# MoraLink: Bridging Morals and Narrative Fables for Retrieval

**Lior Livyatan[1,2], Kai Golan Hashiloni[1,2], Asif Amar[1,2], Roei Rahamim[1,2], Kfir Bar[1,2]**
[1]Data Science Institute, Reichman University, Herzliya, Israel
[2]Efi Arazi School of Computer Science, Reichman University, Herzliya, Israel
lior.livyatan@post.runi.ac.il, {kai.golanhashiloni,asif.amar,roei.rahamim}@post.runi.ac.il
kfir.bar@runi.ac.il

## Abstract

A fable is a short narrative whose purpose is to convey a moral lesson. While semantically related, morals and fables differ fundamentally in form: morals are abstract and propositional, while fables express the same idea through characters and plot. In this work, we ask how well this conceptual relationship is captured in the latent representations of embedding models. To study this, we introduce **MoraLink**, a benchmark for moral-to-fable retrieval in which models must retrieve the correct fable for a given moral. We find that off-the-shelf embedding models struggle with this task, and that augmenting fable documents with LLM-generated moral-style summaries substantially improves retrieval, highlighting document representation as a major bottleneck. Contrastive fine-tuning further improves performance, and we release MoraLink as a challenging testbed for probing how well embedding models capture the abstraction gap between a moral principle and the narrative that embodies it.

**Query moral:** Do not trust flatterers.

**Relevant fable:** A Fox once saw a Crow fly off with a piece of cheese in its beak and settle on a branch of a tree. "That's for me, as I am a Fox," said Master Reynard, and he walked up to the foot of the tree. "Good-day, Mistress Crow," he cried. How well you are looking to-day: how glossy your feathers; how bright your eye. I feel sure your voice must surpass that of other birds, just as your figure does; let me hear but one song from you that I may greet you as the Queen of Birds. The Crow lifted up her head and began to caw her best, but the moment she opened her mouth the piece of cheese fell to the ground, only to be snapped up by Master Fox.

Figure 1: Example of a moral-to-fable retrieval scenario.

## 1 Introduction

Humans can connect abstract principles to concrete narratives even when the two share almost no lexical overlap. Consider the moral "Do not trust flatterers" and Aesop's fable of the Fox and the Crow (adapted version in Figure 1)—a reader immediately recognizes that this principle is embodied in the story, despite sharing almost no common words. Specifically, the fox tricks the crow into dropping its cheese through insincere praise.

The ability to align abstract principles with concrete narrative realizations is core to semantic understanding and to analogical reasoning.

A natural question is whether large language models (LLMs) internalize similar conceptual links, and whether these representations can be leveraged for tasks such as moral reasoning, story generation, and analogical inference. Understanding the extent to which LLMs capture the relationship between abstract principles and their narrative realizations may thus open new avenues for exploiting the rich conceptual knowledge latent in pretrained models.

Modern dense retrieval models have advanced well beyond keyword matching, but most benchmarks focus on topical or paraphrastic similarity. Retrieval under extreme abstraction mismatch, where a concise abstract query must match documents conveying the same concept through narrative, has received little attention.

In this work, we introduce **MoraLink**, a benchmark for conceptual semantic retrieval under extreme lexical divergence. The task is formulated as *moral-to-fable retrieval*: given a short moral statement, the system must retrieve fables that convey the same moral. MoraLink is built on the MORABLES dataset (Marcuzzo et al., 2025),



*Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 44855–44885
October 24-29, 2026 ©2026 Association for Computational Linguistics

which contains 709 paired fables and morals drawn from Western literary traditions, including the collections of Aesop, Phaedrus, and Abstemius. Fables are much longer than their morals, with minimal lexical overlap, so retrieval cannot rely on surface matching and instead requires conceptual alignment between abstract principles and narrative content.

We evaluate a wide range of embedding models spanning multiple architectures and scales. Such models often rely on lexical overlap rather than deeper semantic equivalence (Wang et al., 2022a; Miao et al., 2023; Shi et al., 2023), and indeed struggle to retrieve the original fables directly despite strong results on conventional benchmarks. To better understand the underlying challenge, we investigate several representation-level interventions: document-side abstraction via LLM-generated moral summaries, instruction-augmented embeddings, and contrastive fine-tuning on human-written and synthetic fables. These interventions substantially improve retrieval in the zero-shot setting, and fine-tuning further narrows the gap, yet a large discrepancy remains, with the best configuration achieving 0.479 MRR@10.

Together, these findings suggest that the primary challenge in MoraLink is semantic-level alignment rather than lexical divergence alone. More broadly, this setting reflects a broader class of retrieval problems in which queries and documents are conceptually related but operate at different levels of abstraction, including analogy-based search, description-based retrieval, and RAG systems that must connect abstract intents with concrete evidence.

Our contributions are as follows:

- We introduce **MoraLink**, a new moral-to-fable retrieval benchmark to evaluate the alignment between abstract moral statements and concrete narrative realizations.
- We show that embedding models struggle to retrieve original fables directly, while LLM-generated moral summaries and task-specific contrastive fine-tuning substantially improve performance.
- We provide a systematic evaluation of modern embedding-based retrieval systems across diverse architectures and scales, revealing a significant persistent gap.

To facilitate further research, we release the MoraLink[1] benchmark and generated summaries under the CC-BY-NC-4.0 license, together with the prompts, moral-grouping guidelines, evaluation code, and reproduction instructions under the Apache 2.0 license.

## 2 Related Work

### 2.1 Moral Narratives as a Challenge for Language Models

Moral reasoning remains a longstanding challenge in natural language processing (NLP), as it requires models to map concrete narratives onto abstract principles. Early benchmarks such as ETHICS (Hendrycks et al., 2020) and Delphi (Jiang et al., 2021) evaluate whether language models can predict human moral judgments over short descriptions of everyday scenarios, rules, or social norms. Related resources, such as Social Chemistry 101 (Forbes et al., 2020) and Moral Stories (Emelin et al., 2021), further examine norms, intentions, actions, and consequences in socially grounded situations. Together, prior results on these benchmarks suggest that language models can approximate moral judgments in constrained settings, but they usually operate over short, explicit situations rather than extended narratives whose lesson must be inferred indirectly.

Related to our work, several studies examine moral meaning in stories. STORAL (Guan et al., 2022) extends this direction by linking stories with moral lessons and enabling moral story generation and evaluation. More recently, MORABLES (Marcuzzo et al., 2025) focuses specifically on fables and abstract moral reasoning, asking models to select the correct moral for a given fable from a set of candidate morals, including adversarial distractors designed to probe robustness. MoraLink extends this setting by replacing closed moral selection with retrieval over an open corpus.

A related challenge appears in figurative language. Recent work shows that LLMs can identify idiomatic expressions through prompting (Hashiloni et al., 2025; Hefetz et al., 2025), but remain vulnerable in challenging contexts (Hashiloni et al., 2026b). IdioLink (Hashiloni et al., 2026a) formulates idiom understanding as a retrieval problem, showing that embedding models struggle to retrieve meaning-equivalent texts when the query

[1]Code: https://github.com/Intellexus-DSI/MoraLink; data: https://huggingface.co/datasets/Intellexus/MoraLink.

and target differ lexically. MoraLink asks a similar question in a different domain: can models connect texts that share the same meaning despite little surface vocabulary overlap? Unlike IdioLink, MoraLink operates at the level of full narratives, linking a short moral principle to the fable that expresses it.

### 2.2 Text Retrieval and Embedding Models

Neural embedding models map queries and documents into a shared vector space, enabling retrieval beyond lexical matching (Reimers and Gurevych, 2019; Izacard et al., 2022). Large-scale benchmarks such as BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023) have standardized evaluation across diverse retrieval and embedding tasks. Modern models such as E5 (Wang et al., 2022b, 2024b), BGE (Xiao et al., 2024), NV-Embed (Lee et al., 2025), and Linq-Embed-Mistral (Choi et al., 2024) improve performance through contrastive training, curated data, instruction tuning, and hard-negative mining. Instruction-following retrievers use natural-language task descriptions to steer models toward task-specific notions of relevance (Su et al., 2023), and FollowIR (Weller et al., 2025) evaluates whether retrieval models can follow such instructions in realistic settings.

Despite these advances, most retrieval benchmarks focus on settings where relevant queries and documents remain semantically close. More recently, BRIGHT (Su et al., 2025) shows that strong retrieval models still struggle when relevance requires deeper reasoning rather than surface semantic matching. MoraLink is closest to this line of work, targeting a specific form of semantic mismatch: a short moral states a general principle, while the relevant fable expresses it indirectly through characters, events, and consequences.

## 3 The MoraLink Dataset

### 3.1 Problem Formulation

MoraLink is a moral-to-fable retrieval benchmark. We denote by $Q$ the set of queries (moral statements) and by $C$ the corpus of documents (fables). Figure 1 shows an example query and document from our dataset. We define $R_q \subseteq C$, the relevance set associated with a query $q \in Q$, consisting of all fables that express the same underlying moral principle as $q$. Given a query $q \in Q$, a retrieval system produces a ranked list over all fables $f \in C$. The system is evaluated according to how highly it ranks the fables in $R_q$.

The benchmark, therefore, tests whether models can retrieve concrete narrative realizations of abstract moral principles from a corpus of fables. Since moral queries and fable documents share almost no lexical overlap, success requires embedding models to internalize the conceptual lesson underlying each fable.

### 3.2 Source Data and Conversion

MoraLink is built from the MORABLES dataset (Marcuzzo et al., 2025), which contains 709 fable-moral pairs drawn from Western fable traditions, including Aesop, Phaedrus, and Abstemius. MORABLES was originally designed as a moral-selection task: given a fable, a model must select the correct moral from a set of five candidate moral statements.

MoraLink, our new dataset, reformulates the MORABLES task as a retrieval problem (§3.1): given a moral, a model must retrieve all fables that express that moral from the fable corpus. A natural construction would pair each fable with its moral in MORABLES, place all fables in the index, and treat the morals as queries. However, distinct fables are frequently paired with semantically equivalent morals that differ only in surface form. While acceptable in MORABLES, this creates a retrieval evaluation problem: correct retrievals may be penalized when morals are not explicitly paired in the annotations. To address this, we introduce a dedicated construction pipeline, described below.

MoraLink poses a significant retrieval challenge: fables average 116.7 words compared to 10.2 for morals, and the mean content-word intersection-over-union is only 0.008. We compute content-word overlap after lower-casing, removing punctuation, and excluding the English stop-words list distributed with scikit-learn.[2]

### 3.3 Moral Deduplication

We observe that MORABLES contains multiple instances of the same moral statement with near-identical phrasing. To avoid querying the system multiple times with functionally identical text, which would artificially inflate dataset size and evaluation difficulty, we deduplicate these morals and reduce the query set $Q$ to unique representa-

[2] https://scikit-learn.org/stable/modules/feature_extraction.html#using-stop-words

tives. During this process, we also correct occasional typos and minor formatting inconsistencies.

We first identify morals with identical normalized text after lowercasing, removing punctuation, and collapsing whitespaces. We then identify near duplicates that differ only in minor wording, spelling, or punctuation. For example, *Enemies' promises are made to be broken* and *Enemies promises are made to be broken*. To identify such candidate pairs, we compute pairwise cosine similarity scores between all morals using $\text{BGE}_{\text{large-en-v1.5}}$ and consider pairs with similarity above 0.95. These candidates are then manually reviewed together by two of the authors to decide whether they should be merged.

This deduplication process reduces the query set $Q$ from 709 to 668 unique morals. Each surviving moral inherits the fable associations of all the morals it subsumes, extending its relevance group to potentially include more than one fable.

### 3.4 Moral Grouping

Some fables in MORABLES convey the same underlying moral lesson, even though they are phrased differently. For example, *Little by little does the trick* and *Slow and steady wins the race*. To ensure that a model retrieving any fable expressing a given moral receives appropriate credit, we cluster semantically equivalent morals and treat all fables paired with morals in the same cluster as relevant. Formally, a cluster is defined as $K_i = (Q_i, C_i)$, where $Q_i \subset Q$ represents the set of equivalent moral queries and $C_i \subset C$ represents their combined relevance set of fables.

**Semantic equivalents.** Identifying semantic equivalents requires careful judgment about when two (or more) morals convey the same principle rather than merely address related themes such as greed or trust. An additional challenge arises when two morals convey similar lessons but should not necessarily be grouped together for our purposes, making it difficult to draw the line between equivalence and distinction. For example, *Do not blame others* and *Do not blame others for the circumstances of their upbringing* are closely related, but the latter narrows the general principle to a specific causal context; hence, they should not be merged.

To address this subtle task, we assemble a committee of three annotators (details in Appendix A) who review the full corpus of morals to identify semantic equivalences for grouping. The annotation process follows an iterative approach, in which the committee repeatedly reviews candidate clusters until all assignments stabilize.

In each iteration, we incorporate Claude Sonnet 4.5 as an in-the-loop assistant, prompted with the full moral set, the fable corpus, and the pairwise similarity scores computed as described in §3.3. We begin by placing each moral in its own cluster. Claude is then prompted (prompt in Appendix C) to suggest which clusters should be merged into semantic equivalence groups and which should be separated, along with justifications for each decision. The three annotators review every proposed grouping separately and jointly, including decisions to merge or not merge, and reach a shared decision through discussion. The process produces a single agreed grouping rather than separate annotations. Details about the annotators are provided in Appendix A.

This grouping process yields 558 moral clusters covering the 668 unique morals from the deduplication phase. Morals within a cluster share the same relevance set for retrieval, defined as the union of fables associated with any moral in that cluster.

### 3.5 Final Benchmark

Overall, MoraLink contains 668 unique moral queries, 709 fables, and 558 moral clusters. Of these, 469 queries have a single relevant fable, while 199 queries have two or more relevant fables, with an average of 1.62 relevant fables per query. The full distribution of relevance-set sizes is shown in Figure 3 (Appendix B).

## 4 Benchmarking Embedding Models

We evaluate embedding models in zero-shot (§ 4.2) and fine-tuning (§ 4.3) settings, on our MoraLink benchmark. We generate embeddings with and without task-specific instructions, as well as different configurations for query and document preprocessing and augmentation (§ 4.1).

### 4.1 Retrieval Configurations

**Instructions.** We experiment with two query formulation configurations: with and without task instructions provided alongside the query. We compare several query instruction formats and discuss their types, full implementation, and optimization details in Appendix E. Following best practice, we provide instructions for queries only, unless specific models require them for documents as well.

In the main experiments, + *instr.* denotes that instructions are given alongside the query.

**Document representation.** We consider three document representation strategies for each fable $f$ prior to embedding. (i) The `fable-only` representation passes the original fable text as-is to the embedding model. (ii) The `summary-only` representation replaces the fable with an LLM-generated moral-style summary $s$. (iii) The `summary+fable` representation concatenates the original fable with a generated summary, denoted as $[s \oplus f]$. We use $[s \oplus f]$ order in our main experiments, and report the reversed order $[f \oplus s]$ as a layout ablation in Appendix F. For summary-based representations, we generate moral-style summaries using Gemini 3 Flash, a strong yet cost-efficient LLM suitable for large-scale document processing. We additionally compare against Gemma 4 variants (E2B, E4B, 31B) as a summary-generator ablation; the smaller open-source E2B and E4B models are particularly relevant for retrieval tasks, where corpus-scale summarization typically requires models under 9B parameters to remain practical. We use the 4-bit quantized version of the models to run comfortably on commonly available hardware.

All summaries are generated once and reused across zero-shot and fine-tuned evaluations to ensure that differences between retrieval models reflect document representation rather than summary variation. For the main experiments, we use the summaries generated by Gemini 3 Flash. Summary-generation prompts and results are provided in Appendix F; detailed results for smaller summary generators are provided in Appendix G. The summary+fable representation is motivated by an oracle experiment in which we pair each fable with its ground-truth moral, yielding strong performance; full results are reported in Appendix I.

**Indexing and similarity calculation.** For each configuration, fables are embedded under the selected document representation method, and morals are embedded with or without query instructions, as determined by experiments. Fables are ranked by cosine similarity between the query and document embeddings.

### 4.2 Zero-shot Setting

We evaluate all retrieval configurations in a zero-shot setting. This setting tests whether off-the-shelf embedding models can align abstract moral statements with the narrative fables that express them.

### 4.3 Fine-Tuning Settings

We evaluate whether embedding models can be adapted to the moral-to-narrative alignment required by MoraLink through fine-tuning on real-life as well as synthetic datasets of similar tasks. We use an in-batch contrastive objective inspired by dense retrieval and recent LLM-generated embedding-data frameworks (Karpukhin et al., 2020; Lee et al., 2024). Full training hyperparameters and model-selection details are reported in Appendix M; software, hardware, and runtime details are reported in Appendix N.

**Training data.** We use two fine-tuning sources: STORAL (Guan et al., 2022), a dataset of real human-authored stories with moral statements, and TF1-EN-3M (Nadas et al., 2025), a synthetic corpus of LLM-generated fables. Both require preprocessing to avoid lexical leakage with MoraLink. After filtering, STORAL yields 1,595 story–moral pairs (1,530 unique morals), while TF1-EN-3M yields 1,900 low-overlap fables across 95 morals. Full preprocessing details are in Appendix L.

For every fine-tuning experiment, we reserve 10% for validation. Splits are constructed at the moral level, ensuring that examples sharing the same moral remain in the same split.

Inspired by Gecko (Lee et al., 2024), our contrastive objective compares each moral query with both fable documents and other moral queries in the mini-batch. We construct training tuples of the form:

$$\mathcal{D} = \{(q_i, d_i^+, y_i)\}_{i=1}^{N},$$

where $q_i$ is a moral query, $d_i^+$ is a fable document in the relevance group of $q_i$, and $y_i$ is a cluster label (§ 3.4) used to mask false negatives during training. Since the loss function uses in-batch negatives, documents from other examples in the same mini-batch are treated as negatives by default. However, multiple examples may be in the same moral cluster, meaning their positive documents are mutually valid positives rather than negatives. The label $y_i$ identifies such cases and excludes them from the negative set. In other words, the negatives are constructed within each mini-batch: for a query moral $q_i$, documents paired with examples $j$ such that $y_j \neq y_i$ serve as document negatives, and other morals with $y_j \neq y_i$ serve as moral-query negatives. Examples with the same label are excluded from the negative set, since they express the same moral and may be valid positives.

**Training objective.** We fine-tune the embedding models with an in-batch contrastive objective.

For a mini-batch

$$\mathcal{B} = \{(q_i, d_i^+, y_i)\}_{i=1}^{B},$$

$q_i$ denotes a moral query, $d_i^+$ denotes its paired positive document, and $y_i$ denotes the moral label used for false-negative masking. The model maps each moral and document to a normalized embedding vector. We define $s(x, z)$ as the cosine similarity between the embeddings of texts $x$ and $z$.

