arXiv:2601.04395v1 [cs.IR] 7 Jan 2026

多语言稠密检索中被忽视的分级相关性阈值作用

TomerWullach、Ori Shapira 和 Amir
DNCohenOriginAI{tomerw,oris,amirc}@originai.co

摘要

稠密检索模型通常通过对比学习目标进行微调，而这类目标要求使用二元相关性判断，尽管相关性本质上是分级的。我们分析分级相关性分数以及将其转换为二元标签时所用的阈值如何影响多语言稠密检索。我们使用一个由大语言模型标注相关性分数的多语言数据集，考察单语、多语言混合及跨语言检索场景。研究结果表明，最佳阈值会随语言和任务而系统性变化，且往往反映出资源水平的差异。选择恰当的阈值可以提升检索效果、减少所需的微调数据量，并缓解标注噪声；而阈值选择不当则会降低性能。我们认为，分级相关性是稠密检索中一种宝贵但尚未得到充分利用的信号，而阈值校准应被视为微调流水线中一个有原则依据的组成部分。

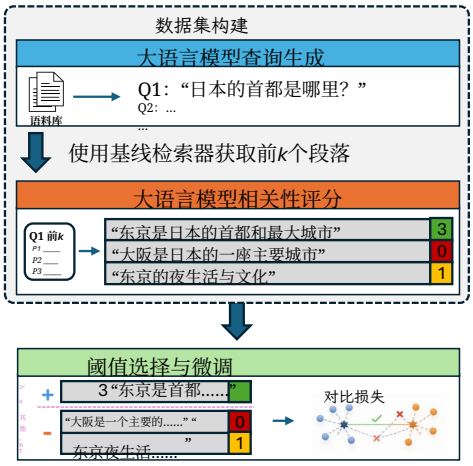


图1：稠密检索模型主要通过对比学习进行训练。与使用二元标签相比，使用分级相关性分数标注数据，能够为正负样本对选择合适的阈值。阈值的选择会显著影响对比式微调的效果，尤其是在多语言环境下。

1 引言

稠密检索模型已成为信息检索的标准方法，如今广泛应用于多语言搜索领域（Wang 等人，2024c；Chen 等人，2024年）。这类模型通常通过对比学习目标进行微调，其依赖于区分正（"相关"）与负（"不相关"）查询-段落对（Gao 等人，2021年；Izacard 等人，2022年）。尽管这种监督信号被构造成二元判定，但在实际应用中，相关性很少是二元的（Vakkari 和 Sormunen，2004；Sakai，2021年）。不同段落对查询的回应程度可能存在显著差异，将这种差异压缩为单一阈值会丢失在训练过程中可能发挥重要作用的信息。

无论相关性标注来自人工标注者还是大语言模型，它们自然都反映了相关程度。问题在于如何利用这种分级结构，而模型最终训练时却使用二元区分。分级标签通过在相关性等级上选定一个阈值，转换为正例和负例。这一步决定了模型将哪些样本视为相关，以及会学着将哪些样本彼此拉远。即使已有分级相关性，阈值应设在何处这一选择，也常被忽视其作为重要设计决策的意义（Scholer 和 Turpin，2009；Radlinski 和 Craswell，2010年）。然而，正如我们的分析所示，它会影响正例的多样性、所学决策边界的锐利程度，甚至还会影响获得性能提升所需的微调数据量。

本文探讨了分级相关性和阈值选择在多语言稠密检索中的作用。我们使用基于大语言模型的流程构建带有细粒度相关性标签的多语言数据集，如图1所示，但我们的分析同样适用于由人工标注

或通过任何多级相关性方案生成的数据集。在三种检索场景中：单语言检索（§4）、多语言混合微调（§5）以及跨语言检索（§6），我们观察到的规律表明，配置阈值具有优势。[1] 最佳阈值因语言和任务而异，通常与语言资源水平、预训练覆盖范围以及查询与语料库之间的语言匹配度等因素相吻合。低资源语言往往受益于更宽松的阈值，以保留更广泛的相关性；而高资源语言在微调期间通常更适合采用更严格的阈值。在微调期间混合多种语言时，阈值之间的相互作用会变得更加显著。

我们的分析还凸显了标注变异性的影响（§4.1）。我们观察到大语言模型标注者之间存在分歧，在资源较少的语言中尤其如此。阈值化提供了一种缓解此类噪声影响的简单方法，同时仍能有效利用分级标签。

综合来看，我们分析中的发现支持一个明确的观点。分级相关性为稠密检索模型提供了有价值的信息，但只有与经过审慎选择的阈值配合使用时，其优势才能得到发挥。因此，在多语言环境中，阈值校准应被视为微调流程的重要组成部分；它能够提升有效性、稳定训练，并减少获得强大检索性能所需的数据量。

2 背景与相关工作

2.1 稠密检索与对比学习

由于能够以语义方式表示文本，稠密检索已成为现代信息检索的主流范式（Zhao 等人，2024年）。该范式下的模型通过对比学习来学习查询和段落嵌入，使相关的配对彼此靠近，并将不相关的配对彼此分开（Karpukhin 等人，2020年；Izacard 等人，2022年；Hambarde 和 Proença，2023年）。尽管存在其他目标函数，包括生成式检索（Li 等人，2025年）和晚期交互模型（Khattab 和 Zaharia，2020年），对比学习仍是标准的训练方法。这促使我们关注相关性信号如何转化为对比监督。

[1]我们将发布用于分析的标注数据和代码。

2.2 二元相关性信号与分级相关性信号

相关性本质上是分级的（Saracevic，2007），在信息检索评估中长期以来也一直如此处理，其中，归一化折损累计增益（Järvelin 和 Kekäläinen，2002年）和期望倒数排名（Chapelle 等人，2009）等指标会明确刻画相关性的不同程度（Roitero 等人，2018年）。相比之下，稠密检索的训练数据大多被二元化。尽管先前研究已探索在训练中使用更细粒度相关性判断的益处（Liu 等人，2021b，a，2022年；Tsirigotis 等人，2025年），出于简便和可扩展性的考虑，二元标注仍是主流做法。

多项研究指出了二元相关性标注的局限性。在用于基于大语言模型的标注时，二元提示会迫使人们在自然存在梯度的情况下作出单一判断，因此可能导致含糊或不一致的判断（Meng 等人，2025年）。相比之下，分级标签能够提供更细致的监督信号（Arabzadeh 和 Clarke，2025年）。人工相关性研究也得出了相同结论。Roi-tero 等人（2018年）表明，二元判断与人工排名的一致性最低，并强调适当的相关性方案取决于数据集和任务要求。

这些研究表明，分级相关性具有信息价值，而将其简化为二元形式会丢失有用信号。然而，既有工作主要关注如何收集分级标签，而不是理解对这些标签进行阈值化会如何影响模型训练；本文将对此展开研究。

