arXiv:2602.01763v1 [cs.LG] 2 Feb 2026

可证明 ble混合线性-全注意力中的表达能力层级

Xiaowei Ye [*1] **Xiaoyu He** [*2] **Chao Liao** [2] **Chen Wu** [2] **Pinyan Lu** [2 3]

摘要

Transformer是大多数现代大语言模型的基础。为缓解标准全注意力的二次复杂度，人们提出了多种高效注意力机制，例如线性注意力和混合注意力。一个根本性的缺口仍然存在：它们相对于全注意力的表达能力缺乏严格的理论刻画。在本工作中，我们从理论上刻画了这些注意力机制之间的性能差异。我们的理论适用于所有可表述为递推的线性注意力变体，包括Mamba、DeltaNet等。具体而言，我们建立了一个表达能力层级：对于序列函数复合——一项必须在模型前向传播中完成的多步推理任务，一个$(L+1)$层全注意力网络是充分的，而任何将 $L-1$层全注意力与数量大得多的(2^{3L^2})个线性注意力层交错的混合网络都无法解决该任务。这一结果表明两类注意力在表达能力上存在明确的分离。我们的工作提供了混合注意力与标准全注意力之间的首个可证明分离结果，为理解不同注意力机制的基本能力与局限性提供了理论视角。

1. 引言

Transformer架构（Vaswani等人，2017年）是现代大语言模型（LLM）的基石，众多近期模型便是例证（OpenAI，2023年；DeepSeek-AI，2024年；MiniMax，2025b；Team 等人，2025年；Yang等人，2025a）。其核心组件是自注意力单元：在标准的全注意力机制中，它将输入元素之间的交互建模为低维嵌入之间的内积，并通过 softmax加权和计算输出。

尽管标准全注意力取得了成功，其二次计算复杂度与线性内存复杂度仍然是显著的瓶颈。因此，人们提出了各种新的注意力机制，包括用于 Mamba（Gu & Dao，2024年）、Minimax-M1（MiniMax，2025a）、RWKV（Peng 等人，2025年）和门控 DeltaNet（Yang等人，2025b）等模型的线性注意力（Katharopoulos 等人，2020年；Kasai等人，2021年；Schlag等人，2021年；Peng 等人，2021年；Yang等人，2024年）；用于混元-TurboS（Team，2025b）、通义千问3-Next（QwenTeam，2025）和 Kimi Linear（Team，2025a）等模型的线性-全混合注意力；以及对数线性注意力（Guo 等人，2025年）和稀疏注意力机制（Guo 等人，2019年；Qiu 等人，2020年；Zaheer 等人，2020年；Beltagy 等人，2020年；Guo 等人，2022年；Yuan 等人，2025年；Lu 等人，2025年）。一个自然的问题随之出现：这些注意力机制能否比全注意力表现得更好？本文旨在通过对这些注意力机制与全注意力进行理论比较来回答这一问题。

我们首先分析了采用混合注意力机制的 Transformer 在（Chen 等人，2025年）提出的序贯函数复合任务上的表现。我们的结果表明，即使线性注意力层的数量相对于全注意力层的数量呈指数增长，性能增益仍然十分有限。

此外，我们考察了稀疏注意力机制。我们在单层稀疏注意力设定下，为 2-Sum 任务建立了一项困难性结果，从而给出了稀疏注意力与全注意力之间的首个可证明分离结果。

1.1. 我们的结果

考虑一个具有 H 个注意力头、注意力头维度为 d、精度为 p、提示长度为 n 的 Transformer。我们针对一类混合注意力机制建立了如下下界。

定理 1.1。 对于任意 L，只要 $Hdp \le n^{2^{-4L-2}}$，一个$(L-1, 2^{3L^2}, \cdots, 2^{3L^2})$混合 *Transformer* 就无法解决 L-顺序函数组合问题。

[*]同等贡献；工作于华为实习期间完成 [1]´巴黎综合理工学院，帕莱索，法国 [2]华为Taylor实验室，华为 [3]计算与经济学交叉研究重点实验室，上海财经大学，上海，中国。通讯作者：Xiaoyu He <hexiaoyu12@huawei.com>, Pinyan Lu<lu.pinyan@mail.shufe.edu.cn>。

预印本。*2026*年*2*月*3*日。

我们的分析涵盖了广泛的一类可用递归形式表述的线性注意力机制，包括 Mamba（Gu & Dao，2024年）、Minimax-M1（MiniMax，2025a）、RWKV（Peng 等人，2025年）和门控 DeltaNet（Yang等人，2025b），从而建立了一个通用的比较框架。

形式化定义如下：一个 $(L,\ a_1,\ \cdots,\ a_L)$-混合变换器和 L-序列函数复合分别见定义 2.3 和 2.5。结合（Chen 等人，2025年）中的发现，我们的结果表明，加入线性注意力层并不能带来显著的性能提升。详细比较见表1。

表 *1*。 不同架构下 L-FuncComp 的复杂度。

模型/任务	L-序列组合
全注意力 - L 层	Ω(多项式 n)(Chen 等人，2025年)
全 - $L+1$ 层	O(多项式对数 n) (Chen 等人，2025年)
混合 - $(L-1,\ 2^{3L^2})$ 层	Ω(多项式 n)(定理 1.1)

我们对本文讨论的 L-序贯函数组合任务作出一些评论。该任务形式化地刻画了多步推理（例如多跳检索）的本质，而这类推理必须在模型的前向传播过程中完成：解决方案需要按顺序组合函数，其中一步的输出定义了下一步的输入上下文。我们的理论实例在每一步都采用经过精心约束的“检索范围”，以便进行严格分析。这是一项理论性任务，我们的结果在现实世界中的意义仍需通过实践验证。

我们还分析了稀疏注意力机制。

定理 1.2. 任何能解决 2-*Sum* 的单层(B, k)-稀疏注意力，其求解 2-*Sum* 所需的计算量必须满足 $Hdp = \Omega(B \log n)$；而单层全注意力只需 $H = 1, d = 3$ 和 $p = \log n$ 即可解决该问题。

该结果展示了大规模B之间的严格效率差距。$(B,\ k)$-稀疏注意力机制和2-Sum 问题分别在定义2.4 和2.6中进行了形式化定义。我们的证明采用基于通信复杂度的方法。表2概述了我们的结果。

表 *2*。 2-Sum 的全注意力与稀疏注意力复杂度。

模型	复杂度
Full	O(对数 n) （Sanford 等人，2023年）
稀疏（块）	$\Omega(B \log n)$

1.2. 组织与概述

第 2 节介绍数学背景，包括注意力机制的正式定义以及所研究的任务：序列函数复合和 2-Sum。第3 节介绍用于解决序列函数复合问题的混合注意力机制通信协议。第 4 节中，我们改编了（Chen 等人，2025年）中的方法，推导出混合架构的下界，从而证明定理1.1。最后，第 5 节证明定理 1.2，阐明稀疏注意力的理论局限。

1.3. 相关工作

Transformer 与 RNN 的比较。 关于 Transformer 与循环神经网络之间的理论比较，已有若干研究成果（Elman，1990年）。Jelassi 等人，2024年）证明了循环神经网络与 Transformer 在重复长序列方面存在表征差距。（Sanford 等人，2023年）证明了循环神经网络与 Transformer 在稀疏平均任务上具有线性—常数分离结果，随后又证明了在k跳归纳注意力头任务上存在分离结果（Sanford 等人，2024a）。

（Bhattamishra 等人，2024年）在索引查找、字符串相等性、最近邻和联想回忆等任务上，建立了 Transformer 与循环神经网络之间的计算分离结果。此外，他们证明了单层 Transformer 在有界 Dyck 语言上的线性下界（ Hahn，2020年），从而形成了分离结果，因为常数规模 RNN 可以解决该任务（Bhattamishra 等人，2020年；Hewitt 等人，2020年）。然而在实践中，单层 Transformer 很少使用。已有研究表明，多层 Transformer 能够在有界 Dyck 语言上取得成功（Wen 等人，2023年；Yao 等人，2021年），这表明 Transformer 在语言识别方面能够达到与循环神经网络相当的水平，同时在检索方面表现出色。

Wen 等人（Wen 等人，2025年）的后续研究将比较范围扩展到加入思维链（CoT）的单层 RNN。尽管思维链提升了 RNN 的性能，但仍无法弥合其与 Transformer 之间的表示能力差距：在温和的复杂性假设下，带思维链的 RNN 严格强于普通 RNN，但在算法问题上仍比增强了思维链的 Transformer 弱几个数量级。（Wen 等人，2025年）还指出，基于检索增强生成的改进可以实现图灵完备性，为缩小这一差距提供了途径。一种提出的方法采用 RNN-Transformer 混合机制，该机制已应用于混元-TurboS（团队，2025b）、Qwen3-Next（QwenTeam，2025年）和 Kimi Linear（团队，2025a）等模型中。

我们对循环模型与基于 Transformer 的模型在表达能力方面的比较作出了一些贡献。具体而言，我们针对序列函数复合任务，为混合架构建立了一项新的分离结果。

Transformer 的理论研究。关于 Transformer 的理论研究主要沿着两个方向发展：表达能力和固有局限性。表达能力方面的研究始于（P´erez 等人，2019年）对其模拟图灵机能力的考察，随后又开展了关于

对抗鲁棒性（Hsieh 等人，2019年）、通用逼近（Yun 等人，2020年；Wei 等人，2022a）以及图灵完备性（P´erez 等人，2021年；Wei 等人，2022a）的研究。

关于局限性，Hahn 2020年开创了这一研究方向，证明硬注意力 Transformer 无法识别奇偶性或 Dyck 语言。后续研究通过单层模型的通信复杂度（Peng 等人，2024年；Sanford 等人，2023年；2024a；b）或电路复杂度猜想（Sanford 等人，2024a；Peng 等人，2024年；Merrill 和 Sabharwal，2023年）进一步揭示了其局限性。Chen 等人，2025年取得了一项重大进展：他们引入不可区分分解的概念，证明了首个针对多层 Transformer 的无条件下界（针对序贯函数复合）。本文将他们的技术加以改造，从而为混合架构推导出一个下界。

思维链（CoT）。思维链（CoT）（Wei 等人，2022b）通过引导模型生成逐步推理轨迹来增强推理能力，从而为 Transformer 提供扩展的计算工作空间。这种扩展带来了显著的表达能力优势：已知具有 CoT 的常数规模 Transformer 能够模拟任何多项式时间算法（P´erez 等人，2021；Merrill 和 Sabharwal，2024；Feng 等人，2023；Li 等人，2024；Li 和 Wang，2025）。鉴于人们认为 Transformer（不使用 CoT）受限于复杂度类 TC^0（Merrill 和 Sabharwal，2023），这些结果表明（假设 $\mathrm{P} \not\subset \mathrm{TC}^0$ 等标准猜想成立），CoT 严格提升了 Transformer 的计算能力。值得注意的是，Chen 等人（2025）近期证明了首个关于 CoT 的无条件分离结果，并提出了一种用于此类结果的关键证明技术。他们证明，带有 CoT 的单层小型 Transformer 可以解决 L-序贯函数复合任务，而 L 层小型 Transformer 则无法解决。

与这些表达能力结果相辅相成，另一条研究路线确立了特定任务所需思维链步骤数量的下界。针对单层 Transformer 在迭代函数复合上的下界已被证明（Peng 等人，2024年）。其他工作将所需步骤数与 Ehrenfeucht-Haussler 秩等结构复杂度度量联系起来（Barcelo 等人，2025年），或在受限注意力模式下为各类问题（例如奇偶性、乘法、中位数、图可达性）建立了界（Amiri 等人，2025年）。

稀疏注意力机制。 全注意力的二次复杂度给长上下文处理带来了显著的计算瓶颈。稀疏注意力机制利用注意力矩阵中固有的稀疏性（Ge 等人，2024年；Jiang 等人，2023年），已成为提升效率的关键方法。这类机制已应用于 MoBA（Lu 等人，2025年）、NSA（Yuan 等人，2025年）、DSA（DeepSeek-AI，2025年）等众多模型中。

尽管稀疏注意力因计算效率高而得到广泛应用，我们仍揭示了（单层）稀疏注意力的一项根本局限：对于要求对所有输入词元进行均匀注意的任务，任何稀疏机制都可证明弱于全注意力。我们为基于压缩和选择的稀疏策略给出了形式化下界；根据（Yuan 等人，2025年）的分类，这两类策略属于三种典型类别中的两种（第三类滑动窗口在功能上类似于循环神经网络层）。

2. 预备知识

我们首先形式化本研究的关键组成部分。本节首先介绍我们所分析的注意力机制：全注意力、线性注意力、对数线性注意力以及混合架构。随后，我们正式定义主要任务：函数求值、置换复合和序列函数复合。

符号约定。 对于整数 $n \geq 0$ 与 $n_2 \geq n_1$，我们使用符号 $[n] = \{1, 2, \ldots, n\}$ 与 $[n_1 : n_2] = \{n_1, n_1+1, \ldots, n_2\}$，并约定 $[0] = \varnothing$。

2.1. 采用全注意力的 Transformer

我们考虑仅解码器 Transformer 架构。遵循（Chen 等人，2025年）中的符号约定，设 L 为注意力层的数量，H 为每层注意力头的数量，p 为精度，d 为模型维度，并且n 为（输入）提示长度。通常假设 $Hdp \geq O(\log(n))$。

一个 L 层的仅解码器 Transformer 由交替排列的注意力层和 MLP 层组成：

$$f_{\text{tran}} = f^{(L)}_{\text{mlp}} \circ f^{(L)}_{\text{attn}} \circ \cdots \circ f^{(1)}_{\text{mlp}} \circ f^{(1)}_{\text{attn}}$$

给定输入序列 $x^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)}) \in (\mathbb{R}^{dH})^n$，Transformer 以递归方式计算第 ℓ 个注意力层 $y^{(\ell)} = (y_1^{(\ell)}, \ldots, y_n^{(\ell)})$ 的输出，以及第ℓ个 MLP 层 $x^{(\ell)} = (x_1^{(\ell)}, \ldots, x_n^{(\ell)})$ 的输出。对于层 $\ell \in [L]$。

• **注意力层** $f^{(\ell)}_{\text{注意力}}$: 对于每个注意力头 $h \in [H]$和位置 $i \in [n]$，输出的计算方式如下

$$y_i^{(\ell,h)} = \sum_{j \leq i} \alpha_{i,j}^{(\ell,h)} V^{(\ell,h)} x_j^{(\ell-1)} \in \mathbb{R}^d \tag{1}$$

其中，$\{\alpha_{i,j}^{(\ell,h)}\}_{j \leq i}$ 是第 h 个注意力头的注意力分数，由 softmax运算 得出

$$\alpha_{i,j}^{(\ell,h)} = \frac{\exp((x_i^{(\ell-1)})^\top (Q^{(\ell,h)})^\top K^{(\ell,h)} x_j^{(\ell-1)})}{\sum_{j \leq i} \exp((x_i^{(\ell-1)})^\top (Q^{(\ell,h)})^\top K^{(\ell,h)} x_j^{(\ell-1)})}.$$

这里，$Q^{(\ell,h)}$，$K^{(\ell,h)}$，$V^{(\ell,h)} \in \mathbb{R}^{d \times dH}$ 表示第 h 个注意力头的查询、键和值矩阵，每个元素均采用 p位精度表示。

最后，第 ℓ 个注意力层的输出是各个注意力头的拼接结果，

$$y_i^{(\ell)} = (y_i^{(\ell,1)}, \ldots, y_i^{(\ell,H)}) \in \mathbb{R}^{dH} \quad \forall i \in [n]$$

• **多层感知机层** $f_{(\ell)\ \text{mlp}}$：第 ℓ 层的输出（同时也是第 $(\ell+1)$ 层的输入）是任意函数$g_{(\ell)}$： $\mathbb{R}^{2dH} \to \mathbb{R}^{dH}$ 按位置应用：

$$x_i^{(\ell)} = g^{(\ell)}(x_i^{(\ell-1)}, y_i^{(\ell)}) \in \mathbb{R}^{dH}.$$

请注意，这里我们修改了多层感知机层的定义，以适应各种类型的 Transformer架构和残差连接（Zhu等，2025年；Xie等，2025年）。

2.2. 线性注意力、循环神经网络和混合注意力

对于线性注意力，注意力概率变为

$$\alpha_{i,j}^{(\ell,h)} = \frac{\varphi(Q^{(\ell,h)}x_i^{(\ell-1)})^\top \varphi(K^{(\ell,h)}x_j^{(\ell-1)})}{\sum_{j\le i}\varphi(Q^{(\ell,h)}x_i^{(\ell-1)})^\top \varphi(K^{(\ell,h)}x_j^{(\ell-1)})}$$

