LG-ANNA-EMBEDDING 技术报告

LG AI Research*

June 10, 2025

ABSTRACT

本报告提出了一个统一的基于指令的框架，用于学习优化信息检索（IR）和非 IR 任务的广义文本嵌入。我们的方法基于一个仅解码的大型语言模型（Mistral-7B）构建，结合了上下文学习、软监督和自适应的困难负样本挖掘，生成上下文感知的嵌入，而无需针对任务的微调。结构化的指令和少量示例用于指导模型跨越多样化的任务，使其在分类、语义相似性、聚类和重排序基准测试中表现出色。为了改善语义区分，我们采用了一个软标签框架，其中从高性能密集检索器和重排序器中提取的连续相关性评分被用作细粒度的监督信号。此外，我们引入了自适应基于边距的困难负样本挖掘，该方法根据样本与正样本的相似性过滤出语义模糊的负样本，从而增强训练的稳定性和检索的鲁棒性。我们的模型在新引入的 MTEB（英语，v2）基准测试中进行了评估，覆盖了七个类别下的 41 个任务。结果表明，我们的方法具有强大的泛化能力，并在 Borda 评分中名列表现最佳的模型之一，优于一些更大或完全微调的基线。这些发现突出了结合上下文提示、软监督和自适应采样以进行可扩展、高质量嵌入生成的有效性。

1 引言

文本嵌入被用作多种自然语言处理（NLP）任务的基础表示，包括检索、分类、重排序和语义相似性。随着大型语言模型（LLM）的兴起，最近的研究越来越多地集中在将 LLM 适应为通用文本编码器。现有的方法如 Llama2Vec [1] 、E5-Mistral [2] 和 NV-Embed [3] 通过使用合成数据、架构修改以及精心设计的池化策略对仅解码器 LLM 进行微调，展示出明显的性能提升。然而，这些模型通常依赖于全模型微调或结构性变化，这可能会限制可扩展性并忽视 LLM 固有的泛化能力。

针对这些限制，人们对使用上下文学习（ICL）生成嵌入而不修改模型权重表现出了浓厚的兴趣。最近的方法如 BGE-en-icl [4] 表明，结构化提示和少量示例可以有效地指导 LLMs 在各种任务中生成高质量的嵌入。这一范式转变强调了灵活性和资源效率，使得模型能够通过最小的调整适应多种使用场景。

另一个挑战涉及到用于监督模型训练的标签和信号的可靠性。传统的对比学习方法经常受到噪音或不充分的标签的影响，特别是由于存在错误的负例。为了解决这个问题，我们通过从高性能检索管道中提取连续值相关性评分来结合软标签策略。这些软目标提供了细致的语义信号，指导模型更好地与人类的相似性判断对齐。此外，我们采用自适应硬负挖掘技术，根据负例与对应正例的相似度动态选择信息量大的负例，减少语义重叠样本的影响。

在本研究中，我们提出了 LG-ANNA-Embedding，一种建立在 Mistral-7B 基础上的通用嵌入模型，并通过上下文学习、软监督和自适应硬负采样进行增强。我们的方法在不修改底层模型架构的情况下，实现了对广泛的检索和非检索任务的强大泛化能力。评估基于新引入的 MTEB（英语，v2）基准 [5] ，我们的模型在 41 项不同任务中表现出竞争力，在 Borda 得分排名中位居前列。我们的研究结果强调了结合上下文大语言模型功能与高质量训练信号以实现可扩展且准确嵌入生成的有效性。

最近的研究越来越多地探索使用大型语言模型（LLMs）作为文本嵌入任务的骨干编码器。这一转变在 Llama2Vec 等模型中得到了体现，该模型引入了两个预训练目标，以更好地使 LLMs 与嵌入任务对齐，从而在 BEIR 等检索基准上取得了显著的性能提升。然而，其在 MTEB 排行榜上的表现仍然相对温和。其他模型如 E5-Mistral、Linq 和 Gecko 利用大规模合成数据有效地微调 LLMs，在检索和非检索任务中都取得了优异

*The complete list of authors who contributed to this work can be found in Appendix A.