2.3 多语言与跨语言检索

稠密检索也推动了多语言检索的发展，支持跨语言的共享训练，并实现了跨语言检索（Nair 等人，2022年；Wang 等人，2024c）。然而，这些模型在不同语言之间表现出显著差异，而这种差异通常与资源水平和预训练接触程度相关。MIRACL（Zhang 等人，2023年）和 CLIRMatrix（Sun 和 Duh，2020年）等基准进一步揭示了多语言和跨语言检索质量的广泛差异，而这种差异受到特定语言资源限制和标注可靠性的影响。

尽管存在这种差异，多语言信息检索中的微调数据几乎总是采用二元标注。鉴于已有证据表明，多语言检索对标注噪声尤其敏感，这一点值得关注。例如，Shen 等人 (2022年) 展示了多样化负例实例的重要性，而 Huang 等人 (2025年) 则发现

在低资源语言中，谨慎选择负例至关重要。这些发现间接表明，捕捉相关性程度具有价值，尽管它们并未直接探讨阈值校准。

我们的分析拓展了这项工作，表明最优相关性阈值取决于语言资源水平、标注一致性和任务设置，凸显了在多语言稠密检索中灵活校准阈值的必要性。

3 分析设置

在段落检索任务中，输入一个查询，并从语料库中返回最相关的段落（Karpukhin 等人，2020年；Chang 等人，2020年）。稠密检索模型会为查询和段落生成向量表示（嵌入），并将查询与最接近的段落进行匹配，通常通过余弦相似度进行衡量。

稠密检索模型采用对比学习目标进行训练，该目标促使模型在嵌入空间中拉近正向（“相关”）查询—段落对，同时推远负向（“不相关”）查询—段落对（Gao 等人，2021年；Wang 等人，2024年）。因此，要针对特定语言或领域微调模型，就需要一组合适的正例和负例。越来越多的从业者依赖大语言模型（LLM）来生成或标注此类数据（Wang 等人，2024年；Lee 等人，2024年），从而实现可扩展且经济高效的数据集构建。我们采用类似的基于大语言模型的流水线来生成用于微调稠密检索模型的实例，如图1所示，具体步骤如下：

1. **查询生成：**给定一个语料库，提示大语言模型为抽样选取的段落生成合成查询。生成的查询涵盖多种信息寻求方式，包括自然语言问题和基于关键词的查询。

2.**候选检索：**对于每个生成的查询，基线稠密检索模型会从语料库中检索排名前-k 的候选段落（当结果较少时，其中一些段落可能不相关）。

3. **相关性标注：**将每个查询-段落候选对呈现给大语言模型，由其按照预先定义的尺度（可能为二元尺度）赋予一个相关性分数。这种排序通常不同于上一步得到的排序，而且通常更可靠，这一点也由Lee 等人（2024年）报告过。

4. **数据集构建：**最终得到的微调数据集包含若干实例（q，p_i，s_i），其中 q 是一个查询，p_i 是检索到的段落之一，k 是分配的相关性分数。

传统上，相关性标注被视为二元判定：相关或不相关。然而，引入分级相关性尺度可以减少标注的任意性，并更好地反映检索任务中常见的连续相关性程度（§2）。高度相似的查询-段落对自然会被标记为明确相关，例如：

> Q：细胞凋亡过度和不足的健康影响
>
> P：“……细胞凋亡过多会导致萎缩，而细胞凋亡不足则会造成细胞增殖失控……”

而较不明显但仍有意义的匹配，可能适合赋予中间分数，而不是一概归为二元的“无关”，例如：

> Q：骑士与兵
>
> P: “邓萨尼国际象棋……一方拥有标准的国际象棋棋子，另一方则有 *32* 个兵。”

由于对比目标最终需要二元监督，因此必须应用一个阈值，将分级相关性分数转换为正负样本：达到或超过阈值的分数归为正例，低于阈值的则归为负例。任意设定这一阈值，仍无法摆脱纯二元标注的局限。后续章节的分析将表明，阈值的选择会带来影响。

较高的阈值会形成较狭隘的相关性概念，使模型在训练期间只接触高度相似的配对。降低阈值则会增加正例的多样性，使模型能够学习更灵活或更远距离的相关性关系。正如我们将看到的，这一决策会显著影响模型在不同语言和任务上的表现。

3.1 我们的合成数据集

为支持我们的分析，我们使用上述流水线构建了一个包含分级相关性分数的合成微调数据集。

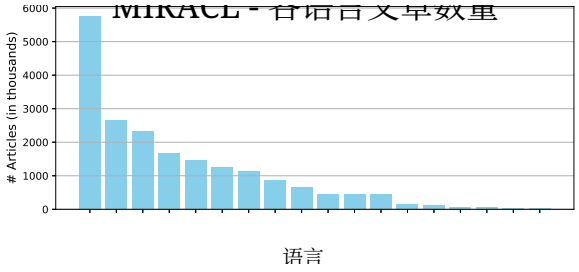


图 2：MIRACL 数据集中各语言的文档数量（维基百科文章）。在分析中，我们使用芬兰语（fi）、阿拉伯语（ar）、日语（ja）、俄语（ru）、西班牙语（es）和英语。

源语料库。 MIRACL[2] (Zhang 等人，2023年) 是一个涵盖 18 种语言的大型语料库，用于检索任务。为开展涵盖不同资源水平的分析，我们选取芬兰语和阿拉伯语作为低资源语言，日语作为中等资源语言，俄语和西班牙语作为高资源语言，英语作为超高资源语言。 图 2 展示了 MIRACL 中各语言的数据分布。对于每种语言，我们从常用开发集之外随机抽取了 50 万个段落。这六个语料库构成了合成数据的基础。[3]

查询生成。 我们使用 `GPT-4o`[4] (OpenAI, 2024年) 生成与目标语言匹配、旨在模拟网络搜索意图的查询。所有语言均使用同一固定提示词，仅修改目标语言说明（见附录 A）。

基线稠密检索模型。 在候选检索中，我们使用 `multilingual-e5-large`[5]嵌入模型（`me5`；Wang 等人，2024c）为每个查询生成覆盖整个语料库的排序，并选取排名前 5 的段落。该模型使用 MIRACL 及相关数据集进行预训练；我们假设其对各语言的熟悉程度大致遵循图 2所示的分布。

相关性分数标注。每个查询—段落对均使用 GPT-4o 进行标注，分配 0 至 3 的相关性分数，其中 3 表示高度相关，0 表示不相关。

研究表明，四点量表在此情境下效果良好（McDonnell 等人，2016年；Roitero 等人，2018年）。提示词和量表定义见附录A。

最后，为了确保分析的公平性，我们将实例数量从 500K 下采样至 300K，同时保持六种语言的分数分布相似。图 3（左侧）展示了所得的分数分布，附录 A 介绍了下采样流程。

