图像、词语与思想：迭代学习模型面临的一项更具挑战性的语言任务

Hyoyeon Lee、Seth Bullock 与 Conor Houghton

布里斯托大学科学与工程学院，英格兰
conor.houghton@bristol.ac.uk

2025年12月23日

arXiv:2601.02911v1 [cs.CL] 6 Jan 2026

摘要

迭代学习模型模拟语言代际传递，以探究语言传递施加的限制如何促进语言结构的涌现。尽管每位建模的语言学习者起初都如同一张白纸，但限制学习者接触的话语数量的瓶颈，能够促使一种语言涌现：这种语言没有歧义，受语法规则支配，并且在连续世代间保持一致；也就是说，它具有表达性、组合性和稳定性。近期，一种计算上更易处理且具有更高生态效度的半监督迭代学习模型问世，该模型在自编码器架构中结合了监督学习与无监督学习，使研究者得以探索意义—信号空间大得多的语言传递动态。在此，我们首次成功将该模型应用于一项语言学习任务，其中涉及更复杂得多的意义的通信：七段显示图像。该模型中的智能体能够学习并传递一种具有表达性的语言：128种字形均使用不同的编码；这种语言具有组合性：信号成分与意义成分之间存在一致的映射；同时具有稳定性：语言不会代代改变。[1]

引言

人类语言具有独特的演化特征，因为它源于至少三种适应过程之间的相互作用：（i）语言生物学基础的缓慢进化适应；（ii）通过社会学习实现的快速个体适应，这有助于语言传递；以及（iii）文化—语言演化的“世界3”中间层次，它体现了语言自身所经历的适应（Popper，1972年；Popper 与 Eccles，1977年；克里斯蒂安森与柯比，2003年）。鉴于这些并行的适应过程发生在截然不同的时间尺度上，其中一些延续了许多千年（迪肯，1997年；克里斯蒂安森与 Chater，2008年），历史语言的记录并不完整（布龙菲尔德，1933年；Kirby 和 Hurford，2002年），语言既是研究对象，也是我们借以交流相关观点的媒介（Tomasello，1999年），而当前的语言行为则由一种极其复杂且尚未得到充分理解的神经基质实现（珀佩尔与希科克，2004年；克里斯蒂安森与 Chater，2008年），因此，要理解这些适应过程之间的相互作用颇具挑战。由此，人类语言的起源与本质之争涵盖了广泛的观点，从强调文化与互动的论述，到着重探讨符号认知的共同演化的主张，以及对以句法为中心的解释提出的多元主义批评（斯蒂尔斯，2011年；迪肯，1997年；博克斯，2021年；霍顿，2025年）。

用于理解语言变化基本属性的综合性计算方法

[1]这是即将发表于 EvoLang2026 的论文扩展版，其中补充了数值实验的详细信息。

颇具吸引力，因为它们为研究演化语言现象提供了另一扇窗口。尤其是，基于计算智能体的语言学习模型能够让我们深入了解语言学习约束可能如何影响、塑造或决定语言属性，并通过探索“语言可能呈现的样貌”（Kirby，2002年），为自然语言某些属性据称具有普遍性这一现象提供解释，例如特定语法形式的存在，或特定语言特征的共现。在模型中，通常会用非常简单的语言属性来替代真实语言更丰富的属性；一个常见的属性集合——本文采用的集合——包括表达性、组合性和稳定性（Kirby，2001）。

迭代学习模型（ILM；Kirby 和 Hurford，2002年）是一个关于语言传递与演化的模型，已成功应用于简单的人工语言学习任务。在这些任务中，意义和信号通常仅限于简短的比特串。本文将该模型拓展到一项更复杂的任务：智能体发展出一种稳定、富有表达性且具有组合性的语言，并借此成功交流有关结构化图像空间的信息。这些图像基于七段 LED 显示屏，常见于计算器、数字手表和数字时钟。

据我们所知，这是首次证明迭代学习模型能够应用于如此复杂的任务。这得益于本文所用的迭代学习模型纳入了自编码器学习。与以往的交互式语言模型相比，该模型在计算上更易处理，也具有更高的生态效度；此外，半监督迭代学习模型的计算架构似乎还模仿了现实世界学习的不同方面。将模型从比特串拓展到字形图像，需要增加一个专门用于图像编码与解码的训练阶段，这呼应了低层次感官表征学习与基于这些表征开展的高层次学习之间的分离。

学习还包含一种认知机制，要求智能体在内部运用自己的语言，将其作为一种“思维语言”。这种简单的智能体模型所能解释的范围有限，但一种可能的解读是，语言与思维的这种联系对应着认知演化中的一个重要步骤：语言的内化使多步骤推理或思维链推理成为可能，而这些推理形式最近已被证明在机器学习中十分重要（Wei 等，2022）。

归根结底，我们的目标是理解语言结构与认知过程之间的关系。诸如迭代学习模型（ILM）这样的简单智能体模型，只能被视为灵感来源，而非对现实的字面描述。从这一视角来看，半监督迭代学习模型的行为表明，基于自编码器的学习可能在支持结构化语言的涌现与传递方面发挥着特定且重要的作用，这也促使我们构建更丰富的迭代学习模型变体，例如本文所探讨的模型。未来的研究应纳入更精细的智能体模型和现实世界实验；我们希望本文提出的模型能代表朝这一目标迈出的进展。