www.xueshuxiangzi.com

的结果。NV-Embed 通过引入潜在注意力池化机制和两阶段训练策略以减轻假阴性，进一步推进了这项工作，从而显著提高了检索的鲁棒性。

尽管基于 LLM 的嵌入模型表现强劲，之前的方法通常依赖于架构的修改——例如用双向机制替换单向注意力——或是对整个模型进行微调。这些策略尽管有效，但往往忽视了 LLM 固有的泛化能力，并需要大量的计算资源。然而，最近已经转向利用上下文学习（ICL）作为一种更高效的替代方法。像 BGE-en-icl [4] 这样的模型表明，可以使用特定任务的提示和演示来调节 LLM 进行嵌入生成，而无需修改模型权重。这一新兴的范式凸显了 ICL 的潜力，可以构建灵活和自适应的嵌入系统，使其在任务间具有良好的泛化能力，同时尽量减少训练和部署的成本。

近来在嵌入模型训练方面的进展强调了软标签和硬负采样的重要性。特别是，从高容量重排序模型中获得的软监督信号（通常通过教师-学生框架实现）已被证明在引导嵌入模型更好地与语义相似性目标对齐方面是有效的。同时，硬负采样策略在缩小正负对之间的语义差距和减轻对比学习过程中假负样本的风险方面起到了关键作用。基于这些见解，我们的目标是开发一种广义嵌入模型，通过利用上下文学习来保留 LLM 的固有优势。我们的方法无需进行架构修改或大规模微调，旨在实现对检索和非检索任务的强适应性，同时保持训练和部署的效率。

2 训练数据集

检索数据集 根据在英文 MTEB 排行榜上表现优秀的模型所采用的常见方法 [5]，我们使用了一组公开可用的检索数据集，包括 MSMARCO [6]、HotpotQA [7]、Natural Questions [8]、SQuAD [9]、ELI5 [10]、ArguAna [11]、FiQA [12]、FEVER [13]、Quora Duplicate Questions [14]。这些检索数据集不仅标注了硬标签，如正负对，还通过软标签进行了增强，软标签将在下节详细描述 **??**。此外，它们还通过复杂的硬负采样策略进行了处理，该策略也将在随后章节中详细说明 2.1。

与其他模型 [3, 4] 类似，我们也结合了来自非检索基准的数据集，特别是那些与分类、聚类、重排序和语义文本相似度（STS）相关的 MTEB 任务。重要的是，我们仅使用训练数据并省去测试数据，以保持评估的完整性。

对于分类任务，我们使用了一系列基准数据集，包括 AmazonCounterfactualClassification [15]、AmazonReviewsClassification [16]、Banking77Classification [17]、EmotionClassification [18]、ImdbClassification [19]、MTOPIntentClassification [20]、ToxicConversationsClassification [21] 和 TweetSentimentExtractionClassification [22]。

我们使用了一系列聚类数据集，例如 ArxivClustering [2]，BiorxivClustering [3]，MedrxivClustering [4]，TwentyNewsgroupsClustering [23]，RedditClustering [24]，StackExchangeClustering [24]。

对于重新排序任务，我们将 SciDocsRR [25] 和 StackOverflowDupQuestions [26] 纳入我们的训练语料库。此外，我们还包括三种广泛使用的语义文本相似性数据集的训练划分——STS12 [27]、STS22 [28] 和 STS-Benchmark [28]——以增强模型捕捉细粒度语义关系的能力。

我们在表 **??** 中总结了训练数据集及每个数据集使用的指令，这些指令改编自之前的工作 [2, 4]。

数据转换 由于公开发布的语义文本相似性（STS）训练数据集有限，我们在这项研究中重新利用自然语言推理（NLI）数据作为 STS 监督的替代。

如图 1 所示，我们从一组 NLI 句子对开始，每个句子对被标记为“蕴含”、“中性”或“矛盾”。根据原始 NLI 标签，每个句子对被映射到一个近似的相似度得分：蕴含句子对被分配高相似度，矛盾句子对则被视为具有低相似度的负例，而中性句子对则由于其语义不明确被舍弃。

