打开黑箱：通过语义共振架构实现可解释的 LLMs

Ivan Ternovtsii
Department of Information Technologies
Uzhhorod National University, Ukraine
HengeBytes
ivan.ternovtsii@uzhnu.edu.ua
https://github.com/ITernovtsii/semantic-resonance

Abstract

大型语言模型(LLMs)表现出卓越的性能,但仍然难以解释。专家混合(MoE)模型通过稀疏激活提高效率，但通常依赖于不透明的学习门控函数。虽然基于相似性的路由（余弦路由器）已被探索用于训练稳定化，其固有可解释性的潜力仍未得到充分利用。我们引入了语义共振架构（SRA），这是一种设计为确保路由决策具有内在可解释性的 MoE 方法。SRA 用语义共振室（CSR）模块替换了学习门控，并基于与可训练的语义锚点的余弦相似性进行令牌路由。我们还介绍了一种新颖的离散损失，以鼓励锚点之间的正交性，从而强制执行多样化的专业化。在 WikiText-103 上的实验表明，SRA 实现了验证困惑度为 13.41，在匹配的活跃参数约束（29.0M）下优于密集基线（14.13）和标准 MoE 基线（13.53）。至关重要的是，SRA 表现出更优的专家利用率（1.0%死亡专家对比标准 MoE 中的 14.8%），并发展出不同的、语义连贯的专业化模式，与在标准 MoEs 中观察到的噪声专业化不同。这项工作确立了语义路由作为一种建立更加透明和可控的语言模型的强大方法论。

1 介绍

大型语言模型（LLMs）的迅速发展彻底改变了自然语言处理 [1, 3]。然而，这些模型的不透明性在需要可解释和可控决策的关键领域中部署时带来了重大挑战 [12]。

专家混合（MoE）架构通过有条件地激活参数子集来解决密集模型的计算低效问题 [13]。然而，标准方法，如 Switch Transformer[6] 和 GShard[10]，依赖于学习到的门控函数（通常是简单的线性层），其决策过程仍然不透明。

最近的工作探讨了替代路由机制，包括基于余弦相似度的路由（余弦路由），主要是为了提高训练稳定性和负载均衡 [2]。虽然这些方法在稳定化方面有效，但它们并未侧重于利用这种机制实现固有的可解释性或强制一致的专业化。

我们提出了语义共鸣架构（SRA），这是一种利用语义相似性进行路由的方法，特别设计用于增强可解释性。我们的关键技术——语义共鸣室（CSR）根据标记表示与每个专家相关联的可学习语义锚点之间的余弦相似性将标记路由到专家。为了确保这些锚点捕捉多样化的概念，我们引入了分散损失，以积极促进锚点间的正交性。

Preprint.

该机制提供了显著的优势：

1. **内在可解释性：**路由决策可以直接通过语义相似度分数来解释，消除了学习门限的不透明性。
2. **一致特化：**在色散损耗的帮助下，专家们开发了稳定且语义连贯的专业化模式，这与标准 MoEs 中经常观察到的嘈杂专业化不同。
3. **改进的利用率：**语义路由与训练稳定技术（如渐进路由）表现出优异的协同作用，从而实现更好的负载均衡并减少不活跃专家。

我们的贡献总结如下：

- 我们介绍了语义共振架构（SRA）和一种新颖的分散损失，证明了可以通过设计使用语义路由来强制实现可解释的专业化。
- 我们证明了 SRA 在 WikiText-103 上优于密集和标准 MoE 基线，同时保持相同的活跃参数预算并表现出更优的专家利用。
- 我们提供了一个全面的分析，显示 SRA 专家比标准 MoE 基线开发出更具意义的语义专业化。

2 相关工作

2.1 专家混合与路由机制

MoE 模型范式可以追溯到 [9]。最近的研究兴趣集中在高效扩展变压器模型上。Shazeer 等人 [13] 引入了稀疏门控 MoE 层，而 Switch Transformer[6] 将其简化为 Top-1 路由。