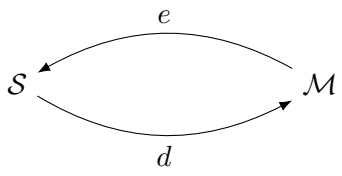
迭代学习模型

迭代学习模型是一种基于主体的模拟，旨在探究语言在何种条件下会逐渐变得具有表达性、组合性且稳定（Kirby，2001；Kirby 和 Hurford，2002年）。迭代学习模型由一个使用语言的“导师”智能体组成，该智能体通过向一个起初毫无经验的“学习者”智能体展示有限数量的话语来对其进行教学；每条话语都将某种含义与导师语言中表达该含义的信号配对。经过这一教学阶段后，学习者便成为导师，并以同样的方式教导一位新的、起初毫无经验的学习者。语言学习瓶颈是该模型的一项关键特征：由于每位学习者在学习过程中接触到的导师语言仅为其中一部分，因此当其自身成为导师时，就需要为学习经历中未曾涉及的含义生成信号。对泛化能力的这一要求，对迭代学习模型的行为至关重要。

尽管可能存在许多不同的语言，其中大多数以任意方式将意义映射为信号，但在适当条件下，迭代学习模型中会出现这样的语言：它们具有表达性、没有歧义，并且具有组合性，即信号的组成部分会以受规则支配的方式，始终编码意义的组成部分；

而且稳定， successive 学习者尽管在学习过程中只接触到语言中一小部分随机片段，却能学会同一种语言。迭代学习模型揭示的核心见解是，只有组合语言在经过学习瓶颈传播后才能保持稳定且具有表达性，因为任何缺乏组合语法的意义—信号映射，都无法可靠地推广到瓶颈集合中未提供的意义。

迭代学习模型（ILM）有多个版本，其中一个由 Kirby 和 Hurford（2002年）提出的典型示例，设定了一个简单可能语言的空间。每种语言都将有限的意义集合 $\mathcal{M}$ 映射到有限的信号集合 $\mathcal{S}$，其中每个意义和信号都由长度为 n 的二进制向量表示。每个语言智能体自身的语言由其解码器映射 $d : \mathcal{S} \to \mathcal{M}$ 表示；它是一个简单的神经网络，将每个可能的信号映射到该信号所传达的意义。其编码器映射 $e : \mathcal{M} \to \mathcal{S}$ 则是一个查找表，将每个意义映射到传达该意义的信号：



在学习过程中，会随机选取一部分意义 $\mathcal{B} \in \mathcal{M}$ ，导师将每个意义与其语言中表达该意义的信号配对 $\{(e(b)\ b)|b \in \mathcal{B}\}$。这一组意义—信号对用于训练学习者的解码器。训练结束后，学习者会对编码器进行逆向工程，使其与解码器保持一致，之后便可成为新创建的天真学习者的导师。遗憾的是，逆向工程，也就是逆向构造这一步并不现实，因为自然语言学习中没有与之相当的步骤。此外，这种方法难以随语言规模扩大而扩展，计算成本会急剧增加，以至于对于位长度超过十的语言，迭代学习模型模拟都无法进行。

半监督 ILM

在半监督迭代学习模型（Bunyan 等人，2025）中，智能体拥有两个神经网络：一个编码器，将意义映射为信号（一种二进制向量）；另一个是解码器，将信号映射为意义。移除将编码器输出离散化为二进制的步骤后，编码器和解码器也可以串联起来，构成一个自编码器。这个自编码器支持学习，并用于耦合编码器与解码器。

半监督 ILM 将语言传递重新表述为监督式学习情境与半监督式 学习情境的结合，从而不再需要逆向构造。监督式学习情境模拟导师指着一张孩子偷饼干的图片并说：“看，一个孩子正在偷饼干！”这有助于学习者训练其解码器映射和编码器映射。无监督式 学习情境则对应于学习者思考：“我会如何描述这幅图像？这个描述又会让我想到什么图像？”这一过程还有两个作用。首先，它确保解码器与编码器保持一致，使意义被编码为信号后，解码时能够还原出原始意义。其次，它能防止语言学习中固有的过度泛化倾向。

过度泛化倾向的一个经典现实例子，是事物一词。这个词已经从指代立法集会，演变为任何物体或概念的通用称呼。只有为立法集会引入新的词语，例如国会、爱尔兰众议院 或议会，人们才仍能轻松地讨论，甚至思考这类集会。过度泛化会侵蚀意义表征（奥威尔，1949年），但它也是通过隐喻进行延伸这一做法的自然结果，而这种做法对于讨论新概念十分有用。过度泛化是迭代学习模型面临的一项特殊挑战。如果没有某种机制加以阻止，语言很快就会崩溃，变成一种缺乏表现力的语言。通过模拟内部反思，并要求保留意义以便重构输入，半监督 ILM 避免了这一危险。需要注意的是，迭代学习模型并不是一种交流模型，而这一点也可能在防止过度泛化方面发挥作用。不过，



图 1：**基础字形示例**。 每个基础字形都是基于经典七段式 LED 显示屏的 28×28 图像。所有 128 个字形都会用到，其中许多并不对应数字或字母。此处展示的三种字形分别对应数字 1、5 和 8。图像以展平后的 784 维向量表示，每个分量的取值介于零和一之间。此处使用的颜色仅用于示意，旨在模仿数字时钟显示屏。

迭代学习模型并非交流模型，而这也可能在防止过度泛化方面发挥作用。

每个监督学习阶段都会从导师的语言中随机选取一个意义—信号对，呈现给学习者的编码器网络或解码器网络，并通过反向传播逐步提升每个网络将意义映射为信号、或将信号映射为意义的能力。这样，学习者便能学会导师提供的语言示例；这些示例只是所有可能语言示例的一部分，而且导师所教授的示例集合与该导师自己曾学过的示例集合并不相同。相比之下，无监督学习会将一个随机意义同时呈现给自编码器网络的输入层和输出层；该网络由学习者的编码器网络和解码器网络串联而成。这种迭代学习模型变体已成功用于探索迭代学习模型在位串语言上的行为，其规模比以往通常使用的语言大一千倍（Bunyan 等，2025）；该变体也已被用于初步探索语言接触等现象对语言结构的影响（Bullock 和 Houghton，2024）。

