SURE or Not? Investigating Semantic Understanding in Dense Retrieval Models

Lingdi Kong[1,2], Xuanang Chen[2]*, Ben He[1,2], Le Sun[2]
[1]School of Computer Science and Technology, University of Chinese Academy of Sciences
[2]Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences
konglingdi24@mails.ucas.ac.cn, chenxuanang@iscas.ac.cn
benhe@ucas.ac.cn, sunle@iscas.ac.cn

Abstract

Dense retrieval has become a core technique in applications like web search and retrieval-augmented generation. Despite their empirical success, it remains unclear whether these models truly understand semantics, and to what degree they can represent semantic consistency and distinguish subtle semantic differences. To address this gap, this paper conducts a systematic investigation by introducing **SURE**, a benchmark for **S**emantic **U**nderstanding in dense **RE**trieval built upon the MSMARCO, NQ, and FiQA datasets. SURE characterizes semantic understanding in dense retrieval along three dimensions: semantic precision, semantic abstraction, and semantic equivalence. We evaluate ten representative models ranging from 110M to 8B parameters, including both general-purpose and domain-specific models. Results show that current dense retrievers struggle to distinguish fine-grained semantic differences across texts with varying information density, and to recognize semantic consistency under lexical paraphrasing. Moreover, larger models do not necessarily exhibit stronger semantic understanding, and diverse training data generally enhances semantic understanding on challenging retrieval tasks. https://github.com/icip-cas/SURE.

1 Introduction

Dense retrieval has advanced rapidly in recent years and become a foundational technique in natural language processing (Zhao et al., 2024), supporting tasks such as question answering (Karpukhin et al., 2020) and web search (Chen et al., 2022). By encoding queries and documents into continuous vectors via pre-trained language models , dense retrieval models overcome the vocabulary mismatch problem of sparse methods like BM25 (Robertson and Zaragoza, 2009), which rely on exact lexical overlap. With the emergence of large language models (LLMs), LLM-based dense retrieval models (Nie et al., 2024; Luo et al., 2024) have achieved state-of-the-art results on comprehensive benchmarks such as MTEB (Muennighoff et al., 2023) and generalize well across domains and languages. These developments highlight their empirical success and growing importance as the backbone of modern information access systems.

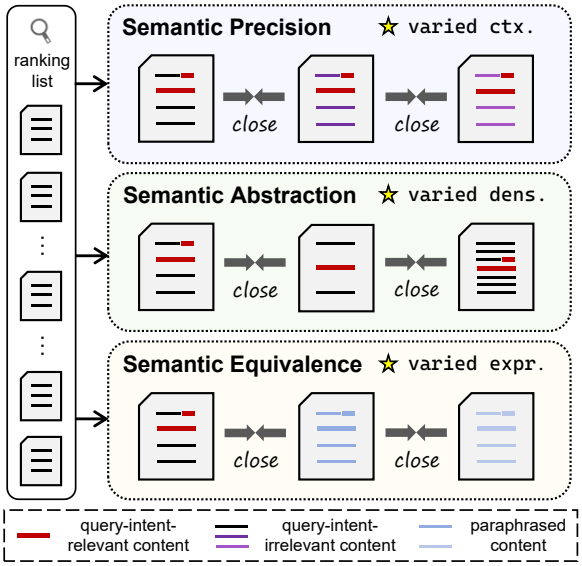

Figure 1: Illustration of the three dimensions of semantic understanding capability for dense retrieval models in retrieval settings. A retriever with strong semantic understanding should rank semantically similar passages close to each other, while remaining sensitive to subtle differences introduced by variations in answer contexts (ctx.), information density (dens.), and lexical expression (expr.).

Despite these advances, it is still unclear whether dense retrieval models possess true semantic understanding, a core capability often assumed to explain their superiority over sparse retrieval methods (Guo et al., 2022). Analyses of LLMs as statistics-of-occurrence models suggest that strong performance may stem from statistical correlations rather than genuine semantic understanding (Titus, 2024), as

*Corresponding Author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45873–45887
July 2-7, 2026

they still struggle with tasks like ambiguity resolution (Yang et al., 2023) and concept composition from words (Riccardi et al., 2024). This raises a natural question for PLM/LLM-based dense retrieval models that rely on semantic relevance: *to what extent do they truly understand semantics?* While most existing studies mainly focus on retrieval performance (Zhang et al., 2025; Shao et al., 2025), this question remains largely unexplored.

To address this, we propose **SURE**, a benchmark for **S**emantic **U**nderstanding in dense **RE**trieval, built on three widely-used datasets including MS-MARCO (Nguyen et al., 2016), NQ (Kwiatkowski et al., 2019), and FiQA (Maia et al., 2018). SURE is designed to evaluate semantic capabilities of dense retrieval models in retrieval settings along three complementary dimensions: *semantic precision*, the ability to capture query-intent–relevant information within passages under diverse contexts; *semantic abstraction*, the ability to recognize the same underlying meaning across passages of varying information density; and *semantic equivalence*, the ability to identify passages conveying the same meaning despite differences in wording or expression, as shown in Figure 1. These dimensions capture core aspects of semantic understanding, with *precision* measuring recognition of key semantic information, *abstraction* measuring information alignment across densities, and *equivalence* measuring robustness to varied expressions.

We systematically evaluate ten representative dense retrieval models from 110M to 8B parameters, spanning domain-specific and general-purpose models such as BGE (Xiao et al., 2024), RepLLaMA (Ma et al., 2024), and Qwen3-Embedding (Zhang et al., 2025). Experimental results show that these models face challenges in detecting subtle semantic differences when texts vary in information density, and are sensitive to lexical variation, which, to some extent, limits models' ability to capture semantic consistency; larger models tend to exhibit stronger semantic understanding, but increasing model size does not necessarily lead to an improvement in the capability; models' semantic understanding on challenging retrieval tasks generally benefits from diverse training data.

Our contributions are three-fold: 1) We introduce SURE, the first benchmark for evaluating semantic understanding in dense retrieval models across three carefully designed complementary dimensions of semantic understanding; 2) We evaluate ten widely-used dense retrieval models, quantifying their semantic abilities across datasets and domains. 3) We reveal key limitations and influencing factors of current dense retrieval models, like model scale and dataset diversity, providing directions for future research.

2 Related Work

Dense Retrieval Models. Dense retrieval models encode texts into continuous vector representations that capture semantic meaning, enabling retrieval beyond surface-level word overlap. This addresses the limitations of traditional sparse models like BM25 (Robertson and Zaragoza, 2009) and TF-IDF (Ramos et al., 2003), which rely on lexical matching. Early dense retrieval models, typically built on pre-trained language models like BERT (Devlin et al., 2019), rely on supervised fine-tuning to learn semantic query–document matching (Karpukhin et al., 2020; Hofstätter et al., 2021). Later studies (Izacard et al., 2022; Wang et al., 2022; Xiao et al., 2024) explore unsupervised and contrastive training on large-scale unlabeled corpora to enhance semantic generalization. With the advance of LLMs, recent work has further leveraged their stronger semantic representation capabilities and rich world knowledge to develop general-purpose embedding models (Muennighoff et al., 2025; Lee et al., 2025; Zhang et al., 2025). Overall, dense retrieval has become a key paradigm for enabling semantically aware and effective information access (Chen et al., 2024, 2025a,b).