这些标准技术依赖于学习到的门控网络（例如，线性变换后跟 softmax）。虽然高效，但这些路由器对路由决策背后的原因提供的见解有限。

与我们的工作最相关的是基于相似性的路由探索，通常被称为余弦路由器。这些在模型如 XMoE[2] 中主要被用于稳定训练。SRA 建立在此机制之上，但将重点从单纯的稳定性转移到了内在可解释性。我们引入分散损失以主动强制语义多样性与正交性，这是之前基于相似性的方法中缺失的连贯专业化的重要组成部分。

2.2 神经网络的可解释性

可解释性研究包括注意力可视化 [16, 4]、探测分类器 [15] 和机制可解释性 [5]。

对于 MoE 架构，标准 MoEs 中专家的专业化 [7] 的后验分析通常发现这些专家并未专门针对可识别的语义类别。这些方法是在事后提供解释性而非设计时就考虑进去的。相比之下，我们的语义共振机制确保了解释性和连贯的专业化本身就是路由过程固有的属性。

3 方法

3.1 架构概述

语义共振架构基于标准的变压器解码器架构，用我们的语义共振室（CSR）模块替换了每一层中的前馈网络（FFN）。该架构利用了标准的学习嵌入和权重绑定、旋转位置嵌入（RoPE）[14] 以及预归一化配置。

形式上，对于输入序列 $X \in \mathbb{R}^{B\times L\times D}$，其中 B 是批处理大小，L 是序列长度，D 是模型维度，每个 SRA 块计算：

$$X' = \text{LayerNorm}(X) \tag{1}$$
$$H = X + \text{MultiHeadAttention}(X') \tag{2}$$
$$H' = \text{LayerNorm}(H) \tag{3}$$
$$Y = H + \text{CSR}(H') \tag{4}$$

3.2 语义共振室

CSR 模块根据令牌与可学习语义锚点的“共鸣”（相似性）来路由令牌。

3.2.1 语义锚点和初始化

我们初始化一组可学习的语义锚点 $A \in \mathbb{R}^{N\times D}$，其中 N 是专家的数量。我们使用正交初始化来最大化语义空间中的初始分散性（当 $N \le D$ 时）：

$$A = \text{orth}(\text{randn}(N, D)) \tag{5}$$

我们将此与 Kaiming 均匀初始化 [8] 进行比较。

3.2.2 共振计算

对于每个标记表示 $\boldsymbol{h} \in \mathbb{R}^D$，我们使用余弦相似度计算与所有锚点的共振得分。该计算在 FP32 中进行以确保数值稳定性。

$$r_i = \cos(\boldsymbol{h}, \boldsymbol{a}_i) = \frac{\boldsymbol{h}\cdot\boldsymbol{a}_i}{\|\boldsymbol{h}\|_2 \cdot \|\boldsymbol{a}_i\|_2 + \epsilon} \tag{6}$$

其中 $\boldsymbol{a}_i$ 是第 i 个语义锚点，而 $\epsilon = 1e-8$。

3.2.3 Top-k 专家选择与执行

我们选择具有最高共鸣分数的前 k 位专家。在训练过程中，可以选择性地添加高斯噪声 $\eta \sim \mathcal{N}(0, \sigma^2)$（带噪前 k 位）[13]。

$$\text{indices}, \text{scores} = \text{top\_k}(\boldsymbol{r} + \eta, k) \tag{7}$$
$$\text{weights} = \text{softmax}(\text{scores}) \tag{8}$$

最终输出是选定专家输出的加权组合：

$$\boldsymbol{y} = \sum_{i\in\text{indices}} w_i \cdot \text{Expert}_i(\boldsymbol{h}) \tag{9}$$

每个专家都是一个使用 GELU 激活函数的两层前馈网络。

3.3 训练目标

总损失函数结合了主要的语言建模目标（$\mathcal{L}_{\text{LM}}$）和辅助损失：

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \alpha \cdot \mathcal{L}_{\text{balance}} + \beta \cdot \mathcal{L}_{\text{dispersion}} + \gamma \cdot \mathcal{L}_{\text{z}} \tag{10}$$

