MambaFormer：词元级引导路由混合专家，实现准确高效的临床辅助

哈马德·汗[1], 萨达姆·侯赛因·汗[1]*

[1]工程与应用科学大学（UEAS）计算机系统工程系人工智能实验室，巴基斯坦斯瓦特 19060**电子邮箱：** hengrshkhan822@gmail.com

摘要

大语言模型（LLM）在现实世界临床应用中的部署，受到计算成本与线性时间模型效率之间根本性权衡的制约。为解决这一问题，我们提出一种基于大语言模型的 MambaFormer 混合专家（MoE）框架，用于高效的医学问答（QA）和临床辅助。MambaFormer 采用轻量级门控机制，针对短而复杂的查询，将词元级动态路由至定制的 Transformer 专家（E_{T5}）；针对长序列、高吞吐量任务，则路由至状态空间模型专家（E_{Mamba}）。定制的 E_{Mamba} 和 E_{T5} 模型根据输入序列维度、嵌入结构、序列长度和特定目标输出头进行调整，并通过迁移学习在新建的定制 DentalQA 数据集上微调。此外，路由决策由词元嵌入的上下文复杂度、归一化序列长度和领域感知特征共同驱动，从而在推理延迟与预测准确率之间实现帕累托最优权衡。此外，一种新颖的效用引导多目标损失，通过自适应调节词元级专家激活，联合优化决策、路由器参数、路由行为、专家利用率和计算成本。最后，我们在新建的定制 DentalQA 和 PubMedQA 数据集上，对提出的 MambaFormer 的医学问答性能进行交叉验证（留出法），并将其与当前最先进的技术进行比较。提出的 MambaFormer 表现更优（BERTScore = 0.9180），且延迟极低（0.077 秒），速度比 T5-Large 快 24.4× 倍，为资源受限环境下的临床部署提供了可扩展的解决方案。

关键词： 混合专家、路由、状态空间模型、Mamba、Transformer、自然语言处理、效用引导优化

1. 引言

随着大语言模型的兴起，自然语言处理取得了显著进展，这些模型通过大规模预训练提升性能 [1]。受限于 grounding 能力不足和计算成本高，自然语言处理在复杂推理方面的效用仍然有限 [2]。这也阻碍了持续知识更新，并引发了对可持续性的担忧 [3], [4]。尽管存在这些挑战，自然语言处理在风险评估、治疗支持和疾病预测等临床应用中发挥着日益重要的作用 [5], [6]。然而，自然语言处理在治疗领域面临着特殊困难，例如医学语言的复杂性，以及计算资源方面的限制。

目前已有多种卓越的语言模型问世。传统模型如 Transformer [7] 已被提出，但 Transformer 使用自 - 注意力机制，其计算公式为 $O(N^2)$，其中 N 为序列长度。虽然该模型在准确率方面表现令人满意，但在内存和速度方面 s 存在不足，导致需要大量计算资源。详细的临床报告对其进行了测试 [8] ，而且这也同样阻碍了它在其他数据密集型 t 域中的部署，例如医学图像分割 [9]。在实时应用和医院中，时间是一个非常重要的因素；由于 o 上述情况，例如，系统处理 1000 个词的临床信息可能需要 1.8 秒，这太慢了；对于医院而言， d 对于诊所而言，快速处理至关重要。我们需要一种处理速度快的系统，同时还必须保护所有患者的隐私 ' 信息，无论是局部还是全局。其他 Transformer 模型，例如 BioBERT [10], 和 ClinicalBERT [5] ，在生物医学任务中也取得了成功，BER T 得分达到 90%，但这会导致高昂的成本。

除了缓解这一限制外，LongFormer [11] 和 BigBird [12] 等模型还采用了稀疏注意力，通过注意力模式将复杂度降至 O(N)。Performer [13], 等基于核的方法也得到了应用，其效率提升优于 Transformer。然而，在长篇医疗记录上进行测试时，其准确率会下降 5% 至 10%，因此不适合在临床环境中使用。近年来，状态空间模型（ SSM ）受到广泛关注，尤其是 Mamba 架构 [14], 。近期（ 2024年）的进展推出了 StripedHyena [15] 和 MambaByte [16], 等先进的 SSM 变体，进一步优化了选择性扫描机制。与此同时，MoE-Mamba [17] 和块状态Transformer [18] 等混合架构探索了注意力机制与循环之间的不同融合策略，但通常采用固定路由或块级路由，而非词元级自适应。鉴于选择性状态空间模型（SSM）具有线性扩展能力和更强的长距离依赖关系建模能力，因此被视为更好的替代方案，如公式（1）所示，其中 L 表示输入序列长度。Mamba 能够以极高的效率分析长序列，并降低训练和推理成本 [19]。然而，其关联学习能力和双任务泛化能力表现不稳定，而且难以捕捉文本中局部的重要信息。同样，Transformer 在准确率方面表现出色，但处理较短序列时需要大量计算资源，成本也较高 [7]。这

在临床应用中，这造成了一个重大矛盾，因为准确率和速度都至关重要。然而，医生又面临着一个问题：要么选择准确率较高的 Transformer，却承担较高的计算成本；要么坚持使用 SSM，却导致性能下降。

帕累托权衡尚未得到充分解决，任何单一模型乃至混合模型都难以克服这一挑战。Jamba 和混合专家 [20], [21], 等混合模型整合了 Transformer 与状态空间模型，展现出改善这一权衡的积极迹象。然而，即使是最新的混合方法，在自适应路由方面仍面临挑战，它们依赖预先设定的架构模式，而非上下文感知的令牌级决策。尽管 Mistral 7B 和 Phi-3 [22], [23], 等高效大语言模型取得了进展，证明混合专家设计具有实际可行性，但它们缺乏适用于临床文本的细粒度路由机制，这一差距依然存在。然而，这些模型的设计基于固定路由，也没有包含任何智能规则或惩罚机制来展示 Transformer 或状态空间模型的计算报告。即使是 Mamba-3 [24] 和 RetNet-MoE [25]等最先进的模型，仍在使用预先设定的路由模式，而非上下文感知的令牌级决策；我们的 MambaFormer 正是为弥补这一不足而提出。此外，这些模型没有提供有关哪些专家将用于提升准确率和速度的信息，而我们的模型克服了上述所有局限，提供了一个兼顾效率与成本的完整模型。因此，我们需要一种更智能的路由机制，以动态平衡准确率与效率。为此，我们提出了 MambaFormer，这是一种用于实现生物医学自然语言处理领域帕累托最优效率的效用引导的动态路由框架。我们提出的新方法作出了以下贡献：

- 我们提出的新型混合专家（MoE）框架“MambaFormer”，集成了定制状态空间模型 Mamba（E_{Mamba}）和 Transformer（E_{T5}），并采用一种新型引导式门控机制，为临床辅助动态路由每个输入词元。此外，MambaFormer 引入了一种新的多目标函数（$L_{MambaFormer}$），可自动优化自适应路由参数，从而系统地将词元分配给最优专家。

- 根据输入词元维度、嵌入结构和长度，对定制状态空间模型 Mamba（E_{Mamba}）和 Transformer（E_{T5}）的输入层进行了修改，并添加了用于特定目标问答和描述的输出层。此外，这些模型通过迁移学习，在一个新设计的定制 DentalQA 数据集（包含 5,000 个问答对，即 10,000 个词元）以及 PubMedQA 数据集上进行微调。

- MambaFormer 通过提出的智能路由器，将词元特征与变分序列系统地结合起来，并将其路由至 E_{T5}专家（复杂短文本、高准确率场景）或高效的 E_{Mamba}专家（长序列、高吞吐量场景）。此外，受引导的路由器能够学习域感知特征和序列长度特征，并根据准确率与延迟之间的帕累托最优权衡，动态地将词元分配给最优专家。因此，96.2% 的词元被分配给快速高效的 E_{Mamba}专家，而只有 3.8% 的词元被分配给高准确率的 E_{T5} 专家，以适应现有数据集的多样性。

- 提出的 MambaFormer 框架成功超越了传统模型的速度（响应时间），响应时间缩短了 0.077 秒（比传统模型快 24.4 倍），同时保持 BERTScore F1 为 0.9180，准确率为 96.2%。

本文其余部分安排如下：第2节概述以往研究，第3节介绍方法论。第 n

4节介绍实验设置，第5节呈现结果，第6节总结本研究。

2. 相关工作

N 许多研究探讨了高效生物医学自然语言处理的复杂性，包括Transformer模型和其他复杂的混合深度学习模型。当前的自然语言处理系统基于Transformer扩展，这一方法最初由采用缩放乘法注意力的Vaswani等人提出 [7], ，使模型能够记录全局依赖关系而无需重复。随后，人们采用了BERT [26] ，其展现了

b通过双向预训练，在命名实体识别基准测试中取得了90%的F1值。随后，人们使用文本到文本的任务来表示

N 利用T5框架表示自然语言处理任务 [27], ，GPT系列则表明，特征扩展对于数据规模和模型规模具有重要意义 [28]。然而，底层的自注意力机制具有二次计算复杂度O(N^2)，这严重限制了其在生物医学应用中的适用性

b 由于临床记录规模庞大，已有研究表明，处理长度为 1,000 个词元的临床笔记最多需要 200MB 内存和 1.8 秒，因此难以实现实时部署 [5]。这一可扩展性问题并非自然语言处理领域独有；正如近期的综合性综述所详述，视觉Transformer（ViT）架构在处理高分辨率医学图像时也面临类似挑战 [29]。尽管准确率很高，Transformer 仍因二次注意力复杂度而难以部署于资源受限的临床环境。

LongFormer [11] 和 BigBird [12]等稀疏模型与核化模型在长篇生物医学文本上的准确率可达 85–90%，而 Performer [13]

的 F1值达到 88%，Linformer [30] 则通过低秩投影降低注意力机制的计算量。然而，这些方法在复杂的 x

生物医学任务 [31] 上会导致 5–10% 的准确率下降，且对细粒度临床模式的建模不足，包括脑肿瘤磁共振成像分析 [32]。近期的高效序列模型 ，

包括 StripedHyena [15] 和 MambaByte [16], ，它们通过状态空间模型和字节级处理提升可扩展性。混合专家-Mamba 等混合设计 a

[17] 达到0.902的BERTScore，而块状态Transformer [18] 采用固定的计算方案。高效大语言模型（例如Mistral 7B [22]

和Phi-3 [23]）通过混合专家架构降低成本，但缺乏适用于临床部署的细粒度词元自适应能力。然而，这些方法要么 r

保持固定路由模式，要么以粗粒度（层级或块级别）运行，错失了进行细粒度、词元级自适应的机会 n

而这种自适应对于复杂度混合的临床文本至关重要。

这种性能下降揭示了速度与准确率之间明显的权衡，这对于关键临床任务而言是不可接受的。因此，我们需要一种能够智能调用高准确率专家的系统，尤其是在生物活动方面；较新的混合注意力状态空间模型展现出了希望 [33]。作为以线性扩展为核心的序列模型的替代方案，状态空间模型应运而生。利用模型处理长程依赖的潜在成果包括