为了确保与 STS 任务格式的一致性，由于语义的模糊性，中性对被丢弃，而保留了蕴含和矛盾对。然后，将得到的句子对重新格式化以符合标准 STS 训练数据的结构，并被纳入对比学习的训练语料库中。还加入了柔性相似度分数，如 **??** 节所述。

为了处理单个查询与多个正段落相关联的情况，每个查询-正例对被单独分割成单独的训练实例。例如，如果一个查询有两个正段落，`["A1", "A2"]`，它被转化为两个不同的例子：`query: "A", positive: "A1"` 和 `query: "A", positive: "A2"`。之后，应用去重步骤以从数据集中去除冗余的 `(query, positive)` 对。这个过程通过减少不必要的数据冗余，消除了重复条目，并有助于提高训练效率。

[2] https://www.kaggle.com/datasets/Cornell-University/arxiv

[3] https://api.biorxiv.org/

[4] https://api.medrxiv.org/

www.xueshuxiangzi.com

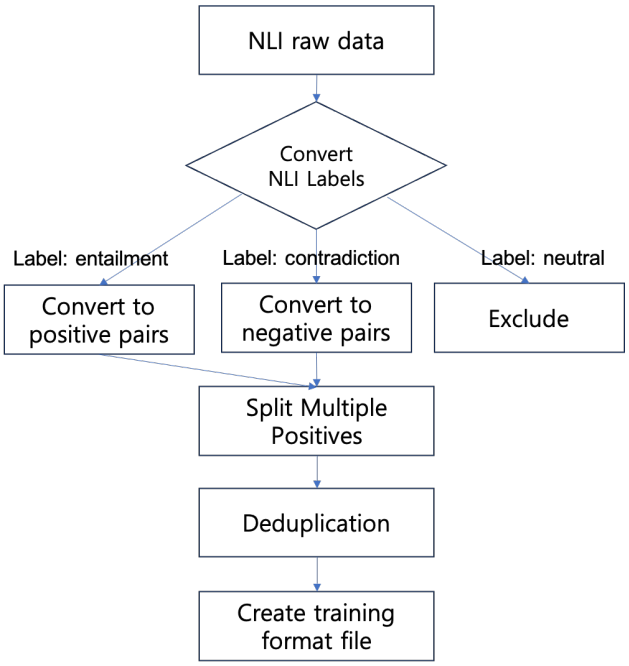

Figure 1: 流程图说明了从自然语言推理（NLI）数据转换为 STS 风格训练实例的过程。标记为“蕴含”的句子对被保留并分配高相似度分数，而“矛盾”对则被转换为具有低相似度分数的负对。由于语义对齐不明确，标记为“中性”的对被丢弃。生成的实例被重新格式化以匹配 STS 输入格式，并用作相似性建模的训练数据。

为了增强我们的嵌入模型的语义区分能力，我们使用检索管道系统生成的相似性评分进行软标签监督。这些连续值的相关信号作为软目标，指导嵌入模型通过对比学习来学习细微的语义关系。

与提供二进制相关性判断的硬标签不同，软标签通过捕捉教师模型在不同候选项上的置信度，提供了更丰富的信息。这种方法与最近的理论见解一致：Mandal 等人表明，与硬标签训练目标相比，软标签监督使得学生模型更有效地泛化，并且需要更少的神经元来逼近教师模型的决策边界。

在本研究中，为了进行软标签处理，我们采用了内部开发的检索系统，即 ANNA IR 管道，如图 2 所示。该管道集成了词汇检索和语义检索组件，并配有一个重排序器以生成相关性评分。ANNA IR 管道是一个两阶段检索系统。在第一阶段，使用词汇搜索和语义检索（ANNA 检索）XMAHTX_1 来检索候选段落。该混合方法确保了对表面层次和语义匹配的覆盖。在第二阶段，使用 ANNA 重排序器对检索到的候选段落进行重新排序，依据与查询的语境对齐分配细粒度相关性评分。最终排序列表用于后续任务或软标签监督。

