从 **Test-Time Training** 到大语言模型的 **Test-Time Learning**：以“在线优化与目标函数匹配”视角理解输入困惑度适配

Zijun Mao（毛子鋆） 学号：**2025439109**
天津大学福州国际联合学院，天津 300072
The International Joint Institute of Tianjin University, Fuzhou, Tianjin University, Tianjin 300072, China
zijunmao@tju.edu.cn

Abstract

本文精读 ICML 2025 的 TTL 工作，并将其核心贡献理解为：把“部署期域偏移”转写成一个预算受限的在线优化问题——在无标签测试流上，通过输入困惑度最小化让模型对目标域文本分布进行“快速似然对齐”。我进一步将方法拆成三条部署逻辑：(i) 困惑度充当“域不匹配”的可微代理信号；(ii) SEL 把有限反传预算集中到“最不匹配”的样本；(iii) LoRA 用受限更新空间换取稳定性并降低遗忘风险。为刻画该工作相较既有研究的增量与代价，本文进一步对比 ICML 2020 的 Test-Time Training（TTT）、ICLR 2021 的 Tent、NeurIPS 2022 的 MAE-TTT，以及 ICML 2025 在 LLM few-shot 场景下的 TTT，并构建“目标函数—更新子集—优化过程—任务与偏移类型”的比较分析框架。最后，本文结合课程所涵盖的优化方法、深度网络，以及 Transformer/生成式建模等内容，讨论测试时学习作为“新型机器学习范式”的方法论意义与开放问题。项目代码见：https://github.com/dogmao/zijun-ml-course-tlm-ttl-quantile-SEL. 同时，附录给出我在复现中提出的 Quantile-SEL 小改动及其“单位反传收益”分析。

1. 引言

1.1. 背景与动机：为什么要把“推理”当成“继续优化”

经典监督学习通常默认训练数据与测试数据同分布，但在真实部署中往往存在不可忽视的分布偏移：训练阶段采样得到的联合分布 $P_{\text{train}}(x, y)$ 与部署阶段所遇到的 $P_{\text{test}}(x, y)$ 不一致，进而导致通过经验风险最小化得到的模型在目标环境中性能显著退化 (Quinonero-Candela et al., 2008)。分布偏移可呈现多种结构形式。其中，协变量偏移（covariate shift）强调输入边缘分布发生变化而条件分布保持不变，即 $P_{\text{train}}(x) \neq P_{\text{test}}(x)$ 且 $P_{\text{train}}(y \mid x) = P_{\text{test}}(y \mid x)$。在该设定下，可通过对对数似然进行重要性加权以改进预测推断 (Shimodaira, 2000)。更一般地，域适配理论给出了目标域误差的上界，其由源域误差与域间差异度量共同决定，从而揭示了仅依赖源域监督信号难以保证目标域泛化的根本原因 (Ben-David et al., 2010)。为便于后文在不同测试时自适应范式下讨论“偏移来源”与“可用假设”，Table 1对上述典型偏移设定及其常见处理思路进行了对照整理。

偏移类型 / 设定	形式化描述与常见假设（来自引用文献）
Dataset shift（总称）	训练与测试分布不一致：$P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y)$；真实部署中普遍存在并导致性能退化 (Quinonero-Candela et al., 2008).
Covariate shift	输入边缘分布变化但条件分布不变：$P_{\text{train}}(x) \neq P_{\text{test}}(x)$ 且 $P_{\text{train}}(y \mid x) = P_{\text{test}}(y \mid x)$；可对对数似然做重要性加权以改进预测推断 (Shimodaira, 2000).
Domain adaptation（误差上界视角）	目标域误差可由源域误差与域间差异度量共同控制的上界分解，揭示仅靠源域监督难以保证目标域泛化 (Ben-David et al., 2010).

Table 1. 分布偏移常见设定与典型假设（基于 (Quinonero-Candela et al., 2008; Shimodaira, 2000; Ben-David et al., 2010)）。

从应用视角看，部署阶段通常伴随多重约束：标注代价高、分布变化难以预先枚举、数据以流式或批式到达，且系统需在不中断服务的条件下保持性能稳定。在此背景下，仅在训练期一次性完成离线学习往往不足；相反，需要利用测试阶段可获得的未标注数据，在测试时对模型进行轻量更新，从而实现对目标环境的自适应与鲁棒泛化。围绕这一核心诉求，测试阶段自适应（test-time adaptation / training / learning）逐渐形成一条从判别模型到生成模型、从视觉到语言模型的研究主线。早期工作提出测试时训练（Test-Time Training, TTT），将单个未标注测试样本构造为自监督学习任务，并在预测前更新模型参数；该方法也可自然扩展到在线数据流 (Sun et al., 2020)。随后，完全测试时自适应（Tent）在不改变训练过程的前提下，通过最小化预测熵驱动在线更新，并仅更新归一化层统计量与通道仿射参数，以获得更稳定的增益 (Wang et al., 2021)。在更强的生成式自监督方向上，MAE-TTT 以掩码自编码作为单样本自监督目标，在多种分布偏移的视觉基准

上进一步提升泛化能力 (Gandelsman et al., 2022)。在我的理解里，这类方法的共同点不是“更新”本身，而是把部署期不确定性显式纳入优化过程：用可计算的代理损失在测试流上做小步修正，从而把“静态模型”变成“可适配的算法过程”。

当研究对象从视觉判别模型扩展到大语言模型（LLM）时，这一问题的现实紧迫性进一步凸显。尽管大规模预训练语言模型在少样本设定下已展现出较强的任务适应能力 (Brown et al., 2020)，但当专业领域知识、语言变体或目标场景分布发生变化时，其性能仍可能显著下降。近期 ICML 2025 的工作提出面向 LLM 的测试时学习（Test-Time Learning, TTL）范式：在测试阶段仅利用未标注测试数据，通过输入困惑度最小化实现自监督增强，并基于“高困惑度样本更具信息量”的观察提出样本高效策略。为缓解在线更新可能引发的灾难性遗忘，作者进一步采用参数高效的 LoRA 更新而非全参数优化，以提升适配稳定性 (Hu et al., 2025; 2022; Kirkpatrick et al., 2017)。与之互补，另一项 ICML 2025 工作系统研究了少样本推理场景下使用 TTT 进行临时参数更新的有效性，表明在 ARC 与 BBH 等基准上，结合 in-context 示例进行测试时训练可显著优于标准少样本提示策略 (Akyürek et al., 2025)。这些进展共同表明：在分布偏移不可避免且监督信号稀缺的部署场景中，测试阶段自适应正成为提升鲁棒泛化与系统可用性的关键机制。

1.2. 问题定义与术语解释

为避免不同文献对“测试时更新”范式的混用，本文对 TTT、TTA 与 TTL 三类设定作如下区分。

Test-Time Training (TTT). TTT 的核心思想是：在推理阶段利用可由测试输入构造的自监督/辅助损失对模型进行更新，再输出主任务预测 (Sun et al., 2020)。在计算粒度上，TTT 可针对单样本执行（将每个测试样本转化为一次自监督训练后再进行预测），也可在在线流式数据上按批次进行增量更新 (Sun et al., 2020)。在具体实现上，MAE-TTT 选择掩码自编码作为单样本自监督目标，将生成式重建任务用于测试时优化 (Gandelsman et al., 2022)。在语言模型的少样本推理场景中，TTT 亦可实例化为“对 in-context 示例进行临时训练”的机制，从而提升推理性能与规则泛化能力 (Akyürek et al., 2025)。

Test-Time Adaptation (TTA). TTA 通常强调在不访问源域训练数据且不引入额外标注的条件下，仅依赖目标域测试数据对模型进行自适应。Tent 属于“完全测试时自适应”设定：模型在测试阶段仅拥有测试数据与自身参数，通过最小化预测熵进行在线更新，并以更新归一化层统计量与通道仿射参数的方式实现批次级自适应 (Wang et al., 2021)。与 TTT 相比，TTA 的目标函数更偏向无监督的置信度/一致性驱动，更新范围也更强调稳定、可控的参数子集更新 (Wang et al., 2021)。

Test-Time Learning (TTL). TTL 在近期 LLM 研究中被更明确地表述为一种测试阶段学习范式：在部署阶段仅利用未标注测试数据，实现对目标域的动态适配。与以判别置信度为核心的 TTA 不同，Hu et al. 将 LLM 的 TTL 过程形式化为输入困惑度最小化，并提出样本高效策略，强调高困惑度样本在优化中的信息量更大。同时，为提升适配稳定性并缓解遗忘风险，作者采用 LoRA 进行轻量更新 (Hu et al., 2025; 2022)。由于在线更新可能干扰既有知识，连续学习中的经典难题——灾难性遗忘——也为 TTL 的稳定优化提供了重要参照 (Kirkpatrick et al., 2017)。