在分类测试中取得了 87% 的准确率4 [34]。凭借近线性扩展能力 O(L)，Mamba 得以提供硬件感知优化和选择性扫描；在生物医学检索中，其 F1值达到 85%，体现了良好的扩展性。Mamba-2[14] 显著改进了这一方法，速度提升 2–8 倍，准确率达到 92%，表明混合系统或许切实可行。在此基础上，2025年出现了进一步改进的状态空间模型，例如 Mamba-3 [24] ，采用分层选择性扫描，以及 HyenaDNA 2。0 [35] 它们面向基因组分辨率，但仍保留了固定的架构模式，不适用于临床级别的令牌适配。

然而，尽管 Mamba 凭借线性时间复杂度 O(L) 在处理长序列时表现出色，但有时难以理解短篇技术文本中的复杂细节。在这些情况下，自注意力机制表现更佳 [7]。在处理复杂临床笔记时，它们无法像 Transformer 那样表现出色，这表明二者能力互补，也凸显了混合解决方案的价值。混合方法在

N LP 已在多个域中展现出良好的效果。SwitchTransformer 和 Glam [36], [37],（统称为专家混合（MOE））在问答任务中实现了 90% 的 F1值，并通过稀疏激活提供速度—准确率权衡。路由 Transformer [38] 实现了精确的 91% 令牌级交互，但它们仅适用于 Transformer 框架，且通常旨在均衡分配工作负载，而非关注计算效率。即使是 2025年的混合架构，如 RetNet-MoE [39] 和 StripedHyena-2 [15] ，仍采用粗粒度或固定路由，忽略了在复杂程度各异的临床文本中进行逐令牌效用引导决策的需求。

生物医学自然语言处理面临着来自冗长医疗记录、研究论文和复杂术语的独特挑战。虽然 PubMedBERT [33] 在问答任务上具有 91% 的 F1值，但在生物医学语料库上训练的 BioBERT [10] 模型在命名实体识别任务上的 F1值为 90%。ClinicalBERT [5] 以 88% 的曲线下面积预测死亡率，展现出较强的临床预测能力。然而，这些模型巨大的计算负载（延迟 1.9 秒）使其不适用于临床应用。BioMamba [40] 在长上下文中以 95% 的准确率展现出良好前景，但在简短临床文本上的表现仍不理想。实时诊所需要一种高效的方法，能够轻松适应不同类型的文本，无论文本密集、简短还是冗长。进入 2025年，这一需求依然存在；近期调查显示，自适应路由和部署可行性仍是尚未解决的挑战，而我们的 MambaFormer 正面应对了这些问题。根据这篇综述，目前生物医学自然语言处理存在以下不足

p实践:

- 现有混合模型依赖固定路由，无法动态平衡推理速度与准确率，从而限制了其实时临床可用性。
- 当前方法无法应对高度不均衡的计算需求，也缺乏针对不同输入类型的自适应专家选择。
- 大多数研究未能评估模型在真实临床条件下的部署可行性，包括其延迟和鲁棒性。

表 1：生物医学自然语言处理现有研究。

模型	架构	主要贡献	指标	速度（秒）	内存（MB）	主要局限
StripedHyena [15]	状态空间模型-卷积	混合状态空间模型	0.895 F1值	0.110	155	固定模式
MoE-Mamba [17]	混合专家+状态空间模型	稀疏路由	0.902 BERT	0.095	165	块路由
Mistral 7B [22]	MoE-Trans	高效混合专家	0.910 MMLU	0.350	14000	高内存占用
BioBERT [10]	Transformer	生物医学预训练	0.900 命名实体识别	1.900	195	高延迟
LongFormer [11]	稀疏 Transformer	长上下文	0.880 F1值	0.450	180	局部性较弱
Mamba [14]	SSM	线性扫描	0.850 F1值	0.100	142	短上下文
Mamba-2 [14]	SSM	硬件优化	0.870 F1值/曲线下面积	0.085	138	局部能力有限
Jamba [20]	混合专家模型	状态空间模型+Trans	0.890 F1值	0.450	210	固定路由
BioMamba [40]	状态空间模型（生物学）	领域状态空间模型	0.880 F1值/曲线下面积	0.092	145	短上下文
CE-RS-SBCIT[32]	卷积神经网络-Transformer	医学视觉	~0.95 Dice系数	N/A	N/A	纯视觉
Drilling-Hybrid[41]	Trans-LSTM	时间序列	~0.92 R2	N/A	N/A	任务特定

3. 方法论

提出的 MambaFormer 旨在定制设计的 DentalQA 和 PubMedQA 数据集上，实现医学问答（QA）中准确率与效率的平衡。MambaFormer 融合了上下文预处理、令牌级自适应专家路由，以及混合专家架构的联合优化，从而实现准确的低延迟推理。提出的 MambaFormer 框架采用基于BERT的分词，将 PubMedQA 和定制设计的 DentalQA 数据集中的输入编码为保留上下文的医学表征。随后，效用引导的动态路由器将语义信息融入每个词元的序列嵌入，并将其分配至轻量级 E_{Mamba} 模块进行低复杂度处理，或分配至基于Transformer的模块进行高精度推理。该路由策略明确优化延迟–准确率权衡，从而在计算预算受限的情况下实现高效推理。由于数据稀缺且标注成本高，混合专家；E_{Mamba} 和 E_{T5} 骨干网络采用任务特定适配层进行定制，并通过基于迁移学习的方式在医学数据上调优，以增强从有限样本中获取领域知识的能力。提出的 MambaFormer 架构的完整推理流程如图1所示。

3.1 分词与特征工程

PubMedQA 和新设计的 DentalQA 数据集呈现出词元分布特征：较长的生物医学序列（~96.2%）占主导，超过较短的临床问答对（~3.8%），凸显了真实世界临床自然语言处理场景的特点。这也促使我们开发一种自适应技术，同时适用于简洁且对准确率要求极高的查询，以及冗长且对吞吐量要求较高的文档。所有文本均经过标准化处理，以便在分词前统一领域专用术语、缩写和异构句子序列。使用 BERT-base 分词器对输入序列进行分词和截断，以便统一批次处理。在 512 词元的限制内保留子词级临床语境，在计算效率与保留医学相关信息以供下游路由使用之间取得平衡。

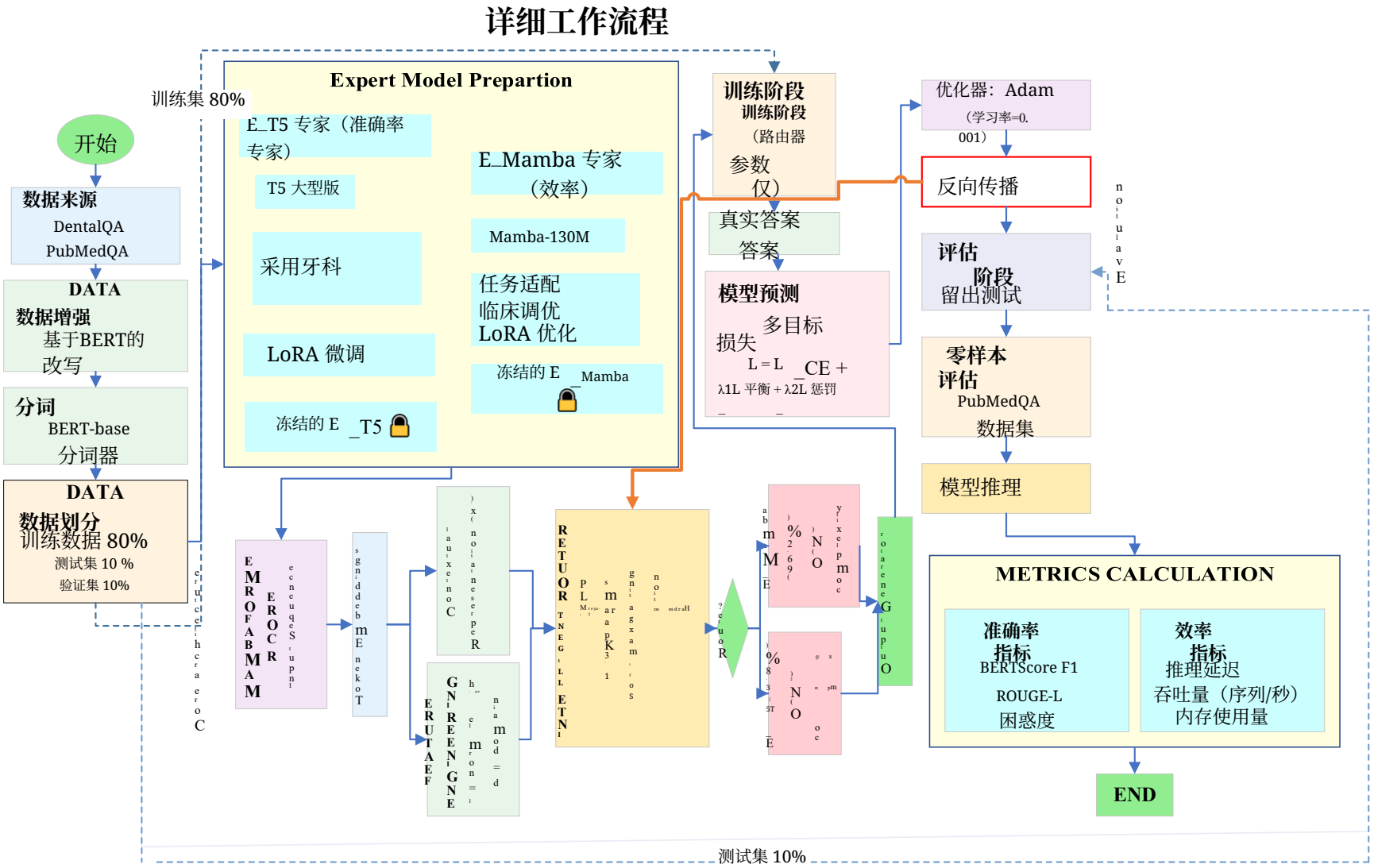


图1. 完整的MambaFormer框架流程。

3.2 数据预处理

基于BERT的改写将 5,000 个问答对扩充至 13,000 个，以应对 DentalQA 规模有限的问题；通过 Sentence-BERT 相似度阈值（<0.95）筛选样本，并保留语义保真度。扩充后的 DentalQA 按训练集/验证集/测试集（80/10/10%）划分。与此同时，使用 PubMedQA 进行零样本评估，以提升模型在长篇生物医学测试数据上的泛化能力。

