

IdioLink: Retrieving Meaning Beyond Words Across Idiomatic and Literal Expressions

Kai Golan Hashiloni^{1,2}, Daniel Fadlon^{1,2}, Lior Livyatan^{1,2}, Ofri Hefetz^{1,2},
Jiahuan Pei³, Kfir Bar^{1,2}

¹Data Science Institute, Reichman University, Herzliya, Israel

²Efi Arazi School of Computer Science, Reichman University, Herzliya, Israel

³Vrije Universiteit Amsterdam, The Netherlands

kai.golanhashiloni@post.runi.ac.il, {daniel.fadlon, lior.livyatan, ofri.hefetz}@post.runi.ac.il,
j.pei2@vu.nl, kfir.bar@runi.ac.il

Abstract

Idioms pose a fundamental challenge for language models, as their meaning cannot be inferred from surface form alone. Understanding such expressions, therefore, requires semantic abstraction beyond lexical overlap. We introduce **IdioLink**, a retrieval benchmark designed to test whether models can link idiomatic expressions to conceptually equivalent meanings expressed in literal or paraphrased forms. IdioLink comprises 10,700 documents and 2,140 queries, spanning 107 idioms with both literal and figurative uses. Each document and query is annotated with spans that convey the core meaning. Evaluating strong embedding baselines (e.g., BGE, E5, Contriever, and Qwen), we show that current models struggle to retrieve equivalent meanings across divergent surface realizations, relying instead on topical and shallow semantic cues. IdioLink exposes key gaps in idiom-aware semantic retrieval and provides a challenging testbed for future models.

1 Introduction

Understanding language goes beyond mapping words to meanings: it involves constructing conceptual interpretations that often depart from surface form. Figurative language, including idioms, metaphors, and other conventionalized expressions, illustrates this challenge. Potential idiomatic expressions (PIEs) like “break the ice” or “spill the beans” convey meanings that cannot be compositionally derived from their constituent words, yet are effortlessly understood by humans based on the context they appear in, which determines whether the expression is intended figuratively or literally. The degree to which an idiom’s parts contribute to its figurative meaning—how compositional or non-compositional it is—varies across expressions and shapes how flexibly they can be used and understood by readers (Nunberg et al., 1994; Gibbs et al., 1989a,b). Decades of work in linguistics and cognitive science show these expressions are central

Literal	The ship must break the ice to go north.
Idiomatic	I’ve got a fun game to break the ice .
Simplification	I’ve got a fun game to help everyone loosen up .
Sense	If you spill your drink, it will ease the tension for sure.

Figure 1: The four different document types for a given potential idiomatic expression in IdioLink.

to human communication and draw on rich conceptual, cultural, and contextual knowledge (Pollio et al., 1977; Weinreich, 1969; Jackendoff, 1997; Gibbs and Colston, 2012), making them a crucial benchmark for evaluating semantic understanding beyond words.

Recent advances in large language models (LLMs) have substantially improved performance on modeling figurative language, including idiom identification (Tedeschi et al., 2022; Hashiloni et al., 2025) and figurative-literal expression disambiguation (Arslan et al., 2025; Mi et al., 2025). However, it remains unclear whether LLMs can reliably align expressions with the same meanings across figurative, literal, or paraphrased forms. Embedding and contrastive models often rely on lexical overlap rather than deeper semantic equivalence (Wang et al., 2022; Miao et al., 2023; Shi et al., 2023), causing retrieval failures when query meaning is expressed figuratively or paraphrased, e.g., “break the ice” vs. “ease the tension”. This gap is especially problematic in real-world scenarios where users encounter unfamiliar idioms and seek semantically equivalent expressions to understand their usage.

In this work, we introduce **IdioLink**, a benchmark designed to evaluate concept-level retrieval across idiomatic, literal, simplification, and sense document types, as shown in Figure 1. Each instance is annotated with a marked span representing the core semantic unit of a PIE and paired

with documents containing a span expressing the same meaning from different types.

We formulate a retrieval task in which, given a query containing a marked span with a PIE, used either figuratively or literally, the model retrieves all documents whose marked spans convey the same meaning, without access to the documents’ span annotations. To the best of our knowledge, this is the first setup that enables evaluation of whether embedding models retrieve documents based on shared conceptual meaning, whether expressed figuratively or literally, rather than superficial lexical or topical similarity.

We evaluate 24 embedding models and their variants, comparing zero-shot and fine-tuned settings, encoding queries w/ or w/o instructions, and employing standard *sentence* or *span* embeddings methods. Our results reveal that even state-of-the-art embedding models struggle to consistently retrieve semantically equivalent expressions when form diverges from meaning. These findings highlight a persistent gap between surface-level semantic matching and genuine conceptual understanding.

In summary, our contributions are as follows:

- We introduce a new retrieval task centered on idiomatic expressions, enabling controlled analysis of conceptual similarity beyond surface form.
- We release IdioLink, a new benchmark for evaluating concept-level retrieval across figurative, literal, and paraphrased expressions.
- We provide an extensive evaluation of modern embedding models, revealing systematic limitations in their ability to represent and retrieve shared meaning.

The IdioLink dataset, prompts, and evaluation code are publicly released to ensure transparency and reproducibility under the CC-BY-4.0 (dataset) and Apache-2.0 (code) licenses: <https://github.com/Intellexus-DSI/IdioLink>.

2 Related Work

2.1 Idioms as a Core Challenge in NLP

Multi-word expressions (MWEs) pose a longstanding challenge in NLP due to their syntactic variability and non-compositional semantics (Constant et al., 2017). Earlier approaches relied on rule-based, statistical, and distributional methods (Cook et al., 2007; Fazly et al., 2009; ?), while recent work predominantly uses Transformer-based architectures for MWE identification and interpretation

(Taslimipoor et al., 2020; Zeng and Bhat, 2021; Savary et al., 2023).

Idioms form a particularly challenging subclass of MWEs because their meanings cannot be compositionally inferred from their constituent words (Baldwin and Kim, 2010). In NLP, idiom *identification* focuses on detecting idiomatic spans in context, whereas idiom *classification* assumes candidate expressions are pre-specified. A wide range of datasets has been proposed for these tasks, including VNC-Tokens (Cook et al., 2008), IDIX (Sporleder et al., 2010), SemEval benchmarks (Korontzelos et al., 2013; Tayyar Madabushi et al., 2022), and larger modern resources such as MAG-PIE (Haagsma et al., 2020) and EPIe (Saxena and Paul, 2020).

While our work focuses on English, several datasets extend idiom and MWE evaluation to multilingual settings. PARSEME (Ramisch et al., 2020; Savary et al., 2023) and AlphaMWE (Han et al., 2020) provide broad multilingual MWE coverage beyond idioms, with PARSEME 2.0 (Scholivet et al., 2026) additionally supporting paraphrasing tasks. Other multilingual resources include ID10M (Tedeschi et al., 2022), Dodiom (Eryigit et al., 2022), MIDAS (Kim et al., 2025), G-IdiomAlign (Ye et al., 2026), and MIDI (Almheiri et al., 2026), while Hefetz et al. (2025) study cross-lingual fine-tuning transfer for idiom identification. Recent work further explores figurative-language understanding, idiom disambiguation, and cross-lingual semantic retrieval (Park et al., 2025; Mi et al., 2025; Aslantaş and Gungor, 2026).

With a similar motivation to ours, a separate line of work probes whether representation models capture idiomaticity directly. Garcia et al. (2021) and He et al. (2025) test contextual word representations (e.g., ELMo, BERT) on noun-compound minimal pairs using similarity judgments, while Wold et al. (2026) evaluates text embedding models using a continuous compositionality measure. Unlike IdioLink, none of these frame the task as retrieval over a corpus of meaning-equivalent realizations.

None of these resources provides parallel, meaning-equivalent realizations of the same expression at sufficient scale or with sufficient control, which is why we build IdioLink rather than adopting an existing resource.

Neural approaches have been applied to both idiom classification and identification (Briskilal and Subalalitha, 2022; He et al., 2024). More recently, LLMs have achieved strong performance

through prompt-based inference, often matching or surpassing fine-tuned systems (Hashiloni et al., 2025; De Luca Fornaciari et al., 2024; Mi et al., 2025; Phelps et al., 2024), though Hashiloni et al. (2026) show this performance is not bulletproof under adversarial contexts. LLMs have also been used to support the construction of idiom datasets and corpora (Arslan et al., 2025).

Another related direction is idiomatic sentence paraphrasing (ISP), which focuses on transforming idiomatic expressions into literal paraphrases or vice versa. Prior work has explored unsupervised, weakly supervised, and controllable paraphrasing settings, while consistently highlighting data-scarcity challenges (Zhou et al., 2021a; Qiang et al., 2023; Zhou et al., 2021b).

No existing work targets the idiom-centered retrieval task we address here.

2.2 Text Retrieval and Embedding Models

Text retrieval has undergone rapid evolution with the advent of neural embedding models, which map queries and documents into a shared semantic space. Early dense retrieval approaches such as Sentence-BERT (Reimers and Gurevych, 2019) and subsequent contrastive learning frameworks (Gao et al., 2021; Izacard et al., 2022) demonstrated strong performance by learning semantically meaningful sentence representations. More recent models scale this paradigm through larger architectures, multilingual training, and instruction-based supervision (Wang et al., 2024a,c; Su et al., 2023; Muennighoff et al., 2025).