1.3. 本文贡献与结构

本文以测试阶段自适应为主线，围绕课程“分布偏移下的鲁棒泛化”主题，开展以下工作：(1) 精读主文（**ICML 2025**）：系统梳理 Hu et al. 提出的 LLM 测试时学习（TTL）范式及其关键设计（输入困惑度目标、样本高效策略与 LoRA 稳定更新），并讨论其面向部署场景的工程含义 (Hu et al., 2025)；(2) 顶会对比综述：以 TTT（ICML 2020）、Tent（ICLR 2021）、MAE-TTT（NeurIPS 2022）与 Few-shot TTT（ICML 2025）为参照，从可用数据、优化目标、更新粒度与在线假设等维度比较不同测试时更新范式的共性与差异 (Sun et al., 2020; Wang et al., 2021; Gandelsman et al., 2022; Akyürek et al., 2025)；(3) 课程映射与心得反思：将上述方法与“优化方法/泛化/Transformer 与大模型/新型学习范式”等课程内容对应，给出对测试时学习在未来大模型部署与持续适配中的个人理解与展望。

全文结构如下：第2章给出问题背景与形式化设定；第3章回顾相关工作；第4章精读 TTL 主文并解析方法细节；第5章进行顶会方法对比与综合讨论；第6章完成课程映射与心得；最后总结并提出开放问题。此外，附录对精读工作中提出的核心算法思想进行了复现，并通过补充实验进一步验证与阐释本文的理解。

2. 预备知识与课程内容映射

2.1. 优化视角：从训练期到测试期的参数更新

从优化角度看，深度模型训练通常对应在源分布 $p_{\text{src}}(x, y)$ 上最小化经验风险（或其正则化形式），并通过随机梯度下降类方法迭代得到参数 $\theta^\star$。在大规模学习情境下，随机梯度下降及其变体因其可扩展性而被广泛采用；其核心是以小批量样本近似总体梯度并执行一阶更新 (Bottou, 2010)。在工程实现中，自适应梯度方法（如 Adam）通过对一阶矩与二阶矩的指数滑动平均构造逐坐标自适应步长，从而在许多深度学习任务中表现出良好的数值性质与收敛性能 (Kingma & Ba, 2015)。

当测试分布 $p_{\text{test}}(x, y)$ 与训练分布发生偏移时，固定参数 $\theta^\star$ 的静态推断可能出现性能劣化。测试阶段自适应（Test-Time Adaptation/Learning/Training）的共同形式是：在不访问源训练数据（或仅访问极有限缓存）的

约束下，针对到达的测试样本（通常无标签）构造一个可在测试时计算的代理目标，并据此对参数（或其子集）执行少量迭代更新。令 x_t 表示第 t 个测试批次（或样本流片段），ϕ 表示允许更新的参数子集（例如仅更新归一化层参数，或以低秩形式更新部分权重），则可写为

$$\phi_t^{(k+1)} = \phi_t^{(k)} - \eta \nabla_\phi \mathcal{L}_{\text{adapt}}\left(x_t;\, \phi_t^{(k)}, \theta^\star\right), \quad k = 0, 1, \ldots, K-1, \tag{1}$$

其中 $\mathcal{L}_{\text{adapt}}$ 为测试时代理目标，K 为测试时步数预算（step budget），η 为步长。该框架覆盖多类经典设定。例如，TTT 在测试阶段通过自监督辅助任务损失驱动参数更新 (Sun et al., 2020)；Tent 通过最小化预测分布熵实现完全测试时自适应；TTL 则在大语言模型中利用与输入困惑度（perplexity）相关的目标进行测试时学习。在部署假设上，不同方法对更新频率及是否跨批次累积作出不同选择：既可以在每个测试片段内进行若干步更新，并将更新累积到后续片段（online/continual），也可以在每个片段前重置到 $\theta^\star$ 并仅做片段内更新（episodic）。由此可在适应性与稳定性之间形成可控权衡 (Wang et al., 2021; Hu et al., 2025)。

2.2. 表征学习与架构背景：CNN 到 Transformer/LLM

在视觉任务中，TTT、Tent 与 MAE-TTT 等方法多以卷积网络及其现代变体作为基础分类器。以残差网络为代表的深层 CNN 通过引入残差连接显著缓解深层网络的优化困难，并在多种视觉基准上成为强有力的表征学习骨干 (He et al., 2016)。与之配套的归一化机制同样关键：批归一化（Batch Normalization, BN）通过对小批量激活进行标准化，并学习可训练的仿射变换参数来改善训练动态 (Ioffe & Szegedy, 2015)。在测试时自适应中，Tent 的核心是在测试阶段以熵最小化目标驱动模型更新（实现上通常聚焦于归一化相关参数），从而在不依赖标签的前提下提升分布偏移下的鲁棒性 (Wang et al., 2021)。

TTT 则强调以自监督任务作为测试时学习信号：其在测试阶段对输入执行数据增强并优化自监督辅助目标，从而在目标域上对表征进行快速校正 (Sun et al., 2020)。在该范式中，旋转预测是经典自监督预训练/表征学习信号之一，能够在无标签条件下促使网络捕捉与语义相关的几何结构；相关工作系统展示了通过预测旋转角度学习判别性表征的可行性 (Gidaris et al., 2018)。沿着更强自监督信号的方向，MAE 通过对输入图像块进行高比例随机遮蔽，并以重建缺失像素为目标来学习表征；其采用非对称编码器—解码器结构以提高训练效率与可扩展性 (He et al., 2022)。MAE-TTT 进一步将该类遮蔽重建目标迁移到测试时训练范式中，使测试阶段更新由重建误差直接驱动，从而在视觉分布偏移情境下增强鲁棒性 (Gandelsman et al., 2022)。

当讨论对象从判别式视觉模型扩展到生成式大语言模型（LLM）时，目标函数与评估指标也随之变化。Transformer 以自注意力为核心构件，摒弃递归与卷积依赖，成为现代序列建模（尤其是 LLM）的主流架构 (Vaswani et al., 2017)。自回归语言模型通常以最大化序列对数似然（等价于最小化 token 级交叉熵）作为训练目标；在语言建模中，困惑度（perplexity）是该目标的常用可解释化度量之一，可视为平均负对数似然的指数化形式，用以反映模型对序列预测的不确定性 (Bengio et al., 2003)。这为后续 TTL 将“输入困惑度相关目标”作为测试时学习信号提供了直接的概念与数学铺垫：即使在无标签测试阶段，生成式模型也能为输入序列提供可微的似然/困惑度信号，从而支持基于梯度的一阶更新 (Hu et al., 2025)。

2.3. 参数高效适配：LoRA 与可部署性

在部署环境中，测试时自适应除追求鲁棒性提升外，还必须满足显存、算力与时延等约束。因此，参数高效（parameter-efficient）的更新机制成为将测试时学习落地到大模型的重要路径。LoRA 的基本思想是在保持预训练权重矩阵 W 冻结的前提下，仅学习一个低秩增量 ΔW，并将其参数化为两个低秩矩阵的乘积：

$$W' = W + \Delta W, \quad \Delta W = BA, \qquad \text{rank}(\Delta W) = r \ll \min(d_{\text{out}}, d_{\text{in}}), \tag{2}$$

其中 $A \in \mathbb{R}^{r \times d_{\text{in}}}$、$B \in \mathbb{R}^{d_{\text{out}} \times r}$，从而将可训练参数规模由 $O(d_{\text{out}} d_{\text{in}})$ 降至 $O(r(d_{\text{out}} + d_{\text{in}}))$ (Hu et al., 2022)。该低秩增量构造使模型在不改动主干参数的情况下即可获得有效的适配能力，尤其适合资源受限或需要频繁更新的场景中进行轻量级学习。进一步地，TTL 将 LoRA 纳入测试时学习框架：在不依赖源数据、以输入困惑度相关目标为信号的测试时更新过程中，LoRA 提供了可部署的参数更新接口，使 LLM 的测试时自适应能够在计算与稳定性约束下运行 (Hu et al., 2025)。

3. 相关工作

3.1. TTT：以自监督任务驱动的测试时训练（ICML 2020）

Test-Time Training（TTT）将分布偏移下的部署推理视为一个可在测试阶段继续学习的问题：面对单个无标签测试样本，方法将其构造为自监督学习任务，并在输出主任务预测之前对模型参数进行少步优化更新，从而使模型参数显式依赖测试输入而不依赖其未知标签。该思想并不要求训练期预先“覆盖所有未来分布”，而是强调从测试样本自身所携带的分布信息中获取反馈信号，并通过优化实现快速适配 (Sun et al., 2020)。