七段式语言任务

七段显示屏的使用历史出人意料地悠久；20 世纪 20 年代，它们被用于结合人工与自动路由的电话交换局（Clark，1929）随着数字手表和计算器日渐普及，七段显示屏也广泛应用起来。后来，更先进的显示技术相继问世，使得七段显示屏不再像从前那样无处不在。不过，大概仍无需解释什么是七段显示屏：它是一种通过点亮七条短垂直或水平线段的某种组合来显示数字和部分字母的设备；图 1 展示了三个示例。

A -噪声 1

B- 噪声 2

C -噪声 3

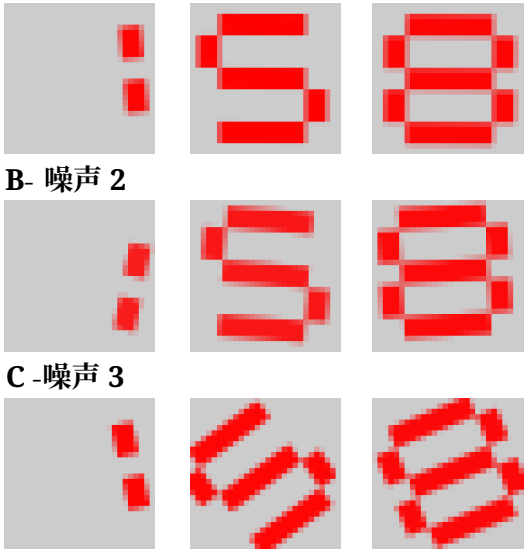


图 2：**示例图像**。对于每个基础字形，通过使用小幅随机旋转、平移和强度变化，生成了 100 张变体图像。此处展示了文中所述三种噪声条件下每种字形的示例图像：**A**：**噪声 1**、**B**：**噪声 2** 和**C**：**噪声 3**。

这里，128 种可能的七段显示字形（包括空白字形）被用作具有组合结构的意义空间，也就是说，字形呈现出内部结构。每个基础字形都是一个 28×28 数组，其组件位于 $[0, 1]$ 中；其尺寸经过选择，以匹配 MNIST 数据集。为了为每个基础字形生成一组变体图像，首先将每个字形按十倍比例上采样，然后采用了三种操作。

1. 强度调整：开启像素的值被更改为 $1 - \epsilon$ 其中

$$\epsilon \sim \text{指数分布}(\mu) \tag{1}$$

2. 旋转：随后图像会旋转一个角度 θ ，其中

$$\theta \sim \text{Normal}(0, \rho) \tag{2}$$

角度以弧度为单位测量。

3. 平移：图像平移了 δx,δy ，其中

$$\delta x, \delta y \sim \text{Normal}(0, \sigma). \tag{3}$$

最后，图像被下采样回 $28 \times$28。这里考虑了三个不同级别的噪声：

噪声 1： $\mu = 0.03$, $\rho = 0.05$, $\sigma = 4.0$；图 2**A**

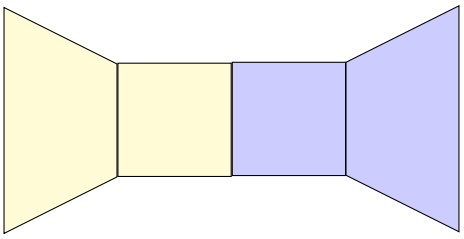
噪声 2： $\mu = 0.1$, $\rho = 0.1$, $\sigma = 5.0$；图 2**B**

噪声 3： $\mu = 0.1$, $\rho = 0.25$, $\sigma = 8.0$；图 2**C**

图像数据集及用于生成它的代码可在线获取，详见下文“数据与代码”部分。

网络与模拟

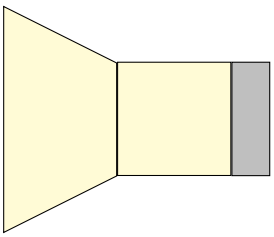
每个语言学习智能体都由一个包含四个部分的人工神经网络表示。这些部分组合在一起构成一个自编码器配置 A，可表示为



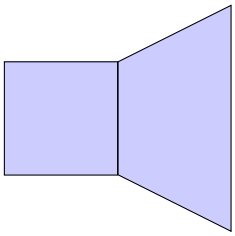
E_i E_w D_w D_i

第一部分 E_i 是图像编码器，这是一个 $784 \times 128 \times n_l$ 前馈全连接神经网络。第一层的大小当然由图像的维度决定。中间层的维度可以变化，但本文报告的所有模拟都采用相同的值。潜在维度 n_l 是 E_i 输出层的大小，也是词语编码器 E_w输入层的大小。这是一个 $n_l \times n_l \times n_l$ 前馈全连接神经网络。接下来， D_w词语解码器 是另一个 $n_l \times n_l \times n_l$ 前馈全连接神经网络。最后，图像解码器D_i是一个 $n_1 \times 128 \times 784$ 前馈全连接神经网络。本文报告的大多数模拟中，$n_l = 7$ 或 $n_l = 10$；不过，有一种情况下，词语编码器和词语解码器的不同层大小各不相同。作为一种实用的约定，网络用大写字母表示，相应的映射用小写字母表示。

智能体的网络并非以 A 中所示的特定自编码器配置进行训练；事实上，训练中使用了四种不同的组合。智能体的编码器为 $E = \delta \circ E_w \circ E_i$:

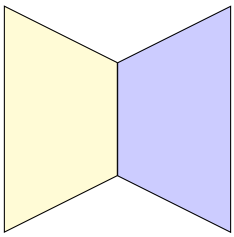


映射 δ 是离散化映射 $\delta z \to (z > 0.5 \,?\, 1 : 0)$，因此 E 将图像映射为二进制 n_l 向量，在这里这些向量被视为词。智能体的解码器为 $D = D_i \circ D_w$:



