通过大语言模型训练建模双语学习：关键期探索

Uriel Benymon, [1] Tristan Whipps, [1] 和
Armita Ghobadi[21]哥伦比亚大学通识教育学院[2]哥伦
比亚大学巴纳德学院

通信地址：

Uriel Benymon 哥伦比亚大学通识教育学
院计算机科学系udb2101@columbia.edu

利益冲突：作者声明不存在利益冲突。**数据可用性：**支持本研究结论的所有数据均

摘要

关于第二语言（L2）习得的关键期理论认为，学习新语言的能力可能取决于此前习得的第一语言（L1）知识能否轻易地被修改。本探索性计算研究将神经语言模型方法用于研究关键期效应，并考察内部表征，而非行为。研究使用一个小型双层 GPT-2 模型，先以英语训练，再通过顺序训练或弹性权重巩固（EWC）训练德语（相对近缘的第二语言）或芬兰语（相对远缘的第二语言）；EWC 会限制对英语学习中重要参数的改动。经过一个和两个第二语言训练轮次后，研究人员分别在嵌入层和两个 Transformer 层计算每种语言 20 个留出探测句的平均池化隐藏状态，并与纯英语模型的结果进行余弦相似度比较。英语相似度在 Transformer 层中始终较高，但在嵌入层中变为负值。第二语言相似度在 Transformer 层中有所提高，而且德语与芬兰语在第二个 Transformer 层中的差异大于在第一个 Transformer 层中的差异。EWC 将费舍尔加权参数漂移降低了约 6%，但使 Transformer 层的相似度变化不到 .001。由于每种条件仅进行了一次训练运行，这些发现仍属初步结果。

关键词：双语现象、关键期、第二语言习得、神经语言模型、弹性权重巩固、表征相似性

引言

双语发展涉及第一语言（第一语言）和第二语言（第二语言）知识、接触与使用的持续变化，由此引发了这样的问题：神经语言模型的内部结构是否也表征了类似的变化（Constantinescu 等，2025；Grosjean，2013）。Grosjean（2013）并未将双语者视为两个单语者的结合体，而是将双语语言知识描述为一个不断变化的系统，其中各种语言依据其使用历史相互关联地发展。因此，语言知识随时间的增强与减弱表明，双语发展是一个组织结构不断变动的过程，而非一个固定的终点。

在这一不断变化的系统中，关键期被定义为：进入青春期时，个体失去通过沉浸式学习习得第二语言的能力，并且在遭受剥夺的情况下，倾向于不会遗忘第一语言。这一概念引出了另一个问题，即语言接触的时间安排和日益下降的可塑性如何影响第一语言保持和二语习得（Constantinescu 等，2025）。Constantinescu 等（2025）通过神经语言模型研究了这些效应。神经语言模型是一种从语言数据中学习统计模式的计算系统；他们先用第一语言训练模型，再在不同的学习条件下引入第二语言。其研究将模型更新参数能力的变化视为神经可塑性变化的计算类比，从而能够在受控条件下研究那些在人类发展过程中一直难以单独考察的学习条件。因此，他们的框架为通过从一语到二语的训练研究关键期效应提供了路径，同时保留了对语言学习时间安排和灵活性的控制。

限制这种灵活性的一种方法是弹性权重巩固（EWC），它会减少对先前习得信息至关重要的参数变动

（Kirkpatrick 等，2017）。Kirkpatrick 等（2017）提出了 EWC，作为限制灾难性遗忘的一种方法；灾难性遗忘是指神经网络在学习新任务后丢失先前知识的现象。随后，Constantinescu 等（2025）将 EWC 应用于语言习得，将其作为后期学习中可塑性降低的一种计算类比。这一关联使从L1到L2的序贯训练有助于考察的不仅是是否习得了第二语言，还包括在新的语言信息进入网络时，先前的第一语言知识有多少仍保持稳定。

不过，仅凭行为表现还不足以全面描述这一过程中模型内部发生了哪些变化。内部表征，即语言输入经过网络时产生的神经激活模式，为考察语言学习提供了另一个层面。Liu 等人（2019年）发现，神经语言模型在不同内部层中编码了不同形式的语言信息，这表明表征在网络各处并非始终一致。因此，在第二语言训练过程中追踪多个层，可以探究与语言相关的变化是否均匀发生，还是在模型不同深度呈现出不同的表征轨迹。