我们使用三个评分函数来估计查询 q 与文档 d 之间的相关性：

- 词汇搜索（BM25）：
  BM25 分数是基于精确的标记重叠计算的，使用的公式为：

$$\text{Lexical score}(d,q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,d)\cdot(k_1+1)}{f(q_i,d)+k_1\cdot\left(1-b+b\cdot\frac{\text{len}}{\text{avglen}}\right)}$$

- 语义搜索（密集检索）：
  语义相似性是通过密集查询和文档嵌入之间的点积来估算的：

$$\text{Semantic score}(q,d) = q\cdot d = \sum_{i=1}^{n} q_i \cdot d_i$$

- 重排序器（交叉编码器）：
  细粒度的语义相关性是通过一个交叉编码器模型 f_θ 捕获的，该模型将查询和文档都作为输入：

$$\text{Reranker score}(q,d) = f_\theta([\text{CLS}], q, d)$$

www.xueshuxiangzi.com

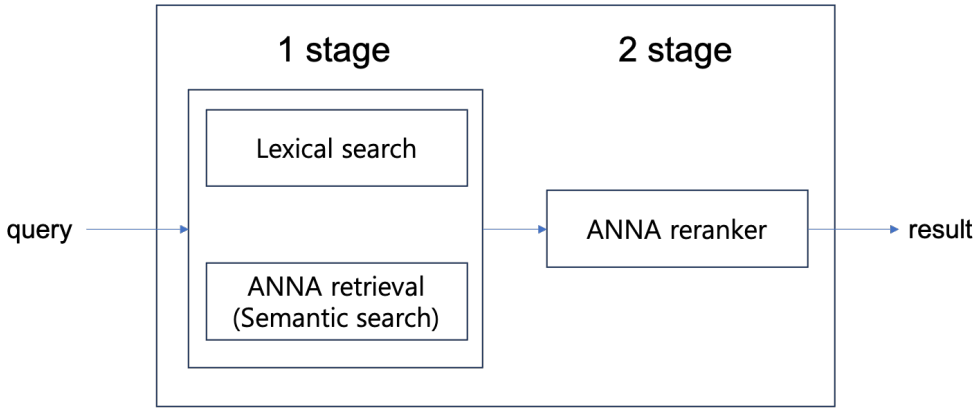

Figure 2: ANNA IR 流程的概述。该系统采用两阶段检索框架。在第一阶段，使用词汇搜索和语义搜索（ANNA 检索）来检索候选段落。在第二阶段，交叉编码器重新排序器通过对查询-文档对进行评分来优化结果以进行最终排名。

为了整合这些组件的各自排序——分别表示为 L （词汇）、S （语义）和 R （重排序器）——我们采用了互惠排名融合（RRF）算法 [29] 。文档 d 的综合得分计算如下：

$$\text{RRF}_{\text{score}_{\text{total}}} = \sum_{l \in L} \frac{1}{k + l(d)} + \sum_{s \in S} \frac{1}{k + s(d)} + \sum_{r \in R} \frac{1}{k + r(d)}$$

其中 k 是一个常数超参数，通常按照原始论文的建议设为 60。

在我们的实现中，我们采用标准的 InfoNCE 损失 [30] 作为对比训练目标。给定一个查询批次 q_i ，它们对应的正向段落 p_i ，以及对应每个查询的一组难例负样本 $\{n_{i,j}\}$ ，损失计算如下：

$$\mathcal{L} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(q_i, p_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(q_i, n_{i,j})/\tau) + \exp(\text{sim}(q_i, p_i)/\tau)}$$

其中 $\text{sim}(\cdot)$ 表示余弦相似度，τ 是温度参数。

通过使用来自强大教师模型的软目标，我们的嵌入模型能够内化教师输出中编码的细粒度语义线索，从而在多个任务中提高检索性能和鲁棒性。

传统的嵌入模型通常通过使用编码器或语言模型直接对输入文本进行编码来生成表示。虽然这种方法简单且计算效率高，但缺乏捕捉多样化用户意图和特定领域细微差别的灵活性，从而限制了泛化能力。相比之下，只有解码器的大型语言模型（LLMs），如 Mistral-7B，已展示出充分利用结构化输入中上下文信息的强大能力。特别是，只有解码器的 LLMs 具有强大的上下文学习（ICL）能力，最近的研究探索了有效利用只有解码器的 LLMs 的上下文学习能力的微调方法。