Evaluation of Dense Retrieval Models. A variety of studies have explored the evaluation of dense retrieval models from different perspectives. BEIR (Thakur et al., 2021) evaluates the generalization ability of retrieval models in zero-shot semantic retrieval, while MTEB (Muennighoff et al., 2023) has become the standard benchmark for assessing the overall representation quality of text embedding models across retrieval and other text understanding tasks. Apart from comprehensive benchmarks, some works examine robustness across multiple dimensions, including query variations (typos, synonyms, paraphrases) (Penha et al., 2022; Chen et al., 2022; Hagen et al., 2024) and factual fidelity (Wu et al., 2025). Beyond robustness, several studies assess logical reasoning abilities such as boolean logic (Zhang et al., 2024) and negation (Petcu et al., 2025), while others explore the capacity to handle complex queries, including reasoning-based questions (Su et al., 2025), flexible-grained

queries (Li et al., 2025) and multi-condition scenarios (Lu et al., 2025). Despite these studies that evaluates high-level composite capabilities of dense retrieval models, there remains a lack of fine-grained, systematic evaluation methods specifically designed to assess foundational and essential semantic understanding ability.

3 SURE: Semantic Understanding in Dense Retrieval Models

Evaluating whether dense retrieval models truly understand semantics remains a critical but underexplored challenge. Existing benchmarks emphasize end-to-end accuracy yet conflate multiple factors, offering limited insight into semantic understanding. To address this, we propose SURE, a benchmark that isolates and measures semantic understanding through controlled evaluation data and three core capabilities grounded in real-world retrieval needs.

3.1 Semantic Understanding Capabilities

To enable systematic analysis, we decompose semantic understanding into three fine-grained and complementary dimensions that reflect core requirements of effective retrieval: (1) precisely aligning query intent with answer-bearing key content, (2) abstracting consistent meaning across varying information density, and (3) recognizing equivalence across different surface expressions. We refer to these as Semantic Precision, Semantic Abstraction, and Semantic Equivalence.

Semantic Precision is the ability of a dense retrieval model to align query intent with the core information in passages. A semantically precise model should identify key information that directly supports answering the query in passages. For example, when asked about the cause of an event, passages explicitly describing causal relations should be ranked higher and appear more clustered than passages merely describe the event itself.

Semantic Abstraction reflects a dense retrieval model's ability to recognize the same underlying meaning across passages with different levels of information density. In practice, information may appear as concise summaries or detailed elaborations. A model with strong semantic abstraction should identify their semantic similarity regardless of length or granularity, ensuring stable retrieval performance across diverse content expression.

Semantic Equivalence denotes the ability of a dense retrieval model to identify passages that convey the same meaning despite variations in wording or expression. A model with strong equivalence consistently treats paraphrased or reworded passages as equally relevant, making retrieval decisions grounded in meaning rather than surface lexical overlap.

3.2 Evaluation Data Construction

Figure 2 illustrates the overall procedure of our evaluation data construction. For each evaluation of semantic precision, semantic abstraction and semantic equivalence, we generate text variants and perform applicable data validation to construct the corresponding testbed.

3.2.1 Semantic Precision

To evaluate the semantic precision of dense retrieval models, we construct three kinds of passage variants per query. The first contains only key sentences representing the answer-bearing information that directly aligns with the query intent. The second is the full original passage, including key sentences plus additional thematically related content. The third combines key sentences with unrelated, off-topic content. By constructing varying contextual information, these samples probe whether dense retrieval models can accurately discern the essential information embedded within.

Specifically, given a query q and a relevant passage p, we first employ an LLM, by default DeepSeek-R1 (DeepSeek-AI, 2025) unless otherwise specified, to extract key sentences from p.

$$p_{ks} = \text{LLM}(I_{ks}, q, p) \tag{1}$$

where the prompt I_{ks} directs the LLM to perform key sentence extraction, and the resulting passage p_{ks} should retain sufficient and precise information to answer the query q.

Then, we prompt the LLM to generate two sentences, s_1 and s_2, without direct answers to the query q, and concatenate them with the key sentence passage p_{ks} to create a passage p_{ksn} containing both useful information and distracting content.

$$s_1, s_2 = \text{LLM}(I_{na}, q) \tag{2}$$

where the prompt I_{na} instructs the LLM to generate sentences that serve as additional non-answering context.

To mitigate the impact of sentence ordering on the ability evaluation, the key sentence passage p_{ks}

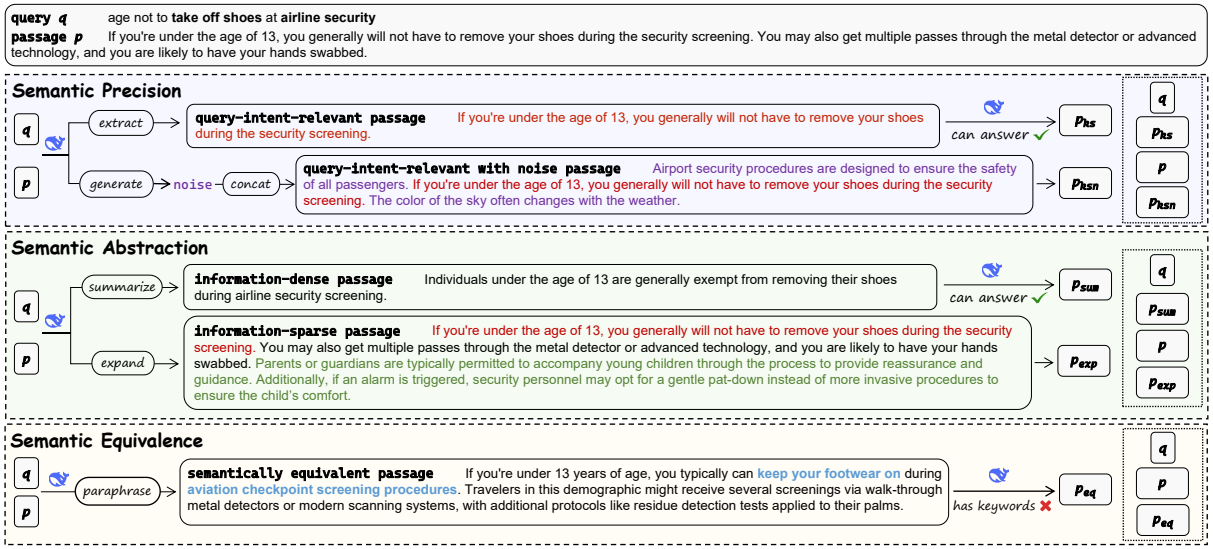

Figure 2: Overview of the data construction pipeline. Passage variants are generated and filtered for constructing semantic precision, abstraction, and equivalence testbeds. The red sentence refers to information aligned with the query intent, while purple, green, and blue sentences introduce semantic variations to test each capability.

is randomly positioned at the beginning, middle, or end of passage p_{ksn}.

$$p_{ksn} \in \left\{ \begin{array}{l} p_{ks} \oplus s_1 \oplus s_2, \\ s_1 \oplus p_{ks} \oplus s_2, \\ s_1 \oplus s_2 \oplus p_{ks} \end{array} \right\} \tag{3}$$

where $\oplus$ denotes the text concatenation operation.