使用该数据集，我们接下来考察不同的阈值选择会如何影响微调结果。通过三个用例，我们认为，分级相关性与经过审慎选择的阈值相结合，相较于单独使用二元标注，能够提供更灵活且信息更丰富的监督信号。调整阈值可以让我们探索模型学习“相关”内容时特异性与覆盖率之间的权衡。尽管我们的分析侧重于合成数据，但有关分级相关性和阈值校准的考量同样适用于任何微调数据集，无论其由人工整理还是由大语言模型生成。

4 用例 1：单语检索

单语段落检索是指检索与查询语言相同的语言的段落，仍然是最常见的检索场景。使用多语言稠密检索模型，可以在不同语言之间保持架构和训练条件不变，同时受益于共享的多语言表示。因此，单语检索是分析采用分级相关性进行微调如何影响不同语言表现的理想起点，尤其是考虑到数据可用性、标注质量和预训练覆盖范围方面的差异。

4.1 实验设置

为评估相关性感知对比微调的效果，我们使用了一个数据集，其中的查询-段落对均标注了分级相关性分数。基于该数据集，我们构建了多个微调划分：(i) 改变用于界定正样本（$s \geq \tau$）和负样本（$s < \tau$）对的阈值 $\tau \in \{1, 2, 3\}$，以及 (ii) 改变训练实例的数量，从 50K 到 300K，每次增加 50K。通过这种方式，我们得以分析监督信号的数量和正负样本边界的严格程度如何影响下游性能。

[2] https://huggingface.co/datasets/miracl/miracl-corpus[3]选择这些语言也是为了与跨语言评估中可用的测试语言保持一致，详见 §6。[4]我们使用了 `GPT-4o`，因为对于大规模标注而言，它在多语言能力与成本之间实现了切实可行的平衡。[5] https://huggingface.co/intfloat/multilingual-e5-large

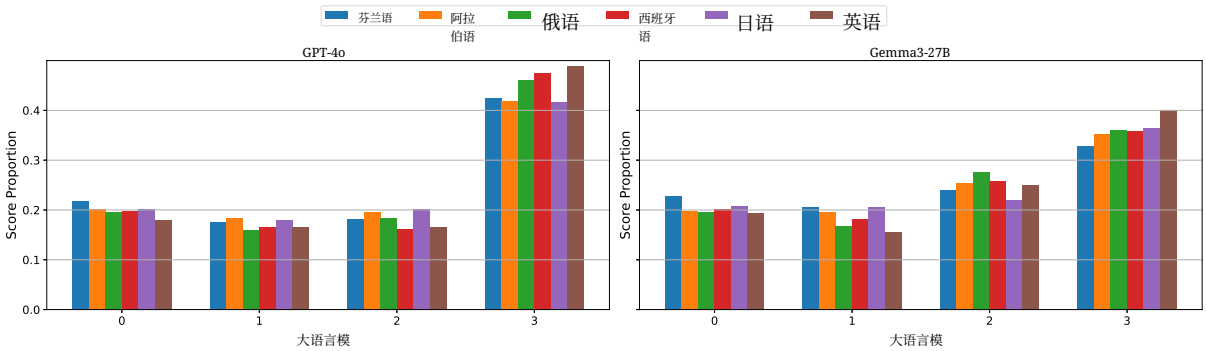


图 3：两种大语言模型标注器（GPT-4o 和 Gemma-3-27B）提供的六种语言相关性分数分布。相关性分数范围为 0（段落与查询无关）至 3（段落与查询高度相关）。主要差异体现在标签 2 和 3 上。

我们对me5 多语言稠密检索器进行了微调，每种配置分别针对六种语言进行微调（技术细节见附录 A）。我们使用 nDCG@10 作为主要指标，在 MIRACL 开发集对应语言的子集上评估了所得模型。

为了进一步评估阈值相对于相关性标注器的影响，我们使用 Gemma-3-27B (Gemma，2025年) 生成的相关性分数，重复了标注和微调流程。与另一种模型相比，它是一个较小的大语言模型。

GPT 与Gemma 的相关性分数。 图 3 展示了两位大语言模型标注者之间的分布差异。尽管类别 0 和 1 的分数分布总体相似，但对于相关性更高的标签，分布差异则更为明显。 GPT-4o 更倾向于标注为 3，而Gemma-3-27B则倾向于选择标签 2 而非 3。这可能表明，GPT-4o 在识别高度相关段落时采用了更宽松的标准；或者，Gemma-3-27B 在做出高置信度的相关性判断时更为谨慎。

图4 比较了不同语言的标注者一致性。低资源语言的一致性明显较低，这与人工标注中常见的模式相符：语言熟练度降低会导致判断更加嘈杂。由于大语言模型能力因语言而异，尤其是在长尾语言中，相关性评分的一致性降低也在情理之中。标注一致性方面的这些差异，进一步促使我们考察阈值选择如何影响微调数据的准备。在这一用例中，我们在实验中纳入了两种标注来源，并比较由此产生的趋势。

4.2 结果

图5 展示了使用不同阈值和数据规模时，各语言的结果；其中使用GPT-4o 作为标注器（附录中的图6 展示了相应的Gemma-3-27B 结果）。我们发现，增加微调数据量通常能提升单语检索性能。起初，较小的训练集会导致性能相较于未经微调的me5 基线明显下降，这表明监督信号不足可能会扭曲模型的内部表示，而非加以改善。随着使用的标注配对增多，性能逐渐恢复，并最终开始趋于平稳。

阈值设定策略至关重要。使用选择不当的相关性阈值训练的模型，往往无法达到基线模型的性能。也许增加微调数据有助于超越基线。相比之下，校准良好的阈值能以更少的微调数据带来明显的性能提升。这凸显了对比式微调需要经过审慎考量的二值化，而非简单地进行阈值化。

不同语言之间呈现出一个显著的规律。对于低资源语言，较低的阈值对模型更有利，因为这能保留更广泛多样的相关性关系。对于高资源语言，则呈现相反趋势：更严格的阈值能带来更好的性能。我们推测，这种现象源于预训练语料库的差异。对于低资源语言，模型可能见过的示例较少，正向信号的范围也较窄，因此在微调期间增加多样性会带来益处。然而，对于高资源语言，模型可能已经编码了丰富多样的相关性区分；注入过多异质正例可能

