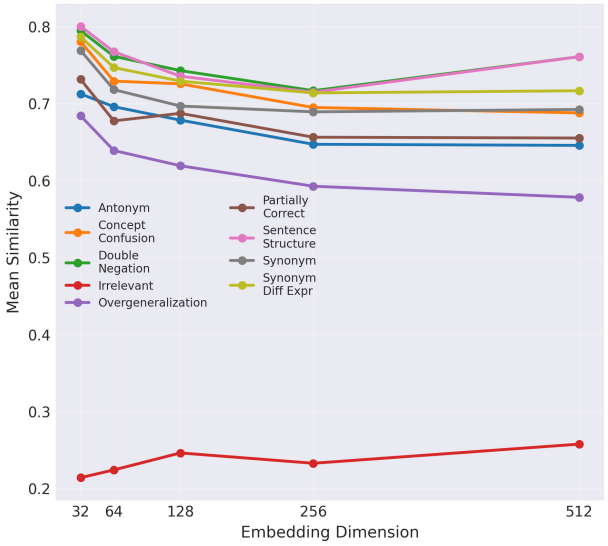


Figure 3: qwen8B Similarity changes of relation types across different dimensions.

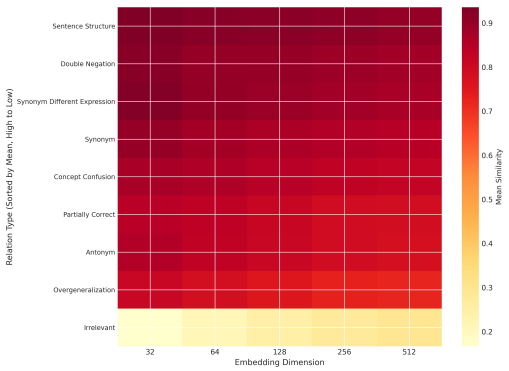
3.3 Comparative Analysis of Different Embedding Models

To verify the universality of our conclusions, we conducted comparative experiments using multiple embedding models of different series and architectures to examine the robustness of similarity distributions. Experiments were first performed on models of the same series but different scales, then extended to multiple publicly available Chinese and English embedding models.

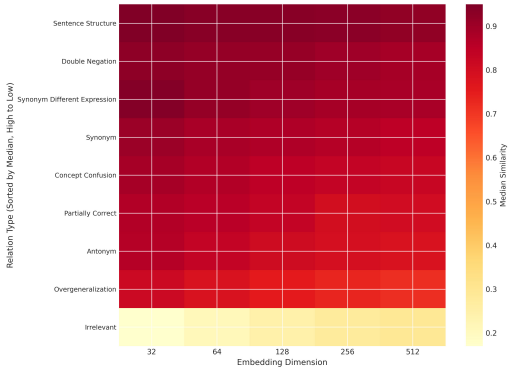
3.3.1 Same-Series Model Verification (Qwen3-Embedding-4B)

Figure 4 shows that conclusions on the Qwen3-Embedding-4B model highly consistent with the main experiment (Qwen3-Embedding-8B):

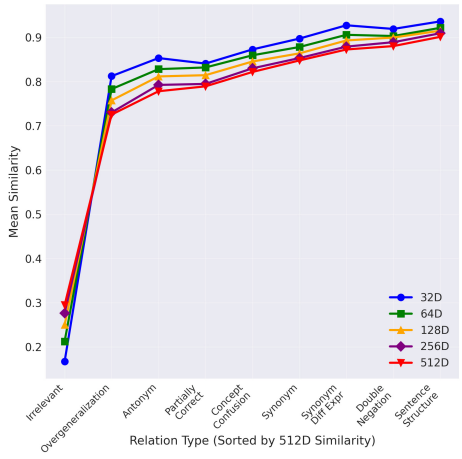
- **Ranking consistency**: The three relation types to be filtered—"synonymous different expression," "double negation," "syntactic variation"—rank top three in similarity; "conceptual confusion" and "synonyms" form the high-difficulty distractor tier; "partially correct" and "antonyms" form the medium-difficulty tier; "overgeneralization" is low-difficulty; "irrelevant item" has the lowest similarity.
- **Dimensionality trend consistency**: The slight declining trend of similarity with increasing dimensionality is the same as in the 8B model, stabilizing around 256 dimensions.
- **Main difference**: The overall similarity values output by the 4B model are generally higher than those of the 8B model, indicating that while ranking, specific screening thresholds need recalibration based on model scale.