除了定位这些表征之外，既有研究还确立了用于比较不同模型或学习阶段之间神经表征差异的方法。Kornblith 等人（2019）评估了多种表征相似性的定量方法，其中包括中心核对齐（CKA），并指出，如果不加审慎地应用，更简单的相似性度量可能会产生误导。本研究采用余弦相似度而非 CKA，因为相较于第二语言训练前基线，余弦相似度能够直接捕捉单个句子表征的方向，这更适合纵向的逐句比较。这一选择也带来了代价，正如 Kornblith 等人（2019）的批评所明确指出的：余弦相似度无法区分真正的

语言重组与表征空间几何结构的整体变化；讨论部分将直接探讨这一局限。

语言学习也取决于相关语言之间的关系，因此，选择第一语言和第二语言条件对于解读表征变化至关重要（Grosjean，2013）。Littell 等（2017）开发了 URIEL 和 lang2vec[1]，通过类型学、地理和系统发生学信息等特征来表征语言，从而能够以计算方式表示跨语言关系。在这一框架中，类型学信息为语言比较提供了有原则的依据，而非仅凭直觉将语言对标记为相关或不相关。因此，这类语言特征在 Grosjean（2013）关于双语语言系统相互作用的论述，与衡量不同第二语言条件如何影响一个经第一语言训练的共享网络之间，搭建起了桥梁。

尽管取得了这些进展，既有研究仍未充分厘清一个重要交叉问题：顺序第二语言习得过程中内部表征如何变化，这些变化是否因层而异，以及语言亲缘关系是否对应着不同的变化轨迹。Constantinescu 等（2025）提出了关键期训练框架，而 Liu 等人（2019）和 Kornblith 等人（2019）则表明，内部表征包含可测量、可供比较的结构。同样，Littell 等（2017）提供了一种依据语言特征来确立跨语言关系的方法。综合来看，这些研究思路促使研究从探讨神经语言模型是否保留或习得某种语言，进一步转向考察其内部表征在学习过程中于何处以及如何发生变化。

本研究通过一个英语第一语言模型探讨这一空白：该模型分别在顺序学习和 EWC约束训练条件下学习德语和芬兰语作为第二语言

[1]一个数据库，提供从类型学、地理学和系统发育数据库中提取的、信息丰富的语言向量识别信息，主要呈现语言的二元事实（Littell等，2017）。

。选择德语和芬兰语是为了与 Constantinescu 等（2025）的研究保持一致，并遵循相同的标准：德语与英语同属一个语系，芬兰语与二者均无亲缘关系，且三种语言都属于资源丰富的语言。研究在嵌入层和两个 Transformer 层上测量内部变化，时间点包括早期第二语言检查点和最终训练阶段；同时，通过与仅英语（第二语言训练前）模型的余弦相似度，追踪英语探测句表征的变化（本文称为第一语言表征保留），以及德语或芬兰语探测句表征的变化；这两项指标都不是语言熟练度的直接测量。如表1所示，德语和芬兰语构成了不同的二语条件，可据此考察表征轨迹的差异，并结合跨语言亲缘关系进行分析（Littell等，2017）。通过这种方式，本实验在一项受控的计算设计中，将双语发展、受限可塑性、语言亲缘关系和内部神经变化联系起来。

本研究探讨三个研究问题。首先，经过英语训练的神经语言模型在习得德语或芬兰语第二语言的早期阶段与最终阶段之间，其内部表征会如何变化？其次，在嵌入层和 Transformer 层中，德语与芬兰语条件下的表征变化有何不同？最后，与不受约束的顺序训练相比，EWC 约束的第二语言训练在多大程度上能以不同方式保留英语第一语言表征？我们假设，第二语言训练期间内部表征会发生变化，但变化的幅度和方向会因表征层和第二语言条件而异。更具体地说，德语和芬兰语会产生不同的表征相似性变化轨迹，这与以下预期一致：跨语言差异会影响新语言信息在模型内部的组织方式（Littell等，2017）。最后，EWC 约束的模型将比顺序训练的模型保留更高的相似性，

相较于顺序训练的模型，其表征将更接近英语第一语言表征，因为 EWC 会限制对早期学习中重要参数的改动（Constantinescu 等，2025；Kirkpatrick 等，2017）。这些预测具有明确的方向性且可证伪：如果不同语言、层或训练条件下的变化轨迹相同，就会构成反对相应假设的证据。第 3.3 节报告的 EWC 参数层面分析是在提出这些假设之后进行的，并作为探索性分析呈现。

