arXiv:2601.00366v1 [cs.CL] 1 Jan 2026

BERT-JEPA：重 组织 CLS 嵌入，实现语言不变的语义 s

Taj Gillin [*1] **Adam Lalani** [*1] **Kenneth Zhang** [*1] **Marcel Mateos Salles** [*1]

摘要

联合嵌入预测架构（JEPA）是一种新颖的自监督训练技术，近期已在多个领域展现出应用前景。我们提出 BERT-JEPA（BEPA），这是一种训练范式，通过为 BERT 风格模型添加 JEPA 训练目标，旨在解决坍塌的 [CLS]嵌入空间问题，并将其转变为语言无关空间。这种新结构提升了模型在多语言基准测试上的表现。

1. 引言

像 BERT 这样的语言模型（Devlin 等人，2019年）被用于为自然语言处理任务生成丰富的嵌入。对于许多下游任务，BERT 嵌入可以直接取用，并用于取得有竞争力的性能。然而，这些嵌入未能捕捉语言背后的真实表征。其 [CLS] 标记是最常用于下游任务的标记之一，但研究发现，它无法充分捕捉句子相似度任务中句子的含义（Reimers 和 Gurevych，2019年）。语言是思想的具体呈现形式，即便句法和结构各异，仍能保留其含义。西班牙语和英语之间的基本翻译就体现了这一点：“El gato es rojo”和“The cat is red”。它们通过不同的语言系统承载并传达相同的意图和思想。因此，我们提出这样一个问题：能否教会 BERT 表示不依赖于语言系统的信息，只表征句子和信息背后的真实含义？我们转向新兴的联合嵌入预测架构（JEPA）研究，这类架构专注于预测两个不同样本之间的潜在表示（LeCun，2022）。近期，这类架构已在视觉（Assran 等人，2023年；2025年；Bardes 等人，2024）和语言任务中展现出潜力，尤其是在代码文本任务中（Huang 等人，2025年）。这自然可以扩展到通用语言，让模型学会根据一个样本预测另一个含义相同样本的嵌入，从而促成一个共享的嵌入空间，我们将其称为模型的“思想空间”。这个丰富的嵌入空间对应于人类所拥有的语言之上的抽象——我们的思想。我们介绍了新的训练范式 BERT-JEPA（BEPA），旨在利用 JEPA 构建丰富的上下文理解能力。我们得到以下结果：

1. **BEPA 微调彻底重组了 [CLS]嵌入空间，使其形成语义优先的结构。**
2. **BEPA 微调将主成分分析表示从低秩转变为更高秩。** 3. **[CLS] 嵌入重组几乎不会损失英语表现。** 4. **BEPA 微调提升了多语言任务的表现。**

2. 背景

过去六年间，自然语言处理领域发生了重大变化。BERT 的发布（Devlin 等人，2019年）借助新兴的 Transformer架构（Vaswani 等人，2023年）彻底改变了语言理解领域，并构建出能够为下游任务生成丰富嵌入的深度网络。BERT 推出了新的预训练任务，其中最重要的是掩码语言建模（MLM）任务，该任务使 BERT 的双向注意力能够正常发挥作用。

后续模型，例如 *RoBERTa* (Liu 等人，2019年）和 *XLM-RoBERTa* (Conneau 等人，2018a）是对 BERT 的改进，采用了更长的训练时间，并移除了下一句预测（NSP）和翻译语言建模（TLM）等不必要的训练任务。这些模型专注于构建强大的潜在表示，因此与基础版 BERT 相比，在各项基准测试中的表现显著提升。我们的实验将使用*RoBERTa* 和 *XLM-RoBERTa* 作为基线模型和比较对象。

尽管取得了这些进展，语言模型仍未能实现语言不变性（Ribeiro 等人，2018年），因为对示例进行细微且不改变语义的修改，就可能导致模型

[*]贡献相同 [1]美国罗得岛州普罗维登斯布朗大学计算机科学系。联系邮箱：Marcel Mateos Salles <marcel_mateos_salles@brown.edu>。

改变其行为和预测。这意味着语言模型往往仍会编码特定于语言的特征。

联合嵌入预测架构（JEPA）改变了模型的训练方式。不同于在输入空间中对齐模型，LeCun (2022) 提出了一种在嵌入空间中对齐模型的架构，从而形成更稳健的表示和理解。在实践中，这些架构具有良好的可扩展性，在视觉任务中表现出色，在各类下游任务中与现有方法相比毫不逊色，甚至更胜一筹（Assran 等, 2023年; 2025年;Bardes 等, 2024）。这些模型生成的嵌入包含丰富的上下文信息，并由最具影响力的特征构成（Littwin 等, 2024）。最近，有研究针对语言任务提出了一种 JEPA 变体（Huang 等,2025年），取得了令人瞩目的表现，也非常适用于 NL-Regex 和 NL-SQL 场景。受这篇论文启发，我们将 JEPA 应用于 BERT 及其变体，以构建更丰富的嵌入空间。

为评估我们新嵌入的质量，我们计划采用多种基准测试和技术。通常，BERT 及其变体会在 SST-2（Socher 等人，2013a）上进行基准测试。此外，STS 跨语言数据集（Cer 等人，2017；Enevoldsen 等人，2025年；Muennighoff 等人，2022）可用于了解不同语言之间的语义表征是否对齐，以及 BEPA 是否能够摆脱语言的影响。除了下游评估，我们还可以使用 t-SNE 可视化（Cai 和 Ma，2022）和各向同性（Mickus 等人，2024）等诊断方法来研究嵌入的特性。我们计划使用所有这些技术来评估 BEPA 的性能。

3. 方法：一种新颖的 JEPA风格BERT微调框架

本节介绍使用 BERT-JEPA（BEPA）训练 BERT 风格模型所采用的训练方案。有关框架的完整细节，请参见第 A.5 节。训练细节和所用超参数见第 A.3 节，对应的代码库和相关细节见第 A.4 节。我们首先介绍训练框架，然后深入探讨所用的损失函数。

3.1. BEPA 框架

我们提出了一种新颖的 BERT 训练方法，可转换 BERT 模型当前的嵌入空间，使其不受语言系统影响，并修复已知的 CLS标记坍缩问题（Reimers 和 Gurevych，2019年）。我们首先使用一个经过预训练的基础 *RoBERTa* 模型（*xlm-roberta-base*）（Conneau 等人，2019年），但任何 BERT 风格模型都适用。我们将两个句子并排嵌入，并将它们打包为一个 Transformer 输入。添加特殊标记后，文本将被分词，如图 4。我们分别测试了两种适用于此设置的方法： 单语 和双语。其中，单语 只将同一语言的样本并排打包，而双语 则允许将不同语言的样本打包在一起。论文正文重点介绍双语 打包策略的结果；单语及其他消融实验的结果见第A.9节。随后，该设置采用 BERT 经典的 MLM 训练（ Devlin 等人，2019 年）。这样做是为了保留 BERT 风格的 Transformer 强大的语言能力。

接下来，我们介绍 BEPA 框架的新颖部分。由于 BERT 风格的模型将最终层 [CLS]词元视为序列的潜在表示，因此我们使用[CLS] 词元计算对齐损失，以鼓励跨语言不变性。如图 5所示，我们进行了两次独立的前向传播，每次都将一个句子单独遮蔽。随后提取每个句子的 [CLS] 词元的潜在表示，并通过$L_{Alignment}$进行对齐；更多信息请参见第3.2节。在对齐 [CLS]词元之前，我们采用了恒等预测器，但未来的工作应测试其他类型和风格的预测器。图 6展示了此步骤的示意图。