在算法实现上，TTT 采用多任务学习范式：训练阶段同时优化主任务监督损失与自监督损失，并在网络结构上形成共享特征提取器与两条任务分支的“Y 形结构”。具体而言，自监督任务与主任务共享底部特征提取器参数 θ_e，并分别拥有自监督分支参数 θ_s 与主任务分支参数 θ_m。训练时联合最小化 $\mathbb{E}\big[l_m(x, y; \theta_m, \theta_e) + l_s(x; \theta_s, \theta_e)\big]$，以确保测试期仅用自监督目标进行更新时仍与主任务表征保持兼容。在测试阶段，标准 TTT

对单个测试样本 x 最小化自监督损失 $l_s(x;\theta_s,\theta_e)$，并主要对共享特征提取器 θ_e 做少步微调，得到更新后的参数 θ_e^*，再用 (θ_e^*,θ_m) 进行主任务预测。为增强自监督信号的稳定性，TTT 在测试时对同一测试样本施加与训练期一致的数据增广，以“单样本多视角”的方式构造仅由该样本增广副本组成的小批量来执行优化 (Sun et al., 2020)。

在自监督任务选择上，原始 TTT 采用旋转预测：将输入图像旋转 0/90/180/270 度，并预测旋转角度作为四分类任务 (Gidaris et al., 2018)。此外，TTT 指出该框架可自然扩展到在线流式测试数据：当测试样本按序到达且分布平稳或缓慢变化时，可保留上一时刻的更新状态作为下一样本的初始化，从而累积利用历史测试数据所蕴含的分布信息。在工程细节上，由于测试时批量仅包含单一样本的增广副本，批归一化在小批量统计下可能不稳定，因此 TTT 在实验中以与批量大小无关的组归一化替代批归一化，以减轻小批量统计估计误差的影响 (Wu & He, 2018; Ioffe & Szegedy, 2015)。从效果角度看，TTT 在多个面向分布偏移鲁棒性的图像分类基准上取得改进，并在保持原分布性能的同时提升了对分布偏移的泛化能力 (Sun et al., 2020; Hendrycks & Dietterich, 2019)。

3.2. Tent：以熵最小化实现完全测试时自适应（ICLR 2021）

与 TTT 强调“训练期联合自监督以确保测试期更新兼容”不同，Tent 聚焦一个更严格的设定：完全测试时自适应（fully test-time adaptation），即在测试阶段仅可访问目标域无标签数据与模型自身参数，不依赖源域数据与监督信号。在该设定下，Tent 提出以预测分布的熵作为无标签优化目标，通过测试熵最小化促使模型在目标域输出更高置信度的预测，从而降低泛化误差 (Wang et al., 2021)。

Tent 的核心目标函数是最小化预测熵 $H(\hat{y})$。其强调直接对单一样本预测执行熵最小化可能产生退化解，因此通过在共享参数下对批量样本联合优化来约束该问题。在参数更新策略上，Tent 选择对归一化层执行“特征调制”（feature modulation）：一方面在线估计归一化统计量，另一方面仅优化通道级仿射变换参数（scale/shift），以低维、计算友好的方式支持测试期的批次级在线更新。这种仅更新少量关键参数的做法，使 Tent 能够在不改变训练流程的前提下，以较低的测试期优化成本实现持续适配 (Wang et al., 2021)。

在经验结果上，Tent 在 ImageNet-C 与 CIFAR-10/100-C 等分布偏移与腐蚀鲁棒性基准上显著降低分类错误率，并在 ImageNet-C 上取得更优表现 (Hendrycks & Dietterich, 2019)。此外，Tent 也覆盖源不可用（source-free）的域自适应与语义分割等任务场景，展示了该目标与更新策略在多任务、多结构下的适用性。总体而言，Tent 提供了一条以“无标签目标（熵）+ 低维参数更新（归一化仿射）”为主线的测试期自适应路径，并与 TTT 的“显式自监督预任务 + 单样本/在线更新”形成互补视角 (Sun et al., 2020; Wang et al., 2021)。

3.3. MAE-TTT：以掩码自编码强化自监督信号（NeurIPS 2022）

MAE-TTT 延续了 TTT “对每个测试样本进行自监督适配”的基本范式，但在自监督信号构造上选择了更贴近视觉统计结构的目标：以掩码自编码（Masked Autoencoders, MAE）的重建任务替代旋转预测，以缓解旋转预测在不同场景下可能“过易或过难”而导致自监督梯度难以有效迁移至主任务的问题。该工作从自然图像普遍满足的空间平滑性/局部冗余出发，将“遮挡部分输入并预测被遮挡内容”的空间自编码任务视为更通用、信息密度更高的自监督信号来源，并指出空间自编码是多种成功自监督方法的重要基础 (Gandelsman et al., 2022)。

在方法上，MAE-TTT 将 MAE 重建目标用于单样本测试时训练（one-sample learning）：测试阶段对每个测试图像执行少步重建优化，并在该过程中实现对目标分布的快速适配。其网络结构同样遵循 Y 形共享编码器与双头设计：共享特征提取器对应 MAE 编码器，自监督头对应 MAE 解码器，主任务头用于对象识别等下游任务。训练与初始化依托 MAE 在 ImageNet-1k 上的重建式预训练，并在下游组合方式上系统比较不同训练/冻结策略，以获得与测试期自监督更新更匹配的表征共享 (He et al., 2022)。在经验层面，MAE-TTT 在多个视觉分布偏移基准上提升了对象识别的泛化性能，并在四个对象识别数据集上报告显著收益 (Gandelsman et al., 2022)。

除经验改进外，该工作还提供了理论刻画：在简化的线性模型下，将测试时自监督更新带来的收益解释为偏差—方差权衡的变化，从而为“为何对单样本执行测试期自监督优化能够降低主任务错误”提供了可分析的视角。因此，MAE-TTT 不仅在预任务设计上对 TTT 做出了关键替换，也尝试从统计学习视角为其有效性提供理论支撑 (Sun et al., 2020; Gandelsman et al., 2022)。

3.4. LLM few-shot 场景的测试时训练（ICML 2025）

在大语言模型（LLM）的 few-shot 推理语境中，Aky”urek 等人系统研究了测试时训练（TTT）作为增强模型适应性与推理能力的机制：在推理阶段基于输入数据（尤其是 in-context 示例）构造损失，并临时更新模型参数，从而超越仅依赖提示学习（prompting）的标准 few-shot 范式。该研究强调，尽管 LLM 在训练分布内任务上表现突出，但在结构新颖（structurally novel）的任务上，即便给定少量示例仍可能受限；TTT 则提供了一条通过显式参数更新提升适应性的路径 (Akyürek et al., 2025)。

在基准评测上，该工作在 Abstraction and Reasoning Corpus（ARC）与 BIG-Bench Hard（BBH）上展示了 TTT 的显著增益。在 ARC 上，使用 in-context 示例进行 TTT 可相较微调基线带来最高约 6× 的准确率提升，并在

8B 参数规模的语言模型上于公开验证集达到 53.0%；进一步与程序综合方法集成可达 61.9%，并报告与平均人类水平相当 (Chollet, 2019)。在 BBH 上，TTT 在 10-shot 设定下相对标准 few-shot prompting 带来 7.3 个百分点的提升（50.5% → 57.8%）(Suzgun et al., 2022)。这些结果共同表明：当任务结构超出纯提示所能实现的内隐对齐能力时，推理期的少步参数更新可成为提升 few-shot 学习与抽象推理的有效补充机制 (Akyürek et al., 2025)。此外，BBH 作为从 BIG-bench 中筛选出的高难任务集合，其“挑战性”的定义与构建方式在相关基准文献中已有讨论 (Suzgun et al., 2022; Srivastava et al., 2022)。

4. 精读主文：大语言模型的测试时学习（ICML 2025）

4.1. 问题设定与符号系统

Hu *et al.*（ICML 2025）提出面向大语言模型（LLM）的测试时学习（Test-Time Learning, TTL）范式：在部署阶段面对目标域分布偏移，仅依赖无标签测试数据对模型进行轻量更新，从而提升模型在目标域上的鲁棒性与泛化能力 (Hu et al., 2025)。

形式化地，设训练（或对齐）阶段得到的基座模型参数为 Θ，目标测试分布为 $Q(x)$，测试时可获得来自 Q 的无标签样本 x。TTL 的一般目标可写为（对应原文式 (1)）

$$\min_{\Theta' \subseteq \Theta} \mathcal{L}(x; \Theta'), \quad x \sim Q(x), \tag{3}$$

其中 $\Theta' \subseteq \Theta$ 表示测试时允许被更新的参数子集；$\mathcal{L}(\cdot)$ 可取困惑度（perplexity）等自监督目标。该设定强调：(i) 不引入 y 标签；(ii) 更新发生在推理阶段；(iii) 需在训练开销、稳定性与遗忘风险之间权衡。