Finally, the LLM filters and refines the generated samples, verifying that key sentences p_{ks} indeed answer the query. The resulting passage set $\{p_{ks}, p, p_{ksn}\}$, together with the query q, constitute the semantic precision evaluation testbed.

3.2.2 Semantic Abstraction

To evaluate semantic abstraction in dense retrieval models, we construct two passage variants with different information densities for each query. This allows us to test whether models can capture the same underlying content across passages that vary in conciseness or elaboration.

Specifically, given a query q and a relevant passage p, we employ the LLM to generate a concise summary p_{sum} and a stylistic expansion p_{exp} that are relevant to q.

$$p_{sum} = \text{LLM}(I_{sum}, q, p) \tag{4}$$

$$p_{exp} = \text{LLM}(I_{exp}, p) \tag{5}$$

wherein I_{sum} and I_{exp} are prompts instructing the LLM to summarize and expand the passage.

The summary p_{sum} preserves all information necessary to answer q while presenting it concisely, forming a high-information-density variant. In contrast, the expansion p_{exp} only adds stylistic elaboration without additional relevant answer content, representing the case of low-information-density. The original passage p lies between these two, serving as a medium-information-density baseline.

Finally, the LLM is further used to verify that p_{sum} retains all key information required to answer the query. The resulting passage set $\{p_{sum}, p, p_{exp}\}$, together with the query q, forms the semantic abstraction evaluation testbed.

3.2.3 Semantic Equivalence

To evaluate dense retrieval models' semantic equivalence capability, we construct semantically equivalent samples by substituting keywords in the original passage with their synonyms, definitions or descriptions. The paraphrased passage shares the identical meaning with the original passage while exhibiting no overlap in keywords, thereby forcing the model to assess semantic equivalence without relying on lexical matching.

Specifically, given a query q and a relevant passage p, we instruct the LLM to identify keywords in passage p with respect to the query q, and perform keyword substitution to generate a paraphrased passage p_{eq}.

$$p_{eq} = \text{LLM}(I_{eq}, q, p) \tag{6}$$

wherein I_{eq} is the prompt instructing the LLM to generate a paraphrase. The resulting passage p_{eq} maintains the same semantic content as p but ex-

hibits minimal lexical overlap, jointly representing alternative expressions of the same semantics.

Finally, the LLM filters passages to ensure all keyword substitutions are complete. The resulting passage set $\{p_{eq},\ p\}$, together with the query q, forms the semantic equivalence evaluation testbed.

3.3 Evaluation Metrics

Through the above construction process, we obtain test data in the form of $(q, p_1, \cdots, p_n)$, where each passage variant preserves the core meaning of the original passage while introducing controlled semantic differences. Since dense retrieval models adopt different score normalization methods and have varied score distributions, the raw similarity scores of retrieved passages are not directly comparable. To address this, we reinsert the constructed passages into the top-2000 retrieved passage list of each dense retriever and evaluate models based on the rankings of the constructed passages rather than their absolute scores.

Ideally, the constructed passages should be ranked close to one another, with their relative ordering further considered, thereby reflecting both semantic similarity and semantic difference. While the relative ordering of constructed passages can be determined by humans, measuring ranking proximity requires exact rank positions, which is infeasible for manual annotation due to the heavy workload as well as potential unreliability in human judgments. Therefore, a reference model is employed to generate reference rankings for the constructed passages within each testbed, and retrievers whose rankings of the constructed passages closely align with the reference rankings can be regarded as exhibiting stronger semantic understanding.

We apply a strong reranker, Qwen3-Reranker-4B (Zhang et al., 2025), to rerank the same set of top-2000 passages combined with constructed passages, and regard the positions of constructed passages in this list as reference rankings. We then define two metrics: Ranking Deviation Consistency (RDC), measuring models' ability to recognize semantic consistency via quantifying ranking deviations, and Ranking Order Consistency (ROC), measuring models' ability to detect semantic differences by checking ranking order.

Ranking Deviation Consistency. Let Q denote the set of queries in a testbed, and n the number of constructed passages per query (e.g., in the semantic precision testbed, each query has three corresponding passages: p_{ks}, p, and p_{ksn}, hence $n = 3$). For a query $q \in Q$, we denote the rankings of the constructed passages by a dense retrieval model $\mathcal{M}$ as $R_{\mathcal{M}} = [r_1, r_2, \cdots, r_n]$ and the corresponding reference rankings as $R_{\mathcal{G}} = [r_1^g, r_2^g, \cdots, r_n^g]$. We compute the standard deviations of rankings in $R_{\mathcal{M}}$ and $R_{\mathcal{G}}$, denoted $\sigma_{\mathcal{M}}^{(q)}$ and $\sigma_{\mathcal{G}}^{(q)}$, to quantify the overall deviation pattern within each ranking set.

The RDC is then defined as the average relative consistency of standard deviation of constructed passages' rankings $R_{\mathcal{M}}$ and $R_{\mathcal{G}}$ across all queries.

$$\mathrm{RDC} = 1 - \frac{1}{|Q|} \sum_{q \in Q} \frac{\left|\sigma_{\mathcal{M}}^{(q)} - \sigma_{\mathcal{G}}^{(q)}\right|}{\max\left(\sigma_{\mathcal{M}}^{(q)}, \sigma_{\mathcal{G}}^{(q)}\right)} \quad (7)$$

Ranking Order Consistency. For a query $q \in Q$, given a model ranking $R_{\mathcal{M}}$ and the reference ranking $R_{\mathcal{G}}$ of constructed passages, we compute the proportion of concordant pairs, i.e., the relative order for pairs $\{(p_i, p_j), 1 \le i < j \le n\}$ in $R_{\mathcal{M}}$ matches that in $R_{\mathcal{G}}$ using a method similar to Kendall's Tau (Kendall, 1938).

$$C^{(q)} = \left\{(i,j) \middle| \operatorname{sgn}(r_i - r_j) = \operatorname{sgn}(r_i^g - r_j^g)\right\} \quad (8)$$

where $\operatorname{sgn}(\cdot)$ indicates the sign of expressions, and $C^{(q)}$ denotes the set of concordant pairs.

The ROC is then defined as the average proportion of concordant pairs across all queries:

$$\mathrm{ROC} = \frac{1}{|Q|} \sum_{q \in Q} \frac{|C^{(q)}|}{\frac{1}{2}n(n-1)} \quad (9)$$

4 Experiments

4.1 Experimental Setups

This section details the SURE benchmark and the dense retrieval models used for evaluation.