其中 φ： $\mathbb{R}^d \to \mathbb{R}^d$ 是任意函数。线性注意力的这一形式导致线性运行时间复杂度。关键在于假设我们维护累积状态

$$S_i^{(\ell,h)} = S_{i-1}^{(\ell,h)} + V^{(\ell,h)}x_i^{(\ell-1)} \otimes \varphi(K^{(\ell,h)}x_i^{(\ell-1)}),$$
$$Z_i^{(\ell,h)} = Z_{i-1}^{(\ell,h)} + \varphi(K^{(\ell,h)}x_i^{(\ell-1)})$$

有了 $S_0^{(\ell,h)} = 0$和 $Z_0^{(\ell,h)} = 0$，我们得到

$$y_i^{(\ell,h)} = \frac{\varphi(Q^{(\ell,h)}x_i^{(\ell-1)})^\top S_i^{(\ell,h)}}{\varphi(Q^{(\ell,h)}x_i^{(\ell-1)})^\top Z_i^{(\ell,h)}}.$$

事实上，线性注意力可以视为循环神经网络。

定义 2.1（循环神经网络（RNN））。一个 RNN 层以序列 $x = (x_1, \cdots, x_n) \in (\mathbb{R}^d)^n$ 作为输入，并生成输出序列 $y = (y_1, \ldots\ldots, y_n) \in (\mathbb{R}^d)^n$，其计算方式如下。先选择 $\mathrm{h}_0 \in \mathbb{R}^m$，然后递归计算 $\mathrm{h}^i = g_{(i)}(x_i, \mathrm{h}_{i-1})$ 和 $y_i = f_{(i)}(x_i, \mathrm{h}_i)$，其中 $i = 1, \ldots\ldots, n$，且 $g_{(i)}$： $\mathbb{R}^{d+m} \to \mathbb{R}^d$ 和$f_{(i)}$： $\mathbb{R}^{d+m} \to \mathbb{R}^d$ 是任意函数。

这些 h_i 称为 RNN 层的隐藏状态。两个关键因素决定了 RNN 的容量：隐藏维度 m，如定义2.1中所定义，以及精度 p，也就是说， h_i 由 p位数字表示。

引理 2.2（线性注意力作为循环神经网络）。维度为 d、精度为 p的单层单注意力头线性注意力可视为隐藏维度为 $d^2 + d$、精度为 p的*RNN*层。更一般地，注意力头数为 H、层数为 L、维度为 d、精度为 p的线性注意力可视为层数为 L、隐藏维度为 $H(d^2 + d)$、精度为 p的多头多层循环神经网络。

混合注意力架构的定义如下。

定义 2.3（混合 Transformer 架构）。一个 $(L, a_1, \cdots, a_L)$-混合 Transformer 是一个 $(L + a_1 + \cdots + a_L)$ 层的 Transformer，由 L 个全注意力层组成，每个全注意力层后分别接有 $a_1, \cdots, a_L$ 个线性注意力层。即，

$$(T_{linear}^{(L,a_L)} \circ \cdots \circ T_{linear}^{(L,1)} \circ T_{softmax}^{(L)})$$
$$\circ \cdots \circ (T_{linear}^{(1,a_1)} \circ \cdots \circ T_{linear}^{(1,1)} \circ T_{softmax}^{(1)}).$$

2.3. 稀疏注意力

我们考虑采用块压缩和选择策略的稀疏注意力机制。

定义 2.4. 单层单头 (B, k)-稀疏注意力定义如下：将输入词元划分为大小为 B 个词元的块，然后应用压缩映射 $f: (\mathbb{R}^d)^B \to \mathbb{R}^d$ 以获得压缩词元$x_{[1:B]}, \cdots, x_{[tB+1,(t+1)B]} \in \mathbb{R}^d$，其中 $t = \lfloor \frac{i}{B} \rfloor$。

随后应用块选择函数 g： $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$，为每个块计算其相对于当前词元的分数。选取得分最高的 k 个块。设 $s_1, \cdots, s_k$ 为这些选定块的索引，对应的词元范围为 $[s_1 B: (s_1 + 1)B], \cdots, [s_k B: (s_k + 1)B]$。然后，我们计算压缩输出和选定输出

$$y_i^{compress} = \sum_{(j+1)B \le i} \alpha_{i,j} V x_{[jB+1:(j+1)B]}$$
$$y_i^{select} = \sum_{k' \in [k]} \sum_{j \in [s_{k'}B+1:(s_{k'}+1)B]} \alpha_{i,j} V x_{[jB+1:(j+1)B]}$$

且最终输出通过加权组合计算得到

$$y_i = \lambda y_i^{compress} + (1-\lambda) y_i^{select}.$$

2.4. 任务

我们正式定义本文分析的任务：序列函数复合（Chen 等人，2025年）和 2-Sum。

定义 2.5（L-序列函数复合）。给定一个整数 $L \ge 2$，L-序列函数复合任务记为 L-FuncComp$(w, z_0, z_1, \ldots\ldots, z_L)$，其由正整数 $m, n_1, \cdots, n_{L-1}$、一系列函数$z_0, z_1 \ldots, z_L$ 以及查询 $w = (w_1, \ldots\ldots, w_{L-1}) \in [n_1] \times \cdots \times [n_{L-1}]$。其中，$z_0 \in [m]$ 是初始输入，且对于满足 $\ell \in [L]$ 的 $z_\ell \in A_\ell := \{[N_{\ell-1}] \to [N_{\ell-1}]\} \simeq [N_{\ell-1}^{N_{\ell-1}}]$

$$N_\ell = m \cdot \prod_{\ell' \in [\ell]} n_\ell \quad \forall \ell \in [0 : L-1]. \tag{2}$$

计算 $i_0 = z_0 \in [m]$，$i_1 = z_1(i_0) \in [N_0]$，并对 $\ell = 1, 2, \ldots\ldots, L-1$进行归纳： $i_{\ell+1} = z_{\ell+1}(w_\ell, i_\ell) \in [N_\ell]$。最终输出为 L-函数复合$(w, z_0, z_1, \ldots\ldots, z_L) = i_L$。

为解决 L-序贯函数复合任务，我们假设 Transformer 接收以下格式的输入提示：首先列出 L 函数 z_L, z_{L-1}、……、 z_0，随后给出查询 w。为简单起见，我们假设函数 z_ℓ 的每个条目（对于 $\ell \in [0: L-1]$）都由单个词元编码（因此 z_ℓ需要 $N_{\ell-1}$ 个词元），查询 w 也由单个词元编码。

定义 2.6（2-Sum 任务）。给定输入序列 $(x_i)_{i\in[n+1]} \in [M]^{n+1}$，其中 $M = O(n)$，目标是输出序列$(y_i)_{i\in[n]}$，其中

$$y_i = \begin{cases} 1, & \exists j < i, x_i + x_j \equiv 0 \mod M \\ 0, & \text{otherwise} \end{cases}.$$

3. 混合通信模型

为证明定理 1.1，我们引入一种 $(L, a_1, \cdots, a_L)$-混合通信模型，用于 L-顺序函数组合，该模型扩展了（Chen 等人，2025年）的框架。模型由 $L+2$ 名参与者组成，每人持有以下内容之一：函数、初始输入或查询。通信分为 L 轮次。在每一轮次中，先用一轮模拟一个全注意力层，随后用若干轮实现后续的线性注意力层。

（L，a_1，$\cdots$，a_L）-混合模型

设置。模型运行 L 轮次，共有 $L+2$ 个参与者，索引为 $[-1: L]$。

输入。参与方 $i \in [L]$ 接收 z_i (根据定义2.5)，编码为 $m_{(i)} = N_{i-1}$ 个词元。参与方 $i \in [L]$ 0和参与方 -1 分别接收 z_0 和 w，且各自编码为单个词元（即 $m_{(-1)} = m_{(0)} = 1$）。

通信。对于 $\ell \in [0: L]$，令 $X_i^{(\ell)}$ 表示参与方 $i \in [-1: L]$ 在经过 ℓ 轮次后收集的消息。对于 $\ell \in [L]$，通信的第 ℓ 轮次按如下方式进行。对于参与方 $i \in [-1: L]$，

- 参与方 i 将其信息 $X_{(\ell-1)\,i}$ 发送给所有参与方 $[i+1: L]$。
- 每个参与方 $j \in [i+1: L]$ 根据自身信息 $X_{(\ell-1)\,j}$ 以及来自参与方 i 的信息 $X_{(\ell-1)\,i}$，向参与方 i 发送消息 $\Pi^{(\ell)}_{j,\,i}$（称为“软转录”，对应于全注意力层）。消息长度满足

$$|\Pi^{(\ell)}_{j,i}| = 2Hdp \cdot m_{(i)}.$$

- 参与方 i 中间更新其信息收集信息的提取方式，例如

$$X_i^{(\ell),0} := X_i^{(\ell-1)} \cup \bigcup_{j>i} \Pi^{(\ell)}_{j,i}.$$

- 随后，参与者进行 a_ℓ 轮次的通信（对应于线性注意力层）。对于每一轮次 $m \in [0: a_\ell - 1]$，参与方 L 发送消息 $\Sigma_{(\ell,m)\,L}$，其长度为$Hd(d+1)p$ 比特。依次地，每个参与方 $i = L-1, \cdots, 1,0$接收来自参与方 $i+1$的消息$\Sigma_{(\ell,m)\,i+1}$（我们称之为“线性通信记录”，对应于线性注意力层），并更新其信息为

$$X_i^{(\ell),m+1} := X_i^{(\ell),m} \cup \Sigma_{i+1}^{(\ell),m},$$

然后向参与方 $i-1$发送一条长度为 $Hd(d+1)p$比特的消息 $\Sigma_i^{(\ell,m)}$。最后，参与方 -1 更新其信息为

$$X_{-1}^{(\ell),m+1} := X_{-1}^{(\ell),m} \cup \Sigma_0^{(\ell),m}.$$

经过 a_ℓ 个线性轮后，该轮次结束，即对所有 $i \in [-1, L]$设置 $X_i^{(\ell)} = X^{(\ell)}_{\cdot,i}{}^{a_\ell}$。

输出。在第 L轮次之后，参与方 -1 基于 $X_{-1}^{(L)}$ 产生最终输出。

我们指出，这些参与者是健忘的，参与方 j 不记得先前参与者发送的任何内容。

4. 混合通信下界

我们展示了将全注意力与高效线性注意力层相结合的混合 Transformer 架构在求解深度序列组合任务上的局限性。我们的主定理表明，即使充足的线性注意力也无法替代全注意力在某些层次任务上的表达能力。

我们证明定理1.1 借助通信复杂度论证。我们证明的核心是对不可区分输入集的归纳构造，尽管这些集合需要不同的输出，模型却无法将它们区分开。

我们从各层之间信息流的角度分析混合 Transformer 的计算。每个全注意力层都能对来自所有先前位置的信息进行并行聚合，而线性注意力层仅允许顺序的循环传播。后者无法形成深度组合所需的丰富交互。

L-序贯函数复合任务专门

旨在控制线性注意力层对通信记录空间的贡献，并满足技术要求。任务参数的具体选择见附录B。

该证明技术为分析结构化 Transformer 中的层级计算提供了一个通用框架。它表明，尽管线性注意力具有效率优势，但其表达能力不足以支持深层的组合推理，即使加入了有限的全注意力也依然如此。

本节余下部分将概述我们的证明；详细证明见附录B。

4.1. 参数与策略

为证明定理 1.1，我们构造该任务的参数，使得输入长度满足 $n \le (Hdp)^{4\cdot 16^L}$，但该任务无法由一个 $L-1(L-1,\ 2^{3L^2},\ \cdots,\ 2^{3L^2})$-混合 Transformer 解决。我们假设 $Hdp \ge 2$。

事实上，我们证明了一个更强的结果：即使允许先由单层 Transformer 进行预处理，L-序贯函数复合任务仍无法由一个小型的（$L-1,\ 2^{3L^2},\ \cdots,\ 2^{3L^2}$）-混合 Transformer 解决。等价地，我们证明了一个小型的（$L,\ 0,\ 2^{3L^2},\ \cdots,\ 2^{3L^2}$）-混合 Transformer 无法解决 L-序贯函数复合。

符号约定。为方便书写，我们交替使用 z_{-1} 和 w 来表示参与方 -1 的输入。下面我们将阐述若干对于证明至关重要的关键定义。

- （软转录 $\Pi^{(\ell)}_{j,i}$）对于任意 $i \in [-1:\ L-1]$，$j \in [i+1:\ L]$，$\ell \in [L]$，回顾一下，$\Pi^{(\ell)}_{j,i}$ 表示在第 ℓ 个通信轮次中从参与方 j 发送给参与方 i 的软转录。其值由参与者 $[i:\ L]$（即 $z_L, \ldots, z_i$）的输入决定，并且独立于参与者 $[-1:\ i-1]$（即 $z_{i-1}, \ldots, z_{-1}$）的输入。对于任意固定输入 $\widetilde{z}_L \in [N_{L-1}]^{N_{L-1}}, \ldots, \widetilde{z}_i \in [N_{i-1}]^{N_{i-1}}$，令 $\Pi^{(\ell)}_{j,i}(\widetilde{z}_L, \ldots, \widetilde{z}_i)$ 表示参与方 t 接收输入 $z_t = \widetilde{z}_t (t \in [i:\ L])$ 时的软转录。

- (线性转录 $\Sigma^{(\ell)}_{i+1}{}^{m}$) 对于任意 $i \in [-1, L-1], l \in [L]$ 和 $m \in [0, a_\ell - 1]$，回顾一下，$\Sigma^{(\ell)}_{i+1}{}^{m}$ 是参与方 $i+1$ 在第 ℓ 个通信轮次的第 $m+1$ 个线性轮中发送给参与方 i 的线性转录。其值由参与者 $[i+1$ 的输入决定：$L]$（即 z_L、……、z_{i+1}），且与参与者 $[-1$ 的输入无关：$i]$（即 z_i、……、z_{-1}）。对于任意固定的输入 $\widetilde{z}_L \in [N_{L-1}]^{N_{L-1}}$、……、$\widetilde{z}_i \in [N_{i-1}]^{N_{i-1}}$，令 $\Sigma^{(\ell)}_{i+1}{}^{m}(\widetilde{z}_L$、……、$\widetilde{z}_i)$ 表示参与方 t 接收输入 $z_t = \widetilde{z}_t$（$t \in [i:\ L]$）时的转录。

- (部分组合值) 对于任意 $\ell \in [0:\ L]$，i_ℓ 的值由 w，z_0、……、z_ℓ 决定。我们用 $i_\ell(\widetilde{w}$，$\widetilde{z}_0$、……、$\widetilde{z}_\ell)$ 表示当 $w = \widetilde{w}$，$z_0 = \widetilde{z}_0$、……、$z_\ell = \widetilde{z}_\ell$ 时它的值。

不可区分分解。我们证明的核心概念是 Chen 等人，2025年提出的不可区分分解。不可区分分解由两个集合 $R_{\ge\ell}$ and $Z_{<\ell}$ 构成，其中 $R_{\ge\ell}$ 是参与方 $[\ell$ 的输入赋值集合：$L]$；$Z_{<\ell}$ 是参与方 $[-1$ 的输入赋值集合：$\ell-1])$。其关键性质是，对于第一个 ℓ 个参与方的任意固定输入 $z_{<\ell} \in Z_{<\ell}$，在经过 ℓ 轮次后，$R_{\ge\ell}$ 的所有赋值对于输入为 $z_{<\ell}$ 的参与方 $[-1$ 都不可区分，因为它们会产生完全相同的通信记录。形式上，我们将 Chen 等人，2025年中的定义改写如下。

定义 4.1（不可区分分解）。设 $\ell \in [2:\ L]$，不可区分分解是满足以下条件的一对集合 $R_{\ge\ell} \subseteq A_L \times A_{L-1} \times \cdots \times A_\ell$ 和 $Z_{<\ell} = Z_{-1} \times \cdots \times Z_{\ell-1}$：$Z_{-1} = A_{-1}$，$Z_0 \subseteq A_0$，$\cdots$，$Z_{\ell-1} \subseteq A_{\ell-1}$，使得对于每个 $\widetilde{z}_{<\ell} \in Z_{<\ell}$，以及每个 $\widetilde{\alpha}_{\ge\ell}$，$\widetilde{\beta}_{\ge\ell} \in R_{\ge\ell}$，均有：

$$\Pi^{(\ell')}_{j,i}(\widetilde{z}_{<\ell}, \widetilde{\alpha}_{\ge\ell}) = \Pi^{(\ell')}_{j,i}(\widetilde{z}_{<\ell}, \widetilde{\beta}_{\ge\ell})$$