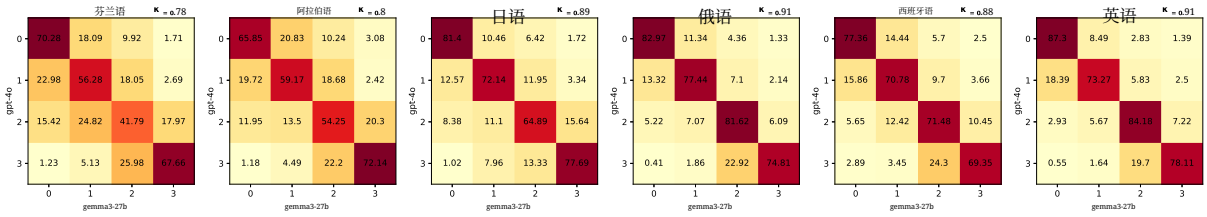


图4：两位大语言模型标注者（GPT-4o 和 Gemma-3-27B）在不同语言上的一致性热图。每个单元格显示以 GPT-4o 的分数为条件的 Gemma-3-27B 分数行归一化频率。高资源语言的对角线一致性更强（加权二次 κ更高），而低资源语言的非对角线噪声更大（ κ更低）。标注一致性的这些差异进一步促使我们考察阈值选择如何影响微调数据的准备。

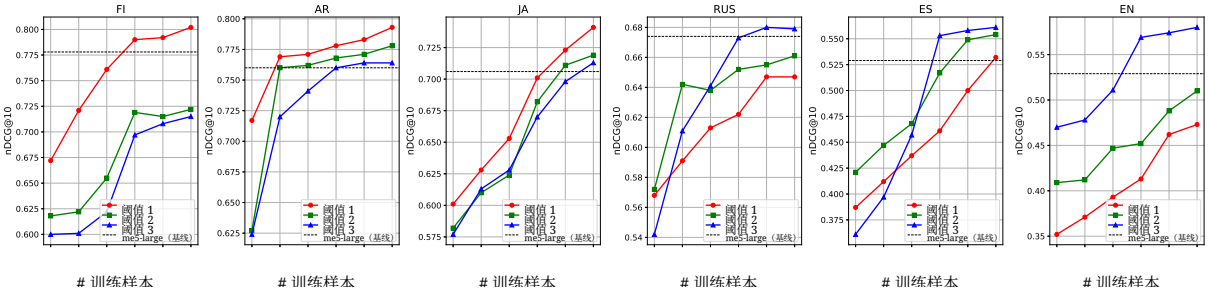


图5：me5 在 MIRACL 上的单语检索性能（nDCG@10，水平虚线）以及微调后的性能，涵盖六种语言。每个子图展示了在不同相关性阈值下，性能随训练集规模的变化情况；这些阈值用于将分级的大语言模型分数转换为对比训练对。低资源语言往往受益于较低的阈值，而高资源语言在较高阈值下表现更好。

引入噪声并降低可区分性。

不同语言之间的差异也可能反映评估数据的特性。我们观察到，某些低资源语言的 nDCG@10 分数更高，这可能表明检索相关段落相对容易，或许是因为查询风格更偏模式化，或者查询与语料库文本之间存在更多词汇重叠。无论这些影响源自训练数据、测试数据、语言差异，还是多种因素共同作用，这些发现都凸显了针对目标语言和领域调整正负样本边界的重要性。

比较不同的大语言模型标注者后，我们发现，基于 Gemma 相关性分数微调的模型，表现不如基于 GPT 相关性分数训练的模型。在本分析范围内，我们只能推测其背后的原因，例如语义敏感度或语言覆盖率。不过，重要的是，不同标注者的定性趋势大体一致，只有少数例外，例如西班牙语略微偏好阈值 2。尽管设置阈值可以减轻一些标注噪声，但更重要的是，相关性判断的差异性也为选择合适的阈值增添了一个考量因素。

鉴于基于 GPT 的相关性分数始终能取得更高的性能，我们在接下来的两个用例中采用这些分数。

5 用例 2：混合语言微调

在多语言稠密检索中，我们通常希望利用其他语言的监督信号，以弥补目标语言标注数据有限的问题，或丰富模型的表征空间（例如，Thakur 等人，2024年）。因此，我们将上一节中的单语段落检索设置扩展到多语言混合数据，考察在混合语言上进行微调时相关性阈值的表现。具体而言，我们研究最优阈值是否仅取决于目标语言，还是也受到微调数据中所包含的特定语言组合的影响。

5.1 实验设置

我们采用第4节所述的相同微调流程，但不再仅使用单一语言进行训练，而是构建混合微调数据集。对于每种目标语言，我们从该语言中选取 150K 个实例，并将其与另

目标语言。	附加语言。	$\tau = 1$	$\tau = 2$	$\tau = 3$
FI	-	0.761	0.655	0.622
	AR	-0.123 (-16.2%)	-0.064 (-9.8%)	-0.100 (-16.1%)
	RUS	+0.049（+6.43%）	-0.017 (-2.59%)	-0.170 (-27.3%)
	ES	+0.056 (+7.4%)	-0.022 (-3.4%)	-0.052 (-8.4%)
	EN	+0.070（+9.2%）	+0.000（+0.0%）	-0.025（-4.0%）
AR	-	0.771	0.762	0.741
	FI	-0.119（-15.4%）	-0.185（-24.3%）	-0.202（-27.3%）
	RUS	-0.058（-7.5%）	-0.138（-18.1%）	-0.141（-19.0%）
	ES	-0.094（-12.2%）	-0.155（-20.3%）	-0.157（-21.2%）
	EN	-0.049（-6.35%）	-0.142 (-18.63%)	-0.15 (-20.2%)
RUS	-	0.613	0.638	0.641
	FI	-0.029 (-4.7%)	-0.093 (-14.6%)	-0.084 (-13.1%)
	AR	-0.021 (-3.4%)	-0.11 (-17.2%)	-0.073 (-11.4%)
	ES	+0.012（+1.95%）	+0.0172（+2.7%）	+0.01（+1.5%）
	EN	+0.022（+3.48%）	+0.019（+2.97%）	+0.013（+2.0 %）
ES	-	0.437	0.468	0.457
	FI	+0.073 (+16.7%)	+0.036（+7.7%）	+0.029 (+6.3%)
	AR	+0.054 (+12.4%)	+0.010（+2.1%）	+0.004 (+0.9%)
	RUS	+0.06 (+13.72%)	+0.041 (+0.87%)	+0.063 (+13.7%)
	EN	+0.064（+14.6%）	+0.044（+9.4%）	+0.065（+14.2%）

表1：在 MIRACL 上使用 150K 个目标语言实例对 me5 进行微调时的单语检索性能（ nDCG@10 ）（基线），以及加入来自另一种语言的 150K 个实例时的相对变化。阈值（τ）会显著影响不同语言组合对性能的影响。