3.3.1 负载均衡损失

为了确保均衡使用，我们采用基于路由概率 [6] 的变异系数（CV）平方的负载平衡损失。

我们首先计算每个专家在批次中被选中的平均概率：

$$P_{\text{mean}} = \frac{1}{B \cdot L} \sum_{b,l} \text{softmax}(\boldsymbol{r}_{b,l}) \tag{11}$$

损失是这些平均概率在 N 个专家之间的平方 CV（变异系数）：

$$\mathcal{L}_{\text{balance}} = N \cdot \frac{\text{Var}(P_{\text{mean}})}{\text{Mean}(P_{\text{mean}})^2 + \epsilon} \tag{12}$$

3.3.2 色散损耗

为了鼓励语义锚点学习多样且不同的概念，我们引入了一个分散损失 。该损失通过最小化所有唯一成对组合的平均成对余弦相似性来惩罚锚点之间的相似性 。通过在语义空间中将锚向量相互推开，这一目标鼓励每个专家发展独特的专长。

$$\mathcal{L}_{\text{dispersion}} = \frac{1}{N(N-1)} \sum_{i \neq j} \cos(\boldsymbol{a}_i, \boldsymbol{a}_j) \tag{13}$$

3.3.3 路由器 Z 损耗

为了数值稳定性，我们探索了路由 z 损失 [17]：

$$\mathcal{L}_{\text{z}} = \frac{1}{B \cdot L} \sum \left(\log \left(\sum_j \exp(r_{ij}) \right) \right)^2 \tag{14}$$

3.4 逐步路由

为了减轻从一开始就使用 Top-k>1 训练时经常观察到的专家崩溃现象，我们采用了一种渐进路由策略。模型在初始时期使用 Top-1 路由进行训练，允许专家建立不同的专长。随后，将路由切换到 Top-2，使模型能够利用专门化专家的组合。

4 实验

4.1 实验设置

4.1.1 数据集

我们在 WikiText-103 上评估了我们的方法 [11]。我们使用字节配对编码（BPE），词表大小为 32,000 个标记。

4.1.2 模型配置

我们将 SRA 与标准密集变压器和具有学习门控的标准 MoE 模型进行了比较。所有模型的每个令牌的活跃参数数量相匹配（≈29.0M）。所有模型使用 $D = 512$，4 层和 8 个注意力头。

- **SRA（我们的）：**半导体路由（CSR）。每层有 N=128 个专家（总共 512 个）。Top-k=2。专家 FFN 维度 (D_{ff})=1024。总参数量：558.5M。活跃参数量：29.0M。
- **标准 MoE：**学习的线性门控。N=128。Top-k=2。D_{ff}=1024。总参数：558.5M。活跃参数：29.0M。
- **密集基线：**标准 FFN。D_{ff}=2048。总参数：29.0M。活跃参数：29.0M。

密集基线 FFN 维度（2048）是 MoE 专家维度（1024）的两倍，以匹配来自 Top-2 路由的有效参数数量。

4.1.3 训练详情

所有模型使用 AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$) 和渐进式路由策略进行训练（第 1-5 轮：Top-1；第 6-10 轮：Top-2）。学习率：3e-4，采用线性预热（4000 步）和余弦衰减。总批次大小：128。序列长度：256。dropout：0.1。