Details of SURE. We build our benchmark based on widely-used passage retrieval datasets MSMARCO (Nguyen et al., 2016), Natural Questions (Kwiatkowski et al., 2019), and FiQA (Maia et al., 2018). For each query in datasets, one annotated relevant passage is chosen for data generation. The statistics of the SURE benchmark is summarized in Table 1. Detailed instructions used for the LLM during data construction is shown in Appendix A.

To assess data quality, we randomly sample 50 queries from MSMARCO used in SURE, each paired with five constructed passages p_{ks}, p_{ksn}, p_{sum}, p_{exp}, p_{eq}, resulting in a total of 250 samples for quality evaluation. Two independent human

Dataset	#Semantic Precision	#Semantic Abstraction	#Semantic Equivalence
MSMARCO	2,287	2,287	1,323
NQ	1,997	1,997	1,357
FIQA	191	191	137

Table 1: Statistics of our SURE benchmark built upon MSMARCO, NQ and FIQA datasets. The data represents the number of queries in each testbed.

annotators, along with Qwen3-32B (Yang et al., 2025), assess whether each passage variant meets its intended requirement. For example, they assess whether summaries in the semantic-abstraction testbed can adequately answer the given queries, and whether expanded passages avoid introducing additional answer-relevant information. A passage variant is labeled as "qualified" if it satisfies the corresponding requirement. As shown in Table 2, 95% of the samples are consistently judged as high-quality by all evaluators.

Annotators	Qualified	Unqualified
H1	244	6
H2	246	4
M	241	9
H1 & H2	242	8
H1 & M	239	11
H2 & M	239	11
H1 & H2 & M	237	13

Table 2: Statistics of instances labeled as qualified or unqualified by individual annotators and their combinations. H1 and H2 denote two human annotators, and M denotes the Qwen3-32B model. For the last four rows, "qualified" indicates that all annotators agree, while "unqualified" indicates that at least one annotator assigns an unqualified label.

We additionally conduct feature-level statistics and distributional analyses on the MSMARCO test samples used in SURE to examine potential biases introduced by LLM-generated passages. We treat summaries as LLM-generated text, while original passages and their key sentences as real text, and compare them in terms of lexical diversity (TTR), lexical distribution (Shannon Entropy), and sentence length. As shown in Table 3, LLM-generated texts are similar to real texts in lexical diversity and distribution, but have more concentrated sentence-length patterns. These minor stylistic differences do not affect our evaluation, which focuses on semantic understanding.

Evaluated Models. We evaluate 10 prevalent dense retrieval models on the SURE benchmark, including four traditional dense retrievers, ANCE (Xiong et al., 2021), TCT-ColBERT-v2 (Lin et al., 2021), bge-base-en-v1.5 (Xiao et al., 2024), and e5-large-v2 (Wang et al., 2022), as well as six LLM-based embedding models, gte-Qwen2-1.5B-instruct (Li et al., 2023), Qwen3-Embedding-4B (Zhang et al., 2025), RepLLaMA (Ma et al., 2024), gte-Qwen2-7B-instruct (Li et al., 2023), e5-mistral-7b-instruct (Wang et al., 2024), and Qwen3-Embedding-8B (Zhang et al., 2025), spanning general-purpose and domain-specific tasks with model sizes from 110M to 8B parameters.

Source	TTR	Ent.	Avg.	Max
Original P	0.1194	11.5805	18.5	163
Key sent. P_{ks}	0.1807	11.2218	21.0	93
Summary P_{sum}	0.1748	11.6446	26.7	74

Table 3: Lexical and sentence-level statistics of real and LLM-generated texts. TTR (Type-Token Ratio) and Shannon Entropy (Ent.) capture lexical features, while average and maximum sentence length reflect sentence-level characteristics.

4.2 Results for Capability Evaluation

Figure 3 provides a detailed view of model performance on the three sub-capability evaluation testbeds, while Table 4 summarizes the overall performance of different models on the benchmark and their results on the three constituent datasets. The results reveal several findings:

When texts vary in information density, dense retrievers tend to capture shared semantics but struggle to discriminate fine-grained semantic differences. As shown in Figure 3, in the semantic abstraction testbed, models achieve the highest mean Avg-RDC score of 0.59 but the lowest mean Avg-ROC score of 0.71 among the three testbeds. This indicates that when texts all contain the core semantic information that satisfies the query intent but differ in information density, ranging from information-dense summaries where most content is query-relevant to information-sparse expansions where the core answer occupies a smaller portion of the text, models are generally able to capture the shared core information across these texts and identify their relevance to the query. However, they exhibit limited sensitivity to the semantic differences of texts introduced by variations in information density, which may limit their effectiveness in applications requiring fine-grained semantic discrimination such as passage ranking.

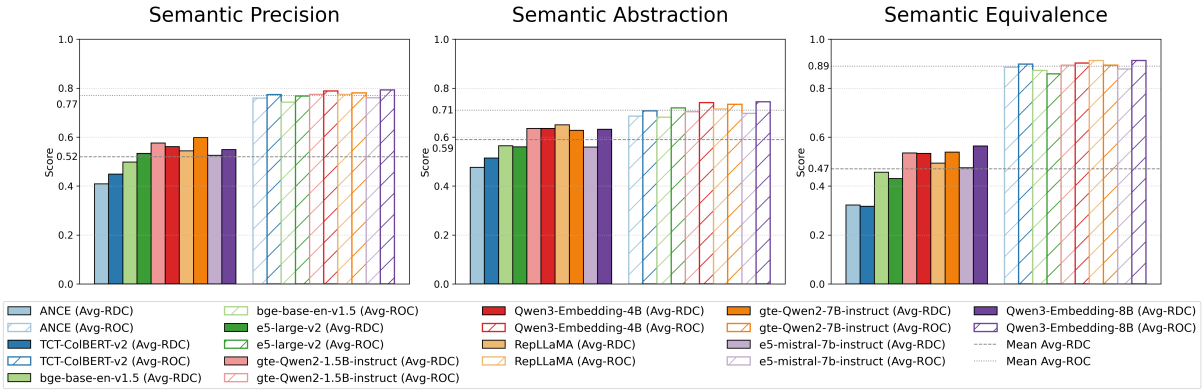

Figure 3: Performance of different models on three capability evaluations. The bar heights in each sub-figure represent the average RDC and ROC scores of the model across MSMARCO, NQ and FiQA datasets in each sub-capability evaluation testbed. The dashed line and the dotted line indicate the mean Avg-RDC score and the mean Avg-ROC score of all evaluated models.