外一种语言的 150K 个实例混合并打乱。在混合数据集中，两种语言统一采用同一个相关性阈值。所有微调后的模型均使用目标语言的 MIRACL 开发集进行单语段落检索评估，并以 nDCG@10 作为主要指标。

5.2 结果

表1 报告了使用仅来自各目标语言的 150K 个实例进行微调时每种目标语言的结果，以及相对 性能变化：加入来自另一种语言的 150K 个实例后观察到的变化。结果展示了不同阈值 $\tau \in \{1, 2, 3\}$ 下的情况。

我们发现，多语言混合会呈现不同的行为，具体取决于目标语言、扩充语言和所选阈值。与单语设置一致，低资源目标语言受益于较低阈值，即使扩充语言属于高资源语言也是如此。相比之下，更高阈值会导致性能下降加剧。对于阿拉伯语，加入任何其他语言的微调数据都无法改善结果，而且往往会使结果变差；阈值越高，这种影响就越明显。相比之下，俄语在更高阈值下表现更好，并且在所有阈值下都能以相近程度受益于高资源扩充语言。然而，当附加语言属于低资源语言时，

	语料库	低至中资源		高资源		
查询		AR	JA	RUS	ES	EN
低-中资源	AR	-2.4%	-15.58%	-15.88%	-11.44%	-6.54%
	JA	-17.71%	-4.2%	-11.73%	-11.31%	-6.21%
High 资源	RUS	-14.93%	-10.03%	+3.2%	+8.71%	+3.7%
	ES	-17.71%	-8.36%	+6.02%	+5.1%	+3%
	EN	+0.90%	+5.15%	+2.90%	+2.14%	+7.23%

表2：将相关性阈值从 1 提高到 3 时，跨语言检索性能的相对变化（负值表示阈值 1 更优，正值表示阈值 3 更优）。在 CLIRMatrix MULTI-8 基准上进行评估。行对应查询语言，列对应语料库语言；模型在查询语言上进行微调。低资源语言通常偏好较低阈值，而高资源语言对之间则受益于更高阈值。完整结果请参见附录中的图7。

性能会下降，但在较低阈值下，下降幅度较小。西班牙语则呈现出更为复杂的模式。使用高资源语言进行扩充时，阈值为 2 时的收益较为有限，而阈值为 1 和 3 时的收益大致相当。然而，使用芬兰语或阿拉伯语（两者均为较低资源语言）进行扩充时，随着阈值降低，性能提升幅度显著增加，尽管西班牙语单独使用时在阈值较高的情况下基线表现更强。

这些复杂（且有些令人困惑）的模式表明，在多语言混合设置中，采用单一全局阈值显然不是最优选择。适用于目标语言的阈值可能不同于适用于增强语言的阈值，而混合语言可能会以复杂的方式改变正例和负例的分布，并与语言特定属性相互作用。总体而言，结果凸显了谨慎选择相关性阈值的重要性，尤其是在多语言微调场景中，因为目标语言和训练混合的构成会共同影响模型行为。

6 用例 3：跨语言检索

在跨语言段落检索中（例如，Litschko 等人，2018年；Asai 等人，2021年），查询以一种语言撰写，而相关段落可能以另一种语言呈现。这种情况在多语言搜索中很常见：即使语料库使用另一种语言，用户仍倾向于或只能用自己的母语构造查询。跨语言检索旨在通过让模型能够检索不同语言中的语义相关内容，弥合这一语言鸿沟。

6.1 实验设置

针对这一应用场景，我们评估了与单语实验相同的微调模型，具体来说，是使用 30 万个实例训练的模型，因为这一配置效果更好。我们在 CLIRMatrix MULTI-8 基准（Sun 和 Duh，2020年）上测试这些模型，该基准支持跨语言查询-段落检索。性能使用 nDCG@10 衡量。评估涵盖查询语言和段落语言均出现在微调数据中的所有语言对，但不包括芬兰语，因为该语言未出现在此测试集中。

6.2 结果

首先，我们的结果表明，选择合适的相关性阈值通常能够提升微调后的跨语言检索性能。当阈值与语言对匹配良好时，微调模型的表现始终优于未微调基线；然而，阈值选择不当往往会导致性能低于基线。这些影响体现在附录中的曲线图7 中，这些曲线展示了模型在不同查询-语料库组合下对不同阈值的响应。虽然增加微调数据最终可能会缓解因阈值不当造成的问题，但选择合适的阈值可以从根本上避免其中大多数问题。

表2总结了这些图表，并通过呈现从阈值 1 调整为阈值 3 时的相对性能差异，对这些影响进行了量化。红色负值表示阈值 1 带来的性能更好，而绿色正值则表示阈值 3 更有利。

总体而言，表2揭示了阈值选择与语言特征之间的系统性依赖关系。当查询语言或语料库语言属于低资源语言时，较低阈值往往表现更佳，这表明在此类情境下，保留相关性对中更大的变异性是有益的。英语作为语料库语言是一个显著例外：较高阈值可能更具优势，这或许得益于英语在预训练语料库中的主导地位和广泛表征。

当查询语言和语料库语言均为高资源语言时（表格右下区域），较高阈值会带来更好的结果，这反映出这些语言中原本就经过良好校准的相关性区分。

尽管评估使用的是 CLIRMatrix，而非 MIRACL，但总体趋势与单语研究结果一致，英语再次成为主要例外。这些结果进一步表明，查询和段落之间相关性差异的程度（通过阈值进行操作化）必须根据语言对和检索任务谨慎选择。

7 结论性讨论

我们的分析表明，分级相关性标注为训练多语言稠密检索器提供了有价值的细微信息，但其有效性在很大程度上取决于如何将这些标注转换为正向和负向监督信号。由于不同语言和检索任务在数据可用性、语义变异以及预训练覆盖范围方面存在差异，严格与宽松相关性信号之间的最佳平衡并非普适。因而，阈值校准至关重要，因为它决定了模型应内化多少相关性变异，使微调能够适应不同的语言和任务条件。

我们还观察到大语言模型标注者之间存在显著分歧，在低资源语言中尤为明显。标注差异是采用受控阈值设定的另一个原因，同时也进一步凸显了需要能够缓冲标注噪声的机制。通过控制分级标签映射为训练信号的方式，可以在保留分级相关性所提供的丰富监督信号的同时，减轻不一致性。

除了这些方法层面的考量，我们的发现还引出了关于预训练和评估数据作用的重要问题。正例偏好变异程度方面极端的跨语言差异，以及若干低资源语言中出乎意料的高分，表明多语言检索器可能继承了不均衡的相关性结构，而这些结构受到其预训练环境或基准测试伪影的影响。我们的分析揭示了这些动态，但要全面理解并解决这些差异，还需要系统研究预训练语料库、标注行为和数据集构建。