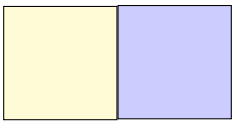
它将词映射为图像。

在半监督 ILM 中，网络也会作为自编码器进行训练。这里指的是训练两种不同配置的自编码器。外部 自编码器， $O = D_i \circ E_i$:



旨在模拟初级视觉系统的自发发育，也就是初级感觉系统中通过神经过程发生的那种精细化。一个例子是哺乳动物视觉系统中视网膜拓扑图的活动驱动的精细化：甚至在眼睛睁开之前，自发的视网膜波就会在新生期视网膜上传播，并在没有任何高阶反馈的情况下，促使投射区和眼优势柱都变得更加精细（Meister 等人，1991；Shatz 和 Stryker，1978；Cang 和 Feldheim，2013）。

内部自编码器，



旨在模拟内部反思；视觉系统的输出由 E_w映射到潜在层，即产生信号的那一层，然后再由 D_w 解码，以恢复原始信号。如果模型要发展出表达能力，就需要一个内部自编码器；没有它，模型的动态会将语言推向越来越小的词集。可以将这个第二自编码器视为对内部认知反思的建模，类似于被假设在语言学习和认知发展中发挥关键作用的内部排演或内言语过程（维果茨基，1986；Carruthers，2002年）。

在每位学习者开始接受教育时， $b = 40$字形会从全部 128 个字形中随机选取；随后再从与这 40 个字形对应的全部图像中随机挑选 400 张图像。导师的编码器 E 用于构建由 400 对图像—信号组成的瓶颈集。另有一组“日常”数据集，由从全部 128×100 变体图像中随机选取 1200 张图像。

一次训练轮次包含400次迭代，瓶颈集中的每个元素对应一次；在每次迭代中，使用来自 $\mathcal{B}$ 的一对样本训练学习者的编码器 E，并使用另一对样本训练学习者的解码器 D；随后从日常集中选取 $r = 15$ 图像来训练外部编码器 O，并使用另外 $r = 15$ 训练内部编码器 I；学习者自身的图像编码器用于将这些图像转换为 I的输入。所有训练均采用随机梯度下降，学习率为 $\eta = 15$，每个学习者训练 $n_e = 15$ 轮次；损失函数为均方误差损失；所有非线性函数均为Sigmoid函数。

评估

模型的语言通过三个量进行评估： x、 c和 s，分别对应表达性、组合性和稳定性。所有基础字形的集合 $\mathcal{G}$ 用于计算 x和 s的度量。令 $s_i = e(g_i)$，其中 $g_i \in \mathcal{G}$，而 e是训练结束时对应于学习者编码器的映射；表达性即该信号集合的大小 $x = |\{s_i | i \in 1 : 128\}|/128$。语言稳定性通过比较学习者使用的信号与导师使用的信号来计算，采用显而易见的记法 $s = |\{s_i^t == s_i^{t-1} | i \in 1 : 128\}|/128$。

组合性总是很难定义。这里采用了一种代理度量。这种度量的优点是相对简单。它使用七个基础字形，这些字形中只有一个片段处于激活状态，我们将这组字形称为独热字形。在最大程度具有组合性的代码中，每个独热字形都应由信号中的一个唯一比特编码。为了计算这一度量，我们计算与这七个独热字形相对应的信号，并将它们组成一个矩阵 $C = [s_1|s_2|\ldots|s_7]$。如果$n_l = 7$这是一个方阵。若某个信号比特对于某个片段而言，仅在该片段对应的位置为 1、其他位置为 0，或在该片段对应的位置为 0、其他位置为 1，那么它就可以编码该片段。为考虑这一点，我们检查 C 的每一行；如果该行中有四个或更多的 1，就将该行中的值取反，即将 1 改为 0，将 0 改为 1。接下来，

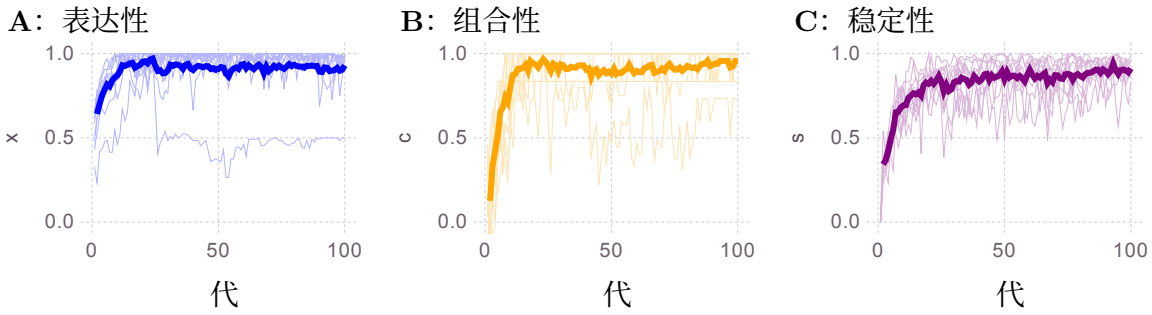


图3：**一种具有表达性、组合性且稳定的语言逐渐形成**。细线表示十个各自进行100代的实例中表达性、组合性和稳定性的变化，$n_l = 7$。粗线表示平均值；这三项度量的定义范围均为零到一。

计算行和与列和，$r_a = \sum_i C_{ai}$ 以及 $c_a = \sum_i C_{ia}$。使用一种修正倒数$\phi(r) = (r > 0 ? 1/r : 0)$对这些值进行评分，随后计算组合性

$$\tilde{c} = \frac{1}{14}\left(\sum_a [\phi(r_a) + \phi(c_a)]\right) \qquad (4)$$