3.2. BEPA 损失

我们选择的损失函数不能是基础损失，因为它必须兼顾 BEPA 所采用的双重训练目标。BERT 风格的 MLM 与基于 CLS 词元的 JEPA 对齐共同构成了一个新的训练目标:

$$L_{BEPA} = L_{MLM} + \lambda L_{Alignment}$$

其中 λ 是一个超参数，用于确定应赋予 $L_{Alignment}$多大权重。我们采用交叉熵，与典型的 BERT 风格模型类似，如 L_{MLM} 以及对于$L_{Alignment}$，我们发现 InfoNCE（van den Oord 等人，2019年）效果最佳。未来工作应测试不同目标函数的影响，例如 SigREG（Balestriero 和 LeCun，2025年）。该框架具有充分的可扩展性，当前的损失函数可替换为用户所需的任意损失函数。我们发现，均方误差（MSE）和余弦相似度在对齐任务中的效果不够理想，因为它们无法抑制 [CLS]坍塌。

4. BEPA 训练在语义上重组 CLS 嵌入空间

在本节中，我们研究 BEPA 微调后得到的 [CLS]空间。微调之前，BERT 风格的模型呈现出坍塌的嵌入空间。这一点可以在第 A.6 节中看到，尤其是表 4 和 5：无论语义是否相似，任意一对样本都表现出较高的余弦相似度。这一趋势在常规的 *RoBERTa* 和 *XLM-RoBERTa* 中均有体现。

然而，经过 BEPA 微调后，情况就不同了。我们的模型开始呈现出这样的趋势：语义相似的样本对（无论是同一语言内还是跨语言）具有较高的对齐度，而语义不同的样本对（无论是同一语言内还是跨语言）对齐度较低。要了解这一趋势，请参阅图 10 和附录第 A.8 节。

为进一步研究重组后的 [CLS] 嵌入，我们进行了主成分分析，以解释空间中各个主成分的方差。我们发现，单语和双语模型将方差分布在更多的主成分上，这表明更多的嵌入维度包含有效信息，因此其 [CLS] 表征的冗余度更低。这一发现可见于图1。

此外， [CLS]嵌入的 t-SNE 可视化显示，*RoBERTa*和*XLM-RoBERTa*呈现明显的语言特定簇，而我们的单语和双语模型在不同语言之间表现出更多重叠，表明其嵌入空间具有更强的语言不变性。这些簇见图2。

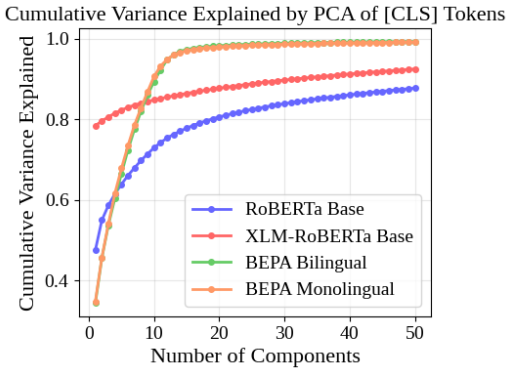


图*1*。展示 [CLS] 嵌入空间各主成分方差的主成分分析图。**BEPA 微调使高方差主成分的数量增加**（↑）。与此同时，标准*RoBERTa* 和*XLM-RoBERTa* 的方差主要集中在单个**主成分**上，分别为 47% 和 78%。相比之下，两个 BEPA 模型的第一主成分占比均为 34%。

5. 语义组织的词元空间提升跨语言迁移

本节介绍我们经过 BEPA 微调的模型在各项基准测试上的表现。我们在 GLUE、XNLI 和 MLQA 上进行了测试。这些基准测试的完整结果见第 A.9 节。

我们使用 XNLI 基准测试，检验 BEPA 微调是否能帮助 BERT 风格模型理解句子级语义，并将这种理解应用于不同语言。如果这些经过 BEPA 微调的模型是在语义对齐的“思想空间”中进行推理，那么它们应该能够能够更准确地跨语言执行。为运行 XNLI 基准测试，我们在英语上微调了两个训练轮次，然后对所有其他语言进行了零样本评估。我们的模型结果见表 1和 9。我们发现，BEPA 微调（具体来说，采用 双语打包设置）能够持续提升所有语言的性能，凸显出更好的语言不变性和语义对齐的 [CLS]嵌入空间。

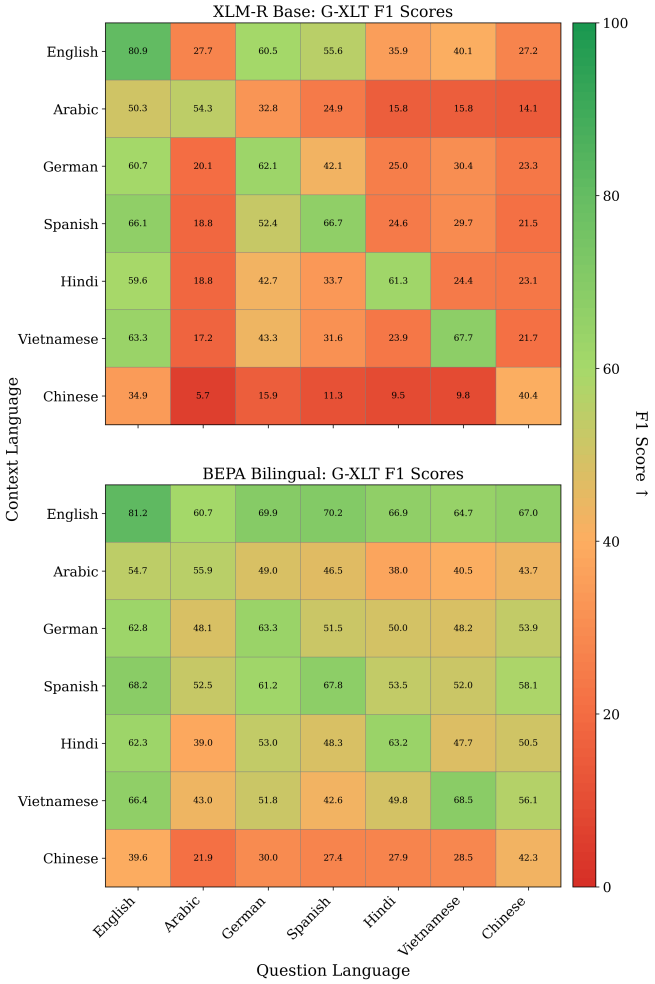


图 *3*。MLQA 基准测试上的 G-XLT 矩阵表现。49 个语言对（7 种上下文语言 ×7 问题语言）的 F1 分数。对角线条目表示单语表现，非对角线条目则显示跨语言迁移。（上）*</style*：*XLM-RoBERTa*Base 在单语任务上表现出色，但在跨语言场景中表现明显较差，尤其是那些不涉及英语的场景。（下）：*BEPA*双语模型展现出更强的跨语言迁移能力，在保持单语表现出色的同时，各语言对上的表现也更加均衡。这表明**BEPA 在表示空间中学习了词元级语义对齐**。