Model	#Params	Training Data	MSMARCO		NQ		FiQA		AVG	
			RDC	ROC	RDC	ROC	RDC	ROC	RDC	ROC
ANCE	110M	MS+NQ	0.3975	0.8253	0.4573	0.7652	0.3531	0.7400	0.4026	0.7768
TCT-ColBERT-v2	110M	MS	0.4268	0.8405	0.5083	0.7742	0.3453	0.7643	0.4268	0.7930
bge-base-en-v1.5	110M	MS+NQ+others	0.4577	0.7964	0.5600	0.7641	0.5011	0.7360	0.5063	0.7655
e5-large-v2	335M	MS+NQ+others	0.4830	0.8344	0.5739	0.7683	0.4672	0.7419	0.5080	0.7815
gte-Qwen2-1.5B-instruct	1.5B	MS+NQ+others	0.5108	0.8317	0.6201	0.7766	0.6157	0.7640	0.5822	0.7907
Qwen3-Embedding-4B	4B	MS+NQ+others	0.4975	0.8368	**0.6407**	0.8000	0.5923	0.7955	0.5768	0.8108
RepLLaMA	7B	MS	**0.5217**	0.8385	0.6324	0.7902	0.5336	0.7741	0.5626	0.8009
gte-Qwen2-7B-instruct	7B	MS+NQ+others	0.5196	**0.8486**	0.6380	0.8014	0.6075	0.7588	**0.5884**	0.8029
e5-mistral-7b-instruct	7B	MS+others	0.4626	0.8239	0.5317	0.7637	0.5650	0.7483	0.5198	0.7786
Qwen3-Embedding-8B	8B	MS+NQ+others	0.4737	0.8410	0.6367	**0.8040**	**0.6341**	**0.8052**	0.5815	**0.8167**

Table 4: Overall performance of different models on the SURE benchmark. The results under MSMARCO, NQ and FIQA are the average scores over three capabilities: semantic precision, semantic abstraction, and semantic equivalence. RDC and ROC denote Rank Deviation Consistency and Rank Order Consistency respectively. "MS" in training data refers to MSMARCO, and "others" refers to additional training data without explicit mention of FiQA.

When texts vary in lexical expression, dense retrievers are sensitive to surface-level differences and struggle to recognize semantic equivalence. As illustrated in Figure 3, the evaluated models achieve the highest mean Avg-ROC score of 0.89 in the semantic equivalence testbed, while the corresponding mean Avg-RDC score is the lowest among the three testbeds at 0.47. This suggests that even when texts differ only through synonym-level substitutions and their overall meaning is largely preserved, dense retrieval models can still easily detect semantic differences introduced by lexical discrepancy. Meanwhile, their ability to judge semantic invariance is degraded under such lexical variation, suggesting that these models are more sensitive to lexical-level changes than to maintaining stable representations of semantics, and that dense retrieval models still partially depend on lexical matching signals.

Larger retrievers demonstrate stronger semantic understanding, but their capability is not strictly correlated with model size. Table 4 shows that larger retrievers ($\geq$ 1.5B) generally achieve higher overall RDC and ROC scores than smaller models, indicating stronger semantic understanding. However, models with 1.5B or 4B parameters attain performance comparable to or even surpassing 7B and 8B models, as exemplified by gte-Qwen2-1.5B-instruct, which outperforms all larger models except gte-Qwen2-7B-instruct in overall RDC score. These results suggest that semantic understanding in dense retrieval models does not scale monotonically with model size.

Training data diversity primarily enhances a retriever's semantic understanding on challenging tasks. We summarize the training and fine-tuning data used for these models from their original publications, as detailed in Table 4. On MSMARCO, models trained or fine-tuned on more diverse data exhibit only modest improvements in

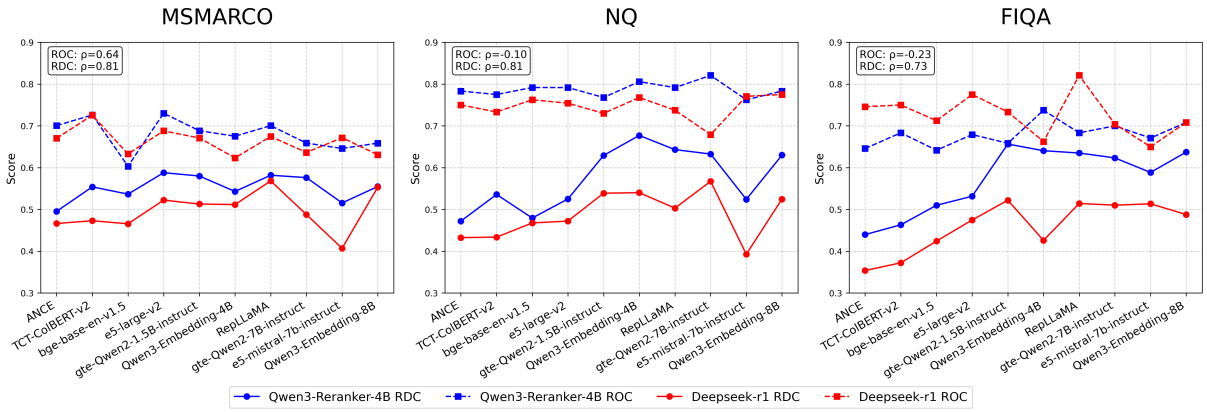

Figure 4: RDC and ROC results of dense retrieval models on MSMARCO, NQ, and FiQA, using both Qwen3-Reranker-4B and Deepseek-r1 to generate reference rankings. Experiments are conducted on 240 samples from the semantic abstraction testbed. Spearman's rank correlation coefficients (ρ) of RDC and ROC are also reported.

RDC over models trained with data of lower diversity. Notably, RepLLaMA, which is fine-tuned solely on MSMARCO, attains the highest RDC score. This suggests that for relatively straightforward, factoid-style queries such as those in MSMARCO, increasing training data diversity may offer limited gains in semantic understanding. On NQ and FiQA, where queries often involve deeper document-level reasoning, precise answer localization, or intent-driven and domain-specific relevance, models trained or fine-tuned on a mixture of data sources tend to achieve stronger RDC performance, whereas those trained exclusively on MSMARCO, such as RepLLaMA and TCT-ColBERT-v2, exhibit noticeable RDC performance drops on FiQA. This pattern implies that exposure to richer linguistic structures, broader domain knowledge, and more complex query–document relationships can help models generalize beyond surface-level matching and better capture semantic relationships, indicating that diverse training data can benefit models in handling more challenging or specialized tasks.

4.3 Further Analysis

Reranker Validation We validate the use of Qwen3-Reranker-4B for generating reference rankings of samples in the testbeds. Given the high computational cost of employing LLMs for reranking, we randomly sample 80 queries each from MSMARCO, NQ, and FiQA (240 queries in total) from the semantic abstraction testbed, and replace Qwen3-Reranker-4B with the stronger Deepseek-R1 to assess whether the evaluation of models' semantic understanding remains consistent. For each query, Deepseek-R1 reranks the top 2002 retrieved passages, with p_{sum}, p_{exp}, and p included, by applying the sliding window strategy (Sun et al., 2023) with a window size of 40 and a step size of 20. The resulting reference rankings are then used to compute models' RDC and ROC scores, which are compared with those obtained from Qwen3-Reranker-4B on the same 240 queries. Figure 4 shows RDC and ROC scores of models derived from both rerankers, with Spearman coefficient quantifying their consistency. RDC correlations exceed 0.7 across all datasets, and the ROC correlation is high on MSMARCO but lower on NQ and FIQA due to locally reversed rankings. Overall, using Qwen3-Reranker-4B yields model semantic understanding evaluations largely consistent with those obtained using Deepseek-R1.