For each query moral $q_i$, the positive target is its paired document $d_i^+$. All other documents in the batch are treated as document negatives, except documents whose label matches $y_i$:

$$\mathcal{D}_i = \{d_i^+\} \cup \{d_j^+ : j \neq i,\ y_j \neq y_i\}. \quad (1)$$

We also include other moral queries with different labels during training to encourage the model to keep distinct lessons apart:

$$\mathcal{Q}_i = \{q_j : j \neq i,\ y_j \neq y_i\} \quad (2)$$

These moral examples provide an additional training signal. At test time, retrieval is still performed only over fable documents. The per-example loss is:

$$\mathcal{L}_i = -\log \frac{\exp(s(q_i, d_i^+)/\tau)}{Z_i} \quad (3)$$

with

$$Z_i = \sum_{d \in \mathcal{D}_i} \exp(s(q_i, d)/\tau) + \sum_{q \in \mathcal{Q}_i} \exp(s(q_i, q)/\tau) \quad (4)$$

Here, $\tau$ is the contrastive temperature, set to $0.05$ in all fine-tuning experiments. The label condition $y_j \neq y_i$ prevents examples with the same moral from being used as negatives, since their documents may also be valid positives.

**Configurations.** We fine-tune embedding models on the vanilla configuration, that is, `fable-only` documents with no instruction templates and no summaries. We experiment with four training dataset combinations: (1) STORAL only, (2) TF1-EN-3M only, (3) a curriculum approach in which the model is trained on TF1-EN-3M and then on STORAL, extending training to the combined size of both, and (4) a shuffled mix of both datasets. Each fine-tuned model is evaluated on all retrieval configurations used in the zero-shot setting.

## 5 Experimental Settings

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

### 5.1 Evaluation Metrics

MoraLink is evaluated using three standard, complementary retrieval metrics (Manning et al., 2008; Thakur et al., 2021): (i) MRR@10, which measures how early the first relevant fable appears in the top-10 ranking, assigning a score of 0 if no relevant fable appears in that range; (ii) MAP@10, which rewards models for ranking multiple relevant fables highly when several valid fables express the same moral; (iii) Recall@10, which measures coverage of the relevant fables retrieved within the top-10 results. More details and examples are provided in Appendix D.

### 5.2 Models

We use BM25 (Robertson and Zaragoza, 2009) as a strong lexical baseline (details in Appendix J).

We evaluate 24 text embedding models with parameter counts ranging from 22M to 8B. The full list of checkpoints and parameter sizes is provided in Table 17 (Appendix P). Following Weller et al. (2025), we group models into two categories:

**No instruction in training.** We evaluate encoder-style embedding models that are not post-trained on instruction-formatted retrieval queries, including $\text{MiniLM}_{\text{L6-v2}}$ (Wang et al., 2021), Contriever (Izacard et al., 2022), E5 (Wang et al., 2024b). We additionally evaluate decoder-based embedding models, such as DRAMA (Ma et al., 2025).

**Instruction in training.** We evaluate models trained or post-trained with retrieval instructions, ranging from fixed task prompts to fully instruction-following embedding models. These include the encoder models: Contriever-based TART (Asai et al., 2023), Multilingual-E5 (Wang et al., 2024b), Nomic-Embed-v2 (Nussbaum and Duderstadt, 2025), and BGE models (Xiao et al., 2024), and their multilingual version BGE-M3 (Chen et al., 2024) also fall into this category, although they are trained with only one instruction for each broad task (retrieval, clustering, etc.) Decoder models include InstructOR (Su et al., 2023), E5-Mistral (Wang et al., 2024a), SFR-Embedding-Mistral,[3] Linq-Embed-Mistral (Choi et al., 2024), Lychee-Embed (Zhang et al., 2025a), $\text{Llama-Embed}_{\text{Nemotron-8B}}$ (Babakhin et al., 2025), and the Qwen3 embedding models (Zhang et al., 2025b).

[3] https://huggingface.co/Salesforce/SFR-Embedding-Mistral

| Model | Param | fable-only | | | summary-only | | | summary+fable | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MRR@10 | MAP@10 | R@10 | MRR@10 | MAP@10 | R@10 | MRR@10 | MAP@10 | R@10 |
| BM25 | – | 0.073 | 0.057 | 0.124 | 0.196 | 0.143 | 0.231 | 0.193 | 0.146 | 0.261 |
| Gemini-3.1-Flash-Lite | N/A | 0.156 | 0.110 | 0.217 | 0.383 | 0.332 | 0.526 | 0.210 | 0.163 | 0.265 |
| GPT-5.4-Nano | N/A | 0.102 | 0.077 | 0.141 | 0.314 | 0.258 | 0.365 | 0.246 | 0.196 | 0.290 |
| $\text{MiniLM}_{\text{L6-v2}}$ | 22M | 0.095 | 0.068 | 0.142 | 0.287 | 0.222 | 0.364 | 0.220 | 0.167 | 0.292 |
| + instr. | | 0.067 | 0.046 | 0.102 | 0.214 | 0.158 | 0.258 | 0.151 | 0.108 | 0.199 |
| $\text{Qwen3-Embedding}_{\text{0.6B}}$ | 596M | 0.123 | 0.084 | 0.164 | 0.337 | 0.266 | 0.424 | 0.325 | 0.251 | 0.421 |
| + instr. | | 0.135 | 0.089 | 0.178 | 0.349 | 0.275 | 0.435 | 0.355 | 0.278 | 0.463 |
| $\text{InstructOR}_{\text{xl}}$ | 1.24B | 0.146 | 0.102 | 0.198 | 0.398 | 0.326 | 0.496 | 0.381 | 0.312 | 0.493 |
| + instr. | | 0.089 | 0.054 | 0.114 | 0.283 | 0.213 | 0.352 | 0.234 | 0.169 | 0.305 |
| $\text{Qwen3-Embedding}_{\text{4B}}$ | 4.02B | 0.181 | 0.126 | 0.230 | 0.378 | 0.315 | 0.494 | 0.358 | 0.290 | 0.468 |
| + instr. | | 0.212 | 0.145 | 0.269 | 0.388 | 0.321 | 0.510 | 0.401 | 0.331 | 0.534 |
| Linq-Embed-Mistral | 7.11B | 0.239 | 0.171 | 0.292 | 0.395 | 0.322 | 0.488 | 0.385 | 0.317 | 0.510 |
| + instr. | | **0.268** | **0.196** | **0.341** | **0.406** | **0.336** | **0.531** | 0.408 | 0.343 | 0.550 |
| $\text{Qwen3-Embedding}_{\text{8B}}$ | 7.57B | 0.198 | 0.136 | 0.273 | 0.392 | 0.320 | 0.499 | 0.410 | 0.339 | 0.517 |
| + instr. | | 0.229 | 0.165 | 0.317 | 0.404 | 0.335 | 0.528 | **0.428** | **0.360** | **0.565** |

Table 1: Zero-shot results on MoraLink for representative embedding models, with BM25 and direct LLM ranking baselines. Embedding models are evaluated with and without query instructions (+ *instr.*); BM25 and direct LLM rows are single-setting baselines. Moral summaries are generated by Gemini 3 Flash. Best values are bolded.

**Fine-tuned models.** We select a representative group of five models for fine-tuning, spanning different parameter scales, architectures and instruction-aware or not: $\text{MiniLM}_{\text{L6-v2}}$ (22M parameters), BGE-M3 (568M), $\text{Qwen3-Embedding}_{\text{0.6B}}$ (596M), $\text{InstructOR}_{\text{xl}}$ (1.24B), and $\text{Qwen3-Embedding}_{\text{4B}}$ (4.02B).

**Direct LLM ranking.** We also evaluate direct LLM ranking as a complementary baseline, following work that prompts LLMs to rank candidate documents for a given query (Sun et al., 2023; Ma et al., 2023; Qin et al., 2024). Unlike embedding models, LLMs receive all the fables with the moral query and are instructed to return the 10 most relevant fables. We evaluate cost-efficient models GPT-5.4 Nano and Gemini 3.1 Flash-Lite, as the full corpus in the prompt incurs substantially higher inference costs. Implementation details and the ranking prompt are provided in Appendix H.

## 6 Results and Discussion

### 6.1 Zero-shot Results

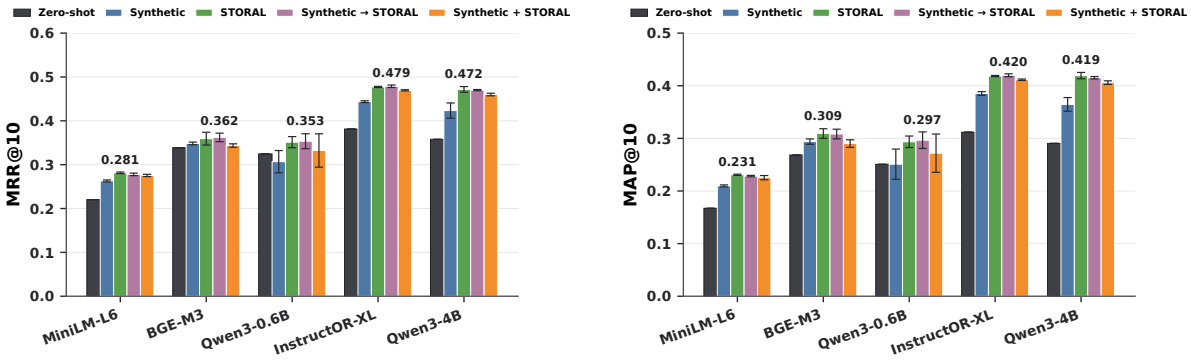
Table 1 summarizes zero-shot results for six representative embedding models spanning different sizes. Full results for all 24 models are reported in Table 11 (Appendix K). Unless stated otherwise, all averages reported in this section are calculated over the full set of 24 embedding models.

We compare three document representations: the original fable text, a Gemini 3 Flash-generated moral summary, and the summary prepended to the fable. For each model, we report results with and without query instructions.

**First, vanilla retrieval is difficult.** Across all models, embedding the fable text naively yields low performance. Even the best `fable-only` configuration, Linq-Embed-Mistral with instructions, reaches only 0.268 MRR@10, 0.196 MAP@10, and 0.341 Recall@10. This supports the central difficulty of MoraLink and encourages the community to explore richer conceptual representations that better capture moral-to-narrative similarity.

**Second, document-side abstraction substantially improves retrieval.** Replacing the fable with a Gemini 3 Flash-generated moral summary (`summary-only`) yields large improvements over `fable-only` representation: +0.210 MRR@10, +0.179 MAP@10, and +0.249 Recall@10 on average without instructions. For example, $\text{InstructOR}_{\text{xl}}$ improves from 0.146 to 0.398 MRR@10, while BGE-en-ICL improves from 0.174 to 0.400.

Generated summaries also share more wording with the queries. Average word overlap is 0.007 for the original fables and 0.076 for the Gemini 3 Flash summaries used in the main experiments; across all 12 summary collections, it ranges from 0.012 to 0.103. Across 159 matched retrieval settings, larger increases in word overlap are associated with larger improvements in rank, with a median Spearman correlation of 0.344. This shows that closer wording helps explain the improvement, although it does not show that word overlap alone caused it. Full results of this analysis are reported




(a) MRR@10

(b) MAP@10

Figure 2: Fine-tuning results: *Synthetic* → *STORAL* is curriculum learning, and *Synthetic* + *STORAL* is mixed training. Inference-time uses the `summary+fable` approach without instructions. Recall@10 is reported in Appendix M.

in Appendix K.2.

Prepending the summary to the fable (`summary+fable`) also improves over `fable-only` (+0.183 MRR@10 on average), but usually underperforms `summary-only`. This difference may reflect document length, how models handle longer inputs, or the greater wording similarity between the summaries and the queries. In the no-instruction setting, the fable-first and summary-first layouts contain exactly the same text in a different order, but reach average MRR@10 scores of 0.275 and 0.306, respectively. Both also outperform the shorter fable-only representation at 0.124, suggesting that document length alone does not explain this pattern (Appendix F). The strongest exception is Qwen3-Embedding$_{8B}$ with instructions, where `summary+fable` achieves the best across-the-board zero-shot result: 0.428 MRR@10, 0.360 MAP@10, and 0.565 Recall@10.

These gains highlight the importance of document-side abstraction: retrieval becomes substantially easier when the fable's moral is stated directly rather than left implicit in the narrative.

**Queries with one or several relevant fables.** We report results separately for the 469 queries with one relevant fable and the 199 queries with several relevant fables in Appendix K.1. Queries with multiple fables generally achieve higher MRR@10 because they provide more opportunities for a first correct result. However, their Recall@10 is normally lower, showing that finding one relevant fable is easier than recovering a larger set. For example, for queries with one fable, Linq-Embed-Mistral with instructions improves from an MRR@10 of 0.208 with `fable-only` to 0.332 with `summary-only` and 0.360 when using `summary+fable`. For queries with several fables, the corresponding scores rise from 0.407 to 0.581 with `summary-only`, while `summary+fable` yields 0.521.

**Third, the effect of query instructions varies across models.** Query instructions improves MRR@10 for 9 models with `fable-only`, 8 with `summary-only`, and 11 with `summary+fable`, including Linq-Embed-Mistral and Qwen3-Embedding$_{8B}$. However, it reduces performance for the remaining models, resulting in average changes of −0.011, −0.033, and −0.038 MRR@10 across the three configurations, respectively. Thus, query instructions can help particular models, but do not provide a consistent benefit across the full model set. Overall, zero-shot gains in MoraLink stem primarily from changes in document representation rather than from query format.

**Fourth, lexical and direct LLM baselines remain limited.** BM25 is weak on raw fables (0.073 MRR@10) but improves with generated moral summaries (0.196), showing that even sparse retrieval benefits from document-side abstraction. Direct LLM ranking follows the same pattern: GPT-5.4 Nano reaches 0.102 on raw fables, 0.314 on generated summaries, and 0.246 with `summary+fable`, while Gemini 3.1 Flash-Lite reaches 0.156, 0.383, and 0.210, respectively. Overall, these baselines reinforce the main zero-shot finding: document representation is a major bottleneck across different ranking methods.

**Fifth, generated summaries remain far from the original human-written morals.** As an upper-bound comparison, we replace each candidate fable with its original MORABLES moral. Using Linq-Embed-Mistral with instructions, MRR@10 is 0.268 for the original fables, 0.406 for the Gemini-generated summaries, and 1.000 for the original morals, where a query equals its document. Full results are reported in Appendix I. The original morals are not available in the actual retrieval task, but this comparison shows that generated summaries do not fully recover the moral information contained in the human-written morals. It is also important to note that such human-written morals are a potentially subjective ground truth and hence should be taken only as a reference point of comparison.

### 6.2 Fine-tuning Results

Figure 2 shows fine-tuning results compared to zero-shot under the `summary+fable` inference-time configuration without instructions; Recall@10 for this configuration and results for the remaining configurations are reported in Appendix M. Each experiment is repeated with three random seeds, and the results are reported as the mean and standard deviation.

Fine-tuning consistently improves over zero-shot retrieval across all dataset combinations. Training on the synthetic TF1-EN-3M data alone already yields meaningful gains, while fine-tuning on real STORAL examples, separately, produces larger improvements across all metrics, suggesting that real moral-story pairs provide a more useful adaptation signal. The strongest individual result is obtained by InstructOR$_{xl}$, which achieves the highest MRR@10 of 0.479 with 0.003 standard deviation.

The synthetic→STORAL curriculum, first training on TF1-EN-3M, then on STORAL, performs competitively with direct STORAL fine-tuning, with slightly higher early-ranking scores (MRR@10, MAP@10) at the cost of a small drop in Recall@10. Simply mixing synthetic and real examples is outperformed by both STORAL-only training and curriculum training, suggesting that the order of exposure and, potentially, the quality or resemblance to MoraLink matter more than the volume of training data alone.

Fine-tuning shifts the optimal document representation: zero-shot retrieval performs best with `summary-only`, whereas fine-tuned models prefer `summary+fable`. This makes sense: fine-tuning on full story documents brings fables and morals closer in the embedding space, and since summaries already improve representation, combining both gives the model the strongest signal.

## 7 Conclusion

We introduced **MoraLink**, a benchmark for moral-to-fable retrieval that probes whether embedding models can bridge the abstraction gap between a concise moral principle and the narrative that expresses it. Our experiments show that off-the-shelf embedding models struggle considerably to retrieve original fables directly, despite strong performance on conventional retrieval benchmarks. A major bottleneck lies in document representation: encoding LLM-generated moral-like summaries instead of the fables substantially improves zero-shot performance. Fine-tuned models, however, perform best when the original fable is retained alongside the summary. Query-side instructions yield smaller, less consistent gains. MoraLink provides a useful testbed for developing retrieval systems capable of reasoning across semantic levels, with broader relevance to analogy-based search, description-based retrieval, and other settings where queries and documents operate at different levels of abstraction.

## Limitations

MoraLink is derived from MORABLES (Marcuzzo et al., 2025), which draws exclusively from Western fable traditions (Aesop, Phaedrus, etc.) and is available only in English. The moral principles encoded in these fables reflect Western cultural values and philosophical traditions, and may not generalize to moral reasoning frameworks from other cultural contexts. This limitation is discussed further in the Ethics Statement. Future work should extend this benchmark to diverse traditions, multiple languages, and cross-lingual retrieval scenarios.

The benchmark contains 709 fables and 668 unique moral queries, which is smaller than large-scale retrieval benchmarks such as BEIR (Thakur et al., 2021) or MTEB (Muennighoff et al., 2023), limiting statistical power for fine-grained comparisons. The average relevance set size is small (1.62 fables per query).

While we carefully preprocess the data in the fine-tuning experiments to avoid leakage, complex overlapping cases may remain. We do not evaluate whether task-specific fine-tuning degrades general-

purpose retrieval capabilities on other benchmarks, and leave this analysis to future work.

Our evaluation treats MORABLES annotations as ground truth, but moral-narrative alignment is inherently interpretive and may vary across readers and cultural contexts. The same holds for our moral grouping stage (§ 3.4), where determining semantic equivalence between morals involves subjective judgments about whether distinct phrasings express the same underlying principle.

Despite these limitations, MoraLink reveals substantial gaps in current retrieval systems and opens promising directions for future research on conceptual-level semantic understanding.