在我们的方法中，我们通过结合结构化的指令以及少量示例来实现上下文感知的嵌入生成。每个查询都通过任务特定的指令和一组示范对（查询-段落）进行扩充，旨在模拟目标任务的语义。输入序列以一个 [EOS] 标记结束，最终的嵌入从最后一个标记的表示中提取出来。该设计允许模型在不需要进行任何参数更新的情况下，隐式学习任务格式、意图和输出模式。

形式上，给定一个任务指令 t ，支持示例 $\{(q_i, p_i)\}_{i=1}^{k}$ ，以及一个目标查询 q ，输入格式如下：

$$\text{Examples}\{(q_i, p_i)\}_{i=1}^{k} + \text{Instruct: \{ task_definition \} + Query}$$

从 [EOS] 标记中提取的嵌入结果反映了查询的语义内容及其与任务规范的上下文关联性。

2.1 自适应基于边际的挖掘策略

受现有方法的启发，我们通过应用自适应基于边界的矿业策略来提高对比学习的鲁棒性，并减轻虚假负例的影响，该策略根据对应正例的相关性得分选择排名前 K 的负片段。该方法动态调整负例选择的阈值，确保语

www.xueshuxiangzi.com

Model	Retrieval	Reranking	Clustering	PairClassification	Classification	STS	Summarization	Mean (Task)
# of tasks	10	2	8	3	8	9	1	41
Seed1.5-Embedding	67.45	50.67	60.83	87.39	89.88	87.23	36.44	74.76
gemini-embedding-001	64.35	48.59	59.39	87.70	90.05	85.29	38.28	73.30
Linq-Embed-Mistral	60.14	49.44	54.07	88.44	83.00	84.69	37.26	69.80
jasper_en_vision_language_v1	56.05	50.00	60.52	88.14	90.27	84.37	37.19	71.41
SFR-Embedding-Mistral	59.33	50.15	54.93	88.59	80.47	84.77	36.32	69.31
NV-Embed-v2	62.84	49.61	47.66	88.69	87.19	83.82	35.21	69.81
text-embedding-005	58.77	48.84	51.91	87.62	86.03	85.18	35.05	69.60
text-embedding-004	59.06	48.48	51.52	87.65	86.03	84.84	36.12	69.53
gte-Qwen2-7B-instruct	58.09	50.47	58.97	85.90	88.52	82.69	35.74	70.72
e5-mistral-7b-instruct	57.62	49.78	51.44	88.42	79.85	84.32	36.57	67.97
LG-ANNA-Embedding(Ours)	66.18	49.13	59.25	88.67	89.97	86.69	38.93	74.12

Table 1: 在 MTEB English (v2) 基准测试中对排名前列的模型的性能比较。模型按博尔达排名进行排序。表格报告了跨越多个类别的 41 个任务的平均分数：检索（10 个任务）、重排序（2 个）、聚类（8 个）、对分类（3 个）、分类（8 个）、语义文本相似性（STS，9 个）和摘要（1 个）。我们的模型在大多数类别中表现出竞争力，尤其是在检索和摘要方面。

义相似但不相同的片段被排除在训练之外，这些片段通常被错误标记为负例。这样一来，它减少了噪声，并保留了对有效的嵌入学习至关重要的语义对比信号。

采用 Moreira 等人提出的策略 [31]，我们利用 ANNA IR 流水线作为教师检索模型，为每个查询识别高质量的困难负面段落。我们将负面段落的最大允许得分定义为相应正面得分的一个固定比例，应用 95 % 的边距。这可以表示为：

$$\text{max_negative_score_threshold} = \text{positive_score} \times \text{percentage_of_margin}$$

在训练过程中，得分低于此阈值的负采样将被排除。随后，从排名前 K 的负样本中随机选取一个子集，以提升训练的多样性并减轻对常见干扰项的过拟合。这些策略简单但有效，可以灵活地结合使用。它们共同确保训练聚焦于难以区分但信息丰富的负样本，从而提高检索准确性和训练稳定性。