对于每个 $j \in [\ell:\ L]$、$i \in [-1:\ \ell-1]$ 和 $\ell' \in [\ell]$，以及

$$\Sigma^{(\ell'),m}_{i+1}(\widetilde{z}_{<\ell}, \widetilde{\alpha}_{\ge\ell}) = \Sigma^{(\ell'),m}_{i+1}(\widetilde{z}_{<\ell}, \widetilde{\beta}_{\ge\ell})$$

对于每个 $i \in [-1:\ \ell-1]$、$\ell' \in [\ell]$ 和 $m \in [0, a_{\ell'} - 1]$。

当 $\ell = L$ 时，不可区分分解的用途便显而易见。在这种情况下，对于从 $Z_{<L}$ 到参与者 $[-1:\ L-1]$ 的每个输入赋值，参与者 -1（最终输出参与方）在经过 L 轮次后（即协议结束时）观察到的通信记录完全相同，无论将哪个输入 $\widetilde{z}_L \in R_{\ge L}$ 分配给参与者 L。因此，对于每个 $\widetilde{z}_{<L} \in Z_{<L}$，协议 L-FuncCo mp($\widetilde{z}_{<L}$，$\widetilde{z}_L$) 的输出必须对每个$\widetilde{z}_L \in R_{\ge L}$都相同。我们将仔细定义集合 $R_{\ge\ell}$ 和 $Z_{<\ell}$，使得满足这一要求会导致矛盾，从而确立所需的下界。

对于子集 $Z_{<\ell}$，我们将 $\mathcal{I}_{\ell-1}(Z_{<\ell})$ 定义为：当参与者 $[-1$ 的输入 $\ell-1]$ 限制为 $Z_{<\ell}$ 时，在第 $\ell-1$ 轮次之后中间复合的所有可能值构成的集合：

$$\{i_{\ell-1}(\widetilde{z}_{-1}, \widetilde{z}_0, \ldots, \widetilde{z}_{\ell-1}) : \widetilde{z}_{-1}, \widetilde{z}_0, \ldots, \widetilde{z}_{\ell-1}) \in Z_{<\ell}\}.$$

如下引理（参见 Chen 等人，2025年）表明，对于 $\ell = L$，只要存在足够好的不可区分配置，就能得到所需的下界。

引理 4.2（引理 B.5）。一个 L 轮次的混合通信协议无法解决 L-FuncComp，前提是存在一个不可区分分解 $R_{\ge L}$ 和 $Z_{<L}$，使得 $|R_{\ge L}|$ 和 $|\mathcal{I}_{L-1}(Z_{<L})|$ 都很大。

由于所有 $z_L \in R_{\geq L}$ 对参与方 -1（输出参与方）而言都不可区分，因此该协议必须对每个 $z_L \in R_{\geq L}$ 输出相同的答案。然而，由于 $\mathcal{I}_{L-1}(Z_{<L})$ 很大，不同的 $z_L \in R_{\geq L}$ 会要求对 $Z_{<L}$ 中的某些输入给出不同的输出。这与协议的正确性相矛盾，从而完成证明。

定理1.1 的证明随后通过归纳论证构造所需的分解而完成。

引理*4.3*（引理 *B.7*）。对于任意 $\ell \in [2:\ L]$，我们可以通过归纳构造一个不可区分分解（$R_{\geq \ell}, Z_{<\ell}$），其中 $R_{\geq \ell} \subseteq A_L \times A_{L-1} \times \cdots \times A_\ell$ 和 $Z_{<\ell} = Z_{-1} \times Z_0 \times \cdots \times Z_{\ell-1}$，且满足 $Z_{-1} = [n_1 \cdots n_{L-1}]$，$Z_0 \subseteq A_0$ 和 $Z_i \subseteq A_i$；该分解还包含参与方 $[\ell:\ L]$ 到 $[-1:\ \ell-1]$ 在前 ℓ 轮次中的软通信记录（当参与方 $[-1:\ \ell-1]$ 接收来自 $Z_{<\ell}$ 的输入时），以及参与方 ℓ 到 $\ell-1$ 在前 ℓ 轮次中的线性通信记录（当参与方 $[-1:\ \ell-1]$ 接收来自 $Z_{<\ell}$的输入时），使得软通信记录和线性通信记录一致，并且 $|R_{\geq \ell}|$ 和 $|\mathcal{I}_{\ell-1}(Z_{<\ell})|$很大。

4.2. 初始步骤

我们首先考虑基本情形 $\ell = 2$。

步骤 1：选择 Z_0，Z_1。我们令 $Z_0 = [x_0]$。下一步是选择集合 $Z_1 \subseteq A_1$。考虑从玩家 1 到参与方 -1的所有可能的第一个轮次消息，此类不同消息模式的总数为 $2^{2Hdp\cdot|Z_{-1}|} = 2^{2Hdp(n_1\cdots n_{L-1}}$)。我们根据鸽巢原理选择一个消息模式 $\widetilde{\Psi}_1^{(1)}{}_{,-1} \in \{0,1\}^{2Hdp\cdot(n_1\cdots n_{L-1})}$，使得输入 $\widetilde{z}_1 \in A_1$ 的一致性集合 S 满足 $|S| \geq |A_1| \cdot 2^{-2Hdp\cdot(n_1\cdots n_{L-1})}$。然后，我们根据以下引理收回一个合适的子集 $Z_1 \subseteq S$。

引理4.4（引理 B.8）。存在一个子集 $Z_1 \subseteq S$使得 $\mathcal{I}_1(Z_{<2})$ 很大。

下一步是固定参与者 $j \in [2:L-1]$ 在前两个轮次发送给参与者 $i = -1,\ 0,\ 1$ 的通信记录。

步骤 2.1：固定发送给参与方 -1的通信记录。 我们从第一个轮次开始。我们的目标是，对每一个 $\widetilde{z}_1 \in Z_1,\ \widetilde{z}_0 \in Z_0,\ \widetilde{z}_{-1} \in Z_{-1}$，固定在第一个轮次中发送给参与方 -1 的软通信记录。参与方 j 在第一个轮次发送给参与方 -1 的消息仅依赖于 $\widetilde{z}_{-1}$ 和参与方j的输入，而独立于 $\widetilde{z}_1,\ \widetilde{z}_0$。因此，此类可能的通信记录模式总数至多为 $2^{2Hdp\cdot|Z_{-1}|} = 2^{2Hdp\cdot(n_1\cdots n_{L-1})}$。根据鸽巢原理，我们可以选择一个固定模式，使其一致性集合达到最大规模 $|C_1| \geq |A_L| \cdots |A_2| \cdot 2^{-2Hdp\cdot(n_1\cdots n_{L-1})\cdot L}$.

对于第二个轮次，目标是为每个 $\widetilde{z}_1 \in Z_1, \widetilde{z}_0 \in Z_0, \widetilde{z}_{-1} \in Z_{-1}$ 固定在第二个轮次发送给参与方 -1 的软通信记录。关键的是，来自参与方 j 的通信记录仅取决于信息状态 $X_{-1}^{(1)}$ 和 $X_j^{(1)}$，而这两者本身都独立于 $z_1 \in Z_1$ 的选择。这种独立性成立，是因为 $X_{-1}^{(1)}$ 和 $X_j^{(1)}$ 中唯一取决于 z_1 的部分，是从玩家 1 发往 -1 的第一个轮次消息，而该消息是固定的。我们选择一致性集合 C_2 大小最大的通信记录。

步骤 2.2：固定发送给参与方 0 和参与方 1。类似地，我们固定前两个轮次中发送给玩家 0 的软转录的值，使一致性集合 C_3 的大小最大。

我们采用相同的策略，固定前两个轮次中发送给玩家 1 的软转录和线性通信记录的值，使一致性集合 C_4 达到最大大小。令 $R_{\geq 2}$ 为 C_4，由此完成基本情形 $\ell = 2$ 的证明。

4.3. 归纳步骤

假设我们已经完成了直到 $\ell \in [2:\ L-1]$ 的部分。我们继续为 $\ell+1$ 进行构造。

关键洞见是，当参与方$[-1:\ \ell-1]$ 从 $Z_{<\ell}$ 获取输入时，经过 ℓ 个轮次后，Z_ℓ 对他们而言不可区分。因此，第 $(\ell+1)$ 个轮次发送给参与方$[-1:\ \ell-1]$ 的通信记录与 $z_\ell \in Z_\ell$ 的选择无关。

步骤 1：选择集合 Z_ℓ。回顾一下，$R_{\geq \ell}$ 的大小很大。我们希望从 $R_{\geq \ell}$中选取一个矩形子集。如（Chen 等人，2025年）中所述，我们有以下引理。

引理 4.5（引理 B.9）。存在一个子集 $S^{(\ell)} = S_1^{(\ell)} \times S_2^{(\ell)}$，其中 $S_1^{(\ell)} \subseteq A_L \times \cdots \times A_{\ell+1}$，$S_2^{(\ell)} \subseteq A_\ell$，使得 $S_1^{(\ell)}$ 和 $\mathcal{I}_{\ell-1}(S_2^{(\ell)})$ 很大。

我们取 $Z_\ell = S_2^{(\ell)}$ 和 $S_{\geq \ell+1} = S_1^{(\ell)}$。接下来，我们将固定通信记录。回顾一下，我们需要固定第一$\ell+1$ 轮次中参与方 $[\ell+1:\ L]$ 到参与方 $[-1:\ \ell]$ 的所有通信记录，此时参与方 $[-1:\ \ell]$ 从$Z_{\leq \ell} = Z_{-1} \times Z_0 \times \cdots \times Z_\ell$接收输入。我们分几个步骤进行。

步骤 2.1：将通信记录固定到参与者 $[-1:\ \ell-1]$在前 ℓ 个轮次。 我们直接采用归纳假设给出的通信记录。引理 B.10 保证了一致性。

步骤 2.2：将通信记录固定给参与者 $[-1:\ \ell-1]$在第 $(\ell+1)$ 轮。 我们的关键洞见是，Z_ℓ 对于参与者 $[-1:\ \ell-1]$ 而言不可区分，因为他们的输入来自 $Z_{\leq \ell-1}$。因此，他们的通信记录与 $z_\ell \in Z_\ell$的选择无关。我们考虑第 $(\ell+1)$ 轮中从参与者 $[\ell+1, L]$ 到参与者 $[-1$ 的软通信记录 $\ell-1]$

$$\Phi^{(\ell+1)} = \left(\Phi_{j,i}^{(\ell+1)}\right)_{j\in[\ell+1:L],i\in[-1:\ell-1]}$$

其中

$$\Phi_{j,i}^{(\ell+1)} = \left(\Phi_{j,i}^{(\ell+1)}(\widetilde{z}_{\ell-1},\ldots,\widetilde{z}_i)\right)_{\widetilde{z}_{\ell-1}\in Z_\ell\ldots,\widetilde{z}_i\in Z_i}$$
$$\text{and} \quad \Phi_{j,i}^{(\ell+1)}(\widetilde{z}_{\ell-1},\ldots,\widetilde{z}_i) \in \mathsf{domain}(\Pi_{j,i}^{(\ell+1)})$$

比较 $\Lambda_{j\,i}^{(\ell+1,\ell+1)}$，与 $\Phi_{j\,i}^{(\ell+1)}$，，我们注意到 $\Phi_{j\,i}^{(\ell+1)}$，与 $\widetilde{z}_\ell \in Z_\ell$无关。对于任意 $\Phi^{(\ell+1)}$，定义 $S(\Phi^{(\ell+1)})$ 为所有与通信记录 $\Phi^{(\ell+1)}$一致的 $(\widetilde{z}_L,\ \ldots\ldots,\ \widetilde{z}_{\ell+1}) \in S_{\geq\ell+1}$ 的集合。

引理 4.6 (引理 B.11)。 我们有

$$\bigcup_{\Phi^{(\ell+1)}} S(\Phi^{(\ell+1)}) = S_{\geq\ell+1}.$$

现在，如同（Chen 等人，2025年）中一样，我们取一致性集合最大的通信记录 $T_{\geq\ell+1} = S\ (\widetilde{\Phi}^{(\ell+1)}) \subseteq S_{\geq\ell+1}$。

步骤 2.3：固定转录以 ℓ。这遵循贪心选择策略。首先，考虑在前 $\ell+1$ 轮次中发送给参与方 ℓ 的软通信记录

$$\Psi = \left(\Psi_{j,\ell}^{(\ell')}(\widetilde{z}_\ell)\right)_{j\in[\ell+1:L],\ell'\in[\ell+1],\widetilde{z}_\ell\in Z_\ell}$$
$$\text{where} \quad \Psi_{j,\ell}^{(\ell')}(\widetilde{z}_\ell) \in \mathsf{domain}(\Pi_{j,\ell}^{(\ell')})$$

定义 T（Ψ）为其一致性集合。我们可以对不同 Ψ 的数量给出上界，并利用鸽巢原理对最大一致性集合的大小给出下界（引理 B.13）。

接下来，我们固定前 ℓ +1 轮次中发给参与方 ℓ 的线性通信记录，以保持最大的 一致性集合。然后我们可以令 $R_{\geq\ell+1}$ 为这个一致性集合，从而完成归纳步骤。

5. 稀疏注意力的下界

现在我们证明定理 1.2。为方便起见，我们假设 B 整除 n。证明通过将一个 (B,k)- 稀疏注意力与以下通信模型联系起来展开。

通过稀疏注意力实现 2-Sum 的模型

设置。 共有 $\frac{n}{B}+1$ 位参与者。**输入**。 前 $\frac{n}{B}$ 位参与者各自接收 B 个词元。最后一位参与者接收一个词元。**通信**。 前 $\frac{n}{B}$ 位参与者各自向最后一位参与者发送一条 Hdp 比特的消息。随后，最后一位参与者可以选择其中的 k 位参与者，并获取他们的全部信息。**输出**。 最后一位参与者根据收集到的信息给出答案。

在此模型中，每个由 B 个词元组成的区块对应前 $\frac{n}{B}$ 位参与者中的一位，最后一个词元对应最终参与者，而压缩表示（大小为 Hdp 比特）对应发送给最终参与者的消息。

可以证明，为了让最后一位参与者输出正确答案，前面的参与者必须各自通过一条 Hdp 比特的压缩消息，传递其 B 个词元组成的集合。这就推出了所需结果；详细内容见附录 A.6。

6. 结论

在本工作中，我们在通信复杂度框架内，为高效注意力机制建立了统一的表达能力层级。通过分析混合架构，我们证明了一个无条件下界：即便拥有大量线性注意力层，也无法替代单个全注意力层在求解深度序列函数复合任务时的组合能力。这些任务形式化地刻画了在单次前向传播中进行多步序列计算的需求，并涵盖多跳检索等问题，其中每一步的输出定义了下一步的输入上下文。我们的理论框架在每个内部计算步骤引入了经过精心约束的“检索范围”，以便进行严格分析。这一结果为如下经验观察提供了理论依据：简单地交错使用线性注意力与全注意力，在深度推理问题上只能带来有限收益。

此外，我们指出了基于块压缩与选择的单层稀疏注意力机制的一个根本局限。对于诸如2-Sum这类需要均匀成对比较的任务，任何此类稀疏机制在可证明意义上都弱于全注意力，除非其有效容量随块大小而扩展。

总的来说，我们的研究结果为理解高效注意力变体提供了一个有理论依据的框架。这些结果表明，未来用于长上下文处理的架构必须采取以下一种方式：通过更智能的近似方法，充分发挥全注意力的表达能力；或者明确针对这些局限性进行设计，例如采用任务充分的稀疏模式，或引入机制来绕过我们所发现的通信瓶颈。本文提出的通信模型和证明技术，或可为结构化 Transformer 的进一步理论分析奠定基础。

参考文献

Amiri, A., Huang, X., Rofin, M., and Hahn, M. 硬注意力 Transformer 中思维链推理的下界。载于第四十二届国际机器学习大会，2025年。

Barcelo, P.、Kozachinskiy, A. 和 Steifer, T. Ehrenfeucht-Haussler 秩与思维链。载于第四十二届国际机器学习大会。arxiv:

2501.12997，2025年。

Beltagy, I., Peters, M. E., 和 Cohan, A. Longformer：长文档 Transformer。*arXiv:2004.05150*，2020年。

Bhattamishra, S., Ahuja, K., 和 Goyal, N. 关于循环神经网络识别层次语言的实际能力。在第*28*届国际计算语言学会议论文集，第1481–1494页。国际计算语言学委员会，2020年12月。

Bhattamishra, S., Hahn, M., Blunsom, P., 和 Kanade, V. Transformer 与循环架构表征能力的分离结果。在第*38*届神经信息处理系统年会，2024年。

