利用多阶段对比学习实现通用文本嵌入

Zehan [Li1], **Xin** [Zhang1], **Yanzhao** [Zhang1], **Dingkun** [Long1], **Pengjun** [Xie1], **Meishan Zhang**
[1]阿里巴巴集团
{lizehan.lzh,linzhang.zx,zhangyanzhao.zyz,
dingkun.ldk,pengjun.xpj}@alibaba-inc.com

摘要

我们介绍的 GTE 是一种通过多阶段对比学习训练的通用文本嵌入模型。为了与最近将各种 NLP 任务统一到一个单一 Mat 中的进展保持一致，我们通过对来自多个来源的数据集的双向混合使用对比学习来训练统一的文本嵌入模型。通过在无监督预训练和有监督微调阶段大幅增加训练数据的数量，我们实现了比现有嵌入模型更高的性能。值得注意的是，即使只有相对适中的 1.1 亿个参数，GTE_{base} 的性能也超过了 OpenAI 提供的黑盒嵌入 API，甚至在海量文本嵌入基准测试中超过了 10 倍大的文本嵌入模型。此外，在不对每种编程语言进行额外微调的情况下，通过将代码视为文本，我们的模型超越了以前类似规模的最佳代码检索器。总之，我们的模型通过有效利用多阶段对比学习取得了令人印象深刻的结果，提供了一个强大而高效的文本嵌入模型，可广泛应用于各种 NLP 和代码相关任务。[1]

1 导言

文本嵌入已成为文本分类、文本检索、问题解答和对话系统等许多自然语言处理任务中不可或缺的组成部分（Karpukhin 等人，2020；Humeau 等人，2020；Choi 等人，2021；Izacard 等人，2022a；Long 等人，2022a；Rajapakse，2023）。这些嵌入模型使用低维向量表示文本，并通过向量运算捕捉文本的相似性。最近出现的大型语言模型（LLMs）（Radford 等人，2018 年；Touvron 等人，2023 年；OpenAI，2023 年）在检索领域引起了相当大的兴趣。

[1]GTE 模型可在 https://huggingface.co/thenlper/gte-large 上公开获取。

图 1：用于训练文本嵌入模型的多阶段对比学习管道示意图。

基于文本嵌入模式的增强系统整合了 LLM 的推理和理解能力（Izacard 等人，2022b；Ram 等人，2023；Shi 等人，2023）。因此，工业界和学术界都越来越关注一般文本的再现。

由于自然语言的格式、领域和下游应用多种多样，长期以来，人们一直在寻求开发一种统一的模型，以适应多种下游任务。预训练语言模型的出现进一步为训练这种通用模型提供了可能。然而，在文本表征研究领域，以往的文本表征模型主要集中在特定任务上，针对单一任务的训练策略或模型可能无法在其他情况下发挥最佳性能。例如，在对称文本对上训练的文本表征模型 SimCSE（Gao 等人，2021 年）在文本检索任务中表现出局限性。同样，某些专为密集检索任务设计的文本再现模型在句子文本相似性任务中也没有表现出强劲的性能。最近，研究重点转向开发更全面的文本再现模型，通过无监督对比预训练，利用大量未标记的网络数据，再加上特定任务的数据、提示或结构，在微调过程中缓解任务冲突（Ni 等人，2022a,b；Neelakantan 等人，2022）；

Wang 等人，2022b；Su 等人，2023）。此外，大量文本嵌入基准（MTEB）（Muennighoff 等人，2023 年）等基准的引入，为评估文本表示模型的通用性奠定了坚实的基础。然而，当前研究的一个重大局限是依赖内部数据进行预训练，这在利用预训练模型权重或 API 方面造成了瓶颈。此外，在实施过程中，为每项任务量身定制提示需要额外的人力（Su 等人，2023 年）。

如图 1 所示，本研究提出了一种在开源数据上仅使用对比学习构建通用文本嵌入（GTE）模型的直接方法。具体来说，我们首先收集了一个大规模数据集，其中包括从各种数据源中提取的无监督文本对，用于对比预训练。令人惊讶的是，我们在该数据集上预训练的模型表现出了可再标记的性能，在零镜头文本再三值任务中超过了 BM25 和 E5 模型（Wang 等人，2022b），并在 MTEB 基准中超过了许多监督模型。为了进一步提高所学文本表征的质量，我们从多个来源获取带有人类标签的高质量文本对，进行对比微调。经过监督微调后，我们基于 110M BERT（Devlin 等人，2019 年）的模型已经超越了 OpenAI 当前的商业嵌入 API，并在 MTEB 基准中名列前茅。此外，由于我们的模型也使用代码数据进行训练，因此我们在包含六种编程语言的 CodeSearchNet 基准上对其代码搜索能力进行了评估。值得注意的是，即使不对每个子集进行特定语言的微调，我们的模型也明显优于针对每种编程语言进行过微调、规模类似的最先进代码检索器。

在本文的其余部分，我们将详细介绍采用的数据源和训练配置。随后，我们介绍了在广受认可的文本阅读基准上的评估结果，并将这些结果与之前针对每项任务进行了专门优化的最先进基准的性能进行了比较。由于我们的模型采用了更多样化的混合训练数据集，因此始终保持了卓越的性能，至少与大型模型的性能相当。我们希望我们的模型能够

作为研究界调查文本和代码嵌入的可靠基准。

2 相关工作

文本嵌入是不同长度文本的低维向量表示，在许多自然语言处理（NLP）任务中都至关重要。与 TF-IDF 等高维稀疏表示法相比，密集文本内嵌能够解决词法不匹配问题，提高文本检索和匹配的效率。

以 BERT（Devlin 等人，2019 年）和 GPT（Radford 等人，2018 年）为代表的预训练语言模型在各种 NLP 任务中都取得了显著的成功。然而，由于遮蔽语言建模目标导致的各向异性嵌入空间的存在，从预训练的语言模型中提取高质量句子嵌入是一项重大挑战。为解决这一问题，后续研究提出了不同的方法，包括监督微调（Reimers 和 Gurevych，2019 年）、归一化流（Li 等人，2020 年）、归一化流（Li 等人，2020 年）、白化（Su 等人，2021 年）或无监督对比学习（Gao 等人，2021 年）。这些研究主要集中在提高语义文本相似性任务的性能上，在这些任务中，两个文本表现出相似的格式。

另一个研究方向是文本再三值问题，在这种情况下，查询和文档通常表现出不对称的关系。在这种情况下，双编码器架构需要同时使用正对和负对进行训练。Lee 等人（2019）提出了 "反向关闭任务"（ICT）作为生成密集检索器的自监督预训练方法。ICT 方法是从段落中随机裁剪一个句子来构建伪查询-文档对。传统上，Chang 等人（2020 年）利用维基百科中的链接结构，在预训练数据中引入进一步的监督信号。与此类似，REALM（Guu 等人，2020 年）提出了一种联合训练方法，即同时训练密集检索器和语言模型。语言模型的学习信号来自遮蔽语言建模，并在检索步骤中加入反向推导。目前，Contriever（Izacard 等人，2022a）和 coCondenser（Gao 和 Callan、

2022）的研究表明，与信息和通信技术任务相比，通过随机段落裁剪来构建正面配对会产生更好的结果。在（Chang 等人，2020 年）提出的观点基础上，一些研究人员还提出了利用网络链接拓扑构建更高质量正对的方法，用于检索器预训练（Zhou 等人，2022 年），这种技术在零镜头场景下证明是有效的。此外，在密集检索领域，大量研究致力于通过设计辅助预训练任务来增强预训练语言模型的文本表示能力（Gao 和 Callan，2021；Xiao 等人，2022；Gao 和 Callan，2022；Wang 等人，2022a；Long 等人，2022b；Li 等人，2023）。

前两项研究可归纳为学习一段文本的向量表征，并以下流任务的类型加以区分。最近，一些研究通过大规模对比学习和基于提示的学习来构建统一的文本表示模型（Neelakantan 等人，2022；Wang 等人，2022b；Su 等人，2023）。传统上，一些研究工作主要集中在构建评估数据集，以更好地评估文本表示模型在不同任务和领域中的稳定性。BEIR（Benchmark-ing IR）（Thakur 等人，2021 年）收集了大量来自不同领域的检索任务，以评估密集检索模型在零点场景下的鲁棒性。同时，MTEB（Massive Text Embedding Benchmark）（Muennighoff 等人，2023 年）对跨越七个类别的 56 个数据集进行了基准测试，对文本嵌入模型进行了全面评估。

本研究旨在通过多阶段训练，建立一个通用的文本电子书模型。

方法。在无监督强制学习的初始阶段，我们利用各种来源的公开数据生成弱监督相关文本对。与之前的研究（Wang 等人，2022b）不同，更多样化，以进一步提高模型的可验证性。此外，我们的模型不包含特定任务的提示，从而提高了可重复性和易用性。