我们利用 Mistral-7B 作为基础模型，并在精选的训练数据集混合上对其进行微调。继 E5-Mistral 和 NV-Embed-v2 等先前的工作之后，我们采用 Mistral-7B 作为我们嵌入框架的主干。

我们对 Mistral-7B 模型进行单个 epoch 的微调，使用对比损失（如 **??** 中描述）和 1e-4 的学习率，采用热身阶段然后线性衰减。为了实现高效的参数微调，我们采用低秩适配（LoRA）[32]，设置 LoRA 阶数为 64，LoRA alpha 为 32。在检索任务中使用批内负采样，而在其他任务类型中不应用。对于检索数据集的每个实例，包含 7 个难负样本。在检索任务中，通过从 ANNA IR 管道中提取相关性分数生成软标签，该管道被作为教师使用。

评估 MTEB(eng, v2) 基准测试是今年新引入的，作为一个仅限英语的评估套件，其设计目的是通过优化任务选择提高计算效率并减少任务间的相关性 [5]。与多语言变体不同，它通过排除常用于微调的任务来强调零样本评估。在这项工作中，我们针对 MTEB(eng, v2) 基准测试评估我们模型的性能。此外，采用 Borda 计数排名方法进行模型比较，遵循 MTEB(eng, v2) 基准测试的评估协议，因为它提供了对不同任务性能的稳健汇总 [33]。对于评估，会报告所有任务的平均得分、每类别的平均值和类别加权平均值。在 Borda 计数框架中，每个任务被视为一个投票者，为模型分配优先级排名。最终排名通过汇总这些投票来计算，在出现并列情况下，应用了锦标赛 Borda 计数变体以确保排名的稳定性。

2.2 MTEB 结果

表 1 展示了七个任务类别的平均性能——分类、聚类、对分类、重新排序、检索、语义文本相似度（STS）和摘要——将我们的模型与截至 2025 年 6 月 7 日在 MTEB 排行榜上报告的最先进方法进行比较。

与现有的在 MTEB English (v2) 基准测试中排名靠前的模型相比，我们提出的 LG-ANNA-Embedding 在多种任务类别中表现出一致的强大性能。值得注意的是，它在检索 (66.18; 第 2 名)、STS (86.69; 第 2 名)、对类分类 (88.67; 第 2 名) 和摘要 (38.93; 第 1 名) 中均取得了高分，在这些类别中名列前茅。特别是，该模型在所有被评估的系统中获得了最高的摘要分数，超过了像 gemini-embedding-001 (38.28) 和 Seed1.5-Embedding (36.44) 这样的更大商业模型。尽管 LG-ANNA-Embedding 报告的平均分 (74.12) 略低于 Seed1.5-Embedding (74.76)，但它在整体 Borda 排名中超越了 Seed1.5-Embedding，反映出在不同任务类型中更强及更一致的性能。这些结果表明我们的模型在泛化和任务特定性能之间取得了良好的平衡，特别是在信息检索和摘要方面表现出色。

www.xueshuxiangzi.com

在这份报告中，我们提出了 LG-ANNA-Embedding，这是一种用于生成高质量、通用化文本嵌入的统一的、基于指令的框架，利用的是仅解码器的大型语言模型。通过结合上下文学习、来自高性能检索-重排序管道的软监督，以及基于自适应边际的硬负样本挖掘，我们的方法有效利用了大型语言模型的内在泛化能力，而无需进行完整的模型微调或架构更改。

通过在 MTEB（英文 v2）基准上的广泛实验，我们的模型在包括检索、分类、聚类、语义相似性和摘要在内的多种任务中表现出强大而一致的性能。值得注意的是，它在 Borda 排名中名列前茅，显示出其在 IR 和非 IR 环境中的强大泛化能力，并且虽然相比更大的商业基准，模型体积较小，却在总结任务中排名第一。

这些发现强调了结构化提示和细粒度监督在可扩展嵌入生成中的价值。我们的方法突出了未来嵌入系统的一个有前途的方向，这些系统优先考虑效率、适应性和泛化，从而为搜索、推荐和语言理解等更广泛的实际应用做出贡献。

我们希望这项工作能够鼓励大家在实际服务中进一步探索统一嵌入模型的应用，为基于 LLM 的解决方案的可扩展和高效部署铺平道路。

References

[1] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024.