4.2. 方法总览：输入困惑度最小化的测试时学习

TLM 的核心主张是：通过最小化无标签测试输入的困惑度，可间接降低模型对输出的困惑度并提升预测质量。作者基于经验观察（见 Figure 1(a)(b)）与理论近似分析（Sec.4.1），将测试时学习具体化为对输入困惑度的优化。算法层面主要由三部分组成：**(1)** 输入困惑度最小化目标、**(2)** 高困惑度样本优先的样本高效学习策略、**(3)** 基于 **LoRA** 的轻量更新以缓解遗忘并降低开销（Algorithm 1）。此外，作者通过对不同困惑度样本参与测试时学习的对比实验（见 Figure 1(c)）进一步论证“高困惑度样本更具信息量”的样本效率动机 (Hu et al., 2025)。

对每个测试样本（或小批）x，TLM 先计算其输入困惑度，并据此得到样本选择/加权分数 $S(x)$；随后在 LoRA 参数空间内执行若干步梯度更新；最后使用更新后的模型生成答案。为兼顾在线部署，作者还提出一种在线协议：每处理 *100* 个样本更新一次模型，以减少反向传播次数并降低推理延迟。

4.3. 关键模块与设计选择

目标函数：输入困惑度最小化　困惑度（perplexity）用于度量语言模型对序列的拟合程度，是语言建模中的常用评估指标 (Bengio et al., 2003; Devlin et al., 2019; Brown et al., 2020)。对 token 序列 $x_1, \ldots, x_T$，原文将输入困惑度定义为平均负对数似然的指数形式（式 (2)）：

$$P(x; \Theta) = \exp\left(-\frac{1}{T} \sum_{t=1}^{T} \log p(x_t \mid x_{<t}; \Theta) \right). \tag{4}$$

在有标签场景下，更直接的测试时学习目标是最小化输出困惑度（式 (3)）：

$$\min_{\Theta} \; P(y \mid x; \Theta), \tag{5}$$

但 TTL 设定下无法获得 y。TLM 的关键转化是：在假设 (x, y) 语义对齐且自回归模型共享参数支撑输入建模与条件生成的前提下，使用一步梯度更新 $\Theta' = \Theta - \eta \nabla_\Theta(-\log P(x; \Theta))$，并对 $\log P(y \mid x; \Theta')$ 做一阶近似（式 (4)）：

$$\begin{aligned} \log P(y \mid x; \Theta') \approx \log P(y \mid x; \Theta) \\ + \eta \left[\nabla_\Theta \log P(x; \Theta)\right]^\top \nabla_\Theta \log P(y \mid x; \Theta) \\ + \mathcal{O}(\eta^2). \end{aligned} \tag{6}$$

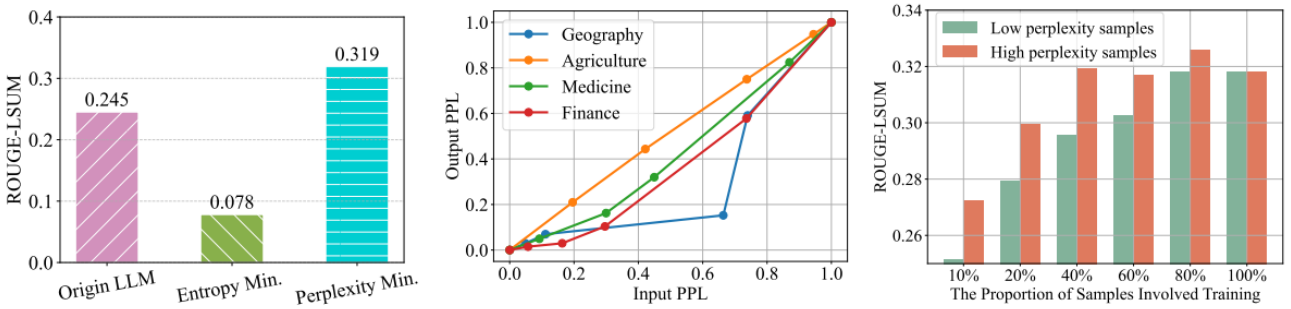
因此，当梯度内积项以非负为主时，最小化输入困惑度会推动输出困惑度下降。作者在 DomainBench 上抽取 400 个 QA batch 计算该内积，报告其中 98.75% 为非负，且均值约为 +5.60，从而为“输入困惑度最小化 ⇒ 更好输出”提供了经验支撑。此外，作者指出将视觉 TTA 的熵最小化目标直接迁移到 LLM 可能因忽略自回归依赖而损害性能 (Hu et al., 2025)。

样本效率：高困惑度样本优先的学习策略　TLM 的第二个观察是：不同测试样本对参数更新的有效性并不均衡，高困惑度样本对测试时适配的贡献更大；相反，持续在低困惑度样本上训练可能贡献有限，甚至产生负面影响（Observation 2, Fig.1(c)）(Hu et al., 2025)。据此，作者提出样本高效学习策略：为每个样本定义主动选择分数 $S(x)$，并使用该分数对输入困惑度最小化目标进行加权（式 (5)–(6)）：

$$\min_{\Theta} \; S(x)\, P(x; \Theta), \tag{7}$$

$$S(x) = \lambda \cdot e^{[\log P(x;\Theta) - \log P_0]} \cdot \mathbb{I}_{\{P(x;\Theta) > P_0\}}(x), \tag{8}$$

其中 $\mathbb{I}$ 为指示函数，P_0 为困惑度阈值，λ 为缩放系数。该设计确保仅当 $P(x; \Theta) > P_0$ 时，“高困惑度”样本才参与反向传播更新；低困惑度样本则被过滤，以节省优化预算并避免无效或潜在有害的更新。作者在实现中取 $\lambda = 0.10$、$P_0 = e^3$，并在消融实验中指出阈值



(a) Effectiveness of Entropy and Perplexity Minimization Strategies.

(b) Trends in perplexity to input and perplexity to output under Llama3.1-8B-Instruct.

(c) Effect of different test samples in test-time perplexity minimization.

Figure 1. TLM/TTL 的关键经验观察与证据：(a) 熵最小化与输入困惑度最小化策略的效果对比；(b) 输入困惑度与输出困惑度在不同领域上的趋势关系；(c) 不同困惑度样本参与测试时困惑度最小化对下游效果的影响。图源（引用自）(Hu et al., 2025)。

过高或过低均会导致性能下降：阈值过高会使可学习样本不足，阈值过低则会引入大量低信息量样本 (Hu et al., 2025)。

稳定性与遗忘：**LoRA** 的轻量更新　第三个关键设计是以 LoRA 替代全参数更新，以缓解测试时学习中的灾难性遗忘风险并降低计算开销。LoRA 将权重更新约束为低秩分解 $\Delta\Theta = BA$，固定原始参数，仅优化低秩增量（式 (7)）：

$$\min_{A,B} S(x)\, P(x;\Theta+\Delta\Theta), \quad \Delta\Theta = BA, \tag{9}$$

其中 A, B 为可训练的低秩矩阵。作者在初始化时采用 $A \sim \mathcal{N}(0,1)$、$B = 0$，并指出相较全参数更新，LoRA 更有助于保留原有知识，且更适用于资源受限的部署场景 (Kirkpatrick et al., 2017)。如 Table 2 所示，主文将测试时学习的可更新集合显式收缩为 LoRA 的低秩增量参数，使模型在不访问标签的测试阶段仍能以参数高效的方式执行梯度更新，并将其作为提升稳定性与可部署性的关键工程选择 (Hu et al., 2025; 2022)。

要点	主文/引用文献中的表述
更新形式	以低秩增量约束更新：$\Delta\Theta = BA$，并在测试时优化 $S(x)\,P(x;\Theta+\Delta\Theta)$（见式 (9)）(Hu et al., 2025).
可训练参数	冻结原参数，仅优化低秩矩阵 A, B（LoRA 参数高效更新机制）(Hu et al., 2025; 2022).
初始化	作者报告初始化采用 $A \sim \mathcal{N}(0,1)$、$B = 0$(Hu et al., 2025).
稳定性/遗忘动机	主文将 LoRA 作为默认测试时更新策略，用于提升适配稳定性并缓解在线更新可能引发的遗忘风险 (Hu et al., 2025)；灾难性遗忘作为连续学习中的经典问题可参见 EWC(Kirkpatrick et al., 2017).
部署含义	LoRA 作为参数高效更新方式，被主文用于降低测试时学习的计算与资源开销，并更贴近资源受限部署场景 (Hu et al., 2025; 2022).