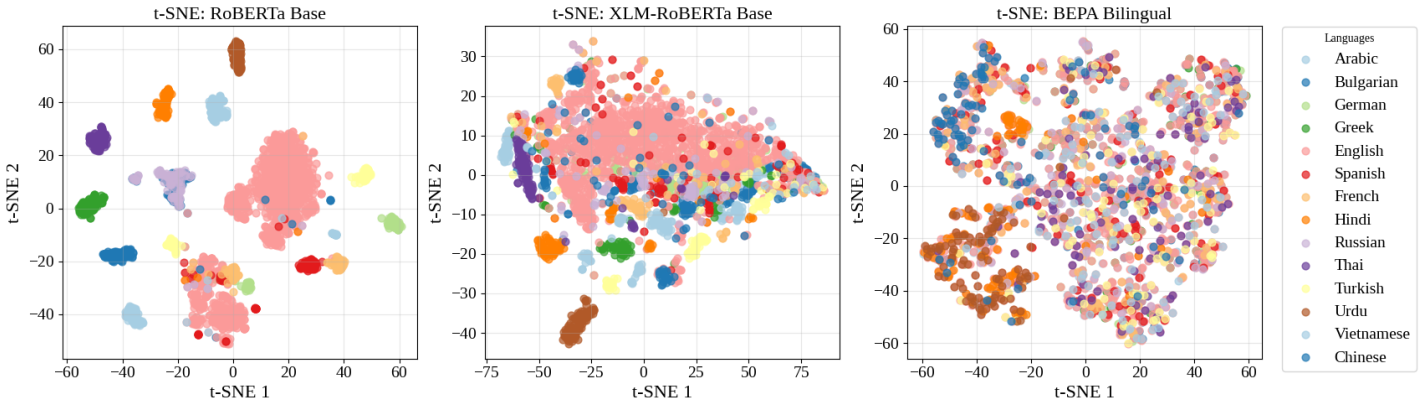


图 *2*。 （左）： t-SNE 图，展示 *RoBERTa* 的结果。样本按语言紧密分布，其中英语明显独特，覆盖范围很大。 （中）： t-SNE 图，展示 *XLM-RoBERTa* 的结果。部分语言彼此孤立，与 *RoBERTa* 类似，但分布不那么紧密。许多语言，尤其是英语，覆盖的空间大得多，并与多种语言共享空间。 （右）： t-SNE 图，展示 *BEPA*双语模型 的结果，各语言均匀分布且彼此重叠。**BEPA 在不同语言之间创建了一个共享且对齐的空间**。我们微调后的 *BEPA* 单语 模型的结果见图 9。

表 *1*。 XNLI 语言迁移基准上 5 个随机种子的平均准确率（小数表示），按语言及总体展示。结果通过在英语上进行微调，然后对每种语言进行零样本评估得出。 ***BEPA***双语模型 **始终优于（↑）*XLM-RoBERTa*，在所有语言上均是如此**。 更多结果和消融实验请参见表 9 和第 A.9节。

模型	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg
XLM-RoBERTa	0.705	0.753	0.753	0.744	0.831	0.783	0.767	0.687	0.740	0.610	0.714	0.706	0.656	0.735	0.741	0.728
BEPA双语模型	**0.717**	**0.765**	**0.771**	**0.763**	**0.842**	**0.794**	**0.780**	**0.693**	**0.750**	**0.673**	**0.726**	**0.725**	**0.657**	**0.749**	**0.751**	**0.744**
BEPA 单语	0.707	0.755	0.755	0.745	0.829	0.782	0.771	0.681	0.738	0.662	0.714	0.718	0.654	0.735	0.735	0.732

我们运行了 MLQA（多语言问答）基准测试，以检验 BEPA 微调是否能帮助 BERT 风格模型跨语言理解并提取词元级信息。与评估句子级分类的 XNLI 不同，MLQA 要求进行抽取式问答，模型必须在上下文段落中识别特定的答案跨度。如果 BEPA 微调模型能够在其表征空间中实现更好的语义对齐，那么它们应当能够更有效地跨语言迁移词元级推理能力。为运行 MLQA 基准测试，我们在英语 SQuAD v1.1（ Rajpurkar 等人，2016）上微调了两个训练轮次，然后对全部 49 个语言对进行了零样本评估（7×7 G-XLT 矩阵）。我们报告了单语（上下文和问题使用相同语言）和跨语言（上下文和问题使用不同语言）配置下的 F1 分数。结果见图 3。对于相同的单语配置，我们发现*XLM-RoBERTa* 基线和我们的*BEPA*双语模型表现都不错且相近；而我们的*BEPA*双语模型在跨语言配置下，相较于基线较差的表现有显著提升。这表明，BEPA 更稳健的对齐有助于跨越语言边界进行抽取式问答。

对于 GLUE（通用语言理解评估）基准，我们评估 BEPA 微调能否在一系列英语句子级理解任务中保持性能。我们使用标准的单线性层和评估流程，在每项 GLUE 任务上对预训练模型进行三个训练轮次的微调（Wang 等人，2019年）。GLUE 包含分类和回归任务，包括自然语言推理、释义检测、情感分析和语义相似度。我们的模型结果见表 8。

6. 结论

我们发布了用于 BERT 风格模型的新训练框架 BERT-JEPA（BEPA）。该框架利用一种新的训练框架，将经典的 BERT 掩码语言模型（MLM）与 JEPA 相结合。在这一新框架下对我们的 BERT 风格的模型进行微调时，我们发现：（1）**BEPA 微调彻底重组了 [CLS]嵌入空间，使其形成语义优先的结构**（2）**BEPA 微调使主成分分析表征从低秩转变为更高秩**（3）**这种重组几乎不会造成英语表现损失**，以及（4）**提升了多语言任务上的表现**。

参考文献

Agirre, E.、M‘arquez, L. 和 Wicentowski, R.（编）。第四届语义评测国际研讨会论文集（*SemEval-2007*）。计算语言学协会，布拉格，捷克共和国，2007年6月。

Assran, M.、Duval, Q.、Misra, I.、Bojanowski, P.、Vincent, P.、Rabbat, M.、LeCun, Y. 和 Ballas, N.《利用联合嵌入预测架构从图像进行自监督学习》，2023年。URLhttps://arxiv.org/abs/2301.08243。

Assran, M.、Bardes, A.、Fan, D.、Garrido, Q.、Howes, R.、Mojtaba, Komeili、Muckley, M.、Rizvi, A.、Roberts, C.、Sinha, K.、Zholus, A.、Arnaud, S.、Gejji, A.、Martin, A.、Hogan, F. R.、Dugas, D.、Bojanowski, P.、Khalidov, V.、Labatut, P.、Massa, F.、Szafraniec, M.、Krishnakumar, K.、Li, Y.、Ma, X.、Chandar, S.、Meier, F.、LeCun, Y.、Rabbat, M. 和 Ballas, N.《V-jepa 2：自监督视频模型实现理解、预测与规划》，2025年。Balestriero, R. 和 LeCun, Y.《Lejepa：无需启发式方法的可证明且可扩展的自监督学习》，2025年。URLhttps://arxiv.org/abs/2511.08544。Bar Haim, R.、Dagan, I.、Dolan, B.、Ferro, L.、Giampiccolo, D.、Magnini, B. 和 Szpektor, I.《第二届PASCAL文本蕴含识别挑战》。2006年。Bardes, A.、Garrido, Q.、Ponce, J.、Chen, X.、Rabbat, M.、LeCun, Y.、Assran, M. 和 Ballas, N.《重新审视用于从视频中学习视觉表征的特征预测》，2024年。URLhttps://arxiv.org/abs/2404.08471。

Bentivogli, L.、Dagan, I.、Dang, H. T.、Giampiccolo, D. 和 Magnini, B.《第五届PASCAL文本蕴含识别挑战赛》。2009年。Cai, T. T. 和 Ma, R.《用于可视化高维聚类数据的 t-SNE 理论基础》，2022。URLhttps://arxiv.org/abs/2105.07536。Cer, D.、Diab, M.、Agirre, E.、Lopez-Gazpio, I. 和 Specia, L.《SemEval-2017任务1：语义文本相似度多语言与跨语言聚焦评测》。载于 Bethard, S.、Carpuat, M.、Apidianaki, M.、Mohammad, S. M.、Cer, D. 和 Jurgens, D.（编），第*11*届语义评测国际研讨会论文集（*SemEval-2017*），第1–14页，温哥华，加拿大，2017年8月。计算语言学协会。DOI: 10.18653/v1/S17-2001。URLhttps://aclanthology.org/S17-2001。

Conneau, A., Lample, G., Rinott, R., Williams, A., Bow- man, S. R., Schwenk, H., 和 Stoyanov, V. XNLI：评估跨语言句子表示，2018a。URLhttps://arxiv.org/abs/1809.05053。