Furthermore, reranking the top-2000 retrieval results of 10 models across three datasets using Qwen3-Reranker-4B improves MRR@2000, and the reference rankings of constructed passages provided by the reranker fall within a narrow range. These results further support the reliability of using Qwen3-Reranker-4B as the reference model. Detailed results are presented in Appendix B.

Capability Dimension Correlations To examine the relationship among the three semantic dimensions, we compute Spearman's rank correlations for semantic precision, abstraction, and equivalence under both RDC and ROC, as shown in Figure 5. Semantic precision exhibits consistently strong correlations with the other two dimensions, with coefficients above 0.76 in both RDC and ROC, indicating that recognizing key semantic information is crucial for supporting other aspects of

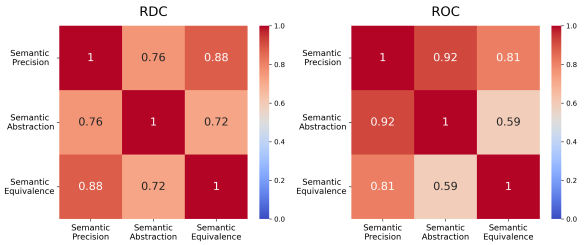

Figure 5: Spearman's rank correlation coefficients among semantic precision, semantic abstraction, and semantic equivalence under RDC and ROC metrics.

Metric Pair	MSMARCO	NQ	FiQA
RDC-nDCG@10	0.86	0.87	0.89
ROC-nDCG@10	0.54	0.63	0.64

Table 5: Spearman's rank correlation coefficients of RDC and ROC with nDCG@10 across MSMARCO, NQ, and FiQA datasets.

retrieval-oriented semantic understanding. In contrast, semantic abstraction and equivalence show lower correlation (e.g., 0.59 under ROC), suggesting that they represent relatively complementary semantic aspects.

Retrieval Performance vs. Semantic Understanding We examine whether retrieval performance measured by nDCG@10 aligns with semantic understanding measured by RDC and ROC. Models' nDCG@10 scores on MSMARCO, NQ, and FiQA are collected from the BEIR leaderboard, with missing results reproduced. Models are ranked based on the three metrics respectively, with Spearman correlations of model rankings computed. Table 5 indicates that nDCG@10 correlates more strongly with RDC than with ROC, suggesting retrieval performance primarily reflects a model's capacity for semantic consistency rather than fine-grained semantic discrimination.

5 Conclusion

In this work, we introduce SURE, a benchmark designed to investigate semantic understanding in dense retrieval models along three dimensions: semantic precision, semantic abstraction, and semantic equivalence. Experiments indicate that current dense retrievers exhibit limitations in semantic understanding, especially under variations in information density and expression, with model scale and training data influencing the capability to different degrees. SURE can serve as a valuable resource for investigating model capabilities and guiding the development of dense retrieval models toward more robust and nuanced semantic understanding.

Limitations

Our study has some limitations that can guide future work. Firstly, While SURE evaluates models along three relatively well-developed dimensions, it may not fully cover the entire spectrum of semantic understanding that could arise in real-world language use. Secondly, our experiments were conducted on three datasets spanning two domains (Web and Finance), but due to limited budget and the high cost of data construction, we were unable to extend our evaluation to additional datasets or domains, such as retrieval in the legal and medical fields, or to datasets in other languages, which might provide a more comprehensive assessment.

Ethical Considerations

This study complies with ethical standards in AI research. In particular, all data utilized for dataset construction are sourced from publicly available repositories, containing no personal or sensitive information, thereby ensuring data privacy and security. Moreover, the instructions employed for model-generated data are fully disclosed and carefully designed to prevent the generation of biased, unsafe, or otherwise harmful content. However, even after data verification and filtering, some automatically generated samples may still fail to meet the task requirements, which may slightly affect the quality of the benchmark.

Acknowledgment

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This work was supported by the Natural Science Foundation of China (No. 62536008, 62506354, 62572456, 62272439), the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20251041.

References

Negar Arabzadeh, Alexandra Vtyurina, Xinyi Yan, and Charles L. A. Clarke. 2022. Shallow pooling for sparse labels. *Inf. Retr. J.*, 25(4):365–385.

Xinran Chen, Xuanang Chen, Ben He, Tengfei Wen, and Le Sun. 2024. Analyze, generate and refine: Query expansion with llms for zero-shot open-domain QA. In *ACL (Findings)*, Findings of ACL, pages 11908–11922. Association for Computational Linguistics.

Xinran Chen, Ben He, Xuanang Chen, and Le Sun. 2025a. Not all terms matter: Recall-oriented adaptive learning for plm-aided query expansion in open-domain question answering. In *ACL (1)*, pages 22139–22151. Association for Computational Linguistics.

Xinran Chen, Yuchen Li, Hengyi Cai, Zhuoran Ma, Xuanang Chen, Haoyi Xiong, Shuaiqiang Wang, Ben He, Le Sun, and Dawei Yin. 2025b. Multi-agent proactive information seeking with adaptive LLM orchestration for non-factoid question answering. In *KDD (2)*, pages 4341–4352. ACM.

Xuanang Chen, Jian Luo, Ben He, Le Sun, and Yingfei Sun. 2022. Towards robust dense retrieval via local ranking alignment. In *IJCAI*, pages 1980–1986. ijcai.org.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Trans. Inf. Syst.*, 40(4):66:1–66:42.

Tim Hagen, Harrisen Scells, and Martin Potthast. 2024. Revisiting query variation robustness of transformer models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4283–4296.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *SIGIR*, pages 113–122. ACM.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781. Association for Computational Linguistics.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models. In *ICLR*. OpenReview.net.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *CoRR*, abs/2308.03281.

Zhuoqun Li, Xuanang Chen, Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2025. Paperregister: Boosting flexible-grained paper search via hierarchical register indexing. *CoRR*, abs/2508.11116.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *RepL4NLP@ACL-IJCNLP*, pages 163–173. Association for Computational Linguistics.

Xuan Lu, Sifan Liu, Bochao Yin, Yongqi Li, Xinghao Chen, Hui Su, Yaohui Jin, Wenjun Zeng, and Xiaoyu Shen. 2025. Multiconir: Towards multi-condition information retrieval. In *EMNLP (Findings)*, pages 13471–13494. Association for Computational Linguistics.

Kun Luo, Minghao Qin, Zheng Liu, Shitao Xiao, Jun Zhao, and Kang Liu. 2024. Large language models as foundations for next-gen dense retrieval: A comprehensive empirical assessment. In *EMNLP*, pages 1354–1365. Association for Computational Linguistics.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *SIGIR*, pages 2421–2425. ACM.