总的来说，我们的研究结果支持这样的观点：经过审慎阈值化的分级相关性，是多语言检索的一项强大资源。对相关性边界进行校准，而非假设存在适用于所有情况的统一阈值，应被视为开发多语言检索系统时的一项核心方法论选择。

局限性

我们的分析涵盖了多语言稠密检索中的若干代表性应用场景，但许多现实配置仍有待探索。检索任务、模型架构、标注流程、数据集和语言可能存在显著差异，我们预计阈值行为也会随之变化。我们正是要强调这种多样性。每种设置都需要单独评估，而不应依赖单一的通用阈值。

另一个需要考虑的因素是，改变相关性阈值会改变正训练实例与负训练实例之间的平衡。尽管两组实例的规模仍然较大，且我们的结果显示不同语言之间存在相反的影响（这表明组规模并非主导因素），但这种不平衡是现实场景所固有的，在选择阈值时应予以考虑。

最后，我们的数据集构建流程涉及在进行标注之前，先使用基线模型检索候选段落。与语料库匹配不佳的查询可能会产生更多不相关的候选段落，从而给最终的训练数据引入潜在偏差。这种偏差存在于通过类似流水线构建的任何数据集中，可能会影响微调行为，值得进一步研究。

参考文献

Negar Arabzadeh 和 Charles L. A. Clarke。2025年。基于大语言模型的相关性判断方法基准测试。载于第*48*届*ACM SIGIR*信息检索研究与发展国际会议论文集，SIGIR '25，第3194–3204页，美国纽约州纽约市。美国计算机协会。

Akari Asai、Xinyan Yu、Jungo Kasai 和 Hanna Ha- jishirzi。2021年。结合跨语言稠密段落检索的多语言问答模型。载于 神经信息处理系统进展，第34卷，第7547–7560页。Curran Associates, Inc.。

Wei-Cheng Chang、Felix X. Yu、Yin-Wen Chang、Yim- ing Yang 和 Sanjiv Kumar。2020年。基于嵌入的大规模检索预训练任务。载于 国际学习表征会议。

Olivier Chapelle、Donald Metlzer、Ya Zhang 和 Pierre Grinspan。2009年。分级相关性的期望倒数排名。载于第*18*届 *ACM* 信息与知识管理会议论文集，CIKM '09，第621–630页，美国纽约州纽约市。美国计算机协会。

Jianlv Chen、Shitao Xiao、Peitian Zhang、Kun Luo、Defu Lian 和 Zheng Liu。2024年。BGE M3-嵌入：通过自知识蒸馏实现多语言、多功能、多粒度的文本嵌入。*CoRR*，abs/2402.03216。

Tianyu Gao、Xingcheng Yao 和 Danqi Chen。2021年。SimCSE：句子嵌入的简单对比学习。载于 *2021*年自然语言处理实证方法会议论文集，第6894–6910页，线上，位于多米尼加共和国蓬塔卡纳。计算语言学协会。

Gemma 团队。2025年。Gemma 3 技术报告。预印本，arXiv:2503.19786。

Kailash A. Hambarde 和 Hugo Proença。2023年。信息检索：近期进展与未来方向。*IEEE Access*，11:76581–76604。

Chao Huang、Fengran Mo、Yufeng Chen、Changhao Guan、Zhenrui Yue、Xinyu Wang、Jinan Xu 和 Kaiyu Huang。2025年。提升多语言密集检索的数据利用率。载于 *2025*年自然语言处理实证方法会议论文集，第12362–12378页，中国苏州。计算语言学协会。

Gautier Izacard、Mathilde Caron、Lucas Hosseini、Sebas- tian Riedel、Piotr Bojanowski、Armand Joulin 和 Edouard Grave。2022年。通过对比学习进行无监督稠密信息检索。 机器学习研究汇刊。

Kalervo Järvelin 和 Jaana Kekäläinen。2002年。基于累积增益的信息检索技术评估。*ACM*信息系统汇刊，20(4):422–446。

Vladimir Karpukhin、Barlas Oguz、Sewon Min、Patrick Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wen-tau Yih。2020年。面向开放域问答的密集段落检索。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第6769–6781页，线上。计算语言学协会。

Omar Khattab 和 Matei Zaharia。2020年。ColBERT：通过基于BERT的上下文化后期交互实现高效且有效的段落搜索。载于第*43*届国际*ACM*信息检索研究与发展会议论文集，SIGIR '20，第39–48页，美国纽约州纽约市。计算机协会。

Jinhyuk Lee、Zhuyun Dai、Xiaoqi Ren、Blair Chen、Daniel Cer、Jeremy R. Cole、Kai Hui、Michael Boratko、Rajvi Kapadia、Wen Ding、Yi Luan、Sai Meher Karthik Duddu、Gustavo Hernandez Abrego、Weiqiang Shi、Nithi Gupta、Aditya Kusupati、Pra- teek Jain、Siddhartha Reddy Jonnalagadda、Ming- Wei Chang 和 IftekharNaim。2024年。Gecko：从大语言模型蒸馏而来的多功能文本嵌入。预印本，arXiv:2403.20327。

Xiaoxi Li、Jiajie Jin、Yujia Zhou、Yuyao Z hang、Peitian Zhang、Yutao Zhu 和 Zhicheng Dou。2025年。从匹配到生成：生成式信息检索综述。*ACM*信息系统汇刊，43(3)。Robert Litschko、Goran Glavaš、Simone Paolo Ponzetto 和 Ivan Vuli´c。2018年。仅使用单语数据的无监督跨语言信息检索。载于第*41*届*ACM*国际信息检索研究与发展会议，SIGIR '18，第1253–1256页，美国纽约州纽约市。美国计算机协会。

Peiyang Liu、Sen Wang、Xi Wang、Wei Ye 和 Shikun Zhang。2021a。QuadrupletBERT：一种用于基于嵌入的大规模检索的高效模型。载于*2021*年北美计算语言学协会分会会议论文集：人类语言技术，第3734–3739页，线上。计算语言学协会。

Peiyang Liu、Xi Wang、Sen Wang、Wei Ye、Xiangyu Xi 和 Shikun Zhang。2021b。通过标签增强改进基于嵌入的大规模检索。载于计算语言学协会：*EMNLP 2021*论文集，第133–142页，蓬塔卡纳，多米尼加共和国。计算语言学协会。

Peiyang Liu、Xiangyu Xi、Wei Ye 和 Shikun Zhang。2022年。用于文本挖掘的标签平滑。载于第*29*届国际计算语言学会议论文集，第2210–2219页，庆州，大韩民国。国际计算语言学委员会。

Tyler McDonnell、Matthew Lease、Mucahid Kutlu 和 Tamer Elsayed。2016年。为什么相关？收集用于相关性判断的标注者理由。*AAAI*人类计算与众包会议论文集，4(1)：139–148。