3 方法

我们模型的训练过程包括两个阶段：无监督预训练和有监督微调。这两个阶段都采用了对比学习的学习目标。首先，我们将介绍模型的基本框架。随后，我们将讨论两个阶段中训练数据的来源和构建方法。最后，我们将介绍一些特殊的优化策略，以提高模型在训练过程中的性能。

3.1 模型架构

我们的嵌入模型的骨干是一个深度 Transformer 编码器（Vaswani 等人，2017 年），它可以通过 BERT（Devlin 等人，2019 年）等预训练语言模型进行初始化。我们的模型采用 vanilla 双编码器架构，在语言模型生成的上下文化标记表示之上进行均值池化。

形式上，给定一段由 n 个标记组成的文本 $x = (x_1, \ldots, x_n)$，一个嵌入模型 E 会将文本垂直嵌入到一个低维密集向量 $\mathbf{x} = E(x) \in R^d$ 中。为了实现 E，我们首先使用一个语言模型来获取深度上下文标记表征

$$\mathbf{h} = \mathrm{LM}(x) \in R^{n \times d} \quad (1)$$

然后，我们对第一个维度进行轻量级均值池化处理，以获得文本表征。

(一)..... $\mathbf{x} = \frac{1}{n}\sum \mathbf{h} \in R^d$ (2)

、

我们只使用开源数据，没有使用任何过滤或清洗方法。在大规模文本对上进行预训练可有效提高领域泛化能力

$$n_{i=1}^{i}$$

文本表征是通过对比目标学习的，将语义相关的文本对与不相关的文本对区分开来。的文本表示模型，并缩小差距

MLM 训练目标和表征模型的强制学习目标之间的关系，使语言模型更适合于文本

表征任务。在有监督的微调阶段，我们方法中的混合训练数据

这种训练过程需要正负文本对，格式为（q, d^+, d^-）。对于查询 q 、一个相关文档 d^+ 、一组无关文档 $D_- = \{d^-, \ldots \ldots, d^-\}$ ，有一种流行的对比性 ob-

这就是 InfoNCE 的损失（van den Oord 等人，2018 年）、

$$L_{\text{cl}} = -\log \frac{es(q,d^+)/\tau}{es(q,d^+)/\tau + \sum_{i=1}^{n} es(q,d_i^-)/\tau}, \qquad (3)$$

其中，$s(q, d)$ 通过 $\mathbf{q} = E(q)$和 $d = E(d)$之间的向量距离来估计两个文本片段 q 和 d 之间的相似性。

为了获得可应用于各种场景的高质量文本嵌入，我们从多种格式和领域中收集了大量文本对数据集。然后，使用改进的对比损失法对该数据集进行多阶段训练。

3.2 无监督预训练数据

弱监督文本相关性数据可随时从公开的网络资源中获取，例如质量保证论坛上的问答之间的内在联系。这些数据可以被广泛收集，无需人工标注，从而有效地帮助训练文本相关性模型。受之前工作（Ni 等人，2022a,b；Neelakantan 等人，2022；Wang 等人，2022b）的启发，我们的模型最初是在从不同来源提取的自然出现的文本对上进行预训练的。为确保嵌入模型的通用性，我们探索了一系列文本对提取资源，包括网页（*如* CommonCrawl、ClueWeb）、科学论文（*如* *arXiv*、SemanticScholarship）和其他资源、arXiv、SemanticScholar）、社区质量保证论坛（*如* StackExchange）、社交媒体（*如* Reddit）、知识库（*如* Wikipedia、DBPedia）和代码库（*如* StackOverflow、GitHub）。此外，我们还利用某些数据集中存在的超链接来促进文本对提取。表 2 举例说明了不同来源的文本对格式。有关数据收集过程的更多详情，请参阅附录 A。在无监督预训练阶段，我们总共使用了 8 亿个文本对。简单的统计数据和数据分布如表 1 所示。

3.3 监督微调数据

在有监督的微调阶段，我们使用规模相对较小的数据集，由人工标注两段文本之间的相关性，并由额外的检索器挖掘出可选的硬否定，从而形成文本三元组。为了处理对称任务（*如*语义文本相似性）和非对称任务（*如*段落检索），我们收集了来自各种任务和领域的数据，包括网络搜索（*如* MS MARCO）、开放域 QA（*如* NQ）、NLI（如 SNLI）、事实验证（如 FEVER）、转述（*如* Quora）。我们总共使用了 ~3M 对数据进行微调，这是一个很好的例子。

资料来源	数据集	道具	尺寸
网页	3	18.7%	147M
学术论文	5	5.7%	45M
超链接	4	13.4%	106M
社交媒体	2	41.5%	327M
知识库	2	4.8%	38M
社区质量保证	7	1.5%	12M
新闻	5	0.4%	3M
代码	2	2.5%	20M
其他	3	11.6%	91M
总计	33	100%	788M

表 1：训练前数据统计。

这些数据结合了之前重新搜索所使用的训练数据（Gao 等人，2021 年；Gao 和 Callan，2022 年；Asai 等人，2023 年；Su 等人，2023 年；Li 等人，2023 年）。更多详情见附录 A。

3.4 培训详情

数据采样 在无超级视图预训练的初始阶段，数据源在训练实例数量上往往存在显著差异。为了解决这种不平衡问题，我们采用多项式分布来对不同数据源的数据批次进行采样，同时考虑到它们各自的大小。假设整个预训练数据集 D 由 m 个不同的子集 $\{D_1, \ldots .D_m\}$，并将每个子集的大小记为 $n_i = |D_i|$，则在每次训练迭代时，从第 i 个子集 D_i 中抽取数据的概率可表示为：

$$p_i = \frac{n^{\alpha}}{\sum_{j=1}^{m} n_j^{\alpha}}, \quad (4)$$

其中，我们将 α 设为 0.5。此外，为了防止模型只学习特定任务的识别捷径，我们确保批次中的所有训练实例都来自同一任务。

改进的对比损失 在使用对比目标时，人们通常会重复使用批内文档作为负候选样本，以提高训练效率（Karpukhin 等人，2020）。本文使用了一种改进的对比学习目标，该目标是双向的，可以通过批内查询和文档来扩大负样本。这可以看作是 Radford 等人（2021 年）、Ren 等人（2021 年）和 Moiseev 等人（2023 年）提出的损失变体的组合。

$$\sum \frac{i}{n}$$

任务类型	文本对格式	查询	文档
网页	(标题、正文)	普罗维登斯房地产 | 普罗维登斯待售住宅	普罗维登斯由罗杰-威廉姆斯（Roger Williams）于 1636 年创建，是美国最古老的城市之一。
学术论文	(标题、摘要)	聚合物量子力学及其连续极限	被公认为该国最古老的城市之一。. . 的一种相当非标准的量子表示。
超级链接	(引用、参考)	1996 年冠军赛之后，美国职业高尔夫球协会	量子力学的典型换向关系 .. . 圆石滩高尔夫林克斯大满贯赛有史以来最大的胜场
社交媒体	(帖子, 评论)	将所持股份增至 50%，并宣布 . 可以肯定的是，任何一支拥有勒布朗-詹姆斯的球队	差，超过了13杆的纪录.我是在讽刺和取笑东方，但
都会成为季后赛球队		竞争者。考虑到 UNC 将在东部.. .	老实说，我真的对此深思熟虑 .
知识库（实体、描述）	动画		动画是通过快速显示 .
社区 QA（ 问题、答案）		人类是如何进化的?	这是一个棘手的问题，因为它涉及科学和神学。因为 你问 "人类是如何进化的？"我假设 . . .
新闻	(尼泊尔反对派欢迎议会回归		尼泊尔反对派联盟正式取消为期数周的议会选举。 捷南德拉国王复职后的民主抗议 .. .
代码	(文本, 代码)	SetMaxRecords 设置 MaxRecords 字段的值。	func (s *DescribeSnapshotCopyGrantsInput) SetMaxRecords (v int64) *DescribeSnapshotCopyGrantsInput { s.MaxRecords

表 2：在预训练数据中挖掘出的（查询、文档）配对示例。

考虑一批正面文本对样本

$B = \{(q_1, d_1), (q_2, d_2), ..., (q_n, d_n)\}$、

我们使用改进的对比损失，其形式为

$$L_{\text{icl}} = -\frac{1}{n}\sum_{i=1}^{n} \text{原木} \frac{es(q_i, d_i)/\tau}{Z} \quad (5)$$

分区函数为

$$Z = \sum_{j} es(q_i, d_j)/\tau + \sum_{j \neq i} es(q_i, q_j) \quad (6)$$

$$+ \sum_{j} es(q_j, d_i)/\tau + \sum_{j \neq i} es(d_j, d_i)/\tau$$

其中，前两个词用于查询与文档的对比，后两个词用于反向对比。在这项工作中，我们使用余弦相似度作为距离度量

$$s(q, d) = \frac{\mathbf{q} \cdot \mathbf{d}}{\|q\|_2 \cdot \|d\|_2}. \quad (7)$$