当 $n_l > 7$ 针对所有可能的七列组合计算该值时，n_l 列和 $\tilde{c}$ 是最大值。最后，通过构造由伯努利分布（0.5）抽取分量的七个 n_l-向量集合，计算背景值 c_0，并将其用于对组合性的计算进行归一化，计算方式为$c = (\tilde{c} - c_0) / (1 - c_0)$。

这个针对 c的代理度量并不完美：虽然组合语言的该度量值为最大值1，但原则上，值为1并不能保证语言具有组合性。对于组合语言，每个独热字形都必须由一个比特编码，并且每个比特只能编码一个字形。根据此处的定义，这确实是 $c = 1$ 所要求的。然而，对于组合语言，在一个字形中编码特定片段的比特，在所有字形中都必须始终编码同一个片段。$c = 1$ 并不要求这一点，因此非组合语言的该度量值也可能为1。例如，目视检查输出（可以使用图7所示类型的图表）可知，这种情况并未发生，而且该度量在实践中似乎能够作为组合性的代理度量发挥作用。

结果

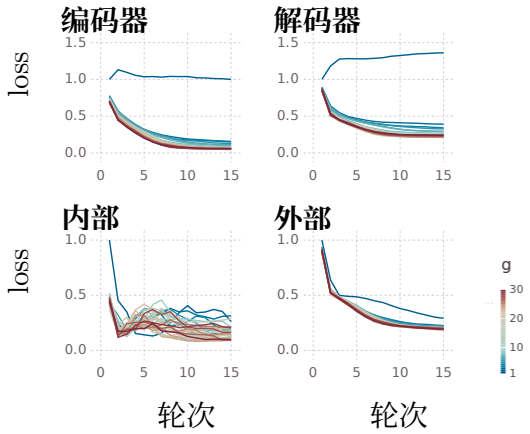


图4：**组合性和表达性语言更容易学习**。图中绘出了学生训练过程中损失随轮次的变化。学生网络的四个不同部分$n_l = 7$ 经过 $g = 30$ 代训练，并在此绘出各部分的损失；不同代以不同颜色区分，颜色越暖表示代数越晚。第一位学习者第一轮的损失用于对图进行归一化，并对十次实例化的结果取平均。

尽管这一任务比迄今为止迭代学习模型所应用过的任何任务都更为复杂，但该模型仍成功发展出一种具有表现力、组合性和稳定性的语言。图3展示了这一点，其中绘制了使用 $n_l = 7$ **噪声1**图像数据集时的演化过程。x、c和 s都迅速接近1附近的值。随着语言的组合性增强，学习也变得更加容易；图4展示了这一点。在后期的

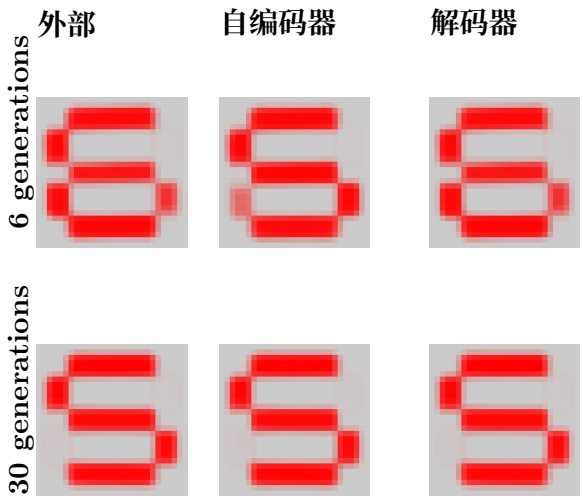


图5：**5个字形的重构图像**。顶行展示经过六代后的重构结果，底行展示经过30代后的重构结果。图中展示了三种不同类型的重构：第一列为**外部**，展示由外部自编码器进行的重构， O；第二列为**自编码器**，展示由整体自编码器进行的重构， A。第三列为**解码器**，与前者类似，即 $D \circ E$，因此它与自编码器的区别在于包含了离散化步骤。经过六代后，网络已经开始出现组合性错误，即包含或遗漏完整片段，但不同类型的自编码结果彼此尚未完全一致。到第30代时，重构结果几乎完全相同。此示例使用了 $n_l = 10$。

代中，损失下降得更快，并达到更低的值；这一效应在内部自编码器中最为明显，因为它受到监督学习与无监督学习之间竞争的影响最大。对于外部自编码器而言，这一效应最不明显，O。

仅有七个比特时，组合性语言无法对噪声的任何方面进行编码，实际重建效果也很差。图5展示了使用更长的 $n_l = 10$词的网络所重建的图像。虽然输入是一个基础字形，但网络是在**噪声 1**图像数据集上训练的，因此重建结果略显模糊，不过显然是成功的。

该模型也成功演化出一种适用于**噪声 2** 数据集的表达性组合语言；然而，这需要更大的 n_l值。在图 **6A** 中，使用具有 $n_l = 7$ 个潜在节点的网络和**噪声 2**数据集， x 和 c 的值基本都低于 0.5。稳定性 s 较高，但仅有高稳定性并不值得称道，因为两个智能体即使使用一种将每种含义都映射到同一信号的语言，也会获得较高的稳定性评分。虽然这里并未出现如此极端的情况，但模型显然未能成功。不过，将潜在节点的数量增加到$n_1 = 10$ 后，便能迅速形成一种表达性组合语言，见图 **6B**。网络足够大时，也能为**噪声 3** 图像形成一种表达性组合语言，见图 **6C**。