## Ethics Statement

We use publicly available datasets and models in accordance with their intended uses, as detailed by the respective publishers and under their licenses; details are provided in Table 17 in Appendix P and Table 15 in Appendix O. No personally identifiable information or offensive data are processed, and annotators ensure that none exists in the released dataset.

As noted above, MoraLink inherits its cultural focus from its underlying source material. Consider this cultural provenance when interpreting the benchmark and its results, particularly the normative assumptions represented in the data.

LLM-generated summaries used in our experiments may introduce annotation artifacts or normative biases; however, they serve only as document representations for retrieval evaluation and do not contribute to the ground-truth annotations.

Following the inherited MORABLES license, we release the generated summaries under CC-BY-NC-4.0. We release the prompts, generation settings, provenance information, and evaluation code under Apache 2.0. The released summaries and evaluation code allow researchers to rerun the retrieval experiments without generating the summaries again or using a paid API. The prompts, generation settings, and provenance information document how the summaries were produced. The filtered TF1-EN-3M subset used for synthetic fine-tuning is released under the dataset's original MIT license. Our work is intended for research purposes only, and we identify no foreseeable risks from its use.

## 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

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.

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.

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.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Jian Guan, Ziqi Liu, and Minlie Huang. 2022. A corpus for understanding and generating moral stories. In *Proceedings of the 2022 Conference of the North*

*American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5069–5087, Seattle, United States. Association for Computational Linguistics.

Kai Golan Hashiloni, Daniel Fadlon, Lior Livyatan, Ofri Hefetz, Jiahuan Pei, and Kfir Bar. 2026a. IdioLink: Retrieving meaning beyond words across idiomatic and literal expressions. *arXiv preprint arXiv:2605.22247*.

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 Mannor, Bar Cohen, and Kfir Bar. 2026b. 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.

Ofri Hefetz, Kai Golan Hashiloni, Alon Mannor, 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.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning AI with shared human values. *arXiv preprint arXiv:2008.02275*.

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.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2021. Can machines learn morality? the Delphi experiment. *Preprint*, arXiv:2110.07574.

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.

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 *International Conference on Learning Representations*, volume 2025, pages 79310–79333.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Vijay Chidambaram, Wenhao Zhou, Claudiu Musat, Gang Chen, Wei Li, Yizhou Zhou, Quoc V. Le, Jun He, and Xinlei Chen. 2024. Gecko: Versatile text embeddings distilled from large language models. *Preprint*, arXiv:2403.20327.

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.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England.

Matteo Marcuzzo, Alessandro Zangari, Andrea Albarelli, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2025. Morables: A benchmark for assessing abstract moral reasoning in LLMs with fables. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27727–27751, Suzhou, China. 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, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Mihai Nadas, Laura Diosan, Andrei Piscoran, and Andreea Tomescu. 2025. TF1-EN-3M: Three million synthetic moral fables for training small, open language models. *arXiv preprint arXiv:2504.20605*.

Zach Nussbaum and Brandon Duderstadt. 2025. Training sparse mixture of experts text embedding models. *Preprint*, arXiv:2502.07972.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu

Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. 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.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

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.

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.

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 O. Arik, Danqi Chen, and Tao Yu. 2025. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. *Preprint*, arXiv:2407.12883.

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 *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 14918–14937.

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

Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao. 2022a. SNCSE: 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. 2022b. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. 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. 2024b. 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 pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

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.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Xin Zhang, Yanzhao Zhang, Wen Xie, Dingkun Long, Mingxin Li, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025a. 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. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

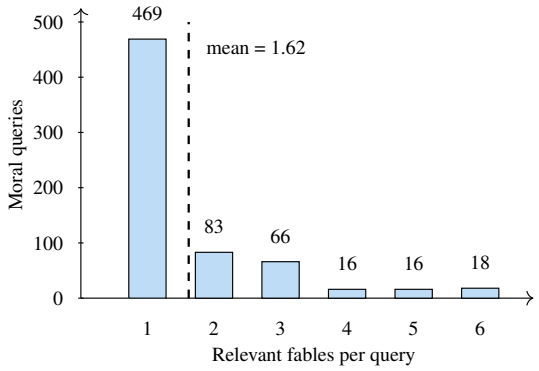



Figure 3: Distribution of relevance-set sizes per query. The dashed line marks the mean.

## A Annotators Information

The data construction process was performed by three authors of this paper, all with academic backgrounds in scientific research: one PhD researcher, one PhD candidate, and a graduate student, all of whom are proficient in English. The authors performed this task as part of their contribution to the paper's development and were therefore not paid specifically for the annotation.

## B Relevance-set Sizes Distribution

Figure 3 shows the distribution of relevance-set sizes across all 668 moral queries in MoraLink. The majority of queries (469, approximately 70%) are singletons with exactly one relevant fable. Among queries with multiple relevant fables, the distribution is long-tailed: 83 queries have two relevant fables, 66 have three, and the remaining queries have four to six. The mean relevance-set size is 1.62 fables per query, indicated by the dashed line in the figure.

## C Semantic Equivalents Prompt

To identify semantically equivalent morals, we prompt an LLM to suggest cluster merges and splits based on whether morals express the same underlying principle. The model is provided with (i) the complete set of moral statements and fable corpus, (ii) pairwise cosine similarity scores computed using $\text{BGE}_{\text{large-en-v1.5}}$ embeddings (§3.3), (iii) the current clustering state. The full prompt is shown in Figure 4.

**Output structure.** The model returns a JSON object containing: (i) `clusters`: An array of moral groups, where each cluster contains semantically equivalent morals (ii) `justification`: A brief explanation for each proposed cluster (iii) `singleton_morals`: Morals that should remain in their own cluster

## D Evaluation Metrics

We evaluate retrieval quality using MRR@10, MAP@10, and Recall@10, three standard information retrieval metrics that capture complementary aspects of ranking quality under the MoraLink setting, where a moral query may correspond to multiple relevant fables.

**Mean Reciprocal Rank at 10 (MRR@10).** Measures how early the first relevant fable appears among the top-10 retrieved results. For a query $q$, the Reciprocal Rank (RR) is defined as:

$$\text{RR}(q) = \frac{1}{\text{rank}_q},$$

where $\text{rank}_q$ is the rank position of the first relevant fable within the top 10; if no relevant fable appears in the top 10, $\text{RR}(q) = 0$. The final MRR@10 score is the mean reciprocal rank across all queries.

MRR@10 emphasizes early precision, rewarding systems that place at least one correct narrative realization near the top of the ranking.

**Example.** Suppose the first relevant fable for a query appears at rank 4. The reciprocal rank is:

$$\text{RR} = \frac{1}{4} = 0.25.$$

**Mean Average Precision at 10 (MAP@10).** Evaluates how well systems rank multiple relevant fables within the top-10 retrieved results. For a query $q$, Average Precision (AP) at 10 is defined as:

$$\text{AP@10}(q) = \frac{1}{|R_q|} \sum_{k=1}^{10} P(k) \cdot \text{rel}(k),$$

where $R_q$ is the set of relevant fables for query $q$, $P(k)$ is the precision at rank $k$, and $\text{rel}(k)$ indicates whether the document at rank $k$ is relevant. Note that relevance is binary, meaning $\text{rel}(k) \in \{0, 1\}$.

MAP@10 rewards systems that retrieve several relevant fables early in the ranking, which is particularly important in MoraLink because multiple fables may express the same moral lesson.

**Example.** Suppose a query has three relevant fables retrieved at ranks 1, 3, and 7. The precision values at these ranks are:

$$P@1 = \frac{1}{1} = 1.0, \quad P@3 = \frac{2}{3} \approx 0.67,$$
$$P@7 = \frac{3}{7} \approx 0.43.$$

The AP@10 score is therefore:

$$\text{AP@10} = \frac{1.0 + 0.67 + 0.43}{3} \approx 0.70.$$

**Recall@10.** Measures the fraction of relevant fables retrieved within the top-10 results:

$$\text{Recall@10}(q) = \frac{|R_q^{10}|}{|R_q|},$$

where $R_q^{10}$ denotes the relevant fables retrieved within the top 10, and $|R_q|$ is the total number of relevant fables for query $q$.

Recall@10 measures coverage of the relevance set regardless of the ordering among relevant fables.

**Example.** Suppose a query has four relevant fables in total, and the system retrieves three of them within the top-10 results. The Recall@10 score is:

$$\text{Recall@10} = \frac{3}{4} = 0.75.$$

**Why @10?** We use a cutoff of 10 because each query is ranked against all 709 candidate fables, while the average relevance set contains only 1.62 relevant fables per query. In this setting, retrieving the relevant fables within the top results already requires strong semantic alignment, whereas larger cutoffs would place less emphasis on practically useful ranking quality.

# E Retrieval Instructions

## E.1 Model-specific Instruction Interfaces

Different embedding models expose different mechanisms for incorporating retrieval instructions. We implement all zero-shot embedding runs using the SentenceTransformers interface and normalize all embeddings before computing dot-product similarities. Instruction formatting is applied before encoding, using one of two mechanisms.

First, for most models, the instruction or retrieval prefix is concatenated directly to the input text. For query-side formatting, the encoded string has the form:

```
{query_prefix}{query}
```

where `{query}` is the moral statement. When a model also requires document-side formatting, documents are encoded as:

```
{doc_prefix}{document}
```

where `{document}` is the selected fable representation, such as the raw fable, a generated moral summary, or their concatenation.

Second, for models that expose named SentenceTransformers prompts, we use the model's prompt interface rather than manually concatenating a textual prefix. Concretely, the query encoder is called with the appropriate `prompt_name`; for example, DRAMA-1B uses `prompt_name="query"`. These named prompts are applied only on the query side; document texts are encoded without a prompt unless a document prefix is specified by the model's default format.

The `default` variant follows each model family's documented retrieval format when available. For example, E5 encoders use `query:` for moral queries and `passage:` for documents, while Nomic-v2 uses `search_query:` and `search_document:`. Models without a documented default retrieval format use raw text on both sides under `default`. Table 2 summarizes the model-specific default formats used in our experiments.

## E.2 Instructions Optimization

To understand how task framing affects retrieval performance, we evaluate up to four retrieval instruction formats across all embedding models.

(i) **No instruction**: the query is passed as raw text, without any additional framing.

(ii) **Default format**: the query is provided alongside the specific model's recommended retrieval instruction format (Table 2).

(iii) **General retrieval**: the query is provided alongside a task-agnostic instruction, prompting the model to retrieve the passage most relevant to the query.

(iv) **Moral-specific**: the query is provided alongside an instruction explicitly framing the query as a moral statement.

Following best practices, we provide instructions for queries only; however, some models also require document prefixes (e.g., "`passage:`" for E5 encoders, "`search_document:`" for Nomic). The `no_instr`, `general`, and `moral_specific` variants are applied across all models by prepending the corresponding text to the moral query. The `default` variant uses each model family's documented retrieval format when available; for models without such a format, it is equivalent to raw input. Table 3 summarizes these variants and their application to query and document sides.

The fixed `general` instruction template is:

```
Instruct: Given a text, retrieve the
most relevant passage that answers the
```

| Model / family | Default query prefix | Default document prefix |
|---|---|---|
| E5 encoders | `query:` | `passage:` |
| Nomic-v2 | `search_query:` | `search_document:` |
| BGE en v1.5 | `Represent this sentence for searching relevant passages:` | none |
| BGE-en-ICL | `Given a moral statement, retrieve the fable that best conveys this moral.` | none |
| InstructOR | `Represent the moral statement for retrieving relevant fables:` | none |
| TART-Contriever | `Retrieve a fable that illustrates the following moral [SEP]` | none |
| DRAMA-1B | SentenceTransformers `prompt_name="query"` | none |

*Models not listed above use raw text on both sides under* `default`.

Table 2: Model-specific formats used by the `default` variant.

| Variant | Purpose | Query | Document |
|---|---|---|---|
| `no_instr` | Control with no task signal | raw query | raw |
| `default` | Model-specific default formatting | model-specific (Tab. 2) | model-specific (Tab. 2) |
| `general` | Task-agnostic retrieval framing | `general` instr. + query | `passage:` for E5; else raw |
| `moral_specific` | Task-aware moral-to-fable framing | `moral_specific` instr. + query | `passage:` for E5; else raw |

Table 3: Instruction variants evaluated in the zero-shot setting.

```
query.
Query: {query}
```

The fixed `moral_specific` instruction template is:

```
Instruct: Given a moral statement,
retrieve the fable that best conveys
this moral.
Query: {query}
```

**Results.** Across our experiments, we find that query instructions have a smaller and less consistent effect compared to document-side representation changes. Adding the `general` instruction changes performance by −0.011, −0.033, and −0.038 MRR@10 for `fable-only`, `summary-only`, and `summary+fable` retrieval, respectively, averaged over all 24 zero-shot embedding models. While instructions improve some models (e.g., Linq-Embed-Mistral and Qwen3-Embedding$_{8B}$), they reduce performance for others, with no clear model category consistently benefiting. These results indicate that the main zero-shot gains in MoraLink come from changing the document representation rather than from query instruction formatting alone. Accordingly, the main tables report the fixed `general` instruction for all models, alongside the no-instruction results.

# F Summary Generation Prompts

We use generated moral summaries as document-side abstractions: each summary rewrites a fable's narrative into the kind of compact moral language used by the queries. This lets us test whether retrieval improves when the document representation makes the latent moral explicit, rather than requiring the embedding model to infer it from the full narrative.

**Generation setup.** For each fable, we send the generator a fixed system prompt and a user message of the form `Fable: {fable}`. The resulting summary is stored once and reused across all retrieval models. This keeps the generated document representation fixed, so differences in retrieval performance reflect the retriever and document layout rather than repeated stochastic generation.

**Prompt variants.** We evaluate three prompt styles, shown in Figure 5. The variants differ in how directly they ask the generator to express the moral. `direct_moral` asks for a concise moral sentence, `conceptual_abstract` asks the model to reason through the fable before outputting an abstract principle, and `cot_proverb` uses the same reasoning scaffold but asks for proverb-like phrasing.

| Subset | Instruction | fable-only | | | summary-only | | | fable+summary | | | summary+fable | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MRR | MAP | R | MRR | MAP | R | MRR | MAP | R | MRR | MAP | R |
| *Universal instruction variants, averaged over all 24 zero-shot embedding models* | | | | | | | | | | | | | |
| All models | no_instr | **0.124** | **0.085** | **0.169** | **0.334** | **0.264** | **0.418** | **0.275** | **0.213** | **0.365** | **0.306** | **0.241** | **0.395** |
| All models | general | 0.113 | 0.077 | 0.152 | 0.301 | 0.234 | 0.377 | 0.246 | 0.191 | 0.330 | 0.269 | 0.208 | 0.352 |
| All models | moral_specific | 0.091 | 0.056 | 0.114 | 0.256 | 0.190 | 0.318 | 0.217 | 0.160 | 0.277 | 0.240 | 0.176 | 0.298 |
| *Default-format comparison, restricted to the 11 models with model-specific default formats* | | | | | | | | | | | | | |
| Default-capable | no_instr | **0.112** | **0.078** | **0.157** | **0.327** | **0.256** | **0.406** | **0.255** | **0.197** | **0.338** | **0.311** | **0.245** | **0.401** |
| Default-capable | default | 0.107 | 0.074 | 0.148 | 0.307 | 0.237 | 0.379 | 0.244 | 0.186 | 0.322 | 0.309 | 0.242 | 0.395 |
| Default-capable | general | 0.087 | 0.059 | 0.121 | 0.274 | 0.207 | 0.338 | 0.202 | 0.151 | 0.271 | 0.252 | 0.190 | 0.322 |
| Default-capable | moral_specific | 0.079 | 0.051 | 0.101 | 0.243 | 0.178 | 0.299 | 0.182 | 0.131 | 0.228 | 0.234 | 0.169 | 0.282 |

Table 4: Zero-shot instruction ablation across document representations. Summary-based columns use Gemini 3 Flash `direct_moral` summaries. The first block averages instruction variants available for all 24 zero-shot embedding models. The second block restricts evaluation to the 11 models with a model-specific `default` format, so that `default` is compared against the other instruction variants on the same model subset. **R** denotes Recall@10.

**Prompt, generator, and layout ablation.** Table 5 reports the full ablation over prompt style, summary generator, document layout, and query instruction setting. We compare Gemini 3 Flash with Gemma 4 variants (E2B, E4B, and 31B), and evaluate each generated summary as a standalone document, appended after the fable, and prepended before the fable. Values are macro-averages over the 24 embedding models in the zero-shot evaluation.

The results show that the summary generator matters more than the specific prompt. Gemini 3 Flash summaries consistently outperform Gemma summaries, especially in the `summary-only` setting. Among Gemini summaries, `direct_moral` and `conceptual_abstract` are close, with `conceptual_abstract` slightly stronger in most document-layout columns. The more elaborate `cot_proverb` prompt does not improve retrieval, suggesting that proverb-like phrasing is less useful than directly expressing the moral principle.

**Summary placement.** Table 6 isolates the order of the concatenated document representation. Placing the generated moral-style summary before the original fable generally improves performance over appending the summary after the fable. The effect is strongest for Gemini 3 Flash summaries, where `summary+fable` improves over `fable+summary` by +0.026 MRR, +0.021 MAP, and +0.026 Recall on average. This suggests that when the generated abstraction is useful, presenting it first helps the embedding model anchor the document representation before processing the longer narrative text.

## G Summary-generator Ablation

We report the results for Gemma 4 variants as a summary generator: E2B in Table 7, E4B in Table 8, and 31B in Table 9.

## H Direct LLM Ranking Prompt

For the direct LLM ranking baseline, each model receives the moral query and the full fable corpus in a single prompt. The corpus is formatted as one document per entry, using the form `[fable_XXXX] text`. Models are instructed to return the top-10 fable identifiers ranked by relevance.

**System prompt.**

```
You are a retrieval system.
Return only the requested JSON.
```

**User prompt.**

```
Moral: {moral}

Corpus:

{corpus}

Return a JSON array of the {top_k}
fable IDs most relevant to this
moral, ranked by relevance. IDs
are in the format fable_XXXX.
```

## I Oracle Document-Abstraction Control

We report an oracle document-abstraction control in which each fable is paired with its human-annotated MORABLES moral. This setting uses label information that is unavailable at test time, and is therefore not a deployable retrieval setting. Instead, it estimates how much performance improves when the moral abstraction underlying each fable is made explicit.

Table 10 reports results using the oracle moral either as the document representation alone or prepended to the fable text.

## J BM25 Baseline

### J.1 BM25 Setup

BM25 serves as a lexical control for MoraLink. Because the task requires matching a short moral statement to the fable that expresses it, BM25 lets us estimate how much of the benchmark can be solved through exact surface-form overlap, without learned semantic representations.