3.3 混合设计：专家模型选择与表征

提出的 MambaFormer 框架采用令牌级混合专家，将高准确率的基于迁移学习微调的 Transformer（T5-Large）[27] 与高效率的状态空间模型（Mamba-130M）[14], 相结合，根据序列长度、域和内容自适应地路由词元，以平衡准确率与效率。相较之下，Mamba-130M 展现出卓越的推理速度（在 PubMedQA 上每个序列仅需 0.055 秒），证实了二者分别作为准确率专家和效率专家的互补作用。图 2 描述了如何通过牙科词汇扩展，并在 DentalQA 数据集上进行低秩适配（LoRA）[42] 微调，将预训练的 T5 和 Mamba 模型适配为 MambaFormer 的领域专属冻结专家。

$$C_{\mathrm{T5}}(L) = O(L^2), \quad C_{\mathrm{Mamba}}(L) = O(L) \tag{1}$$

$$W' = W + \frac{\alpha}{r} BA \tag{2}$$

$$e_i' = e_i + P_i + W_d \cdot d_s \tag{3}$$

公式（1）对它们的计算复杂度进行了形式化描述：$\mathrm{E_{T5}}$ 随序列长度呈二次增长，而 $\mathrm{E_{Mamba}}$ 呈线性增长。公式（2）定义了用于高效微调的低秩适配（LoRA）[42] 。公式（3）利用领域专属知识对词元嵌入进行适配。这些公式为下文所述的定制过程奠定了基础。

3.3.1 $\mathrm{E_{T5}}$ (T5-Large) 的定制

我们通过保留 T5-Large [27] 的 Transformer 架构，同时针对临床问答进行调整，将其定制为生物医学精度专家。我们使用公式 (3) 添加了医学词汇，并将最后一层替换为问答专用头。我们使用 LoRA[42] (公式 2)进行高效微调，以学习临床模式，同时避免灾难性遗忘。

定义 $\mathrm{E_{T5}}$ 的数学运算如下

$$h_l^{\mathrm{T5}} = \mathrm{LayerNorm!}\left(h_{l-1} + \mathrm{MultiHead}(h_{l-1})\right) \tag{4}$$

$$h_l^{\mathrm{T5}} = \mathrm{LayerNorm!}\left(h_l^{\mathrm{T5}} + \mathrm{FFN}\left(h_l^{\mathrm{T5}}\right)\right)$$

$$\mathcal{L}_{\mathcal{E}_{\mathcal{T}_5}} = \mathcal{L}_{\mathcal{CE}}\left(y_{pred}, y_{true}\right) + \lambda \mathcal{L}_{\mathcal{LM}}(X) \tag{5}$$

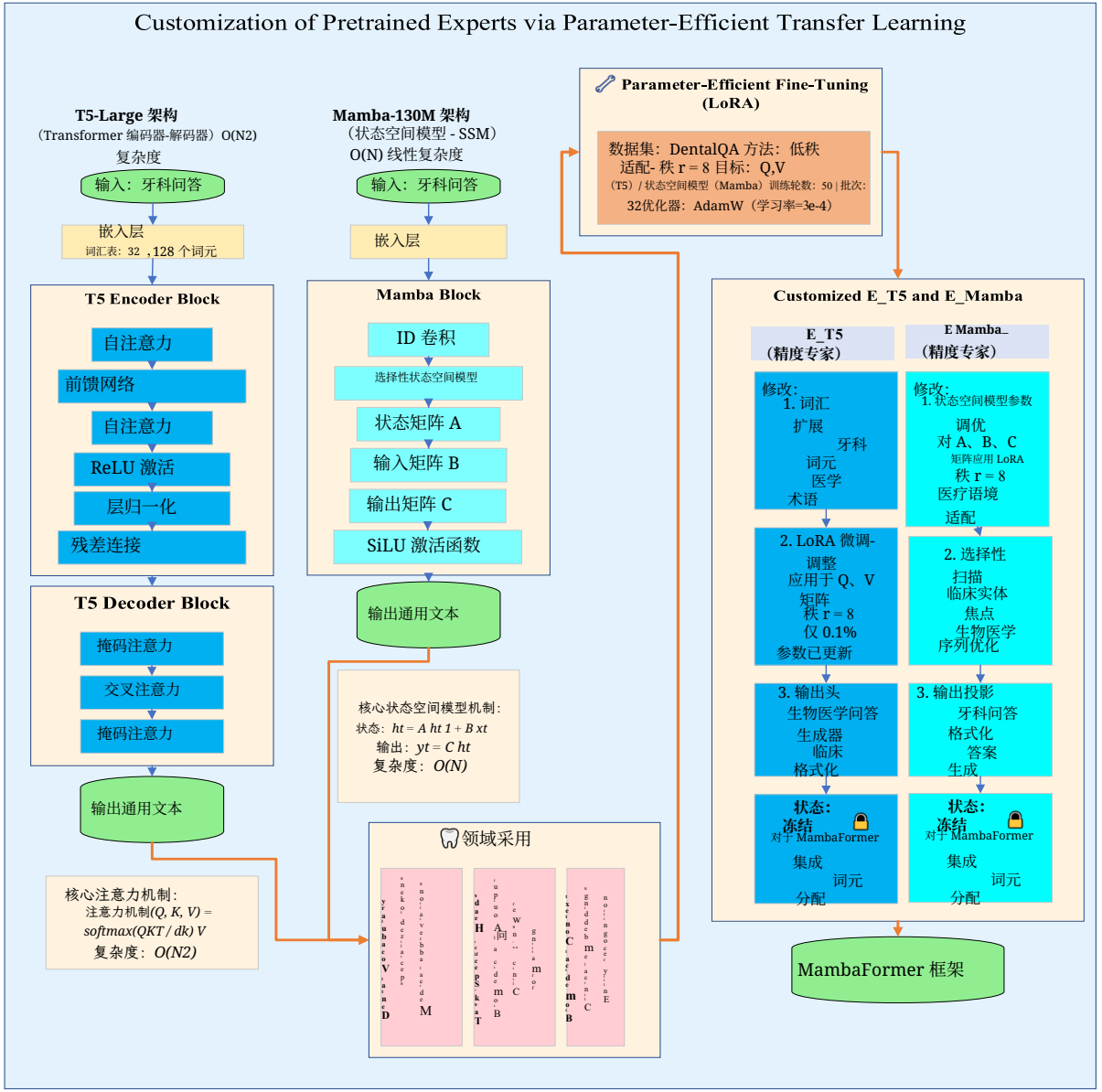


图 2：通过参数高效迁移学习进行专家定制。

上述公式（4）描述了 Transformer 层的更新方式，使 E_{T5} 具备强大的上下文理解能力。公式（5）定义了其训练目标，将用于提升问答准确率的交叉熵损失与用于保留通用知识的语言建模项结合起来。定制完成后，E_{T5} 被冻结并集成为精度专家。

3.3.2 E_{Mamba} （Mamba-130M）的定制

我们通过增大 Mamba-130M [14] 的状态维度，使其能够更好地进行长序列建模，同时保持其线性时间递归，从而将其适配为高效率专家。词汇表对齐遵循公式（3），并且仅对输入/输出投影应用选择性 LoRA[42] 微调（公式 2），以保留核心状态空间模型的效率。

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t, \quad y_t = \bar{C}h_t \tag{6}$$

$$\mathcal{L}_{\mathcal{E}_{Mamba}} = \mathcal{L}_{\mathcal{CE}}(y_{pred}, y_{true}) + \beta||\bar{A} - I||_F^2 \tag{7}$$

公式（6）给出了状态空间模型的循环，使 E_{Mamba} 具备线性计算复杂度（如公式 1 所形式化）。公式（7）定义了其训练损失，将交叉熵与稳定性正则化项结合起来，以防止状态空间模型参数在微调过程中发散。适配完成后，E_{Mamba} 被冻结并集成为效率专家。

表 2：专家模型选择的实证评估

模型	参数量	DentalQA F1值	DentalQA 困惑度	PubMedQA 速度（秒）	PubMedQA F1值	选择依据
GPT2-Small[43]	124M	0.9129	2.91	0.716	0.808	良好的牙科
GPT2-[43]中型	345M	0.8577	98767.56	2.169	0.523	不稳定
DistilGPT2 [43]	82M	0.5575	9.82	0.485	0.620	准确率低
T5-Small [44]	60M	0.8646	16.76	0.920	0.832	权衡
T5-Large[44]	770M	0.9362	1.41	1.949	0.864	最佳准确率
Mamba-130M [14]	130M	0.8541	1.93	0.055	0.840	最佳速度

3.3.2 集成式混合专家架构

此外，提出的 MambaFormer 将两个专家集成为一个混合专家层，并按照公式 (6) 将其定义为并行的专用前馈网络。这种设计能够实现自适应计算，根据上下文特征将每个词元路由至最合适的专家。嵌入 Transformer 前馈模块中的混合专家层通过门控函数 G_k（公式 (8)）实现词元级自适应路由，将词元分配给 E_{T5} 进行高准确率处理，或分配给 E_{Mamba} 进行高效的线性计算。两者

专家保持冻结，利用各自的互补优势（公式 1），同时轻量级路由器根据上下文线索动态分配词元。完整的定制与集成流程如图 2 所示，混合专家集成的数学定义如下：

$$\mathrm{MoE}(h_i) = \sum_{k\in\{M,T\}} G_k(h_i)E_k(h_i) \quad (8)$$

表 3：战　略角色及 MambaFormer 中定制专家的目标利用率　　.