有趣的是，在这些例子中，语言从未趋于稳定。起初，这似乎出乎意料。组合性语言的结构意味着，它能够从一组瓶颈数据推广到下一组；这让语言更容易学习，也理应有助于稳定性。然而，当 $n_l = 10$ 例如，组合性只体现在十个信号位中的七个上，而图像的组合结构也仅由这七个位来体现；同一字形的图像之间的其他差异则源于噪声。可以设想，额外的三个比特能够提供一些信息，说明噪声如何影响某个特定字形的不同图像。然而，迭代学习模型似乎无法在噪声中找到组合结构，因此这些额外的比特并不稳定。图7对此作了说明：同一字形的不同变体图像之间发生变化的比特，也正是代与代之间发生变化的比特。在这种情况下，尽管从先前语言使用者那里学习的机会有限，仍然形成了一种共享语言，能够以跨代稳定的方式传达每个字形的身份；然而，虽然智能体也逐渐能够表示这些字形所受到的扰动，语言的这一特征却始终带有鲜明的个体性。

讨论

此处呈现的结果表明，半监督 ILM 能够成功演化出具有表达性、组合性且稳定的

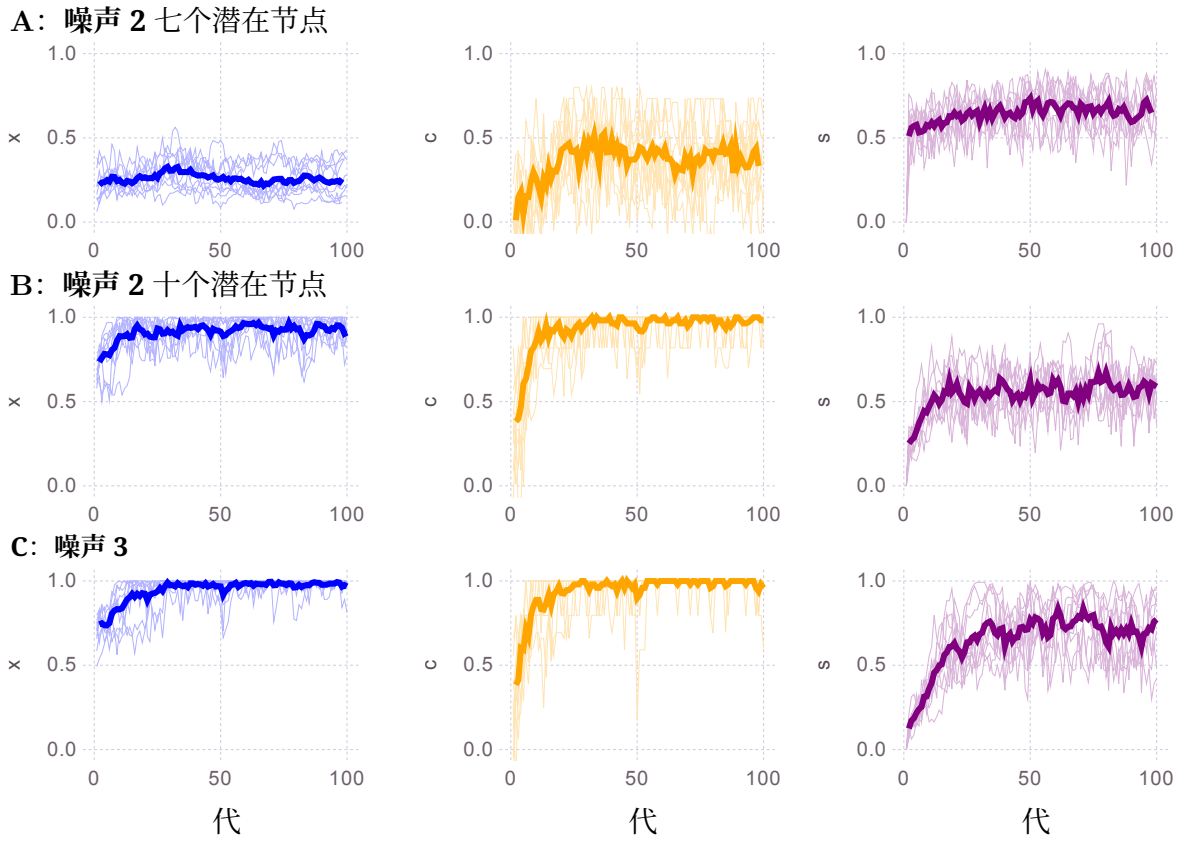


图6：**迭代学习模型在噪声更大的图像上的表现**。此处各列的顺序与图3相同，x、 c 和 s 。在前两行中，**A**和**B**使用的是**噪声2**图像。**A** 中绘制的网络（包含 $n_l = 7$ ）未能建立良好的语言。然而，扩展潜在维度的数量，使 $n_l = 10$ ，就能产生具有表达性和组合性的语言。其稳定性仍然有限。只要潜在节点的数量足够多，即使对于**噪声3**图像，模型也能实现具有表达性和组合性的语言；**C** 展示了 E_w 为$20 \times 18 \times 15$时的结果。

语言，方法是从带噪声的七段显示图像中发现组合结构。这标志着意义空间复杂度方面的一项进展。该模型的训练包括显而易见的监督训练（使用导师提供的示例），以及先前在（Bunyan等，2025）中介绍的无监督自编码器学习。为了在这个更复杂的情形下取得成功，自编码器被拆分为两种配置：外部自编码器使网络能够学习数据的主要结构，这类似于在 V1 中学习视觉数据的基本结构；内层自编码器则学习数据更高层次的组合结构。

显然，这里报告的结果只是初步的：我们尚未通过改变瓶颈大小来探索参数空间，也没有系统地探究网络的架构。更重要的是，我们尚未描绘组合性的轮廓。我们的结果中最有趣的一点是，网络似乎能够利用七段图像的组合结构，却无法利用噪声的结构；未来可以研究图像在位置或方向等方面的变化，因为这些变化介于可能被视为图像属性的因素与可能被视为噪声的因素之间。