方法

2.1 研究设计与实验概述

本研究采用计算实验，考察语言模型的内部表征在二语习得过程中如何变化，以及这些变化是否因语言和训练条件而异。研究设计扩展了 Constantinescu 等（2025）的神经语言模型框架。该框架通过操控模型接触另一种语言的时间，以及模型在后续训练中参数变化的自由程度，来模拟关键期效应。本研究未改变引入第二语言的时间，而是固定从一语到二语的学习进度安排，并在此框架内考察内部表征变化，因此本研究本身并不检验关键期边界。英语作为第一语言，德语和芬兰语作为两种第二语言条件，旨在提供与英语不同的关系。因此，实验比较了普通顺序训练与通过弹性权重巩固（EWC）加以约束的训练中，英语表征保持和二语表征的变化。

2.2 模型与训练框架

本实验采用 GPT-2 语言模型架构，以及改编自 Constantinescu 等（2025）的训练框架。模型先学习英语，随后继续

进行第二语言训练阶段，使用德语或芬兰语数据，因此仅英语检查点可作为第二语言训练前基线。研究在第1轮和第 2 轮考察表征，以捕捉习得轨迹中的早期和后期阶段。这样的纵向比较提供了一种可控方法，用于考察随着第二语言接触的持续，表征是否保持稳定或发生变化。模型采用小型 GPT-2 架构，包含两层 Transformer 层、两个注意力头和 128 维隐藏状态（共 6,960,768 个参数，输入和输出嵌入权重共享），并搭配自定义的字节级 BPE 分词器，词元数量为 8,016；由于模型配置保留了标准 GPT-2 词汇表大小 50,257，分词器词汇表之外的嵌入行从未被使用。英语第一语言模型从头开始训练，四个第二语言训练运行（德语或芬兰语，顺序学习或弹性权重巩固）均从同一个已保存的英语检查点初始化，相关信息记录在各第二语言检查点的配置元数据中。所有训练运行均使用随机种子 42，训练两轮，学习率为 0.001，预热比例为 0.07，每设备批量大小为 4，且不使用梯度累积；每轮结束时进行评估并保存检查点。由于语料库按句子数量而非词元数量进行匹配，各语言完成的优化步数有所不同：英语第一语言为 70 步，每个德语第二语言训练运行均为 170 步，每个芬兰语第二语言训练运行均为 138 步。第1轮表征取自第二语言训练第一轮结束时保存的检查点（例如德语在第 85 步时保存的检查点）；第 2 轮（最终）表征则取自训练结束时保存的模型，未以最佳验证检查点替代。检查点元数据记录的 Transformers 版本为 4.30.0.dev0；此处报告的评估使用 Python 3.10.20、PyTorch 2.0.1 和 Accelerate 0.19.0 运行。

表1用于选择第二语言条件的 *URIEL/lang2vec* 度量

第二语言	句法距离 与英语的	共享的谱系特征 与英语	实验 分类
德语	0.0777	4	相对近缘的第二语言
芬兰语	0.2136	0	相对远缘的第二语言

注。 句法距离和谱系特征数据取自 URIEL/lang2vec（Littell等，2017）。句法距离指在 lang2vec 的 k近邻法估算的 103 项 URIEL句法特征中，该语言与英语不同的特征所占比例（德语为 8/103，芬兰语为 22/103）；比例越高，表示与英语的句法差异越大。该比例并非 lang2vec 预计算的句法距离（德语为 .42，芬兰语为 .53），不过两者得出的排序相同。共享的谱系特征指与英语共享的 URIEL语系特征数量（德语共享的特征为：印欧语系、日耳曼语族、西北日耳曼语支和西日耳曼语支）。德语被归类为相对近缘的第二语言条件，而芬兰语被归类为相对远缘的第二语言条件。

2.3 语言条件与类型学比较

德语和芬兰语作为对比性的第二语言条件，其依据是通过 URIEL/lang2vec 获得的句法和谱系测量指标（Littell等，2017）。德语与英语的句法距离为 0.0777，并与英语共享四项谱系特征；而芬兰语与英语的句法距离较大，为 0.2136，且与英语没有任何共同的谱系特征（见表1）。由于类型学距离并非这两种第二语言条件之间唯一的系统性差异，因此还使用模型采用的自定义字节级 BPE 分词器，而非原始 GPT-2