在本文中，温度 τ 被固定为 0.01。

训练和评估 我们的嵌入模型的训练包括两个阶段。在第一阶段的对比预训练中，只使用批这样可以降低内存成本，并将批量规模扩大到数万以上。我们运行了 50,000 步预训练，这大致相当于在整个预训练数据上运行一个历元。我们只对学习率进行了调整，以确保较大的学习率也能收敛。

我们采用的 AdamW 优化器具有线性学习率衰减和初始 5% 训练步骤的热身期。我们将

在三种不同的模型尺度（小型、基本型和大型）上进行了实验。这些模型是使用小型 MiniLM（Wang et al、2020 年）模型和基本模型和大型模型的 BERT（Devlin 等人，2019 年）模型。更多详情见表 3。

在使用有监督数据和硬阴性数据进行对比微调的第二阶段，由于硬阴性数据已经可以对学习目标进行可靠的梯度估计，因此不需要很大的批次规模（Xiong 等人，2021 年；Li 等人，2021 年）、2023).因此，全局批量规模为 128，列车组规模为 16，其中一个正组为："....."。

内负片，使用大的批量对于提高模型性能至关重要，因为这样可以减少训练和推理之间的差距，同时包含更多的负片，并提供对底层学习

目标的更好近似。为此，我们在预训练期间将最大序列长度限制为 128，并在所有 GPU 上分配使用底片。自动混合精度训练（Micikevicius 等人，2018 年）与 fp16、deepspeed ZeRO（Rajb- handari 等人，2020 年）第 1 阶段和梯度检查指导（Chen 等人，2016 年）等流行技术也被联合用于

而剩下的要么是硬否定，要么是随机否定。相反，我们将最大序列长度增加到 512，以便更好地处理长度更长的文本。在微调过程中，学习率降低了 10 倍。模型在收集的数据集上进行单次微调。批次内文本也会作为否定候选文本，使用等式 5 中描述的增强对比度损失。

训练结束后，我们直接取最后一个检查点进行评估。我们最多在 8 个配备 80GB 内存的英伟达 A100 GPU 上运行模型训练，最多在 8 个配备 32GB 内存的英伟达 Tesla V100 GPU 上运行模型评估。模型使用 fp16 混合精度进行训练，并使用半精度 fp16 进行评估。

模型	参数	LR	图形处理器	BS	基地 LM
GTEsmall	30M	3×10^{-4}	2	16384	microsoft/MiniLM-L12-H384-uncased
GTEbase	110M	2×10^{-4}	4	16384	伯特基不分区
GTElarge	330M	5×10^{-5}	8	16384	bert-large-uncased

表 3：不同大小模型的预训练配置。

4 实验

在本节中，我们对我们的嵌入模型进行了广泛的评估，并就每项任务与最先进的模型进行了比较。请注意，由于不同模型使用不同的内部数据进行预训练，而且基础语言模型差异很大，因此很难进行苹果与苹果之间的比较。我们主要使用模型参数的数量作为性能比较的标准，因为它与推理速度密切相关。

4.1 零镜头文本分类

模型	参数	提示	准确性
E5base	110M	✓	81.3
E5 大	330M	✓	85.3
cpt-text	6B		88.1
cpt-text	6B	✓	89.1
GTEbase	110M		85.1
GTEbase	110M	✓	87.2

表 4：SST-2 上的零镜头文本分类性能。所有比较模型均为微调模型。

评估学习到的表征质量的一种方法是进行零点分类（zero-shot classifica-tion）。(Radford 等人，2021；Neelakantan 等人，2022；Wang 等人，2022b）。我们将文本分类转换为基于嵌入的相似性匹配问题。在这种情况下，输入文本被直接转换为嵌入，标签被口头化为相应的文本，从而得到标签嵌入。输入嵌入和标签嵌入之间的距离用它们的内积来衡量，与输入文本嵌入距离最接近的标签被视为分类结果。SST-2 二元情感分类任务就是一个例子。我们考虑使用两种标签动词化器进行评估。普通版本使用感性词 "积极 "或 "消极 "来表示相应的标签。提示版本使用模糊提示模板，例如 "这是一个正面/负面电影评论的例子"。

表 4 显示了 SST- 2 的零镜头文本分类准确率。在 vanilla 设置中，我们的 110M 模型的性能已经可以与带有 330M 参数的提示 $E5_{large}$ 相媲美。使用提示策略可进一步显著提高成绩，缩小与大型模型的差距。即使在训练过程中没有明确的提示或指导，我们的模型在某种程度上也能更好地理解自然语言文本格式的标签上下文。

4.2 无监督文本检索

文本检索需要从大规模候选集中检索出最相关的文档。我们使用 BEIR（Thakur 等人，2021 年）作为零点无监督文本检索的评估基准。BEIR 是一个异构信息再评估基准，包含不同格式和不同领域的检索任务。我们使用公开的 15 个数据集进行评估。

我们将我们的无监督预训练检查点与 Contriever（Izacard 等人，2022a）和 E5（Wang 等人，2022b）等最新的无监督密集检索器进行了比较。根据表 5，我们发现我们的基本模型明显优于规模相当的模型，如 SimCSE、Con- triever 和 E5。在不使用人工监督的情况下，我们的基本模型可与 $E5_{large}$ 相媲美。

4.3 海量文本嵌入基准

大规模文本嵌入基准（MTEB）是一个全面的半监督基准，其中包含有限的监督数据用于评估。在本文中，我们对包含 56 个英文数据集的英文子集进行了评估，其中包括文本分类 (Class.)、文本聚类 (Clust.)、成对分类 (Pair.)、文本重排 (Rerank.)、文本还原 (Retr.)、语义文本相似性 (STS) 和摘要 (Summ.)等七项不同的任务。MTEB 采用的评价指标分别是准确率、v-measure、平均精度、MAP、nDCG@10 和 Spear- man 系数。更多详情

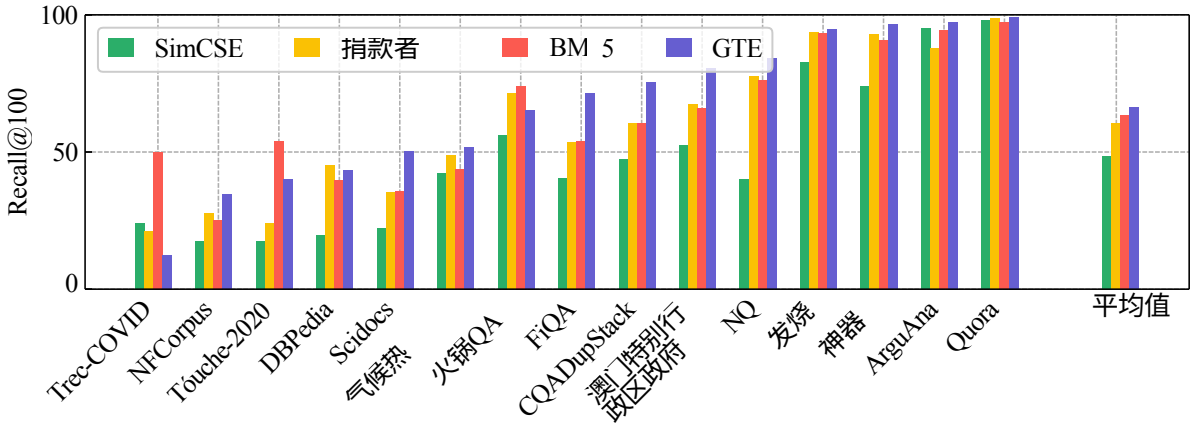


图 2：无监督文本检索方法在 BEIR 基准（Thakur 等人，2021 年）上的 Recall@100。我们将不使用任何注释数据的 GTE_{base} 模型（基于 $BERT_{base}$ ）与 SimCSE（Gao 等人，2021 年）（基于 $RoBERTa_{large}$ ）、Contriever（Izacard 等人，2022a）（基于 $BERT_{base}$ ）和 BM25 进行了比较。基线结果借自 Contriever 论文（Izacard 等人，2022a），点积是相似性函数。

数据集	BM25	SimCSE	Contriever	CPT-S	$E5_{small}$	$E5_{base}$	$E5_{large}$	GTE_{small}	GTE_{base}	GTE_{large}
马可号	22.8	9.4	20.6	19.9	25.4	26.0	26.2	31.3	31.8	31.7
Trec-Covid	65.6	26.2	27.4	52.9	52.0	61.0	61.8	61.8	64.0	64.8
NFCorpus	32.5	9.9	31.7	32.0	29.3	35.8	33.7	34.9	36.2	38.1
NQ	32.9	11.7	25.4	-	37.3	39.0	41.7	32.0	35.3	34.5
火锅QA	60.3	19.8	48.1	51.5	46.0	52.4	52.2	49.3	50.8	49.2
FiQA	23.6	9.8	24.5	34.1	38.3	40.0	43.2	37.0	36.9	40.6
ArguAna	31.5	38.3	37.9	38.7	42.5	42.2	44.4	41.6	41.0	41.3
Touche-2020	36.7	8.9	19.3	21.0	19.9	16.9	19.8	17.7	18.2	18.5
CQADupStack	29.9	13.2	28.4	-	35.0	35.4	38.9	38.1	39.9	39.8
Quora	78.9	78.0	83.5	68.1	85.8	85.7	86.1	86.1	85.0	84.8
DBPedia	31.3	15.0	29.2	27.2	34.5	35.4	37.1	33.5	33.2	33.6
Scidocs	15.8	5.5	14.9	-	19.9	21.1	21.8	21.5	22.5	22.7
发烧	75.3	21.1	68.2	57.1	62.5	63.4	68.6	71.3	72.7	70.5
气候热	21.3	11.8	15.5	15.8	14.5	15.4	15.7	21.4	21.0	25.4
神器	66.5	25.7	64.9	65.4	68.5	73.7	72.3	72.7	74.1	74.1
平均	41.7	20.3	36.0	-	40.8	42.9	44.2	43.4	44.2	44.6