[2] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2023.

[3] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

[4] Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. Making text embedders few-shot learners. *arXiv preprint arXiv:2409.15700*, 2024.

[5] Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025.

[6] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human-generated machine reading comprehension dataset. 2016.

[7] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

[8] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[9] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[10] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*, 2019.

[11] Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, 2018.

[12] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942, 2018.

[13] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[14] Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. Natural language understanding with the quora question pairs dataset. *arXiv preprint arXiv:1907.01041*, 2019.

[15] James O'Neill, Polina Rozenshtein, Ryuichi Kiryo, Motoko Kubota, and Danushka Bollegala. I wish i would have loved this one, but i didn't–a multilingual dataset for counterfactual detection in product reviews. *arXiv preprint arXiv:2104.06893*, 2021.

www.xueshuxiangzi.com

[16] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172, 2013.

[17] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*, 2020.

[18] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3687–3697, 2018.

[19] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

[20] Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. *arXiv preprint arXiv:2008.09335*, 2020.

[21] cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. Jigsaw unintended bias in toxicity classification. `https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification`, 2019. Kaggle.

[22] Maggie, Phil Culliton, and Wei Chen. Tweet sentiment extraction. `https://kaggle.com/competitions/tweet-sentiment-extraction`, 2020. Kaggle.

[23] Ken Lang. Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995*, pages 331–339. Elsevier, 1995.

[24] Gregor Geigle, Nils Reimers, Andreas Rücklé, and Iryna Gurevych. Tweac: Transformer with extendable qa agent classifiers. *arXiv preprint*, abs/2104.07081, 2021.

[25] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*, 2020.

[26] Xueqing Liu, Chi Wang, Yue Leng, and ChengXiang Zhai. Linkso: a dataset for learning to retrieve similar question answer pairs on software development forums. In *Proceedings of the 4th ACM SIGSOFT International Workshop on NLP for Software Engineering*, pages 2–5, 2018.

[27] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity.* sem 2012: The first joint conference on lexical and computational semantics—. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), Montréal, QC, Canada*, pages 7–8, 2012.

[28] Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott A. Hale, David Jurgens, and Mattia Samory. SemEval-2022 task 8: Multilingual news article similarity. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States, July 2022. Association for Computational Linguistics.

[29] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759, 2009.

[30] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021.

[31] Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*, 2024.

[32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[33] Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stéphan Clémençon. What are the best systems? new perspectives on nlp benchmarking. *Advances in neural information processing systems*, 35:26915–26932, 2022.

A

A 贡献者 所有作者按姓氏字母顺序排列。

www.xueshuxiangzi.com

核心贡献者　崔周勇、金贤、张汉率、全昌旭

贡献者　Kyunghoon Bae, Hyewon Choi, Stanley Jungkyu Choi, Honglak Lee, Chulmin Yun

B

B　模型许可证

B.1

LG-ANNA-嵌入模型许可证：Apache 2.0 和 CC BY-NC 4.0

本许可协议（“协议”）由您（“被许可方”）与 LG 管理开发研究院有限公司（“许可方”）签订，管理对 LG-ANNA-Embedding 模型（“模型”）的使用。通过下载、安装、复制或使用该模型，您同意遵守并受本协议条款的约束。如果您不同意所有条款，则不得下载、安装、复制或使用该模型。此协议构成被许可方与许可方之间具有约束力的法律协议。

LG-ANNA-Embedding 模型的软件组件遵循 Apache License 2.0 许可协议发布，数据和模型权重遵循知识共享署名-非商业性使用 4.0 国际许可协议（CC BY-NC 4.0）发布。详细信息请参考完整的许可文本：

- `https://www.apache.org/licenses/LICENSE-2.0`
- `https://creativecommons.org/licenses/by-nc/4.0/`

www.xueshuxiangzi.com