分词器，考察了探测句长度。与“芬兰语的黏着语形态学会产生长得多的词元序列”这一预期相反，德语和芬兰语探测集的平均分词长度几乎相同。德语每句平均为 59.50 个词元（标准差 = 38.98），芬兰语每句平均为 57.50 个词元（标准差 = 24.07）；每种语言均基于第 2.5 节所述的 20 个探测句。由此看来，平均序列长度的简单差异不太可能解释结果中德语与芬兰语之间的任何差异，但这并不能证明两种语言在子词层面的切分方式等同。在大多数大规模训练语料库中，芬兰语的资源也少于英语或德语；此处的训练语料库按句子数量匹配，而非按经核实的词元数量匹配（第 2.5 节），因此在当前设计中，语言亲缘关系与训练数据接触量无法完全区分。

2.4 顺序学习与弹性权重巩固训练条件

每种语言条件都包含顺序训练和弹性权重巩固训练两种变体。顺序训练允许模型从英语训练继续学习第二语言，且不额外采用保护与既有知识相关参数的机制。相反，弹性权重巩固会根据参数对既有信息的估计重要性，对参数变化施加惩罚；这种方法最初是为减少神经网络中的灾难性遗忘而提出的（Kirkpatrick 等，2017）。Constantinescu 等（2025）将这一约束用作后续语言学习中可塑性降低的计算操控。因此，对这些条件进行比较，可以考察限制参数变化是否保留了第一语言表征，或改变了第二语言表征的发展。在改编自 Constantinescu 等（2025）的训练代码中，弹性权重巩固条件在语言模型损失中加入了二次惩罚项，$L_{\text{total}} = L_{\text{LM}} + \lambda \, \Sigma_i \, F_i(\theta_i - \theta^*_i)2$，其中 θ^* 表示英语第一语言模型的参数，F_i 是参数 i 的估计费希尔信息，而 $\lambda = 20$ 是该代码中的默认值，

该代码不提供命令行覆盖选项。此惩罚项省略了传统的二分之一因子，且仅在训练模式下应用，因此训练期间计算的评估损失不包含该惩罚项。在英语第一语言训练完成后，利用英语验证数据，通过从模型自身的预测分布中采样目标词元（每个批次采样 10 次）并累积梯度平方，对角费希尔信息仅估计一次；累积值除以样本数量，但未除以批次数量，因此其绝对尺度取决于所处理的验证批次数量。保存的 EWC检查点使用带惩罚项的模型类，与英语检查点一同保存的费 Fisher 矩阵在结构上与该模型兼容，但训练期间未记录惩罚项的大小。

2.5 探测材料与隐状态提取

训练材料和探测材料取自 Helsinki-NLP OPUS Books 平行语料库（Tiedemann，2012）。每种语言的语料库均包含 3,000 个经过清理且互不重复的句子，分为 2,490 个训练句、255 个验证句和 255 个测试句。英语和德语句子取自德英平行语料库中互不重叠的行，而芬兰语句子则独立取自英芬平行语料库，因此三个探测集并不是彼此对应的平行译文。语料库取自 Hugging Face 发布的上游 OPUS Books 版本，未锁定特定的数据集修订版本，因此如果上游数据发生变化，仅使用固定采样种子并不能保证语料库完全相同。共享分词器是在包含留出的验证集和测试集文本的数据上训练的；这属于分词器层面的评估文本接触，而不是在测试样例上进行基于梯度的训练，但这意味着该分词器并未与探测句完全独立。我们从每种语言留出的测试集中，使用固定随机种子（42）以可复现的方式抽取了 20 个探测句，得到贯穿全文使用的英语、德语和芬兰语探测集

，用于表征分析。使用同一分词器测量的探测句长度为：英语平均每句 19.80 个词元（SD = 16.56），德语平均 59.50 个词元（SD = 38.98），芬兰语平均 57.50 个词元（SD = 24.07）。英语探测集用于测量第二语言接触后第一语言表征的变化，而德语和芬兰语探测集则追踪正在习得的语言的表征。对于每个句子，我们在每个选定的检查点，从嵌入层、Transformer第1层和Transformer第2层提取隐藏状态。注意力掩码感知均值池化将词元级激活合并为每一层的一个句子表征，同时，评估模式和禁用梯度计算避免了提取过程改变模型。探测句使用最多 256 个词元的长度进行分词，这一长度足以容纳所有探测句；其中最长的是一个德语句子，包含 177 个词元。此前的一次提取使用了 128 个词元的长度限制，导致该句被截断，因此我们从同一批已保存的检查点中以完整长度重新提取了所有德语表征；其余 19 个德语句子的向量未发生变化，本文报告的所有结果均采用修正后的提取结果。隐藏状态提供了神经网络如何编码信息的中间视角，而不是完全依赖其最终的行为输出，因此，对这些表征进行比较有助于考察模型变化（Liu 等，2019）。