Chuan Meng、Negar Arabzadeh、Arian Askari、Mohammad Aliannejadi 和 Maarten deRijke。2025年。使用大语言模型生成的相关性判断进行查询性能预测。*ACM*信息系统汇刊，43(4)。

Suraj Nair、Eugene Yang、Dawn Lawrie、Kevin Duh、Paul McNamee、Kenton Murray、James Mayfield 和 Douglas W. Oard。2022年。用于构建跨语言稠密检索模型的迁移学习方法。载于信息检索进展，第 382–396 页，查姆。施普林格国际出版公司。OpenAI 团队。2024年。GPT-4o 系统卡。预印本，arXiv:2410.21276。Filip Radlinski 和 Nick Craswell。2010年。信息检索指标敏感性的比较。载于第 *33* 届 *ACM* 国际信息检索研究与开发会议论文集，SIGIR '10，第 667–674 页，美国纽约州纽约市。美国计算机协会。

五点相关性评分二值化。Kevin Roitero、Eddy Maddalena、Gianluca Demartini 和 Stefano Mizzaro。2018年。关于细粒度相关性等级。载于第 *41* 届 *ACM* 国际信息检索研究与开发会议，SIGIR '18，第 675–684 页，美国纽约州纽约市。美国计算机协会。酒井哲也。2021年。分级相关性，第 1–20 页。新加坡施普林格，新加坡。Tefko Saracevic。2007年。相关性：文献综述及关于信息科学中相关性概念的思考框架。第二部分：相关性的本质与表现形式。美国信息科学与技术学会会刊，58(13):1915–1933。

Falk Scholer 和 Andrew Turpin。2009年。检索评估中的指标与相关性不匹配。载于信息检索技术，第50–62页，柏林，海德堡。施普林格柏林海德堡。

Tianhao Shen、Mingtong Liu、Ming Zhou 和熊德意。2022年。从黑沙中淘金：利用困难负样本与错误负样本进行多语言密集段落检索。载于*2022*年自然语言处理实证方法会议论文集，第10659–10670页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

孙硕和 Kevin Duh。2020年。CLIRMatrix：一个规模庞大的双语和多语言数据集集合，用于跨语言信息检索。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第 4160–4170 页，线上。计算语言学协会。

NandanThakur、JianmoNi、Gustavo Hernandez Abrego、John Wieting、Jimmy Lin 和 Daniel Cer。2024年。利用大语言模型合成适用于多种语言的训练数据，用于多语言密集检索。载于北美计算语言学协会分会：人类语言技术会议 *2024* 年论文集（第 *1* 卷：长篇论文），第 7699–7724 页，墨西哥城，墨西哥。计算语言学协会。

Christos Tsirigotis、Vaibhav Adlakha、Joao Mon- teiro、Aaron Courville 和 Perouz Taslakian。2025年。BiXSE：通过概率分级相关性蒸馏改进稠密检索。载于第二届语言建模会议。Pertti Vakkari 和 Eero Sormunen。2004年。相关性等级对交互式信息检索成效的影响。美国信息科学与技术学会期刊，55(11)：963–969。

Liang Wang、Nan Yang、Xiaolong Huang、Binxing Jiao、Linjun Yang、Daxin Jiang、Rangan Majumder 和 Furu Wei。2024年。通过弱监督对比预训练获得文本嵌入。预印本，arXiv:2212.03533。

Liang Wang、Nan Yang、Xiaolong Huang、Linjun Yang、Rangan Majumder 和 Furu Wei。2024年。借助大语言模型改进文本嵌入。载于第*62*届计算语言学协会年会论文集（第*1*卷：长篇论文），第11897–11916页，泰国曼谷。计算语言学协会。

Liang Wang、Nan Yang、Xiaolong Huang、Linjun Yang、Rangan Majumder 和 Furu Wei。2024年。多语言 E5 文本嵌入：技术报告。预印本，arXiv:2402.05672。

Xinyu Zhang、Nandan Thakur、Odunayo Ogundepo、Ehsan Kamalloo、David Alfonso-Hermelo、Xiaoguang Li、Qun Liu、Mehdi Rezagholizadeh 和 Jimmy Lin。2023年。MIRACL：涵盖18种多样语言的多语言检索数据集。计算语言学协会会刊，11:1114–1131。

Wayne Xin Zhao、Jing Liu、Ruiyang Ren 和 Ji-Rong Wen。2024年。基于预训练语言模型的稠密文本检索：综述。*ACM*信息系统汇刊，42(4)。

A 实现细节

大语言模型标注提示。准备合成数据集的第一阶段，是为从 MIRACL 语料库中抽取的段落生成查询。我们向 GPT-4o 提供了如下提示：

> 你的任务是撰写一个 <LANG> 查询，以查找给定上下文段落中包含的信息。
>
> 查询应与该段落相关，但不必涵盖段落的全部内容。查询应使用<LANG>，简短、简单，并采用网络搜索查询的风格，例如基于问题或关键词的形式。
>
> 段落内容如下：<PASSAGE>

随后会提示一个大语言模型（GPT-4o 或 Gemma-3-27B）为查询与段落对打出相关性分数。提示词如下：

> 你是一名搜索质量评估员，负责评估段落的相关性。给定一个查询和一个段落，你必须按照 0 到 3 的整数刻度给出分数，并将分数放在括号内 `[]` ，具体含义如下：
>
> `[3]` =高度相关：该段落专门针对查询，并包含确切答案。
>
> `[2]` = 相关：该段落为查询提供了部分答案，但缺少所需的确切信息，因此不能视为高度相关。
>
> `[1]` =有关联：该段落似乎与查询相关，但并未回答查询。
>
> `[0]` = 不相关：该段落与查询毫无关系。
>
> 查询：<QUERY>
>
> 段落：<PASSAGE>

大语言模型标注成本。 我们使用 Azure OpenAI API 调用 GPT-4o，在查询生成阶段花费了约 3000 美元，在相关性评分标注上花费了 6000 美元。使用 Gemma-3-27B 进行标注耗用了数小时的 GPU 服务器时间。

对已标注实例进行下采样。 共有 500K 个查询-段落对完成了标注，但为了使六种语言的相关性分数分布大致保持一致，我们按以下方式下采样至 300K：

1. 确定非零分数（即 1、2 和 3）数量最少的语言。2. 然后，分别从其余语言中抽取相同数量的实例。3. 对于每种语言，抽取分数为 0 的实例，直至总数达到 300K。