表 5：不同无监督方法在 BEIR 基准（Thakur 等人，2021 年）上的 nDCG@10。SimCSE 基于 $BERT_{base}$ 骨干网。CPT-S（Neelakantan 等人，2022 年）的规模与 $BERT_{large}$ 相似。基线结果借自 E5 论文（Wang 等人，2022b）。请注意，Contriever 使用点积作为相似性度量，而其他模型使用余弦相似性。

关于 MTEB 基准所涵盖的任务，请参阅附录 B。

比较考虑了两种情况：无监督情况和有监督情况。在无监督环境下，模型使用无标签数据进行训练，而有监督环境下的模型则使用带有人类标签的高质量数据集进行微调。表 6 列出了强基准模型的结果。

在无监督的情况下，我们的模型比

在所有考虑的任务中，在不使用特定任务提示的情况下，它以明显的优势超过了之前的最佳模型 E5。这一进步可归功于加入了更多的训练数据格式和各种自我监督信号源。此外，值得注意的是，我们的无监督预训练模型与 GTR 和 Sentence-T5 等较大的监督基线相比，差距进一步缩小。在超级监督环境下，我们的模型超越了 OpenAI 的结果

	参数	班级	集群。	一对	重新排名	Retr.	STS	总结	平均值
# 数据集数量 →		12	11	3	4	15	10	1	56
无监督模型									
手套	120M	57.3	27.7	70.9	43.3	21.6	61.9	28.9	42.0
伯特	110M	61.7	30.1	56.3	43.4	10.6	54.4	29.8	38.3
SimCSE	110M	62.5	29.0	70.3	46.5	20.3	74.3	31.2	45.5
$E5_{small}$	30M	67.0	41.7	78.2	53.1	40.8	68.8	25.2	54.2
$E5_{base}$	110M	67.9	43.4	79.2	53.5	42.9	69.5	24.3	55.5
E5 $330M_{large}$		69.0	44.3	80.3	54.4	44.2	69.9	24.8	56.4
GTE_{small}	30M	71.0	44.9	82.4	57.5	43.4	77.2	30.4	58.5
GTE_{base}	110M	71.5	46.0	83.3	58.4	44.2	76.5	29.5	59.0
GTE $330M_{large}$		71.8	46.4	83.3	58.8	44.6	76.3	30.1	59.3
监督模型									
SimCSE	110M	67.3	33.4	73.7	47.5	21.8	79.1	23.3	48.7
捐款者	110M	66.7	41.1	82.5	53.1	41.9	76.5	30.4	56.0
GTR $330M_{large}$		67.1	41.6	85.3	55.4	47.4	78.2	29.5	58.3
Sentence-T5 $330M_{large}$		72.3	41.7	85.0	54.0	36.7	81.8	29.6	57.1
$E5_{small}$	30M	71.7	39.5	85.1	54.5	46.0	80.9	31.4	58.9
$E5_{base}$	110M	72.6	42.1	85.1	55.7	48.7	81.0	31.0	60.4
E5 $330M_{large}$		73.1	43.3	85.9	56.5	50.0	82.1	31.0	61.4
指导员$_{base}$	110M	72.6	42.1	85.1	55.7	48.8	81.0	31.0	60.4
指导员 $330M_{large}$		73.9	45.3	85.9	57.5	47.6	83.2	31.8	61.6
$OpenAI_{ada-001}$	n.a.	70.4	37.5	76.9	49.0	18.4	78.6	26.9	49.5
$OpenAI_{ada-002}$	n.a.	70.9	45.9	84.9	56.3	49.3	81.0	30.8	61.0
GTE_{small}	30M	72.3	44.9	83.5	57.7	49.5	82.1	30.4	61.4
GTE_{base}	110M	73.0	46.1	84.3	58.6	51.2	82.3	30.7	62.4
GTE $330M_{large}$		73.3	46.8	85.0	59.1	52.2	83.4	31.7	63.1
较大型号									
指导员$_{xl}$	1.5B	73.1	44.7	86.6	57.3	49.3	83.1	32.3	61.8
GTR_{xxl}	4.5B	67.4	42.4	86.1	56.7	48.5	78.4	30.6	59.0
句子-$T5_{xxl}$	4.5B	73.4	43.7	85.1	56.4	42.2	82.6	30.1	59.5

表 6：MTEB 的结果（Muennighoff 等人，2023 年）（英语子集中的 56 个数据集）。比较的模型包括 SimCSE（Gao 等人，2021 年）、Sentence-T5（Ni 等人，2022a 年）、GTR（Ni 等人，2022b 年）、Contriever（Izacard 等人，2022a 年）、OpenAI 文本嵌入 API（Neelakantan 等人，2022 年）、E5（Wang 等人，2022b 年）和 InstructOR（Su 等人，2023 年）。OpenAI ada 模型的确切参数量尚不详，但预计在 3 亿以上，与 BERT 大尺寸模型相当。

尽管使用的模型大小适中，但仍有很大差距。GTE_{small} 与 $E5_{large}$ 相当，但体积小 10 倍。GTE_{large} 在 MTEB 基准测试中取得了新的一流性能，平均比多任务指令调整嵌入模型 $InstructOR_{large}$ 高出 1.5 个百分点。

4.4 代码搜索

编程语言可被视为一种不同形式的文本。为了评估我们的方法在代码搜索中的有效性，我们与其他基于代码的语言模型进行了比较分析、

例如 CodeBERT（Guo 等人，2021 年）和 Graph- CodeBERT（Guo 等人，2021 年）。我们还将我们的方法与最新的代码语言模型 UniXcoder（郭等人，2022 年）进行了比较，后者旨在将各种预训练任务整合到一个统一的模型中。CodeRetriever（Li等人，2022年）是从GraphCodeBERT初始化而来，在启发式挖掘和清理的大规模多模态代码文本对上进行预训练。值得注意的是，基准模型是针对每种编程语言进行单独训练和评估的，而我们的模型则是直接针对所有语言进行评估的。

与最近的研究结果一致（Guo 等人，2021 年，2022 年；

模型	参数	红宝石	联署材料	转到	Python	Java	PHP	平均值
代码ERT	110M×6	67.9	62.0	88.2	67.2	67.6	62.8	69.3
GraphCodeBERT	110M×6	70.3	64.4	89.7	69.2	69.1	64.9	71.3
UniXcoder	110M×6	74.0	68.4	91.5	72.0	72.6	67.6	74.4
代码重设器	110M×6	77.1	71.9	92.4	75.8	76.5	70.8	77.4
GTEbase	110M	76.1	73.6	88.1	95.9	80.1	85.3	83.2

表 7：CodeSearchNet 的结果。6 种编程语言的代码搜索（Husain 等人，2019 年）与 CodeBERT（Feng 等人，2020 年）、GraphCodeBERT（Guo 等人，2021 年）、UniXcoder（Guo 等人，2022 年）和 CodeRetriever（Li 等人，2022 年）的比较。这种设置要求从开发集和测试集中的所有候选代码中找到相应的候选代码。

我们主要在代码语料包括开发集和测试集中的所有代码，而不是随机抽样的 1k 个代码的挑战性设置上进行评估。[2]结果见表 7。出乎意料的是，我们的模型超过了先对代码进行预训练，然后再针对每种编程语言分别进行微调的模型。这一发现表明，通过调整数据量和计算资源，语言模型可以直接从代码标记序列中获取高质量的代码表示，而无需结合人类关于代码结构信息的知识（Guo 等，2021 年）。我们在 Python 中观察到了明显的改进，这可能是由于 Python 与自然语言的相似性。我们的模型在跨越不同领域的大量文本对上进行了预训练，证明了从文本检索到代码检索的有效跨任务知识转移。

5 分析

在本节中，我们将分析影响模型性能的关键因素，并介绍一系列消融实验。除非另有说明，实验均使用具有 1.1 亿个参数的 BERT 基准模型。所有消融实验的训练步骤和历时保持一致。