A growing body of work explores instruction-aware retrieval, where natural language instructions guide embeddings toward task-specific representations. Models such as InstructOR (Su et al., 2023) and FollowIR (Weller et al., 2025) show that instructions can substantially improve retrieval, motivating our use of instruction-augmented inputs to steer models toward idiomatic or literal interpretations without changing model parameters.

Another relevant direction concerns the granularity of representation and context modeling. Standard retrieval pipelines typically embed entire documents or passages, which may obscure fine-grained semantic distinctions. To address this, prior work has explored chunking strategies based on fixed windows, sentence boundaries, or semantic similarity (Lewis et al., 2021; Kamradt, 2024). More recently, methods such as late interaction (Santhanam et al., 2022; Jha et al., 2024) and late

Chunking (Günther et al., 2025) aim to preserve contextual information while enabling localized matching. These approaches prove that retaining broader context during representation learning can substantially improve retrieval quality, especially when meaning depends on subtle contextual cues.

Our work builds on these insights while addressing a distinct challenge: retrieving semantically equivalent documents based on underlying conceptual meaning. Unlike prior work focused on topical relevance or general similarity, IdioLink tests whether models can distinguish literal from idiomatic meanings despite similar surface forms, situating the benchmark at the intersection of information retrieval and figurative language understanding.

3 The IdioLink Dataset

3.1 Problem Formulation

We formulate IdioLink as a retrieval task, consisting of a non-overlapping index $\mathcal{I}$ and queries $\mathcal{Q}$. Both $\mathcal{I}$ and $\mathcal{Q}$ contain similarly structured texts, each associated with a target PIE p and containing a marked span that conveys the meaning of p , either idiomatically or literally. We define four types of documents (examples are in Figure 1): (i) *literal*, where p is used in its compositional meaning; (ii) *idiomatic*, where p appears in its figurative meaning; (iii) *simplification* is a copy of idiomatic, but replaces p with an explicit explanation of its meaning that fits the context, and; (iv) *sense* expresses the same idea as the idiomatic usage of p , similarly to simplification but as a standalone, independent document.

We design literal and idiomatic documents to test whether models can distinguish literal and idiomatic usages PIEs. The simplification documents evaluate whether models capture the equivalence between an idiom and its literal paraphrase, while sense documents probe retrieval based on shared abstract meaning beyond lexical overlap.

Queries are generated exclusively as either literal or idiomatic document types. Given a query $q \in \{\text{literal}, \text{idiomatic}\}$, which contains a PIE p , the goal is to retrieve all documents in $\mathcal{I}$ containing spans expressing the same underlying meaning of p based on its usage in q . If q is literal, the relevant documents are other literal instances of the same p . If q is idiomatic, we expect all three other types (a.k.a., idiomatic types) to be retrieved.

3.2 Dataset Construction

We construct $\mathcal{I}$ and $\mathcal{Q}$ synthetically rather than from naturally occurring text, since no existing corpus offers real-world parallel, meaning-equivalent realizations of the same PIE we seek. Moreover, our settings offer a very controlled construction: document types and their quantities, document style and domains. This is central to isolating the representational capability we target.

All models used to construct the dataset are selected based on preliminary exploratory experiments and are set to the default temperature value of 1. All prompts are first optimized to reduce disagreement with human annotations during a calibration phase. The prompts are released in the project repository¹. The dataset construction process, including prompt design and optimization, required approximately 6,500 API calls and incurred a total cost of approximately \$60.

3.2.1 PIE Collection

We collect our set of PIEs from the MAGPIE corpus (Haagsma et al., 2020). To ensure both linguistic validity and contextual diversity, we filter PIEs using three criteria. First, the PIE must appear at least 30 times in MAGPIE to ensure sufficient contextual variability and reflect more common usage in the language. Second, we require a balanced distribution between literal and idiomatic usages, constraining the proportion of idiomatic instances to lie between 25% and 75%. This ensures that each expression meaningfully supports both usage types and remains genuinely ambiguous and context-dependent. Third, we retain only PIEs annotated with maximal annotator confidence (1.0), ensuring high annotation reliability. This yields 118 candidate PIEs. During manual inspection, we also removed 11 PIEs whose literal usage was judged highly unnatural or artificially constructed (e.g., “for keeps”), resulting in 107 PIEs in the final dataset. To simulate a realistic low-resource fine-tuning setting and to test whether contrastive fine-tuning improves alignment, rather than to maximize training scale, we design a relatively small training split. The much larger index, by contrast, simulates retrieval over a broad corpus. Specifically, we partition the 107 PIEs into training (22), validation (10), and test (75) splits.

¹Given their length, we omit them from the paper.

3.2.2 Data Generation and Indexing

To promote semantic diversity and reduce topical bias, we compile a set of ten domains: *Politics*, *Sport*, *Technology*, *History*, *Medicine*, *Culture*, *Entertainment*, *Food*, *Business*, and *Environment*.

To construct the index $\mathcal{I}$, for every pair of a PIE and a domain (p, d) , we generate ten documents: four literal documents and two documents for each of the idiomatic types. As a result, each p is associated with 100 documents in total—40 literal and 60 belonging to the idiomatic types. The query set, $\mathcal{Q}$, contains one literal and one idiomatic document for each (p, d) pair. Specifically, we use Gemini 2.5 Pro to generate $\mathcal{I}$ and Gemini 3 Flash to generate $\mathcal{Q}$. This separation encourages novel query generation and reduces lexical and semantic overlap between queries and documents. Each query and document consists of one or two sentences. Both $\mathcal{I}$ and $\mathcal{Q}$ are constructed independently for each data split, ensuring no PIEs are shared across them.

3.2.3 Human Annotation

To establish a gold-standard evaluation set, we manually annotate 13 randomly sampled PIEs from the test split, denoted $\text{test}_{\text{gold}}$, which serves as the test set in all our experiments. This set is denoted as $\text{test}_{\text{gold}}$ (1,300 samples), and the remaining portion of the test set, 62 PIEs, as $\text{test}_{\text{silver}}$ (6,200 samples). The index and queries of $\text{test}_{\text{gold}}$ are manually validated and corrected when needed. As described in § 3.2.4, the $\text{test}_{\text{gold}}$ subset serves as a calibration set to validate our automatic validation pipeline against human judgment; once validated, the pipeline is then used to assess the quality of the remaining silver-standard data at scale.

Three annotators (demographics in Appendix A) are provided with the documents and queries via a shared Google Sheet and instructed to follow the annotation guidelines, summarized below. The full guidelines are released in the project’s repository.

Each document is accompanied by its corresponding (p, d) pair, its type, and the marked span, and is considered valid if all of the following conditions hold: (i) It is fluent, coherent, and fully interpretable to a human reader. (ii) It contains no idiomatic expressions other than possibly p , which must appear at most once, and its usage, idiomatic or literal, is unambiguous. (iii) The content of the document relates to d . (iv) The marked span correctly and minimally captures the conceptual realization of p and appears in the document.

Annotators may mark a document as valid, correct the span, and/or revise the text. Each query and document receives a single annotation, with ambiguous or complex cases resolved through discussion among the annotators.

Only 125 corrections or revisions were required, with most cases (66) requiring only a span correction. There were also literal–idiomatic mismatches (21), ambiguous realizations (16), and sentence-level rewrites for improved clarity or validity (9). Most corrections occurred in sense documents (67) and literal documents (47).

3.2.4 Automatic and Human Validation

To scale validation while maintaining quality, we combine automatic validation with human review.

We prompt GPT-4o mini with instructions based on the annotation guidelines (§ 3.2.3). The model is provided with a set of documents or queries, each represented as (p, d) , and is tasked with determining whether each of them is valid or invalid. We apply self-consistency (Wang et al., 2023), running each prompt three times and assigning the final label via majority vote.

To assess the performance of the automatic validation model, we compute agreement between human annotators and the automatic validation model using the index $\mathcal{I}$ of test_{gold} . It achieves an accuracy of 72.08% and a macro-averaged F1 score of 82.54%, where both metrics are calculated for the binary valid/not-valid classification task. We observe only 42 cases in which the model labeled a document as valid, even though the annotator found it invalid, which may introduce a flaw in the dataset. In most of those cases, the usage of the PIE was ambiguous. This implies that the model may miss only 3.23% of the invalid cases, which are more important to us.

When applied to test_{silver} texts (11,280 documents and queries), this process flags 2,010 documents (17.81%) as invalid. Our annotators then reviewed the flagged 2,010 texts and found that only 316 (15.72%) were confirmed as problematic and subsequently corrected or replaced. Overall, we obtain 3,570 gold-standard documents and queries, accounting for approximately 27.8% of the full IdioLink dataset. An overview of the dataset is provided in Table 1.

4 IdioLink Retrieval Benchmark Models

We evaluate models under zero-shot and fine-tuning settings, provided with and without task-

Split	# PIE	# Query	# Document
Train	22	440	2,200
Validation	10	200	1,000
Test_{silver}	62	1,240	6,200
Test_{gold}	13	260	1,300
Test_{total}	75	1,500	7,500
Total	107	2,140	10,700

Table 1: Overview of the IdioLink Dataset; test_{total} is the sum of test_{gold} and test_{silver} .

specific instruction and explore a span-aware embedding-extraction method to control how q is encoded and represented.