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S. R., Schwenk, H., 和 Stoyanov, V. XNLI：评估跨语言句子表示。载于 *2018*年自然语言处理实证方法会议论文集。计算语言学协会，2018b。Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., 和 Stoyanov, V. 大规模无监督跨语言表示学习。 *CoRR*，abs/1911.02116，2019年。URL http://arxiv.org/abs/1911.02116。Dagan, I., Glickman, O., 和 Magnini, B. PASCAL文本蕴含识别挑战赛。载于机器学习挑战：评估预测不确定性、视觉对象分类和文本蕴含识别，第177–190页。施普林格，2006年。Devlin, J., Chang, M.-W., Lee, K., 和 Toutanova, K. BERT：面向语言理解的深度双向变换器预训练，2019年。URL https://arxiv.org/abs/1810.04805。Dolan, W. B. 和 Brockett, C. 自动构建句子释义语料库。载于 国际释义研讨会论文集，2005年。Enevoldsen, K., Chung, I., Kerboua, I., Kardos, M., Mathur, A., Stap, D., Gala, J., Siblini, W., Krzemi´nski, D., Winata,G. I., Sturua, S., Utpala, S., Ciancone, M., Schaeffer,M.,Sequeira, G.,Misra,D.,Dhakal,S., Rystrøm,J.,Solo-matin, R., ÖmerÇa˘gatan, Kundu, A., Bernstorff, M.,Xiao, S., Sukhlecha, A.,Pahwa, B., Po´swiata, R., GV, K. K., Ashraf, S., Auras, D., Plüster, B., Harries, J. P., Magne, L., Mohr, I., Hendriksen, M., Zhu, D., Gisserot- Boukhlef, H., Aarsen, T., Kostkan, J., Wojtasik, K., Lee, T., Šuppa, M., Zhang, C., Rocca, R., Hamdy, M., Michail, A., Yang, J., Faysse, M., Vatolin, A., Thakur, N., Dey, M., Vasani, D., Chitale, P., Tedeschi, S., Tai, N., Sne- girev, A., Günther, M., Xia, M., Shi, W., Lù, X. H., Clive, J., Krishnakumar, G., Maksimova, A., Wehrli, S., Tikhonova, M., Panchal, H., Abramov, A., Ostendorff, M., Liu, Z., Clematide, S., Miranda, L. J., Fenogenova, A., Song, G., Safi, R. B., Li, W.-D., Borghini, A., Cas- sano, F., Su, H., Lin, J., Yen, H., Hansen, L., Hooker, S., Xiao, C., Adlakha, V., Weller, O., Reddy, S., 和 Muennighoff, N. MMTEB：大规模多语言文本嵌入基准。*arXiv* 预印本 *arXiv:2502.13595*，2025 年。doi: 10.48550/arXiv.2502.13595。URLhttps://arxiv.org/abs/2502.13595。

Foundation，W. ACL 2019第四届机器翻译会议（WMT19），共享任务：新闻机器翻译。
U
RLhttp://www.statmt.org/wmt19/translation-task.html。

Francisco, G.，Peng-Jen, C.，Myle, O.，Juan, P.，Guillaume, L.，Philipp, K.，Vishrav, C. 和 Marc'Aurelio, R. 两个用于低资源机器翻译的新评估数据集：尼泊尔语-英语和僧伽罗语-英语。2019年。

Garrido, Q.，Balestriero, R.，Najman, L. 和 Lecun, Y. Rankme：根据排名评估预训练自监督表征的下游性能，2023年。网址 https://arxiv.org/abs/2210.02885。

Giampiccolo, D.，Magnini, B.，Dagan, I. 和 Dolan, B. 第三届 PASCAL 文本蕴含识别挑战赛。收录于*ACL-PASCAL*文本蕴含与释义研讨会论文集，第 1–9 页。计算语言学协会，2007年。

Huang, H.、LeCun, Y. 和 Balestriero, R. LLM-JEPA：大语言模型与联合嵌入预测架构相遇，2025年。网址 https://arxiv.org/abs/2509.14252。

LeCun, Y. 通往自主机器智能的道路，0.9.2 版，2022。网址 https://openreview.net/forum?id=BZ5a1r-kVsf。

Levesque, H. J.、Davis, E. 和 Morgenstern, L. Wino-grad schema 挑战赛。载于 *AAAI*春季研讨会：常识推理的逻辑形式化，第 46 卷，第 47 页，2011 年。

Lewis, P.、Oguz, B.、Rinott, R.、Riedel, S. 和 Schwenk, H. MLQA：评估跨语言抽取式问答。*arXiv* 预印本 *arXiv:1910.07475*，文章编号 arXiv：1910.07475，2019年。

Littwin, E.、Saremi, O.、Advani, M.、Thilak, V.、Nakkiran, P.、Huang, C. 和 Susskind, J. JEPA 如何避免噪声特征：深度线性自蒸馏网络的隐式偏置，2024年。网址https://arxiv.org/abs/2407.03475。

Liu, Y.、Ott, M.、Goyal, N.、Du, J.、Joshi, M.、Chen, D.、Levy, O.、Lewis, M.、Zettlemoyer, L. 和 Stoyanov, V. RoBERTa：一种经过稳健优化的BERT预训练方法，2019年。网址 https://arxiv.org/abs/1907.11692。

Mickus, T.、Grönroos, S.-A. 和 Attieh, J. 各向同性、簇和分类器，2024。网址 https://arxiv.org/abs/2402.03191。

Muennighoff, N., Tazi, N., Magne, L. 和 Reimers, N. Mteb：大规模文本嵌入基准。*arXiv* *arXiv*预印本：*2210.07316*，2022。doi：10.48550/ARXIV.2210.07316。网址 https://arxiv.org/abs/2210.07316。Naman, G., Cynthia, G., Vishrav, C., Peng-Jen, C., Guil- laume, W., Da, J., Sanjana, K., Marc'Aurelio, R., Fran- cisco, G., 和 Angela, F. flores-101：面向低资源和多语言机器翻译的评测基准。2021。NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., El- bayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., 和 Wang, J. 不让任何语言掉队：扩展以人为中心的机器翻译。2022。

Rajpurkar, P.、Zhang, J.、Lopyrev, K. 和 Liang, P. SQuAD：100,000+ 个用于文本机器阅读理解的问题。载于*EMNLP*论文集，第2383–2392页。计算语言学协会，2016年。

Reimers, N. 和 Gurevych, I. Sentence-BERT：使用孪生BERT网络的句子嵌入，2019年。网址 https://arxiv.org/abs/1908.10084。

Ribeiro, M. T.、Singh, S. 和 Guestrin, C. 用于调试自然语言处理模型的语义等价对抗规则。载于 Gurevych, I. 和 Miyao, Y.（编），计算语言学协会第*56*届年会论文集（第*1*卷：长篇论文），第856–865页，澳大利亚墨尔本，2018年7月。计算语言学协会。DOI：10.18653/v1/P18-1079。网址 https://aclanthology.org/P18-1079/。

Rogendo. Eng-Swa-Dataset：英语–斯瓦希里语句对数据集。https://github.com/Rogendo/Eng-Swa-Dataset，2023年。访问日期：2025年12月13日。

Socher、R.、Perelygin、A.、Wu、J.、Chuang、J.、Manning、C. D.、Ng、A. 和 Potts、C. 基于情感树库的语义组合性递归深度模型。载于*2013*年自然语言处理经验方法会议论文集，第 1631–1642 页，美国华盛顿州西雅图，2013年10月a。计算语言学协会。网址：
h
ttps://www.aclweb.org/anthology/D13-1170。

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. 和 Potts, C. 基于情感树库的语义组合性的递归深度模型。载于*EMNLP* 会议论文集，第1631–1642页，2013b年。