Chen, L., Peng, B. 和 Wu, H. 多层 Transformer 的理论局限性。载于 第*66*届 *IEEE* 计算机科学基础年会（*FOCS*），2025年。DeepSeek-AI。Deepseek-v3 技术报告。*arXiv*，2412.19437，2024年。DeepSeek-AI。Deepseek-v3.2-exp：利用 DeepSeek 稀疏注意力提升长上下文效率，2025年。Elman, J. L. 探索时间中的结构。认知科学，14(2):179–211，1990年。Feng, G., Zhang, B., Gu, Y., Ye, H., He, D. 和 Wang, L. 探索思维链背后的奥秘：理论视角。载于第*37*届神经信息处理系统大会，2023年。Ge, S., Zhang, Y., Liu, L., Zhang, M., Han, J. 和 Gao, J. 模型告诉你该丢弃什么：大语言模型的自适应 KV 缓存压缩。载于 第十二届国际学习表征会议，2024年。网址 https://openreview.net/forum?id=uNrFpDPMyo。Gu, A. 和 Dao, T. Mamba：采用选择性状态空间的线性时间序列建模。载于 首届语言建模会议，2024年。Guo, H., Yang, S., Goel, T., Xing, E. P., Tri, D. 和 Kim, Y. 对数线性注意力。*arXiv*，2506.04761，2025年。Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y.-H. 和 Yang, Y. LongT5：适用于长序列的高效文本到文本 Transformer。载于 计算语言学协会论文集：*NAACL 2022*，第724–736页，美国西雅图，2022年7月。计算语言学协会。

Guo, Q., Qiu, X., Liu, P., Shao, Y., Xue, X. 和 Zhang, Z. 星形变换器。载于 《*2019*年北美计算语言学协会分会年会论文集：人类语言技术》，第*1*卷（长篇与短篇论文），第1315–1325页，美国明尼苏达州明尼阿波利斯，2019年6月。计算语言学协会。

Hahn, M. 自注意力在神经序列模型中的理论局限性。计算语言学协会会刊，8:156–171，2020年。

Hewitt, J.、Hahn, M.、Ganguli, S.、Liang, P. 和 Manning, C. D. 循环神经网络能够以最优记忆生成有界层次语言。载于 *2020*年自然语言处理实证方法会议论文集，第 1978–2010 页，2020年11月。

Hsieh, Y.-L., Cheng, M., Juan, D.-C., Wei, W., Hsu, W.-L., and Hsieh, C.-J. 关于自注意力模模型的鲁棒性。在计算语言学协会第*57*届年会论文集，第1520–1529页，意大利佛罗伦萨，2019年7月。计算语言学协会。Jelassi, S., Brandfonbrener, D., Kakade, S. M., and Malach, E. 跟我重复：Transformer在复制任务上优于状态空间模型。载于第*41*届国际机器学习大会，2024年。Jiang, H., Wu, Q., Lin, C.-Y., Yang, Y., and Qiu, L. LLMLingua：压缩提示词以加速大语言模型推理。载于*2023*年自然语言处理实证方法会议论文集，第13358–13376页，新加坡，2023年12月。计算语言学协会。Kasai, J., Peng, H., Zhang, Y., Yogatama, D., Ilharco, G., Pappas, N., Mao, Y., Chen, W., and Smith, N. A. 将预训练变换器微调为循环神经网络。载于*2021*年自然语言处理实证方法会议论文集，第10630–10643页，线上及多米尼加共和国蓬塔卡纳，2021年11月。计算语言学协会。Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. Transformer就是循环神经网络：采用线性注意力的快速自回归变换器。载于第*37*届国际机器学习会议论文集，第119卷机器学习研究论文集，第5156–5165页。PMLR，2020年7月13日至18日。Li, Q. and Wang, Y. 恒定位数的Transformer是图灵完备的。载于第*39*届神经信息处理系统年会，2025年。网址：https://openreview.net/forum?id=RBWnyDEBKf。

Li, Z., Liu, H., Zhou, D., and Ma, T. 思维链赋能Transformer解决本质上具有串行性质的问题。载于第*12*届国际学习表征会议，2024年。

Lu, E., Jiang, Z., Liu, J., Du, Y., Jiang, T., Hong, C., Liu, S., He, W., Yuan, E., Wang, Y., HUang, Z., Yuan, H., Xu, S., Xu, X., Lai, G., Chen, Y., Zheng, H., Yan, J., Su, J., Wu, Y., Zhang, Y., Yang, Z., Zhou, X., Zhang, M. 和 Qiu, J. Moba：长上下文大语言模型的块混合注意力。*arXiv* 预印本 *arXiv:2502.13189*，2025年。

Merrill, W. 和 Sabharwal, A. 并行性权衡：对数精度 Transformer 的局限性。计算语言学协会汇刊，11：531–545，2023年。

Merrill, W. 和 Sabharwal, A. 具有思维链的 Transformer 的表达能力。载于第十二届国际学习表征会议，2024年。

MiniMax。Minimax-M1：通过闪电注意力高效扩展测试时计算。*arXiv*，2506.13585，2025a。

MiniMax。Minimax-M2。*GitHub*，https://github.com/MiniMax-AI/MiniMax-M2，2025b。

OpenAI。GPT-4技术报告。*arXiv*，2303.08774，2023年。

Peng, B.、Narayanan, S. 和 Papadimitriou, C.，论Transformer架构的局限性。载于首届语言建模会议，2024年。

Peng, B.、Zhang, R.、Goldstein, D.、Alcaide, E.、Du, X.、Hou, H.、Lin, J.、Liu, J.、Lu, J.、Merrill, W. 等。具有表现力动态状态演化的RWKV-7“Goose”。*arXiv*预印本 *arXiv:2503.14456*，2025年。

Peng, H., Pappas, N., Yogatama, D., Schwartz, R., Smith, N. 和 Kong, L. 随机特征注意力。载于国际学习表征会议，2021年。

P´erez, J., Marinkovi´c, J. 和 Barcel´o, P. 论现代神经网络架构的图灵完备性。载于国际学习表征会议，2019年。

P´erez, J., Barcel´o, P. P. 和 Marinkovic, J. 注意力机制是图灵完备的。机器学习研究期刊，22(75)：1–35，2021年。

Qiu, J., Ma, H., Levy, O., Yih, W.-t., Wang, S., 和 Tang, J. 面向长文档理解的分块自注意力。载于计算语言学协会论文集：*EMNLP 2020*，第 2555–2565 页。计算语言学协会，2020年11月。

QwenTeam。《Qwen3-next：迈向极致的训练与推理效率》。技术报告，https://qwen3-next.com/ - 技术文档，2025年。Sanford, C., Hsu, D., 和 Telgarsky, M. Transformer 的表征优势与局限性。载于第三十七届神经信息处理系统大会，2023年。Sanford, C., Hsu, D., 和 Telgarsky, M. Transformer、并行计算与对数深度。载于国际机器学习会议，2024a。Sanford, C., Hsu, D., 和 Telgarsky, M. 单层 Transformer 无法解决归纳头任务。*arXiv*，2408.14332，2024b。

Schlag, I., Irie, K. 和Schmidhuber, J. 线性Transformer实际上是快速权重编程器。载于第*38*届国际机器学习会议论文集，机器学习研究论文集第139卷，第9355–9366页。PMLR，2021年7月18日至24日。

Stirling, J. 《微分法》：关于无穷级数求和与插值的论著。作者 *Jacobo Stirling，R.S.S.* 由 Ric. Manby 于伦敦 Ludgate-Hill 街“王子徽记”处发行，1753年。

Team, K. Kimi Linear：一种富有表现力且高效的注意力架构。*arXiv*，2510.26692，2025a。

Team, K., Bai, Y., Bao, Y., Chen, G., Chen, J., Chen, N., Chen, R., Chen, Y., Chen, Y., Chen, Y. 等。Kimi k2：开放式智能体智能。*arXiv* 预印本 *arXiv:2507.20534*，2025年。

Team, T. H. 混元-TurboS：通过 Mamba-Transformer协同与自适应思维链推进大语言模型。*arXiv*，2505.15431，2025b。

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. 和 Polosukhin, I. 注意力就是你所需要的一切。载于神经信息处理系统进展，第 30 卷。Curran Associates, Inc., 2017年。

Wei, C., Chen, Y. 和 Ma, T. 统计有意义的近似：用 Transformer 近似图灵机的案例研究。载于神经信息处理系统进展，第 35 卷，第 12071–12083 页，2022a。

Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V. 和 Zhou, D. 思维链提示能够激发大语言模型的推理能力。载于 Oh, A. H., Agarwal, A., Belgrave, D. 和 Cho, K.（编），神经信息处理系统进展，2022b。

Wen, K., Li, Y., Liu, B. 和 Risteski, A. 使用短视方法无法解释 Transformer：有界 Dyck 文法案例研究。载于第 *37* 届神经信息处理系统国际会议论文集，NIPS '23，美国纽约州雷德胡克，2023年。Curran AssociatesInc.

Wen, K.、Dang, X. 与 Lyu, K. 循环神经网络还不是 Transformer：上下文检索的关键瓶颈。见第十三届国际学习表征会议，2025年。Xie, Z.、Wei, Y.、Cao, H.、Zhao, C.、Deng, C.、Li, J.、Dai, D.、Gao, H.、Chang, J.、Zhao, L. 等。mhc：流形约束超连接。*arXiv* 预印本 *arXiv:2512.24880*，2025年。Yang, A.、Li, A.、Yang, B.、Zhang, B.、Hui, B.、Zheng, B.、Yu, B.、Gao, C.、Huang, C.、Lv, C. 等。通义千问3技术报告。*arXiv* 预印本 *arXiv:2505.09388*，2025a。Yang, S.、Wang, B.、Zhang, Y.、Shen, Y. 与 Kim, Y. 沿序列长度用增量规则并行化线性 Transformer。见 第*38*届神经信息处理系统年会，2024年。Yang, S.、Kautz, J. 与 Hatamizadeh, A. 门控增量网络：用增量规则改进 Mamba2。见第十三届国际学习表征会议，2025b。Yao, S.、Peng, B.、Papadimitriou, C. 与 Narasimhan, K. 自注意力网络可以处理有界层次语言。见 第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长文），第3770–3785页。计算语言学协会，2021年8月。Yuan, J.、Gao, H.、Dai, D.、Luo, J.、Zhao, L.、Zhang, Z.、Xie, Z.、Wei, Y.、Wang, L.、Xiao, Z.、Wang, Y.、Ruan, C.、Zhang, M.、Liang, W. 与 Zeng, W. 原生稀疏注意力：硬件对齐且原生可训练的稀疏注意力。见第*63*届计算语言学协会年会论文集（第*1*卷：长文），第23078–23097页，奥地利维也纳，2025年。计算语言学协会。Yun, C.、Bhojanapalli, S.、Rawat, A. S.、Reddi, S. 与 Kumar, S. Transformer 是序列到序列函数的通用逼近器吗？见 国际学习表征会议，2020年。Zaheer, M.、Guruganesh, G.、Dubey, K. A.、Ainslie, J.、Alberti, C.、Ontanon, S.、Pham, P.、Ravula, A.、Wang, Q.、Yang, L. 等。Big Bird：面向更长序列的 Transformer。神经信息处理系统进展，33，2020年。

Zhu, D.、Huang, H.、Huang, Z.、Zeng, Y.、Mao, Y.、Wu, B.、Min, Q. 和 Zhou, X.《超连接》。载于第十三届国际学习表征会议，2025年。

A. 线性注意力、对数线性注意力和稀疏注意力的下界

本节将介绍不同注意力机制在各种任务上的通信模型。我们将通过线性注意力和对数线性注意力解决此类任务的过程抽象为某些通信协议，并通过通信复杂度论证证明定理A.3、A.6、A.5、A.8和A.9 中的下界。

A.1. 定义

A.1.1. 对数线性注意力

在（Guo 等人，2025年）提出的对数线性注意力机制中，第 ℓ个注意力层的输出变为

$$y_i^{(\ell,h)} = \sum_{r=0}^{R-1} \lambda_i^{(r,\ell,h)} S_i^{(r,\ell,h)} Q x_i^{(\ell-1,h)} \tag{3}$$

其中 $\lambda_i^{(r,\ell,h)}$ 是仅依赖于 $x_i^{(\ell-1,h)}$ 的权重， $R = \lceil \log_2 i + 1 \rceil + 1$ 和 $S_i^{(r,\ell,h)}$ 是隐藏状态。在（Guo 等人，2025年）中，这些状态通过递归计算

$$S_i^{(r,\ell,h)} = \begin{cases} V x_i^{(\ell-1,h)} (K x_i^{(\ell-1,h)})^\top & \text{if } r{=}0 \\ 0 & \text{if } 0{<}r{\le}\,\mathrm{lssb}(i) \\ \sum_{r'=0}^{r-1} S_{i-1}^{(r',\ell,h)} & \text{if } r{=}\,\mathrm{lssb}(i){+}1 \\ S_{i-1}^{(r,\ell,h)} & \text{if } r{>}\,\mathrm{lssb}(i){+}1 \end{cases},$$

其中 $\mathrm{lssb}(t) = \max\{\ell \in \mathbb{N} \mid 2^\ell$ 可整除 $t\}$。在此设置下，维度为d 的单层单头对数线性注意力可以视为隐藏维度为 $3d^2$的循环神经网络，其隐藏状态包含 $S_i^{(r)}$ 中可能非零的 3 个值，即 $\mathbf{h}_i = (S_i^{(0)}$ 、 $S_i^{(\mathrm{lssb}(i))}$ 、 $S_i^{(\mathrm{lssb}(i)+1)})$。

我们考虑一种更一般的设置

$$S_i^{(r,\ell,h)} = f^{(r,\ell,h)} \left((S_{i-1}^{(r',\ell,h)})_{r' \in [0,r-1]}, x_i^{(\ell-1,h)} \right) \tag{4}$$

其中 $f^{(r,\ell,h)}$ 是某些预训练函数。这使得计算可以在对数时间和对数内存内完成，因为只需维护一组大小关于序列长度呈对数增长的隐藏状态。

A.1.2. 函数求值与置换复合

定义 A.1 （函数求值）。 在函数求值任务 $\mathrm{Eva}(f$、 $x)$ 中，输入由一个函数 f： $[n] \to [n]$构成，该函数由编码 $f(1)$ $\cdots$、 $f(n)$ 的 n个词元表示；输入还包括一个由单个词元描述的元素 $x \in [n]$，目标是输出 $\mathrm{Eva}(f$、 $x) = f(x) \in [n]$。

置换　复合任务可视为 n 函数求值任务的并行实例 .