5.1 扩大规模的影响

我们研究了调整数据源数量、批量大小和模型参数对所学文本嵌入质量的影响。评估是在 MTEB 基准上进行的。

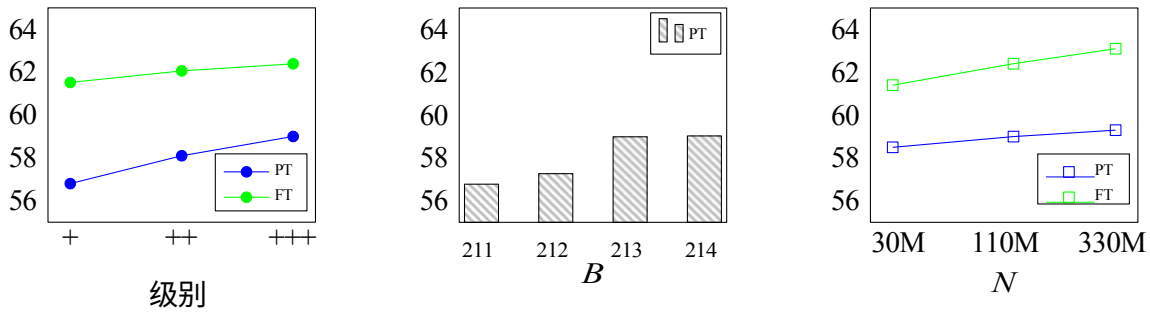
训练数据集的数量 首先，我们对训练数据集的数量进行了消融研究。

[2]对原始设置的评估见附录 C。

用于预训练的数据集。模型训练是从所有可用数据集中随机抽取一个子集进行的。在预训练阶段，第一组只包括按大小排序的五个最大的数据集。第二组包括另外 10 个随机抽样的数据集，即 15 个混合数据集。第三组在预训练过程中使用了所有 33 个数据集。为了进行微调，我们首先使用了 E5（Wang 等人，2022b）微调中使用的三个数据集，然后逐渐加入了 MEDI（Su 等人，2023）和 BERRI（Asai 等人，2023）的数据集，以研究其潜在优势。图 3a 中显示的结果表明，在预训练和微调阶段，纳入更多样化的数据源始终能提高模型性能。

预训练批次大小 我们在保持训练步骤固定的情况下，将批次大小逐渐增加 2 倍，以研究批次大小对嵌入模型预训练的影响。根据图 3b，模型性能在批量规模达到一万时达到饱和。当进一步扩大批量大小时，没有观察到任何性能提升。

模型参数数 我们通过训练不同大小的语言模型（包括 30M、110M 和 330M，分别对应 BERT 模型的小尺度、基本尺度和大尺度）来研究扩展行为。图 3c 展示了预训练模型和微调模型的性能。可以看出，随着模型大小呈指数增长，模型性能也呈线性提高。



(a) 训练数据集的数量。 (b) 批量大小 。 (c) 模型参数数量。

图 3：对比预训练和微调期间不同因素的比例分析。模型性能以 MTEB 的平均性能来衡量。

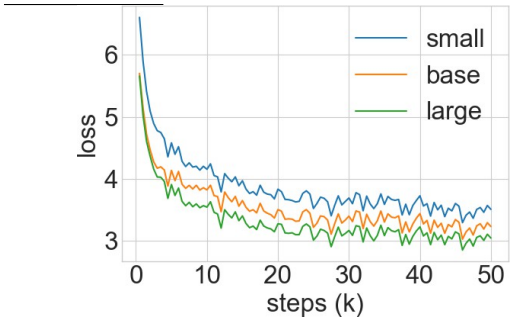


图 4：对不同大小的模式进行对比预训练时的损失 。

5.2 培训行为

我们在图 4 中绘制了不同大小的模型在对比预训练过程中的训练损失。较大的模型在学习区分正负配对方面能力更强。在所有规模的模型中，训练损失都会出现轻微的波动，这表明每批数据的质量和难度存在差异。[3]

我们还评估了不同训练步长下的模型性能。结果表明，模型性能在 20k 步时达到饱和，这与训练收敛大致对应。

步骤	10k	20k	30k	40k	50k
MTEB	56.4	59.0	57.8	57.7	59.0

表 8：无监督对比预训练中不同训练步骤的模型性能。

[3]在模型训练过程中，我们使用固定的随机种子进行数据采样，确保每个模型以相同的顺序遇到数据批次。

5.3 不同训练阶段的影响

为了检验多阶段对比学习的效果，我们对训练策略进行了分析。我们比较了三种设置：a) 仅在从不同来源提取的无监督文本对上进行预训练；b) 仅在超级可见数据集上进行微调；c) 对比预训练后再进行微调。所有模型均从原始 BERT 基础模型初始化。

设置	PT	FT	全职	MTEB
59.0	57.8	62.4		

表 9：不同训练阶段的模型性能。PT 表示仅运行无监督预训练。FT 仅使用监督数据进行模型训练。完全按顺序应用两个阶段。

从表 9 中可以看出，仅仅依靠有监督的数据进行微调不足以获得高质量的文本嵌入模型，这可能是由于数据规模有限。相反，使用网络规模的文本对进行非监督预训练，与仅依赖标记数据进行微调相比，能获得更优的文本嵌入效果。尽管如此，在无监督预训练的基础上以多阶段的方式纳入监督数据，仍然有助于完善所获得的文本嵌入模型。

5.4 训练数据混合

我们研究了预训练数据采样分布中使用的混合比对模型性能的影响。

表 10 报告了重新三元组和 STS 两类任务的成绩以及 MTEB 的平均成绩。我们发现，无论是从每个预

α	检索	STS	MTEB
0	36.7	73.2	55.4
0.3	44.6	75.9	58.9
0.5	44.2	76.5	59.0
1	42.0	75.5	58.3

表 10：预训练数据采样中使用的比率 α 的影响。

$\alpha = 0$）或直接结合所有数据源（$\alpha = 1$）都不是最佳选择。将 α 设为 0.5 可以改善所有任务的结果。

5.5 消融对比目标

这项工作使用了一种改进的对比度目标，它可以在固定批次大小的情况下有效地扩大负值池。我们将其与在预训练和微调阶段仅使用批内负值的虚构对比损失进行了比较。

设置	PT	FT
香草	57.3	61.8
改进	57.8	62.4

表 11：使用批内负值的 vanilla 对比损失与使用扩大负值池的改进对比损失的比较。为了降低计算成本，我们对消融进行了 30k 步的预训练 (PT)。我们报告了 MTEB 的平均得分。

表 11 显示，在预训练和微调阶段，使用改进的弹性损耗始终能提高模型性能。

6 讨论

尽管在英语任务中表现出色，但我们目前的模型只能处理长度小于 512 的文本，因为它是从 BERT 初始化的，缺乏多语言功能。因此，较长的文本必须截断或分割后才能编码。不过，如果有更多的数据工程和计算资源，所述的训练方法可以很容易地扩展到多语言版本，并适应更长的上下文。

另一个问题是对互联网数据进行大规模预训练所带来的数据污染问题。目前，我们只根据文本对的精确匹配进行推断，这是一种过于严格的过滤。Brown 等人（2020）在训练大规模生成语言模型时也强调了这一问题。我们

我们怀疑这是其他模型也会遇到的一个常见问题，但如果没有关于训练数据源的详细信息，量化这个问题就更加具有挑战性（Neelakantan 等人，2022 年）。

此外，本研究中训练的模型是基于非因果结构的双直向上下文关注模型。为因果或前缀语言模型探索类似的预训练方法将很有意义，因为这些模型可以联合优化生成和检索，并将它们统一到一个模型中。

7 结论

本文提出了一种多阶段对比学习方法，用于开发可应用于各种任务的文本嵌入模型。我们的模型得益于多样化的训练数据混合物，使其在单一向量嵌入方面实现了良好的泛化性能。通过对多个基准的广泛评估，我们证明了我们的文本嵌入模型的有效性和通用性。我们未来的工作重点是扩展模型以支持更长的上下文，将其扩展到支持多语言和多模式应用，以及探索提示和说明的优势。

参考资料

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih.2023.带指令的任务感知检索。In *Findings of the Association for Computational Linguistics：ACL 2023*，第 3650-3675 页，加拿大多伦多。计算语言学协会。