2.6 表征相似性度量

我们使用余弦相似度量化表征变化，即比较训练检查点与其第二语言训练前基线之间对应句向量的方向。对于探测句 s 和层 l，相似度的计算方式为 $\cos(h_t, h_0) = (h_t \cdot h_0) / (\| h_t \| \, \| h_0 \|)$，其中，$h_t$ 是检查点 t 处对 s 进行均值池化得到的表征，h_0 是纯英语模型对同一句子生成的表征。数值范围从 −1 到 1，也可能为负值，例如英语嵌入层中的情况（表 2）。对于第一语言保持，

我们将第1轮和第 2 轮的英语表征与纯英语模型的表征进行比较；因此，相似度越高，表示英语表征的方向与其原始纯英语表征越接近。在第二语言分析中，我们将相同阶段的德语和芬兰语表征分别与第二语言训练前的纯英语模型生成的相应表征进行比较。因此，相似度越低，表示与基线表征的偏离越大；而从第1轮到第 2 轮的变化则揭示了第二语言轨迹的方向。

2.7 数据分析

在考察主要对比之前，先按语言条件、训练条件、阶段和层汇总句子级余弦相似度。首先，通过第1轮到第 2 轮的变化来衡量第二语言训练中的表征发展。随后，弹性权重巩固 – 顺序学习对比检验可塑性操控是否改变了英语保持或二语表征，而芬兰语 – 德语对比则考察与两种语言条件相关的差异。95%置信区间通过百分位自助法获得，使用10,000个重抽样样本（随机种子 42），并以探测句作为重抽样单位。在英语探测集上，阶段对比、弹性权重巩固 – 顺序学习对比以及芬兰语 – 德语对比均在模型之间比较相同的句子，因此对句子级差异进行重抽样，并报告配对 Cohen’s d_z效应量。由于德语和芬兰语探测集包含不同且不平行的句子，在二语探测集上的芬兰语 – 德语对比将两种语言视为独立组，分别进行重抽样，并报告 Hedges’ g。这些统计量描述的是当前训练模型运行中探测句之间的变异；因此，它们不表示跨独立训练模型随机种子的不确定性，仍属于对所观察表征模式的探索性测量。不包含零的区间

因此表明，在本次运行的这20个探测句中，该差异是一致的，而不是预期能够在独立训练的模型之间重复出现的差异。对于计算出的72个对比（涵盖层、阶段、训练条件和对比类型），未针对多重比较进行调整，因此应将不包含零的区间理解为描述性结果，而非验证性结果。

表 2不同第二语言训练条件下第一语言与第二语言表征的平均余弦相似度

分析	L2	训练条件	阶段	嵌入	Transformer 第1层	Transformer 第2层
英语第一语言						
保持	德语	顺序学习	第1轮	-0.0174	0.7721	0.8624
	德语	顺序学习	第 2 轮	-0.1339	0.7633	0.8609
	德语	EWC	第1轮	-0.0080	0.7726	0.8625
	德语	EWC	第 2 轮	-0.1207	0.7638	0.8616
	芬兰语	顺序学习	第1轮	0.0037	0.8058	0.8593
	芬兰语	顺序学习	第 2 轮	-0.1091	0.7945	0.8492
	芬兰语	EWC	第1轮	0.0096	0.8067	0.8587
	芬兰语	EWC	第 2 轮	-0.0982	0.7945	0.8490
L2						
表示 德语		顺序学习	第1轮	0.6780	0.7915	0.8765
	德语	顺序学习	第 2 轮	0.6674	0.8497	0.9148
	德语	EWC	第1轮	0.6781	0.7911	0.8761
	德语	EWC	第 2 轮	0.6675	0.8500	0.9147

分析	L2	训练条件	阶段	嵌入	Transformer 第1层	Transformer 第2层
	芬兰语	顺序学习	第1轮	0.5149	0.7800	0.7039
	芬兰语	顺序学习	第 2 轮	0.5071	0.8162	0.8103
	芬兰语	EWC	第1轮	0.5150	0.7800	0.7038
	芬兰语	EWC	第 2 轮	0.5072	0.8162	0.8099