对于 SRA 超参数：α（平衡）=0.4；β（离散度）=0.6；γ（z 损失）=0.0。

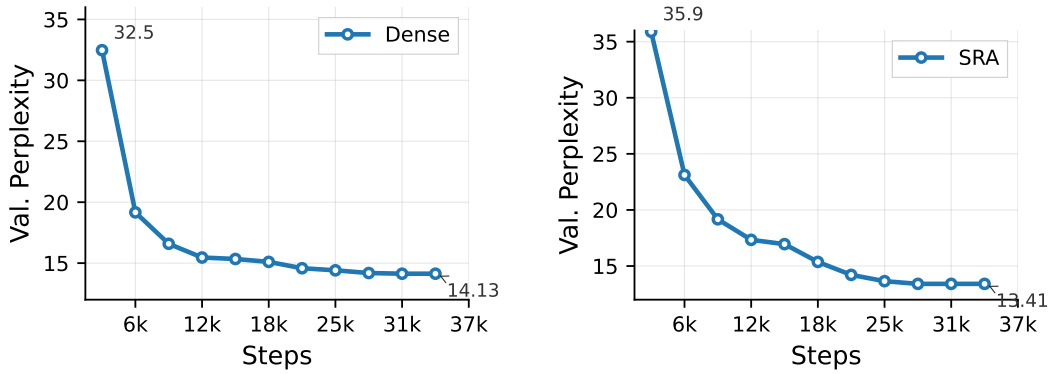
4.2 主要结果

表 1: 困惑度比较在 WikiText-103 验证集上。活跃参数（≈29.0M）在所有模型中匹配。总专家数=512。

Model	Routing	Total Params	Val PPL	Dead Experts (%)
Dense Baseline	N/A	29.0M	14.13	N/A
Standard MoE	Learned	558.5M	13.53	76 (14.8%)
SRA（我们的）	**语义**	**558.5M**	**13.41**	**5 (1.0%)**

SRA 的性能优于密集基线和标准 MoE 模型（表 1）。关键的是，SRA 达到了与标准 MoE 相当的性能（略高于标准 MoE），同时在可解释性方面提供了显著优势。

此外，SRA 展示了显著优越的专家利用率和训练稳定性。尽管使用了相同的稳定技术，SRA 的结果只有 1.0%的失效专家，相比之下标准 MoE 有 14.8%。这表明语义路由促进了模型容量更有效的利用。



(a) 密集基线（前馈网络）验证困惑度。

(b) SRA 验证困惑度。

图 1: 训练期间的验证困惑度曲线。SRA 模型（b）显示了一个与渐进路由策略相对应的独特模式。

4.3 训练动态

SRA 模型展现出与渐进路由策略相对应的独特模式（图 1）。在第 5 个纪元的转变（约 16k 步，见图 1b）对应从 Top-1 到 Top-2 路由的过渡。在 Top-1 阶段（0-16k 步），SRA 发展了初步的专业化但表现不及密集模型。转向 Top-2（从 16k 步开始）立即提升了性能，使 SRA 能够利用专家组合并最终超越基准。密集基准（图 1a）显示稳定收敛至 14.13 困惑度。

4.4 消融研究

我们进行了消融研究以了解路由策略、初始化和辅助损失的影响（表 2）。

表 2: 消融研究结果在 WikiText-103 验证集上。总专家数=512。

Configuration	Strategy/Setting	Val PPL	Dead Experts (%)
SRA（完整）	**逐步的 (1→2)**	**13.41**	**5 (1.0%)**
消融研究：路由策略（10 个纪元）			
SRA	Top-1 only	17.93	0 (0.0%)
SRA	Top-2 only	12.97	176 (34.4%)
SRA	Top-6 only	11.76	96 (18.8%)
消融研究：初始化（Top-1，5 个周期）			
SRA	Orthogonal (Baseline)	16.95	0
SRA	Kaiming Uniform	19.99	0
消融研究：辅助损失（Top-1，5 个 epoch）			
SRA	All enabled (Baseline)	16.95	0
w/o Balance Loss ($\alpha = 0$)		20.72	0
w/o Dispersion Loss ($\beta = 0$)		19.47	0

关键发现来自消融研究：

- **Top-k 的权衡：** 增加 k 可以提高困惑度（Top-6 PPL 11.76）。然而，从一开始就使用 Top-k>1 进行训练会导致专家崩溃严重（例如，在 Top-2 中为 34.4%）。

- **逐步路由是必不可少的：**进步策略实现了最优平衡（困惑度 13.41，1.0% 死专家）。
- **初始化的影响：**正交初始化显著优于 Kaiming 均匀初始化（5 个周期时的 PPL 为 16.95 对比 19.99），验证了最大化初始分散对于发展有效的语义专业化至关重要的假设。
- **辅助损失很重要：**负载均衡损失和色散损失都对性能至关重要。