汤姆-布朗、本杰明-曼、尼克-莱德、梅兰妮-苏比亚、贾里德-D-卡普兰、普拉富拉-达里瓦尔、阿文德-尼拉坎坦、普拉纳夫-希亚姆、吉里什-萨斯特里、阿曼达-阿斯凯尔、桑迪尼-阿加瓦尔、阿里埃尔-赫伯特-沃斯、格雷琴-克鲁格、汤姆-亨尼根、雷旺-查尔德Aditya Ramesh、Daniel Ziegler、Jeffrey Wu、Clemens Winter、Chris Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020.语言模型是少量学习者。In *Ad- vances in Neural Information Processing Systems*, volume 33, pages 1877-1901.Curran Associates, Inc.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yim- ing Yang, and Sanjiv Kumar.2020.基于嵌入的大规模检索的预训练任务。In *Inter- national Conference on Learning Representations*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser、穆罕默德-巴伐利亚、克莱门斯-温特、菲利普-蒂莱、费利佩-佩特罗斯基-苏奇、戴夫-康明斯、马蒂亚斯-普拉珀特、福蒂奥斯-钱茨斯、伊丽莎-贝丝-巴恩斯、阿里埃尔-赫伯特-沃斯、威廉-赫伯根-古斯、亚历克斯-尼科尔、亚历克斯-帕诺、尼古拉斯-特扎克、唐杰、伊戈尔-巴布什金、苏奇尔-巴拉吉、尚塔努-詹恩、威廉-桑德斯、克里斯托弗-赫塞、安德鲁-N.卡尔、扬-雷克、乔什-阿奇亚姆、维丹特-米斯拉、埃文-森川、亚历克-拉德福德、马修-奈特、迈尔斯-布伦戴奇、米拉-穆拉提、凯蒂-梅尔、彼得-韦林德、鲍勃-麦克格鲁、达里奥-阿莫代、萨姆-麦坎德利什、伊利亚-苏茨克沃尔和沃伊切赫-扎伦巴。2021.评估基于代码训练的大型语言模型。

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin.2016.以亚线性内存成本训练深度网络

Hyunjin Choi, Judong Kim, Seongho Joe, and Youngjune Gwon.2021.在下游 NLP 任务上评估 Bert 和 Albert 句子嵌入性能。*2020年第25届国际模式识别大会（ICPR）*，第5482-5487页。

Alexis Conneau、Douwe Kiela、Holger Schwenk、Loïc Barrault 和 Antoine Bordes。2017.从自然语言推理数据中监督学习通用句子表征。*2017年自然语言处理实证方法大会论文集》*，第670-680页，丹麦哥本哈根。计算语言学协会。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019.BERT：语言理解深度双向变换器的预训练。In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Tech- nologies, Volume 1 (Long and Short Papers)*, pages 4171-4186, Minneapolis, Minnesota.计算语言学协会。

Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xi- aocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, and Ming Zhou.2020. 代码 BERT：编程和自然语言的预训练模型。In *Findings of the Association for Computational Linguistics：EMNLP 2020*，第 1536-1547 页，在线。计算语言学协会。

Luyu Gao 和 Jamie Callan。2021.Condenser：用于密集检索的预训练架构。*自然语言处理经验方法会议*。

Luyu Gao 和 Jamie Callan。2022.用于密集词条检索的无监督感知语言模型预训练。In *Proceedings of the 60th Annual Meeting of the Association for Computational Lin- guistics (Volume 1: Long Papers)*, pages 2843-2853, Dublin, Ireland.计算林学协会。

Tianyu Gao, Xingcheng Yao, and Danqi Chen.2021.SimCSE: Simple contrastive learning of sentence em- beddings.In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Process- ing*, pages 6894-6910, Online and Punta Cana, Do- minican Republic.计算语言学协会。

Daya Guo, Shuai Lu, Nan Duan, Yanlin Wang, Ming Zhou, and Jian Yin.2022.UniXcoder：代码表示的统一跨模态预训练。In *Pro- ceedings of the 60th Annual Meeting of the Associa- tion for Computational Linguistics (Volume 1: Long Papers)*, pages 7212-7225, Dublin, Ireland.计算语言学协会。

Daya Guo, Shuo Ren, Shuai Lu, Zhangyin Feng, Duyu Tang, Shujie LIU, Long Zhou, Nan Duan, Alexey Svyatkovskiy, Shengyu Fu, Michele Tufano, Shao Kun Deng, Colin Clement, Dawn Drain, Neel Sundaresan, Jian Yin, Daxin Jiang, and Ming Zhou.2021.Graphcode{bert}：用数据流预训练代码表示。在*学习表示国际会议上*。

Kelvin Guu、Kenton Lee、Zora Tung、Panupong Pasupat 和 Mingwei Chang。2020.检索增强语言模型预训练。*第 37 届国际机器学习大会论文集》*，《*机器学习研究论文集》第* 119 卷，第 3929-3938 页。PMLR。

Samuel Humeau、Kurt Shuster、Marie-Anne Lachaux 和 Jason Weston。2020.多编码器：快速准确多句子评分的架构和预训练策略。*学习表征国际会议*。

Hamel Husain、Ho-Hsiang Wu、Tiferet Gazit、Miltiadis Allamanis 和 Marc Brockschmidt。2019.Code- searchnet challenge: Evaluating the state of semantic code search.*CoRR*，abs/1909.09436。

Gautier Izacard、Mathilde Caron、Lucas Hosseini、Sebas-tian Riedel、Piotr Bojanowski、Armand Joulin 和 Edouard Grave。2022a.使用对比学习的无监督密集信息检索。*机器学习研究论文集》*。

Gautier Izacard、Patrick Lewis、Maria Lomeli、Lucas Hosseini、Fabio Petroni、Timo Schick、Jane Dwivedi- Yu、Armand Joulin、Sebastian Riedel 和 Edouard Grave。2022b.使用检索增强语言模型的少量学习。

Vladimir Karpukhin、Barlas Oguz、Sewon Min、Patrick Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和

Wen-tau Yih.2020.用于开放域问题解答的密集段落检索。In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769-6781, Online.计算语言学协会。

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova.2019.用于弱监督开放领域问题解答的潜在检索。In *Proceedings of the 57th Annual Meeting of the Association for Computa- tional Linguistics*, pages 6086-6096, Florence, Italy.计算语言学协会。

李博涵、周浩、何俊贤、王明轩、杨一鸣和李磊。2020.关于预训练语言模型的句子嵌入。*自然语言处理实证方法（EMNLP）2020 年会议论文集》*，第 9119-9130 页，在线。计算语言学协会。

李晓楠、龚叶云、沈叶龙、邱希鹏、张航、姚博伦、齐维珍、蒋大新、陈伟柱、段楠。2022.CodeRetriever：用于代码搜索的大规模对比预训练方法。*自然语言处理实证方法 2022 年会议论文集》*，第 2898-2910 页，阿联酋阿布扎比。计算语言学协会。

Zehan Li, Yanzhao Zhang, Dingkun Long, and Pengjun Xie.2023.挑战解码器有助于掩码自动编码器预训练的密集通道检索 .*CoRR*，abs/2305.13197.

Kyle Lo、Lucy Lu Wang、Mark Neumann、Rodney Kin- ney 和 Daniel Weld。2020.S2ORC：语义学者开放研究语料库。In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969-4983, Online.Association for Computational Linguistics.

龙定坤、高琼、邹宽、徐光伟、谢鹏军、郭瑞杰、徐剑锋、蒋冠军、邢璐茜和杨平。2022a.Multi-cpr：用于通道检索的多域中文数据集。*第45届国际信息检索研究与发展大会（ACM SIGIR Conference on Research and Development in Information Retrieval）论文集*。

Dingkun Long, Yanzhao Zhang, Guangwei Xu, and Pengjun Xie.2022b.用于密集语段检索的检索导向屏蔽预训练语言模型。*ArXiv*，abs/2210.15133。

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gre- gory Diamos, Erich Elsen, David Garcia, Boris Gins- burg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu.2018.混合精度训练。*学习表征国际会议*。

Fedor Moiseev、Gustavo Hernandez Abrego、Peter Dorn-bach、Imed Zitouni、Enrique Alfonseca 和董哲。2023.SamToNe：改善对比损失

同塔否定的双编码器检索模型。In *Findings of the Association for Compu- tational Linguistics：ACL 2023*，第 12028-12037 页，加拿大多伦多。计算语言学协会。

Niklas Muennighoff、Nouamane Tazi、Loic Magne 和 Nils Reimers。2023.MTEB：大规模文本嵌入基准。*第 17 届计算语言学协会欧洲分会会议论文集》*，第 2014-2037 页，克罗地亚杜布罗夫尼克。计算语言学协会。

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power、Tyna Eloundou Nekoul、Girish Sastry、Gretchen Krueger 、 David Schnurr 、 Felipe Petroski Such 、 Kenny Hsu 、 Madeleine Thompson、 Tabarak Khan、 Toki Sherbakov、Joanne Jang、 Peter Welinder 和 Lilian Weng。2022.通过强制预训练进行文本和代码嵌入。*CoRR*，abs/2201.10005。

Jianmo Ni 、 Gustavo Hernandez Abrego 、 Noah Con-stant、 Ji Ma、 Keith Hall、 Daniel Cer 和 Yinfei Yang。2022a.句子-t5：来自预训练文本到文本模型的可扩展句子编码器。In *Findings of the Association for Computational Linguistics：ACL 2022*，第 1864-1874 页，爱尔兰都柏林。计算语言学协会。

Jianmo Ni 、 Chen Qu 、 Jing Lu 、 Zhuyun Dai 、 Gustavo Her- nandez Abrego 、 Ji Ma 、 Vincent Zhao、 Yi Luan、 Keith Hall、 Ming-Wei Chang 和 Yinfei Yang。2022b.大型双编码器是可泛化的检索器。*自然语言处理经验方法 2022 年会议论文集》*，第 9844-9855 页，阿联酋阿布扎比。As- sociation for Computational Linguistics.