注。 数值表示 20 个探测句的平均余弦相似度。对于英语母语保持，将第二语言训练后的英语表征与英语母语基线中相应的表征进行了比较。对于第二语言表征，将第二语言训练后的德语和芬兰语表征与纯英语模型在第二语言训练前针对相同探测句生成的表征进行了比较。数值越高，表示与相应的第二语言训练前基线的表征相似性越高。Sequential 指普通的第二语言训练；弹性权重巩固 = 。第 2 轮表示最终的第二语言检查点。嵌入输出和两个 Transformer 层均由 128 维隐藏状态表征组成。

结果

3.1 英语母语保持

在 Transformer 层中，在两种语言条件下，英语表征在整个第二语言训练过程中都与第二语言训练前基线保持较高的相似度（介于 .76 和 .87 之间；见表 2）；而嵌入层的相似度在第1轮时已接近零，到第 2 轮时变为负值（顺序训练条件下，德语为 -.1339，芬兰语为 -.1091）。在每一层和每种条件下，平均相似度都从第1轮到第 2 轮有所下降，其中嵌入层的降幅最大。在顺序训练的第 2 轮，芬兰语在 Transformer 第1层的英语保持相似度高于德语，

平均差异为 .0311，95% 置信区间为 [.0153, .0478], ；而 Transformer 第2层的置信区间包含零，平均差异为 -.0118，95% 置信区间为 [-.0240, .0025]。这些层级特异性模式表明，英语表征相似度在网络中的变化并不均一。

3.2 二语表征变化

在 Transformer 层中，从第1轮到第 2 轮，德语和芬兰语表征都与各自的第二语言训练前基线更加相似（见表 2）；而在嵌入层，两种语言的相似度均略有下降（德语约下降 .011，芬兰语约下降 .008）。由于与该基线的相似度越高，表示某个表征越接近纯英语模型针对同一句子所生成的表征，因此，这一模式表明，随着训练推进，从余弦取向来看，Transformer层的二语表征是朝着而非背离第二语言训练前基线移动；但这本身并不表明二语习得有所提升。芬兰语在第2层 Transformer 层中的变化最大：在整个训练过程中，相似度增加了约 .11（顺序训练下从 .7039 增至 .8103；平均变化 = .1064，95% 置信区间 [.0946，.1176]）。

3.3 训练条件（弹性权重巩固 vs. 顺序学习）

对于英语第一语言保持和二语表征而言，弹性权重巩固与顺序学习在 Transformer 层中产生了几乎重叠的轨迹（见表 2）：跨语言和阶段来看，Transformer 层平均相似度中最大的绝对弹性权重巩固 – 顺序学习差异小于 .001。这些差异虽小但并非不存在；在英语探测集的八项 Transformer 层比较中，有七项的句子级 95% CI 排除了配对弹性权重巩固 – 顺序学习差异为零的可能性；在二语探测集的八项比较中，有六项如此。对于英语探测集，嵌入层显示出更大的差异：在两个阶段，弹性权重巩固与英语

基线的相似度都高于顺序学习，其中以第 2 轮最为明显（德语高出 .0132，芬兰语高出 .0110）。因此，训练条件之间的隐藏状态差异虽小却并非不存在，并且在英语嵌入层中最大。

在原始分析之后进行的另一项参数层面分析，将保存的最终第二语言参数与英语第一语言参数进行比较，并根据估计的费希尔信息对每个平方参数变化进行加权。与顺序训练相比，弹性权重巩固使德语（0.000328 对 0.000350）和芬兰语（0.000349 对 0.000371）的 Fisher加权漂移均降低了约 6%；而与英语参数之间的未加权欧几里得距离差异不到 1%（德语：141.38 对 142.18；芬兰语：146.41 对 146.99）。这一模式与弹性权重巩固惩罚项对其设计上要保护的参数施加约束的预期一致，但这是一项不同于隐藏状态相似度的参数层面指标：它并不能说明英语熟练度得到了更好的保留，也不能证明该惩罚项在每个优化步骤中都处于激活状态。