5 分析

5.1 专家专业化模式

SRA 的一个主要优势是其路由机制的可解释性。

5.1.1 语义共振架构专化

SRA 模型的分析揭示了明显的语义和句法专化。表 3 提供了来自第一个 CSR 层的例子。

表 3: 专家专业化的示例在 SRA（第 0 层）中。专家学习不同的且一致的类别。

Expert	Category	Top-10 Tokens
E_1	Temporal/Months	May, March, February, June, April, July, December...
E_{10}	Prepositions	to, To, for, towards, in, than, of, against, as, into
E_{20}	Time Periods	years, months, weeks, later, then, year, decades...
E_{35}	Past Tense Verbs	was, were, same, became, also, reported, came, is...
E_{42}	Media (Nouns)	film, game, book, games, movie, novel, films, books...
E_{58}	Proper Names	Henry, Edward, Peter, Scott, Robert, Richard, David...

专家显然学会对具有相似属性的标记作出反应，这些标记跨越连贯的语义类别（例如，E42，媒体）和句法角色（例如，E10，介词）。

5.1.2 与标准 MoE 的比较

我们将此与标准的 MoE（学习门控）进行对比（表 4）。

表 4: 专家专业化的示例在标准 MoE（第 0 层）中。类别通常是嘈杂的且一致性较差。

Expert	Top-10 Tokens
E_{27}	ars, Jordan, orn, steam, tank, Br, bodies, con, semin, iner
E_{32}	San, ink, withstand, despite, pieces, été, Buccaneers, Dog, onica, Sem
E_{36}	mon, Bo, comm, his, being, some, the, making, against, military
E_{38}	De, than, American, 12, or, Super, still, Red, for, are

标准的 MoE 导致了嘈杂的专业化，这表明依赖于低层次的统计特征而不是连贯的语言概念。

5.2 专家利用

图 2 说明了最终 SRA 模型（使用渐进路由训练）的四层中专家利用模式。该策略成功缓解了专家崩溃问题，总共只有 5 个失效专家。

图 2: 专家在验证数据集上的利用（最终的 SRA 模型）。

逐步策略允许专家在 Top-1 阶段建立专业化（图 3），这防止了从一开始就使用 Top-2 训练时观察到的专家急剧崩溃现象（如表 2 所示）。

我们还将 SRA 和标准 MoE 的利用率指标进行了比较（表 5）。

虽然标准的 MoE 可能在活跃专家中达到相似的变化系数（CV），但它面临着显著更高的专家崩溃率（76 个死亡专家对比 SRA 中的 5 个）。这突显了 SRA 的一个关键优势：将路由基于语义空间的几何约束，本质上促进了更好的整体利用率和训练稳定性。

图 3: 顶级第 1 阶段（第 5 个周期）中专家的利用。

5.3 语义锚点分散

语义空间的可视化（例如，t-SNE 投影，见附录）证实了正交初始化和分散损失的组合成功地保持了锚点之间的分离（平均成对相似性：0.073±0.065），防止冗余专家。

表 5: 专家利用指标对比（CV 和失效专家）（图 2 和图 5。

Layer	SRA (CV)	Std MoE (CV)	SRA (Dead)	Std MoE (Dead)
0	1.84	1.66	3	28
1	1.65	1.33	0	5
2	1.77	0.89	0	7
3	2.25	1.20	2	36
总数	–	–	**5**	**76**

6 讨论

6.1 可解释性和连贯性

语义共振机制通过设计提供了内在的可解释性。路由决策可以追溯，专业化是一致的，专家专门从事可识别的语义类别。这些特性使得 SRA 适用于理解模型内部逻辑至关重要的应用场景。

6.2 效率和利用率

SRA 维持与密集基线相同的活跃参数数量（29.0M），确保推理过程中的计算平等。一个关键发现是，与标准 MoE 相比，SRA 的专家利用率更优（1.0% 对比 14.8% 死亡专家），表明训练稳定性增强。