BM25 ranks documents using exact term overlap, term-frequency saturation, inverse document frequency, and document-length normalization (Robertson and Zaragoza, 2009). We implement BM25 with lowercase regex tokenization and use $k_1 = 1.2$ and $b = 0.75$. For each moral query, BM25 ranks the full set of candidate fables under the selected document representation.

We evaluate BM25 over the same document representations used in the zero-shot experiments: the raw fable text, the generated moral summary alone, and the concatenation of the fable and generated summary. Since BM25 is order-insensitive, the two concatenated layouts, `fable+summary` and `summary+fable`, receive the same lexical score. We also evaluate surface-only controls using the fable title, first sentence, and first 50 words.

### J.2 BM25 Results Analysis

BM25 performs poorly on raw fables, reaching 0.073 MRR@10, 0.057 MAP@10, and 0.124 Recall@10. This confirms that exact lexical overlap between moral queries and fable narratives is limited, and that MoraLink cannot be solved by surface matching alone.

Generated moral summaries substantially improve BM25, raising performance to 0.196 MRR@10, 0.143 MAP@10, and 0.231 Recall@10 in the `summary-only` setting. This mirrors the embedding-model results: retrieval improves when the document representation makes the fable's latent moral explicit. The concatenated `fable+summary` representation performs similarly, reaching 0.193 MRR@10, 0.146 MAP@10, and 0.261 Recall@10.

The surface-only controls are weaker than full-fable retrieval. BM25 reaches 0.022 MRR@10 using only titles, 0.024 using first sentences, and 0.041 using the first 50 words. These controls indicate that MoraLink is not driven primarily by titles, openings, or short positional cues; successful retrieval requires representing the moral content of the full narrative.

## K Full Zero-shot Results

Table 11 reports the full zero-shot embedding models results, together with BM25 and direct LLM ranking baselines.

**Moral Clustering Prompt**

**System prompt**

You are a linguist specializing in moral philosophy and semantic analysis. Your task is to identify semantically equivalent moral statements that express the same underlying principle, even when phrased differently.

Two morals should be grouped together if they convey the same core lesson or principle. Surface-level similarity is not sufficient. Focus on whether the morals teach the same fundamental idea.

For example:

- "Union is strength" and "Union gives strength" → **should be clustered** (same principle)
- "Do not trust flatterers" and "Beware of false friends" → **may be separate** (related themes, but different lessons)

**User prompt**

You are provided with:

(i) A complete list of moral statements
(ii) The associated fables corpus
(iii) Pairwise similarity scores between morals (cosine similarity based on BGE-large-en-v1.5 embeddings)

Current cluster state: *[Current clustering configuration;json]*
Similarity scores: *[Pairwise similarity matrix];json*
Morals and fables: *[Moral-to-fable mappings;json]*

Task: Suggest which clusters should be **merged** (because they express the same principle) and which should be **split** (because they express distinct lessons). Provide justification for each decision.

Output your suggestions in the following JSON format:

```
{
  "merge_suggestions": [{
      "cluster_ids": [1, 5, 12],
      "representative_moral": "Union is strength",
    "justification": "All express the principle that collective action is..."
    }],
  "split_suggestions": [{
      "cluster_id": 3,
      "split_into": [
        {
          "morals": ["Do not trust flatterers"],
          "justification": "Focuses on deception through insincere praise"
        },
        {
          "morals": ["Beware of false friends"],
          "justification": "Focuses on betrayal in relationships"
        }]}]
}
```

**Assistant**

Output: *[JSON clustering suggestions]*

Figure 4: Prompt used for iterative moral clustering.

**Summary-Generation Prompt Variants**

**`direct_moral`** *Direct moral principle*

You are an expert in fables, parables, and moral philosophy. When given a fable, extract its moral lesson as a single sentence. Be as concise as possible.

---

**`conceptual_abstract`** *Reasoned moral principle*

You are a moral philosopher. When given a fable, reason step by step: (1) What is the central conflict or situation? (2) What does the outcome reveal about human nature? (3) What abstract principle does this illustrate? After reasoning, output only the abstract moral principle as a single sentence on the last line. Do not include your reasoning in the output. Be as concise as possible.

---

**`cot_proverb`** *Reasoned proverb/maxim*

You are an expert in fables and proverbs. When given a fable, reason step by step: (1) What is the central conflict or situation? (2) What does the outcome reveal about human nature? (3) What abstract principle does this illustrate? After reasoning, state the moral as a proverb or maxim. Output only the proverb on the last line. Do not include your reasoning in the output. Be as concise as possible.

---

**User message for all variants:** `Fable: {fable}`

Figure 5: Summary-generation prompts evaluated in the ablation. Each variant uses the system message; the user message is fixed across variants.

| | | summary-only | | | fable+summary | | | summary+fable | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Prompt** | **Summary source** | **MRR@10** | **MAP@10** | **R@10** | **MRR@10** | **MAP@10** | **R@10** | **MRR@10** | **MAP@10** | **R@10** |
| direct moral | Gemini 3 Flash | 0.334 | **0.264** | **0.418** | 0.275 | 0.213 | 0.365 | 0.306 | 0.241 | 0.395 |
| | + instr. | 0.301 | 0.234 | 0.377 | 0.246 | 0.191 | 0.330 | 0.269 | 0.208 | 0.352 |
| | Gemma 4 E2B | 0.136 | 0.092 | 0.187 | 0.170 | 0.122 | 0.237 | 0.173 | 0.123 | 0.241 |
| | + instr. | 0.120 | 0.079 | 0.165 | 0.154 | 0.109 | 0.214 | 0.154 | 0.109 | 0.217 |
| | Gemma 4 E4B | 0.165 | 0.119 | 0.216 | 0.192 | 0.139 | 0.257 | 0.201 | 0.146 | 0.265 |
| | + instr. | 0.143 | 0.102 | 0.195 | 0.170 | 0.124 | 0.232 | 0.175 | 0.126 | 0.234 |
| | Gemma 4 31B | 0.280 | 0.208 | 0.341 | 0.242 | 0.182 | 0.321 | 0.265 | 0.200 | 0.345 |
| | + instr. | 0.246 | 0.180 | 0.304 | 0.215 | 0.162 | 0.291 | 0.230 | 0.172 | 0.308 |
| conceptual abstract | Gemini 3 Flash | **0.335** | 0.263 | 0.412 | **0.279** | **0.218** | **0.372** | **0.308** | **0.243** | **0.404** |
| | + instr. | 0.301 | 0.235 | 0.377 | 0.248 | 0.192 | 0.334 | 0.272 | 0.211 | 0.356 |
| | Gemma 4 E2B | 0.112 | 0.068 | 0.147 | 0.163 | 0.114 | 0.222 | 0.162 | 0.113 | 0.225 |
| | + instr. | 0.096 | 0.057 | 0.129 | 0.144 | 0.100 | 0.203 | 0.140 | 0.097 | 0.198 |
| | Gemma 4 E4B | 0.143 | 0.090 | 0.182 | 0.177 | 0.126 | 0.245 | 0.184 | 0.129 | 0.247 |
| | + instr. | 0.124 | 0.076 | 0.158 | 0.160 | 0.113 | 0.222 | 0.163 | 0.113 | 0.223 |
| | Gemma 4 31B | 0.251 | 0.178 | 0.301 | 0.234 | 0.174 | 0.311 | 0.254 | 0.188 | 0.327 |
| | + instr. | 0.225 | 0.158 | 0.275 | 0.210 | 0.155 | 0.279 | 0.223 | 0.162 | 0.288 |
| cot proverb | Gemini 3 Flash | 0.308 | 0.247 | 0.374 | 0.271 | 0.212 | 0.358 | 0.297 | 0.234 | 0.384 |
| | + instr. | 0.279 | 0.221 | 0.339 | 0.244 | 0.190 | 0.322 | 0.265 | 0.207 | 0.346 |
| | Gemma 4 E2B | 0.135 | 0.090 | 0.176 | 0.166 | 0.118 | 0.230 | 0.169 | 0.120 | 0.232 |
| | + instr. | 0.115 | 0.076 | 0.151 | 0.148 | 0.106 | 0.209 | 0.149 | 0.105 | 0.208 |
| | Gemma 4 E4B | 0.093 | 0.057 | 0.121 | 0.153 | 0.110 | 0.215 | 0.156 | 0.111 | 0.216 |
| | + instr. | 0.081 | 0.048 | 0.105 | 0.136 | 0.097 | 0.192 | 0.137 | 0.097 | 0.192 |
| | Gemma 4 31B | 0.194 | 0.141 | 0.250 | 0.217 | 0.162 | 0.294 | 0.231 | 0.171 | 0.308 |
| | + instr. | 0.172 | 0.122 | 0.223 | 0.195 | 0.144 | 0.261 | 0.204 | 0.149 | 0.272 |

Table 5: Summary-generator and prompt ablation results averaged over the 24 embedding models in the zero-shot evaluation. Each summary source is evaluated without a query instruction and with the general retrieval instruction (+ *instr.*). Values are macro-averages across models for the corresponding prompt, summary source, document layout, and instruction setting. Best values in each metric column are bolded.

| | fable+summary | | | summary+fable | | | Δ | | |
|---|---|---|---|---|---|---|---|---|---|
| **Summary source** | **MRR** | **MAP** | **R** | **MRR** | **MAP** | **R** | **MRR** | **MAP** | **R** |
| Gemini 3 Flash | 0.261 | 0.203 | 0.347 | **0.286** | **0.224** | **0.373** | +0.026 | +0.021 | +0.026 |
| Gemma 4 E2B | 0.158 | 0.111 | 0.219 | 0.158 | 0.111 | 0.220 | +0.000 | -0.000 | +0.001 |
| Gemma 4 E4B | 0.165 | 0.118 | 0.227 | **0.169** | **0.120** | **0.230** | +0.005 | +0.002 | +0.002 |
| Gemma 4 31B | 0.219 | 0.163 | 0.293 | **0.235** | **0.174** | **0.308** | +0.016 | +0.010 | +0.015 |

Table 6: Average effect of summary placement in concatenated document representations. Values are averaged over the prompt variants and instruction settings in Table 5. Δ is computed as `summary+fable` minus `fable+summary`.

| | | fable-only | | | summary-only | | | summary+fable | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Param** | **MRR@10** | **MAP@10** | **R@10** | **MRR@10** | **MAP@10** | **R@10** | **MRR@10** | **MAP@10** | **R@10** |
| $\text{MiniLM}_{\text{L6-v2}}$ | 22M | 0.095 | 0.068 | 0.142 | 0.120 | 0.082 | 0.161 | 0.131 | 0.090 | 0.191 |
| + instr. | | 0.067 | 0.046 | 0.102 | 0.083 | 0.052 | 0.111 | 0.092 | 0.062 | 0.137 |
| $\text{BGE}_{\text{base-en-v1.5}}$ | 109M | 0.075 | 0.049 | 0.105 | 0.111 | 0.075 | 0.156 | 0.130 | 0.092 | 0.195 |
| + instr. | | 0.050 | 0.031 | 0.058 | 0.078 | 0.044 | 0.121 | 0.097 | 0.066 | 0.122 |
| all-mpnet-base-v2 | 110M | 0.096 | 0.067 | 0.157 | 0.144 | 0.097 | 0.204 | 0.152 | 0.107 | 0.219 |
| + instr. | | 0.042 | 0.027 | 0.073 | 0.070 | 0.038 | 0.092 | 0.075 | 0.050 | 0.118 |
| Contriever | 110M | 0.101 | 0.070 | 0.159 | 0.111 | 0.075 | 0.163 | 0.140 | 0.097 | 0.208 |
| + instr. | | 0.078 | 0.056 | 0.126 | 0.099 | 0.064 | 0.146 | 0.096 | 0.064 | 0.157 |
| $\text{E5}_{\text{base-v2}}$ | 110M | 0.106 | 0.072 | 0.137 | 0.122 | 0.081 | 0.173 | 0.161 | 0.118 | 0.215 |
| + instr. | | 0.103 | 0.071 | 0.144 | 0.112 | 0.073 | 0.164 | 0.159 | 0.115 | 0.217 |
| $\text{InstructOR}_{\text{base}}$ | 110M | 0.097 | 0.066 | 0.148 | 0.128 | 0.085 | 0.180 | 0.163 | 0.115 | 0.234 |
| + instr. | | 0.056 | 0.035 | 0.086 | 0.090 | 0.058 | 0.134 | 0.121 | 0.083 | 0.171 |
| TART-Contriever | 110M | 0.103 | 0.070 | 0.143 | 0.116 | 0.079 | 0.180 | 0.132 | 0.093 | 0.191 |
| + instr. | | 0.085 | 0.059 | 0.141 | 0.094 | 0.061 | 0.170 | 0.124 | 0.086 | 0.177 |
| $\text{BGE}_{\text{large-en-v1.5}}$ | 335M | 0.081 | 0.053 | 0.123 | 0.116 | 0.075 | 0.173 | 0.152 | 0.107 | 0.223 |
| + instr. | | 0.053 | 0.036 | 0.076 | 0.064 | 0.036 | 0.098 | 0.099 | 0.066 | 0.154 |
| $\text{E5}_{\text{large-v2}}$ | 335M | 0.110 | 0.081 | 0.163 | 0.123 | 0.080 | 0.170 | 0.165 | 0.122 | 0.246 |
| + instr. | | 0.106 | 0.077 | 0.153 | 0.113 | 0.073 | 0.158 | 0.165 | 0.116 | 0.222 |
| Nomic-v2 | 475M | 0.126 | 0.090 | 0.182 | 0.118 | 0.079 | 0.177 | 0.169 | 0.123 | 0.248 |
| + instr. | | 0.064 | 0.042 | 0.084 | 0.087 | 0.051 | 0.114 | 0.107 | 0.069 | 0.143 |
| BGE-M3 | 568M | 0.109 | 0.076 | 0.160 | 0.115 | 0.075 | 0.160 | 0.176 | 0.126 | 0.252 |
| + instr. | | 0.079 | 0.051 | 0.108 | 0.113 | 0.074 | 0.140 | 0.144 | 0.098 | 0.210 |
| $\text{Multilingual-E5}_{\text{large}}$ | 560M | 0.114 | 0.078 | 0.147 | 0.116 | 0.076 | 0.155 | 0.173 | 0.125 | 0.244 |
| + instr. | | 0.113 | 0.077 | 0.137 | 0.130 | 0.083 | 0.184 | 0.171 | 0.120 | 0.218 |
| $\text{Multilingual-E5}_{\text{large-instruct}}$ | 560M | 0.099 | 0.068 | 0.154 | 0.139 | 0.090 | 0.178 | 0.167 | 0.120 | 0.243 |
| + instr. | | 0.114 | 0.077 | 0.164 | 0.151 | 0.101 | 0.200 | 0.173 | 0.126 | 0.259 |
| $\text{Qwen3-Embedding}_{\text{0.6B}}$ | 596M | 0.123 | 0.084 | 0.164 | 0.144 | 0.097 | 0.193 | 0.191 | 0.134 | 0.256 |
| + instr. | | 0.135 | 0.089 | 0.178 | 0.144 | 0.099 | 0.199 | 0.209 | 0.147 | 0.282 |
| $\text{Drama}_{\text{1B}}$ | 1.24B | 0.103 | 0.072 | 0.151 | 0.122 | 0.082 | 0.162 | 0.163 | 0.112 | 0.238 |
| + instr. | | 0.052 | 0.033 | 0.077 | 0.075 | 0.044 | 0.095 | 0.081 | 0.055 | 0.143 |
| $\text{InstructOR}_{\text{xl}}$ | 1.24B | 0.146 | 0.102 | 0.198 | 0.169 | 0.121 | 0.227 | 0.223 | 0.167 | 0.310 |
| + instr. | | 0.089 | 0.054 | 0.114 | 0.113 | 0.071 | 0.156 | 0.135 | 0.085 | 0.183 |
| Lychee-Embed | 1.54B | 0.132 | 0.087 | 0.160 | 0.145 | 0.101 | 0.193 | 0.192 | 0.139 | 0.259 |
| + instr. | | 0.137 | 0.090 | 0.166 | 0.149 | 0.104 | 0.200 | 0.202 | 0.148 | 0.293 |
| $\text{Qwen3-Embedding}_{\text{4B}}$ | 4.02B | 0.181 | 0.126 | 0.230 | 0.160 | 0.110 | 0.207 | 0.216 | 0.155 | 0.287 |
| + instr. | | 0.212 | 0.145 | 0.269 | 0.176 | 0.120 | 0.235 | 0.241 | 0.181 | 0.341 |
| BGE-en-ICL | 7.11B | 0.174 | 0.123 | 0.223 | 0.170 | 0.114 | 0.228 | 0.212 | 0.151 | 0.273 |
| + instr. | | 0.191 | 0.134 | 0.255 | 0.163 | 0.111 | 0.215 | 0.231 | 0.168 | 0.305 |
| $\text{E5}_{\text{Mistral}}$ | 7.11B | 0.088 | 0.048 | 0.100 | 0.159 | 0.108 | 0.197 | 0.100 | 0.056 | 0.112 |
| + instr. | | 0.107 | 0.061 | 0.121 | 0.167 | 0.111 | 0.207 | 0.114 | 0.066 | 0.140 |
| Linq-Embed-Mistral | 7.11B | 0.239 | 0.171 | 0.292 | 0.170 | 0.117 | 0.228 | 0.241 | 0.179 | 0.339 |
| + instr. | | **0.268** | **0.196** | **0.341** | **0.185** | 0.126 | 0.235 | **0.271** | **0.208** | 0.382 |
| SFR-Embedding-Mistral | 7.11B | 0.199 | 0.145 | 0.252 | 0.171 | 0.117 | 0.236 | 0.219 | 0.160 | 0.301 |
| + instr. | | 0.233 | 0.165 | 0.298 | 0.176 | 0.120 | 0.236 | 0.244 | 0.180 | 0.328 |
| $\text{Qwen3-Embedding}_{\text{8B}}$ | 7.57B | 0.198 | 0.136 | 0.273 | 0.174 | 0.122 | 0.242 | 0.253 | 0.183 | 0.333 |
| + instr. | | 0.229 | 0.165 | 0.317 | 0.180 | **0.127** | **0.260** | 0.270 | 0.205 | **0.390** |
| $\text{Llama-Embed}_{\text{Nemotron-8B}}$ | 7.50B | 0.074 | 0.046 | 0.097 | 0.110 | 0.070 | 0.139 | 0.128 | 0.084 | 0.161 |
| + instr. | | 0.044 | 0.024 | 0.067 | 0.079 | 0.045 | 0.092 | 0.077 | 0.049 | 0.123 |