4.1 Retrieval Configurations

Using instructions in query construction. We consider two approaches to submitting queries to the model: a standard approach without additional instructions and one that provides explicit instructions to guide the model. The instruction specifies the marked span of the PIE, and encourages the model to infer whether the PIE is used literally or figuratively and to retrieve documents that entail conceptually equivalent meanings. Implementation details are described in Appendix B.

Using span embedding. We evaluate two methods for obtaining query representations. The first is the *sentence* embedding, which uses the model’s default text-embedding output. The second is *span* embedding, inspired by the recent work on late chunking (Günther et al., 2025) and prior findings suggesting that span-level representations can better capture fine-grained semantic alignment than sentence-level similarity alone (Kanerva et al., 2025). While standard dense retrieval typically embeds entire documents as single vectors, such representations may obscure fine-grained semantic information—particularly when the retrieval target is a localized meaning expressed as a specific span. In contrast, late chunking first extracts token-level contextual embeddings as the model’s last-layer outputs, then aggregates the embeddings of tokens corresponding to the annotated span using mean pooling. This yields a span-level representation that remains context-aware while focusing on the segment conveying the intended meaning. Note that span embedding narrows only the aggregation strategy, not the contextual encoding, as the span tokens are encoded within the full sentence context before pooling. An illustration is shown in Figure 2.

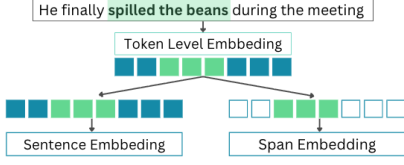


Figure 2: Sentence and span embedding. The latter employs mean-pooling over the embeddings of the span tokens only.

Index and similarity calculation. At retrieval time, queries are encoded w/ or w/o instructions and represented using sentence or span embeddings, while indexed documents are encoded without instructions as full text, reflecting a realistic retrieval setting without access to annotated spans.

4.2 Zero-shot Setting

We evaluate all models in a zero-shot setting, with and without instructions, and with sentence and span embeddings, yielding four zero-shot configurations per model. Importantly, these methods require no additional training and can be used uniformly across all architectures.

4.3 Fine-tuning Setting

To improve retrieval of conceptual meaning beyond surface form, we fine-tune embedding models using a contrastive objective inspired by recent dense retrieval frameworks (Lee et al., 2024; Karpukhin et al., 2020). Our goal is to encourage alignment between semantically equivalent expressions, across literal and idiomatic realizations, while separating semantically mismatched ones. Hyperparameters are reported in Table 4 (Appendix F).

Training data. We construct a fine-tuning dataset $\mathcal{D}$ consisting of structured tuples:

$$\mathcal{D} = \{(q, p, d^+, \{d_i^-\}_{i=1}^{\alpha_N}, \{h_j^-\}_{j=1}^{\alpha_H})\}, \quad (1)$$

where each tuple contains a query q , its associated PIE p , one positive document d^+ , α_N “soft” negative documents $\{d_i^-\}_{i=1}^{\alpha_N}$, and α_H “hard” negative documents $\{h_j^-\}_{j=1}^{\alpha_H}$.

Each q produces a single training tuple, so $\mathcal{D}$ contains 440 training samples, as elaborated in §3.2.1. A positive document d^+ contains a span that conveys the same underlying meaning as p in q . If q is of type *literal*, the corresponding d^+ must also be *literal*, that is, another document in which p is used literally. If q is *idiomatic*, the corresponding d^+ is drawn from the corresponding idiomatic types, that is, a document with a span that conveys

the figurative meaning of p . A “hard” negative h^- is a document of the same p , but representing the opposing interpretation: For *literal* queries, a “hard” negative is a document drawn from the idiomatic types, whereas for idiomatic queries a hard negative is a *literal* document. We define “soft” negatives d^- as documents of any type sampled from the documents generated for other PIEs. This construction ensures that each query q_i ranks its corresponding positive example d_i^+ above both hard and soft negatives. While hard negatives target the literal/idiomatic contrast for the same PIE, soft negatives push the model to distinguish conceptual relevance from mere topical or domain similarity, discouraging it from conflating “same domain” with “same meaning”. This design choice closely follows the loss formulation of Gecko (Lee et al., 2024).

Training objective. We optimize the model for the following contrastive learning objective:

$$\mathcal{L} = -\log \frac{e^{s(q_i, d_i^+)}}{\sum_{j=1}^{\alpha_N} e^{s(q_i, d_{i,j}^-)} + \sum_{j=1}^{\alpha_H} e^{s(q_i, h_{i,j}^-)}} \quad (2)$$

where $s(\cdot)$ denotes cosine similarity. This objective encourages the model to align semantically equivalent expressions while explicitly separating misleading or contrastive realizations.

Configurations. We fine-tune a subset of models using four training configurations that mirror the zero-shot configurations from §4: training is performed on queries with or without instructions, paired with sentence- or span-level embeddings. During inference, queries are formulated and embedded following the same procedure used during training. Models trained with instructions are evaluated using both sentence- and span-level embeddings, whereas models trained with span-level embeddings are evaluated using span embeddings only.

5 Experimental Settings

For details about software, hardware, running times, etc., refer to Appendix C.

5.1 Evaluation Metrics

As a retrieval benchmark, IdioLink is evaluated using two standard, complementary metrics (more in Appendix D): (i) nDCG@10 (binary relevance), which emphasizes ranking quality at top positions (Lee et al., 2024; Günther et al., 2025); (ii) R-Precision, which measures precision at rank R, the

number of relevant documents per query (Wu et al., 2024; ?). In IdioLink, we set $R=60$ for idiomatic queries and $R=40$ for literal queries. These values are determined by design and reflect the number of relevant documents generated for each query type during dataset construction (§ 3.2.2). These metrics capture both early precision and overall retrieval correctness, providing a sufficient and interpretable evaluation of ranking quality.

5.2 Models

IdioLink is designed to evaluate the embedding-based semantic retrieval methods for figurative language. We evaluate 24 text-embedding models, with parameter counts ranging from 100 million to 9 billion, and provide the exact checkpoint details and licenses in Table 7 (Appendix H). Following Weller et al. (2025), we group models into two categories:

No instructions in training. We evaluate models that were not post-trained on queries containing instructions. Such encoder models include MiniLM_{L6-v2} (Wang et al., 2021), SimCSE (Gao et al., 2021), Contriever (Izacard et al., 2022) and E5 (Wang et al., 2024a). We also include decoder models: Stella (Zhang et al., 2025a) and DRAMA (Ma et al., 2025).

Instructions in training. Models that are post-trained with instructions given with the queries. This ranges from minimal, constant instruction to fully flexible instruction-following embedding models. We assess encoder models: Contriever-based TART (Asai et al., 2023), the instruction-tuned multilingual E5 (Wang et al., 2024c), and Nomic-Embed-v2 (Nussbaum and Duderstadt, 2025). BGE models (Xiao et al., 2024) and their multilingual version M3 (Dense) (Chen et al., 2024) also fall into this category, although they are trained with only one instruction for each broad task (retrieval, clustering, etc.) We also evaluate decoder models: InstructOR (Su et al., 2023), GritLM-7B (Muennighoff et al., 2025), E5-Mistral (Wang et al., 2024b), the GTE models (Li et al., 2023), which are based on Qwen-2 (Yang et al., 2024), the SFR-Embedding-Mistral (Meng et al., 2024), and the Linq-Embed-Mistral (Choi et al., 2024), Lychee-embed (Zhang et al., 2025b), which is based on Qwen-2.5 (Yang et al., 2025), the Gemma-based BGE model (Chen et al., 2024), Llama-Embed_{Nemotron-8B} (Babakhin et al., 2025) and the Qwen3 Embedding models (Zhang et al., 2025c).

6 Results and Discussion

Table 2 summarizes the zero-shot results, from which we derive the following main findings.

First, augmenting queries with instructions substantially improves retrieval performance.

Across all models, both instruction-tuned and non-instruction models, augmenting the query with explicit instructions that focus the model on the target span leads to consistent gains in both R-Precision and nDCG (+11.21 R-Precision, +14.83 nDCG on average). For example, GTE_{Qwen2-7B} improves by 33.5 points in R-Precision and 55.10 in nDCG. These gains highlight the benefit of explicit inquiry semantic guidance for task-specific retrieval. Only one model, BGE-M3 (Dense), shows no measurable improvement from instruction augmentation, though its performance does not degrade.

Second, span embedding substantially improves retrieval performance.

For most models, replacing sentence embedding with span embedding yields large improvements in both R-Precision and nDCG (+12.13 and +13.49 on average, respectively). For example, InstructOR_{xl} gains 32.35 points in R-Precision and 32.05 in nDCG. This highlights the importance of modeling the localized span that carries the intended meaning for figurative language retrieval. Only BGE_{multilingual-gemma2} and Lychee-embed show notable drops under span embedding, indicating this approach does not benefit all embedding architectures equally.

Third, combining instructions and span embedding yields the strongest gains.

Adding instructions and using span embedding yields the largest gains (+21.99 R-Precision, +22.71 nDCG on average). For example, GritLM_{7B} improves by 43.50 points in R-Precision and achieves substantial gains in nDCG. Models that perform poorly when using span embeddings in isolation also underperform in a combined setting, though the degradation is notably smaller.