3.4 语言条件（德语 vs. 芬兰语）

在第 2 轮中，德语和芬兰语在更深的 Transformer 层中的差异最为明显。在顺序训练条件下，芬兰语 – 德语的差异较小，其自身置信区间在第1层中跨越零点，平均差异为 -0.0335，95% 置信区间为 [-0.0729,0.0037], Hedges' g = -0.5181；但在第 2 层中，该差异明显增大且不包含零点，平均差异为 -0.1045，95% 置信区间为 [-0.1186, -0.0909], Hedges' g = -4.4256。由于德语和芬兰语探针使用的是不同句子，这些组间比较除了反映语言条件外，还可能反映句子内容和分词的差异；而第 2 层中的较大标准化效应，既反映了约 0.10 的绝对差异，也反映了不同探针之间第 2 层相似度的低变异性；不应将其解读为

较大的行为效应。结合表 2 可以看出，第二语言表征在训练过程中变得更加接近第二语言训练前基线，而德语和芬兰语的轨迹在更深的 Transformer 层中分化得最为明显。

4. 讨论

第一个研究问题探讨的是，在德语或芬兰语二语习得的早期阶段和最终阶段，经英语训练的模型的内部表征如何变化。英语第一语言表征在各 Transformer 层中总体保持相对稳定，但嵌入层除外；与此同时，正在习得的第二语言的表征在同一时期逐渐接近第二语言训练前基线（表 2）。从余弦取向来看，这表明随着训练继续，德语和芬兰语的句子表征越来越接近纯英语模型可能生成的表征，而非与之渐行渐远。一个可能但尚未验证的解释是，这反映了训练的一般特性，而不是特定语言的学习信号：如果随着训练推进，隐藏状态表征变得更加集中于共享方向，那么无论所处理的语言内容是什么，与基线的余弦相似度都可能上升。本研究并未测量表征的离散程度；此外，由于余弦相似度能够捕捉表征对齐，却无法单独说明隐藏状态编码了哪些语言信息（Kornblith 等，2019），本研究设计无法将这种解释与二语特异性变化区分开来，因为研究中没有纳入对照比较。通过无关句子或打乱的句子进行对照比较，可以揭示这种增幅是否也会在与模型实际学习德语或芬兰语的内容无关时出现。尽管如此，这种对表征变化的计算视角仍补充了以往关键期建模的行为研究重点（Constantinescu 等，2025）；后者主要通过模型性能而非内部表征结构考察从一语到二语的

训练效应。

第二个研究问题探讨了德语和芬兰语在嵌入层与 Transformer 层中的表征变化有何不同。两种条件在较深层的差异最为明显：芬兰语在 Transformer 第2层的增幅更大，而德语在 Transformer 第1层的增幅更大。芬兰语与德语之间的差异较小，在 Transformer 第1层中，其置信区间跨越零，均值差为 -0.0335，95% 置信区间为 [-0.0729，0.0037]，，Hedges' g = -0.5181；但在 Transformer 第2层中，差异明显增大且不再包含零，均值差为 -0.1045，95% 置信区间为 [- 0.1186，-0.0909]，，Hedges' g = -4.4256。这种差异与语言亲缘关系影响网络不同深度整合新材料的方式这一观点相符，因为研究依据 URIEL/lang2vec框架，选取德语和芬兰语来代表与英语类型学距离迥异的语言（Littell等，2017）。然而，德语和芬兰语的差异并不止于类型学距离。两个探测集的分词后句子平均长度相近（分别为59.50和57.50个词元），但平均值相近并不能排除基于分词的解释，因为这两个非平行探测集的长度分布、子词切分和词汇内容各不相同。训练暴露量也有所不同：德语训练运行完成了170个优化步骤，芬兰语训练运行则完成了138步。此外，在典型训练语料库中，芬兰语仍比德语更低资源；而且，两种语言细粒度的子词构成以及完整训练语料库的词元总数都尚未经过独立核实，因此目前还无法将语言亲缘关系确定为第2层差异的唯一原因。这一限定与Grosjean（2013）更广泛的审慎观点一致：不应将双语