Table 7: Zero-shot summary-generator ablation results on MoraLink using Gemma 4 E2B generated direct moral summaries. Each model is evaluated without a query instruction and with the general retrieval instruction (+ *instr.*). `summary-only` and `summary+fable` columns use Gemma 4 E2B generated moral summaries. Best values in each metric column are bolded.

| Model | Param | fable-only | | | summary-only | | | summary+fable | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **MRR@10** | **MAP@10** | **R@10** | **MRR@10** | **MAP@10** | **R@10** | **MRR@10** | **MAP@10** | **R@10** |
| $\text{MiniLM}_{\text{L6-v2}}$ | 22M | 0.095 | 0.068 | 0.142 | 0.131 | 0.090 | 0.183 | 0.147 | 0.102 | 0.196 |
| + instr. | | 0.067 | 0.046 | 0.102 | 0.098 | 0.063 | 0.135 | 0.099 | 0.064 | 0.135 |
| $\text{BGE}_{\text{base-en-v1.5}}$ | 109M | 0.075 | 0.049 | 0.105 | 0.152 | 0.111 | 0.199 | 0.151 | 0.107 | 0.220 |
| + instr. | | 0.050 | 0.031 | 0.058 | 0.092 | 0.060 | 0.152 | 0.095 | 0.062 | 0.140 |
| all-mpnet-base-v2 | 110M | 0.096 | 0.067 | 0.157 | 0.164 | 0.117 | 0.217 | 0.182 | 0.131 | 0.254 |
| + instr. | | 0.042 | 0.027 | 0.073 | 0.088 | 0.060 | 0.124 | 0.092 | 0.065 | 0.148 |
| Contriever | 110M | 0.101 | 0.070 | 0.159 | 0.151 | 0.100 | 0.193 | 0.198 | 0.138 | 0.238 |
| + instr. | | 0.078 | 0.056 | 0.126 | 0.134 | 0.090 | 0.175 | 0.130 | 0.088 | 0.162 |
| $\text{E5}_{\text{base-v2}}$ | 110M | 0.106 | 0.072 | 0.137 | 0.153 | 0.112 | 0.202 | 0.192 | 0.144 | 0.240 |
| + instr. | | 0.103 | 0.071 | 0.144 | 0.156 | 0.113 | 0.201 | 0.184 | 0.135 | 0.252 |
| $\text{InstructOR}_{\text{base}}$ | 110M | 0.097 | 0.066 | 0.148 | 0.153 | 0.109 | 0.211 | 0.182 | 0.135 | 0.258 |
| + instr. | | 0.056 | 0.035 | 0.086 | 0.127 | 0.089 | 0.173 | 0.137 | 0.094 | 0.182 |
| TART-Contriever | 110M | 0.103 | 0.070 | 0.143 | 0.150 | 0.105 | 0.207 | 0.177 | 0.128 | 0.234 |
| + instr. | | 0.085 | 0.059 | 0.141 | 0.130 | 0.088 | 0.168 | 0.152 | 0.108 | 0.204 |
| $\text{BGE}_{\text{large-en-v1.5}}$ | 335M | 0.081 | 0.053 | 0.123 | 0.153 | 0.110 | 0.212 | 0.193 | 0.139 | 0.278 |
| + instr. | | 0.053 | 0.036 | 0.076 | 0.091 | 0.058 | 0.127 | 0.104 | 0.068 | 0.154 |
| $\text{E5}_{\text{large-v2}}$ | 335M | 0.110 | 0.081 | 0.163 | 0.149 | 0.105 | 0.181 | 0.197 | 0.147 | 0.266 |
| + instr. | | 0.106 | 0.077 | 0.153 | 0.145 | 0.099 | 0.188 | 0.185 | 0.139 | 0.262 |
| Nomic-v2 | 475M | 0.126 | 0.090 | 0.182 | 0.157 | 0.112 | 0.214 | 0.206 | 0.148 | 0.269 |
| + instr. | | 0.064 | 0.042 | 0.084 | 0.097 | 0.065 | 0.144 | 0.126 | 0.085 | 0.154 |
| BGE-M3 | 568M | 0.109 | 0.076 | 0.160 | 0.153 | 0.106 | 0.199 | 0.210 | 0.153 | 0.276 |
| + instr. | | 0.079 | 0.051 | 0.108 | 0.134 | 0.091 | 0.184 | 0.180 | 0.122 | 0.234 |
| $\text{Multilingual-E5}_{\text{large}}$ | 560M | 0.114 | 0.078 | 0.147 | 0.152 | 0.105 | 0.184 | 0.207 | 0.151 | 0.271 |
| + instr. | | 0.113 | 0.077 | 0.137 | 0.153 | 0.106 | 0.198 | 0.198 | 0.143 | 0.251 |
| $\text{Multilingual-E5}_{\text{large-instruct}}$ | 560M | 0.099 | 0.068 | 0.154 | 0.170 | 0.118 | 0.217 | 0.197 | 0.145 | 0.264 |
| + instr. | | 0.114 | 0.077 | 0.164 | 0.182 | 0.130 | 0.238 | 0.207 | 0.156 | 0.285 |
| $\text{Qwen3-Embedding}_{\text{0.6B}}$ | 596M | 0.123 | 0.084 | 0.164 | 0.171 | 0.123 | 0.223 | 0.217 | 0.160 | 0.280 |
| + instr. | | 0.135 | 0.089 | 0.178 | 0.175 | 0.126 | 0.239 | 0.236 | 0.176 | 0.309 |
| $\text{Drama}_{\text{1B}}$ | 1.24B | 0.103 | 0.072 | 0.151 | 0.150 | 0.104 | 0.189 | 0.206 | 0.146 | 0.262 |
| + instr. | | 0.052 | 0.033 | 0.077 | 0.095 | 0.062 | 0.129 | 0.105 | 0.065 | 0.150 |
| $\text{InstructOR}_{\text{xl}}$ | 1.24B | 0.146 | 0.102 | 0.198 | 0.195 | 0.146 | 0.272 | 0.261 | 0.200 | 0.349 |
| + instr. | | 0.089 | 0.054 | 0.114 | 0.111 | 0.077 | 0.170 | 0.145 | 0.102 | 0.211 |
| Lychee-Embed | 1.54B | 0.132 | 0.087 | 0.160 | 0.155 | 0.111 | 0.209 | 0.214 | 0.158 | 0.278 |
| + instr. | | 0.137 | 0.090 | 0.166 | 0.168 | 0.124 | 0.240 | 0.238 | 0.178 | 0.306 |
| $\text{Qwen3-Embedding}_{\text{4B}}$ | 4.02B | 0.181 | 0.126 | 0.230 | 0.189 | 0.145 | 0.246 | 0.242 | 0.178 | 0.305 |
| + instr. | | 0.212 | 0.145 | 0.269 | 0.200 | 0.153 | 0.268 | 0.284 | 0.217 | 0.356 |
| BGE-en-ICL | 7.11B | 0.174 | 0.123 | 0.223 | 0.200 | 0.150 | 0.253 | 0.223 | 0.167 | 0.294 |
| + instr. | | 0.191 | 0.134 | 0.255 | 0.184 | 0.134 | 0.237 | 0.245 | 0.184 | 0.310 |
| $\text{E5}_{\text{Mistral}}$ | 7.11B | 0.088 | 0.048 | 0.100 | 0.185 | 0.134 | 0.234 | 0.113 | 0.066 | 0.121 |
| + instr. | | 0.107 | 0.061 | 0.121 | 0.185 | 0.136 | 0.239 | 0.124 | 0.073 | 0.138 |
| Linq-Embed-Mistral | 7.11B | 0.239 | 0.171 | 0.292 | 0.191 | 0.143 | 0.247 | 0.275 | 0.208 | 0.370 |
| + instr. | | **0.268** | **0.196** | **0.341** | 0.204 | 0.156 | 0.287 | 0.292 | 0.226 | **0.418** |
| SFR-Embedding-Mistral | 7.11B | 0.199 | 0.145 | 0.252 | 0.201 | 0.153 | 0.260 | 0.237 | 0.175 | 0.317 |
| + instr. | | 0.233 | 0.165 | 0.298 | **0.209** | **0.161** | 0.285 | 0.259 | 0.195 | 0.360 |
| $\text{Qwen3-Embedding}_{\text{8B}}$ | 7.57B | 0.198 | 0.136 | 0.273 | 0.198 | 0.149 | 0.256 | 0.266 | 0.202 | 0.348 |
| + instr. | | 0.229 | 0.165 | 0.317 | 0.203 | 0.157 | **0.289** | **0.295** | **0.230** | 0.395 |
| $\text{Llama-Embed}_{\text{Nemotron-8B}}$ | 7.50B | 0.074 | 0.046 | 0.097 | 0.130 | 0.092 | 0.176 | 0.133 | 0.086 | 0.169 |
| + instr. | | 0.044 | 0.024 | 0.067 | 0.080 | 0.048 | 0.096 | 0.092 | 0.057 | 0.108 |

Table 8: Zero-shot summary-generator ablation results on MoraLink using Gemma 4 E4B generated direct moral summaries. Each model is evaluated without a query instruction and with the general retrieval instruction (+ *instr.*). `summary-only` and `summary+fable` columns use Gemma 4 E4B generated moral summaries. Best values in each metric column are bolded.

| | | fable-only | | | summary-only | | | summary+fable | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Param** | **MRR@10** | **MAP@10** | **R@10** | **MRR@10** | **MAP@10** | **R@10** | **MRR@10** | **MAP@10** | **R@10** |
| $\text{MiniLM}_{\text{L6-v2}}$ | 22M | 0.095 | 0.068 | 0.142 | 0.232 | 0.164 | 0.284 | 0.186 | 0.134 | 0.258 |
| + instr. | | 0.067 | 0.046 | 0.102 | 0.148 | 0.099 | 0.204 | 0.132 | 0.091 | 0.189 |
| $\text{BGE}_{\text{base-en-v1.5}}$ | 109M | 0.075 | 0.049 | 0.105 | 0.249 | 0.178 | 0.294 | 0.207 | 0.150 | 0.283 |
| + instr. | | 0.050 | 0.031 | 0.058 | 0.152 | 0.098 | 0.205 | 0.131 | 0.088 | 0.198 |
| all-mpnet-base-v2 | 110M | 0.096 | 0.067 | 0.157 | 0.271 | 0.200 | 0.339 | 0.233 | 0.171 | 0.327 |
| + instr. | | 0.042 | 0.027 | 0.073 | 0.157 | 0.109 | 0.209 | 0.101 | 0.069 | 0.159 |
| Contriever | 110M | 0.101 | 0.070 | 0.159 | 0.233 | 0.169 | 0.292 | 0.243 | 0.177 | 0.306 |
| + instr. | | 0.078 | 0.056 | 0.126 | 0.214 | 0.157 | 0.273 | 0.145 | 0.105 | 0.214 |
| $\text{E5}_{\text{base-v2}}$ | 110M | 0.106 | 0.072 | 0.137 | 0.268 | 0.196 | 0.314 | 0.266 | 0.207 | 0.350 |
| + instr. | | 0.103 | 0.071 | 0.144 | 0.249 | 0.176 | 0.299 | 0.259 | 0.199 | 0.346 |
| $\text{InstructOR}_{\text{base}}$ | 110M | 0.097 | 0.066 | 0.148 | 0.267 | 0.195 | 0.328 | 0.250 | 0.183 | 0.329 |
| + instr. | | 0.056 | 0.035 | 0.086 | 0.216 | 0.152 | 0.271 | 0.196 | 0.145 | 0.265 |
| TART-Contriever | 110M | 0.103 | 0.070 | 0.143 | 0.270 | 0.201 | 0.322 | 0.218 | 0.162 | 0.312 |
| + instr. | | 0.085 | 0.059 | 0.141 | 0.242 | 0.174 | 0.299 | 0.187 | 0.133 | 0.263 |
| $\text{BGE}_{\text{large-en-v1.5}}$ | 335M | 0.081 | 0.053 | 0.123 | 0.264 | 0.191 | 0.316 | 0.271 | 0.207 | 0.358 |
| + instr. | | 0.053 | 0.036 | 0.076 | 0.127 | 0.083 | 0.197 | 0.152 | 0.110 | 0.240 |
| $\text{E5}_{\text{large-v2}}$ | 335M | 0.110 | 0.081 | 0.163 | 0.258 | 0.186 | 0.312 | 0.267 | 0.203 | 0.338 |
| + instr. | | 0.106 | 0.077 | 0.153 | 0.246 | 0.174 | 0.292 | 0.262 | 0.203 | 0.352 |
| Nomic-v2 | 475M | 0.126 | 0.090 | 0.182 | 0.269 | 0.198 | 0.316 | 0.285 | 0.217 | 0.367 |
| + instr. | | 0.064 | 0.042 | 0.084 | 0.178 | 0.121 | 0.215 | 0.180 | 0.119 | 0.217 |
| BGE-M3 | 568M | 0.109 | 0.076 | 0.160 | 0.272 | 0.197 | 0.324 | 0.283 | 0.217 | 0.368 |
| + instr. | | 0.079 | 0.051 | 0.108 | 0.245 | 0.177 | 0.294 | 0.241 | 0.173 | 0.314 |
| $\text{Multilingual-E5}_{\text{large}}$ | 560M | 0.114 | 0.078 | 0.147 | 0.269 | 0.193 | 0.313 | 0.271 | 0.203 | 0.328 |
| + instr. | | 0.113 | 0.077 | 0.137 | 0.279 | 0.203 | 0.324 | 0.272 | 0.201 | 0.331 |
| $\text{Multilingual-E5}_{\text{large-instruct}}$ | 560M | 0.099 | 0.068 | 0.154 | 0.295 | 0.225 | 0.374 | 0.261 | 0.201 | 0.339 |
| + instr. | | 0.114 | 0.077 | 0.164 | 0.316 | 0.244 | 0.383 | 0.280 | 0.218 | 0.351 |
| $\text{Qwen3-Embedding}_{\text{0.6B}}$ | 596M | 0.123 | 0.084 | 0.164 | 0.289 | 0.218 | 0.356 | 0.286 | 0.211 | 0.356 |
| + instr. | | 0.135 | 0.089 | 0.178 | 0.297 | 0.220 | 0.337 | 0.311 | 0.232 | 0.397 |
| $\text{Drama}_{\text{1B}}$ | 1.24B | 0.103 | 0.072 | 0.151 | 0.263 | 0.192 | 0.311 | 0.282 | 0.212 | 0.363 |
| + instr. | | 0.052 | 0.033 | 0.077 | 0.183 | 0.120 | 0.226 | 0.159 | 0.109 | 0.230 |
| $\text{InstructOR}_{\text{xl}}$ | 1.24B | 0.146 | 0.102 | 0.198 | 0.312 | 0.241 | 0.400 | 0.324 | 0.258 | 0.433 |
| + instr. | | 0.089 | 0.054 | 0.114 | 0.235 | 0.174 | 0.286 | 0.189 | 0.127 | 0.261 |
| Lychee-Embed | 1.54B | 0.132 | 0.087 | 0.160 | 0.282 | 0.202 | 0.318 | 0.282 | 0.212 | 0.358 |
| + instr. | | 0.137 | 0.090 | 0.166 | 0.287 | 0.210 | 0.339 | 0.301 | 0.232 | 0.406 |
| $\text{Qwen3-Embedding}_{\text{4B}}$ | 4.02B | 0.181 | 0.126 | 0.230 | 0.327 | 0.250 | 0.388 | 0.314 | 0.246 | 0.409 |
| + instr. | | 0.212 | 0.145 | 0.269 | 0.331 | 0.255 | 0.409 | 0.350 | 0.276 | 0.465 |
| BGE-en-ICL | 7.11B | 0.174 | 0.123 | 0.223 | 0.328 | 0.256 | 0.414 | 0.306 | 0.241 | 0.395 |
| + instr. | | 0.191 | 0.134 | 0.255 | 0.327 | 0.251 | 0.387 | 0.328 | 0.257 | 0.419 |
| $\text{E5}_{\text{Mistral}}$ | 7.11B | 0.088 | 0.048 | 0.100 | 0.274 | 0.205 | 0.346 | 0.137 | 0.086 | 0.161 |
| + instr. | | 0.107 | 0.061 | 0.121 | 0.288 | 0.220 | 0.364 | 0.157 | 0.097 | 0.172 |
| Linq-Embed-Mistral | 7.11B | 0.239 | 0.171 | 0.292 | 0.317 | 0.246 | 0.400 | 0.333 | 0.267 | 0.455 |
| + instr. | | **0.268** | **0.196** | **0.341** | 0.331 | 0.256 | 0.430 | 0.368 | 0.297 | 0.486 |
| SFR-Embedding-Mistral | 7.11B | 0.199 | 0.145 | 0.252 | 0.330 | 0.256 | 0.425 | 0.292 | 0.227 | 0.388 |
| + instr. | | 0.233 | 0.165 | 0.298 | **0.338** | 0.257 | 0.416 | 0.319 | 0.251 | 0.427 |
| $\text{Qwen3-Embedding}_{\text{8B}}$ | 7.57B | 0.198 | 0.136 | 0.273 | 0.334 | 0.258 | 0.413 | 0.357 | 0.285 | 0.461 |
| + instr. | | 0.229 | 0.165 | 0.317 | **0.338** | **0.262** | **0.437** | **0.382** | **0.308** | **0.516** |
| $\text{Llama-Embed}_{\text{Nemotron-8B}}$ | 7.50B | 0.074 | 0.046 | 0.097 | 0.246 | 0.176 | 0.292 | 0.194 | 0.133 | 0.245 |
| + instr. | | 0.044 | 0.024 | 0.067 | 0.192 | 0.125 | 0.190 | 0.129 | 0.084 | 0.167 |

Table 9: Zero-shot summary-generator ablation results on MoraLink using Gemma 4 31B generated direct moral summaries. Each model is evaluated without a query instruction and with the general retrieval instruction (+ *instr.*). `summary-only` and `summary+fable` columns use Gemma 4 31B generated moral summaries. Best values in each metric column are bolded.