Table 2. TTL/TLM 中 LoRA 轻量更新的关键要点（表内陈述均来自对应引用：(Hu et al., 2025; 2022; Kirkpatrick et al., 2017)）。

4.4. 实验协议与基准：**AdaptEval**

为系统评估 TTL，作者构建 AdaptEval 基准，并划分为三类场景：**DomainBench**（地理、农业、医学、金融等领域知识问答）、**InstructionBench**（多来源指令跟随数据：Alpaca-GPT4、Dolly-15k、InstructionWild）、**ReasoningBench**（GSM8K、MetaMath、LogiQA 等推理任务）(Hu et al., 2025)。为直观呈现该基准的层次化构成，Figure 2 给出了 AdaptEval 的三大子基准及其对应数据来源/任务集合示意：DomainBench 由多个专业领域知识问答子域组成，InstructionBench 汇集多来源指令跟随数据集，ReasoningBench 则覆盖多类推理任务。如 Table 3 所示，在解码策略上作者采用贪心解码并设温度 $T = 0$ 以获得稳定结果；优化器使用 Adam，batch size 设为 1；学习率在不同子基准上分别设为 5×10^{-5}（DomainBench/InstructionBench）与 1×10^{-6}（ReasoningBench）。

评价指标方面，DomainBench 与 InstructionBench 采用 ROUGE-Lsum (Lin, 2004)，ReasoningBench 采用 Exact Match（EM）(Chang et al., 2024)，从而在生成式输出质量与推理题精确匹配两类任务形态下形成一致的评测口径。

在对比方法方面，作者将若干代表性测试时自适应方法（如 Tent (Wang et al., 2021)、EATA (Niu et al., 2022)、COME (Zhang et al., 2025)）适配到离线 TTL 协议下进行公平比较，并在在线协议下额外报告更新频率与反向传播次数等部署相关的量化指标。因此，Table 3 不

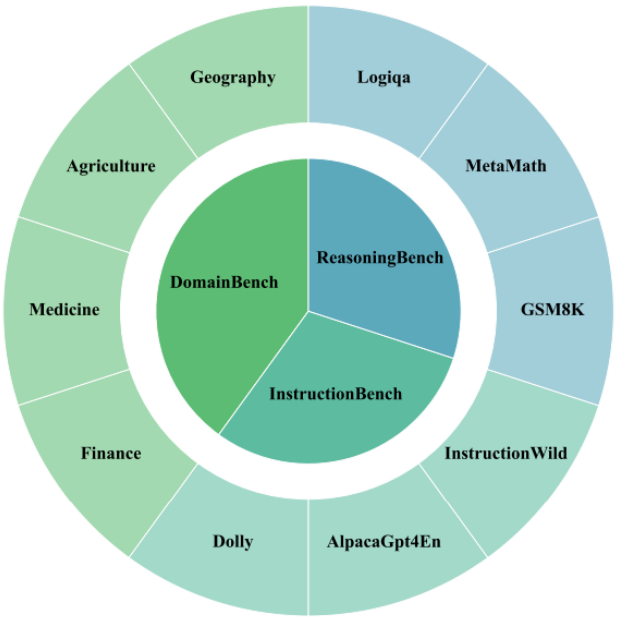


Figure 2. AdaptEval 基准的组成示意：由 DomainBench、InstructionBench 与 ReasoningBench 三类场景构成，并进一步细分为领域知识子域、指令跟随数据来源与推理任务集合。图源（引用自）(Hu et al., 2025)。

仅概括了 AdaptEval 的三类场景划分与指标选择，也汇总了用于复现实验协议的关键解码与优化设定，以及对比评测时离线/在线的主要报告要点 (Hu et al., 2025)。

为便于复现与理解，本文将 TLM/TTL 的离线与在线实现流程总结为伪代码（见 Algorithm 1）。其中，核心信号来自输入困惑度 $P(x;\Theta)$（式 (4)）。样本高效学习策略（SEL）通过式 (8) 对高困惑度样本进行选择或加权，并在 LoRA 参数子空间内执行测试时更新（式 (9)）。在更适于部署的在线协议下，算法采用“每 M 个样本更新一次”的策略（主文示例为 M=100）以降低反向传播频率，从而在性能与开销之间取得权衡 (Hu et al., 2025)。

4.5. 实证结果与消融分析

总体而言，TLM 在 DomainBench 的领域知识适配上相较原始 LLM 取得显著增益。作者在摘要与贡献点总结中进一步指出，其在领域知识适配上“至少提升 20%”。为更直观呈现关键消融结果与部署代价，本文将原文表格中的数值重绘为图：Figure 3 汇总了 DomainBench 上样本高效学习策略（SEL）的消融结果（对应原文 Table 4），Figure 4 则刻画了在线设定下“效果（ROUGE-Lsum）—反向传播次数”的权衡关系（对应原文 Table 5）(Hu et al., 2025)。

从重绘结果可见，Figure 3 主要体现两点：其一，相较不进行测试时更新的 LLM 基线，测试时学习本身即可

Algorithm 1 TLM/TTL：输入困惑度最小化的测试时学习（SEL + LoRA），含在线更新协议

Require: Base LLM parameters Θ (frozen); LoRA parameters (A, B); test stream $\{x_t\}_{t=1}^{T}$; threshold P_0, scale λ (Eq.(8)); learning rate η, update steps K; update interval M (online: M=100, offline: M=1).
Ensure: Predictions $\{y_t\}_{t=1}^{T}$ and updated LoRA parameters (A, B).

Initialize LoRA as in (Hu et al., 2025): $A \sim \mathcal{N}(0,1)$, $B = 0$.
Initialize buffer $\mathcal{B} \leftarrow \emptyset$.
for $t = 1$ to T **do**
  Compute input perplexity $P_t \leftarrow P(x_t; \Theta, A, B)$ via Eq.(4).
  Compute SEL score $s_t \leftarrow \lambda \exp(\log P_t - \log P_0) \cdot \mathbb{I}[P_t > P_0]$.
  Buffer the sample: $\mathcal{B} \leftarrow \mathcal{B} \cup \{(x_t, s_t)\}$.
  if $|\mathcal{B}| = M$ **then**
    Update every M samples (online: M=100; offline: M=1).
    $\mathcal{L} \leftarrow \sum_{(x,s)\in\mathcal{B}} s \cdot \log P(x; \Theta, A, B)$.
    Weighted objective (Eq.(7)) in log-space; minimizing $\mathcal{L}$ reduces perplexity.
    for $k = 1$ to K **do**
      Update only LoRA params (A, B) with Adam: $(A, B) \leftarrow (A, B) - \eta \nabla_{A,B} \mathcal{L}$.
    end for
    Clear buffer: $\mathcal{B} \leftarrow \emptyset$.
  end if
  Generate answer: $y_t \leftarrow \textsc{Generate}(x_t; \Theta, A, B)$.
  Greedy decoding with temperature 0 (Hu et al., 2025).
end for

在四个领域带来明显提升（w/o SEL 相对 LLM 的增益最为显著），说明“输入困惑度最小化 + LoRA 更新”确实提供了有效的无标签适配信号。其二，在此基础上加入 SEL（即式 (8) 的高困惑度样本筛选/加权）还能在多个子域上进一步提升 ROUGE-Lsum（w/ SEL 相对 w/o SEL 的增益虽更小但更稳定），与主文 Observation 2 “高困惑度样本对更新贡献更大”的经验结论一致。结合 Algorithm 1 的流程，SEL 的作用可理解为：在固定测试时预算下减少低信息样本的无效反向传播，将梯度预算集中于更“域不匹配”的样本，从而提高单位反传的收益。

进一步地，Figure 4 将原文 Table 5 的在线结果转化为“质量—反传成本”的坐标系，突出部署层面的关键权衡：（i）LLM 基线无需反向传播（#Backward= 0），但在线场景下平均 ROUGE-Lsum 受限；（ii）若直接将部分视觉侧/判别侧的在线自适应目标迁移到 LLM（如 Tent/EATA 的在线版本），往往需要大量反向传播，但在该基准上的质量增益并不稳定，体现了目标函数与生成式建模机制错配可能带来的风险 (Wang et al., 2021;

子基准	场景/数据构成	指标（**Metric**）	解码（**Decoding**）	优化设定（**Optimization**）
DomainBench	领域知识问答（地理、农业、医学、金融等）.	ROUGE-Lsum(Lin, 2004).	贪心解码，$T = 0$.	Adam，batch size=1，lr=5×10^{-5}.
InstructionBench	多来源指令跟随：Alpaca-GPT4、Dolly-15k等.	ROUGE-Lsum(Lin, 2004).	贪心解码，$T = 0$.	Adam，batch size=1，lr=5×10^{-5}.
ReasoningBench	推理任务：GSM8K、MetaMath、LogiQA 等.	Exact Match (EM)(Chang et al., 2024).	贪心解码，$T = 0$.	Adam，batch size=1，lr=1×10^{-6}.
对比与报告口径	将 Tent(Wang et al., 2021)、EATA(Niu et al., 2022)、COME(Zhang et al., 2025) 等方法适配到离线 TTL 协议进行比较；在线协议额外报告更新频率与反向传播次数等部署指标.	（同上三类子基准）	（同上）	（同上；并区分离线/在线报告）