专家	**战略角色**	**复杂度**	**目标利用率**
E_{Mamba}（效率专家）	用于长序列的线性时间状态空间模型	O(N)	≈ 96.2%
E_{T5}（精度专家）	用于短而复杂查询的自注意力	$O(N^2)$	≈ 3.8%

3.4 智能路由器设计与选择策略

提出的 MambaFormer 利用序列长度和域特定信息，以最小开销实现引导式令牌级路由。评估中的帕累托分配将大部分输入词元路由至高效的 Mamba 专家（≈96.2%），并将少量词元路由至注重准确率的 E_T5 专家（≈3.8%），从而保持较低的计算成本，如表 3 所示。

3.4.1 路由器架构

路由器设计包含一个具有约 1.3K 个可训练参数的两层多层感知机，用于控制延迟。对于每个专家 *i*（其中 j∈ E_{Mamba}，E_{T5}），路由器根据输入词元的上下文表示 x_i，使用 softmax 函数计算门控分数 $S_{i,j}$，如公式（9）所定义。

$$S_{i,j} = \frac{\exp\left((x_i W_g + b_g)_j\right)}{\sum_{k\in\{Mamba,T5\}} \exp\left((x_i W_g + b_g)_k\right)} \quad (9)$$

$$\text{where} j \in \{Mamba,T5$$

$$m^* = \arg\max_{m\in\{E_{Mamba},E_{T5}\}} S_{i,m} \quad (10)$$

其中，W_g 和 b_g 是门控网络学习得到的权重和偏置，并在训练期间计算专家输出的加权组合，以实现梯度流。路由器根据这些分数的 argmax 执行硬选择（公式（10）），仅激活选定的专家 $m*$。该硬选择激活专家 $m*$，为词元 x_i的前向传播选定该专家。此外，整体计算成本是两个专家复杂度的加权和。完整的推理时路由过程在**算法 1**中进行了形式化描述。

算法 1：速度约束门控的动态路由（推理）

1：过程 DYNAMICROUTING (z_t, x_f)

2: // 输入：z_t （令牌表示），x_f= [l，d] （输入特征）

3: // 输出：y_t （专家输出） 5: // 1. 路由器输入特征融合6: R ← CONCATENATE (z_t, x_f) 8: // 2. 计算软路由分数9: S ← SOFTMAX(MLP®) ▷ 基于公式 2 11: // 3. 硬选择（效用引导）12: $m^* \leftarrow \text{ARGMAX}(S_m)$ for $m \in \{E_{Mamba}, E_{T5}\}$ 14: // 4. 专家执行 15: $y_t \leftarrow m^*(z_t)$ ▷ 单个专家前向传播 17: 返回 y_t18: 结束过程

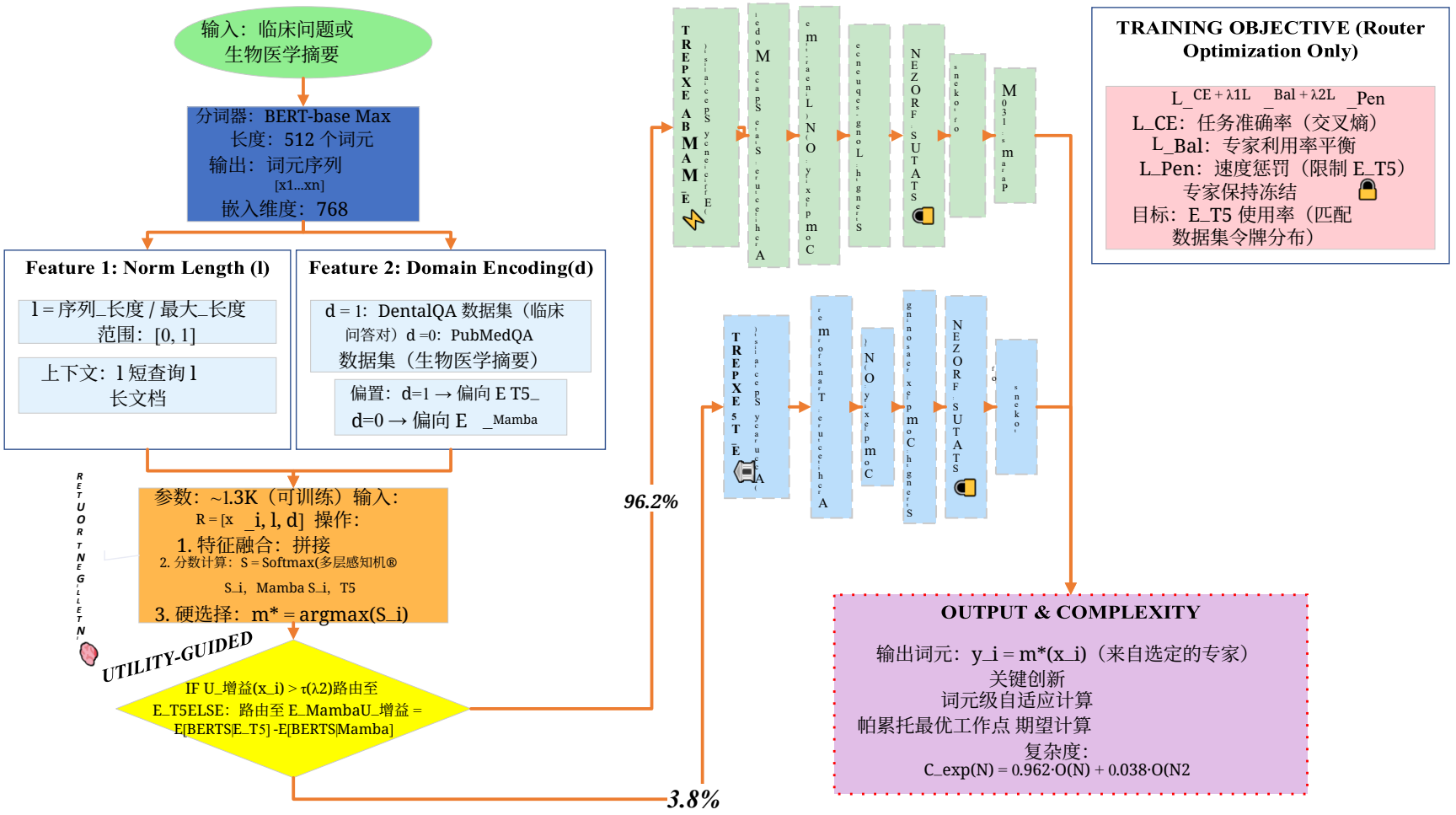
该算法简洁地概括了推理过程：特征融合、分数计算、效用引导的专家选择以及执行。端到端的路由流水线也如图 3 所示。

3.4.2 路由器的输入特征

路由器的决策基于原始词元嵌入 x_i、序列长度和领域感知特征，以表示词元的上下文信息，从而可能同时为各专家提供指导。

a) **归一化序列长度 (l)：** 当前输入序列特征长度被归一化到范围 [0, 1]。它直接编码了较长序列的计算上下文（l→1），将其直接路由至高效的 E_{Mamba}；而较短序列（l→0）则可利用 E_{T5} 的准确率，并实现最优延迟成本。

b) **二元领域编码 (d)：** 此特征表示输入序列的源数据集：如果序列来自 DentalQA 数据集，则 $d = 1$ ；如果序列来自 PubMedQA 数据集，则为 0。该编码提供领域上下文：DentalQA 信息 (d=1) 更短，且临床问答要求精度更高，因此偏向 E_{T5}。此外，PubMedQA 序列 (d=0) 以较长的生物医学摘要为特征，因此偏向 E_{Mamba}，以实现高效的长上下文处理。这两个特征 [l, d] 与其上下文表示 x_i 拼接，‘R= Concatenate(x_i, l, d) ’，构成路由器多层感知机的最终输入。这种融合表示使路由器能够根据语义内容和结构/领域元数据共同做出决策。



Fi图3：采用令牌级动态路由的MambaFormer核心架构 .

3.5 效用引导路由与帕累托优化

路由器的选择策略由学习和优化目标函数驱动，该目标函数旨在为整个系统实现准确率与延迟之间的最优平衡，并达到帕累托最优工作点。路由器会为每个词元估算预测准确率效用增益 (U_{gain})。该增益值表示答案质量（如 BERTScore）的预期提升，如公式 (11) 所示。词元随后由高-

精度专家 E_{T5}，或者由效率专家 E_{Mamba}。最终的路由决策通过将此效用与动态阈值 'τ'进行比较来确定，该阈值由训练目标中的速度惩罚权重 λ_2 调节（见第3.6节）：

$$\text{Route to } E_T5 \quad \text{if } U_{gain} > \tau(\lambda_2) \tag{11}$$

否则，路由至 E_{mamba}。

$$U_{gain}(x_i) = E[\,\text{BERTScore} \mid E_{T5}, x_i\,] - E[\,\text{BERTScore} \mid E_{Mamba}, x_I\,] \tag{12}$$

$$C_{expected}(N) = 0.962 \cdot O(N) + 0.038 \cdot O(N^2) \tag{13}$$

此阈值'τ'直接体现了准确率-延迟权衡。较高的阈值会保守地限制速度较慢的 E_{T5} 的使用，从而优先考虑速度。较低的阈值则允许更频繁地使用 E_{T5}，从而优先考虑准确率。参数 λ_2 经过调整，使学习到的阈值 τ 达到帕累托最优平衡；在该平衡点上，准确率每提升一点，都会付出不成比例的延迟代价。这种路由分布带来的期望计算复杂度如公式 12 所示。该框架能够学习并收敛到稳定的路由分布，其中 ≈96.2 % 的词元被分配至 E_{mamba}，仅有 ≈3.8% 被分配至 E_{T5}, 如公式 13 所示。此分布与生物医学数据集的内在特性相符：大多数词元属于较长、对吞吐量敏感的序列（理想情况下由 E_{mamba}处理）。相比之下，少部分词元来自较短、对准确率要求较高的查询，这类查询可从深入分析中受益。最后，效用引导机制确保仅在必要时才调用 E_{T5} 计算成本高昂的二次复杂度，且预期的准确率收益足以证明其计算成本是合理的。图 3 展示了这一决策过程，说明了从特征提取、效用计算到最终专家选择的流程。

3.6 速度约束训练目标

通过优化一种新颖的多目标损失函数 $L_{MamabaFormer}$，学习引导式路由操作。该损失函数专门设计用于综合提升任务准确率，同时保持专家利用率均衡。该函数施加了严格的推理速度约束，并直接塑造帕累托最优工作点。总损失定义为三个组成部分的加权和，如公式 (14) 所述。第一项 L_{CE} 是预测答案与实际临床答案之间的标准交叉熵任务损失。它确保医学问答准确率这一首要目标得以保留，如公式 (15) 所定义。此外，公式 (16) 中的平衡损失（L_{Bal}）也采用基于库尔贝克-莱布勒散度的负载均衡损失，以防止路由器崩溃，即路由器忽略某个专家的情况。此外，它还鼓励路由器按照软均匀目标分布的比例使用两个专家。其中，L 为序列长度，E 为专家数量（2），Si,j 为词元 i 与专家 j 对应的软门控分数。

$$L_{MambaFormer} = L_{CE} + \lambda_1 L_{Bal} + \lambda_2 L_{Pen} \tag{14}$$

$$L_{CE} = -\sum_c y_c \log(p_c) \tag{15}$$

$$L_{\text{Bal}} = \sum_{i=1}^{L} \sum_{j=1}^{|E|} S_{i,j} \log!\left(\frac{1/|E|}{S_{i,j}}\right) \quad (16)$$

$$L_{\text{Pen}} = \sum_{i \in \text{tokens}} \max!\left(0, S_{i,E_{T5}} - T_{u_i}\right) \quad (17)$$

最后，关键创新组件速度惩罚（λ_2·，$L_{惩罚}$）如公式 (17) 所示，它明确地实现了速度—准确率权衡。它会惩罚对较慢的 E_{T5} 专家的过度使用，前提是其利用率超过预设阈值 Tui≈0.08。超参数 λ_2 控制该惩罚的强度；将其设为 $\lambda_2 = 0.5$，可将 E_{T5} 的使用率限制在约 3.8%，从而使延迟达到以 E_{Mamba} 为主的系统水平。在两个专家均保持冻结的情况下，对轻量级路由器（1.3K 个参数）进行优化，同时确保路由稳定。由此得到的 $L_{MambaFormer}$ 损失能够学习一种兼顾准确率、利用率和效率的路由策略。