归根结底，我们的目标是利用迭代学习模型研究语言结构、语言传递以及思维结构对语言的影响之间的关系。在这方面，成功将迭代学习模型应用于要求更高的输入数据颇有意义；值得注意的是，这需要将训练拆分为——

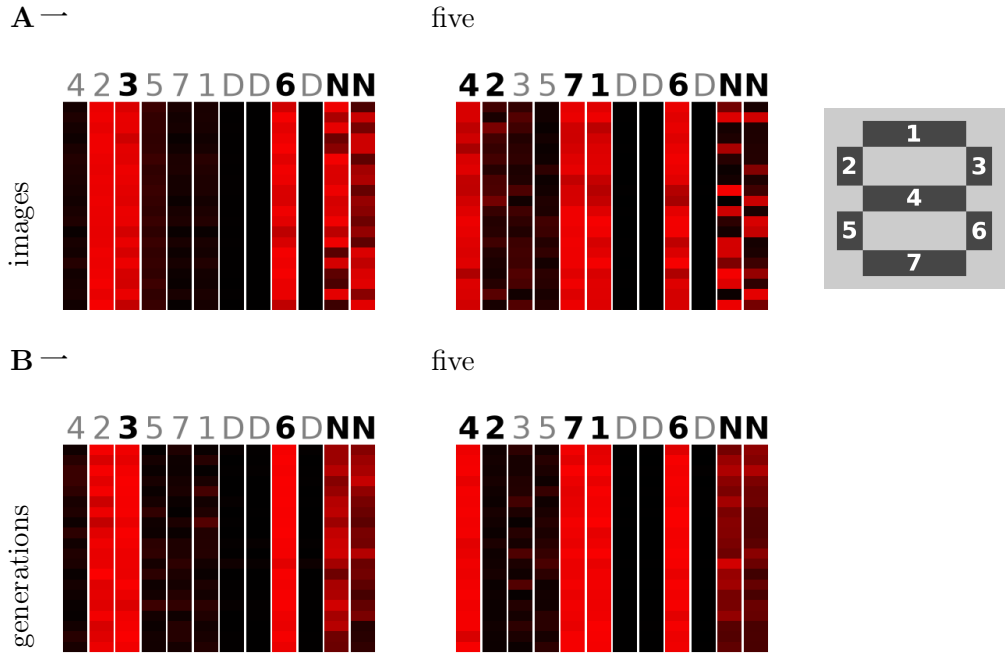


图7： **潜在激活**。这些展示了两个字形的潜在激活。在每种情况下，各列对应潜在节点。其中一些列已根据其响应的字形片段进行标注；处于激活状态的片段以粗体标示。对于 **A**，考察第60代，此时 $x = 1.0$和 $c = 1.0$。对于每个字形，每一行展示20个**噪声 2 1**或5幅图像中的一幅如何通过编码器 E映射为潜在激活。未应用离散化，但潜在激活通常接近0或1。许多激活在变体图像之间保持不变；这些激活大概对应于字形片段的组合编码。对于 **B** ，考察第50代至第69代，此时 $s = 0.496$。每一行对应于不同代中基础字形的潜在值，较低的行对应较晚的代。这里显示变化的列，与**A**中跨图像显示变化的列相同。对于这些模拟，使用了a$16 \times 14 \times 12$ E_w 和噪声3数据；这个较大的潜在空间有助于网络快速学习，但看起来其中一些标记为D的潜在节点不再处于激活状态。在未标记的列中，只有最后两列（标记为N）在表示噪声时显示出激活。在此示例中，只有第二个潜在节点以“反向”方式对字形片段进行编码，即较低的激活表示片段2的存在。

外部自编码器和内部自编码器两部分：前者负责发现输入的结构，后者则找出一种以该结构为模式、具有组合性的语言。这展现了半监督 ILM 等计算模型在阐明语言信号与其认知表征之间相互作用方面的潜力。未来的研究可以在此基础上，将模型扩展到更丰富的表征领域，并探究在什么条件下，层级句法或语用推理等其他自然语言属性可能自发涌现。

结论

以往的迭代学习模型研究探讨了几种重要且有趣的语言变化现象，例如社会阶层分化、空间分布各异或语言多样化群体中的语言学习（ Brace et al., 2015; Brace and 布洛克, 2016; Sains et al., 2023; 布洛克 和 Houghton, 2024）。然而，这些研究往往局限于考察简单的意义—信号空间或人工设计语法，这两者都无法让模型探索复杂自然语言特征的涌现，例如层级句法结构、词序、一致关系和屈折变化。

本文采用了一类半监督迭代学习模型

来探索语言传递的动态过程，所研究的任务比迭代学习模型文献中以往探讨的任务更具挑战性。本文报告的结果表明，当需要围绕一组具有内部组合结构的噪声图像进行交流时，半监督迭代学习模型的智能体能够发展出自身具有表达性和组合性的语言，并成功地使其通过语言学习瓶颈传播。这为构建语言变化模拟研究奠定了重要基础，使研究者能够探索复杂语法的涌现与传播，并解释导致自然语言变化动态的多种适应过程之间的相互作用。

迭代学习模型是一个非常简单的模型，但在本文采用的半监督形式下，它支持一个关于语言起源的引人注目的假说：智能体不仅要完全致力于对共享语言中的外部话语进行编码和解码，还必须依靠自编码器驱动的学习来习得自身的“思维语言”（Vygotsky，1986；Carruthers，2002）。显然，模型的简洁性与思维语言理念的微妙性之间存在相当大的差距；为了弥合这一差距，需要将该模型应用于更复杂的例子。本文是这一计划的第一步，并带来了一个有趣的启示：同一种对语言学习有用且对语言演化至关重要的自编码器机制，对于学习路径中更外围的部分同样有用，类似于初级视觉处理。