定义 A.2 （置换复合）。 在置换复合任务 $\mathrm{PerCom}(\sigma, \tau)$ 中，输入由两个双射 σ, τ: $[n] \to [n]$组成（每个双射各占用 n个词元来描述 $\sigma(1) \cdots, \sigma(n)$，然后是 $\tau(1) \cdots, \tau(n)$），目标是以序列 $\sigma(\tau(1)) \cdots, \sigma(\tau(n))$ 的形式输出它们的复合。

A.2. Eva 上线性注意力的下界

在本小节中，我们针对不同任务建立线性注意力和对数线性注意力的通信模型。

定理 A.3. L层线性注意力无法解决 Eva，只要 $LHd(d+1)p < n$对数 n；而单层全注意力可用 $Hdp = O($多项式对数 $n)$解决该任务。对带有思维链的线性注意力，结论同样成立。

定理 A.3 可由以下结果推出，因为线性注意力可以视为循环神经网络（引理 2.2）。

定理 A.4. 一个具有 H 个注意力头、隐藏维度为 m 、精度为 p 的 L 层循环神经网络无法求解 Eva，只要 $LHmp < n\log\, n$。对于带有思维链的循环神经网络，结论也成立。

为了证明这个定理，我们首先介绍用于求解 Eva 问题的无思维链循环神经网络通信模型。

通过循环神经网络与Eva通信模型

设置。 有 2 位参与者：Alice 和 Bob。他们通过 L 轮次进行通信。如果启用思维链，还会加入第三位参与者 Charles。

输入。Alice 接收 n 个词元作为输入（对应于 $f(1)\cdots, f(n)$），Bob 接收一个词元作为输入（对应于 x），而 Charles 处理思维链步骤中生成的词元。**通信**。对于 $\ell \in [L]$，在第 ℓ 轮通信中，Alice 向 Bob 发送一条包含 Hmp 比特的消息 M_ℓ。如果启用思维链，Bob 随后会向 Charles 发送一条包含 $H(m+d)p$ 比特的消息 M_ℓ。**输出**。在第 L 轮结束时，如果启用思维链，Charles 会根据其掌握的信息输出一条消息；否则，Bob 会根据其掌握的信息输出一条消息。

L 个通信轮次对应于循环神经网络的 L 层。在每一轮 ℓ中，Alice 发送给 Bob 的消息 M_ℓ 对应于第 ℓ 层第 h 个注意力头在位置 n 的隐藏状态 $\mathrm{h}_n^{(\ell,h)}$。随后，Bob 可以利用这些信息更新隐藏状态 $\mathrm{h}_{n+1}^{(\ell,h)}$，并计算 $y_{n+1}^{(\ell,h)}$，即第 ℓ 层在位置 $n+1$ 的输出。

在这个模型的变体中，我们用带思维链（CoT）的循环神经网络来描述函数求值的计算，并引入第三位参与者 Charles，以模拟思维链（CoT）步骤中生成额外词元所涉及的计算。在这个扩展模型中，Bob 必须将隐藏状态和输出 $y_{n+1}^{(\ell,h)}$ 都发送给 Charles。该输出 $y_{n+1}^{(\ell,h)}$ 作为 Charles 处理的第一个词元的初始值。因此，消息的大小（以比特计）变为$Hmp + Hdp$。

关键洞见在于，Charles 的所有知识都来源于 Bob 发送的信息。因此，如果 Charles 能够计算出最终输出，那么掌握至少相同信息的 Bob 也必然能够计算出该输出。

证明的关键在于，通信比特总数为 $LHmp$；要让 Bob 区分所有函数，该数量至少应为 $\log(n^n) = n\log n$。

我们还为 PerCom和 2-Sum 建立了通信模型，详细证明见附录 A。

定理 *A.4* 的证明。 我们首先考虑没有思维链的情况。在 L 轮次中，Alice 总共向 Bob 发送 $LHmp$ 比特，而 Bob 不向 Alice 发送任何消息。这些消息最多能让 Bob 区分 2^{LHmp} 个不同的函数。

如果 $2^{LHmp} < n^n$，或者等价地，如果 $LHmp < n\log n$，那么就存在两个不同的函数 f_1，f_2：$[n] \to [n]$ $(f_1 \neq f_2)$，它们对 Bob 来说不可区分；也就是说，对于 f_1 和 f_2，Alice 发送的消息序列完全相同。由于 $f_1 \neq f_2$，存在某个元素 $x \in [n]$，使得 $f_1(x) \neq f_2(x)$。如果 Bob 的输入是 x，那么他就不可能输出正确答案，因为 Bob 无法判断应该输出 $f_1(x)$ 还是 $f_2(x)$。

在思维链的情况下，Charles 的所有信息都来自 Bob。因此，如果 Charles 能确定答案 $f(x)$，那么 Bob 也必须能够计算出该答案。这与无思维链情况下得到的下界相矛盾。 □

A.3. 对数线性注意力在Eva上的下界

在本小节中，我们建立对数线性注意力求解Eva的通信模型，从而得到一个下界。**定理 A.5.** 当 $LHd^2p < \Theta(n)$时，L 层对数线性注意力无法求解Eva，而单层全注意力以 $Hdp = O(\mathrm{poly}\log n)$求解该任务。对于带思维链的对数线性注意力，结论同样成立。

由（3）和（4）可知，为计算输出 $y^{\ell}_{n+1}{}^{h}$，Bob需要对所有 $r \in [0, \lceil\log(n+1)+1\rceil]$获得隐藏状态 $S_{n+1}^{(r,\ell,h)}$。这些状态又依赖于先前的隐藏状态 $S_n^{(r,\ell,h)}$，对应于 $r \in [0, \lceil\log(n) + 1\rceil]$。因此，Alice只需将这些 $O(\log n)$个隐藏状态发送给Bob。每个注意力头需要 $O(d^2p\log n)$比特，从而使每层的消息总大小为$O(Hd^2p\log n)$比特。

基于对数线性注意力的Eva通信模型

设置。两位参与方 Alice 和 Bob 在 L 轮次中进行通信。启用思维链时，会加入第三位参与方 Charles。

输入。 Alice 接收 n 个词元作为输入，Bob 接收一个词元作为输入。Charles 接收与

思维链步骤相对应的词元。
通信。 对于 $\ell \in [L]$，在第 ℓ 轮，Alice 向 Bob 发送一条包含 $O(Hd^2p\log\ n)$ 比特的消息 M_ℓ。如果启用了思维链，Bob 随后向 Charles 发送一条包含 $O(Hd^2p\log(n+1)\ + Hdp)$ 比特的消息 $\tilde{M}_\ell$。**输出**。 在第 L 轮结束时，如果启用了思维链，Charles 会根据收到的所有信息生成输出；否则，Bob 会根据自身信息输出一条消息。

在带有思维链的对数线性注意力模型中，我们引入一位新的参与方 Charles，用来处理思维链步骤中的词元。在该模型中，Bob 将输出 $y_{n+1}^{(\ell,h)}$ 以及相关的隐藏状态发送给 Charles，形成一条大小为 $Hd^2p\log(n+1)\ + Hdp$ 比特的消息。然而，正如定理 A.3的证明所示，Charles 的整个信息状态都源自 Bob。因此，引入 Charles 不会改变无思维链情形下得到的下界。

现在我们证明 Eva 在对数线性注意力下的下界。

定理的证明 *A.5*。 对于不使用思维链的情况，在这场进行 L 轮的通信过程中，Alice 总共向 Bob 发送了 $O(LHd^2p\ \log\ n)$ 比特，而 Bob 没有向 Alice 发送任何消息。这种通信最多只能区分 $N = 2^{O(LHd^2p\log\ n)}$ 个不同的对象。

如果 $N < n^n$，或者等价地说，如果 $LHd^2p < \Theta(n)$，那么应该存在两个不同的函数 $f_1,\ f_2 : [n] \to [n]$，它们对于 Bob 来说不可区分（也就是说，当输入为 f_1 时 Alice 发送给 Bob 的消息，与输入为 f_2 时发送的消息相同）。由于 $f_1 \neq f_2$，存在一个元素 $x \in [n]$，使得 $f_1(x) \neq f_2(x)$。如果 Bob 的输入是 x，那么就不可能得出正确的输出，因为 Bob 无法判断答案应该是 $f_1(x)$ 还是 $f_2(x)$。
在思维链情形下，Charles 的全部知识都来自 Bob。因此，如果 Charles 能够计算 $f(x)$，那么 Bob 也必然能够计算它。这与非思维链情形下得到的下界相矛盾。 □

A.4. PerCom 上对数注意力和对数线性注意力的下界

与前面的 结果类似，我们建立了循环神经网络和对数线性注意力解决 PerCom 的通信模型 .

定理 A.6. *L 层线性注意力无法解决 PerCom，只要 $LHd(d+1)p < \log(n!) = \Theta(n\log\ n)$；而单层全注意力只需 $Hdp = O(\mathrm{poly}\log n)$ 即可解决该问题。对于带有思维链的线性注意力，结论同样成立。*

对于循环神经网络，我们证明以下定理；结合引理2.2 以及第C节中的构造，可推出定理 A.6。

定理 A.7. *一个具有 H 个注意力头、隐藏维度 m 和精度 p 的 L 层循环神经网络，在 $LHmp < \log(n!)$ 时无法求解 PerCom 。同样的结果也适用于带思维链的线性注意力。*

通过循环神经网络实现 PerCom 的通信模型

设定。 两名参与者 Alice 和 Bob 通过 L 个轮次进行通信。若启用思维链，则纳入第三名参与方 Charles。

输入。 Alice 接收 n 个词元作为输入（对应于 $\sigma(1)\ \cdots, \sigma(n)$），Bob 接收 n 个词元作为输入（对应于 $\tau(1)\ \cdots, \tau(n)$）。若允许思维链，Charles 接收对应于思维链步骤的词元。**通信**。 对于 $\ell \in [L]$，在第 ℓ 轮次中，Alice 向 Bob 发送一条 Hmp 比特的消息。若允许思维链，Bob 随后向 Charles 发送一条 $H(m+d)p$ 比特的消息。

输出。 在第 L 轮次结束时，若启用思维链，则由 Charles 产生输出；否则由 Bob 产生输出。

定理证明 *A.7*。 类似地，在 L 轮次通信过程中，Alice 总共发送 $LHmp$ 比特，而 Bob 不向 Alice 发送任何消息。为了让 Bob 区分所有可能的排列 σ，应满足 $2^{LHmp} \geq n!$ 。应用斯特林公式（斯特林，1753年）

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \quad \text{as} \quad n \to \infty$$

由此可得 $LHmp = \Omega(n\log\ n)$ 是循环神经网络解决PerCom所必需的。对于思维链情形，采用与定理A.4的证明相同的论证即可证明所需结果。 □

定理 A.8。 L层对数线性注意力无法解决PerCom，只要 $LHd^2p < \log(n!)\ \log n = \Theta(n)$；而单层全注意力可用 $Hdp = O(\text{poly}\log n)$解决该任务。对数线性注意力结合思维链时，结论也成立。

定理的证明 *A.8*。 我们考虑如下通信模型：

基于对数线性注意力的PerCom通信模型

设定。 两名参与者 Alice 与 Bob 在 L 个轮次中进行通信。若启用思维链，则纳入第三名参与方 Charles。

输入。 Alice 接收 n 个词元作为输入，Bob 接收 n 个词元。若允许思维链，Charles 接收与思维链步骤对应的词元。**通信**。 对于 $\ell \in [L]$，在第 ℓ 轮次中，Alice 向 Bob 发送一条包含 $O(Hd^2p\log n)$ 比特的消息。若允许思维链，Bob 随后向 Charles 发送一条包含 $O(Hd^2p\log(2n) + Hdp)$ 比特的消息。**输出**。 在第 L 轮次结束时，若允许思维链，则 Charles 根据其信息输出一条消息，否则 Bob 根据其信息输出一条消息。

为得到正确输出，Bob 必须识别排列 σ。为区分所有可能情况，必须有 $2^{O(Hd^2p\log n)} \geq n!$。这意味着 $Hd^2p \geq \Omega\left(\frac{\log(n!)}{\log n}\right) = \Theta(n)$ 正如定理 A.7 的证明中所述。 □

A.5. 对数注意力和对数线性注意力在 2-Sum 上的下界

定理 A.9. *L 层线性注意力无法解决 2-Sum，只要 $LHd(d+1)p < \Theta(n\log n)$；$L$ 层对数线性注意力无法解决 2-Sum，只要 $LHd^2p < \Theta(n)$；而单层全注意力以 $Hdp = O(\text{poly}\log n)$ 解决该任务。对于带思维链的线性注意力和对数线性注意力，同样成立。*

定理 *A.9* 的证明。我们首先证明线性注意力的下界。考虑输入序列长度为 n 的 2-Sum 任务。我们聚焦于最后一个令牌 x_n 及其对应的输出 y_n。 y_n 的值取决于是否存在一个索引 $j < n$，使得 $x_j + x_n \equiv 0$ 模 n。

我们在任务中取 $M = n^2$。我们采用如下通信模型。

通过线性注意力实现 2-Sum 的通信模型

设置。 两名参与者 Alice 与 Bob 通过 L 轮次进行通信。若启用思维链，则加入第三名参与方 Charles。

输入。 Alice 接收前 $n-1$ 个词元（x_1，……，x_{n-1}），Bob 接收最后一个令牌 x_n。若允许思维链，Charles 接收与思维链步骤相对应的词元。**通信**。 对于 $\ell \in [L]$，在第 ℓ 轮次中，Alice 向 Bob 发送一条 $H \cdot (d^2+d) \cdot p$ 比特的消息（对应于位置 $n-1$ 处第 ℓ 层的隐藏状态）。若允许思维链，Bob 随后向 Charles 发送一条 $H \cdot (d^2+d) \cdot p + H \cdot d \cdot p$ 比特的消息（位置 n 处第 ℓ 层的隐藏状态与输出）。**输出**。 在第 L 轮次结束时，若允许思维链，Charles 根据其所掌握的信息输出 y_n；否则 Bob 根据其所掌握的信息输出 y_n。

若 $LHd(d+1)p < \Theta(n\log n)$ 满足 $2^{LHd(d+1)p} < \binom{n^2}{n-1}$，则存在两个由前 $n-1$ 个词元组成的序列，分别记为 S_1 和 S_2，它们产生相同的隐藏状态，但构成的值集合不同： S_1 和 S_2。因此，存在一个值 v，它出现在 S_1 中，却不出现在 S_2 中（反之亦然）。不失一般性，假设 v 出现在 S_1 中，却不出现在 S_2 中。将 Bob 的词元设为 $x_n = -v$ 模 M。

对于 S_1，我们有 $y_n = 1$，因为存在 j，满足 $x_j = v$，使得 $x_j + x_n \equiv 0$ 模 n。对于 S_2，由于 v 并未出现，因此不存在这样的 j，所以 $y_n = 0$。然而，由于两个序列的隐藏状态相同，模型必须产生相同的输出，导致至少一种情况下出错。因此，只要 $LHd(d+1)p < \Theta(n\log n)$，线性注意力就无法解决 2-Sum。

对于对数线性注意力，我们采用类似的通信模型。与定理 A.5的证明中一样，隐藏状态 e

从 Alice 传递给 Bob 的大小为 $O(LHd^2p\log\ n)$ 比特。如果 $LHd^2p\log\ n < \Theta(n\log\ n)$，即 $LHd^2p < \Theta(n)$，那么通过类似的论证，存在两个序列 S_1 和 S_2，它们由前 $n-1$ 个词元构成，却产生相同的隐藏状态，并对应不同的值集合。选择一个在 S_1 中出现、但不在 S_2 中出现的值 v，令 $x_n = -v \mod n$，便会导致矛盾：模型必须为两个序列输出相同的 y_n，但正确输出却不同。因此，只要 $LHd^2p < \Theta(n)$，对数线性注意力就无法解决 2-Sum。

通过对数线性注意力实现2-Sum的通信模型

设置。 两名参与者 Alice 和 Bob 通过 L 轮次进行通信。如果启用思维链，则加入第三名参与者 Charles。

输入。 Alice 接收前 $n-1$ 个词元（x_1，……，x_{n-1}），Bob 接收最后一个词元 x_n。如果允许思维链，Charles 接收与思维链步骤对应的词元。**通信。** 对于 $\ell \in [L]$，在第 ℓ轮中，Alice 向 Bob 发送一条长度为 $O(Hd^2p\log\ n)$ 比特的消息。如果允许思维链，Bob 随后向 Charles 发送一条长度为 $O(Hd^2p\log\ n) + Hdp$ 比特的消息。**输出。** 在第 L轮结束时，如果允许思维链，Charles 根据其信息输出 y_n；否则，Bob 根据其信息输出 y_n。

上界——单层 Transformer 解码器利用全注意力以 $Hdp = O(\log n)$解决 2-Sum——源自（Sanford 等人，2023年），定理6。

对于思维链变体，同样适用这些下界，因为 Charles 接收到的所有信息都源自 Bob（如果 Charles 能够计算出输出，Bob 也能做到，这与不使用思维链时得出的下界相矛盾）。由此完成定理A.9的证明。

□

A.6. 稀疏注意力的下界

定理 *1.2* 的证明。 不失一般性，我们假设 B 整除 n，并证明参数较小的稀疏注意力机制无法正确输出位置 $n+1$处的期望值。我们定义如下通信模型，用于计算 (B,k)-稀疏注意力的这一输出。

通过稀疏注意力实现 2-Sum 的通信模型

设置。 共有 $\frac{n}{B}+1$ 位参与者。**输入。** 前 $\frac{n}{B}$ 位参与者各自接收 B 个词元，最后一位参与者接收一个词元。**通信。** 前 $\frac{n}{B}$ 位参与者各自向最后一位参与者发送一条 Hdp 比特的消息；随后，最后一位参与者选择 k 位参与者，并获取他们的全部信息。**输出。** 最后一位参与者根据其掌握的信息输出一条消息。

在此模型中，每个块对应一个参与方，而压缩词元则构成发送给最后一个词元的消息 .