6.3 限制和未来工作

我们的研究突出了未来研究的领域：

1. **缩放行为：** 实验在适度的规模下进行（558.5M 参数）。SRA 在十亿参数规模下的行为仍然是一个重要领域。
2. **定量可解释性：** 虽然定性结果很强，但未来的工作必须纳入定量指标，如标准化点互信息（NPMI），以严格测量和比较专业一致性。这是验证可解释性主张的重要下一步。
3. **可控输出（模型引导）：** 探索通过在推理过程中干预特定专家权重来操纵生成文本的方法（例如，用于减少偏差或控制风格）。

6.4 更广泛的影响和可控性

SRA 增强的可解释性对模型可控性具有重要影响。由于 SRA 专家专注于连贯的概念，因此在推理过程中进行干预以调整模型行为成为可能。这种精确且语义基础的控制在密集模型或路由不透明的标准 MoE 中大多不可用。

7 结论

该项研究介绍了语义共振架构（SRA），表明基于语义相似性的路由，结合一种新颖的扩散损失，为 MoE 模型中的学习门控函数提供了有效、高效且可解释的替代方案。我们的实验显示，SRA 达到了标准 MoE 基线的性能水平，同时实现了显著更高的专家利用率，并表现

出更一致的语义专业化。SRA 提供的透明度代表了向开发更具可控性和可解释性的 AI 系统迈出的重要一步。

致谢

本研究作为乌日霍罗德国立大学博士学位学习的一部分进行。HengeBytes 慷慨提供了计算资源。我们感谢乌日霍罗德国立大学信息技术系的学术支持，以及 HengeBytes 团队维护基础设施的支持。我们还承认 PyTorch、HuggingFace（Accelerate）、微软 DeepSpeed 团队和 wandb.ai 的贡献。

参考文献

[1] Tom Brown et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[2] Jiaao Chi et al. Xmoe: Scaling mixture-of-experts with adaptive routing for multitask learning. *arXiv preprint arXiv:2305.14704*, 2023.

[3] Aakanksha Chowdhery et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[4] Kevin Clark et al. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP*, 2019.

[5] Nelson Elhage et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

[6] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research (JMLR)*, 2022.

[7] Suchin Gururangan et al. Demix layers: Disentangling domains for modular language modeling. In *Proceedings of NAACL*, 2022.

[8] Kaiming He et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2015.

[9] Robert A Jacobs et al. Adaptive mixtures of local experts. *Neural Computation*, 1991.

[10] Dmitry Lepikhin et al. Gshard: Scaling giant models with conditional computation and automatic sharding. In *Proceedings of ICLR*, 2020.

[11] Stephen Merity et al. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[12] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.

[13] Noam Shazeer et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of ICLR*, 2017.

[14] Jianlin Su et al. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

[15] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of ACL*, 2019.

[16] Ashish Vaswani et al. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[17] Barret Zoph et al. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.

附录

A.1 可解释性案例研究：详细路由

为了说明 SRA 的可解释性，表 6 提供了句子"The film was released in December 1995 and received positive reviews."的路由决策详细情况。每个词对应的前两名专家及其权重都被展示出来。路由决策与之前识别出的专业功能（例如 E42 用于媒体、E35 用于过去时动词）很好地对应，为模型内部处理提供了清晰的解释。

表 6: 详细路由决策示例句子在 SRA（第 0 层）。

令牌	**该**	**电影**	**是**	**发布**	**在**	**十二月**	**1995**	**和**	**接收**	**正的**	**评论**
专家 1	E111	E42	E35	E119	E10	E1	E57	E46	E9	E74	E84
(Weight)	(0.505)	(0.551)	(0.569)	(0.525)	(0.575)	(0.523)	(0.540)	(0.528)	(0.518)	(0.502)	(0.524)
专家 2	E126	E26	E55	E9	E118	E57	E1	E80	E74	E30	E74
(Weight)	(0.495)	(0.449)	(0.431)	(0.475)	(0.425)	(0.477)	(0.460)	(0.472)	(0.482)	(0.498)	(0.476)

A.2 语义锚点分散可视化

图 4: t-SNE 投影的语义锚点

A.3 专家利用率混合专家系统

图 5: 专家利用验证数据集（基线 MoE）。