系统中的两种语言视为彼此孤立、能够清晰分离的条件，而应将其看作共同发展的整体。

第三个研究问题探讨了EWC约束训练在多大程度上比无约束的顺序训练更能保留英语第一语言表征。EWC与顺序训练在Transformer层中产生的变化轨迹几乎相同，平均相似度的最大差异不到 .001；而英语嵌入层的差异稍大，EWC与英语基线的相似度更高（表 2）。EWC限制对那些被估计为对先前习得信息至关重要的参数所做的更改。这一机制最初旨在减少神经网络中的灾难性遗忘（Kirkpatrick 等，2017）；Constantinescu 等（2025）则将同一约束用作第二语言学习期间降低可塑性的一种计算操作。在本研究的训练方案下，观察到的微小差异表明，EWC在参数层面的约束仅使所考察的Transformer层隐藏状态表征略有不同，而英语嵌入层的差异稍大。不过，在参数层面，EWC使两种语言的英语参数所受的Fisher加权漂移都减少了约 6%（第 3.3 节），因此，这项约束确实改变了其目标参数；但这种参数层面的差异并未转化为本文所用句子层面余弦度量的显著变化。鉴于$\lambda = 20$ 以及第 2.4 节所述的Fisher缩放，这个小型模型中的惩罚力度相对较弱，可能是原因之一。此外，该约束也可能集中作用于某些参数或层，而基于探测的方法分析采用了Liu 等人（2019）的层间比较方法，未必能够捕捉到这些影响。

这些发现仍属探索性：分析基于单个训练模型随机种子、单个小型双层架构，以及每种语言 20 个探测句，而余弦相似度

本身无法区分真正的语言重组与表征空间中的普遍漂移（Kornblith 等，2019）。此外，与仅英语基线的余弦相似度并非衡量第一语言保持或第二语言熟练度的行为指标；研究设计也未改变第二语言习得的年龄或开始时间，因此研究结果无法直接说明人类关键期效应。其他局限还包括：分词器是在包含留出数据划分的文本上训练的，上游数据集修订版本未固定，语料库按句子数量而非词元数量进行匹配，以及缺少留出损失或困惑度等直接行为指标。未来若将研究设计扩展至多个随机种子、更多语言对，并加入乱序句子对照，将有助于判断本文报告的模式反映的是第二语言特定学习，还是更广泛的训练动态。

研究现状与局限

数据包括五种模型条件（英语第一语言，以及在顺序学习和弹性权重巩固训练条件下的德语和芬兰语第二语言）各进行一次训练运行（随机种子 42）；每种语言的训练语

[2]本项目是以下课程的一部分：双语心智：语言、大脑与发展中的文化，该课程是巴纳德学院、哥伦比亚大学开设的一门四学分研讨课。更多课程信息请参阅哥伦比亚大学课程目录。

因此无法检验人类关键期效应。这些结果应视为进一步研究的基础，而非确凿结论。

参考文献

Constantinescu, I., Pimentel, T., Cotterell, R., & Warstadt, A.（2025年）。通过神经语言模型研究语言习得中的关键期效应。计算语言学协会汇刊，*13*，96–120。https://doi.org/10.1162/tacl_a_00725

Grosjean, F.（2013年）。双语现象：简要介绍。载于 F. Grosjean & P. Li（编），双语心理语言学（第 5–25 页）。约翰·威立父子出版公司。

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R.（2017年）。克服神经网络中的灾难性遗忘。美国国家科学院院刊，*114*（13），3521–3526。https://doi.org/10.1073/pnas.1611835114

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G.（2019年）。神经网络表征相似性再探。第*36*届国际机器学习会议论文集，3519–3529。https://proceedings.mlr.press/v97/kornblith19a.html

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., & Levin, L.（2017年）。URIEL 和 lang2vec：以类型学、地理和系统发育向量表示语言。载于 M. Lapata, P. Blunsom, & A. Koller（编），欧洲计算语言学协会分会第*15*届会议论文集：第*2*卷，短篇论文（第8–14页）。计算语言学协会。https://aclanthology.org/E17-2002/

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A.（2019年）。上下文表征的语言学知识与可迁移性。载于 J. Burstein、C. Doran 和

T. Solorio（编），《*2019*年北美计算语言学协会年会论文集：人类语言技术》，第 *1* 卷（长篇与短篇论文）（第 1073–1094 页）。计算语言学协会。https://doi.org/10.18653/v1/N19-1112

Tiedemann, J.（2012）。OPUS 中的平行数据、工具和接口。载于 N. Calzolari、K. Choukri、T.Declerck、M. U. Doğan、B. Maegaard、J. Mariani、A. Moreno、J. Odijk 和 S. Piperidis（编），第八届国际语言资源与评估会议论文集（*LREC'12*）（第 2214–2218 页）。欧洲语言资源协会（ELRA）。https://aclanthology.org/L12-1246/