4. 实验设置

4.1 数据集

使用了两个生物医学问答数据集来支持领域专门化和泛化。我们专门设计了 DentalQA，这是一个经过整理的临床数据集，包含 5,000 组由专家标注的牙科材料与设备问答对，并附有丰富的结构化元数据；通过基于BERT的改写，将其扩充至 13,000 个训练样本。此外，PubMedQA（长答案）包含 2,600 个生物医学问题及完整的研究摘要，并将其合并为连续的长上下文序列，长度最多达 4,096 个词元，从而构建了一个具有挑战性的泛化基准。表 5 总结了数据集在训练和评估阶段的使用情况。

表 4：DentalQA 数据集统计

属性	值
原始问答对总数	5,000
增强问答对总数	13,000
平均问题长度	12.4 个词
平均回答长度	27.8 个词
独特产品	850
涵盖的制造商	47

表 5：数据集使用情况汇总

数据集	域	规模（原始/增强）	平均序列长度（l）	Role
DentalQA	牙科问答	5,000 / 13,000	≈512 词元	训练集与验证集。（按 80/10/10 划分）
PubMedQA	生物医学摘要 2,600	/ 2,600	2,000–4,000 个词元	留出测试集（零样本评估）

4.2 实现与训练配置

所有实验均在单台 NVIDIA A100 GPU（40GB VRAM）上进行，使用 PyTorch 和 Hugging Face Transformers 库。训练采用混合精度（FP16）以提高效率。用于专家选择的候选模型（GPT-2 各变体、T5-Small、T5-Large、Mamba-130M）均在定制的 DentalQA 训练集划分上进行了微调。通过低秩适配（LoRA）[37]，以秩 r=8高效微调模型参数。训练配置包括 50 个训练轮数、Adam优化器，以及交叉熵损失；根据 DentalQA 验证集和 PubMedQA 测试集上的 BERTScore F1（准确率）和推理延迟（速度）评估并选择专家。在完成专家选择和 MambFormer 训练后，选定的专家（E_{T5} 和 E_{Mamba}）被冻结。仅更新路由器的 1.3K 个参数。所用路由器由一个 2 层多层感知机组成（输入：2 个特征，隐藏层：16 个神经元）。我们在 DentalQA 数据集（扩增后，按 80/10/10 划分）上进行了训练/留出交叉验证。训练数据：DentalQA 训练集划分（扩增数据的 80%）；优化配置包括 Adam（学习率=0.001）、批量大小=64，训练 20 个轮数。关键超参数汇总于表 6。

表 6：MambaFormer 超参数

参数	值	描述
门控网络	2 层多层感知机	隐藏层大小为 16（~1.3K 个参数）
可训练组件	仅路由器	专家被冻结
输入特征	[l，d]	归一化长度（l）和域（d）
批量大小	64	-
学习率	1e-3	Adam优化器
损失权重	λ1 = 1.0，λ2 = 0.5	平衡与速度惩罚
目标利用率（Tui）	0.08	E_{T5} 的最大软使用量

4.3 评估指标

模型采用涵盖三类指标的综合指标体系进行评估：性能、效率和路由分析。性能方面，使用 BERTScore F1（公式 18；P、R）衡量语义相似度，使用 ROUGE-L（公式 19；R_{LCS} 和 P_{LCS}）衡量流畅度，并使用困惑度评估语言建模。效率方面，通过推理延迟（秒/序列）、吞吐量（公式 20；序列/秒）和 GPU内存（公式 21；MB）进行评估。路由行为则通过令牌路由分布和帕累托最优性（公式 22）进行评估。所有指标均在留出集 PubMedQA 和 DentalQA 数据集上计算，并对多次运行的结果取平均，以确保鲁棒性。关键指标的数学表达式定义如下：

$$\boldsymbol{F} = \frac{\boldsymbol{2 \cdot P \cdot R}}{\boldsymbol{P+R}} \quad (18)$$

$$\text{ROUGE-L} = \frac{(\boldsymbol{1+\beta^2})\, \boldsymbol{R}_{\text{LCS}} \cdot \boldsymbol{P}_{\text{LCS}}}{\boldsymbol{R}_{\text{LCS}} + \boldsymbol{\beta^2} \cdot \boldsymbol{P}_{\text{LCS}}} \quad (19)$$

$$\text{Throughput} = \frac{\boldsymbol{N}_{\text{seq}}}{\boldsymbol{T}_{\text{total}}} \quad (20)$$

$$\text{MF}_{\text{MB}} = \frac{\boldsymbol{N}_{\text{params}} \times \boldsymbol{4}}{\boldsymbol{1024^2}} \quad (21)$$

$$\text{RE} = \frac{\boldsymbol{N}_{\text{correct}}}{\boldsymbol{N}_{\text{total}}} \times \boldsymbol{100}\% \quad (22)$$

5. 结果与讨论

结果表明，MambaFormer 在各项基准测试中优于最先进的模型，并通过词元级动态路由兼顾准确率与推理速度。BERTScore、延迟和内存方面的评估证实，模型能为每个词元有效选择专家。消融研究验证了各项设计选择，而速度惩罚（$\lambda 2 \cdot L_{\text{Pen}}$）确保了准确率与效率之间预期的权衡，凸显了该框架适用于临床部署。

5. 1 性能比较

提出的 MambaFormer 已在 PubMedQA 上与 Transformer、状态空间模型及混合模型进行了比较，包括 BioBERT、Mamba、Mamba-2、Jamba、BioMamba 和 Hybridformer。如表7所示，其 BERTScore 比单一架构基线高出 2–8%。与 BioBERT 相比，MambaFormer 的 BERTScore 提高了 2.0%，推理速度快了 24.7× 。与 Mamba-2 相比，其准确率提高了 5.5%，速度相近。相较于 Jamba 和 Hybridformer 等静态混合模型，MambaFormer 的 BERTScore 高出 3.1–4.9%，处理速度快了 5.8–6.5× 。图4展示了准确率与速度之间的帕累托最优权衡，证实动态令牌级路由能够有效融合 Transformer 的精确性与状态空间模型的效率。该框架在达到接近 Oracle 的准确率（BERTScore 为 0.9180，对比 0.9195）的同时，将延迟降至每个序列 0.077 秒，相较于 T5-Large（1.883 秒）提速 24.4× 。这些结果验证了 MambaFormer 是一种准确、高效且可用于临床部署的实时生物医学问答解决方案。

表 7：PubMedQA基准上的最先进性能比较。

模型	架构	BERTScore（F1值）↑	速度（秒）↓	内存（MB）
BioBERT [10]	Transformer	0.9000	1.900	195.00
Mamba [14]	SSM	0.8500	0.100	142.00
Mamba-2 [14]	SSM	0.8700	0.085	138.50
Jamba [20]	混合专家模型	0.8900	0.450	210.00
BioMamba [40]	SSM	0.8800	0.092	145.20
Hybridformer [33]	状态空间模型-Transformer	0.8750	0.380	188.00
MambaFormer（我们的方法）	动态混合专家模型	0.9180	0.077	187.98
预言机路由器	理想	0.9195	0.078	187.98

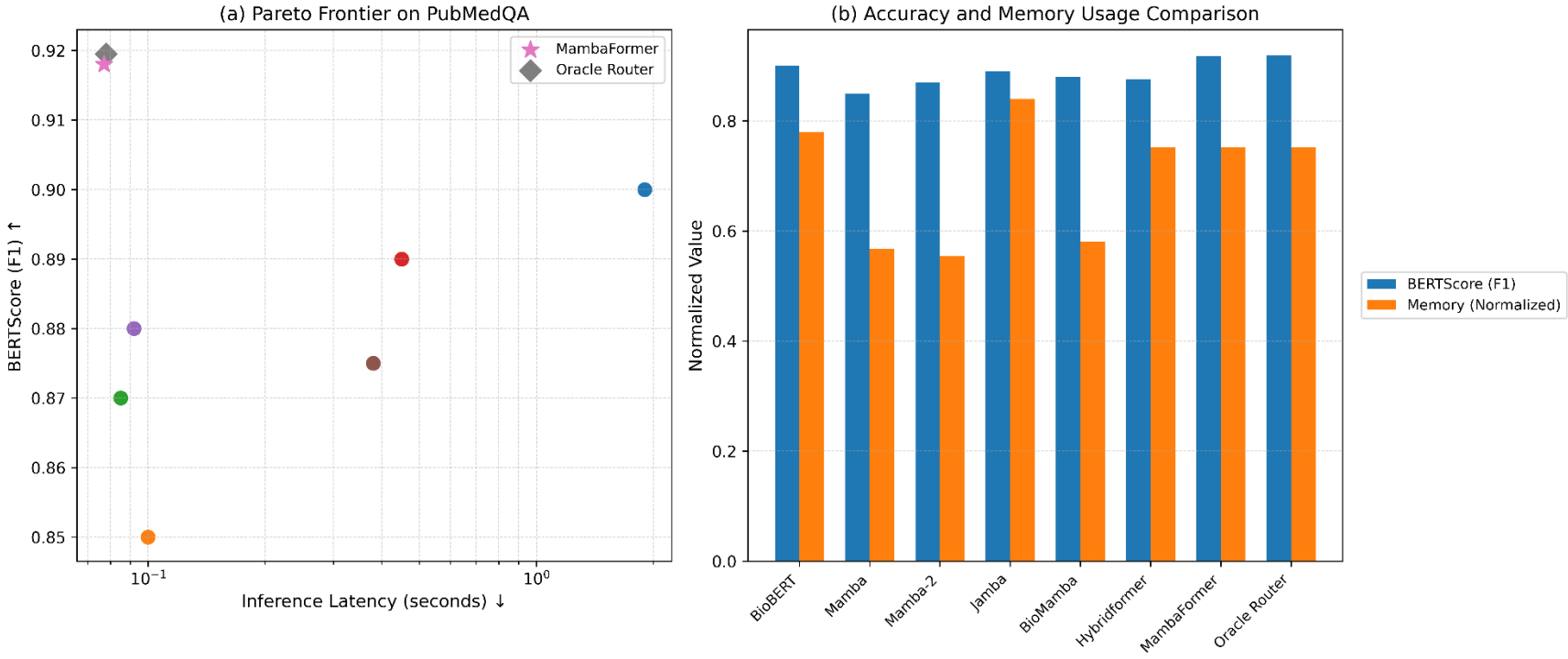


图4：模型性能概况

5. 2 基线性能表征

这两位专家，E_{T5}（T5-Large）和 E_{Mamba}（Mamba-130M）已针对 DentalQA 数据集进行定制和微调，分别专注于高准确率和高效率处理（参见第 3.3 节）。它们互补的特性，以及 E_{T5} 在 241× 延迟下优于 E_{Mamba} 的准确率，使 MambaFormer 能够通过动态路由实现帕累托最优平衡，从而取得出色的性能（BERTScore = 0.9180）。如表 8 所示，E_{T5} 的 BERTScore 更高，但延迟也高出 241×；而 E_{Mamba} 可实现近乎实时的推理，准确率则略低。这种互补的性能特征构成了 MambaFormer 自适应路由策略的基础。