| Model | Param | fable-only | | | summary-only | | | summary+fable | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **MRR@10** | **MAP@10** | **R@10** | **MRR@10** | **MAP@10** | **R@10** | **MRR@10** | **MAP@10** | **R@10** |
| $\text{MiniLM}_{\text{L6-v2}}$ | 22M | 0.095 | 0.068 | 0.142 | **1.000** | 0.929 | 0.940 | 0.723 | 0.636 | 0.759 |
| + instr. | | 0.067 | 0.046 | 0.102 | 0.956 | 0.865 | 0.893 | 0.562 | 0.480 | 0.614 |
| $\text{BGE}_{\text{base-en-v1.5}}$ | 109M | 0.075 | 0.049 | 0.105 | **1.000** | 0.934 | 0.948 | 0.856 | 0.765 | 0.867 |
| + instr. | | 0.050 | 0.031 | 0.058 | 0.993 | 0.903 | 0.921 | 0.674 | 0.592 | 0.714 |
| all-mpnet-base-v2 | 110M | 0.096 | 0.067 | 0.157 | **1.000** | 0.940 | 0.958 | 0.830 | 0.747 | 0.859 |
| + instr. | | 0.042 | 0.027 | 0.073 | 0.869 | 0.786 | 0.870 | 0.463 | 0.409 | 0.573 |
| Contriever | 110M | 0.101 | 0.070 | 0.159 | **1.000** | 0.934 | 0.948 | 0.933 | 0.835 | 0.893 |
| + instr. | | 0.078 | 0.056 | 0.126 | **1.000** | 0.926 | 0.936 | 0.736 | 0.629 | 0.742 |
| $\text{E5}_{\text{base-v2}}$ | 110M | 0.106 | 0.072 | 0.137 | **1.000** | 0.939 | 0.952 | 0.965 | 0.881 | 0.916 |
| + instr. | | 0.103 | 0.071 | 0.144 | **1.000** | 0.939 | 0.951 | 0.961 | 0.880 | 0.928 |
| $\text{InstructOR}_{\text{base}}$ | 110M | 0.097 | 0.066 | 0.148 | **1.000** | 0.937 | 0.953 | 0.947 | 0.863 | 0.911 |
| + instr. | | 0.056 | 0.035 | 0.086 | 0.995 | 0.920 | 0.937 | 0.820 | 0.729 | 0.847 |
| TART-Contriever | 110M | 0.103 | 0.070 | 0.143 | **1.000** | 0.938 | 0.948 | 0.929 | 0.828 | 0.902 |
| + instr. | | 0.085 | 0.059 | 0.141 | **1.000** | 0.930 | 0.946 | 0.866 | 0.762 | 0.856 |
| $\text{BGE}_{\text{large-en-v1.5}}$ | 335M | 0.081 | 0.053 | 0.123 | **1.000** | 0.940 | 0.952 | 0.931 | 0.849 | 0.901 |
| + instr. | | 0.053 | 0.036 | 0.076 | 0.958 | 0.865 | 0.906 | 0.777 | 0.689 | 0.802 |
| $\text{E5}_{\text{large-v2}}$ | 335M | 0.110 | 0.081 | 0.163 | **1.000** | 0.937 | 0.950 | 0.940 | 0.854 | 0.894 |
| + instr. | | 0.106 | 0.077 | 0.153 | **1.000** | 0.939 | 0.947 | 0.953 | 0.869 | 0.905 |
| Nomic-v2 | 475M | 0.126 | 0.090 | 0.182 | **1.000** | 0.938 | 0.951 | **0.993** | **0.913** | 0.935 |
| + instr. | | 0.064 | 0.042 | 0.084 | 0.998 | 0.915 | 0.931 | 0.893 | 0.790 | 0.866 |
| BGE-M3 | 568M | 0.109 | 0.076 | 0.160 | **1.000** | 0.945 | 0.958 | 0.966 | 0.889 | 0.930 |
| + instr. | | 0.079 | 0.051 | 0.108 | **1.000** | 0.935 | 0.950 | 0.914 | 0.829 | 0.903 |
| $\text{Multilingual-E5}_{\text{large}}$ | 560M | 0.114 | 0.078 | 0.147 | **1.000** | 0.946 | 0.957 | 0.952 | 0.864 | 0.908 |
| + instr. | | 0.113 | 0.077 | 0.137 | **1.000** | 0.944 | 0.957 | 0.961 | 0.875 | 0.910 |
| $\text{Multilingual-E5}_{\text{large-instruct}}$ | 560M | 0.099 | 0.068 | 0.154 | **1.000** | 0.951 | 0.964 | 0.914 | 0.829 | 0.896 |
| + instr. | | 0.114 | 0.077 | 0.164 | **1.000** | 0.951 | 0.963 | 0.930 | 0.849 | 0.913 |
| $\text{Qwen3-Embedding}_{\text{0.6B}}$ | 596M | 0.123 | 0.084 | 0.164 | **1.000** | 0.945 | 0.958 | 0.903 | 0.827 | 0.911 |
| + instr. | | 0.135 | 0.089 | 0.178 | **1.000** | 0.947 | 0.963 | 0.955 | 0.884 | 0.933 |
| $\text{Drama}_{\text{1B}}$ | 1.24B | 0.103 | 0.072 | 0.151 | **1.000** | 0.934 | 0.954 | 0.981 | 0.904 | 0.937 |
| + instr. | | 0.052 | 0.033 | 0.077 | 0.999 | 0.912 | 0.928 | 0.844 | 0.741 | 0.831 |
| $\text{InstructOR}_{\text{xl}}$ | 1.24B | 0.146 | 0.102 | 0.198 | **1.000** | 0.958 | 0.968 | 0.940 | 0.867 | 0.930 |
| + instr. | | 0.089 | 0.054 | 0.114 | 0.999 | 0.930 | 0.946 | 0.717 | 0.632 | 0.782 |
| Lychee-Embed | 1.54B | 0.132 | 0.087 | 0.160 | **1.000** | 0.944 | 0.960 | 0.830 | 0.752 | 0.871 |
| + instr. | | 0.137 | 0.090 | 0.166 | **1.000** | 0.943 | 0.956 | 0.897 | 0.817 | 0.900 |
| $\text{Qwen3-Embedding}_{\text{4B}}$ | 4.02B | 0.181 | 0.126 | 0.230 | **1.000** | 0.959 | 0.970 | 0.800 | 0.727 | 0.849 |
| + instr. | | 0.212 | 0.145 | 0.269 | **1.000** | 0.958 | 0.971 | 0.910 | 0.836 | 0.923 |
| BGE-en-ICL | 7.11B | 0.174 | 0.123 | 0.223 | **1.000** | 0.955 | 0.971 | 0.842 | 0.771 | 0.884 |
| + instr. | | 0.191 | 0.134 | 0.255 | **1.000** | 0.956 | 0.970 | 0.936 | 0.863 | 0.923 |
| $\text{E5}_{\text{Mistral}}$ | 7.11B | 0.088 | 0.048 | 0.100 | **1.000** | 0.947 | 0.962 | 0.266 | 0.212 | 0.311 |
| + instr. | | 0.107 | 0.061 | 0.121 | **1.000** | 0.950 | 0.966 | 0.321 | 0.258 | 0.389 |
| Linq-Embed-Mistral | 7.11B | 0.239 | 0.171 | 0.292 | **1.000** | 0.961 | 0.974 | 0.923 | 0.855 | 0.938 |
| + instr. | | **0.268** | **0.196** | **0.341** | 1.000 | 0.959 | 0.972 | 0.973 | 0.910 | **0.951** |
| SFR-Embedding-Mistral | 7.11B | 0.199 | 0.145 | 0.252 | **1.000** | **0.964** | **0.977** | 0.782 | 0.709 | 0.858 |
| + instr. | | 0.233 | 0.165 | 0.298 | **1.000** | 0.961 | 0.975 | 0.879 | 0.811 | 0.911 |
| $\text{Qwen3-Embedding}_{\text{8B}}$ | 7.57B | 0.198 | 0.136 | 0.273 | **1.000** | 0.961 | 0.976 | 0.883 | 0.820 | 0.908 |
| + instr. | | 0.229 | 0.165 | 0.317 | **1.000** | 0.963 | 0.975 | 0.934 | 0.876 | 0.939 |
| $\text{Llama-Embed}_{\text{Nemotron-8B}}$ | 7.50B | 0.074 | 0.046 | 0.097 | 0.996 | 0.921 | 0.941 | 0.806 | 0.715 | 0.843 |
| + instr. | | 0.044 | 0.024 | 0.067 | 0.983 | 0.879 | 0.890 | 0.505 | 0.443 | 0.577 |

Table 10: Oracle document-abstraction results on MoraLink. Each model is evaluated without a query instruction and with the general retrieval instruction (+ *instr.*). `summary-only` and `summary+fable` columns use the human-annotated MORABLES moral associated with each fable as an oracle summary. This setting uses label information and is intended as an upper-bound control, not as a valid test-time input. Best values in each metric column are bolded.

| Model | Param | fable-only | | | summary-only | | | summary+fable | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **MRR@10** | **MAP@10** | **R@10** | **MRR@10** | **MAP@10** | **R@10** | **MRR@10** | **MAP@10** | **R@10** |
| BM25 | – | 0.073 | 0.057 | 0.124 | 0.196 | 0.143 | 0.231 | 0.193 | 0.146 | 0.261 |
| Gemini-3.1-Flash-Lite | N/A | 0.156 | 0.110 | 0.217 | 0.383 | 0.332 | 0.526 | 0.210 | 0.163 | 0.265 |
| GPT-5.4-Nano | N/A | 0.102 | 0.077 | 0.141 | 0.314 | 0.258 | 0.365 | 0.246 | 0.196 | 0.290 |
| $\text{MiniLM}_{\text{L6-v2}}$ | 22M | 0.095 | 0.068 | 0.142 | 0.287 | 0.222 | 0.364 | 0.220 | 0.167 | 0.292 |
| + instr. | | 0.067 | 0.046 | 0.102 | 0.214 | 0.158 | 0.258 | 0.151 | 0.108 | 0.199 |
| $\text{BGE}_{\text{base-en-v1.5}}$ | 109M | 0.075 | 0.049 | 0.105 | 0.281 | 0.210 | 0.334 | 0.254 | 0.193 | 0.328 |
| + instr. | | 0.050 | 0.031 | 0.058 | 0.217 | 0.153 | 0.259 | 0.165 | 0.116 | 0.213 |
| all-mpnet-base-v2 | 110M | 0.096 | 0.067 | 0.157 | 0.326 | 0.253 | 0.406 | 0.273 | 0.208 | 0.362 |
| + instr. | | 0.042 | 0.027 | 0.073 | 0.192 | 0.141 | 0.252 | 0.126 | 0.092 | 0.205 |
| Contriever | 110M | 0.101 | 0.070 | 0.159 | 0.301 | 0.230 | 0.377 | 0.285 | 0.214 | 0.347 |
| + instr. | | 0.078 | 0.056 | 0.126 | 0.279 | 0.207 | 0.334 | 0.187 | 0.133 | 0.247 |
| $\text{E5}_{\text{base-v2}}$ | 110M | 0.106 | 0.072 | 0.137 | 0.315 | 0.244 | 0.393 | 0.306 | 0.243 | 0.392 |
| + instr. | | 0.103 | 0.071 | 0.144 | 0.295 | 0.226 | 0.378 | 0.307 | 0.239 | 0.385 |
| $\text{InstructOR}_{\text{base}}$ | 110M | 0.097 | 0.066 | 0.148 | 0.321 | 0.248 | 0.391 | 0.287 | 0.225 | 0.378 |
| + instr. | | 0.056 | 0.035 | 0.086 | 0.257 | 0.188 | 0.314 | 0.213 | 0.158 | 0.284 |
| TART-Contriever | 110M | 0.103 | 0.070 | 0.143 | 0.309 | 0.233 | 0.389 | 0.262 | 0.196 | 0.351 |
| + instr. | | 0.085 | 0.059 | 0.141 | 0.295 | 0.222 | 0.368 | 0.232 | 0.168 | 0.299 |
| $\text{BGE}_{\text{large-en-v1.5}}$ | 335M | 0.081 | 0.053 | 0.123 | 0.314 | 0.244 | 0.387 | 0.309 | 0.246 | 0.399 |
| + instr. | | 0.053 | 0.036 | 0.076 | 0.170 | 0.113 | 0.211 | 0.207 | 0.154 | 0.271 |
| $\text{E5}_{\text{large-v2}}$ | 335M | 0.110 | 0.081 | 0.163 | 0.311 | 0.243 | 0.398 | 0.311 | 0.244 | 0.405 |
| + instr. | | 0.106 | 0.077 | 0.153 | 0.302 | 0.231 | 0.369 | 0.321 | 0.251 | 0.399 |
| Nomic-v2 | 475M | 0.126 | 0.090 | 0.182 | 0.318 | 0.246 | 0.378 | 0.315 | 0.250 | 0.412 |
| + instr. | | 0.064 | 0.042 | 0.084 | 0.242 | 0.175 | 0.287 | 0.202 | 0.143 | 0.252 |
| BGE-M3 | 568M | 0.109 | 0.076 | 0.160 | 0.309 | 0.248 | 0.403 | 0.339 | 0.268 | 0.426 |
| + instr. | | 0.079 | 0.051 | 0.108 | 0.281 | 0.219 | 0.353 | 0.281 | 0.217 | 0.371 |
| $\text{Multilingual-E5}_{\text{large}}$ | 560M | 0.114 | 0.078 | 0.147 | 0.321 | 0.252 | 0.409 | 0.322 | 0.253 | 0.404 |
| + instr. | | 0.113 | 0.077 | 0.137 | 0.311 | 0.245 | 0.385 | 0.315 | 0.247 | 0.393 |
| $\text{Multilingual-E5}_{\text{large-instruct}}$ | 560M | 0.099 | 0.068 | 0.154 | 0.347 | 0.276 | 0.442 | 0.316 | 0.255 | 0.424 |
| + instr. | | 0.114 | 0.077 | 0.164 | 0.366 | 0.295 | 0.462 | 0.328 | 0.264 | 0.445 |
| $\text{Qwen3-Embedding}_{\text{0.6B}}$ | 596M | 0.123 | 0.084 | 0.164 | 0.337 | 0.266 | 0.424 | 0.325 | 0.251 | 0.421 |
| + instr. | | 0.135 | 0.089 | 0.178 | 0.349 | 0.275 | 0.435 | 0.355 | 0.278 | 0.463 |
| $\text{DRAMA}_{\text{1B}}$ | 1.24B | 0.103 | 0.072 | 0.151 | 0.309 | 0.243 | 0.386 | 0.327 | 0.258 | 0.422 |
| + instr. | | 0.052 | 0.033 | 0.077 | 0.258 | 0.191 | 0.293 | 0.207 | 0.150 | 0.261 |
| $\text{InstructOR}_{\text{xl}}$ | 1.24B | 0.146 | 0.102 | 0.198 | 0.398 | 0.326 | 0.496 | 0.381 | 0.312 | 0.493 |
| + instr. | | 0.089 | 0.054 | 0.114 | 0.283 | 0.213 | 0.352 | 0.234 | 0.169 | 0.305 |
| Lychee-Embed | 1.54B | 0.132 | 0.087 | 0.160 | 0.333 | 0.258 | 0.410 | 0.323 | 0.254 | 0.421 |
| + instr. | | 0.137 | 0.090 | 0.166 | 0.345 | 0.274 | 0.423 | 0.342 | 0.275 | 0.460 |
| $\text{Qwen3-Embedding}_{\text{4B}}$ | 4.02B | 0.181 | 0.126 | 0.230 | 0.378 | 0.315 | 0.494 | 0.358 | 0.290 | 0.468 |
| + instr. | | 0.212 | 0.145 | 0.269 | 0.388 | 0.321 | 0.510 | 0.401 | 0.331 | 0.534 |
| BGE-en-ICL | 7.11B | 0.174 | 0.123 | 0.223 | 0.400 | 0.329 | 0.505 | 0.340 | 0.274 | 0.424 |
| + instr. | | 0.191 | 0.134 | 0.255 | 0.388 | 0.317 | 0.501 | 0.366 | 0.298 | 0.479 |
| $\text{E5}_{\text{Mistral}}$ | 7.11B | 0.088 | 0.048 | 0.100 | 0.343 | 0.277 | 0.417 | 0.152 | 0.093 | 0.166 |
| + instr. | | 0.107 | 0.061 | 0.121 | 0.356 | 0.290 | 0.454 | 0.174 | 0.111 | 0.193 |
| Linq-Embed-Mistral | 7.11B | 0.239 | 0.171 | 0.292 | 0.395 | 0.322 | 0.488 | 0.385 | 0.317 | 0.510 |
| + instr. | | **0.268** | **0.196** | **0.341** | **0.406** | **0.336** | **0.531** | 0.408 | 0.343 | 0.550 |
| SFR-Embedding-Mistral | 7.11B | 0.199 | 0.145 | 0.252 | 0.386 | 0.313 | 0.495 | 0.328 | 0.265 | 0.437 |
| + instr. | | 0.233 | 0.165 | 0.298 | 0.403 | 0.334 | **0.531** | 0.365 | 0.299 | 0.493 |
| $\text{Qwen3-Embedding}_{\text{8B}}$ | 7.57B | 0.198 | 0.136 | 0.273 | 0.392 | 0.320 | 0.499 | 0.410 | 0.339 | 0.517 |
| + instr. | | 0.229 | 0.165 | 0.317 | 0.404 | 0.335 | 0.528 | **0.428** | **0.360** | **0.565** |
| $\text{Llama-Embed}_{\text{Nemotron-8B}}$ | 7.50B | 0.074 | 0.046 | 0.097 | 0.286 | 0.220 | 0.339 | 0.224 | 0.163 | 0.285 |
| + instr. | | 0.044 | 0.024 | 0.067 | 0.225 | 0.167 | 0.265 | 0.133 | 0.092 | 0.175 |

Table 11: Complete zero-shot results on MoraLink, together with BM25 and direct LLM ranking baselines. Embedding models are evaluated without a query instruction and with the general retrieval instruction (+ *instr.*); BM25 and direct LLM rows are single-setting baselines. `summary-only` and `summary+fable` columns use Gemini 3 Flash generated moral summaries. Best values in each metric column are bolded.

## K.1 Queries with One or Several Relevant Fables

Table 12 reports results separately for the 469 queries with one relevant fable and the 199 queries with several relevant fables. Queries with several relevant fables provide more opportunities for a first correct result, so their higher MRR@10 should not be interpreted as showing that these queries are inherently easier.