Fourth, model scale alone does not determine retrieval performance.

Although larger models often perform well, parameter count alone is not a reliable predictor of retrieval quality. Smaller or mid-sized models can outperform larger counterparts when paired with effective representation strategies. For example, across configurations, the 1B Drama_{1B} achieves R-Precision and nDCG scores of 52.0 and 72.9, respectively, whereas the

			without instructions				with instructions			
			R-Precision		nDCG@10		R-Precision		nDCG@10	
Model	Param		sentence	span	sentence	span	sentence	span	sentence	span
No-Instruct	MiniLM _{L6-v2}	22M	18.34	40.88 [▲]	35.08	58.86 [▲]	32.57	42.40 [▲]	52.76	55.10 [▲]
	Contriever	110M	9.60	38.40 [▲]	21.60	64.80 [▲]	15.10	42.10 [▲]	30.80	64.10 [▲]
	E5 _{base-v2}	110M	20.00	30.20 [▲]	38.50	52.50 [▲]	24.30	35.40 [▲]	45.10	57.60 [▲]
	Drama _{1B}	1B	19.70	52.00 [▲]	37.10	72.80 [▲]	23.90	50.30 [▲]	42.10	64.90 [▲]
	Stella _{en_1.5B_v5}	2B	18.30	27.60 [▲]	32.00	41.10 [▲]	18.30	43.60 [▲]	32.00	58.10 [▲]
Instruction	TART-Contriever	110M	12.10	40.40 [▲]	26.20	64.10 [▲]	17.50	44.60 [▲]	34.30	60.20 [▲]
	BGE _{base-en-v1.5}	109M	20.00	22.80 [▲]	38.20	42.30 [▲]	24.70	26.10 [▲]	42.30	43.90 [▲]
	InstructOR _{base}	335M	19.80	52.96 [▲]	38.70	68.37 [▲]	21.30	47.50 [▲]	40.40	55.70 [▲]
	Nomic-v2	0.5B	17.90	41.20 [▲]	35.40	65.80 [▲]	24.20	45.00 [▲]	46.60	62.50 [▲]
	BGE-M3 (Dense)	0.5B	22.50	30.10 [▲]	42.80	52.50 [▲]	22.60	37.10 [▲]	42.80	55.90 [▲]
	Multilingual-E5 _{large-instruct}	0.6B	25.60	35.60 [▲]	45.80	58.70 [▲]	30.10	40.20 [▲]	52.40	63.30 [▲]
	Qwen3-Embedding _{0.6B}	0.6B	15.70	28.40 [▲]	30.40	45.70 [▲]	25.90	31.70 [▲]	45.20	47.30 [▲]
	InstructOR _{xl}	1.5B	23.60	55.95[▲]	42.50	74.55[▲]	31.80	59.10[▲]	58.40	62.3 [▲]
	Lychee-embed	1.5B	12.40	11.20 [▽]	27.00	18.40 [▽]	20.40	15.80 [▽]	41.20	26.90 [▽]
	GTE _{Qwen2-1.5B}	1.5B	16.40	21.40 [▲]	33.00	33.20 [▲]	48.00	50.40 [▲]	69.80	73.10 [▲]
	Qwen3-Embedding _{4B}	4B	20.00	27.00 [▲]	36.40	43.70 [▲]	26.50	39.00 [▲]	44.90	59.20 [▲]
	Linq-Embed-Mistral	7B	34.60	37.60 [▲]	55.00	58.30 [▲]	52.80	51.50 [▽]	75.50	72.50 [▽]
	SFR-Embedding-Mistral	7B	31.40	33.60 [▲]	50.40	53.70 [▲]	45.00	44.70 [▽]	66.40	65.50 [▽]
	E5 _{Mistral}	7B	31.50	32.40 [▲]	52.00	53.10 [▲]	42.10	44.20 [▲]	64.40	64.90 [▲]
	GritLM _{7B}	7B	13.90	32.80 [▲]	32.90	63.00 [▲]	49.50	57.40 [▲]	72.90	77.10 [▲]
	GTE _{Qwen2-7B}	7B	2.90	6.00 [▲]	6.60	11.20 [▲]	36.40	40.10 [▲]	61.70	64.00 [▲]
	Qwen3-Embedding _{8B}	8B	25.10	31.60 [▲]	43.70	50.50 [▲]	31.80	49.50 [▲]	52.10	69.50 [▲]
	Llama-Embed _{Nemotron-8B}	8B	23.80	41.00 [▲]	41.10	62.60 [▲]	41.62	54.20 [▲]	66.13	78.53[▲]
	BGE _{multilingual-gemma2}	9B	37.50	12.70 [▽]	58.90	15.30 [▽]	38.90	28.50 [▽]	61.60	44.30 [▽]

Table 2: Zero-shot results on the IdioLink dataset. Embedding method is indicated by sentence and span. The best result in each configuration is highlighted in bold. The “▲” symbol indicates an increase and “▽” decrease compared to the respective sentence embedding setting.

9B BGE_{multilingual-gemma2} reaches only 38.9 and 61.6. Similarly, the 1.5B InstructOR_{xl} reaches an R-Precision of 59.1, making it the best-performing model on this metric in the zero-shot setting, and achieves an nDCG of 74.55 in another configuration. Overall, differences attributable to query formulation and embedding extraction frequently exceed those attributable to model size, highlighting the central role of representation design.

A fine-grained analysis is provided in Appendix E. Notably, idiomatic queries are consistently harder than literal ones (§E.1). We also report subject-level retrieval (§E.2) and lexical overlap control (§E.3).

6.1 Fine-tuning Results

Figure 3 compares fine-tuned models with their zero-shot counterparts across all four inference-time configurations, while the full results are reported in Table 6 (Appendix F). Each fine-tuning experiment is repeated using three different random seeds, and results are reported as the mean with the corresponding standard deviation. Each subplot corresponds to a specific inference-time configuration (§ 4). For example, when span embedding is used at inference time, fine-tuning is performed

either with sentence or with span embedding; in the plot, we report the higher of the two.

Fine-tuning consistently improves performance over zero-shot settings. Consistent with the zero-shot findings, span embedding at inference time yields clear performance gains for fine-tuned models, regardless of the training configuration. In contrast, instructions at inference time typically lead to a slight performance degradation, though it is not substantial.

Standard training configuration wins. Unlike in the zero-shot setting, incorporating instructions during fine-tuning consistently harms performance compared to using raw queries. Similarly, except for BGE-M3 (Dense), models fine-tuned using span embedding underperform those trained with the standard sentence embedding, assuming span embedding is applied at inference time for a fair comparison. Overall, the most effective fine-tuning configuration relies on standard methods, queries with no instructions, and sentence embedding.

7 Conclusions

We introduced IdioLink, a retrieval benchmark that tests whether embedding models can retrieve con-

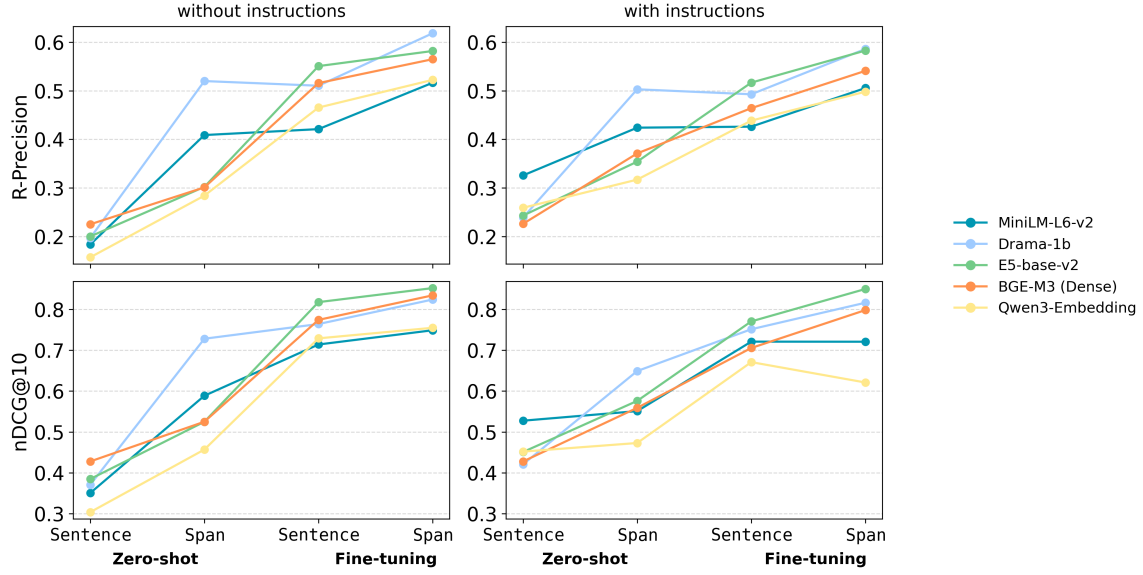


Figure 3: Performance across our four inference-time configurations, where fine-tuning is done with no instructions and with sentence embeddings (§ 4).

ceptually equivalent meanings across literal and idiomatic expressions. Our results show that IdioLink is challenging even for state-of-the-art models, with performance driven more by representation choices than by model scale. In zero-shot settings, both instruction augmentation and span embedding substantially improve retrieval, and their combination yields the strongest results. Fine-tuning further improves performance, but only under standard query and embedding configurations; adding instructions or using span embedding in training proved harmful. At inference time, however, span embedding provides robust gains, highlighting the importance of decoupling training and inference strategies.