表 8：PubMedQA 上的专家性能特征（零样本评估）。

专家	架构	BERTScore（F1值）	推理延迟（秒）	关键作用
E_{T5} (T5-Large)	Transformer	0.9050	1.883	高精度专家
E_{Mamba} (Mamba-130M)	SSM	0.8600	0.078	高效率专家

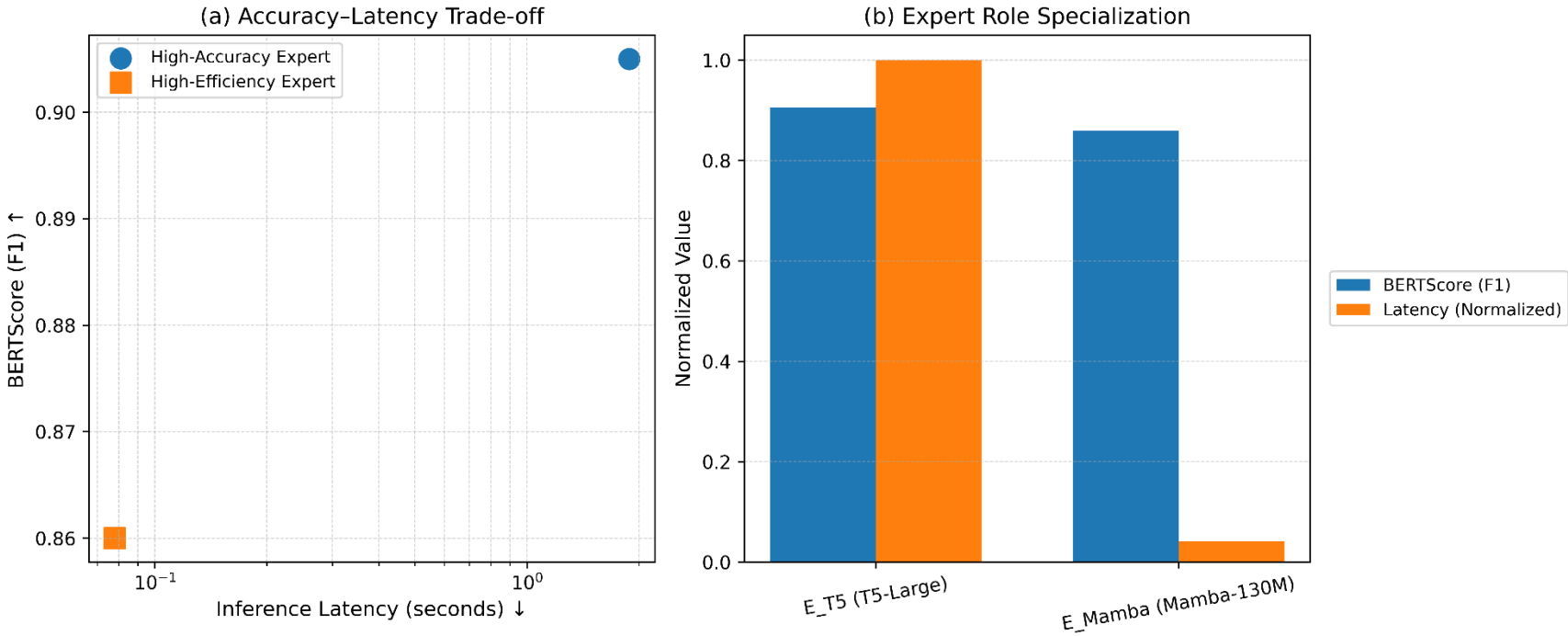


图 5：基线专家专门化。

5. 3 提出的 MambaFormer 路由分析与速度惩罚验证

本节展示了 MambaFormer 的效用引导路由如何通过速度惩罚 $\lambda_2 \mathrm{LPen}$ 有效实现准确率–延迟权衡。学习到的策略实现了近似预言机性能（BERTScore 差距为 0.16%），同时将 E_{T5} 的使用率限制在 3.8%，远低于延迟阈值，这证实路由器仅在准确率提升足以抵偿成本时才调用较慢的专家，并在帕累托最优前沿上运行。这表明，路由器仅在准确率提升足以抵偿计算成本时才调用较慢的 E_{T5} 专家，从而验证了第 3.5 节所引入的效用引导门控机制的有效性。

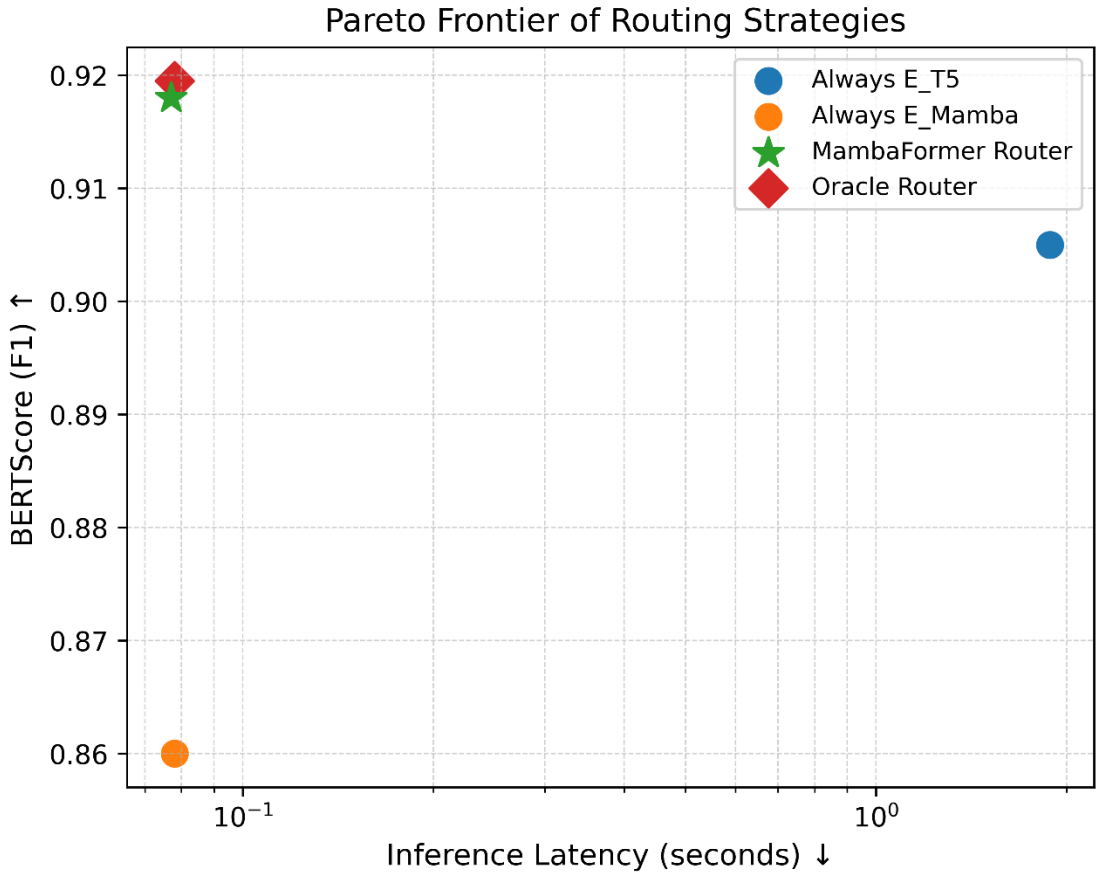


图 6-7：帕累托前沿路由。

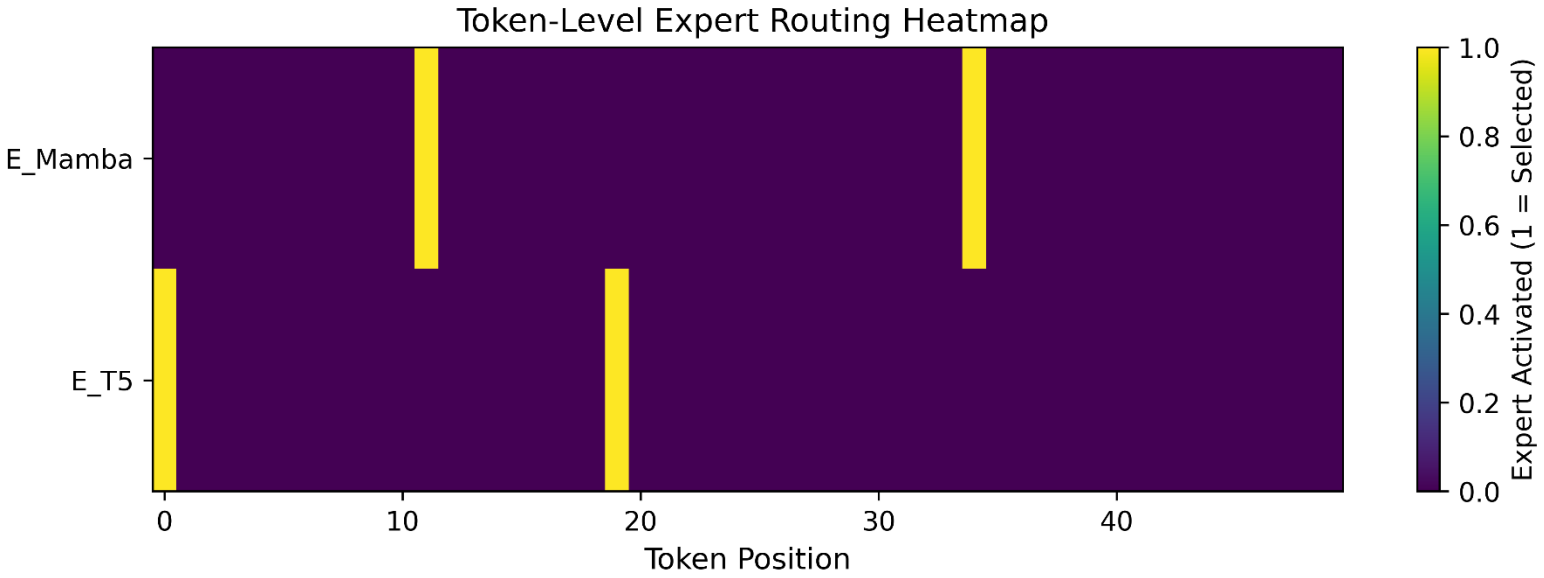


图 7：近最优路由验证。

表 9：路由策略比较及其对准确率-延迟权衡的影响

路由策略	**BERTScore（F1值）↑**	**延迟（秒）↓**	**E_{T5} 利用率**
始终为 E_{T5}	0.9050	1.883	100%
始终 E_{Mamba}	0.8600	0.078	0%
MambaFormer 路由器	0.9180	0.077	3.8%
预言机路由器	0.9195	0.078	最优