Tiedemann, J. OPUS 中的并行数据、工具和接口。载于 Calzolari, N., Choukri, K., Declerck, T., Do˘gan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. 和 Piperidis, S.（编），第八届国际语言资源与评估会议（*LREC'12*）论文集，第2214–2218页，土耳其伊斯坦布尔，2012年5月。欧洲语言资源协会（ELRA）。网址：http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf。

Tiedemann, J. Tatoeba 翻译挑战：面向低资源语言和多语言机器翻译的真实数据集，2020年。网址 https://arxiv.org/abs/2010.06354。

van den Oord, A.、Li, Y. 和 Vinyals, O. 使用对比预测编码进行表征学习，2019年。网址 https://arxiv.org/abs/1807.03748。

Vaswani, A.、Shazeer, N.、Parmar, N.、Uszkoreit, J.、Jones, L.、Gomez, A.N.、Kaiser, L. 和 Polosukhin, I. 注意力就是你所需要的一切，2023年。网址 https://arxiv.org/abs/1706.03762。

Wang, A.、Singh, A.、Michael, J.、Hill, F.、Levy, O. 和 Bowman, S. R. GLUE：自然语言理解的多任务基准测试与分析平台。2019年。发表于 ICLR 会议论文集。

Warstadt, A., Singh, A., 和 Bowman, S. R. 神经网络可接受性判断。*arXiv*预印本*1805.12471*，2018年。

Williams, A., Nangia, N., 和 Bowman, S. R. 一种用于通过推理进行句子理解的广覆盖挑战语料库。载于 *NAACL-HLT*论文集，2018年。

Zhang, B., Williams, P., Titov, I., 和 Sennrich, R. 改进超大规模多语言神经机器翻译和零样本翻译。载于 Jurafsky, D., Chai, J., Schluter, N., 和 Tetreault, J.（编），计算语言学协会第*58*届年会论文集，第1628–1639页，线上，2020年7月。计算语言学协会。DOI：10.18653/v1/2020.acl-main.148。网址：https://aclanthology.org/2020.acl-main.148。

A. 附录

本节将分享与我们的实验和方法相关的更多细节和信息。附录的编排如下：A.1节介绍作者贡献，A.2节介绍生成式人工智能的使用，A.3节介绍所用资源，A.4节介绍我们的BEPA代码库和代码，A.5节介绍完整的BEPA框架，A.6节介绍有关坍塌的BERT[CLS] 嵌入空间的信息，A.7节介绍训练过程相关图表，A.8节展示我们新的BERT嵌入空间的结果，A.9节介绍详尽的结果，A.10节介绍未来工作。

A.1. 贡献

本节概述每位作者的具体贡献。作者的排列顺序是随机决定的，方法是用等概率转动幸运轮盘。

Marcel 的 主要贡献包括协助提出 BEPA 框架及其理论基础、主导论文撰写、在基准测试上进行评估，以及制作图表。

Adam 的 主要贡献包括协助提出 BEPA 框架及其理论基础、主导评估工作、制作图表，以及撰写论文。

Taj 的 主要论文贡献包括协助提出 BEPA 框架及其理论基础，主导编码和模型训练、评估工作，并撰写论文。

Kenneth 的 主要论文贡献包括协助提出 BEPA 框架及其理论基础、进行评估，并撰写论文。

A.2. 生成式人工智能的使用

在本文的撰写过程中，我们在一些任务中使用了生成式人工智能。我们将其用作检查写作语法并在必要时进行纠正的工具。此外，我们还利用它简化代码实现工作，协助调试并实现 BEPA 框架。

A.3. 使用的资源

本节将介绍本文中使用的各类资源。我们首先介绍用于微调的数据集，随后介绍用于对模型进行基准测试的数据集。

表 2。消融实验中训练不同模型所用数据集的信息。语言对的数量表示存在翻译样本的语言组合数量。

Name	语言对数量
Opus-100（Tiedemann，2012；Zhang 等人，2020）	114
WMT19（Foundation）	9
Flores（NLLB Team 等人，2022；Naman 等人，2021；Francisco 等人，2019年）	200
Eng-Swa-Dataset（Rogendo，2023年）	2

我们发现，Opus-100中没有任何英语-斯瓦希里语语对，因此我们使用Eng-Swa-Dataset作为一种引导方法，使我们的模型能够学习这两种语言之间的关系。

我们希望确保在各种下游任务上对BEPA进行评估。因此，我们选择了以下基准测试，将微调后的模型进行比较：

1. XNLI（Conneau等人，2018年b）
2. MLQA（Lewis等人，2019年）
3. GLUE（Agirre 等，2007；Bar Haim 等，2006；Bentivogli 等，2009；Dagan 等，2006；Dolan 和 Brockett，2005；Giampiccolo 等，2007；Levesque 等，2011；Rajpurkar 等，2016；Socher 等，2013b；Wang 等，2019；Warstadt 等，2018；Williams 等，2018）

表 3. 用于评估我们模型的基准测试。这些是标准的 BERT评估基准。 **它们涵盖了多种不同的任务，包括推理、分析、跨语言分类和问答。**

Name	任务
XNLI	跨语言句子分类
MLQA	跨语言问答（QA）
GLUE	语法正确性、情感分析、语义相似度、语义推理

受图形处理器资源限制，本文未进行完整的模型预训练。不过，我们认为，未来工作中，使用我们新的训练框架对模型进行完整预训练至关重要。微调时，我们使用 1 个 NVIDIA RTX A5500 图形处理器训练了 10 个训练轮次。为避免图形处理器内存耗尽，我们采用了 16 的批量大小，并将样本最大长度设为 256 个词元。我们使用 $2x10^{-5}$ 的学习率配置 AdamW优化器，权重衰减为 0.01，预热步数为 500。我们将对齐损失的 λ 设为 1，并在掩码语言建模任务中采用业界标准的 15% 掩码概率。

A.4. BEPA 代码库

我们发布了代码库的匿名版本，其中包含用于训练 BEPA 的代码，包括训练、环境搭建、配置和训练脚本。可通过以下网址访问：https://anonymous.4open.science/r/bert-jepa-translation-3EB8/README.md

A.5. 完整的 BEPA 框架

我们提出的新型 BERT 训练方法能够转换 BERT 模型当前的嵌入空间。该方法由两个主要组件构成：

1. 经典的掩码语言模型训练
2. CLS 标记 JEPA 对齐

我们首先选用一个预训练的 RoBERTa 基础模型（xlm-roberta-base）（Conneau 等人，2019年）。由于我们希望教会 BERT 表示独立于语言系统的信息，因此选择将两个句子并排嵌入，并将它们打包成一个 Transformer 输入。我们使用 XLM-RoBERTa 分词器，并在样本中添加了特殊标记，例如 [CLS], [SEP], 和 [PAD]。输入示例见图 4。

这样一来，打包方式有两种可能：将同一语言中的不同样本打包在一起，或将两种不同语言的样本打包在一起。我们据此设计了两种独立的训练模式：

1. 单语：打包在一起的样本全部来自同一种语言 e.
2. 双语：其中 打包在一起的样本可以来自任意语言对（只要它们含义相同).