Table 12: Zero-shot results for queries with one relevant fable and queries with several relevant fables. Summary-based representations use Gemini 3 Flash summaries. Best values within each query group are bolded.

| Model | Queries | fable-only | | | summary-only | | | summary+fable | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MRR@10 | MAP@10 | R@10 | MRR@10 | MAP@10 | R@10 | MRR@10 | MAP@10 | R@10 |
| $\text{MiniLM}_{\text{L6-v2}}$ | One | 0.087 | 0.077 | 0.158 | 0.213 | 0.201 | 0.350 | 0.184 | 0.173 | 0.309 |
| | Several | 0.114 | 0.046 | 0.104 | 0.463 | 0.271 | 0.398 | 0.306 | 0.154 | 0.251 |
| + instr. | One | 0.063 | 0.054 | 0.117 | 0.155 | 0.143 | 0.243 | 0.124 | 0.111 | 0.211 |
| | Several | 0.075 | 0.026 | 0.067 | 0.352 | 0.195 | 0.292 | 0.216 | 0.099 | 0.172 |
| all-mpnet-base-v2 | One | 0.087 | 0.075 | 0.173 | 0.244 | 0.231 | 0.392 | 0.211 | 0.197 | 0.365 |
| | Several | 0.118 | 0.049 | 0.121 | 0.517 | 0.305 | 0.437 | 0.417 | 0.233 | 0.356 |
| + instr. | One | 0.042 | 0.033 | 0.083 | 0.140 | 0.127 | 0.239 | 0.107 | 0.094 | 0.215 |
| | Several | 0.045 | 0.014 | 0.049 | 0.317 | 0.174 | 0.282 | 0.169 | 0.086 | 0.181 |
| $\text{BGE}_{\text{base-en-v1.5}}$ | One | 0.067 | 0.056 | 0.111 | 0.193 | 0.179 | 0.305 | 0.210 | 0.197 | 0.345 |
| | Several | 0.093 | 0.033 | 0.091 | 0.489 | 0.283 | 0.403 | 0.359 | 0.183 | 0.286 |
| + instr. | One | 0.045 | 0.036 | 0.062 | 0.140 | 0.128 | 0.232 | 0.129 | 0.117 | 0.220 |
| | Several | 0.063 | 0.021 | 0.048 | 0.399 | 0.212 | 0.321 | 0.251 | 0.113 | 0.198 |
| $\text{BGE}_{\text{large-en-v1.5}}$ | One | 0.070 | 0.058 | 0.132 | 0.231 | 0.218 | 0.369 | 0.256 | 0.242 | 0.403 |
| | Several | 0.108 | 0.041 | 0.101 | 0.508 | 0.306 | 0.428 | 0.435 | 0.254 | 0.391 |
| + instr. | One | 0.051 | 0.043 | 0.090 | 0.108 | 0.094 | 0.190 | 0.173 | 0.160 | 0.279 |
| | Several | 0.057 | 0.017 | 0.043 | 0.314 | 0.158 | 0.262 | 0.286 | 0.139 | 0.253 |
| BGE-M3 | One | 0.099 | 0.088 | 0.179 | 0.236 | 0.225 | 0.388 | 0.272 | 0.259 | 0.439 |
| | Several | 0.132 | 0.049 | 0.116 | 0.481 | 0.305 | 0.438 | 0.495 | 0.290 | 0.396 |
| + instr. | One | 0.068 | 0.057 | 0.117 | 0.210 | 0.198 | 0.337 | 0.215 | 0.204 | 0.375 |
| | Several | 0.105 | 0.037 | 0.085 | 0.450 | 0.268 | 0.390 | 0.437 | 0.249 | 0.362 |
| BGE-en-ICL | One | 0.135 | 0.123 | 0.235 | 0.327 | 0.317 | 0.507 | 0.281 | 0.268 | 0.433 |
| | Several | 0.267 | 0.124 | 0.197 | 0.573 | 0.359 | 0.499 | 0.479 | 0.287 | 0.404 |
| + instr. | One | 0.147 | 0.135 | 0.267 | 0.314 | 0.304 | 0.503 | 0.303 | 0.291 | 0.495 |
| | Several | 0.297 | 0.130 | 0.228 | 0.562 | 0.348 | 0.497 | 0.517 | 0.313 | 0.442 |
| Contriever | One | 0.093 | 0.082 | 0.179 | 0.217 | 0.205 | 0.352 | 0.222 | 0.210 | 0.345 |
| | Several | 0.120 | 0.042 | 0.112 | 0.500 | 0.289 | 0.437 | 0.431 | 0.223 | 0.352 |
| + instr. | One | 0.078 | 0.068 | 0.145 | 0.189 | 0.175 | 0.296 | 0.145 | 0.134 | 0.247 |
| | Several | 0.080 | 0.027 | 0.080 | 0.492 | 0.283 | 0.422 | 0.285 | 0.130 | 0.246 |
| TART-Contriever | One | 0.090 | 0.079 | 0.158 | 0.216 | 0.204 | 0.367 | 0.199 | 0.187 | 0.352 |
| | Several | 0.133 | 0.050 | 0.108 | 0.529 | 0.303 | 0.441 | 0.411 | 0.218 | 0.349 |
| + instr. | One | 0.078 | 0.068 | 0.162 | 0.210 | 0.198 | 0.350 | 0.178 | 0.165 | 0.301 |
| | Several | 0.102 | 0.036 | 0.092 | 0.496 | 0.279 | 0.412 | 0.361 | 0.174 | 0.294 |
| $\text{DRAMA}_{\text{1B}}$ | One | 0.096 | 0.084 | 0.173 | 0.232 | 0.220 | 0.369 | 0.262 | 0.249 | 0.422 |
| | Several | 0.121 | 0.043 | 0.101 | 0.492 | 0.299 | 0.426 | 0.482 | 0.279 | 0.420 |
| + instr. | One | 0.050 | 0.039 | 0.087 | 0.192 | 0.179 | 0.279 | 0.168 | 0.154 | 0.267 |
| | Several | 0.055 | 0.018 | 0.052 | 0.414 | 0.221 | 0.324 | 0.300 | 0.142 | 0.249 |
| $\text{E5}_{\text{base-v2}}$ | One | 0.093 | 0.081 | 0.151 | 0.230 | 0.217 | 0.377 | 0.250 | 0.237 | 0.405 |
| | Several | 0.135 | 0.052 | 0.104 | 0.514 | 0.308 | 0.429 | 0.437 | 0.255 | 0.360 |
| + instr. | One | 0.095 | 0.083 | 0.164 | 0.212 | 0.200 | 0.362 | 0.249 | 0.237 | 0.392 |
| | Several | 0.120 | 0.043 | 0.097 | 0.489 | 0.287 | 0.416 | 0.445 | 0.244 | 0.368 |
| $\text{E5}_{\text{large-v2}}$ | One | 0.106 | 0.094 | 0.188 | 0.235 | 0.223 | 0.386 | 0.257 | 0.246 | 0.426 |
| | Several | 0.119 | 0.052 | 0.105 | 0.490 | 0.290 | 0.427 | 0.439 | 0.241 | 0.355 |
| + instr. | One | 0.101 | 0.090 | 0.177 | 0.223 | 0.210 | 0.354 | 0.265 | 0.251 | 0.405 |
| | Several | 0.115 | 0.045 | 0.098 | 0.486 | 0.281 | 0.406 | 0.452 | 0.251 | 0.384 |
| $\text{E5}_{\text{Mistral}}$ | One | 0.049 | 0.039 | 0.081 | 0.270 | 0.258 | 0.394 | 0.091 | 0.079 | 0.149 |
| | Several | 0.180 | 0.069 | 0.144 | 0.514 | 0.322 | 0.469 | 0.294 | 0.126 | 0.207 |
| + instr. | One | 0.057 | 0.047 | 0.098 | 0.284 | 0.274 | 0.439 | 0.104 | 0.093 | 0.175 |
| | Several | 0.226 | 0.096 | 0.174 | 0.524 | 0.328 | 0.487 | 0.337 | 0.154 | 0.237 |
| $\text{Multilingual-E5}_{\text{large}}$ | One | 0.099 | 0.087 | 0.164 | 0.241 | 0.231 | 0.401 | 0.268 | 0.256 | 0.416 |
| | Several | 0.148 | 0.055 | 0.108 | 0.510 | 0.303 | 0.428 | 0.449 | 0.247 | 0.378 |
| + instr. | One | 0.096 | 0.085 | 0.154 | 0.236 | 0.222 | 0.369 | 0.256 | 0.244 | 0.397 |
| | Several | 0.153 | 0.057 | 0.099 | 0.490 | 0.298 | 0.423 | 0.454 | 0.253 | 0.385 |

Continued on next page

Table 12 continued

| Model | Queries | fable-only | | | summary-only | | | summary+fable | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MRR@10 | MAP@10 | R@10 | MRR@10 | MAP@10 | R@10 | MRR@10 | MAP@10 | R@10 |
| Multilingual-E5$_{\text{large-instruct}}$ | One | 0.088 | 0.076 | 0.162 | 0.267 | 0.254 | 0.429 | 0.267 | 0.256 | 0.439 |
| | Several | 0.128 | 0.050 | 0.136 | 0.535 | 0.329 | 0.473 | 0.432 | 0.252 | 0.390 |
| + instr. | One | 0.097 | 0.085 | 0.171 | 0.290 | 0.278 | 0.454 | 0.276 | 0.265 | 0.467 |
| | Several | 0.154 | 0.059 | 0.148 | 0.544 | 0.335 | 0.480 | 0.451 | 0.259 | 0.393 |
| InstructOR$_{\text{base}}$ | One | 0.086 | 0.075 | 0.166 | 0.234 | 0.221 | 0.369 | 0.230 | 0.217 | 0.382 |
| | Several | 0.122 | 0.043 | 0.106 | 0.525 | 0.312 | 0.442 | 0.420 | 0.243 | 0.369 |
| + instr. | One | 0.053 | 0.042 | 0.098 | 0.178 | 0.164 | 0.292 | 0.172 | 0.159 | 0.292 |
| | Several | 0.063 | 0.019 | 0.059 | 0.444 | 0.242 | 0.367 | 0.310 | 0.158 | 0.264 |
| InstructOR$_{\text{xl}}$ | One | 0.127 | 0.114 | 0.215 | 0.323 | 0.310 | 0.495 | 0.321 | 0.309 | 0.507 |
| | Several | 0.191 | 0.073 | 0.157 | 0.577 | 0.363 | 0.498 | 0.524 | 0.318 | 0.458 |
| + instr. | One | 0.067 | 0.056 | 0.115 | 0.210 | 0.197 | 0.339 | 0.171 | 0.158 | 0.303 |
| | Several | 0.139 | 0.049 | 0.112 | 0.454 | 0.249 | 0.383 | 0.381 | 0.195 | 0.309 |
| Nomic-v2 | One | 0.117 | 0.105 | 0.203 | 0.233 | 0.220 | 0.345 | 0.255 | 0.243 | 0.418 |
| | Several | 0.147 | 0.053 | 0.134 | 0.517 | 0.307 | 0.455 | 0.455 | 0.265 | 0.398 |
| + instr. | One | 0.057 | 0.047 | 0.087 | 0.163 | 0.150 | 0.271 | 0.147 | 0.134 | 0.247 |
| | Several | 0.080 | 0.028 | 0.077 | 0.429 | 0.234 | 0.326 | 0.333 | 0.164 | 0.263 |
| Qwen3-Embedding$_{\text{0.6B}}$ | One | 0.108 | 0.096 | 0.181 | 0.251 | 0.240 | 0.401 | 0.253 | 0.240 | 0.422 |
| | Several | 0.159 | 0.055 | 0.123 | 0.537 | 0.327 | 0.478 | 0.492 | 0.275 | 0.418 |
| + instr. | One | 0.110 | 0.097 | 0.188 | 0.261 | 0.248 | 0.416 | 0.277 | 0.264 | 0.465 |
| | Several | 0.194 | 0.069 | 0.157 | 0.557 | 0.338 | 0.481 | 0.539 | 0.312 | 0.457 |
| Qwen3-Embedding$_{\text{4B}}$ | One | 0.145 | 0.131 | 0.237 | 0.315 | 0.305 | 0.490 | 0.301 | 0.290 | 0.480 |
| | Several | 0.266 | 0.115 | 0.215 | 0.529 | 0.338 | 0.503 | 0.491 | 0.291 | 0.441 |
| + instr. | One | 0.158 | 0.143 | 0.269 | 0.320 | 0.308 | 0.514 | 0.344 | 0.332 | 0.552 |
| | Several | 0.340 | 0.149 | 0.271 | 0.548 | 0.352 | 0.502 | 0.535 | 0.330 | 0.492 |
| Qwen3-Embedding$_{\text{8B}}$ | One | 0.145 | 0.132 | 0.277 | 0.314 | 0.304 | 0.497 | 0.351 | 0.339 | 0.531 |
| | Several | 0.324 | 0.145 | 0.264 | 0.573 | 0.358 | 0.505 | 0.548 | 0.338 | 0.486 |
| + instr. | One | 0.180 | 0.166 | 0.316 | **0.333** | **0.323** | 0.529 | **0.372** | **0.361** | **0.593** |
| | Several | 0.344 | 0.162 | 0.321 | 0.571 | 0.364 | **0.525** | **0.559** | **0.358** | **0.501** |
| Linq-Embed-Mistral | One | 0.188 | 0.175 | 0.301 | 0.314 | 0.302 | 0.478 | 0.337 | 0.325 | 0.539 |
| | Several | 0.357 | 0.162 | 0.271 | **0.587** | 0.368 | 0.514 | 0.499 | 0.298 | 0.441 |
| + instr. | One | **0.208** | **0.196** | **0.343** | 0.332 | 0.321 | **0.535** | 0.360 | 0.347 | 0.571 |
| | Several | **0.407** | **0.196** | **0.336** | 0.581 | **0.369** | 0.521 | 0.521 | 0.333 | 0.500 |
| SFR-Embedding-Mistral | One | 0.160 | 0.148 | 0.254 | 0.306 | 0.294 | 0.484 | 0.283 | 0.270 | 0.463 |
| | Several | 0.292 | 0.136 | 0.247 | 0.574 | 0.358 | 0.520 | 0.433 | 0.254 | 0.375 |
| + instr. | One | 0.177 | 0.165 | 0.307 | 0.331 | 0.320 | **0.535** | 0.312 | 0.299 | 0.512 |
| | Several | 0.366 | 0.164 | 0.277 | 0.574 | 0.366 | 0.523 | 0.490 | 0.298 | 0.449 |
| Llama-Embed$_{\text{Nemotron-8B}}$ | One | 0.061 | 0.050 | 0.102 | 0.216 | 0.203 | 0.320 | 0.171 | 0.157 | 0.281 |
| | Several | 0.104 | 0.037 | 0.084 | 0.451 | 0.259 | 0.385 | 0.349 | 0.177 | 0.293 |
| + instr. | One | 0.037 | 0.027 | 0.072 | 0.167 | 0.157 | 0.258 | 0.107 | 0.093 | 0.179 |
| | Several | 0.061 | 0.017 | 0.054 | 0.361 | 0.191 | 0.282 | 0.195 | 0.088 | 0.167 |
| Lychee-Embed | One | 0.099 | 0.087 | 0.162 | 0.246 | 0.234 | 0.397 | 0.268 | 0.255 | 0.439 |
| | Several | 0.211 | 0.088 | 0.154 | 0.539 | 0.315 | 0.441 | 0.452 | 0.252 | 0.379 |
| + instr. | One | 0.102 | 0.090 | 0.168 | 0.265 | 0.251 | 0.416 | 0.287 | 0.275 | 0.480 |
| | Several | 0.219 | 0.089 | 0.159 | 0.534 | 0.328 | 0.439 | 0.470 | 0.274 | 0.412 |

### K.2 Word Overlap Between Queries and Summaries

We lowercase each text and remove punctuation, one-character tokens, and stopwords such as "the" and "and." For every query and relevant document, we calculate the proportion of their distinct remaining words that appear in both texts. Table 13 reports the average word overlap across the 1,085 relevant query-document pairs.

To test whether summaries improve retrieval partly because they reuse words from the query, we compare every summary-based experiment with the corresponding experiment using the original fables. Within each comparison, we keep the retrieval model, query instructions, and all 668 queries unchanged. We change only the candidate texts from original fables to generated summaries.

We make 159 comparisons in total. For Gemini, we evaluate 23 retrieval models with three summary types, giving 69 comparisons. For Gemma, we evaluate 10 retrieval models with three Gemma versions and three summary types, giving 90 comparisons.

For each query, we measure two changes: how much its word overlap increases and how much the rank of its first relevant result improves. We use Spearman correlation to measure whether these changes tend to move together. The correlation is positive in all 159 comparisons, meaning that queries with larger increases in word overlap also tend to receive larger improvements in rank. The median correlation is 0.344, with values ranging from 0.051 to 0.566. These results show that closer wording may help explain the improvements, but they do not show that it caused them. The corresponding retrieval results for these three summary types are reported in Table 5 in Appendix F.

| Source | Summary type | Mean overlap | Median $\rho$ |
|---|---|---|---|
| Original fable | | 0.007 | |
| Gemini 3 Flash | direct moral | 0.076 | **0.491** |
| | conceptual abstract | 0.058 | 0.417 |
| | cot proverb | **0.103** | 0.489 |
| Gemma 4 E2B | direct moral | 0.016 | 0.216 |
| | conceptual abstract | 0.012 | 0.253 |
| | cot proverb | 0.016 | 0.263 |
| Gemma 4 E4B | direct moral | 0.025 | 0.262 |
| | conceptual abstract | 0.015 | 0.327 |
| | cot proverb | 0.015 | 0.290 |
| Gemma 4 31B | direct moral | 0.052 | 0.361 |
| | conceptual abstract | 0.040 | 0.340 |
| | cot proverb | 0.060 | 0.381 |

Table 13: Word overlap between moral queries and relevant documents. $\rho$ is the median Spearman correlation between increased word overlap and improved reciprocal rank across matched retrieval settings. Best values are bolded.

## L Fine-tuning Data Preprocessing

We use two sources for fine-tuning: STORAL, a dataset of real human-authored stories paired with morals, and TF1-EN-3M, a large synthetic corpus of LLM-generated fables. Both datasets require careful preprocessing to avoid lexical leakage and near-duplicate content that could artificially inflate performance on MoraLink. We describe the preprocessing pipeline for each source below.

### L.1 STORAL Preprocessing

STORAL (Guan et al., 2022) provides short stories paired with explicit moral statements. To avoid leakage into the MoraLink evaluation set, we apply two filtering steps.

First, we remove any STORAL pair whose moral exactly matches a moral in MoraLink. Matching is performed after light normalization: lowercasing, trimming whitespace, and collapsing repeated whitespace. This removes 140 pairs.