为了让最后一位参与方输出正确答案，前面的每位参与方都必须通过一条包含 Hdp 比特的压缩消息来传达其 B 词元集合。事实上，若情况并非如此，也就是说，存在两个元组 $(a_i)_{i\in[B]}$ 和 $(b_i)_{i\in[B]}$，使得 $f(a_1, \cdots, a_B) = f(b_1, \cdots, b_B)$，但 $\{a_1,\cdots,a_B\} \neq \{b_1,\cdots,b_B\}$，那么我们可以考虑每个块的输入要么是 $(a_i)_{i\in[B]}$，要么是 $(b_i)_{i\in[B]}$的情况。

在这种情况下，所有压缩词元都会相同，因此每个块的选择分数也相同。令 $n - x_{n+1}$为 $\{a_1,\cdots,a_B\} \setminus \{b_1,\cdots,b_B\}$的一个元素，并考虑所有 k 个选定块的输入均为 $(b_i)_{i\in[B]}$的情况。在这种情况下，对于其余块，最后一位参与方无法区分 $(a_i)_{i\in[B]}$ 和 $(b_i)_{i\in[B]}$。因此，它无法正确输出答案。

因此，Hdp 比特的压缩消息应当传达由 B 个词元组成的集合；该集合是 $[n]$ 的子集，大小至多为 B，因此我们有

$$2^{Hdp} \geq \sum_{j\leq B}\binom{n}{j} \geq \binom{n}{B} \sim \frac{n^B}{B!},$$

这意味着 $Hdp = \Omega(B\log n)$，并完成了证明。 □

B. 混合通信下界

在本节中，我们证明定理1.1。我们按照（Chen等人，2025年）的框架选择参数，并借鉴他们构造不可区分分解的技术来推导结果。

B.1. 参数与策略

为证明定理1.1，我们构造任务参数，使输入长度满足 $n \le (Hdp)^{4\cdot 16^L}$，但该任务无法由一个$(L, 2^{3L^2}, \cdots, 2^{3L^2})$-混合 Transformer 求解。本节中我们始终使用以下参数：

$$K = (HdpL)^8 \cdot 8^{2L^2}, \quad m = K^{\sum_{\ell\in[0:L-1]} 8^\ell + 1}, \tag{5}$$

$$n_\ell = K^{4\cdot 8^{L-\ell-1}}, \forall \ell \in [L-1]. \tag{6}$$

我们假设 $Hdp \ge 2$。根据定义2.5 可知

$$N_\ell = m \cdot \prod_{\ell'\in[\ell]} n_{\ell'} \quad \forall \ell \in [0:L-1]. \tag{7}$$

我们还定义以下辅助参数：

$$x_\ell = K^{8^{L-\ell-1}}, \forall \ell \in [0:L-1] \tag{8}$$

$$A_\ell = \left[N_{\ell-1}^{N_{\ell-1}}\right], \forall \ell \in [L] \tag{9}$$

$$\Delta_\ell = 2^{4\sqrt{K}(x_0 \ldots x_{\ell-2})\cdot(n_1 \ldots n_{L-1})}, \forall \ell \in [2:L]) \tag{10}$$

$$\Theta_\ell = 8^{-L\ell}(x_0 \ldots x_\ell)\cdot(n_1 \ldots n_{\ell-1}), \forall \ell \in [L-1]. \tag{11}$$

为方便符号约定，我们还定义 $A_{-1} = \prod_{i=1}^{L-1}[n_i]$ 和 $A_0 = [m]$。回顾一下，我们将查询 w 记为 z_{-1}。因此，对于每个 $i \in [-1:\ L]$，参与方 i 都会从集合 A_i 中接收一个输入。注意，我们将 $A_\ell = \left[N_{\ell-1}^{N_{\ell-1}}\right]$ 中的一个元素视为一个函数 $[N_{\ell-1}] \to [N_{\ell-1}]$。

我们首先证明，该任务具有所需的输入大小。

引理 B.1。 对于参数在 L-序贯函数复合任务中定义为 (6)的输入提示长度$n = 2 + N_0 + N_1 + \cdots + N_{L-1}$ 满足 $n \le (Hdp)^{4\cdot 16^L}$。

证明。 根据（6）和（7）中的参数定义，可得 $2 \le N_0$ 以及 $2N_{\ell-1} \le N_\ell$ 对于所有 $\ell \in [1, L-1]$均成立。因此，总输入长度可界定如下：

$$\begin{aligned}
n &= 2 + N_0 + N_1 + \cdots + N_{L-1} \\
(2 \le N_0, 2N_{\ell-1} \le N_\ell) \quad &\le 2N_{L-1} = 2m \cdot \prod_{\ell\in[L-1]} n_\ell = 2 \cdot K^{\sum_{\ell\in[0:L-1]} 8^\ell + 1 + \sum_{\ell\in[1:L-1]} 4\cdot 8^{L-\ell-1}} \\
&= 2 \cdot \left((HdpL)^8 \cdot 8^{2L^2}\right)^{\frac{12}{7}\cdot 8^{L-1} + \frac{2}{7}} \\
(Hdp \ge 2) \quad &\le (Hdp)^{1+\left(\frac{12}{7}\cdot 8^{L-1} + \frac{2}{7}\right)(8\log L + 8 + 6L^2)} \\
&\le (Hdp)^{2\cdot 8^{L-1}(8\log L + 8 + 6L^2)} \\
(1 + \log L \le L) \quad &\le (Hdp)^{2\cdot 8^{L-1}(8L + 6L^2)} \\
(2L \le 2^L, L^2 \le 2^L) \quad &\le (Hdp)^{2\cdot 8^{L-1}\cdot 10\cdot 2^L} \le (Hdp)^{4\cdot 16^L}
\end{aligned}$$

□

还需证明该任务无法由 $(L,\ 2^{3L^2},\ \cdots,\ 2^{3L^2})$-混合 Transformer 解决。事实上，我们将证明以下更强的结论。

定理 B.2。 在以下假设下，不存在确定性的$(L, a_1, \cdots, a_L)$混合通信协议能够解决 L-函数组合 。

1. $a_1 \le 1$,
2. 对于所有 $\ell \in [L-1]$，我们有

$$Hd(d+1)p(a_1+\cdots+a_{\ell+1}) \le \sqrt{K}(x_0\cdots x_{\ell-1})\cdot(n_1\cdots n_{L-1}). \tag{12}$$

现在我们证明，我们的 $(L, 1, 2^{3L^2}, \cdots, 2^{3L^2})$ 满足定理 B.2 的假设。因此，定理 B.2 蕴含定理 1.1，因为 $(L-1, 2^{3L^2}, \cdots, 2^{3L^2})$-混合 Transformer 可以视为一个 $(L, 1, 2^{3L^2}, \cdots, 2^{3L^2})$-Transformer，其中首个全注意力层及其后续线性注意力层的多层感知机层 $x_i^{(\ell)} = g(x_i^{(\ell-1)}, y_i^{(\ell)}) := x_i^{(\ell-1)}$ 均为平凡层。

引理 B.3. 式 (12) 中的假设在 $a_1 = 1, a_2 = \cdots = a_L = 2^{3L^2}$.

引理 *B.3* 的证明。 对于任意 $\ell \in [L-1]$，我们有

$$Hd(d+1)p(a_1+a_2+\cdots+a_{\ell+1}) \le 2HdpL2^{3L^2} \le 8^{L^2}(HdpL)^4,$$

而对于右侧，我们有

$$\sqrt{K}(x_0\cdots x_{\ell-1})\cdot(n_1\cdots n_{L-1}) \ge \sqrt{K} = 8^{L^2}(HdpL)^4 \ge Hd(d+1)p(a_1+a_2+\cdots+a_{\ell+1}). \qquad \square$$

符号约定。 为方便表述，我们用 z_{-1} 和 w 互换表示参与方 -1的输入。接下来，我们将详细说明几个对证明至关重要的关键定义。

- (软转录 $\Pi_{(\ell)\,j,i}$) 对于任意 $i \in [-1:\ L-1]$, $j \in [i+1:\ L]$, $\ell \in [L]$，回顾一下， $\Pi_{(\ell)\,j,i}$ 表示在第 ℓ个通信轮次中从参与方 j 发送给参与方 i 的软转录。其值由参与者 $[i:\ L]$的输入决定（即 $z_L, \ldots, z_i$），且与参与者 $[-1:\ i-1]$ 的输入无关（即 $z_{i-1}, \ldots, z_{-1}$）。

  对于任意固定的输入 $\widetilde{z}_L \in [N_{L-1}]^{N_{L-1}}$、……、 $\widetilde{z}_i \in [N_{i-1}]^{N_{i-1}}$，令 $\Pi_{j\,i}^{(\ell)}(\widetilde{z}_L$、……、 $\widetilde{z}_i)$ 表示参与方 t 接收输入 $z_t = \widetilde{z}_t\,(t \in [i:\ L])$ 时的软转录。
- (线性转录 $\Sigma_{\ i+1}^{(\ell)\ m}$) 对于任意 $i \in [-1, L-1]$、 $l \in [L]$ 和 $m \in [0, a_\ell - 1]$，回顾一下， $\Sigma_{\ i+1}^{(\ell)\ m}$ 是参与方 $i+1$ 在通信第 ℓ 个轮次的第 $m+1$ 个线性轮中发送给参与方 i 的线性转录。其值由参与者 $[i+1$ 的输入决定: $L]$（即 z_L、……、 z_{i+1}），并且与参与者 $[-1$ 的输入无关: $i]$（即 z_i、……、 z_{-1}）。

  对于任意固定的输入 $\widetilde{z}_L \in [N_{L-1}]^{N_{L-1}}$、……、 $\widetilde{z}_i \in [N_{i-1}]^{N_{i-1}}$，令 $\Sigma_{\ i+1}^{(\ell)\ m}(\widetilde{z}_L$，……， $\widetilde{z}_i)$ 表示参与方t接收输入 $z_t = \widetilde{z}_t\,(t \in [i:\ L])$ 时的通信记录。
- （部分组合值）对于任意 $\ell \in [0:\ L]$， i_ℓ的值由 w， z_0、……、 z_ℓ决定。我们用$i_\ell(\widetilde{w}$， $\widetilde{z}_0$，……， $\widetilde{z}_\ell)$ 表示当 $w = \widetilde{w}$， $z_0 = \widetilde{z}_0$，……， $z_\ell = \widetilde{z}_\ell$ 时它的值。

不可区分分解。我们的证明中的关键概念是不可区分分解，该概念由 Chen 等人（2025）提出。一个不可区分分解由两个集合 $R_{\ge\ell}$ 和 $Z_{<\ell}$构成，其中 $R_{\ge\ell}$ 是参与方 $[\ell:\ L]$ 的输入赋值集合，而 $Z_{<\ell}$ 是参与方 $[-1:\ \ell-1]$ 的输入赋值集合。其关键性质是，对于前 ℓ 个参与者的任意固定输入$z_{<\ell} \in Z_{<\ell}$， $R_{\ge\ell}$ 的所有赋值对于参与方 $[-1:\ \ell-1]$ 在经过 ℓ 轮次后的输入 $z_{<\ell}$ 上都不可区分，因为它们会产生相同的通信记录。形式上，我们对 Chen 等人（2025）中的定义作如下调整。

定义 B.4 （不可区分分解）。 令 $\ell \in [2:\ L]$，不可区分分解是由 $R_{\ge\ell} \subseteq A_L \times A_{L-1} \times \cdots \times A_\ell$ 和 $Z_{<\ell} = Z_{-1} \times \cdots \times Z_{\ell-1}$ 构成的一对集合，其中 $Z_{-1} = A_{-1}$， $Z_0 \subseteq A_0$， $\cdots$， $Z_{\ell-1} \subseteq A_{\ell-1}$，满足：对于每个 $\widetilde{z}_{<\ell} \in Z_{<\ell}$，以及对于每个 $\widetilde{\alpha}_{\ge\ell}$， $\beta_{\ge\ell} \in R_{\ge\ell}$，均有：

$$\Pi_{j,i}^{(\ell')}(\widetilde{z}_{<\ell}, \widetilde{\alpha}_{\ge\ell}) = \Pi_{j,i}^{(\ell')}(\widetilde{z}_{<\ell}, \widetilde{\beta}_{\ge\ell})$$

对于每个 $j \in [\ell:\ L]$、 $i \in [-1:\ \ell-1]$ 和 $\ell' \in [\ell]$，以及

$$\Sigma_{i+1}^{(\ell'),m}(\widetilde{z}_{<\ell}, \widetilde{\alpha}_{\ge\ell}) = \Sigma_{i+1}^{(\ell'),m}(\widetilde{z}_{<\ell}, \widetilde{\beta}_{\ge\ell})$$

对于每个 $i \in [-1:\ \ell-1]$、 $\ell' \in [\ell]$ 和 $m \in [0, a_{\ell'} - 1]$。

当 $\ell = L$ 时，不可区分分解的作用便显而易见。在这种情况下，对于从 $Z_{<L}$ 分配给参与者 $[-1$ 的每个输入赋值：$L-1]$，无论分配给参与者 L 的输入 $\widetilde{z}_L \in R_{\geq L}$ 是什么，参与者 -1（最终输出参与者）在经过 L 轮次后（即协议结束时）观察到的通信记录都完全相同。因此，对于每个 $\widetilde{z}_{<L} \in Z_{<L}$，协议 L-FuncComp($\widetilde{z}_{<L}$，$\widetilde{z}_L$) 的输出对于每个 $\widetilde{z}_L \in R_{\geq L}$ 都必须相同。我们将仔细定义集合 $R_{\geq \ell}$ 和 $Z_{<\ell}$，使得满足这一要求会导致矛盾，从而确立所需的下界。

对于子集 $Z_{<\ell}$，我们 定义经过第 $(\ell-1)$ 轮后可达的部分复合值集合 $\mathcal{I}_{\ell-1}(Z_{<\ell})$ s

$$\mathcal{I}_{\ell-1}(Z_{<\ell}) := \{i_{\ell-1}(\widetilde{z}_{-1}, \widetilde{z}_0, \ldots, \widetilde{z}_{\ell-1}) | (\widetilde{z}_{-1}, \widetilde{z}_0, \ldots, \widetilde{z}_{\ell-1}) \in Z_{<\ell}\}.$$

换言之，$\mathcal{I}_{\ell-1}(Z_{<\ell})$ 是中间复合所有可能取值的集合，其中参与方 $[-1$ 的输入：$\ell-1]$被限制在 $Z_{<\ell}$内。

如下引理与（Chen 等人，2025年）中的结果一致，表明只要存在一个足够好的不可区分配置，就能推出所需的下界，适用于 $\ell = L$。

引理 B.5。若存在一个不可区分分解 $R_{\geq L}$ 和 $Z_{<L}$，满足 $|R_{\geq L}| \geq |A_L|/\Delta_L$ 和 $|\mathcal{I}_{L-1}(Z_{<L})| \geq \Theta_{L-1}$，则一个 L 轮混合通信协议无法解决 L-FuncComp 。

定理 B.2 的证明随后通过归纳论证构造所需的分解来完成，我们将在下一小节中给出该论证。这一构造在以下引理中予以形式化。

引理 B.6. 对于每一个 $(L,\ a_1,\ \ldots\ldots,\ a_L)$混合通信协议，在 (12) 的假设下，存在不可区分的分解 $R_{\geq L}$ 与 $Z_{<L}$，满足定理 *B.5* 的要求。

本节的其余部分 专门用于证明引理 B.6。我们建立了一个更一般的归纳断言，即定理 B .7（如下）。情形 $\ell = L$直接 蕴含定理 B.6。

引理 B.7 (主要引理)。 对于任意 $\ell \in [2:\ L]$，我们可以构造

- 一对集合$(R_{\geq \ell}, Z_{<\ell})$，其中 $R_{\geq \ell} \subseteq A_L \times A_{L-1} \times \cdots \times A_\ell$ 且 $Z_{<\ell} = Z_{-1} \times Z_0 \times \cdots \times Z_{\ell-1}$，满足 $Z_{-1} = [n_1 \cdots n_{L-1}]$、$Z_0 \subseteq A_0$、$Z_i \subseteq A_i$，并且 $|Z_i| = x_i$ 对于 $i \in [0:\ \ell-1]$。
- 当参与者 $[-1:\ \ell-1]$ 以 $Z_{<\ell}$为输入时，参与者 $[\ell:\ L]$ 至 $[-1:\ \ell-1]$ 在前 ℓ 轮次中的软转录。即，

$$\Lambda^{(\ell)} := \left(\Lambda_{j,i}^{(\ell,\ell')}\right)_{j\in[\ell:L], i\in[-1:\ell-1], \ell'\in[\ell]}$$

其中

$$\Lambda_{j,i}^{(\ell,\ell')} := \left(\Lambda_{j,i}^{(\ell,\ell')}(\widetilde{z}_{\ell-1}, \ldots, \widetilde{z}_i)\right)_{\widetilde{z}_{\ell-1}\in Z_{\ell-1}, \ldots, \widetilde{z}_i \in Z_i} \quad \textit{and} \quad \Lambda_{j,i}^{(\ell,\ell')}(\widetilde{z}_{\ell-1}, \ldots, \widetilde{z}_i) \in \mathsf{domain}(\Pi_{j,i}^{(\ell')})$$

- 当参与者 $[-1:\ \ell-1]$ 以 $Z_{<\ell}$为输入时，参与者 ℓ 至 $\ell-1$ 在前 ℓ 轮次中的线性转录。即，

$$\Xi^{(\ell)} := \left(\Xi_\ell^{(\ell'),m}\right)_{\ell'\in[\ell], m\in[0,a_{\ell'}-1]} \quad \textit{where} \quad \Xi_\ell^{(\ell'),m} \in \mathsf{domain}(\Sigma_\ell^{(\ell'),m})$$