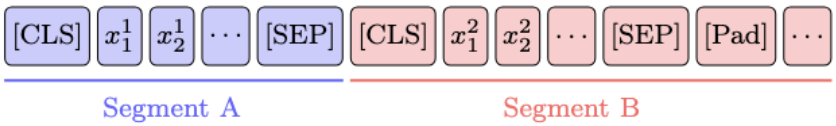


图 4. BEPA 的输入打包策略。两个序列通过两个 [CLS] 词元和一个 [SEP] 边界连接起来。 x_1^i 表示与第一个序列对应的词元， x_2^i 表示与第二个序列对应的词元。片段 A 标记第一个句子，片段 B 标记第二个句子。**整个序列会进行联合编码**。

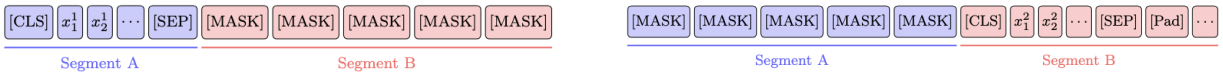


图 *5.* BEPA 框架中 CLS 标记 JEPA 对齐阶段的掩码策略。（左）：屏蔽第二个句子对应的词元，使模型能够构建第一个序列的 [CLS] 标记表示。（右）：**屏蔽第一个句子对应的词元，使模型能够构建 [CLS] 第二个序列的标记表示。**

接下来，这一设置会执行 BERT 论文（Devlin 等人，2019年）中经典的 BERT 掩码语言建模任务。这样做是为了保留 BERT 风格的 Transformer 强大的语言能力。这一步包括随机掩盖一定比例的词元，掩盖比例由一个超参数决定；该参数沿用 BERT 的常见设置，即 15%。然后预测被掩盖的词元，并计算掩码语言建模损失。

接下来，我们介绍 BEPA 框架中的创新部分。由于 BERT 风格的模型将最后一层的 [CLS] 标记视为序列的潜在表示，我们提取 [CLS] 标记并计算对齐损失，以促进跨语言不变性。为此，我们针对每个句子分别进行两次前向传播，每次都屏蔽其中一个句子。掩码策略见图 5。随后提取每个句子的 [CLS] 标记的潜在表示，并通过 $L_{Alignment}$进行对齐。我们使用恒等预测器；不过，未来工作应测试不同预测器的影响。图 6展示了 BEPA 中 CLS 标记对齐阶段的完整示意图。

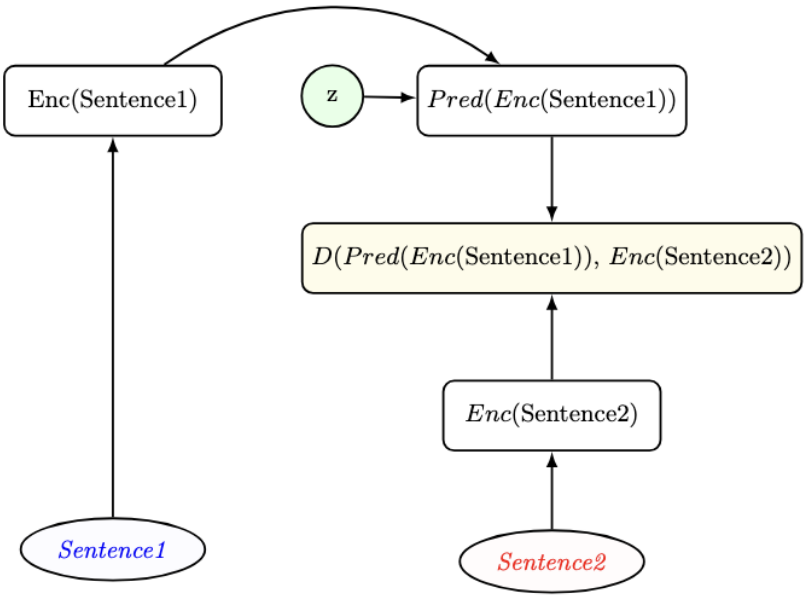


图 *6.* BEPA 框架中 JEPA 部分的架构。具体而言，这是通过对 BERT 风格模型进行两次前向传播来实现的，每次都遮蔽另一个句子。编码后的输出是对应于每个句子的 [CLS] 标记的嵌入，预测器则是单位矩阵，因此潜在表示保持不变。最后一步计算两个 [CLS] 潜在表示之间的距离，并将其最小化。

T这两个步骤，即 BERT 风格 MLM 和 CLS 标记 JEPA 对齐，产生了一种新的训练损失 f:

$$L_{BEPA} = L_{MLM} + \lambda L_{Alignment}$$

其中，λ 是一个超参数，用于确定赋予 $L_{Alignment}$的权重；$L_{Alignment}$是用于 JEPA 部分的损失函数，旨在对齐两个语义相同样本之间的 [CLS] 嵌入。由于 BERT 风格模型的 [CLS] 嵌入空间本身就存在坍塌，我们决定采用一种有助于模型区分负样本对的损失函数

同时使相似样本对的表征更加接近。InfoNCE（vandenOord 等人，2019年）和 SigREG（Balestriero 和 LeCun，2025年）在满足这些要求方面效果最佳，但该框架具有充分的扩展性，可替换为用户所需的任意损失函数。此外，实验使用了恒等函数预测器；不过，我们认为未来工作中有必要测试不同预测器对 BEPA 的影响。

A.6. 一个坍塌的 CLS 标记

本节将探讨 [CLS] 词元在 BERT 模型中如何发生坍塌，导致潜在空间丢失大量有价值的信息。如果大多数下游任务都使用 [CLS] 词元来获取 BERT 嵌入，那么 [CLS] 词元生成的潜在空间就必须具有丰富的信息，而不能发生坍塌。BERT 风格的模型生成的嵌入中，不同点对之间的余弦相似度和欧氏距离充分体现了这种坍塌。这会导致所有句子，即使是不相似的句子对，也具有较高的余弦相似度。这一点可见表 4 和 5。

表 *4*。 对于 *RoBERTa* 在 WMT19 数据集上的跨语言和语言内句子的嵌入之间的余弦相似度。 **所有句子对之间的余弦相似度都很高，即使是不相关的句子对和跨语言句对也是如此。这表明 [CLS] 标记嵌入空间发生了坍塌，因为它们都高度相似。**

语言对	同一语言（无关）	不同语言（相关）	不同语言（不相关）
阿拉伯语–英语	0.9968	0.9918	0.9935
保加利亚语–英语	0.9965	0.9941	0.9937
德语–英语	0.9959	0.9965	0.9926
希腊语–英语	0.9965	0.9923	0.9936
英语–西班牙语	0.9961	0.9973	0.9928
英语–法语	0.9957	0.9964	0.9919
英语–印地语	0.9961	0.9908	0.9909
英语–俄语	0.9965	0.9929	0.9933
英语–泰语	0.9976	0.9904	0.9933
英语–土耳其语	0.9961	0.9952	0.9937
英语–乌尔都语	0.9968	0.9887	0.9917
英语–越南语	0.9963	0.9939	0.9931
英语–中文	0.9959	0.9919	0.9932

表*5.* XLM-RoBERTa 在不同语言之间及语言内部的句子嵌入余弦相似度，*XLM-RoBERTa* 用于 WMT19 数据集。 **所有句子对之间的余弦相似度都很高，包括不相关句对和跨语言句对。这表明 CLS 词元嵌入空间发生了坍塌，因为它们都非常相似。**

语言对	同一语言（不相关）	不同语言（相关）	不同语言（不相关）
阿拉伯语–英语	0.9923	0.9926	0.9979
保加利亚语–英语	0.9957	0.9945	0.9845
德语–英语	0.9923	0.9968	0.9923
希腊语–英语	0.9929	0.9959	0.9925
英语–西班牙语	0.9918	0.9970	0.9939
英语–法语	0.9877	0.9949	0.9907
英语–印地语	0.9970	0.9981	0.9903
英语–俄语	0.9910	0.9975	0.9916
英语–泰语	0.9975	0.9956	0.9927
英语–土耳其语	0.9924	0.9962	0.9891
英语–乌尔都语	0.9984	0.9977	0.9928
英语–越南语	0.9939	0.9961	0.9918
英语–中文	0.9967	0.9981	0.9900