代码与数据

所有代码均使用 Julia 1.12.3 编写，可在以下地址获取：github.com/IteratedLM/2025_05_7Seg。数据使用 Julia 生成，可在以下地址获取：github.com/IteratedLM/2025_12_7Seg_data。该代码库包含用于生成数据的 Julia 和 Python 脚本。

致谢

SB 获得了英国研究与创新署工程与物理科学研究理事会项目资助，项目编号：EP/Y028392/1：集体智能人工智能（*AI4CI*）。模拟使用布里斯托大学高级计算研究中心的计算设施完成，网址：www.bristol.ac.uk/acrc/。我们感谢斯图尔特博士的慷慨资助，他为本项目提供了部分计算资源。

参考文献

布龙菲尔德，L.（1933年）。语言。亨利·霍尔特出版社，纽约。

博克斯，C.（2021年）。《语言演化思考：从极简主义到多元主义》。语言科学出版社。

布雷斯，L. 和布洛克，S.（2016）。理解强化学习智能体重叠代中的语言演化。载于格申森，C.、弗罗泽，T.、西凯罗斯，J. M.、阿吉拉尔，W.、伊斯基耶多，E. J. 和佐山，H. 编，第十五届人工生命国际会议论文集，第492–499页。麻省理工学院出版社。

布雷斯，L.、布洛克，S. 和 Noble，J.（2015）。在迭代学习者群体中实现组合语言。载于 Andrews，P.、Caves，L.、Doursat，R.、Hickinbotham，S.、Polack，F.、Stepney，S.、Taylor，T. 和 Timmis，J. 编，*ECAL 2015*：第十三届欧洲人工生命会议论文集，第349–356页。麻省理工学院出版社。

Bullock, S. 和 Houghton, C.（2024）。使用迭代学习模型对语言接触进行建模。载于 Faíña, A.、Risi, S.、Medvet, E.、Støy, K.、Chan, B.、Miras, K.、Zahadat, P.、Grbic, D. 和 Nadizar, G.（编），*ALIFE 2024*：*2024* 年人工生命会议论文集，第 54–61 页。麻省理工学院出版社。

Bunyan, J.、布洛克, S. 和霍顿, C.（2025年）。一种结合监督学习和无监督学习的语言变化迭代学习模型。*PLOS* 复杂系统，2(3):e0000030。

Cang, J. 和 Feldheim, D. A.（2013）。拓扑映射形成与对齐的发育机制。神经科学年度评论，36:51–77。

Carruthers, P.（2002年）。语言的认知功能。行为与脑科学，25(6):657–674。

Christiansen, M. H. 和 Chater, N.（2008）。语言如何由大脑塑造。行为与脑科学，31(5):489–558。

Christiansen, M. H. 和 Kirby, S.，编者（2003）。《语言演化》。牛津大学出版社，牛津。Clark, E. H.（1929）。呼叫指示系统的演化。《贝尔实验室记录》，8(5):171–173。Deacon, T. W.（1997）。《符号物种：语言与大脑的共同演化》。W. W. Norton，纽约。Houghton, C.（2025年）。不可能的堆积？评 Cedric Boeckx 所著《语言演化思考：从极简主义到多元主义》。《*LINGUIST List*》，36.427。Kirby, S.（2001）。语言结构的自发演化：关于规则性与不规则性涌现的迭代学习模型。《*IEEE* 进化计算汇刊》，5(2):102–110。Kirby, S.（2002）。来自人工生命的自然语言。《人工生命》，8(2):185–215。Kirby, S. 和 Hurford, J.（2002）。语言结构的涌现：迭代学习模型概述。载于 Cangelosi, A. 和 Parisi, D. 编，《模拟语言演化》，第121–148页。伦敦施普林格出版社。Meister, M., Wong, R. O. L., Baylor, D. A. 和 Shatz, C. J.（1991）。发育中哺乳动物视网膜神经节细胞动作电位的同步爆发。《科学》，252(5008):939–943。Orwell, G.（1949年）。《一九八四》。Secker & Warburg，伦敦。Poeppel, D. 和 Hickok, G.（2004年）。迈向语言功能解剖学的新认识。《认知》，92(1-2):1–12。Popper, K. R.（1972年）。《客观知识：一种演化论观点》。牛津克拉伦登出版社。Popper, K. R. 和 Eccles, J. C.（1977年）。《自我与大脑：互动论辩护》。施普林格国际出版社，柏林、海德堡、纽约。Sains, G., Houghton, C. 和 Bullock, S.（2023）。空间社群结构阻碍基于群体的迭代学习模型中的语言融合。载于 Iizuka, H., Suzuki, K., Uno, R., Damiano, L., Spychalav, N., Aguilera, M., 伊斯基耶多, E., Suzuki, R. 和 Baltieri, M. 编，《人工生命会议论文集 *2023*（*ALIFE 2023*）》。麻省理工学院出版社。Shatz, C. J. 和 Stryker, M. P.（1978）。猫视觉皮层第 IV 层的眼优势及单眼剥夺的影响。《生理学杂志》，281:267–283。Steels, L.（2011年）。语言文化演化建模。《生命物理学评论》，8(2):339–356。Tomasello, M.（1999）。《人类认知的文化起源》。哈佛大学出版社，马萨诸塞州剑桥。Vygotsky, L. S.（1986）。《思维与语言》。麻省理工学院出版社，马萨诸塞州剑桥，修订版。Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V. 和 Zhou, D.（2022）。思维链提示可激发大型语言模型的推理能力。*arXiv* 预印本 *arXiv:2201.11903*。