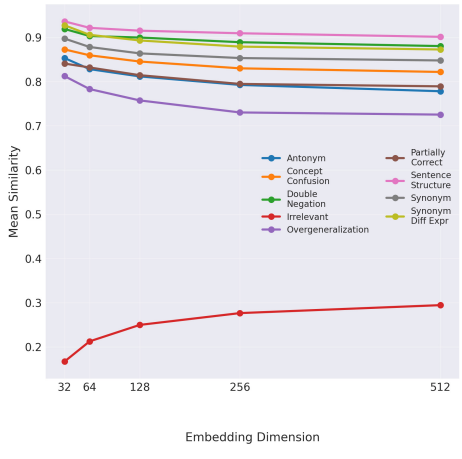
(a) qwen4B Mean Similarity Heatmap



(b) qwen4B Median Similarity Heatmap



(c) qwen4B Impact of embedding dimensions on similarity.



(d) qwen4B imilarity changes of relation types across different dimensions.

Figure 4: Experimental results of Qwen3-Embedding-4B model.

3.3.2 Multi-Model Comparative Verification

To obtain more universal conclusions, experiments were conducted on 5 publicly available models from different series, all widely evaluated on benchmarks like MTEB [5]. Table 4 summarizes the key findings:

1. **Ranking robustness**: Despite significant differences in absolute similarity values output by different models, the relative ranking of semantic relations shows high consistency. Specifically, the ranking patterns of bge-m3 and bge-large-zh-v1.5 are completely identical to those of the Qwen3-Embedding-8B model, while jina-clip-v2, all-mpnet-base-v2, and Youtu-Embedding show largely consistent rankings with only minor variations in ordering.
2. **Threshold model dependence**: Results clearly show that similarity thresholds used for distractor screening (e.g., 0.62–0.73 for Qwen3-8B) strongly depend on the specific embedding model used. For instance, in the bge-m3 model, the effective distractor similarity interval shifts upward overall; in the Youtu-Embedding model, it shifts downward significantly. This indicates that in practical applications, independent threshold calibration must be performed for the selected model; a universal absolute value cannot be directly applied.

Table 4: Similarity comparison across different embedding models (sorted by similarity value for each model).

Model	Relation Type	Value
bge-m3 (dim=1024)	Syntactic Variation	0.902
	Double Negation	0.883
	Synonymous Different Expression	0.849
	Conceptual Confusion	0.809
	Antonyms	0.784
	Synonyms	0.798
	Partially Correct	0.774
	Overgeneralization	0.724
	Irrelevant	0.323
bge-large-zh-v1.5 (dim=1024)	Syntactic Variation	0.871
	Double Negation	0.839
	Synonymous Different Expression	0.838
	Conceptual Confusion	0.802
	Synonyms	0.773
	Antonyms	0.732
	Partially Correct	0.728
	Overgeneralization	0.650
	Irrelevant	0.247
jina-clip-v2 (dim=1024)	Syntactic Variation	0.897
	Synonymous Different Expression	0.852
	Conceptual Confusion	0.809
	Double Negation	0.768
	Partially Correct	0.768
	Synonyms	0.776
	Antonyms	0.763
	Overgeneralization	0.707
	Irrelevant	0.244
all-mpnet-base-v2 (dim=768)	Antonyms	0.821
	Syntactic Variation	0.819
	Double Negation	0.804
	Conceptual Confusion	0.793
	Synonymous Different Expression	0.770
	Overgeneralization	0.728
	Partially Correct	0.744

Continued on next page...

Table 4 – *Continued from previous page*

Model	Relation Type	Value
	Synonyms	0.678
	Irrelevant	0.497
Youtu-Embedding (dim=2048)	Syntactic Variation	0.649
	Conceptual Confusion	0.662
	Double Negation	0.635
	Antonyms	0.604
	Synonymous Different Expression	0.605
	Partially Correct	0.580
	Synonyms	0.540
	Overgeneralization	0.498
	Irrelevant	0.091

Note: For each model, relation types are sorted in descending order of similarity value.

4 Discussion

This study, through systematic experiments, validates the effectiveness of using cosine similarity of text embedding vectors for automated quality evaluation and screening of multiple-choice question distractors. The main contribution lies in providing a quantifiable, interpretable, and operable semantic control dimension for the post-processing of LLM-generated questions.