Table 3. AdaptEval 的场景划分与实验协议要点汇总（表内条目均来自引用文献：(Hu et al., 2025)；指标引用：ROUGE-Lsum(Lin, 2004)，EM(Chang et al., 2024)；对比方法引用：Tent(Wang et al., 2021)、EATA(Niu et al., 2022)、COME(Zhang et al., 2025)）。

Niu et al., 2022)；（iii）相较之下，TLM 以输入困惑度作为自监督信号，并采用低频更新协议（主文示例：每 100 个样本更新一次），在显著降低反向传播频率的同时取得更高的平均 ROUGE-Lsum，从而呈现出更具部署友好性的“性能—成本”折中 (Hu et al., 2025)。

在更细粒度分析上：

- 核心增益（输入困惑度最小化）：作者在 Fig.1(b) 报告输入困惑度与输出困惑度在趋势上高度一致，从而解释了为何以输入困惑度作为自监督目标能够带来输出质量提升。
- 关键组件消融（样本高效策略）：基于式 (8) 的高困惑度样本选择/加权在不引入额外标注的前提下提升更新效率。原文在 DomainBench 上给出 SEL 的消融对比（Table 4）：完整方法（TLM，w/ SEL）相对去除 SEL（w/o SEL）在多个子域上进一步提升 ROUGE-Lsum（见 Figure 3），从而支持“高困惑度样本更具信息量”的经验动机。
- 超参敏感性（阈值与更新频率）：阈值 P_0（或其等价设定）过高会导致可用于更新的样本不足，适配不充分；过低则会引入大量低信息样本，既浪费预算也可能造成性能下降。在线场景下，作者采用“每 100 个样本更新一次”的协议以减少反向传播次数并降低延迟，并在在线比较中同时报告性能与反传开销（Table 5）。Figure 4 将该表格结果重绘为“质量—反传成本”关系，用于更直观呈现部署友好性的权衡。
- 稳定性与遗忘（**LoRA**）：作者将 LoRA 作为默认更新策略，并据此论证其在测试时学习中对缓解灾难性遗忘与降低训练开销的有效性 (Hu et al., 2022)。

4.6. 局限性与适用边界

尽管 TLM 在多个 AdaptEval 子场景上取得了稳定收益，其适用性仍受多方面因素制约。首先，TTL 需要在推理链路中引入梯度更新，从而带来额外计算开销与工程复杂度。为降低反向传播成本，作者采用样本选择策略与在线低频更新（每 100 个样本更新一次），但测试时训练开销仍难以完全消除。其次，方法对阈值 P_0、学习率等超参数较为敏感：若高困惑度样本过少或被错误筛除，适配可能不足；若大量低信息样本参与更新，则可能出现无效更新，甚至导致性能退化（Observation 2）。最后，尽管 LoRA 有助于缓解遗忘风险，其本质仍属于参数更新范式。面对长序列在线数据流或跨域频繁切换等场景，仍需进一步研究更强的稳定性约束与更精细的更新调度策略。

5. 比较式分析框架

为将视觉侧的测试时自适应（TTA/TTT）与语言模型侧的测试时学习（TTL/TTT）置于同一坐标系下，本文从四个维度组织比较：测试时目标函数（Objective）、更新参数子集（What to Update）、优化过程与数据组织（How to Optimize）、任务与分布偏移类型（Where it Works）。统一记号上，设训练后模型参数为 θ_0，测试阶段可访问的数据为 $\mathcal{D}_{\text{test}}$（可能无标签，也可能包含 few-shot 演示），则测试时学习可抽象为

$$\theta^\star \in \arg\min_{\theta\in\Theta(\theta_0)} \mathcal{L}_{\text{adapt}}(\theta;\mathcal{D}_{\text{test}}), \qquad \hat{y} = f_{\theta^\star}(x), \quad (10)$$

其中 $\mathcal{L}_{\text{adapt}}$ 由测试时可计算的信号构造，$\Theta(\theta_0)$ 则刻画“允许更新哪些参数/模块”的部署约束。该抽象覆盖了以自监督/无监督信号驱动测试期参数更新的一系列代表性路线：视觉侧的 TTT (Sun et al., 2020)、Tent (Wang et al., 2021) 与 MAE-TTT (Gandelsman et al., 2022)，以及语言侧面向 LLM 的 TTL/TLM (Hu et al., 2025) 与 few-shot 推理场景下的 TTT (Akyürek et al., 2025)。

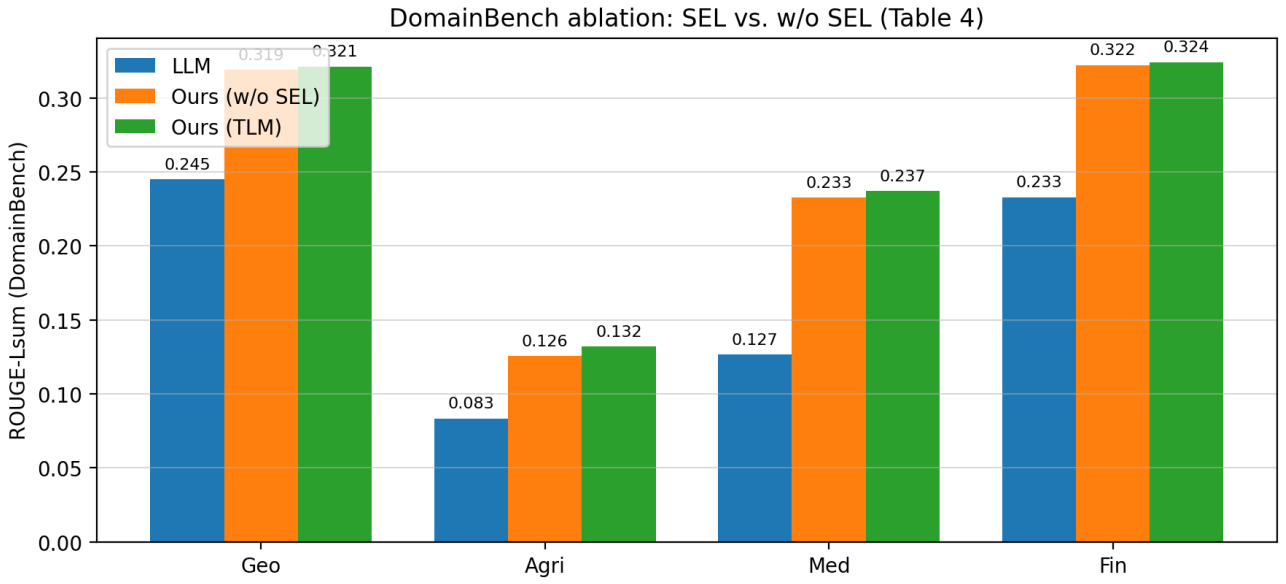


Figure 3. DomainBench 消融：SEL vs. w/o SEL（数据重绘自原文 Table 4）。图中给出了 LLM 基线、去除样本高效学习策略（w/o SEL）以及完整方法 TLM（w/ SEL）在地理/农业/医学/金融四个子域上的 ROUGE-Lsum 表现 (Hu et al., 2025)。

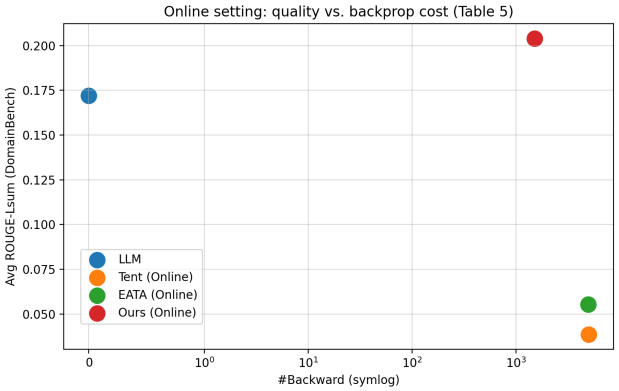


Figure 4. 在线设定下的效果–反向传播成本权衡（数据重绘自原文 Table 5）。横轴为反向传播次数（#Backward，按原文统计口径），纵轴为 DomainBench 平均 ROUGE-Lsum；该图用于直观展示不同在线测试时自适应方法在“质量提升”与“反传开销”之间的权衡 (Hu et al., 2025)。