A.7. 训练过程

本节将考察训练过程，重点关注各训练轮次中的损失函数（图 7）和 RankMe（Garrido 等人，2023年）（图 8）数值。

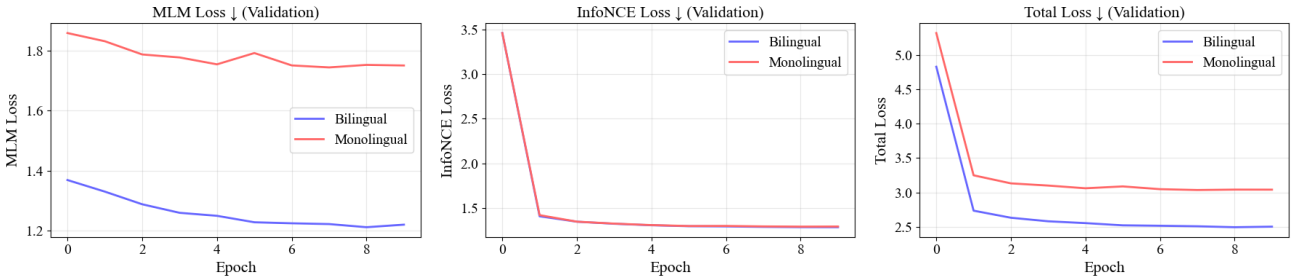


图 7。训练过程中各训练轮次的验证损失。（左）掩码语言建模验证损失。双语模型的损失始终较低，因为它获得了两种语言的句子，可能可以从中找到被掩码词元的直接译文。（中）InfoNCE 验证损失。该损失按每个批次计算，将译文作为正样本，所有其他句子作为负样本。它不依赖于掩码语言建模任务。（右）每个训练轮次的总验证损失。

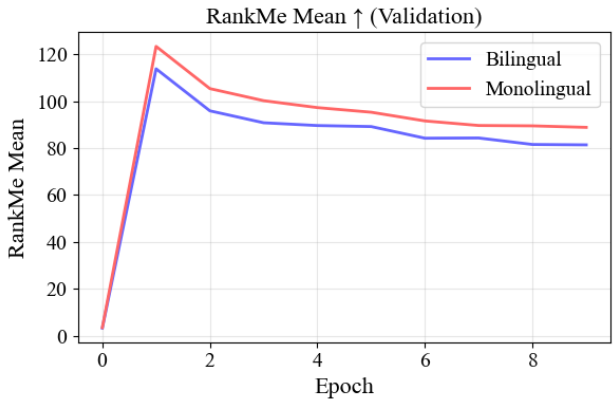


图 8。RankMe（Garrido 等人，2023年）在双语任务中两种语言的 [CLS] 词元的平均值，或在单语任务中单独的 [CLS] 词元的平均值；这些数值是在验证集上按训练轮次计算的。RankMe 给出了嵌入的有效秩。模型起初的嵌入空间发生坍塌，短暂飙升后趋于稳定。与双语训练相比，单语训练期间的 RankMe 更高。

A.8. 更丰富的嵌入空间

本节介绍 BEPA 微调后更丰富的 [CLS] 嵌入空间所取得的结果。表格6 和 7 展示了相关句子与不相关句子之间的余弦相似度，相较于第 [CLS]节中坍塌的词元A.6，有了显著提升。

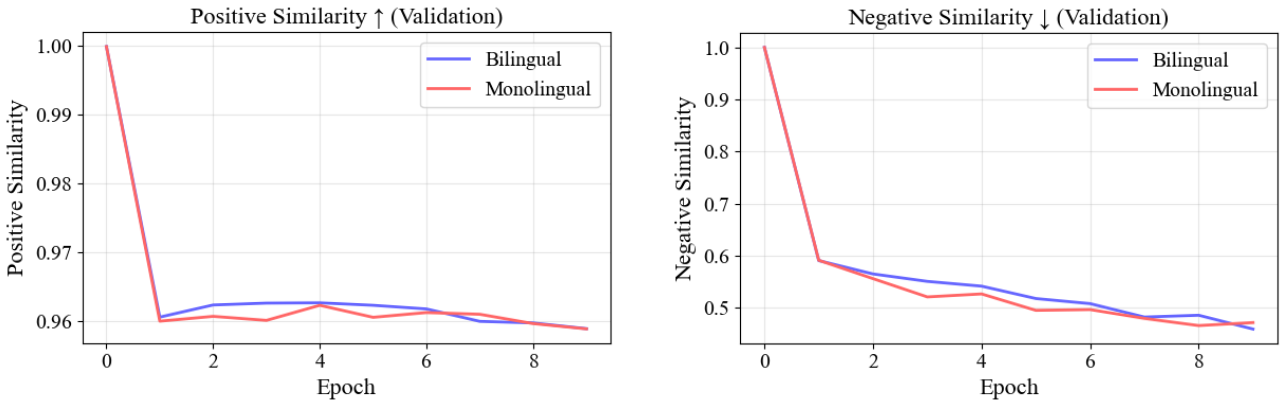


图*10*。 正样本和负样本的平均余弦相似度随训练轮次的变化。 *(左)*：BEPA 微调后，验证集中相关（语义相似）样本的平均余弦相似度仍然很高。 *(右)*：BEPA 微调后，验证集中不相关（语义不相似）样本的平均余弦相似度下降（↓）。 **总体而言，BEPA 微调会重新组织 [CLS] 嵌入空间，并抑制坍塌。相似的样本对彼此靠近，不相似的样本对则彼此远离。** 更多示例请参阅表 6 和 7。

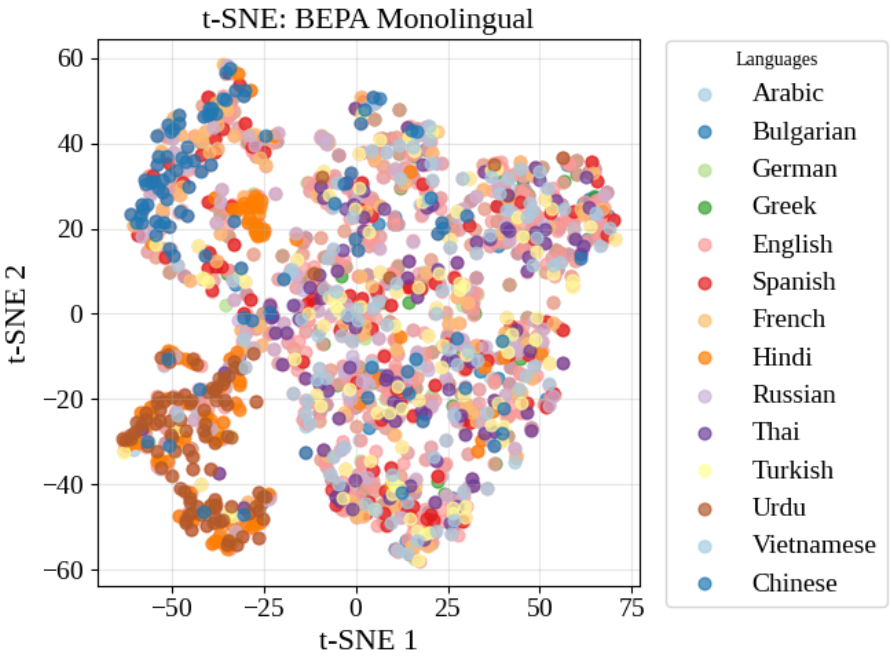


图*9*。 BEPA 单语的 t-SNE 图，显示 CLS 词元嵌入在整个空间中均匀分布。 **BEPA 在不同语言之间创建了一个共享且对齐的空间。**

表*6* WMT19 数据集中*BEPA*-单语 跨语言和同语言句子嵌入的余弦相似度。相似句子对之间的余弦相似度较高（↑），不相关句子对之间则较低。展示的是未坍塌的 [CLS] 词元嵌入空间。**BEPA 训练会降低(↓)不相关句子对之间的余弦相似度，从而形成更丰富的嵌入空间**。