使得我们具有以下性质：

- **（软转录的一致性）**

$$\Pi_{j,i}^{(\ell')}(\widetilde{z}_L, \ldots, \widetilde{z}_i) = \Lambda_{j,i}^{(\ell,\ell')}(\widetilde{z}_{\ell-1}, \ldots, \widetilde{z}_i)$$
$$\forall j \in [\ell:L], i \in [-1:\ell-1], \ell' \in [\ell], \widetilde{z}_{\geq\ell} \in R_{\geq L}, \widetilde{z}_{\ell-1} \in Z_{\ell-1}, \ldots \widetilde{z}_i \in Z_i,$$

- **（线性转录的一致性）**

$$\Sigma_\ell^{(\ell'),m}(\widetilde{z}_L, \ldots, \widetilde{z}_\ell) = \Xi_\ell^{(\ell'),m}, \forall \ell' \in [\ell], m \in [0, a_{\ell'}-1], \widetilde{z}_{\geq\ell} \in R_{\geq L},$$

- $|R_{\geq\ell}| \geq |A_L| \cdots |A_\ell|/\Delta_\ell$和 $|\mathcal{I}_{\ell-1}\ (Z_{<\ell})\ | \geq \Theta_{\ell-1}$。

B.2. 初始步骤

我们首先证明基本情形 B.7 的引理 $\ell = 2$。

步骤 1：选择 Z_0，Z_1。我们设定 $Z_0 = [x_0]$。下一步是选择集合 $Z_1 \subseteq A_1$。考虑从玩家 1 到参与方 -1的所有可能的第一个轮次消息，记为

$$\Psi^{(1)}_{1,-1} = \left(\Psi^{(1)}_{1,-1}(\widetilde{z}_{-1})\right)_{\widetilde{z}_{-1}\in Z_{-1}} \quad \text{where} \quad \Psi^{(1)}_{1,-1}(\widetilde{z}_{-1}) \in \{0,1\}^{2Hdp}.$$

此类不同消息模式 $\Psi^{(1)}_{1,-1}$ 的总数为 $2^{2Hdp\cdot|Z_{-1}|} = 2^{2Hdp(n_1\cdots n_{L-1})}$。根据鸽巢原理，存在一个消息模式 $\widetilde{\Psi}^{(1)}_{1,-1} \in \{0,1\}^{2Hdp\cdot(n_1\cdots n_{L-1})}$ 使得

$$S := \{\widetilde{z}_1 \in A_1 : \Pi^{(1)}_{1,-1}(\widetilde{z}_1, \widetilde{z}_{-1}) = \Psi^{(1)}_{1,-1}(\widetilde{z}_{-1})\ \forall \widetilde{z}_{-1} \in Z_{-1}\} \subseteq A_1 \tag{13}$$

满足 $|S| \geq |A_1| \cdot 2^{-2Hdp\cdot(n_1\cdots n_{L-1})}$。注意，第一个轮次消息仅依赖于 $\widetilde{z}_1$ 和 $\widetilde{z}_{-1}$，因此我们将其记为$\Pi^{(1)}_{1,-1}(\widetilde{z}_1, \widetilde{z}_{-1})$。如下引理，如（Chen 等人，2025年）中所述，使我们能够选择一个合适的子集 $Z_1 \subseteq S$。

引理 B.8 ((Chen 等人，2025年)，引理 4.6)。存在一个子集 $Z_1 \subseteq S$ 满足 $|Z_1| = x_1$ 使得 t

$$|\{\widetilde{z}_1(i_0) : \widetilde{z}_1 \in Z_1, i_0 \in Z_0\}| \geq 8^{-L} x_0 x_1 = \Theta_1. \tag{14}$$

我们取引理 Z_1 给出的子集 B.8。下一步是固定前两个轮次中从参与者 $j \in [2: L-1]$ 到参与者$i = -1$，0、1 的通信记录。

步骤 2.1：固定发给参与者 -1的通信记录。我们从第一个轮次开始。我们的目标是对每个 $\widetilde{z}_1 \in Z_1$，$\widetilde{z}_0 \in Z_0$，$\widetilde{z}_{-1} \in Z_{-1}$，固定 $\Lambda^{(2,1)}_{j,-1}(\widetilde{z}_1, \widetilde{z}_0, \widetilde{z}_{-1})$。参与者 j 发给参与者 -1 的第一个轮次消息仅取决于 $\widetilde{z}_{-1}$ 和参与者 j 的输入，且独立于 $\widetilde{z}_1$，$\widetilde{z}_0$。因此，只需固定一种模式

$$\Phi^{(1)}_{j,-1} = \left(\Phi^{(1)}_{j,-1}(\widetilde{z}_{-1})\right)_{\widetilde{z}_{-1}\in Z_{-1}} \quad \text{where} \quad \Phi^{(1)}_{j,-1}(\widetilde{z}_{-1}) \in \{0,1\}^{2Hdp}.$$

然后定义

$$\Lambda^{(2,1)}_{j,-1}(\widetilde{z}_1, \widetilde{z}_0, \widetilde{z}_{-1}) = \Phi^{(1)}_{j,-1}(\widetilde{z}_{-1}) \quad \forall \widetilde{z}_1 \in Z_1, \widetilde{z}_0 \in Z_0, \widetilde{z}_{-1} \in Z_{-1}$$

这种通信记录模式的可能总数至多为 $2^{2Hdp\cdot|Z_{-1}|} = 2^{2Hdp\cdot(n_1\cdots n_{L-1})}$。根据鸽巢原理，我们可以选定一个固定模式 $\{\Lambda^{(2,1)}_{j,-1}\}_{j\in[2:L]}$，使得

$$C_1 := \left\{ \begin{array}{c} (\widetilde{z}_L, \ldots, \widetilde{z}_2) \in A_L \times \cdots \times A_2 : \\ \Pi^{(1)}_{j,-1}(\widetilde{z}_L, \ldots, \widetilde{z}_0, \widetilde{z}_{-1}) = \Lambda^{(2,1)}_{j,-1}(\widetilde{z}_1, \widetilde{z}_0, \widetilde{z}_{-1})\ \forall \widetilde{z}_1 \in Z_1, \widetilde{z}_0 \in Z_0, \widetilde{z}_{-1} \in Z_{-1}, j \in [2:L] \end{array} \right\}.$$

满足 $|C_1| \geq |A_L| \cdots |A_2| \cdot 2^{-2Hdp\cdot(n_1\cdots n_{L-1})\cdot L}$.

对于第二个轮次，目标是为每个 $\widetilde{z}_1 \in Z_1, \widetilde{z}_0 \in Z_0, \widetilde{z}_{-1} \in Z_{-1}$ 固定 $\Lambda^{(2,2)}_{j,-1}(\widetilde{z}_1, \widetilde{z}_0, \widetilde{z}_{-1})$。关键在于，这段通信记录仅取决于信息状态 $X^{(1)}_{-1}$ 和 $X^{(1)}_j$，而它们本身不受 $z_1 \in Z_1$ 的选择影响。这种独立性成立，是因为 $X^{(1)}_{-1}$ 和 $X^{(1)}_j$ 中唯一取决于 z_1 的部分，是玩家 1 发给 -1 的第一个轮次消息（回顾 $a_1 \leq 1$, (12)），而对于所有 $\widetilde{z}_1 \in Z_1$，它都固定为 $\Psi^{(1)}_{1,-1}(\widetilde{z}_{-1})$ (13)。因此，只需固定

$$\Phi^{(2)}_{j,-1} := \left(\Phi^{(2)}_{j,-1}(\widetilde{z}_0, \widetilde{z}_{-1})\right)_{\widetilde{z}_0\in Z_0, \widetilde{z}_{-1}\in Z_{-1}}$$

然后定义 $\Lambda^{(2,2)}_{j,-1}(\widetilde{z}_1, \widetilde{z}_0, \widetilde{z}_{-1}) = \Phi^{(2)}_{j,-1}(\widetilde{z}_0, \widetilde{z}_{-1})\ \forall \widetilde{z}_1 \in Z_1, \widetilde{z}_0 \in Z_0, \widetilde{z}_{-1} \in Z_{-1}$。通信记录总数至多为 $2^{2Hdp\cdot x_0\cdot(n_1\cdots n_{L-1})}$。因此，我们可以选择 $\{\Lambda^{(2,2)}_{j,-1}\}_{j\in[2:L-1]}$，使得

$$C_2 := \left\{ \begin{array}{c} (\widetilde{z}_L, \ldots, \widetilde{z}_2) \in C_1 : \\ \Pi^{(2)}_{j,-1}(\widetilde{z}_L, \ldots, \widetilde{z}_0, \widetilde{z}_{-1}) = \Lambda^{(2,2)}_{j,-1}(\widetilde{z}_1, \widetilde{z}_0, \widetilde{z}_{-1})\ \forall \widetilde{z}_1 \in Z_1, \widetilde{z}_0 \in Z_0, \widetilde{z}_{-1} \in Z_{-1}, j \in [2:L] \end{array} \right\}.$$

满足 $|C_2| \geq |C_1| \cdot 2^{-2Hdp \cdot x_0 \ (n_1 \cdots n_{L-1}) \ \cdot L} \geq |A_L \cdots A_2| \cdot 2^{-4HdpL \cdot x_0 \ (n_1 \cdots n_{L-1})}$。

步骤 2.2：将通信记录固定为参与方0。 $\{\Lambda^{(2,\ell')}{}_{j,0} \}_{j\in[2:L],\ell'\in[2]}$ 的总数至多为 $2^{2Hdp\cdot x_0 x_1 \cdot 2L}$。我们可以固定其取值，使得

$$C_3 := \left\{ \begin{matrix} (\widetilde{z}_L, \ldots, \widetilde{z}_2) \in C_2 : \\ \Pi_{j,0}^{(\ell')}(\widetilde{z}_L, \ldots, \widetilde{z}_0) = \Lambda_{j,0}^{(2,\ell')}(\widetilde{z}_1, \widetilde{z}_0) \ \forall \widetilde{z}_1 \in Z_1, \widetilde{z}_0 \in Z_0, j \in [2:L], \ell' \in [2] \end{matrix} \right\}.$$

满足 $|C_3| \geq |C_2| \cdot 2^{-2Hdp\cdot x_0 x_1 \cdot 2L} \geq |A_L \cdots A_2| \cdot 2^{-6HdpL\cdot x_0 (n_1 \cdots n_{L-1})}$。

步骤 2.3：将通信记录固定为参与方 1。 $\{\Lambda^{(2,\ell')}{}_{j,1} \}_{j\in[2:L],\ell'\in[2]}$ 的总数至多为 $2^{2Hdpm\cdot x_1 \cdot 2L}$，而且 $\{\Xi^{(\ell')}{}_2{}^{m}\}_{\ell'\in[2],m\in[0,a_{\ell'}-1]}$ 的总数至多为 $2^{Hd(d+1)p(a_1+a_2)}$，我们可以固定其取值，使得

$$C_4 := \left\{ \begin{matrix} (\widetilde{z}_L, \ldots, \widetilde{z}_2) \in C_3 : \quad \Sigma_2^{(\ell')}(\widetilde{z}_L, \ldots, \widetilde{z}_1) = \Xi_2^{(\ell')}, \forall \ell' \in [2] \text{ and} \\ \Pi_{j,1}^{(\ell')}(\widetilde{z}_L, \ldots, \widetilde{z}_1) = \Lambda_{j,1}^{(2,\ell')}(\widetilde{z}_1) \ \forall \widetilde{z}_1 \in Z_1, j \in [2:L], \ell' \in [2], \end{matrix} \right\},$$

满足以下界

$$\begin{aligned} C_4 \geq C_3 \cdot 2^{-2Hdpm\cdot x_1 \cdot 2L} \cdot 2^{-2Hdp} &\geq |A_L| \cdots |A_2| \cdot 2^{-8HdpL\cdot x_0 (n_1\cdots n_{L-1}) - Hd(d+1)p(a_1+a_2)} \\ &\geq |A_L| \cdots |A_2| \cdot 2^{-4\sqrt{K} x_0 (n_1 \cdots n_{L-1})} = |A_L| \cdots |A_2| / \Delta_2 \end{aligned} \tag{15}$$

此处，第二步由参数选择得出（见式（6）（8）（12）），最后一步由 Δ_2 的定义得出（见式（10））。

结合引理 B.8 与式（15），我们完成基本情形 $\ell = 2$的证明。

B.3. 归纳步骤

假设定理 B.7 成立直至 $\ell \in [2: \ L-1]$。我们证明它对于 $\ell + 1$继续成立。

关键洞见在于，当这些参与者从 $Z_{<\ell}$获取输入时，经过 ℓ 轮次后， Z_ℓ 对参与者 $[-1: \ \ell-1]$ 不可区分。因此，发给参与者 $[-1: \ \ell-1]$ 的第（$\ell+1$）轮次通信记录与 $z_\ell \in Z_\ell$的选择无关。

步骤 1：选择集合 Z_ℓ。 回顾一下， $R_{\geq \ell}$的大小满足 $|R_{\geq\ell}| \geq |A_L| \times \cdots \times |A_\ell| / \Delta_\ell$。我们希望从 $R_{\geq\ell}$中选取一个矩形子集（见 Chen 等人，2025年），有如下引理。

引理 B.9 （（Chen 等人，2025年），引理 4.7）。 *存在一个子集 $S^{(\ell)} \subseteq R_{\geq\ell}$ 使得*

- $S^{(\ell)} = S_1^{(\ell)} \times S_2^{(\ell)}$, *where* $S_1^{(\ell)} \subseteq A_L \times \cdots \times A_{\ell+1}$, $S_2^{(\ell)} \subseteq A_\ell$, *such that*

$$|S_1^{(\ell)}| \geq |A_L| \cdots |A_{\ell+1}| / \Delta_\ell^{2x_\ell} \quad \textit{and} \quad |S_2^{(\ell)}| = x_\ell.$$

- $|\{i_\ell : i_\ell = \widetilde{z}_\ell(\widetilde{w}_{\ell-1}, \widetilde{i}_{\ell-1}) \textit{ for some } \widetilde{w}_{\ell-1} \in [n_{\ell-1}], \widetilde{i}_{\ell-1} \in \mathcal{I}_{\ell-1}, \widetilde{z}_\ell \in S_2^{(\ell)}\}| \geq \Theta_\ell$.