4.1 Engineering Practical Significance

The method in this study complements automatic distractor generation research [6], which focuses on generation, while this study focuses on semantic quality screening and difficulty control after generation. Systems can set a dynamic or predefined similarity interval (e.g., 0.62–0.73 for the Qwen3-Embedding-8B model) to filter all candidate distractors by LLMs, automatically eliminating options semantically almost equivalent to the correct answer (high similarity) or completely irrelevant (low similarity), retaining options semantically related but with reasonable deviations, thereby improving question quality at the source.

4.2 Limitations and Future Work

1. **Model dependence**: As mentioned, although semantic relation ranking, specific similarity thresholds vary by model. This provides critical guidance for building robust automated question generation systems: first select the base model, then perform threshold learning and strategy deployment based on data from that model.
2. **Single evaluation dimension**: Similarity primarily evaluates options from the perspectives of "semantic relevance" and "difference," but a high-quality distractor also needs to consider logical reasonableness, typical misconceptions of knowledge points, and consistency with the question stem context. Future work should adopt a multi-dimensional evaluation framework, e.g., combining knowledge graph validation and human evaluation.
3. **Application mode suggestions**: To avoid limitations of a single metric, we suggest adopting a "hybrid generation–screening" pipeline in practical applications:
    - **Phase 1 (Guided generation)**: Use prompt engineering to guide LLMs, explicitly requesting generation of specific distractor types (e.g., "Please generate a conceptually confusing option"), controlling candidate quality at the source.
    - **Phase 2 (Similarity screening)**: Use the method from this study to compute similarity and apply threshold filtering on generated candidate options, ensuring moderate semantic distance.
    - **Phase 3 (Human–AI collaboration)**: Provide AI-screened high-quality candidate options to teachers for final review and fine-tuning. Teachers can adjust option expression or difficulty based on personal teaching experience and student characteristics, thereby enhancing efficiency while preserving teaching personalization and professionalism.

Generality note: This study finds that for quality control of LLM-generated text, semantic vector similarity is a highly generalizable and computationally efficient proxy indicator. Although this study focuses on multiple-choice distractors, the method can be generalized to other scenarios requiring evaluation of "reasonable difference" between text pairs, such as in open-ended Q&A, error type classification in text correction, etc.

5 Conclusion

This study systematically investigates a quantitative evaluation method based on semantic vector spaces and cosine similarity to address the issue of uncontrollable distractor quality in AI-driven teaching scenarios. Through experimental analysis of 9 core semantic relation types and 1800 samples across multiple dimensions and embedding models, the following core conclusions are drawn:

1. Distractor semantic types have clear screening value and efficiency differences: The five semantic relation types—"conceptual confusion," "synonyms," "partially correct," "antonyms," and "overgeneralization"—are reliable sources of high-quality distractors due to maintaining topical relevance with the correct answer while having reasonable semantic deviations (similarity falls in the medium range). In contrast, "synonymous different expression," "double negation," and "syntactic variation" should be automatically filtered due to high semantic overlap with the correct answer.
2. Optimal cost-effectiveness in embedding dimensionality selection: For semantic representation of option-level short texts, 128 or 256 dimensions already provide sufficiently stable and precise similarity judgment. Blindly increasing to 512 dimensions or higher yields limited performance gains but significantly increases computational costs.
3. Methodology exhibits cross-model robustness: Although absolute similarity values output by different embedding models differ, requiring model-specific threshold calibration (e.g., threshold interval of 0.62–0.73 for Qwen3-Embedding-8B), the relative ranking and group of semantic relation types remain consistent across models, demonstrating the universality of the method.

In summary, this study provides a solid semantic computational foundation and a replicable engineering solution for distractor quality control in intelligent question generation systems. By using semantic similarity as a core quantitative indicator and incorporating it into a hybrid workflow of "guided generation–automatic screening–human calibration," the reliability and usability of AI-generated questions can be effectively enhanced, promoting deeper and more practical application of AI in teaching assessment.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[2] Lishan Zhang and Kurt VanLehn. Evaluation of auto-generated distractors in multiple choice questions from a semantic network. *Interactive Learning Environments*, 29(6):1019–1036, 2021.

[3] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[4] Bohan Li, Yuxiang Zhang, Lei Gao, et al. Qwen2.5-1.5b/7b/72b-instruct-embedding technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[5] Niklas Muennighoff. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

[6] Lei Gao, Yuxiang Wang, Yifan Zhang, and Yang Liu. Automatic distractor generation for multiple choice questions in standard tests. *arXiv preprint arXiv:2011.13100*, 2020.