为便于在同一坐标系下快速定位各方法的关键选择，Table 4 将上述代表性工作按四个维度进行统一对照。在 *Objective* 维度上，TTT 以自监督辅助任务损失驱动测试时更新 (Sun et al., 2020)，Tent 以预测熵最小化实现 fully test-time adaptation (Wang et al., 2021)，MAE-TTT 以 MAE 掩码重建增强测试时自监督信号 (Gandelsman et al., 2022)，TTL/TLM 将 LLM 测试时学习形式化为输入困惑度最小化 (Hu et al., 2025)，而 few-shot TTT 则基于 in-context 示例构造训练信号，并在推理时进行临时参数更新 (Akyürek et al., 2025)。在 *What/How* 维度上，各方法通过限制更新子集（如归一化仿射参数或 LoRA 低秩增量），并组织更新粒度（逐输入、逐批次，或降低在线更新频率）来权衡部署成本与稳定性 (Wang et al., 2021; Hu et al., 2025; 2022)。在 *Where* 维度上，各方法分别面向视觉腐蚀/偏移鲁棒性 (Hendrycks & Dietterich, 2019)、视觉对象识别分布偏移 (Gandelsman et al., 2022)，以及 LLM 的领域知识、指令跟随、推理偏移与结构性新颖任务（如 ARC、BBH）(Hu et al., 2025; Akyürek et al., 2025; Chollet, 2019; Suzgun et al., 2022)。后续小节将以该表为索引，分别从四个维度展开更细粒度的比较与讨论。

5.1. 维度一：测试时目标函数（Objective）

熵最小化（Tent）. Tent 在 fully test-time adaptation 设定下仅使用无标签测试数据，提出以预测分布熵作为优化目标（test entropy minimization），并据此驱动测试时适配 (Wang et al., 2021)。从目标类型上看，该策略与半监督学习中的最小熵正则化一脉相承：利用未标注数据促使决策函数在数据流形附近产生更确定的预测 (Grandvalet & Bengio, 2004; Wang et al., 2021)。

自监督任务（TTT/MAE-TTT）. TTT 将单个无标签测试样本转化为自监督学习问题，并在输出预测之前对模型参数进行更新；该过程也可扩展至在线数据流 (Sun et al., 2020)。MAE-TTT 沿用“对每个测试输入进行自监督优化”的测试时训练范式，但将自监督信号替换为 masked autoencoder 的重建目标，并将其明确表述为 one-sample learning 问题，同时从偏差—方差权衡

方法	测试时目标（**Objective**）	更新子集（**What to Update**）	优化组织（**How to Optimize**）	任务/偏移（**Where it Works**）
TTT (ICML'20)	将无标签测试样本构造成自监督辅助任务并在预测前优化该损失；原文采用旋转预测作为自监督信号 (Sun et al., 2020; Gidaris et al., 2018).	主要更新共享特征提取器（与主任务头共享的表示部分）；训练期以主任务 + 自监督多任务联合优化以保证测试期自监督更新与主任务兼容 (Sun et al., 2020).	可对单样本执行短步更新；同一样本可通过增广形成"小批"以稳定自监督更新；可扩展到在线流式累计更新 (Sun et al., 2020).	视觉分类分布偏移；在腐蚀/偏移基准上提升鲁棒泛化 (Sun et al., 2020; Hendrycks & Dietterich, 2019).
Tent (ICLR'21)	完全测试时自适应：在仅有无标签测试数据时最小化预测分布熵（test entropy minimization）(Wang et al., 2021).	在线估计归一化统计量，并仅优化通道仿射参数（feature modulation）以低维稳定更新 (Wang et al., 2021).	按测试批次在线优化；强调 batch 级约束避免单样本熵最小化退化 (Wang et al., 2021).	ImageNet-C/CIFAR-C 等鲁棒性与 source-free 适配任务；显著降低错误率 (Wang et al., 2021; Hendrycks & Dietterich, 2019).
MAE-TTT (NeurIPS'22)	沿用"每个测试输入自监督优化"，但以 MAE 掩码重建替代旋转预测以增强自监督信号 (Gandelsman et al., 2022; He et al., 2022).	对测试输入做重建目标的短步优化（one-sample learning）；共享编码器 + 自监督解码器 + 下游头的结构化复用 (Gandelsman et al., 2022).	逐输入（one-sample）测试时训练；报告并分析其在简化模型下与偏差–方差权衡的关系 (Gandelsman et al., 2022).	多类视觉分布偏移对象识别/分类基准上提升泛化 (Gandelsman et al., 2022).
TTL/TLM (ICML'25)	将 LLM 测试时学习形式化为输入困惑度最小化；并给出一阶近似与梯度内积统计支持"输入困惑度下降促进输出困惑度下降" (Hu et al., 2025).	采用 LoRA 进行参数高效更新，以低秩增量约束更新子空间并缓解遗忘/降低开销 (Hu et al., 2025; 2022; Kirkpatrick et al., 2017).	样本高效策略：优先/加权高困惑度样本；在线协议示例为"每 100 个样本更新一次"以减少反传次数 (Hu et al., 2025).	LLM 的领域知识、指令跟随、推理任务分布偏移；AdaptEval 基准上报告显著收益 (Hu et al., 2025).
Few-shot TTT (ICML'25)	在 few-shot 推理中用 in-context 示例构造训练信号并临时更新参数，超越仅 prompting 的 few-shot 范式 (Akyürek et al., 2025).	推理时临时参数更新（强调 per-instance training 等组件），用于提升结构性泛化 (Akyürek et al., 2025).	关键组件包括 leave-one-out 构造与优化设置；在 ARC/BBH 等任务上报告显著提升 (Akyürek et al., 2025; Chollet, 2019; Suzgun et al., 2022).	结构性新颖任务（ARC/BBH）上的 few-shot 学习与抽象推理提升 (Akyürek et al., 2025).

Table 4. 测试时更新方法的统一比较矩阵（表内陈述均来自对应引用：(Sun et al., 2020; Wang et al., 2021; Gandelsman et al., 2022; Hu et al., 2025; Akyürek et al., 2025) 等）。

角度讨论改进来源 (Gandelsman et al., 2022)。因此，两者的共同点在于：测试时目标函数并不直接依赖任务标签，而是由输入结构诱导的自监督约束所提供 (Sun et al., 2020; Gandelsman et al., 2022)。

输入困惑度最小化（主文 **TLM/TTL**）. 主文提出面向 LLM 的 TTL（TLM），指出在无标签测试数据下可通过最小化输入困惑度实现自监督增强，并据此将 LLM 的测试时学习形式化为 input perplexity minimization (Hu et al., 2025)。与 Tent 的输出分布熵不同，这里使用自回归语言建模的序列似然信号（以困惑度表征）。该信号可直接在无标签输入上计算并反向传播，从而为 LLM 在目标域上的动态适配提供优化驱动。

few-shot 场景损失构造（**ICML 2025 few-shot TTT**）. 在 few-shot 推理语境中，TTT 的测试时损失来自输入数据（尤其是 in-context 演示）所诱导的训练信号。该工作将推理时的临时参数更新作为关键机制，并提出以 leave-one-out 的 in-context 任务构造、优化设置与 per-instance training 等组件形成更稳健的框架 (Akyürek et al., 2025)。因此，该路线对应于在测试时可获得弱监督结构（演示对）时，利用更任务对齐的损失进行短程更新。

5.2. 维度二：更新参数子集与可部署性（**What to Update**)

Tent：归一化统计与通道仿射参数. Tent 报告其方法通过在线估计归一化统计量并优化通道级仿射变换参数，从而支持对每个测试 batch 的在线更新 (Wang et al., 2021)。这种" 低自由度更新集" 将 $\Theta(\theta_0)$ 约束在少量参数上，旨在以较低的测试时开销获得更稳定的适配收益。

TTT/MAE-TTT：对测试输入执行自监督更新的模型参数. TTT 的核心流程是在预测前更新模型参数，并可在在线数据流中持续执行 (Sun et al., 2020)；MAE-TTT 同样强调"为每个测试输入优化模型"，只是其自监督信号由 masked autoencoding 的重建目标提供 (Gandelsman et al., 2022)。与 Tent 相比，这类方法通常允许更大的可更新空间，潜在适配能力更强；但相应地，测试时反向传播成本与稳定性控制成为部署中的关键约束 (Sun et al., 2020; Gandelsman et al., 2022)。

主文 **TLM**：以 **LoRA** 约束更新子空间. 主文明确采用 LoRA 替代全参数优化，以缓解灾难性遗忘并提升适配稳定性，同时降低测试时更新成本 (Hu et al., 2025; 2022)。LoRA 的机制是在冻结预训练权重的前提下引