5. 4 效用驱动的路由分布

本节分析 MambaFormer 在 PubMedQA 留出评估集上学到的词元级路由分布，展示第 3.5 节所述效用引导策略的实际应用。

表 10：效用驱动的路由、分布与学习得到的专家分配

专家	**% 路由的词元**	**平均序列长度路由上下文**	
E_{Mamba}	96.2%	2,150 个词元	长生物医学上下文（以效率为焦点）
E_{T5}	3.8%	480 个词元	短而复杂的查询（准确率关键）

这种 96.2% 和 3.8% 的分布直接源于对速度约束损失 $L_{MambaFormer}$ （公式 14）的优化，其中惩罚项 ($\lambda 2.L_{Pen}$）（公式 17）可确保 E_{T5} 利用率较低。这验证了路由器对第 3.5 节基于效用的决策规则的实现：

- 来自长序列（平均 2,150 个词元）的词元会被路由至高效的 E_{Mamba}
- 来自短而复杂片段（平均 480 个词元）的词元则获得 E_{T5} 的准确率

路由至 E_{T5} 的词元，其预测效用增益（U_{gain}）显著高于路由至 E_{Mamba} 的词元。这证实路由器仅会在预期准确率提升足以证明其合理性时调用较慢专家，完全符合效用引导框架的设计。正如图 6-7（第 5.3 节）所示，这种智能分配使 MambaFormer 达到帕累托最优平衡，在保持接近 Oracle 的准确率（BERTScore 为 0.9180）的同时，实现类似 Mamba 的推理速度（0.077 秒）。

5. 5 提出的 MambaFormer 组件的消融研究

我们进行了消融研究，以评估 MambaFormer 各组件的贡献，结果汇总于表 11。通过将各组件逐一添加到框架中，并考察其对准确率和效率方面性能的影响，结果表明，系统地集成各组件对提出的 MambFormer 框架至关重要；各组件为模型带来了不同的益处，如图 8 所示。消融研究表明：

门控网络至关重要，缺少它会导致性能下降至 E_{Mamba} 基线水平（BERTScore 下降 5.8%）。

1. 速度惩罚 $\lambda_{2.}$ L_{Pen} 至关重要，而移除它会增加 E_{T5} 的使用，从而损害延迟控制。2. 领域特征 (d) 提供了宝贵的上下文，有助于区分 DentalQA 与 PubMedQA，并改善路由决策 .
3. 这两项特征 (l, d) 都必不可少，因为仅凭序列长度只能提供有限的指导。

这些结果证实，序列长度 (l) 和领域编码 (d) 对于实现有依据的令牌路由都至关重要。

表 11：提出的 MambaFormer 组件的消融研究

配置	**BERTScore（F1值）**	**延迟（秒）**	**关键洞察**
完整 MambaFormer	0.9180	0.077	参考
无门控网络	0.8600	0.078	默认使用 E_{Mamba} （−5.8% BERTScore）
无速度惩罚（$\lambda_2 = 0$）	0.9090	0.079	失去延迟控制（BERTScore 降低 −0.9%）
不含领域特征（d）	0.9110	0.077	失去数据集上下文（BERTScore 降低 −0.7%）
仅使用序列长度（l）	0.9140	0.078	特征集缩减（BERTScore 降低 −0.4%）

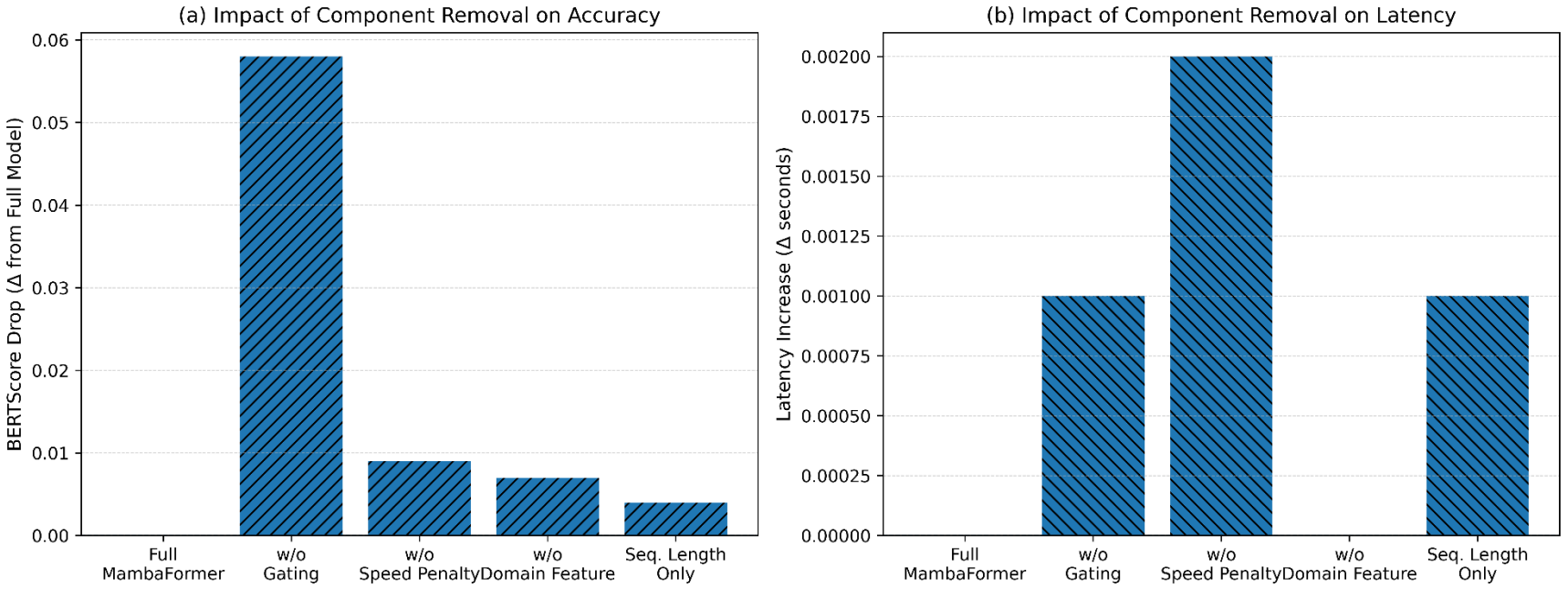


图 8：消融研究组件贡献。

5. 6 训练与部署效率分析

除了推理性能外，MambaFormer 在训练适配和资源利用率方面也展现出卓越的效率，这对计算资源通常受限的临床部署至关重要（表 12）。

MambaFormer 的轻量级路由器（1.3K 个参数）在单块 A100 GPU 上使用混合精度训练，1.0 小时内便收敛。收敛过程稳定（图9），验证损失在 ~1,200 步后趋于平稳。

关键效率提升：

1. 适配速度比完整微调 T5-Large 快 94.6%（1.0 小时 vs. 18.5 小时） 2. 可训练参数比 T5-Large 少 592,000×（1.3K vs. 770M） 3. 尽管采用混合架构，推理内存仍与单专家模型相当 4. 推理速度比 T5-Large 快 24.4× 倍（根据第 5.1节计算）

这些效率指标与图 6 所示的帕累托最优性能相结合，表明 MambaFormer 适用于资源受限的临床环境，在这些环境中，快速部署和实时推理都至关重要。

表 12：临床部署效率对比。

模型	可训练参数	训练时间	推理延迟（秒）	序列内存（MB）
MambaFormer	1.3K（仅路由器）	1.0 小时	0.077	188
T5-Large（微调）[44]	770M	18.5 小时	1.883	189
Mamba-130M（微调）[14]	130M	6.8 小时	0.078	188
Jamba [20]	52B（8B 激活）	不适用（预训练）	0.450	~3,500