Second, we apply near-duplicate fable detection to avoid content overlap between STORAL stories and MoraLink fables. Since duplicated or lightly adapted fables typically share their opening sentences, we compare the beginning of each STORAL story against the beginning of every MoraLink fable. Both texts are lowercased and truncated to their first 400 characters, and a STORAL pair is removed if any MoraLink fable obtains a character-overlap similarity score of at least 0.6. This process flags 44 additional pairs, which we manually verify to confirm near-duplicate story content before removal.

After both filtering steps (exact moral matches and near-duplicate fables), the final STORAL training source contains 1,595 story–moral pairs. These correspond to 1,530 unique normalized morals. The difference arises because some normalized moral statements are paired with multiple STORAL stories: 57 duplicate-moral groups account for 122 stories. We keep these examples and treat stories with the same normalized moral as same-label positives during contrastive training.

For every fine-tuning experiment, 10% of moral groups (not individual pairs) are reserved for vali-

dation, ensuring that stories sharing the same moral remain in the same split.

### L.2 TF1-EN-3M Preprocessing

TF1-EN-3M (Nadas et al., 2025) contains three million LLM-generated fables produced from a fixed pool of 100 moral prompts. Each example contains a generated fable together with the moral prompt used to generate it. We treat this prompt as the intended moral for the generated fable.

Since MoraLink is sensitive to lexical leakage, we construct a low-overlap derivative of TF1-EN-3M. For each moral in TF1-EN-3M, we first remove generated fables that contain the moral verbatim or near-verbatim, including cases where the fable explicitly states the moral as a concluding lesson. Among the remaining fables, we select those with the lowest content-word intersection-over-union (IoU) between the moral and the fable text. This yields 100 morals, each paired with 20 fables, for 2,000 moral–fable pairs.

To avoid direct overlap with MoraLink, we remove entire moral groups whose morals exactly match a MoraLink moral, as well as groups containing any generated fable with an explicit MoraLink moral phrase embedded in its text. This removes 5 moral groups, leaving 95 morals and 1,900 generated fables.

We experiment with multiple training set sizes: 20 morals with 10 fables each (200 pairs), 50 morals with 10 fables each (500 pairs), all 95 morals with 10 fables each (950 pairs), and all 95 morals with 20 fables each (1,900 pairs). As with STORAL, 10% of moral groups are reserved for validation in every experiment, and all fables associated with the same moral remain in the same split to prevent label leakage.

## M Fine-tuning Results and Hyperparameters

**Hyperparameters.** All fine-tuned models are trained on moral–story pairs without query instructions and without generated summaries; generated summaries are used only at evaluation time. We evaluate four training-data settings: real STORAL data, synthetic TF1-EN-3M data, a synthetic-to-real curriculum, and a shuffled mixture of synthetic and real examples. We use the in-batch contrastive objective described in §4.3, with false-negative masking based on exact moral labels. All runs use a 10% validation split at the moral-group level, so examples sharing the same moral remain in the same split. Each experiment is repeated with seeds 42, 43, and 44. Shared optimization hyperparameters are reported in Table 16, and model-specific batch sizes, learning rates, and adaptation methods are reported in Table 14.

**Full Fine-tuning Results.** Figure 6 shows Recall@10 results for the fine-tuning experiments under the same inference-time configuration used in Figure 2. Figure 7 shows the results of the fine-tuned models using the rest of the inference-time configurations.

## N Technical Details

We run the experiments on an NVIDIA RTX 3090 GPU. We report GPU-hours rather than wall-clock time because several runs were executed in parallel across available GPUs. Including experimentation and hyperparameter tuning, we estimate that the zero-shot experiments took approximately 20 GPU-hours, and the fine-tuning experiments approximately 55 GPU-hours. For prompting and accessing LLMs, we use Agno, which provides a modular, reproducible framework for LLM evaluation across multiple providers by wrapping their APIs in a unified layer. Generating the reusable Gemini 3 Flash moral summaries for all 709 fables cost approximately $4 across the three prompt variants reported in the paper. For direct LLM ranking, we evaluate three document configurations for each of two models using the providers' batch APIs. Based on the recorded token usage, these six runs cost approximately $35.5 in total: about $20.0 for Gemini-3.1-Flash-Lite and $15.5 for GPT-5.4-Nano. We use *SentenceTransformer* to load and run embedding models, and *PyTorch* for fine-tuning.

## O Artifacts

We detail the artifacts we use, their usage, and licenses in Table 15.

## P Model Checkpoints

In Table 17, we present the checkpoints (or snapshots) used in this work and the models' sizes and licenses.

## Q AI assistants

We used AI assistants (e.g., ChatGPT) to support code formatting, provide phrasing suggestions, and

| Model | Param | Batch | Grad. accum. | Eff. batch | LR | Adaptation |
|---|---|---|---|---|---|---|
| $\text{MiniLM}_{\text{L6-v2}}$ | 22M | 16 | 1 | 16 | $2\times10^{-5}$ | full fine-tuning |
| BGE-M3 | 568M | 8 | 2 | 16 | $2\times10^{-5}$ | full fine-tuning |
| $\text{Qwen3-Embedding}_{\text{0.6B}}$ | 596M | 4 | 4 | 16 | $1\times10^{-4}$ | LoRA, $r = 64$, $\alpha = 128$ |
| $\text{InstructOR}_{\text{xl}}$ | 1.24B | 4 | 4 | 16 | $2\times10^{-5}$ | LoRA, $r = 16$, $\alpha = 32$ |
| $\text{Qwen3-Embedding}_{\text{4B}}$ | 4.02B | 2 | 8 | 16 | $1\times10^{-4}$ | LoRA, $r = 64$, $\alpha = 128$ |

Table 14: Model-specific fine-tuning hyperparameters. Gradient checkpointing is enabled for all models except $\text{MiniLM}_{\text{L6-v2}}$. LoRA uses dropout 0.05; Qwen models target the attention projection matrices, while $\text{InstructOR}_{\text{xl}}$ uses all linear modules.

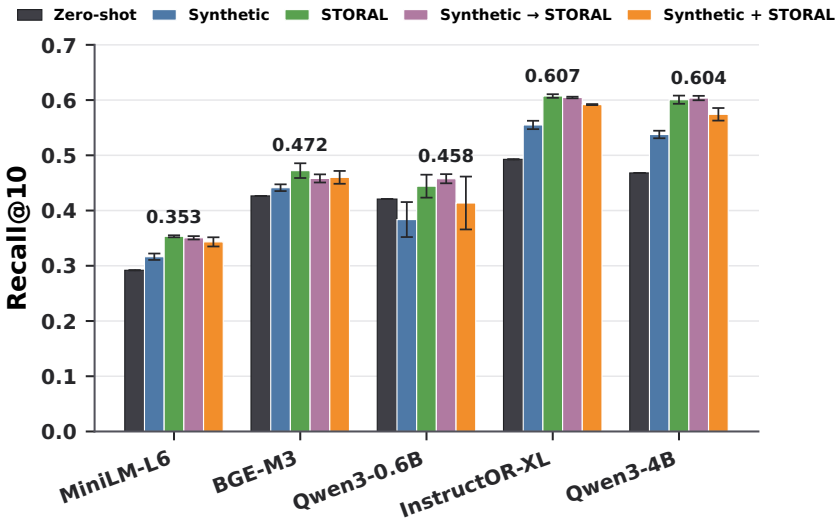



Figure 6: Recall@10 results for the fine-tuning experiments under the same inference-time configuration used in Figure 2. Results are averaged over three random seeds, with error bars denoting standard deviation.

| Artifact | Type | License | Usage |
|---|---|---|---|
| MORABLES[a] | Dataset | CC-BY-NC-4.0 | Benchmark source |
| STORAL[b] | Dataset | Not specified | Fine-tuning data |
| TF1-EN-3M[c] | Dataset | MIT | Synthetic fine-tuning data |
| Agno[d] | Framework | Apache-2.0 | LLM prompting |
| SentenceTransformers[e] | Package | Apache-2.0 | Embedding models |
| PyTorch[f] | Package | BSD-style | Fine-tuning |

[a] https://huggingface.co/datasets/cardiffnlp/Morables
[b] https://huggingface.co/datasets/Jiann/STORAL
[c] https://huggingface.co/datasets/klusai/ds-tf1-en-3m
[d] https://www.agno.com/
[e] https://huggingface.co/sentence-transformers
[f] https://github.com/pytorch/pytorch

Table 15: Artifacts used during experiments, along with their license and usage explanation.

assist with LaTeX styling during writing. All outputs were reviewed and edited by the authors. No content was directly generated or used without human verification.

| Parameter | Value |
|---|---|
| Max epochs | 10 |
| Validation split | 10% |
| Early stopping metric | MRR@10 |
| Early stopping patience | 3 epochs |
| Contrastive temperature $\tau$ | 0.05 |
| Effective batch size | 16 |
| Max sequence length | 512 |
| Seeds | 42, 43, 44 |

Table 16: Shared fine-tuning hyperparameters. The effective batch size is kept fixed at 16 across models using gradient accumulation; model-specific physical batch sizes and learning rates are listed in Table 14.

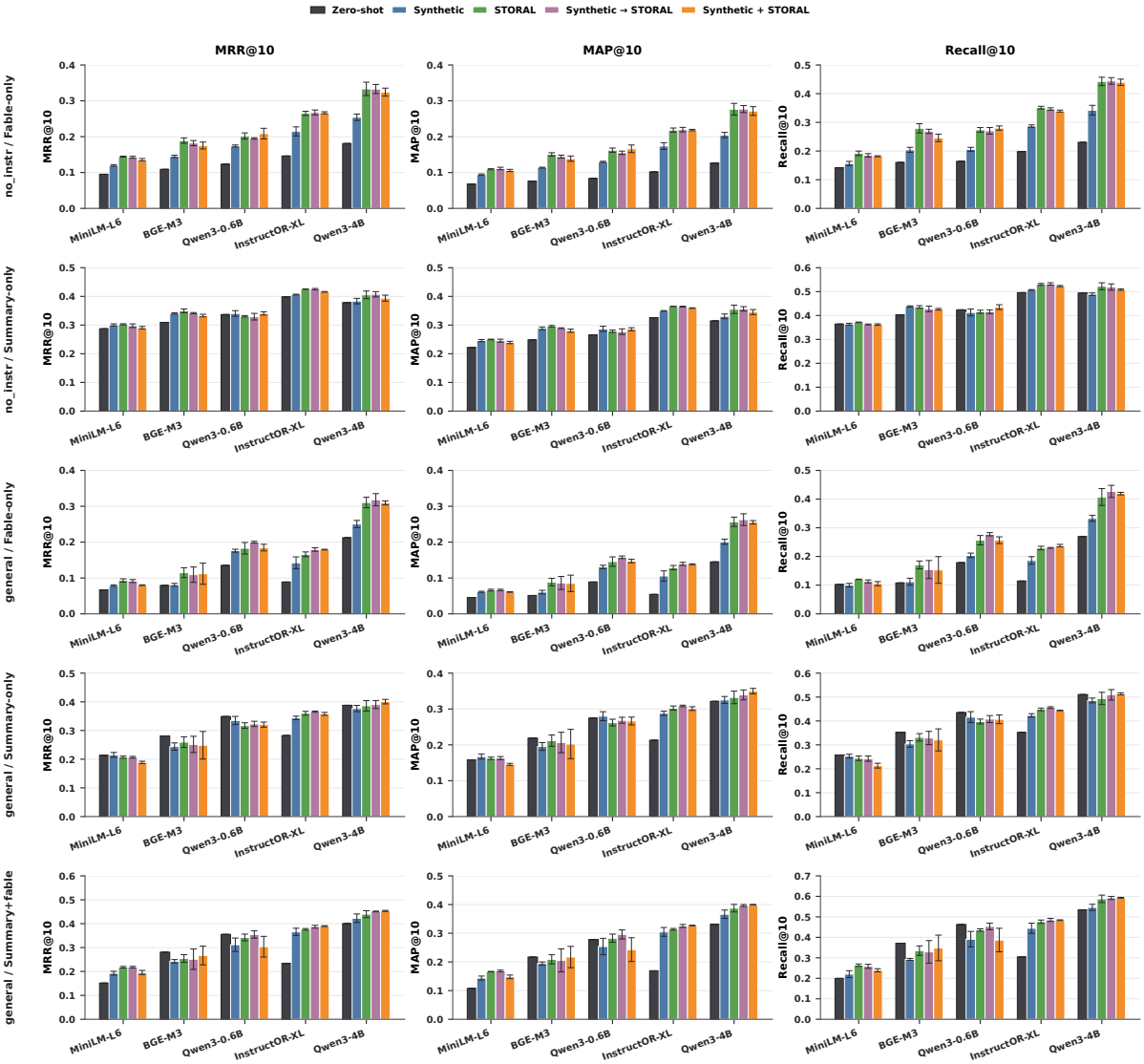



Figure 7: Fine-tuning results under additional fixed inference-time configurations using Gemini-generated `direct_moral` summaries. Rows correspond to document representation and query-instruction settings; columns report MRR@10, MAP@10, and Recall@10. The main paper fine-tuning figure reports the best fixed configuration, while this appendix figure shows complementary configurations.

| Model | Checkpoint | License | # Param |
|---|---|---|---|
| MiniLM$_{\text{L6-v2}}$[a] | sentence-transformers/all-MiniLM-L6-v2 | Apache-2.0 | 22M |
| all-mpnet-base-v2[b] | sentence-transformers/all-mpnet-base-v2 | Apache-2.0 | 110M |
| BGE$_{\text{base-en-v1.5}}$[c] | BAAI/bge-base-en-v1.5 | MIT | 109M |
| BGE$_{\text{large-en-v1.5}}$[d] | BAAI/bge-large-en-v1.5 | MIT | 335M |
| BGE-M3[e] | BAAI/bge-m3 | MIT | 568M |
| BGE-en-ICL[f] | BAAI/bge-en-icl | MIT | 7.11B |
| Contriever[g] | facebook/contriever | CC-BY-NC 4.0 | 110M |
| TART-Contriever[h] | orionweller/tart-dual-contriever-msmarco | CC-BY-NC 4.0 | 110M |
| DRAMA$_{\text{1B}}$[i] | facebook/drama-1b | CC-BY-NC 4.0 | 1.24B |
| E5$_{\text{base-v2}}$[j] | intfloat/e5-base-v2 | MIT | 110M |
| E5$_{\text{large-v2}}$[k] | intfloat/e5-large-v2 | MIT | 335M |
| E5$_{\text{Mistral}}$[l] | intfloat/e5-mistral-7b-instruct | MIT | 7.11B |
| Multilingual-E5$_{\text{large}}$[m] | intfloat/multilingual-e5-large | MIT | 560M |
| Multilingual-E5$_{\text{large-instruct}}$[n] | intfloat/multilingual-e5-large-instruct | MIT | 560M |
| InstructOR$_{\text{base}}$[o] | hkunlp/instructor-base | Apache-2.0 | 110M |
| InstructOR$_{\text{xl}}$[p] | hkunlp/instructor-xl | Apache-2.0 | 1.24B |
| Nomic-v2[q] | nomic-ai/nomic-embed-text-v2-moe | Apache-2.0 | 475M |
| Qwen3-Embedding$_{\text{0.6B}}$[r] | Qwen/Qwen3-Embedding-0.6B | Apache-2.0 | 596M |
| Qwen3-Embedding$_{\text{4B}}$[s] | Qwen/Qwen3-Embedding-4B | Apache-2.0 | 4.02B |
| Qwen3-Embedding$_{\text{8B}}$[t] | Qwen/Qwen3-Embedding-8B | Apache-2.0 | 7.57B |
| Linq-Embed-Mistral[u] | Linq-AI-Research/Linq-Embed-Mistral | CC-BY-NC 4.0 | 7.11B |
| SFR-Embedding-Mistral[v] | Salesforce/SFR-Embedding-Mistral | CC-BY-NC 4.0 | 7.11B |
| Llama-Embed$_{\text{Nemotron-8B}}$[w] | nvidia/llama-embed-nemotron-8b | customized-nscl-v1 | 7.50B |
| Lychee-Embed[x] | vec-ai/lychee-embed | Apache-2.0 | 1.54B |
| Gemini 3 Flash[y] | gemini-3-flash-preview | Proprietary | Undisclosed |
| Gemini 3.1 Flash-Lite[z] | gemini-3.1-flash-lite-preview | Proprietary | Undisclosed |
| GPT-5.4 Nano[aa] | gpt-5.4-nano | Proprietary | Undisclosed |
| Gemma 4 E2B[ab] | mlx-community/gemma-4-e2b-it-4bit | Gemma | 2B |
| Gemma 4 E4B[ac] | mlx-community/gemma-4-e4b-it-4bit | Gemma | 4B |
| Gemma 4 31B[ad] | mlx-community/gemma-4-31b-it-4bit | Gemma | 31B |
| Claude Sonnet 4.5[ae] | claude-sonnet-4-5 | Proprietary | Undisclosed |

[a] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[b] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[c] https://huggingface.co/BAAI/bge-base-en-v1.5
[d] https://huggingface.co/BAAI/bge-large-en-v1.5
[e] https://huggingface.co/BAAI/bge-m3
[f] https://huggingface.co/BAAI/bge-en-icl
[g] https://huggingface.co/facebook/contriever
[h] https://github.com/facebookresearch/tart/tree/main?tab=readme-ov-file#pre-trained-checkpoints
[i] https://huggingface.co/facebook/drama-1b
[j] https://huggingface.co/intfloat/e5-base-v2
[k] https://huggingface.co/intfloat/e5-large-v2
[l] https://huggingface.co/intfloat/e5-mistral-7b-instruct
[m] https://huggingface.co/intfloat/multilingual-e5-large
[n] https://huggingface.co/intfloat/multilingual-e5-large-instruct
[o] https://huggingface.co/hkunlp/instructor-base
[p] https://huggingface.co/hkunlp/instructor-xl
[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/Linq-AI-Research/Linq-Embed-Mistral
[v] https://huggingface.co/Salesforce/SFR-Embedding-Mistral
[w] https://huggingface.co/nvidia/llama-embed-nemotron-8b
[x] https://huggingface.co/vec-ai/lychee-embed
[y] https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview
[z] https://ai.google.dev/gemini-api/docs/models
[aa] https://platform.openai.com/docs/models
[ab] https://huggingface.co/mlx-community/gemma-4-e2b-it-4bit
[ac] https://huggingface.co/mlx-community/gemma-4-e4b-it-4bit
[ad] https://huggingface.co/mlx-community/gemma-4-31b-it-4bit
[ae] https://platform.claude.com/docs/en/about-claude/models/model-ids-and-versions

Table 17: Checkpoints and API models used during experiments, their licenses, and their number of parameters. Parameter counts for proprietary API models are not publicly disclosed.