利用引理B.9，我们取 $Z_\ell = S_2^{(\ell)}$ 和 $S_{\geq \ell+1} = S_1^{(\ell)}$。

接下来，我们将固定通信记录 $\Lambda^{(\ell+1)}$。回顾一下，我们需要在前 $\ell+1$ 轮次中，固定从参与者 $[\ell+1: \ L]$ 到参与者 $[-1: \ \ell]$ 的所有通信记录；在此期间，参与者 $[-1: \ \ell]$ 从 $Z_{\leq\ell} = Z_{-1} \times Z_0 \times \cdots \times Z_\ell$接收输入。我们分几步进行。

步骤 2.1：将前 ℓ 轮次中参与者 $[-1$ 的通信记录固定下来： $\ell-1]$ 我们只需使用 $\Lambda^{(\ell)}$，也就是说，对于任意 $\widetilde{z}_\ell \in Z_\ell$、.....、 $\widetilde{z}_i \in Z_i$,

$$\Lambda_{j,i}^{(\ell+1,\ell')}(\widetilde{z}_\ell, \ldots, \widetilde{z}_i) = \Lambda_{j,i}^{(\ell,\ell')}(\widetilde{z}_{\ell-1}, \ldots, \widetilde{z}_i). \quad \forall j \in [\ell+1:L], i \in [-1:\ell-1], \ell' \in [\ell] \tag{16}$$

与（Chen 等人，2025年）中的论述一样， $S_{\geq\ell+1} \subseteq A_L \times \cdots \times A_{\ell+1}$ 与 $\Lambda^{(\ell+1)}$ 截至目前是一致的。

引理 B.10。 集合 $S_{\geq \ell+1}$ 与 $\{\Lambda^{(\ell,\ell')}{}_{j,\ i}\}_{j\in[\ell+1:L],\ i\in[-1:\ell-1],\ \ell'\in[\ell]}$一致。形式上，对于任意 $\widetilde{z}_{\geq \ell+1}\in S_{\geq \ell+1}$ 和 $\widetilde{z}_{<\ell+1}\in Z_{<\ell+1}$，有

$$\Pi^{(\ell')}_{j,i}(\widetilde{z}_L,\ldots,\widetilde{z}_i)=\Lambda^{(\ell+1,\ell')}_{j,i}(\widetilde{z}_\ell,\ldots,\widetilde{z}_i).$$

对于任意 $j\in[\ell+1:\ L]$、$i\in[-1:\ \ell-1]$、$\ell'\in[\ell]$。

步骤 2.2：将通信记录固定到参与者 $[-1:\ \ell-1]$ 在第 $(\ell+1)$ 轮次。 我们的关键洞见是，当从 $Z_{\leq \ell-1}$ 获取输入时，Z_ℓ 对参与者 $[-1$ 而言是不可区分的：$\ell-1]$。因此，他们的通信记录与 $z_\ell\in Z_\ell$ 的选择无关。我们考虑

$$\Phi^{(\ell+1)}=\left(\Phi^{(\ell+1)}_{j,i}\right)_{j\in[\ell+1:L],i\in[-1:\ell-1]}$$

其中

$$\Phi^{(\ell+1)}_{j,i}=\left(\Phi^{(\ell+1)}_{j,i}(\widetilde{z}_{\ell-1},\ldots,\widetilde{z}_i)\right)_{\widetilde{z}_{\ell-1}\in Z_\ell\ldots,\widetilde{z}_i\in Z_i}\quad\text{and}\quad\Phi^{(\ell+1)}_{j,i}(\widetilde{z}_{\ell-1},\ldots,\widetilde{z}_i)\in\mathsf{domain}(\Pi^{(\ell+1)}_{j,i})$$

比较 $\Lambda^{(\ell+1,\ell+1)}_{j\ i}$、与 $\Phi^{(\ell+1)}_{j\ i}$，我们注意到 $\Phi^{(\ell+1)}_{j\ i}$ 与 $\widetilde{z}_\ell\in Z_\ell$ 无关。对于任意 $\Phi^{(\ell+1)}$，定义

$$S(\Phi^{(\ell+1)}):=\left\{\begin{array}{c}(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1})\in S_{\geq\ell+1}:\\ \Pi^{(\ell+1)}_{j,i}(\widetilde{z}_L,\ldots,\widetilde{z}_i)=\Phi^{(\ell+1)}_{j,i}(\widetilde{z}_{\ell-1},\ldots,\widetilde{z}_i)\\ \forall\widetilde{z}_\ell\in Z_\ell,\ldots,\widetilde{z}_i\in Z_i,j\in[\ell+1:L],i\in[-1:\ell-1]\end{array}\right\}\tag{17}$$

换言之，$S(\Phi^{(\ell+1)})$ 包含所有 $(\widetilde{z}_L$、.....、$\widetilde{z}_{\ell+1})\in S_{\geq\ell+1}$ 与通信记录 $\Phi^{(\ell+1)}$一致的输入。以下引理的证明不同于（Chen 等人，2025年），因为在我们的混合通信模型中，必须考虑线性通信记录。

引理 B.11。我们有

$$\bigcup_{\Phi^{(\ell+1)}}S(\Phi^{(\ell+1)})=S_{\geq\ell+1}.$$

证明。 只需证明，对于任意的 $(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1})\in S_{\geq\ell+1},\widetilde{z}_{\ell-1}\in Z_{\ell-1},\ldots,\widetilde{z}_i\in Z_i,j\in[\ell+1:\ L],i\in[-1:\ \ell-1]$，通信记录 $\Pi^{(\ell+1)}_{j\ i},(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1},z_\ell,\widetilde{z}_{\ell-1},\ldots\widetilde{z}_i)$ 对每个 $z_\ell\in Z_\ell$ 都相同。

为证明这一点，注意到通信记录 $\Pi^{(\ell+1)}_{j\ i}$,由信息状态 $X^{(\ell)}_j$ 和 $X^{(\ell)}_i$ 决定。显然，$X^{(\ell)}_j$ 不会随 $z_\ell\in Z_\ell$ 的选择而改变，因为 $j>\ell$。因此，只需证明 $X^{(\ell)}_i$ 也不会随 $z_\ell\in Z_\ell$ 改变。我们证明，所有信息状态 $\{X^{(\ell')}{}_{r}{}^{m}\}_{r\in[i:\ell-1],\ell'\in[\ell]}$ 都不会随 z_ℓ 改变。

回想一下，我们已固定 $(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1})\in S_{\geq\ell+1}$和 $\widetilde{z}_{\ell-1}\in Z_{\ell-1},\ldots,\widetilde{z}_i\in Z_i$的值。我们通过对 r从 $r=\ell-1$ 到 $r=i$作反向归纳来证明这一点。当 $r=\ell-1$时，信息状态 $X^{(\ell')}{}_{\ell-1}$ 由 $\widetilde{z}_{\ell-1},\Pi^{(\ell'')}{}_{t,\ell-1}(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1},z_\ell,\widetilde{z}_{\ell-1})$，以及 $\Sigma^{(\ell'')}_\ell{}^{m}(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1},z_\ell,\widetilde{z}_{\ell-1})\ (t\in[\ell:\ L],\ell''\in[\ell'],m\in[0,a_{\ell''}-1])$ 确定。

由于对于每个 $z_\ell\in Z_\ell$，$(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1},z_\ell)\in R_{\geq\ell}$，因此我们有 $\Pi^{(\ell'')}{}_{t,\ell-1}(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1},z_\ell,\widetilde{z}_{\ell-1})=\Lambda^{(\ell,\ell'')}{}_{t,\ell-1}(\widetilde{z}_{\ell-1})$ and$_{\Sigma^{(\ell'')}_\ell{}^{m}}(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1},z_\ell,\widetilde{z}_{\ell-1})=\Xi^{(\ell'')}_\ell{}^{m}(\widetilde{z}_{\ell-1})$，而这些对于每个 $z_\ell\in Z_\ell$ 都相同。至此，基本情形的证明完成。现在假设断言对 $r+1$ 成立。那么，对于 $r,X^{(\ell')}_r$ 由 $\widetilde{z}_r,\Pi^{(\ell'')}{}_{t,r}(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1},z_\ell,\widetilde{z}_{\ell-1},\ldots,\widetilde{z}_r)$ 确定，并且 $\Sigma^{(\ell'')}_{r+1}{}^{m}(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1},z_\ell,\widetilde{z}_{\ell-1},\ldots,\widetilde{z}_r)\ (t\in[r+1:\ L],\ell''\in[\ell'],m\in[0,a_{\ell''}-1])$。

对于 $t\in[\ell:\ L]$，由于 $(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1},z_\ell)\in R_{\geq\ell}$，我们有 $\Pi^{(\ell'')}{}_{t,r}(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1},z_\ell,\widetilde{z}_{\ell-1},\ldots,\widetilde{z}_r)=\Lambda^{(\ell,\ell'')}{}_{t,r}(\widetilde{z}_{\ell-1},\ldots,\widetilde{z}_r)$，这对每个 $z_\ell\in Z_\ell$ 都相同。对于 $t\in[r+1:\ \ell-1]$，我们已证明 $X^{(\ell'')}{}_t$ 对每个 $z_\ell\in Z_\ell$ 都相同，因此 $\Pi^{(\ell'')}{}_{t,r}(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1},z_\ell,\widetilde{z}_{\ell-1},\ldots,\widetilde{z}_r)$ 也是如此。此外，$\Sigma^{(\ell'')}_{r+1}{}^{m}(\widetilde{z}_L,\ldots,\widetilde{z}_{\ell+1},z_\ell,\widetilde{z}_{\ell-1},\ldots,\widetilde{z}_r)$ 仅依赖于 $X^{(\ell'')}_{r+1}{}^{m}$，而这对每个 $z_\ell\in Z_\ell$ 都相同。至此完成归纳并结束证明。 □

现在，如同在（Chen 等人，2025年）中一样，我们得到

引理 B.12. 存在 $\widetilde{\Phi}^{(\ell+1)}$ 使得

$$|S(\widetilde{\Phi}^{(\ell+1)})| \geq |A_L| \cdots |A_{\ell+1}| \cdot 2^{-2\sqrt{K}\cdot(x_0\cdots x_{\ell-1})\cdot(n_1\cdots n_{L-1})}$$

集合 $T_{\geq \ell+1} = S(\widetilde{\Phi}^{(\ell+1)}) \subseteq S_{\geq \ell+1}$ 与通信记录 $(\Lambda^{(\ell+1,\ell')}{}_{j,i})_{i\in[-1:\ell-1],j\in[\ell+1:L],\ell'\in[\ell+1]}$ 一致。形式上，对于任意 $(\widetilde{z}_L, \ldots, \widetilde{z}_{\ell+1}) \in T_{\geq \ell+1}$ 和 $\widetilde{z}_{<\ell+1} \in Z_{<\ell+1}$，有

$$\Pi^{(\ell')}_{j,i}(\widetilde{z}_L, \ldots, \widetilde{z}_i) = \Lambda^{(\ell+1,\ell')}_{j,i}(\widetilde{z}_\ell, \ldots, \widetilde{z}_i). \tag{18}$$

对于任意 $j \in [\ell+1 : L], i \in [-1 : \ell-1], \ell' \in [\ell+1]$。此外，我们有

$$|T_{\geq \ell+1}| \geq |A_L \times \cdots \times A_{\ell+1}| \cdot 2^{-2\sqrt{K}\cdot(x_0\cdots x_{\ell-1})\cdot(n_1\cdots n_{L-1})}. \tag{19}$$

步骤 2.3：将通信记录固定给参与方 ℓ。 这遵循贪心选择策略。令

$$\Psi = \left(\Psi^{(\ell')}_{j,\ell}(\widetilde{z}_\ell)\right)_{j\in[\ell+1:L],\ell'\in[\ell+1],\widetilde{z}_\ell\in Z_\ell} \quad \text{where} \quad \Psi^{(\ell')}_{j,\ell}(\widetilde{z}_\ell) \in \mathsf{domain}(\Pi^{(\ell')}_{j,\ell})$$

定义

$$T(\Psi) := \left\{ \begin{array}{c} (\widetilde{z}_L, \ldots, \widetilde{z}_{\ell+1}) \in T_{\geq \ell+1} : \\ \Pi^{(\ell')}_{j,\ell}(\widetilde{z}_L, \ldots, \widetilde{z}_\ell) = \Psi^{(\ell')}_{j,\ell}(\widetilde{z}_\ell) \quad \forall \widetilde{z}_\ell \in Z_\ell, \ell' \in [\ell+1], j \in [\ell+1:L] \end{array} \right\} \tag{20}$$

与 Chen 等人，2025年 中一样，我们可以对不同的 Ψ 数量给出上界，并利用鸽巢原理得到以下引理。

引理 B.13。 Ψ 的总数至多为 $2^{\sqrt{K}\cdot(x_0\cdots x_{\ell-1})\cdot(n_1\cdots n_{L-1})}$。因此，存在 $\widetilde{\Psi}$，使得 $|T(\widetilde{\Psi})| \geq |A_L| \cdots |A_{\ell+1}| 2^{-3\sqrt{K}\cdot(x_0\cdots x_{\ell-1})\cdot(n_1\cdots n_{L-1})}$。

根据引理 B.13，我们固定参与者 $j \in [\ell+1 : L]$ 到参与者 ℓ 在前 $\ell+1$ 轮次中使用的通信记录，采用 Ψ。特别地，我们取

$$\Lambda^{(\ell+1,\ell')}_{j,\ell}(\widetilde{z}_\ell) = \widetilde{\Psi}^{(\ell')}_{j,\ell}(\widetilde{z}_\ell), \quad \forall j \in [\ell+1:L], \ell' \in [\ell+1], \widetilde{z}_i \in Z_i. \tag{21}$$

接下来，我们固定 $\left(\Xi^{\ell'}_{\ell+1}\right)_{\ell'\in[\ell+1]}$。可选方案的总数为 $2^{2Hdp(a_1+a_2+\cdots+a_{\ell+1})}$，因此存在一种选择，使其一致性集合的大小至少为

$$|A_L| \cdots |A_{\ell+1}| \cdot 2^{-3\sqrt{K}\cdot(x_0\cdots x_{\ell-1})\cdot(n_1\cdots n_{L-1}) - Hd(d+1)p(a_1+a_2+\cdots+a_{\ell+1})} \geq |A_L| \cdots |A_{\ell+1}|/\Delta_{\ell+1}$$

by (12)。随后，我们可以将 $R_{\geq \ell+1}$ 设为这个一致性集合，从而完成归纳步骤 p.

C. 全注意力的上界

本节为定理A.3、A.6、A.5和A.8中给出的上界提供了构造性证明。具体而言，我们设计了配备全注意力的 Transformer 解码器，使其能够成功完成相应任务。我们设计的核心是检索头机制，改编自（Chen 等人，2025年）。我们指出，也可以采用利用近乎正交向量的另一种更隐式的构造（Bhattamishra 等人，2024年）。

检索任务

输入。对于前 n 个输入词元 $x_1, \cdots, x_n$，每个词元 x_i 由向量 $a_i \in \{0,1\}^D$ 和 $b_i \in \{0,1\}^D$ 构成，而最后一个词元 x_{n+1} 包含一个查询向量 a。**任务。**找到位置 $i \in [n]$，使得 $a_i = a$，并返回对应的值 b_i。**输出。**如果存在唯一的 i 满足 $a_i = a$，则输出 b_i。否则，输出可以是任意值。

我们现在详细说明如何使用单个注意力头实现这一检索操作。

检索头的实现。我们将值投影 V设为 b_i，并将位置 $i \in [n]$ 的键投影 K 设为$\log^2(n) \cdot (a_i, \vec{1} - a_i)$（针对位置 $i \in [n]$），其中 $\vec{1} \in \{0,1\}^D$ 表示长度为 D的全一向量，$\vec{1} - a_i$ 表示逐元素相减；位置 $n+1$ 处的查询投影 Q 取为$\log^2(n) \cdot (a, \vec{1} - a)$。注意力分数（softmax之前）满足

$$\langle Qx_{n+1}^{(\ell)}, Kx_i^{(\ell)} \rangle = \begin{cases} \log^2(n) D & a_i = a \\ \leq \log^2(n) D - \log^2(n) & a_i \neq a. \end{cases}$$

因此，如果恰有一个位置 $i \in [n]$ 满足 $a_i = a$，那么注意力概率满足

$$\alpha_{n+1,i} \geq \frac{\exp(\log^2(n))}{\exp(\log^2(n)) + n - 1} \geq 1 - \frac{n}{n^{\log(n)}},$$

在精度为 $p = \Theta(\log(n))$ 的假设下，它与 1 不可区分，并且

$$\alpha_{n+1,j} \leq \frac{1}{n^{\log(n)}},$$

对于所有 $j \neq i$，在精度为 $p = \Theta(\log(n))$ 的假设下，它与 0 不可区分。我们得出结论，在 $p = \Theta(\log n)$ 位精度下，注意力头将只关注位置 i，并检索值 b_i。 □

下面，我们将说明这一机制如何实现 Eva 和 PerCom 的求解。

构造 Eva。 函数求值任务可以表述为检索任务。在此，输入的最后一个令牌表示查询 $a = x \in [n]$。对于每个更早的位置 $i \in [n]$（其词元编码值 $f(i)$），我们设定 $a_i = i$和 $b_i = f(i)$。目标是找到满足 $i = x$ 的唯一位置 i，并输出 $f(i)$。直接应用上一小节中的检索头即可得到 $f(x)$。

PerCom 的构造 PerCom。 PerCom 可以视为 n 个检索任务。输入的最后 n 个词元表示元素 $\tau(1) \cdots, \tau(n)$。每个词元都为其所在位置的检索操作提供一个查询 a。对于每个之前的词元 $j \in [n]$，第 j 个词元对应于 $\sigma(j)$，并设置 $a_j = j$ 和 $b_j = \sigma(j)$。为了计算 $\sigma(\tau(i))$，模型执行检索任务，找到唯一的位置 $j \in [n]$，使得 $j = \tau(i)$，然后输出值 $\sigma(j)$。因此，检索头的实现完成了 PerCom。

D. 部分证明缺失

引理的证明2.2。 假设我们有一个维度为 d、精度为 p 的线性注意力头，其查询、键和值矩阵分别为 Q, K 和 V。设 $S_i = S_{i-1} + Vx_i \otimes \varphi(Kx_i)$ $Z_i = Z_{i-1} + \varphi(Kx_i)$，其中 $S_0 = 0$ 且 $Z_0 = 0$，则有

$$y_i = \frac{\varphi(Qx_i)^\top S_i}{\varphi(Qx_i)^\top Z_i}.$$

令 $\mathrm{h}_0 = (S_0, Z_0) \in \mathbb{R}^{2d}$, $g(x, \mathrm{h} = (S, Z)) = (S + Vx \otimes \varphi(Kx)\ Z + \varphi(Kx))$，并且

$$f(x, \mathrm{h} = (S, Z)) = \frac{\varphi(Qx)^\top S}{\varphi(Qx)^\top Z}$$

其中 g 和 f 均与时间步 i 无关，很容易验证，相应的循环神经网络会计算出与给定线性注意力层相同的输出。 □