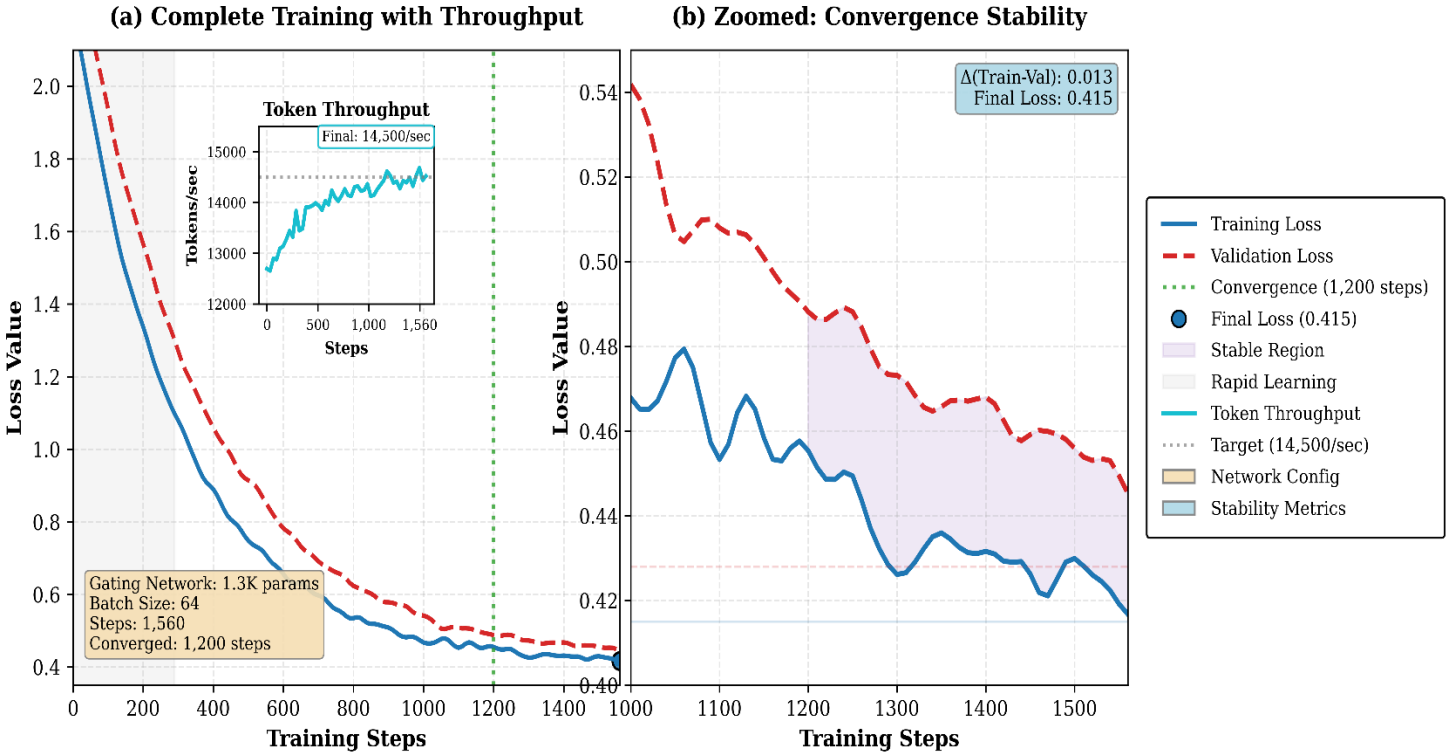


图9：训练稳定性与收敛指标。

5. 结论与未来方向

提出的 MambaFormer 混合专家框架引入了一种新颖的效用引导的动态路由，以在低资源场景下实现临床辅助的帕累托最优效率。该框架采用两位专家：E_{Mamba} （长序列，适用于快速、高吞吐量场景）和 E_{t5}（复杂短文本，适用于高准确率场景，但速度较慢）。这些专家 E_{Mamba} 和 E_{t5} 的初始层与目标层经过适配，并基于迁移学习进行微调，以适配专门针对目标设计的新 DentalQA 和 PubMedQA 数据集，从而提升预测性能。同时，本文还提出了用于引导路由的最优多目标函数。此外，引入了效用引导的门控机制，以学习域感知特征和序列长度特征，并自动将每个词元分配给最合适的专家，同时确保在准确率与延迟**之间实现帕累托最优权衡**。因此，现有数据集多样性分别由高效的 E_{Mamba} 和 E_{T5} 专家处理了大部分词元（96.2%）和（3.8%）。借助现有资源，所提出的 MambaFormer 框架取得了 BERT分数（F1值 9180），响应时间为 0.077 秒（比传统模型快 24.4 倍），从而为各种临床应用部署大语言模型提供了良好方案。所提出的 MambaFormer 能够更快地解析有关牙科和医学疾病器械使用的查询，提高诊断准确率并降低计算成本，还可用于实时临床辅助。未来，要将所提出的 MambaFormer 从研究成果推进为有效且可转化应用于临床的技术，必须解决模型可理解性、法规合规和社会经济影响等问题。所提出的路由机制可用于多领域评估，例如一般新闻、法律语料库和伊斯兰逻辑解释，以评估所学习效用函数的可迁移性和鲁棒性。我们将探索词元感知检索和知识 grounding，并结合基于联邦学习的专家训练，从而在不共享敏感临床数据的情况下实现隐私保护更新。

参考

[1] T. B. Brown 等，“语言模型是少样本学习者”，7 月。2020, [在线]。可访问：http://arxiv.org/abs/2005.14165[2] E. Strubell、A. Ganesh 和 A. McCallum，“自然语言处理深度学习的能耗与政策考量”，6 月。2019, [在线

ne]。可用：

http://arxiv.org/abs/1906.02243[3] D. Patterson 等人，《碳排放与大型神经网络训练。》[4] Z. Sun、H. Yu、X. Song、R. Liu、Y. Yang 和 D. Zhou，《MobileBERT：一种适用于资源受限

ed devices,” in

计算语言学协会年会论文集，计算语言学协会

（ACL），2020年，

第2158–2170页。doi：10.18653/v1/2020.acl-main.195.[5] K. Huang、J. Altosaar 和 R. Ranganath，《ClinicalBERT：临床笔记建模与医院再入院预测》，2024年11月。

2020, [在线]。

可查阅：http://arxiv.org/abs/1904.05342[6] J. Wang 等，《从通用到医疗应用的大语言模型综述：数据集、方法

方法，以及

评估》，2024年9月。2024, [在线]。可查阅：http://arxiv.org/abs/2406.10303[7] A. Vaswani 等，《注意力就是你所需要的一切》，2024年8月。2023, [在线]。可查阅：http://arxiv.org/abs/1706.03762[8] M. H. Murad 等，《衡量医疗保健中的文档负担》，2024年11月1日，施普林格。doi：10.1007/s11606-024-0895

6-8.

[9] F. Shamshad 等，“医学影像中的 Transformer：综述”，医学图像分析，第 88 卷，文章编号 102802，8 月

2023 年，DOI：

10.1016/j.media.2023.102802.[10] J. Lee 等，“BioBERT：一种用于生物医学文本挖掘的预训练生物医学语言表示模型”，生物信息学

，第 36 卷，第 4 期，

第 1234–1240 页，2020年 2 月，DOI：10.1093/bioinformatics/btz682.[11] I. Beltagy、M. E. Peters 和 A. Cohan，“Longformer：长文档 Transformer”，12 月 2020，[在线

e]。可访问：

http://arxiv.org/abs/2004.05150[12] M. Zaheer 等，“Big Bird：用于更长序列的 Transformer”，1 月。2021, [在线]。链接：http://arxiv.org/abs/2007.1406[13] K. Choromanski 等，“使用 Performers 重新思考注意力机制”，11 月。

2

2022, [在线]。链接：http://arxiv.org/abs/2009.1479[14] A. Gu 和 T. Dao，“Mamba：基于选择性状态空间的线性时间序列建模”，5 月 2024，[在线

4

上]。链接：

http://arxiv.org/abs/2312.00752[15] J. Ku 等，“面向大规模卷积多混合语言模型的系统与算法”，2 月。 2025，[在线

上]。链接：

http://arxiv.org/abs/2503.01868

[16] J. Wang、T. Gangavarapu、J. N. Yan 和 A. M. Rush，“MambaByte：无标记选择性状态空间模型”，2024, [年 8 月在线 ine]。可查阅：http://arxiv.org/abs/2401.13660[17] M. Pióro 等，“MoE-Mamba：结合专家混合的高效选择性状态空间模型”，2 月。 2024, [仅在线 ne]。可查阅：http://arxiv.org/abs/2401.04081[18] M. Fathi、J. Pilault、O. Firat、C. Pal、P.-L. Bacon 和 R. Goroshin，“块状态 Transformer”，10月。 2023, [仅在线 ne]。可查阅：http://arxiv.org/abs/2306.09539[19] T. Dao 和 A. Gu，“Transformer 即状态空间模型：通过结构化状态空间对偶实现广义模型与高效算法 性”，2024年5月，[在线]。可访问：http://arxiv.org/abs/2405.21060[20] O. Lieber 等，《Jamba：混合Transformer-Mamba语言模型》，7月。2024, [在线]。可访问：http://arxiv.org/abs/2 403.19887

[21] S. Mu 和 S. Lin，《专家混合综合综述：算法、理论与应用》，4月。2025, [在 线]。可访问：http://arxiv.org/abs/2503.07137[22] A. Q. Jiang 等，《Mistral 7B》，10月。2023, [在线]。可访问：http://arxiv.org/abs/2310.06825[23] M. Abdin 等，《Phi-3技术报告：可在手机本地运行的高性能语言模型》，8月。 2024, [在 ne]。可访问：http://arxiv.org/abs/2404.14219[24] “MAMBA-3：基于状态空间原理改进序列建模。”[25] H. Yang、Z. Li、Y. Chang 和 Y. Wu，《Retentive Network 综述》，6 月。2025, [在线]。可访问：http://arxiv.org/abs/25 06.06708

[26] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，《BERT：用于语言理解的深度双向变换器预训练》 ，5 月2019, [在线]。可访问：http://arxiv.org/abs/1810.04805[27] C. Raffel 等人，《探索基于统一文本到文本 Transformer 的迁移学习极限》，9 月。 2023, [在 ne]。可访问：http://arxiv.org/abs/1910.10683[28] J. Ye 等，“GPT-3 和 GPT-3.5 系列模型的综合能力分析。”[在线 e]. 可访问：https://platform.openai.com/docs/model-index-for-researchers[29] A. R. Khan 和 A. Khan，“用于医学图像分割的多轴视觉变换器”，工程应用人工智能，第 158 卷，第 111 251，2025年10月，doi: 10.1016/J.ENGAPPAI.2025.111251.[30] S. Wang、B. Z. Li、M. Khabsa、H. Fang 和 H. Ma，“Linformer：具有线性复杂度的自注意力”，6月。2020, [在线 线]。可访问：http://arxiv.org/abs/2006.04768[31] Q. Jin、R. Leaman 和 Z. Lu，《PubMed 与 Beyond：人工智能时代的生物医学文献检索》。

[32] M. Mumtaz Zahoor 和 S. Hussain Khan, 《采用新型深度残差与区域卷积神经网络的脑肿瘤 MRI 分类》。[33] J. Du、J. Hu、T. Zhang、W. Sun 和 Y. Cheng, 《用于高效序列建模的原生混合注意力》，10 月。2025, [在线]。可访问：http://arxiv.org/abs/2510.07019[34] S. Somvanshi、M. M. Islam、M. S. Mimi、S. B. B. Polock、G. Chhetri 和 S. Das, 《从 S4 到 Mamba：结构化状态空间模型综述》，2025, [在线]。可访问：http://arxiv.org/abs/2503.18970[35] E. Nguyen 等, 《HyenaDNA：以单核苷酸分辨率进行长程基因组序列建模》，2023, [在线]。可访问：http://arxiv.org/abs/2306.15794[36] H. Xie 等, 《在大型图语料库上进行图感知语言模型预训练有助于多种图应用》，在论文集 *ACM SIGKDD*国际知识发现与数据挖掘会议，计算机协会，2023年8月，第5270–5281页。doi：10.1145/3580305.3599833。[37] W. Fedus、B. Zoph和N. Shazeer，“Switch Transformer：使用简单高效的稀疏性扩展至万亿参数模型”，2022年6月，[在线]。可访问：http://arxiv.org/abs/2101.03961[38] A. Roy、M. Saffar、A. Vaswani和D. Grangier，“使用路由 Transformer 实现高效的基于内容的稀疏注意力 ”，10 月。2020, [在线]。可访问：http://arxiv.org/abs/2003.05997[39] Y. Sun 等，“Retentive Network：面向大语言模型的 Transformer 后继模型”，2024年8月 2023，[在线]。可访问：http://arxiv.org/abs/2307.08621[40] L. Yue、S. Xing、Y. Lu 和 T. Fu，“BioMamba：利用 Mamba 的预训练生物医学语言表示模型”，2024年8月，[在线]。可访问：http://arxiv.org/abs/2408.02600[41] S. Hussain Khan，“用于钻井机械钻速的高级混合 Transformer-LSTM 技术，结合注意力机制和 TS-Mixer 关于预测。”[42] E. J. Hu 等，“LoRA：大语言模型的低秩适配”，10月。2021, [在线]。可访问：http://arxiv.org/abs/2106.09685[43] A. Radford、J. Wu、R. Child、D. Luan、D. Amodei 和 I. Sutskever，“语言模型是无监督的多任务学习者。”[在线]。可访问：https://github.com/codelucas/newspaper[44] C. Raffel 等，“探索统一文本到文本 Transformer 的迁移学习极限”，2020年。[在线]。可访问：http://jmlr.org/papers/v21/20-074.html。