Overall, our findings suggest that effective idiom retrieval requires careful design of representations beyond scaling alone. IdioLink offers a controlled testbed for studying conceptual retrieval under semantic variation and opens avenues for future work on meaning-centric evaluation and figurative language understanding.

Limitations

IdioLink focuses on English and on a curated set of PIEs, enabling controlled experimentation but limiting linguistic diversity and cross-lingual generalization. Future work could extend the benchmark to additional languages, broader figurative phenomena, and more abstract semantic relations. In addition, we collapse multiple figurative senses of the same PIE into a single retrieval target, which

may introduce ambiguity for some expressions.

Although the dataset undergoes both automatic and human validation, part of the corpus is synthetically generated. As with other LLM-generated benchmarks, some artifacts or unnatural distributions may remain despite filtering and manual correction.

Our evaluation is conducted in a controlled retrieval setting with relatively short documents and limited training data (440 training queries). Real-world retrieval systems often involve longer documents, noisier inputs, multi-hop reasoning, and larger-scale supervision. Similarly, experiments are primarily designed to compare architectures and retrieval configurations rather than optimize individual models exhaustively.

In this work, we evaluate how well embedding models capture the figurative meaning of idiomatic expressions, framing this as a retrieval task. However, retrieval metrics such as nDCG are optimized for search relevance rather than semantic similarity, and casting this as a clustering problem may better expose the models’ ability to group expressions by figurative meaning.

Despite these limitations, we believe IdioLink provides a challenging and complementary testbed for evaluating retrieval beyond surface-level similarity.

Ethics Statement

We use publicly available datasets and models in accordance with their intended use, as detailed by the respective publishers and under their licenses. Details are provided in Table 7 (Appendix H) and in Table 5 (Appendix G). No personally identifiable information or offensive data are processed, and annotators ensure that none exists in the released dataset. Our work is intended for research purposes only, and we see no potential risks.

Acknowledgments

This study is supported in part by the European Research Council (Intellexus, Project No. 101118558). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authorities can be held responsible for them.

References

- Saeed Almheiri, Bilal Elbouardi, Salsabila Zahira Pranita, Irina Nikishina, Ashwath Rao B, Parameswari Krishnamurthy, Muhammad Cendekia Airlangga, Rifo Ahmad Genadi, Nguyen Phan Gia Bao, Amir Hossein Yari, Hawau Olamide Toyin, Nurdaulet Mukhituly, Mena Attia, Beshir Hassan, Ahmad Fathan Hidayatullah, Tatsuki Kuribayashi, Haonan Li, Suma Bhat, and Fajri Koto. 2026. [Multilingual idioms in sentences and conversations across high-, medium-, and low-resource languages](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12363–12389, San Diego, California, United States. Association for Computational Linguistics.
- Doğukan Arslan, Hüseyin Anıl Çakmak, Gülşen Eryiğit, and Joakim Nivre. 2025. [Using LLMs to advance idiom corpus construction](#). In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 21–31, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.
- Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. [Task-aware retrieval with instructions](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675, Toronto, Canada. Association for Computational Linguistics.
- Gözde Aslantaş and Tunga Gungor. 2026. [A unified Turkish idiom understanding benchmark: Idiom detection and semantic retrieval across five Turkic languages](#). In *Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026)*, pages 38–51, Rabat, Morocco. Association for Computational Linguistics.
- Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. [Llama-Embed-Nemotron-8B: A universal text embedding model for multilingual and cross-lingual tasks](#). *Preprint*, arXiv:2511.07025.
- Timothy Baldwin and Su Nam Kim. 2010. [Multiword expressions](#). In Nitin Indurkha and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2 edition, pages 267–292. CRC Press, Boca Raton, FL.
- J Briskilal and C.N. Subalalitha. 2022. [An ensemble model for classifying idioms and literal texts using BERT and RoBERTa](#). *Information Processing & Management*, 59(1):102756.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Chanyeol Choi, Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, and Jy yong Sohn. 2024. [Linq-Embed-Mistral technical report](#). *Preprint*, arXiv:2412.03223.
- Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. [Survey: Multiword expression processing: A Survey](#). *Computational Linguistics*, 43(4):837–892.
- Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. [Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context](#). In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 41–48, Prague, Czech Republic. Association for Computational Linguistics.
- Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-Tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*, pages 19–22.
- Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. [A hard nut to crack: Idiom detection with conversational large language models](#). In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

- Gülşen Eryiğit, Ali Şentaş, and Johanna Monti. 2022. [Gamified crowdsourcing for idiom corpora construction](#). *Natural Language Engineering*, 29(4):909–941.
- Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. [Unsupervised type and token identification of idiomatic expressions](#). *Computational Linguistics*, 35(1):61–103.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. [Probing for idiomaticity in vector space models](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.
- Raymond W. Gibbs, Nandini P. Nayak, John L. Bolton, and Melissa E. Keppel. 1989a. [Speakers’ assumptions about the lexical flexibility of idioms](#). *Memory & Cognition*, 17(1):58–68.
- Raymond W. Gibbs, Nandini P. Nayak, and Cooper Cutting. 1989b. [How to kick the bucket and not decompose: Analyzability and idiom processing](#). *Journal of Memory and Language*, 28(5):576–593.
- Raymond W. Gibbs, Jr and Herbert L. Colston. 2012. *Interpreting Figurative Meaning*. Cambridge University Press.
- Michael Günther, Isabelle Mohr, Daniel James Williams, Bo Wang, and Han Xiao. 2025. [Late chunking: Contextual chunk embeddings using long-context embedding models](#). *Preprint*, arXiv:2409.04701.
- Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. [MAGPIE: A large corpus of potentially idiomatic expressions](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.
- Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. [AlphaMWE: Construction of multilingual parallel corpora with MWE annotations](#). In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.
- Kai Golan Hashiloni, Ofri Hefetz, and Kfir Bar. 2025. [Easy as PIE? identifying multi-word expressions with LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23771–23790, Suzhou, China. Association for Computational Linguistics.
- Kai Golan Hashiloni, Lior Livyatan, Ofri Hefetz, Alon Manner, Bar Cohen, and Kfir Bar. 2026. [ID10M-JAM: Stress-testing idiom identification under challenging context](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 20846–20864, San Diego, California, United States. Association for Computational Linguistics.
- Wei He, Marco Idiart, Carolina Scarton, and Aline Villavicencio. 2024. [Enhancing idiomatic representation in multiple languages via an adaptive contrastive triplet loss](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12473–12485, Bangkok, Thailand. Association for Computational Linguistics.
- Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2025. [Investigating idiomaticity in word representations](#). *Computational Linguistics*, 51:505–555.
- Ofri Hefetz, Kai Golan Hashiloni, Alon Manner, and Kfir Bar. 2025. [Not just a piece of cake: Cross-lingual fine-tuning for idiom identification](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2521–2537, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. [Unsupervised dense information retrieval with contrastive learning](#). *Preprint*, arXiv:2112.09118.
- Ray S. Jackendoff. 1997. *The Architecture of the Language Faculty*, volume 28 of *Linguistic Inquiry Monographs*. MIT Press, Cambridge, MA; London, England.
- Rohan Jha, Bo Wang, Michael Günther, Georgios Mastrotras, Saba Sturua, Isabelle Mohr, Andreas Koukounas, Mohammad Kalim Akram, Nan Wang, and Han Xiao. 2024. [Jina-ColBERT-v2: A general-purpose multilingual late interaction retriever](#). In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 159–166, Miami, Florida, USA. Association for Computational Linguistics.
- Greg Kamradt. 2024. [5 levels of text splitting](#). GitHub repository. Accessed: 2025-12-27.
- Jenna Kanerva, Hanna Kitt, Li-Hsin Chang, Teemu Vahtola, Mathias Creutz, and Filip Ginter. 2025. [Semantic search as extractive paraphrase span detection](#). *Language Resources and Evaluation*, 59(1):257–276.
- 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 *Proceedings of the 2020 Conference on Empirical Methods in Natural*

- Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. [Memorization or reasoning? exploring the idiom understanding of LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21678–21699, Suzhou, China. Association for Computational Linguistics.
- Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. [SemEval-2013 task 5: Evaluating phrasal semantics](#). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Praateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftexhar Naim. 2024. [Gecko: Versatile text embeddings distilled from large language models](#). *Preprint*, arXiv:2403.20327.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). *Preprint*, arXiv:2005.11401.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. [Towards general text embeddings with multi-stage contrastive learning](#). *Preprint*, arXiv:2308.03281.
- Xueguang Ma, Xi Victoria Lin, Barlas Oguz, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2025. [DRAMA: Diverse augmentation from large language models to smaller dense retrievers](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30170–30186, Vienna, Austria. Association for Computational Linguistics.
- Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. [SFR-Embedding-Mistral:enhance text retrieval with transfer learning](#). Salesforce AI Research Blog.
- Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. [Rolling the DICE on idiomaticity: How LLMs fail to grasp context](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.
- Pu Miao, Zeyao Du, and Junlin Zhang. 2023. [DebcSE: Rethinking unsupervised contrastive sentence embedding learning in the debiasing perspective](#). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, (CIKM 23), page 1847–1856. ACM.
- Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2025. [Generative representational instruction tuning](#). *Preprint*, arXiv:2402.09906.
- Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. [Idioms](#). *Language*, 70(3):491–538.
- Zach Nussbaum and Brandon Duderstadt. 2025. [Training sparse mixture of experts text embedding models](#). *Preprint*, arXiv:2502.07972.
- Seoyoon Park, Hyeji Choi, Minseon Kim, Subin An, Xiaonan Wang, Gyuri Choi, and Hansaem Kim. 2025. [FLUID QA: A multilingual benchmark for figurative language usage in dialogue across English, Chinese, and Korean](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30280–30294, Suzhou, China. Association for Computational Linguistics.
- Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. [Sign of the times: Evaluating the use of large language models for idiomaticity detection](#). In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.
- Howard R. Pollio, John M. Barlow, Howard J. Fine, and Marilyn R. Pollio. 1977. *Psychology and the Poetics of Growth: Figurative Language in Psychology, Psychotherapy, and Education*. Lawrence Erlbaum, Hillsdale, NJ.
- Jipeng Qiang, Yang Li, Chaowei Zhang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023. [Chinese idiom paraphrasing](#). *Transactions of the Association for Computational Linguistics*, 11:740–754.
- Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archana Bhatia, Uxoia Iñurieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. [Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions](#). In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [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, Hong Kong, China. Association for Computational Linguistics.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Found. Trends Inf. Retr.*, 3(4):333–389.
- Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. [ColBERTv2: Effective and efficient retrieval via lightweight late interaction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.
- Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archana Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurieta, Albert Gatt, and 9 others. 2023. [PARSEME corpus release 1.3](#). In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.
- Prateek Saxena and Soma Paul. 2020. [EPIE dataset: A corpus for possible idiomatic expressions](#). *Preprint*, arXiv:2006.09479.
- Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Mitrofan, and Vasile Pais. 2026. [Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions](#). In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 254–275, Rabat, Morocco. Association for Computational Linguistics.
- Zhan Shi, Guoyin Wang, Ke Bai, Jiwei Li, Xiang Li, Qingjun Cui, Belinda Zeng, Trishul Chilimbi, and Xiaodan Zhu. 2023. [OssCSE: Overcoming surface structure bias in contrastive learning for unsupervised sentence embedding](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7242–7254, Singapore. Association for Computational Linguistics.
- Caroline Sporleder, Linlin Li, Philip Gorinski, and Xaver Koch. 2010. [Idioms in context: The IDIX corpus](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. [One embedder, any task: Instruction-finetuned text embeddings](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.
- Shiva Taslimipour, Sara Bahaadini, and Ekaterina Kochmar. 2020. [MTLB-STRUCT @parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models](#). In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.
- Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. [SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding](#). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.
- Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. [ID10M: Idiom identification in 10 languages](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.
- Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao. 2022. [Snscse: Contrastive learning for unsupervised sentence embedding with soft negative samples](#). *Preprint*, arXiv:2201.05979.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024a. [Text embeddings by weakly-supervised contrastive pre-training](#). *Preprint*, arXiv:2212.03533.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. [Improving text embeddings with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024c. [Multilingual e5 text embeddings: A technical report](#). *Preprint*, arXiv:2402.05672.
- Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. [MiniLMv2: Multi-head self-attention relation distillation for compressing pre-trained transformers](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-Consistency improves chain of thought reasoning in language models](#). *Preprint*, arXiv:2203.11171.

- Uriel Weinreich. 1969. Problems in the analysis of idioms. In *Problems in the Analysis of Idioms*, pages 23–82. University of California Press, Berkeley.
- Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2025. [FollowIR: Evaluating and teaching information retrieval models to follow instructions](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11926–11942, Albuquerque, New Mexico. Association for Computational Linguistics.
- Sondre Wold, Étienne Simon, Erik Velldal, and Lilja Øvrelid. 2026. [Measuring idiomaticity in text embedding models with epsilon-compositionality](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2239–2252, Rabat, Morocco. Association for Computational Linguistics.
- Di Wu, Da Yin, and Kai-Wei Chang. 2024. [KPEval: Towards fine-grained semantic-based keyphrase evaluation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1959–1981, Bangkok, Thailand. Association for Computational Linguistics.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-pack: Packed resources for general chinese embeddings](#). *Preprint*, arXiv:2309.07597.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. [Qwen2 technical report](#). *Preprint*, arXiv:2407.10671.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jixia Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Fengying Ye, Yanming Sun, Runzhe Zhan, Lidia S. Chao, Zheqi Zhang, and Derek F. Wong. 2026. [G-IdiomAlign: A gloss-pivoted benchmark for cross-lingual idiom alignment](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 38720–38739, San Diego, California, United States. Association for Computational Linguistics.
- Ziheng Zeng and Suma Bhat. 2021. [Idiomatic expression identification using semantic compatibility](#). *Transactions of the Association for Computational Linguistics*, 9:1546–1562.
- Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025a. [Jasper and stella: distillation of sota embedding models](#). *Preprint*, arXiv:2412.19048.
- Xin Zhang, Yanzhao Zhang, Wen Xie, Dingkun Long, Mingxin Li, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025b. [Phased training for LLM-powered text retrieval models beyond data scaling](#). In *Second Conference on Language Modeling*.
- 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. 2025c. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *Preprint*, arXiv:2506.05176.
- Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021a. [PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing](#). In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.
- Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. 2021b. [Idiomatic expression paraphrasing without strong supervision](#). *Preprint*, arXiv:2112.08592.

A Annotators Information

Data annotation and validation are performed by three annotators with academic backgrounds in scientific research: one PhD researcher and two PhD candidates, all of whom are proficient in English. Annotators volunteered to support the research activities and were thus not paid specifically for performing the annotation task.

B Instruction Implementation Details

The instruction is prepended to the original query and formulated as:

```
Based on the literal/idiomatic usage of
the span {span} in the query, retrieve
documents that contain a span conveying
the same conceptual meaning.
```

where {span} denotes the target span in the query.

Instructions are provided in accordance with the recommended usage patterns for each model, as specified by the model’s authors or providers. When available, we follow the model-specific instruction formatting guidelines. For full implementation details per model, we refer the reader to the project repository.

In practice, we observe two common instruction integration strategies:

- (i) Some models require the instruction to be passed separately as an additional argument to the encode function, alongside the queries.
- (ii) Another widely used format adopted by many embedding models, in which the final input is constructed as:

```
model.encode(queries, instruction
= instruction).
```

```
Instruct: {instruction}
Query: {query}
```

We additionally apply this instruction-query concatenation for models that do not support a dedicated instruction field or predefined format.

C Technical Details

All of our experiments run on an M4 Pro GPU (20C) with 64GB of memory. Including experimentation and hyperparameter tuning, we estimate that all zero-shot experiments took approximately 24 hours, whereas those with fine-tuning took approximately 90 hours. For prompting and accessing LLMs, we use Agno, which provides a modular,

reproducible framework for LLM evaluation, particularly when working with multiple providers, by wrapping their APIs in a unified layer. We use *SentenceTransformer* to load and use embedding models. For the fine-tuning approach, we utilize *PyTorch*.

C.1 BM25 Setup

Although IdioLink is designed to evaluate embedding-based semantic retrieval for figurative language, we include BM25 (Robertson and Zaragoza, 2009) as a strong lexical control baseline. This allows us to estimate how much retrieval performance can be achieved through surface-form overlap alone, without learned semantic representations.

BM25 is a bag-of-words ranking function based on exact term overlap, term frequency saturation, inverse document frequency, and document length normalization. We use the rank-bm25 implementation of BM25Okapi with lowercase regex tokenization (`(\b\b\w+(?:'\w+)?\b)`). All configurations index the sentence field of documents and retrieve the top 100 candidates.

Query configurations. We evaluate two BM25 configurations: sentence, which uses the full query sentence as the retrieval query, and span, which uses only the annotated PIE span. The span configuration serves as a task-aware lexical variant that explicitly emphasizes the target expression.

Hyperparameter tuning. We tune k_1 and b on the validation split using $\text{nDCG}@10$, searching over $k_1 \in \{0.6, 0.9, 1.2, 1.5, 2.0\}$ and $b \in \{0.2, 0.4, 0.6, 0.75, 0.9\}$. Validation selected $k_1 = 0.9$ and $b = 0.4$, matching standard Lucene/Anserini defaults.

D Evaluation Metrics

We evaluate retrieval quality using R-precision and $\text{nDCG}@10$, two standard information retrieval metrics that measure different aspects of ranking quality.

R-precision. For a query q with K relevant documents, R-precision measures the proportion of relevant documents retrieved within the top- K ranked results. A perfect score of 1.0 means that all relevant documents appear within the first K retrieved items.

In our benchmark, the number of relevant documents is 40 for literal questions and 60 for the

idiomatic ones

Example. Suppose a query has $K = 5$ relevant documents. A model ranks relevant documents at positions $[1, 2, 4, 7, 8]$. Only the top-5 retrieved documents are considered for R-precision. Since 3 of the 5 relevant documents appear within the top-5 positions, the R-precision score is:

$$R\text{-precision} = \frac{3}{5} = 0.6$$

nDCG@10. We also report normalized Discounted Cumulative Gain at rank 10 (nDCG@10), using binary relevance labels. Unlike R-precision, nDCG@10 is sensitive to the ordering of retrieved results, assigning greater weight to relevant documents that appear earlier in the ranking. This is particularly important in our setting because retrieval models may return partially related documents (e.g., documents that share topical context or lexical overlap) while failing to rank true semantic matches in top positions.