语言对	同一语言（不相关）	不同语言（相关）	不同语言（不相关）
阿拉伯语–英语	0.5683	0.9719	0.5083
保加利亚语–英语	0.5986	0.9641	0.4804
德语–英语	0.4940	0.9739	0.4445
希腊语–英语	0.4709	0.9751	0.4687
英语–西班牙语	0.4877	0.9885	0.4506
英语–法语	0.3797	0.9815	0.4163
英语–印地语	0.5388	0.9875	0.4451
英语–俄语	0.4113	0.9716	0.4693
英语–泰语	0.6198	0.9568	0.4954
英语–土耳其语	0.5656	0.9751	0.3966
英语–乌尔都语	0.6758	0.9878	0.3952
英语–越南语	0.5783	0.9695	0.5595
英语–中文	0.5087	0.9824	0.3713

表 *7.* WMT19 数据集中，*BEPA-Bilingual* 不同语言之间和语言内部句子的嵌入之间的余弦相似度。相似句子对之间的余弦相似度较高（↑），而不相关句子对之间的余弦相似度较低。展示的是未发生坍塌的 [CLS] 词元嵌入空间。**BEPA 训练会降低（↓）不相关句子对之间的余弦相似度，从而形成更丰富的嵌入空间**。

语言对	同一语言（不相关）	不同语言（相关）	不同语言（不相关）
阿拉伯语–英语	0.5351	0.9719	0.5035
保加利亚语–英语	0.5786	0.9665	0.5207
德语–英语	0.4706	0.9757	0.4718
希腊语–英语	0.4911	0.9755	0.4336
英语–西班牙语	0.4640	0.9897	0.4158
英语–法语	0.4600	0.9829	0.4147
英语–印地语	0.5080	0.9862	0.4290
英语–俄语	0.4522	0.9712	0.3601
英语–泰语	0.6369	0.9553	0.5377
英语–土耳其语	0.5629	0.9759	0.4996
英语–乌尔都语	0.6757	0.9876	0.3838
英语–越南语	0.6027	0.9695	0.5331
英语–中文	0.4345	0.9825	0.3117

A.9. 消融基准测试结果

本节补充了我们开展的不同消融实验结果，包括以下基准测试：GLUE（表 8）、XNLI（表 9）和 MLQA（表10）。我们研究了数据质量和数据集规模对跨语言迁移性能的影响。我们采用三种不同的训练配置对 XLM-RoBERTa-Base 进行了微调：（1）Opus 模型使用了 Opus-100 数据集和斯瓦希里语-英语翻译数据集中的 10,000 对翻译对（Tiedemann，2012；2020；Rogendo，2023）；（2）Flores 模型使用了高质量 Flores 数据集中的约 2,000 对数据，并增加了 5× 训练轮次，以保持梯度更新步数相当（NLLB Team 等，2022；Naman 等，2021；Francisco 等，2019）；（3）小型模型使用 Opus-100 中的 2,000 个示例进行训练，并将训练轮次增加 5 倍，以保持梯度更新步数相当。所有模型均在 XNLI 基准测试上进行了评估，以考察其零样本跨语言迁移能力（Tiedemann，2012；2020；Rogendo，2023）。

表 *8.* GLUE 基准测试上 5 个随机种子的平均结果。对于所有分类任务，我们报告准确率；对于 STS-B，我们报告皮尔逊相关系数。带有 † 的结果引自 XLM-R 论文（ Conneau 等人，2019 年）。我们还报告了在自有硬件和训练/测试配置下，使用 XLM-RoBERTa 在 GLUE 上取得的结果。 **BEPA 微调对英语任务的性能几乎没有影响，甚至没有影响。**

模型	MNLI-m/mm	QNLI	QQP	SST	MRPC	STS-B	Avg
XLM-R†	88.9 / 89.0	93.8	92.3	95.0	89.5	91.2	91.8
XLM-RoBERTa（我们的模型）	84.64 / 84.59	90.38	90.69	92.55	86.32	87.36	88.07
单语 JEPA	84.31 / 84.47	90.57	90.60	92.87	86.47	87.56	88.12
双语 JEPA	84.97 / 85.04	90.82	90.83	92.20	88.24	88.39	88.64

表 *9*。 XNLI 语言迁移基准上 5 个随机种子的平均准确率（小数），按语言及总体结果展示。列出了我们所有模型、测试过的消融实验以及 *XLM-RoBERTa* 基础版的结果。结果通过在英语上微调后，对每种语言进行零样本评估得出。 ***BEPA***双语模型 **在所有语言上的表现始终优于（↑）*XLM-RoBERTa*** 。 更多结果和消融实验请参见表 1 和第 A.9 节。

模型	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg
XLM-RoBERTa 基础版	0.705	0.753	0.753	0.744	0.831	0.783	0.767	0.687	0.740	0.610	0.714	0.706	0.656	0.735	0.741	0.728
双语 - Opus	0.717	0.765	0.771	0.763	0.842	0.794	0.780	0.693	0.750	0.673	0.726	0.725	0.657	0.749	0.751	0.744
双语 - Flores	0.718	0.764	0.764	0.757	0.841	0.788	0.778	0.693	0.746	0.656	0.718	0.719	0.656	0.742	0.745	0.739
双语 - Small	0.710	0.760	0.769	0.757	0.839	0.792	0.776	0.691	0.750	0.664	0.719	0.729	0.658	0.746	0.747	0.741
单语 - OPUS	0.707	0.755	0.755	0.745	0.829	0.782	0.771	0.681	0.738	0.662	0.714	0.718	0.654	0.735	0.735	0.732
单语 - Flores	0.701	0.751	0.757	0.743	0.829	0.782	0.768	0.687	0.740	0.652	0.710	0.709	0.655	0.735	0.736	0.730
单语 - 小型	0.700	0.750	0.751	0.741	0.830	0.779	0.768	0.685	0.733	0.660	0.718	0.721	0.653	0.737	0.731	0.730

表10. MLQA 问答结果。我们报告在零样本分类中取得的 F1/EM 分数：模型在英语 SQuAD 上微调，并在 7种语言上进行评估。带有 †的结果来自 MLQA论文（Lewis等人，2019年）。

模型	en	es	de	ar	hi	vi	zh	Avg
BERT-Large†	80.2/67.4	-	-	-	-	-	-	-
mBERT†	77.7/65.2	64.3/46.6	57.9/44.3	45.7/29.8	43.8/29.7	57.1/38.6	57.5/37.3	57.7/41.6
XLM-15†	74.9/62.4	68.0/49.8	62.2/47.6	54.8/36.3	48.8/27.3	61.4/41.8	61.1/39.6	61.6/43.5
$XLM\text{-}R_{Base}$	77.1/64.6	67.4/49.6	60.9/46.7	54.9/36.6	59.4/42.9	64.5/44.7	61.8/39.3	63.7/46.3
XLM-R	80.6/67.8	74.1/56.0	68.5/53.6	63.1/43.5	69.2/51.6	71.3/50.9	68.0/45.4	70.7/52.7
XLM-RoBERTa（我们的方法）	80.9/68.2	66.8/45.8	62.4/46.8	55.0/35.9	61.7/44.0	67.4/47.0	40.3/39.6	62.1/46.8
BEPA 单语（我们的方法）	81.0/68.0	67.8/46.1	62.4/46.4	55.5/36.6	64.0/46.0	68.6/47.6	41.3/40.6	62.9/47.3
BEPA 单语（我们的）	81.2/68.2	67.8/46.4	63.3/47.2	55.9/37.0	63.2/45.4	68.5/47.8	42.3/41.6	63.2/47.7

A.10. 未来工作

BEPA 的探索才刚刚开始。还有几项实验有待完成。在此，我们重点介绍感兴趣的未来工作和实验：（1）我们希望了解在微调过程中采用不同对齐损失（例如 SigREG）所产生的影响

。（2）测试对齐损失的不同权重（λ）；（3）将我们的框架扩展到自回归模型，因为自回归机制能够隐式地防止坍塌；（4）尝试为 JEPA 预测器采用不同的预测器，例如预测网络，并探索潜变量的用法。