如图 3 所示，不同语言之间的分数分布非常相似。

测试集规模。对于我们在用例 1 和用例 2 中进行单语分析所涉及的语言，MIRACL 开发集根据语言不同包含 648 至 2896 个实例（查询）（参见: https://huggingface.co/datasets/miracl/miracl）。在用于跨语言检索任务（用例 3）的 CLIRMatrix 基准测试中，每种语言包含 1000 个实例（查询）。

微调稠密检索模型。 我们使用从整理后的数据集中筛选出的（q, p）对，并根据实验要求使用相应数量的实例，对基于编码器的 Transformer 模型 multilingual-e5-large 进行了微调。模型使用批大小为 32、带批内负例的 InfoNCE 损失、初始学习率为 1×10^{-6}、5% 的预热比例，训练两个周期。为确保不同阈值和数据规模配置之间的公平比较，我们在训练周期数、优化器和学习率调度方面保持训练流程不变。尽管不同阈值会导致不同的损失表现，但我们的目标是在相同训练预算下比较不同的监督信号设置，而不是分别对每种配置进行单独优化。

模型在一台配备 24 个 CPU、220 GiB 内存和一块 80GB NVIDIA A100 GPU 的服务器上进行训练。模型完成微调平均耗时约 5 小时。

AI 的使用。 在编写数据集创建流水线的代码时，我们使用以 GPT-4o 为底层模型的 Copilot 对代码进行了少量改进。我们还使用 Copilot，根据实验结果协助准备论文中的图表。最后，在撰写本文时，我们使用 ChatGPT 5.1 协助改写了部分文本，以提升其清晰度和连贯性。每次使用 AI 后，我们都进行了人工核查和编辑，以确保内容准确且真实。

B 补充结果

用例 1（单语检索）。图6展示了单语任务（用例 1，§4）的结果，其中`Gemma-3-27B` 被用作相关性分数的标注器。此图应与图5 进行比较，后者使用`GPT-4o` 作为标注器。正如第4节所述，使用`Gemma` 作为标注器时，效果不如前者，但总体趋势相似，仅有少数差异。这再次说明，根据语言、数据和任务合理设置阈值十分重要。

用例 3（跨语言检索）。图7展示了跨语言实验的结果。图中每个子图均报告了两个检索方向（X→Y 和 Y→X；表示为 <查询语言> → <语料库语言>），并以未经微调的`multilingual-e5-large`模型的基线结果作为参考。图中结果也汇总于表2。讨论请参见第6节。

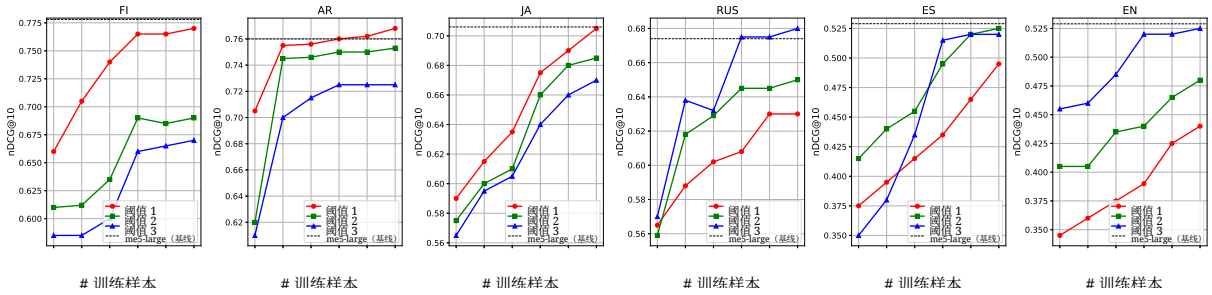
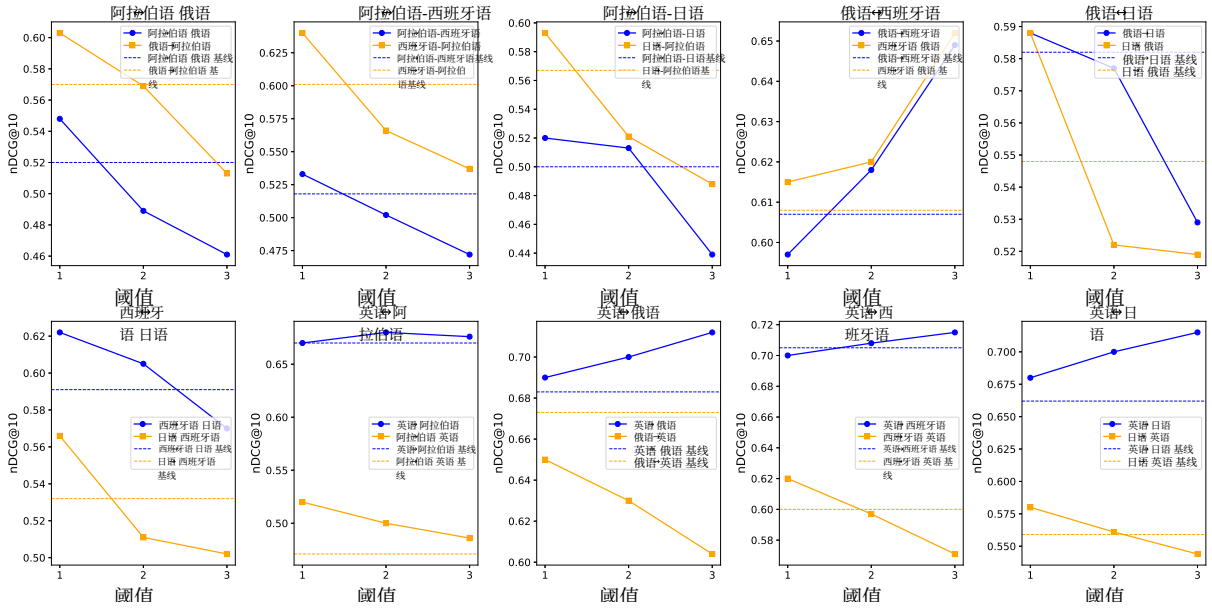


图6：在 MIRACL（水平虚线）上以及微调后的 me5 在六种语言上的单语检索性能（nDCG@10），其中使用了由 Gemma-3-27B标注的相关性分数。每个子图展示了在不同相关性阈值下，性能随训练集规模的变化；这些阈值用于将分级的大语言模型分数转换为对比训练对。低资源语言往往更受益于较低阈值，而高资源语言在较高阈值下表现更佳。使用 图5 查看使用 GPT-4o 作为相关性分数标注器时的结果。



不同语言对在各阈值下的跨语言检索性能（以<查询语言> → <语料库语言>表示）。每个子图对应一个语言对，展示了在 CLIRMatrix MULTI-8 基准上评估的两个检索方向的 nDCG@10 值。水平虚线表示使用 me5 在各方向上达到的基线性能。 表2 总结了结果。