Joel Mackenzie, Matthias Petri, and Alistair Moffat. 2021. A sensitivity analysis of the MSMARCO passage collection. *CoRR*, abs/2112.03396.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *WWW (Companion Volume)*, pages 1941–1942. ACM.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2025. Generative representational instruction tuning. In *ICLR*. OpenReview.net.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: massive text embedding benchmark. In *EACL*, pages 2006–2029. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@NIPS*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Zhijie Nie, Zhangchi Feng, Mingxin Li, Cunwang Zhang, Yanzhao Zhang, Dingkun Long, and Richong Zhang. 2024. When text embedding meets large language model: A comprehensive survey. *CoRR*, abs/2412.09165.

Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the robustness of retrieval pipelines with query variation generators. In *ECIR (1)*, Lecture Notes in Computer Science, pages 397–412. Springer.

Roxana Petcu, Samarth Bhargav, Maarten de Rijke, and Evangelos Kanoulas. 2025. A comprehensive taxonomy of negation for NLP and neural retrievers. In *EMNLP (Findings)*, pages 15511–15533. Association for Computational Linguistics.

Juan Ramos and 1 others. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. New Jersey, USA.

Nicholas Riccardi, Xuan Yang, and Rutvik H Desai. 2024. The two word test as a semantic benchmark for large language models. *Scientific Reports*, 14(1):21593.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, and Luke Zettlemoyer. 2025. Reasonir: Training retrievers for reasoning tasks. *CoRR*, abs/2504.20595.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan Ö. Arik, Danqi Chen, and Tao Yu. 2025. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In *ICLR*. OpenReview.net.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. In *EMNLP*, pages 14918–14937. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *CoRR*, abs/2104.08663.

Lisa Miracchi Titus. 2024. Does chatgpt have semantic understanding? A problem with the statistics-of-occurrence strategy. *Cogn. Syst. Res.*, 83:101174.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *ACL (1)*, pages 11897–11916. Association for Computational Linguistics.

Haoyu Wu, Qingcheng Zeng, and Kaize Ding. 2025. Fact or facsimile? evaluating the factual robustness of modern retrievers. *CoRR*, abs/2508.20408.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *SIGIR*, pages 641–649. ACM.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *ICLR*. OpenReview.net.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Shuguang Yang, Feipeng Chen, Yiming Yang, and Zude Zhu. 2023. A study on semantic understanding of large language models from the perspective of ambiguity resolution. In *JCRAI*, pages 165–170. ACM.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *CoRR*, abs/2506.05176.

Zongmeng Zhang, Jinhua Zhu, Wengang Zhou, Xiang Qi, Peng Zhang, and Houqiang Li. 2024. Boolquestions: Does dense retrieval understand boolean logic in language? In *EMNLP (Findings)*, Findings of ACL, pages 2767–2779. Association for Computational Linguistics.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Trans. Inf. Syst.*, 42(4):89:1–89:60.

A Prompts for Constructing SURE Benchmark

In benchmark data construction, we employ DeepSeek-R1 to both generate and validate testbed samples. For data generation, we use the following prompts to extract query-intent-relevant key sentences from passages and to generate non-answering contents, summaries, expansions, and paraphrased versions:

I_{ks}

Key sentences extraction:
You are a professional and instruction-following text processing model.
query: {query}
passage: {passage}
Please extract the key sentence or sentences in the passage that can answer the query directly. Output the extracted sentences starting with 'key sentences:'.

I_{na}

Non-answering contents generation:
You are a professional and instruction-following text generation model.
query: {query}
Please generate two sentences that can not answer the query. The generated sentences can be topic relevant or topic irrelevant to the query. Output the sentences starting with 'sentences:'.

I_{sum}

Summary generation:
You are a professional and instruction-following text processing model.
query: {query}
passage: {passage}
Please summarize the passage and make sure the summary can answer the query. Output the summary starting with 'summary:'.

I_{exp}

Expansion generation:
You are a professional and instruction-following text generation model.
passage: {passage}
Please expand on the given passage by adding two sentences, ensuring that the added content revolves around the original theme but can not answer the query directly. Output the expanded passage starting with 'expansion:'.

I_{eq}

Paraphrase generation:
You are a professional and instruction-following text processing model.
query: {query}
passage: {passage}
Please first point out the keywords in the query. Then replace all query keywords in the passage with their definition, description or synonyms. Output in the form 'keywords in the query:' and 'paraphrased passage:'.

For the verification of extracted key sentences, generated summaries, and paraphrased passage variants, we adopt the following instructions:

Key sentences verification:
You are a professional and instruction-following text evaluator.
query: {query}
sentences: {key sentences}
Please determine whether the given sentences can answer the given query. Please output 'yes' or 'no' only.

Summary verification:
You are a professional and instruction-following text evaluator.
query: {query}
summary: {summary}
Please determine whether the given summary can answer the given query. Please output 'yes' or 'no' only.

Paraphrase verification:
You are a professional and instruction-

following text evaluator.
query: {query}
passage: {paraphrased passage}
Please determine whether the passage contains the EXACT keywords in the query. Please output 'yes' or 'no' only.

We additionally design a prompt to handle pronouns in extracted key sentences:

Pronouns substitution in key sentences:
You are a professional and instruction-following text processing model.
sentences: {key sentences}
passage: {passage}
Please replace all pronouns in the given sentences with their corresponding referents based on the provided passage. Output the new sentences starting with 'new sentences:'.

B Analysis of Qwen3-Reranker-4B

Qwen3-Reranker-4B demonstrates sota semantic understanding on MTEB-R benchmarks. We further validate Qwen3-Reranker-4B as a reference ranker by evaluating both its reranking effectiveness and the reliability of the reference rankings it generates.

B.1 Reranking Effectiveness

We evaluate the ten models on the MSMARCO, NQ, and FiQA datasets included in the SURE benchmark by retrieving top-2000 passages, and rerank the passage lists using Qwen3-Reranker-4B. The MRR@2000 before and after reranking is presented in Table 6. The results show that reranking with Qwen3-Reranker-4B substantially improves MRR@2000 for all models across the three datasets, indicating that Qwen3-Reranker-4B possesses stronger semantic understanding and can serve as the reference model.

B.2 Reference Rankings reliability

For each evaluation testbed, we construct passage variants that preserve the core semantics while introducing subtle differences. In principle, these variants should remain semantically close, and are expected to receive similar rankings in retrieval results. We analyze the ranking distribution of the constructed variants after reranking with Qwen3-Reranker-4B and quantify ranking compactness using average ranking intervals, in order to assess whether Qwen3-Reranker-4B places these documents close to each other and whether the resulting rankings are suitable as evaluation references.

As shown in Table 7, across all testbeds on the NQ and FiQA datasets, the average ranking intervals of the constructed variants consistently fall between 7 and 43, indicating that Qwen3-Reranker-4B effectively clusters semantically equivalent documents into proximate positions. However for the MSMARCO dataset, the sparsity of relevance annotations leaves many relevant documents unlabeled (Mackenzie et al., 2021; Arabzadeh et al., 2022), which amplifies the impact of semantic variations on ranking and leads to slightly larger intervals.

Overall, the results show that Qwen3-Reranker-4B reliably captures semantic consistency, thus its rankings of the constructed passages can be used as reference rankings.