Example. Consider two models: one ranks these documents at positions $[1, 2, 3]$, while the other ranks them at $[3, 7, 9]$. Both systems retrieve relevant documents, but the first receives a higher nDCG@10 score because semantically correct matches appear earlier in the ranking.

E Fine-Grained Analysis

We further analyze the zero-shot results by stratifying performance according to query type (§E.1), subject-level retrieval (§E.2), and lexical overlap (§E.3).

E.1 Literal vs. Idiomatic Query Performance

Figure 4 shows that literal queries consistently outperform idiomatic queries across all four inference-time configurations: sentence, span, with instructions, sentence, and with instructions, span. Results are averaged across 13 embedding models: BGE_{base-en-v1.5}, BGE-M3, Qwen3-Embedding_{0.6B}, Contriever, Drama_{1B}, InstructOR_{base}, InstructOR_{x1}, E5_{base-v2}, Multilingual-E5_{large-instruct}, Nomic-v2, TART-Contriever, MiniLM_{L6-v2}, and Lychee-embed. This confirms that retrieving figurative meaning is substantially more challenging than retrieving compositional meaning, even when the same target expression is provided.

Span embeddings improve retrieval for both query types, but the gains are larger for literal queries. Average R-Precision increases from 20.87

to 39.03 for literal queries when moving from sentence to span embeddings, compared to 15.62 to 22.10 for idiomatic queries.

E.2 Subject-level Retrieval Trends

In addition, subject-level retrieval follows the opposite trend. As shown in Figure 4, sentence embeddings achieve the highest subject-level nDCG@10, while span-based variants substantially reduce subject-level performance. This indicates that full-sentence embeddings preserve broad topical similarity, whereas span-focused representations shift the retrieval signal toward the localized meaning-bearing expression. Overall, this analysis supports our main finding: IdioLink cannot be solved by topical matching alone, and idiom-aware conceptual retrieval remains challenging for current embedding models.

E.3 Lexical Overlap Control

Reduced lexical overlap means the relevant documents no longer contain the idiom’s surface form, so a model must rely on shared meaning rather than shared words to retrieve them. To isolate this, we construct two overlap-controlled index variants. The *literal-to-idiomatic* index contains literal and idiomatic documents, where relevant idiomatic documents often preserve the PIE’s surface form. The *literal-to-simplification/sense* index instead pairs literal documents with simplification and sense documents, where the figurative meaning is expressed without the original idiom—testing whether models retrieve shared meaning beyond exact expression matching.

Table 3 shows a clear dependence on lexical overlap. BM25 performs strongly in the literal-to-idiomatic index, reaching 40.25 R-Precision and 54.90 nDCG@10 on idiomatic queries, since relevant documents often contain the same PIE. However, BM25 collapses in the literal-to-simplification/sense index, achieving only 0.03 R-Precision and 0.04 nDCG@10 on idiomatic queries with span-based retrieval, demonstrating that lexical overlap alone is insufficient when the relevant figurative meaning is expressed without the idiom surface form.

Dense models also benefit from lexical overlap. Under span embeddings, Contriever improves from 5.27 to 46.98 idiomatic R-Precision when moving from literal-to-simplification/sense to literal-to-idiomatic, and Drama_{1B} improves from 20.99 to 51.71. However, some stronger embedding con-

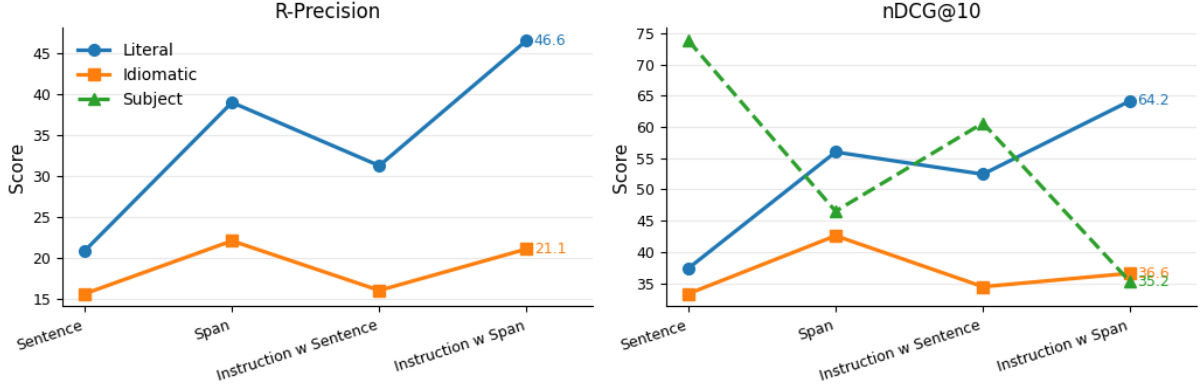


Figure 4: **Fine-grained zero-shot trends across query type and subject.** We report average R-Precision (left) and nDCG@10 (right) for literal queries, idiomatic queries, and subject-level retrieval across 13 embedding models. R-Precision is computed with $R = 40$ for literal queries and $R = 60$ for idiomatic queries, matching the number of relevant documents in the index. Since R is not well defined for subject-level retrieval, subject results are shown only for nDCG@10.

Model	Query mode	literal-to-simplification/sense				literal-to-idiomatic			
		Lit. R-P	Idi. R-P	Lit. nDCG	Idi. nDCG	Lit. R-P	Idi. R-P	Lit. nDCG	Idi. nDCG
BM25	sentence	58.43	6.82	70.17	15.83	45.09	40.25	62.14	54.90
	span	96.84	0.03	97.20	0.04	66.52	28.29	73.44	26.02
BGE-M3	sentence	24.86	16.06	43.33	28.92	26.27	27.24	44.59	39.66
Drama _{1B}	sentence	20.70	15.68	35.61	28.82	21.43	22.12	36.39	33.50
mE5 _{large-inst}	sentence	30.11	18.47	49.90	31.71	30.59	25.73	50.28	37.28
Contriever	span	55.60	5.27	76.60	6.96	50.65	46.98	70.22	60.61
Drama _{1B}	span	65.95	20.99	83.12	26.22	63.09	51.71	81.23	64.21
E5 _{base}	span	38.53	15.42	58.05	27.54	38.20	30.85	57.56	44.88
mE5 _{large-inst}	span	44.55	21.77	65.38	35.71	44.25	35.47	64.98	49.61
Qwen3-4B	w/ instructions, sentence	30.43	21.97	47.68	36.64	31.60	23.11	48.57	33.44
mE5 _{large-inst}	w/ instructions, sentence	40.80	15.29	62.73	26.59	41.30	28.63	62.88	41.47
MiniLM _{L6-v2}	w/ instructions, sentence	51.02	2.29	74.59	2.05	47.10	28.83	67.71	37.78
Qwen3-4B	w/ instructions, span	42.37	30.13	62.55	44.80	43.27	29.35	62.95	40.83
Contriever	w/ instructions, span	64.82	3.57	84.24	3.49	57.15	44.86	74.60	54.21
Drama _{1B}	w/ instructions, span	73.23	11.41	89.36	9.46	68.50	38.99	85.09	47.48
mE5 _{large-inst}	w/ instructions, span	56.13	16.79	77.09	25.68	55.42	35.53	76.16	49.19
TART-Contriever	w/ instructions, span	72.59	2.71	90.14	2.09	65.70	33.10	81.55	39.62

Table 3: Lexical-overlap ablation for literal queries (Table ??). BM25 is a lexical control; dense models cover sentence and span embeddings, with and without instructions.

figurations retain more signal under reduced lexical overlap—for example, Qwen3-Embedding-4B with instructions and span embeddings obtains similar idiomatic R-Precision in both settings. Overall, these ablations confirm that IdioLink contains surface-form signal, but the most challenging part of the benchmark requires retrieving figurative meaning when lexical overlap is reduced or absent.

F Fine-Tuning Results and Hyperparameters

We present the full results of the fine-tuning experiments in Table 6.

We list our fine-tuning hyperparameters in Table 4, which are selected based on preliminary experimentation on the validation set. We use the

AdamW optimizer, a *linear* learning rate scheduler, and early stopping.

With a larger α_H , we notice gradient instability that leads to slow convergence. This could be addressed with additional training data. When using a smaller α_H , the model failed to learn the fine-grained separation of figurative usage that we aim to achieve.

G Artifacts

We detail the artifacts we use, their usage, and licenses in Table 5. For the exact versions and more details, see the *requirements.txt* file in the project’s repository.

Parameter	Value
α_N (“soft” negatives)	3
α_H (“hard” negatives)	2
Max epochs	10
Batch size	32*
Learning rate	2e-5
Warm-up steps (linear)	100
Min-delta (early stopping)	0.001
Patience epochs (early stopping)	3

Table 4: Fine-tuning hyperparameters. * Drama runs with batch size 8.