Barlas Oguz 、 Kushal Lakhotia 、 Anchit Gupta 、 Patrick Lewis、 Vladimir Karpukhin、 Aleksandra Piktus 、 Xilun Chen 、 Sebastian Riedel 、 Scott Yih、 Sonal Gupta 和 Yashar Mehdad。2022.密集检索的领域匹配预训练任务。*计算语言学协会论 文 集 ： NAACL 2022*，第 1524-1534 页，美国西雅图。计算语言学协会。

OpenAI.2023.Gpt-4 技 术 报 告 。 *ArXiv* ， abs/2303.08774。

Alec Radford、 Jong Wook Kim、 Chris Hallacy、 Aditya Ramesh、 Gabriel Goh、 Sandhini Agarwal 、 Girish Sas-try 、 Amanda Askell 、 Pamela Mishkin、 Jack Clark、 Gretchen Krueger 和 Ilya Sutskever。2021.从自然语言监督中学习可转移的视觉模型。*第 38 届国际机器学习大会论文集》*，《*机器学习研究论文集》第* 139 卷，第 8748-8763 页。PMLR。

Alec Radford、 Karthik Narasimhan、 Tim Salimans 和 Ilya Sutskever。2018.通过生成预训练提高语言理解能力。

Thilina C.拉贾帕克塞2023.密集通道检索：架构与增强方法。*第 46 届国际 ACM SIGIR 信息检索研究与发展会议论文集》*。

Samyam Rajbhandari、Jeff Rasley、Olatunji Ruwase 和 Yuxiong He。2020.零：面向训练万亿参数模型的内存优化。*高性能计算、网络、存储和分析国际会议论文集*，SC '20。IEEE Press.

Ori Ram、Yoav Levine、Itay Dalmedigos、Dor Muhlgay、Amnon Shashua、Kevin Leyton-Brown、Yoav Shoham。2023.上下文检索增强语言模型。*ArXiv*, abs/2302.00083.

Nils Reimers 和 Iryna Gurevych.2019.句子 BERT：使用连体 BERT 网络的句子嵌入。*自然语言处理实证方法会议暨第九届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集》*，第 3982-3992 页，中国香港。通用语言学协会。

任瑞阳、吕尚文、曲颖琦、刘静、赵维新、佘巧巧、吴华、王海峰和温继荣。2021.PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval.In *Findings of the Association for Computational Linguistics：ACL-IJCNLP 2021*，第 2173-2183 页，在线。计算语言学协会。

安德鲁-罗森伯格和朱莉娅-赫希伯格。2007.V-measure：基于条件熵的外部宗族评价度量。*2007 年自然语言处理和计算自然语言学习经验方法联合会议（EMNLP-CoNLL）论文集》*，第 410-420 页，捷克共和国布拉格。通用语言学协会。

史维佳、闵绍元、安永道弘、徐敏俊、里奇-詹姆斯、迈克-刘易斯、卢克-泽特尔莫耶和易文韬。2023.Replug：检索增强的黑盒语言模型。*ArXiv*, abs/2301.12652.

苏宏进、史伟佳、葛西淳吾、王义忠、胡玉石、Mari Ostendorf、易文涛、Noah A. Smith、Luke Zettlemoyer 和于涛。2023.一个嵌入器，任何任务：指令调整的文本嵌入。In *Findings of the Association for Compu- tational Linguistics：ACL 2023*，第1102-1121页，加拿大多伦多。计算语言学协会。

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou.2021.为更好的语义和更快的检索而美白句子表征。

Nandan Thakur、Nils Reimers、Andreas Rücklé、Ab- hishek Srivastava 和 Iryna Gurevych。2021.贝尔：

用于信息检索模型零点评估的异构基准。*神经信息处理系统论文集：数据集与基准》*，第 1 卷。Curran.

Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet 、 Marie-Anne Lachaux 、Timothée Lacroix 、 Baptiste Rozière 、 Naman Goyal 、 Eric Hambro 、 Faisal Azhar 、 Aurelien Rodriguez 、 Armand Joulin 、 Edouard Grave 和 Guillaume Lample。2023.Llama：开放而高效的基础语言模型。*ArXiv*，abs/2302.13971。

Aäron van den Oord、Yazhe Li 和 Oriol Vinyals。2018. 对比预测编码的表征学习。*CoRR*，abs/1807.03748。

Ashish Vaswani 、 Noam Shazeer 、 Niki Parmar 、Jakob Uszkoreit 、 Llion Jones 、 Aidan N Gomez 、Ł ukasz Kaiser 和 Illia Polosukhin.2017.注意力就是你所需要的一切。*神经信息处理系统进展》*，第30卷。Curran Associates, Inc.

王亮、杨楠、黄小龙、焦斌兴、杨林军、蒋大新、Rangan Majumder 和魏福如。2022a.Simlm：带代表瓶颈的预训练用于密集通道检索。*计算语言学协会年会*。

王亮、杨楠、黄小龙、焦斌兴、杨林军、蒋大新、Rangan Majumder 和魏福如。2022b.弱监督对比预训练的文本嵌入。*arXiv preprint arXiv:2212.03533*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou.2020.Minilm：深度自我注意力提炼，用于预训练变换器的任务识别压缩。*第 34 届神经信息处理系统国际会议论文集*， NIPS'20 ， 美国纽约红钩。 Curran Associates Inc.

纪尧姆-文泽克、玛丽-安妮-拉肖、亚历克西斯-康诺、维什拉夫-乔杜里、弗朗西斯科-古斯曼、阿 曼德-朱林和爱德华-格拉夫。2020.CCNet：从网络抓取数据中提取高质量单语数据集。*第十二届语言资源与评估大会论文集》*，第 4003-4012 页，法国马赛。欧洲语言资源协会。

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou.2020.MIND：用于新闻推荐的大规模数据集。In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597-3606, Online.计算语言学协会。

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao.2022.Retromae：通过掩码自动编码器预训练面向检索的语言模型。*自然语言处理经验方法会议*。

Yiqing Xie, Xiao Liu, and Chenyan Xiong.2023.使用网络锚点的无监督密集检索训练。*第 46 届国际信息检索研究与发展 ACM SI- GIR 会议论文集*，SIGIR '23，第 2476-2480 页，美国纽约州纽约市。美国计算机协会。

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk.2021.针对密集文本的近似近邻负对比学习（Approximate nearest neigh- bor negative contrastive learning for dense text retrieval.在*学习表征国际会议上*。

周嘉伟、李晓光、尚立峰、罗岚、詹珂、胡恩瑞、张新宇、蒋浩、曹昭、于凡、蒋昕、刘群、陈磊。2022.超链接诱导预训练用于开放域问题解答中的段落检索。In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7135-7146, Dublin, Ireland.计算语言学协会。

A 关于培训数据的更多详情

A.1 训练前数据

网页 在网页中，我们使用标题作为查询，正文作为文档。一些重新来源包括 Common Crawl、Clue Webs 和 MS MARCO 文档。任务的格式可以是：给定一个简短的标题，从一组随机抽样的文本中找出最相关的正文。

学术论文 由于其正式性， 科学文章通常具有较高的质量。对于每篇论文，我们使用标题作为查询，摘要作为文档来构建文本对。我们从不同的网站（如 arXiv、bioRxiv、medRxiv、PubMed 和 Semantic Scholar）挖掘论文，以涵盖广泛的主题。

超链接 互联网上的另一个重要信息是带有文本的超链接，也称为网络锚点。超链接可以为当前论点提供必要的参考。我们将引用论据和参考文献中的文本作为相关文本对进行对比。这类任务更具挑战性，因为它通常涉及多跳推理。我们使用了三种资源来整合链接信息：ClueWeb、维基百科和语义学者论文引文。

社区质量保证 我们还使用了许多来自社区质量保证网站的数据。这类网站的用户界面设计通常采用结构化格式，其中

用户可以用概述性标题和描述性正文的格式来编写问题。这两个字段通常在语义上是一致的。此外，我们还考虑了这类网站的问题答案对。我们使用的数据源包括 StackExchange、Yahoo Answers、Wik- iHow 和 Amazon QA。我们使用文本长度和投票数等简单的启发式方法来过滤低质量数据。

社交媒体 Twitter 和 Reddit 等社交媒体网站通常会让人们就某一事件发布帖子，并有许多网友留言。帖子的结构也包括标题和正文，我们将其视为正对。与社区质量保证类似，帖子评论也被视为数据挖掘的正对。我们从 Reddit 上挖掘数据。

新闻 新闻的结构是标题与正文成对。有些新闻中还有高亮句子。我们利用这些信息构建（query,doc）对。我们使用的数据来自 CCNews、MicrosoftNews、NPR 和 CNNDaily。

知识库 知识库通常存储有关实体或事件的文本描述知识。对（实体、描述）进行挖掘。在这项工作中，我们使用维基百科和 DBPedia 进行文本对挖掘。

代码 代码可被视为文本的另一种形式。自然配对的文本-代码可以作为正对重新使用。我们使用 GitHub 和 StackOverflow 作为两个数据源。我们重复使用从 GitHub 挖掘出的 CodeSearchNet 的训练集。