C Case Study

For **overall retrieval**, models should retrieve passages that contain answers to the query and rank them higher. However, models sometimes prefer passages which are merely topically or lexically relevant to the query but contain no answers.

Overall Retrieval

Query: another name for reaper
Gold passage: This thesaurus page is about all possible synonyms, equivalent, same meaning and similar words for the term reaper. harvester, reaper (noun)
Top retrieved passage: The name Reaper is ranked on the 24,945th position of the most used names. It means that this name is rarely used. We estimate that there are at least 8100 persons in the world having this name which is around 0.001% of the population. The name Reaper has six characters.

For **semantic precision testbed**, models should identify key information in passages that supports answering the query. However, we observe that models sometimes fail to locate such useful information. For the example shown below, the gold rankings of p_{ks}, p, and p_{ksn} given by the reranker model are 1, 2, 11, while for Qwen3-Embedding-8B, the rankings of the samples are 1, 2, 50.

Semantic Precision

Query: alabama central credit union routing number

p_{ks}**:** Alabama Central Credit Union Routing Number 262087502.

p**:** Alabama Central Credit Union Website Home Page. Alabama Central Credit Union Phone Number: 205-510-1300. Alabama Central Credit Union Routing Number 262087502. Routing number is a 9-digit number generally found at the bottom-left corner of the paper check. It is also used to identify a bank uniquely.

p_{ksn}**:** The weather in Alabama can be quite humid, especially during the summer months. Many credit unions offer a variety of financial services to their members, including savings accounts and loans. Alabama Central Credit Union Routing Number 262087502.

For **semantic abstraction testbed**, We find that models can be influenced by information density of passages when capturing the same useful information in passages. For the example shown below, the reference rankings of p_{sum}, p_{exp}, and p are 2, 1, 20, while for Qwen3-Embedding-8B, the rankings of the samples are 1, 14, 204.

Semantic Abstraction

Query: average of rn

p_{sum}**:** The average salary of a registered nurse (RN) in the U.S. is 69,790 dollars annually, with an average hourly wage of 33.55 dollars, varying by location, experience, and other factors.

p**:** According to the Bureau of Labor Statistics latest data, the average salary of a registered nurse in the United States is 69,790 dollars. The average hourly wage of a registered nurse is 33.55 dollars. Keep in mind that these are only averages, and that a registered nurse may make more or less than these amounts based on location, experience, and other factors. In addition, the starting salaries for an RN may be much less than the average salary.

p_{exp}**:** According to the Bureau of Labor Statistics latest data, the average salary of a registered nurse in the United States is 69,790 dollars. The average hourly wage of a registered nurse is 33.55 dollars. Keep in mind that these are only averages, and that a registered nurse may make more or less than these amounts based on location, experience, and other factors. Specialized roles, such as nurse practitioners or those in critical care units, often command higher salaries due to advanced skills and certifications. Additionally, benefits like overtime pay, shift differentials, and healthcare packages can further influence overall compensation beyond base wages. In addition, the starting salaries for an RN may be much less than the average salary.

For **semantic equivalence testbed**, we observe that when the query keywords in the passage are replaced with their synonyms, description, or knowledge-based expressions, models struggle to match the paraphrased passage with the query and recognize the equivalence between the passage and its paraphrase. For the example below, when replacing "Calgary" with "Alberta's largest urban center", models still rank p high but rank p_{eq} much lower.

Semantic Equivalence

Query: calgary population

p**:** When the 2016 census was taken last May 10, the population of the census metropolitan area of Calgary was 1,392,609, compared with 1,214,839 from the 2011 census. (Jeff McIntosh/Canadian Press) Calgary had the highest growth rate of any metropolitan area in Canada over the past five years, despite being in the throes of an economic downturn, according to the latest census data.

p_{eq}**:** When the 2016 census was taken last May 10, the demographic count of the census metropolitan area encompassing Alberta's largest urban center was 1,392,609, compared with 1,214,839 from the 2011 census. (Jeff McIntosh/Canadian Press) This prairie-region municipality exhibited the most rapid expansion pace among Canadian urban corridors during the preceding half-decade, despite being in the throes of an economic downturn, according to the latest census data.

Model	MSMARCO		NQ		FIQA	
	Original	**After Rerank**	**Original**	**After Rerank**	**Original**	**After Rerank**
ANCE	0.388	0.484	0.412	0.680	0.481	0.843
TCT-ColBERT-v2	0.435	0.482	0.483	0.676	0.520	0.841
bge-base-en-v1.5	0.428	0.482	0.515	0.675	0.657	0.843
e5-large-v2	0.452	0.482	0.574	0.674	0.648	0.843
gte-Qwen2-1.5B-instruct	0.445	0.482	0.587	0.674	0.783	0.843
Qwen3-Embedding-4B	0.443	0.482	0.597	0.674	0.803	0.843
RepLLaMA	0.475	0.482	0.611	0.675	0.715	0.843
gte-Qwen2-7B-instruct	0.474	0.481	0.620	0.674	0.804	0.843
e5-mistral-7b-instruct	0.449	0.484	0.514	0.676	0.761	0.843
Qwen3-Embedding-8B	0.449	0.485	0.610	0.674	0.800	0.843

Table 6: Mean Reciprocal Rank (MRR)@2000 of ten evaluated dense retrieval models on MSMARCO, NQ, and FiQA datasets, before and after reranking with Qwen3-Reranker-4B.

Model	MSMARCO			NQ			FIQA		
	Precision	**Abstraction**	**Equivalence**	**Precision**	**Abstraction**	**Equivalence**	**Precision**	**Abstraction**	**Equivalence**
ANCE	52.66	18.45	200.59	8.22	7.66	26.16	20.35	8.17	21.89
TCT-ColBERT-v2	57.57	19.25	228.02	9.26	8.78	34.34	20.96	8.15	25.20
bge-base-en-v1.5	60.82	19.70	241.12	9.66	9.31	37.35	21.29	8.59	26.85
e5-large-v2	61.84	19.83	242.99	10.05	9.75	41.17	20.39	8.29	23.41
gte-Qwen2-1.5B-instruct	63.88	20.32	256.22	10.03	9.87	42.11	22.68	8.61	24.69
Qwen3-Embedding-4B	65.27	20.45	265.63	10.05	9.77	41.20	21.86	8.68	26.72
RepLLaMA	63.56	20.22	253.31	9.53	9.27	34.69	21.62	8.60	25.92
gte-Qwen2-7B-instruct	64.92	20.29	262.39	10.27	10.15	41.95	22.47	8.70	25.47
e5-mistral-7b-instruct	63.73	20.08	257.96	9.75	9.50	39.65	21.73	8.69	25.10
Qwen3-Embedding-8B	65.23	20.48	266.47	10.20	10.08	43.05	22.04	8.71	25.08

Table 7: Average ranking ranges of constructed passages after reranking by Qwen3-Reranker-4B across different testbeds. Precision, Abstraction, and Equivalence refer to semantic precision, semantic abstraction, and semantic equivalence, respectively. The reranking is performed on the top-2000 retrieved passages combined with the constructed passages for each model.