Artifact	Type	License	Usage
MAGPIE ^a	Dataset	CC-BY-4.0	Prompting
Agno ^b	Framework	Apache-2.0	Prompting
SentenceTransformer ^c	Package	Apache-2.0	Embedding models
PyTorch ^d	Package	Own	Embedding models
rank-bm25 ^e	Package	Apache-2.0	BM25 model

^a <https://github.com/hslh/magpie-corpus>

^b <https://www.agno.com/>

^c <https://huggingface.co/sentence-transformers>

^d <https://github.com/pytorch/pytorch>

^e <https://pypi.org/project/rank-bm25/>

Table 5: Artifacts used during experiments, along with their license and usage explanation.

H Model Checkpoints

In Table 7, we present the checkpoints (or snapshots) used in this work and the models’ sizes and licenses.

I AI assistants

We used AI assistants (e.g., ChatGPT) to support code formatting, phrasing suggestions, and LaTeX styling during writing. All outputs were reviewed and edited by the authors. No content was directly generated or used without human verification.

Train	Model	Param	I/N	w/o instructions				w/ instructions			
				R-Precision		nDCG@10		R-Precision		nDCG@10	
				sentence	span	sentence	span	sentence	span	sentence	span
w/o inst - sentence	MiniLM _{L6-v2}	22M	-	42.14±0.05	51.78±0.10 [▲]	71.45±0.07	74.60±0.30 [▲]	-	-	-	-
	Drama _{1B}	1B	-	50.32±1.39	60.73±1.51 [▲]	75.48±1.97	80.15±3.48 [▲]	-	-	-	-
	E5 _{base-v2}	110M	-	55.41±0.43	58.60±0.30[▲]	81.81±0.32	85.50±0.30[▲]	-	-	-	-
	BGE-M3 (Dense)	0.5B	I	51.37±0.79	55.28±0.87 [▲]	77.12±0.63	81.81±0.63 [▲]	-	-	-	-
	Qwen3-Embedding _{0.6B}	0.6B	I	45.97±0.75	51.32±0.89 [▲]	72.02±0.80	74.37±1.01 [▲]	-	-	-	-
w/ inst - sentence	MiniLM _{L6-v2}	22M	-	-	-	-	-	42.26±0.08	50.51±0.16 [▲]	71.53±0.02	71.88±0.03 [▲]
	Drama _{1B}	1B	-	-	-	-	-	49.23±1.20	59.40±0.98 [▲]	74.10±2.46	81.60±1.50 [▲]
	E5 _{base-v2}	110M	-	-	-	-	-	54.67±0.02	58.12±0.03[▲]	78.68±0.05	84.57±0.04[▲]
	BGE-M3 (Dense)	0.5B	I	-	-	-	-	46.98±0.10	53.10±0.06 [▲]	71.11±0.14	79.63±0.07 [▲]
	Qwen3-Embedding _{0.6B}	0.6B	I	-	-	-	-	42.14±1.82	42.46±0.12 [▲]	64.87±0.24	64.87±0.23
w/o inst - span	MiniLM _{L6-v2}	22M	-	-	42.00±0.16	-	69.36±0.44	-	-	-	-
	Drama _{1B}	1B	-	-	49.90±0.13	-	74.35±0.15	-	-	-	-
	E5 _{base-v2}	110M	-	-	55.56±0.06	-	84.03±0.05	-	-	-	-
	BGE-M3 (Dense)	0.5B	I	-	56.36±0.08	-	83.33±0.05	-	-	-	-
	Qwen3-Embedding _{0.6B}	0.6B	I	-	49.71±3.11	-	74.29±3.51	-	-	-	-
w/ inst - span	MiniLM _{L6-v2}	22M	-	-	-	-	-	-	48.03±0.08	-	71.83±0.14
	Drama _{1B}	1B	-	-	-	-	-	-	52.23±0.19	-	69.89±0.08
	E5 _{base-v2}	110M	-	-	-	-	-	-	51.35±0.04	-	78.93±0.04
	BGE-M3 (Dense)	0.5B	I	-	-	-	-	-	53.43±0.07	-	79.60±0.07
	Qwen3-Embedding _{0.6B}	0.6B	I	-	-	-	-	-	49.20±0.19	-	63.93±0.06

Table 6: Fine-tuning results on the IdioLink dataset averaged over three random seeds. Values are reported as mean±std. I/N indicates whether the model was post-trained with instructions (§ 5.2); inst denotes instructions. The best result in each configuration is highlighted in bold. The “▲” symbol indicates an increase and “▼” decrease compared to the respective sentence embedding setting, compared to the corresponding sentence embedding setting.

Model	Checkpoint	License	# Param
BGE _{base-en-v1.5} ^a	BAAI/bge-base-en-v1.5	MIT	109M
BGE _{multilingual-gemma2} ^b	BAAI/bge-multilingual-gemma2	Gemma	9B
BGE-M3 (Dense) ^c	BAAI/bge-m3	MIT	0.5B
Contriever ^d	facebook/contriever	CC-BY-NC 4.0	110M
Drama _{1B} ^e	facebook/drama-1b	CC-BY-NC 4.0	1B
E5 _{Mistral} ^f	intfloat/e5-mistral-7b-instruct	MIT	7B
E5 _{base-v2} ^g	intfloat/e5-base-v2	MIT	110M
GritLM _{7B} ^h	GritLM/GritLM-7B	Apache-2.0	7B
GTE _{Qwen2-1.5B} ⁱ	Alibaba-NLP/gte-Qwen2-1.5B-instruct	Apache-2.0	1.5B
GTE _{Qwen2-7B} ^j	Alibaba-NLP/gte-Qwen2-7B-instruct	Apache-2.0	7B
InstructOR _{base} ^k	hkunlp/instructor-base	Apache-2.0	335M
InstructOR _{xl} ^l	hkunlp/instructor-xl	Apache-2.0	1.5B
Linq-Embed-Mistral ^m	Linq-AI-Research/Linq-Embed-Mistral	CC-BY-NC 4.0	7B
Llama-Embed _{Nemotron-8B} ⁿ	nvidia/llama-embed-nemotron-8b	customized-nscl-v1	8B
Lychee-embed ^o	vec-ai/lychee-embed	Apache-2.0	1.5B
Multilingual-E5 _{large-instruct} ^p	intfloat/multilingual-e5-large-instruct	MIT	0.6B
Nomic-v2 ^q	nomic-ai/nomic-embed-text-v2-moe	Apache-2.0	0.5B
Qwen3-Embedding _{0.6B} ^r	Qwen/Qwen3-Embedding-0.6B	Apache-2.0	0.6B
Qwen3-Embedding _{4B} ^s	Qwen/Qwen3-Embedding-4B	Apache-2.0	4B
Qwen3-Embedding _{8B} ^t	Qwen/Qwen3-Embedding-8B	Apache-2.0	8B
SFR-Embedding-Mistral ^u	Salesforce/SFR-Embedding-Mistral	CC-BY-NC 4.0	7B
MiniLM _{L6-v2} ^v	sentence-transformers/all-MiniLM-L6-v2	Apache-2.0	22M
TART-Contriever ^w	orionweller/tart-dual-contriever-msmarco	CC-BY-NC 4.0	110M
Stella _{en_1.5B_v5} ^x	NovaSearch/stella-en-1.5B-v5	MIT	2B
Gemini 2.5 Pro ^y	gemini-2.5-pro	Proprietary	*200B+
Gemini 3 Flash ^z	gemini-3-flash-preview	Proprietary	*20B+
GPT-4o mini ^{aa}	gpt-4o-mini-2024-07-18	Proprietary	*200B+

^a <https://huggingface.co/BAAI/bge-base-en-v1.5>

^b <https://huggingface.co/BAAI/bge-multilingual-gemma2>

^c <https://huggingface.co/BAAI/bge-m3>

^d <https://huggingface.co/facebook/contriever>

^e <https://huggingface.co/facebook/drama-1b>

^f <https://huggingface.co/intfloat/e5-mistral-7b-instruct>

^g <https://huggingface.co/intfloat/e5-base-v2>

^h <https://huggingface.co/GritLM/GritLM-7B>

ⁱ <https://huggingface.co/Alibaba-NLP/gte-Qwen2-1.5B-instruct>

^j <https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct>

^k <https://huggingface.co/hkunlp/instructor-base>

^l <https://huggingface.co/hkunlp/instructor-xl>

^m <https://huggingface.co/Linq-AI-Research/Linq-Embed-Mistral>

ⁿ <https://huggingface.co/nvidia/llama-embed-nemotron-8b>

^o <https://huggingface.co/vec-ai/lychee-embed>

^p <https://huggingface.co/intfloat/multilingual-e5-large-instruct>

^q <https://huggingface.co/nomic-ai/nomic-embed-text-v2-moe>

^r <https://huggingface.co/Qwen/Qwen3-Embedding-0.6B>

^s <https://huggingface.co/Qwen/Qwen3-Embedding-4B>

^t <https://huggingface.co/Qwen/Qwen3-Embedding-8B>

^u <https://huggingface.co/Salesforce/SFR-Embedding-Mistral>

^v <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

^w <https://github.com/facebookresearch/tart/tree/main?tab=readme-ov-file#pre-trained-checkpoints>

^x https://huggingface.co/NovaSearch/stella_en_1.5B_v5

^y <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro>

^z <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-flash>

^{aa} <https://platform.openai.com/docs/models/gpt-4o-mini>

Table 7: Checkpoints used during experiments, their license, and their number of parameters. * = non-official estimation, as this information is not public.