其他 此外，我们还使用了来自各种网站的数据，如亚马逊上关于商品的评论、关于某个论点的辩论网站、googaq q,a 对（通过搜索日志查询提示 google 搜索框）。

A.2 微调数据

网络搜索 我们使用 MS MARCO 通过重新三重基准。通过从排名靠前的文档检索系统中抽样，挖掘出硬阴性文档，但不包括阳性文档。

开放式质量保证 我们考虑过自然问题、琐事质量保证、网络问题、火锅质量保证等。在开放域问答数据集中，问题及其支持证据段落作为正对提供。在检索系统中排名最靠前的段落是

不包括对问题的回答被视为硬否定。

自然语言推理 之前的工作（Con- neau 等人，2017 年）表明，高质量的句子嵌入可以从有监督的自然语言推理任务中学习到。我们使用包含作为正对，矛盾作为负对来构建训练三元组。在这项工作中，我们使用了 MNLI 和 SNLI 的组合。

事实验证 一个论点及其支持来源（维基百科文档）是正对的。我们使用 FEVER 的训练集作为这项任务的数据源。

转述 意思相近的两个句子被标记为正对。这类数据包括 Quora 和 StackExchangeDupquestion。

其他 除了以前的数据集之外，我们还使用了 MEDI（Su 等人，2023 年）和 BERRI（Asai 等人，2023 年）中发布的不同 NLP 任务和领域的其他数据集。通过这种方法，预训练数据的子采样版本也被纳入微调范围，以避免灾难性遗忘。

A.3 数据来源

预训练数据主要来自之前发布的语料库。由于处理的计算成本较高（Wenzek et al.）由于 Reddit 数据不再免费提供，我们使用了句子转换器和 Oguz et al.[4]和 Oguz 等人（2022 年）的预处理版本进行文本对挖掘。从 hy- perlinks 中挖掘的文本对来自 Zhou 等人（2022 年）和 Xie 等人（2023 年）。我们还包括来自 S2ORC 数据集（Lo 等人，2020 年）的引文对。我们重新使用了来自 BEIR 的 DBPe-dia、辩论论据和 PubMed 语料库（Thakur 等人，2021 年）。维基百科数据来自 Izacard 等人（2022b）。微软新闻数据来自 Wu 等人（2020 年）。Arxiv 数据来自 Kaggle ，medRxiv 和 bioRxiv 数据则是通过请求 2013 年至 2022 年的公共 API 挖掘的。StackExchange 和 StackOverflow 数据来自句子转换器团队维护的预处理版本。[5] 其余的

数据来自嵌入训练数据。[6]除了在某些数据集上使用文本对精确匹配进行训练数据去重之外，训练数据保持原样，不做任何特定过滤。微调数据基本上是以往研究的组合。对于 MS MARCO 数据集，我们使用了 Li 等人（2023 年）的第二阶段检索器挖掘出的硬负数据。对于 NQ 数据集，我们重复使用了 co-Condenser（Gao 和 Callan，2022 年）发布的训练数据。我们使用 SimCSE 发布的 NLI 数据（Gao 等人，2021 年）。其他数据来自 MEDI 和 BERRI（Su 等人，2023 年；Asai 等人，2023 年），但我们舍弃了为每个任务编写的指令，只使用训练三元组。一些随机抽样的例子可以在表 12.

B 海量文本嵌入基准

分类 该任务在线性探测设置中进行评估。嵌入模型被冻结，用于从训练集和测试集中提取每个示例的文本嵌入。训练集的嵌入信息作为输入特征，用于训练一个最大重复次数为 100 次的逻辑回归分类器。测试集的准确率是主要的评估指标。在这种情况下，不同的分类任务只需用少量标注的训练数据训练一个额外的分类头。

聚类 高质量的嵌入模型应将语义相似的文本嵌入到嵌入空间中。通过对测试集中每个句子产生的嵌入模型运行 *k-means* 算法来评估这一特性。使用的是迷你批量 k-means 模型，批量大小为 32 ，k 为标签数。文本被划分为 k *个聚类*。聚类性能通过 v-measure （罗森伯格和赫希伯格，2007 年）来衡量，v-measure 不受聚类标签排列的影响。

重新排序 给定查询和相关与不相关参考文献列表后，重新排序需要根据参考文献列表与查询的相似度对其进行排序。利用嵌入模型可以获得每个查询和参考文本的嵌入，并使用余弦相似度作为排序得分。这种推理设置与文本检索非常相似 ，参考文献集为

[4]https://huggingface.co/datasets/sentence-transformers/reddit-title-body

[5]https://huggingface.co/亚麻句子嵌套

[6]https://huggingface.co/数据集/句子转换器/嵌入训练数据

任务类型	文本三重格式	查询	文档	硬否定
网络搜索	(查询, 通过, 否定)	手指蜂窝织炎症状	以下是最常见的症状 蜂窝组织炎的症状。然而.. .	蜂窝组织炎开始时通常 只有小范围的疼痛和..... 。
开放 QA	(问题, 通过, 否定)	大话西游》第二季 共多少集	大话西游》（电视剧）。 系列获得了多项殊荣。. .	小人物，大世界 第二季的最后几分钟.. .
自然语言推理（句子、蕴涵、矛盾）（	阅读 Slate 的观点	杰克逊的调查结果） 。	Slate 对杰克逊的调	斯莱特对杰克逊的调查结果
事实验证	(论据、证据、其他)	罗曼-阿特伍德是一位 内容创建者。	查结果有自己的看 法。 罗曼-伯纳德-阿特伍德（生于 1983年5月28日）是美国 YouTube 名人。. .	不持任何意见。 第 6 届 Streamy Awards 凯西-尼斯塔特 奖 和 Jesse Wellens，PrankvsPrank . .
转述	(句子、转述、其他)	Lexapro 与 同服会 ？	crestor 有什么反应	dayquil可以与来士普一起服用吗？ 停止服用来士普洛会导致 更长的时间?

表 12：微调数据中的（查询、正向、负向）文本三元组示例。

更小，更难区分。与之前的工作一样，主要的评估指标是 MAP（平均精度）。

检索 我们省略了文本检索评估，因为它与上一章节介绍的内容类似。

文本**对分类** 该任务需要为一对文本指定一个标签。常用的任务包括重复或转述识别，其中的标签是二进制的。相似度得分是两个文本嵌入之间的余弦相似度。使用最佳二进制阈值报告平均精度分数作为主要评估指标。

语义文本相似性 为确定给定句子对之间的相似性，会分配连续的分数，分数越高，相似性越大。采用嵌入模型嵌入句子，并使用余弦相似度计算它们的相似度。估算出的相似度分数与 1 到 5 分的人类标注分数进行比较。我们报告的是斯皮尔曼相关性，它衡量的是等级而不是实际分数，更适合评估句子嵌入的需要。

摘要 这是一项文本生成评估任务，旨在自动评估生成文本的质量。在总结任务中，每个生成的总结的质量都是通过测量其嵌入与基本真实参考文献的嵌入之间的余弦相似度来计算的。在有多个黄金参考文献的情况下，我们使用相似度得分最高的最接近的参考文献进行质量评估。与 STS 任务类似，我们使用文本嵌入模型生成的排序与人工评估之间的 Spearman 相关性进行评估。

C 原始 CodeSearchNet 结果

我们在表 13 中列出了在 Code- SearchNet 上的原始设置结果，其中检索 cor- pus 包含 1k 随机抽样的代码片段。与之前架构和规模相似的开源代码语言模型（CodeBERT（Feng 等人，2020 年）和 GraphCode- BERT（Guo 等人，2021 年））相比，我们的模型在大多数编程语言中都更胜一筹。与 Neelakantan 等人（2022 年）训练的代码嵌入模型相比，我们的模型在性能上仍有差距，该模型以 Codex（陈等人，2021 年）为骨干，在从开源代码中提取的大规模（代码、文本）对上进行训练。如何进一步缩小这一差距值得探讨。

模型	参数	红宝石	联署材料	转到	Python	Java	PHP	平均值
代码ERT	110M × 6	69.3	70.6	84.0	86.8	74.8	70.6	76.0
GraphCodeBERT	110M × 6	84.1	73.2	87.9	75.7	71.1	72.5	77.4
cpt-code S	300M	86.3	86.0	97.7	99.8	94.0	96.7	93.4
cpt-code M	1.2B	85.5	86.5	97.5	99.9	94.4	97.2	93.5
GTEbase	110M	79.6	79.4	84.2	98.8	86.8	86.8	85.9

表 13：CodeSearchNet（Husain 等人，2019 年）的结果。我们将其与 CodeBERT（Feng 等人，2020 年）、GraphCodeBERT（Guo 等人，2021 年）和 cpt-code （Neelakantan 等人，2022 年）进行了比较。这一设置要求在给定自然语言查询的 1K 个候选代码块中找到相关代码块。