入低秩可训练增量，从而显著减少需要更新的参数量；其动机亦源于" 大模型全量微调在部署与存储上不可行" 的工程现实。因此，TLM 在 What-to-Update 维度上的核心增量在于：将 TTL 的可更新集合显式收缩至参数高效子空间，使测试时学习在 LLM 场景中更具可部署性。

few-shot TTT：临时参数更新**.** few-shot TTT 强调临时更新：在推理时利用少量演示构造损失并更新参数，以提升抽象推理与 few-shot 学习能力 (Akyürek et al., 2025)。从部署角度看，这种" 按实例/按任务临时适配" 的思路同样需要可控的更新子集，以避免推理成本失控，并便于在多任务之间快速切换。

5.3. 维度三：优化过程与数据组织（How to Optimize)

逐批次在线更新（**Tent**）**.** Tent 明确采用"online on each batch" 的方式在测试阶段更新，并指出其收益可在一次测试时优化轮次（one epoch of test-time optimization）内获得 (Wang et al., 2021)。这对应于以 batch 为基本单位组织 $\mathcal{D}_{\text{test}}$，并将更新持续累积到后续测试数据的 online 过程。

逐输入更新与在线流式扩展（**TTT/MAE-TTT**）**.** TTT 的核心机制是将单个无标签测试样本转化为自监督问题，并在预测前执行少步更新；该机制可自然扩展到 online stream (Sun et al., 2020)。MAE-TTT 也采用“对每个测试输入进行自监督优化” 的组织方式，并将其表述为 one-sample learning (Gandelsman et al., 2022)。这两类方法都将测试数据组织为“以样本为中心” 的适配单元，其效率与稳定性高度依赖测试时训练预算与自监督信号质量 (Sun et al., 2020; Gandelsman et al., 2022)。

样本选择/加权与预算集中（主文 **TLM**）**.** 主文报告一个关键经验观察：高困惑度样本往往对优化更具信息量；据此提出 Sample Efficient Learning Strategy，在测试时主动选择并强调高困惑度样本以提升更新效率 (Hu et al., 2025)。在 How-to-Optimize 维度上，这等价于在固定测试时预算下对 $\mathcal{D}_{\text{test}}$ 进行非均匀采样或加权，使梯度更新更聚焦于" 域不匹配更强" 的输入区域。

few-shot：**leave-one-out** 任务构造与 **per-instance training.** few-shot TTT 总结其有效性依赖若干关键组件，包括 leave-one-out 的 in-context 任务构造、优化设置与 per-instance training，并将其作为提升 few-shot 学习与抽象推理能力的核心机制 (Akyürek et al., 2025)。因此，该方向的数据组织更接近" 将上下文演示重排为多个训练子任务"，以放大极少样本条件下可用的训练信号。

5.4. 维度四：任务与分布偏移类型（Where it Works)

视觉分布偏移（**TTT/Tent/MAE-TTT**）**.** TTT 面向训练与测试分布不一致的视觉任务，强调在分布偏移下通过测试时自监督训练提升泛化，并支持在线流式扩展 (Sun et al., 2020)。Tent 报告其在 corrupted ImageNet 与 CIFAR-10/100 等鲁棒性基准，以及若干 source-free 适配任务上的改进 (Wang et al., 2021)。MAE-TTT 报告其在多类视觉分布偏移基准上的泛化提升，并将该提升在理论上联系到偏差—方差权衡 (Gandelsman et al., 2022)。总体而言，这一谱系主要处理由感知输入分布变化驱动的偏移 (Sun et al., 2020; Wang et al., 2021; Gandelsman et al., 2022)。

语言域知识/语言变体偏移（主文 **TLM**）**.** 主文将分布偏移具体化为 LLM 在专业领域与多样语言变体上的泛化不足，并提出仅依赖无标签测试数据进行动态适配的 TTL 范式。作者引入 AdaptEval 基准，并报告在领域知识适配上相对原始 LLM 至少提升 20% (Hu et al., 2025)。因此，TLM 的有效区域更偏向于语言输入分布（术语、风格与领域知识表达方式）变化所导致的不匹配。

few-shot 结构性任务（**ICML 2025 few-shot TTT**）**.** few-shot TTT 聚焦于 out-of-distribution 的推理与结构性任务（如 ARC、BBH），并将" 推理时继续训练" 作为提升模型抽象推理与 few-shot 学习能力的关键机制 (Akyürek et al., 2025)。这类任务的偏移更体现为结构组合与规则泛化，而非单纯的风格变化；因此更依赖由演示驱动的损失构造来注入任务约束。

5.5. 小结：主文的增量与代价

在统一框架下，主文的主要增量体现在两个方面。其一，针对 LLM 的生成式建模特性，选择输入困惑度最小化作为测试时自监督目标，并给出经验证据与理论洞见支撑这一选择。其二，通过高困惑度样本优先策略集中预算，并以 LoRA 约束更新子空间，以缓解遗忘并提升适配稳定性 (Hu et al., 2025; 2022)。相应的代价在于：与纯前向推理相比，TLM 与 few-shot TTT 均需要在部署阶段引入梯度更新，从而带来额外计算开销以及对超参数与策略的敏感性。Tent 虽更新参数更少、工程实现更轻量，但其目标函数与模型输出分布直接耦合，适用性与收益在很大程度上取决于模型在偏移域上的初始置信度结构 (Wang et al., 2021; Hu et al., 2025; Akyürek et al., 2025)。从方法学角度看，这四条路线共同揭示：测试时学习的关键并非“是否更新”，而在于目标函数是否与模型生成机制匹配、更新子空间是否可控且可部署，以及数据组织是否能在有限预算下提供高质量梯度信号 (Sun et al., 2020; Wang et al., 2021; Gandelsman et al., 2022; Hu et al., 2025; Akyürek et al., 2025)。

6. 课程视角下的心得与反思

6.1. 从训练期优化到测试期优化：范式迁移

经典经验风险最小化（ERM）隐含训练与测试同分布（或近似同分布）的假设；但在真实部署场景中，训练分布与目标测试分布之间的偏移（dataset shift）更为常

见，这会系统性破坏“训练期一次性优化 → 测试期静态推断”的工作流 (Quinonero-Candela et al., 2008)。从理论视角看，领域自适应的一类代表性结果将目标域风险上界分解为源域风险、分布差异项与不可约误差项，从而揭示：当分布差异增大时，仅靠训练期拟合并不能保证目标域泛化 (Ben-David et al., 2010)。因此，课程中“优化算法—泛化误差—分布假设”三者之间的逻辑链条，在部署情境下自然导向一个问题：能否在测试阶段利用可获得的无标签数据，进行预算受限的继续优化，从而缓解分布偏移带来的性能坍塌?

TTT、Tent 与 MAE-TTT 等工作以不同方式给出了肯定回答：它们将测试时可观测数据视为在线到达的样本流，并通过少步梯度更新在推断前改变模型状态，以获得对目标域的适配能力 (Sun et al., 2020; Wang et al., 2021; Gandelsman et al., 2022)。进一步地，主文（ICML 2025）将这一范式扩展到大语言模型，并在测试阶段以“输入困惑度最小化”为自监督目标进行测试时学习（test-time learning, TTL），将适配更明确地表述为部署期的参数更新过程 (Hu et al., 2025)。从课程的优化视角看，上述方法可抽象为一类预算受限的在线优化（online optimization）：在时间步 t 接收测试样本（或批）$\mathcal{B}_t$，对某个无标签目标 $\mathcal{L}_{\text{test}}(\theta;\mathcal{B}_t)$ 做少步更新

$$\theta_{t+1} \leftarrow \theta_t - \eta_t \nabla_\theta \mathcal{L}_{\text{test}}(\theta_t;\mathcal{B}_t), \tag{11}$$

并以“单位样本/单位时间的增益”来约束步数预算、停止准则与更新频率。该抽象与在线凸优化（OCO）中“迭代决策—事后反馈—遗憾界（regret）”的基本框架在形式上高度一致，为理解测试时学习的收敛—稳定—成本权衡提供了成熟语言。当然，深度模型的非凸性以及无标签目标可能存在的错配风险，使其更接近经验驱动的在线过程，而非可直接套用的凸优化理论 (Shalev-Shwartz, 2011; Hazan, 2015)。

6.2. 生成式建模视角：困惑度作为无标签适配信号

主文将大语言模型的测试时学习建立在生成式建模的自监督信号之上。对输入序列 $x = (x_1, \ldots, x_T)$，其输入困惑度（perplexity）被定义为平均负对数似然的指数形式

$$\text{ppl}(x;\theta) = \exp\left(\frac{1}{T}\sum_{t=1}^{T} -\log p_\theta(x_t \mid x_{<t})\right), \tag{12}$$

并在测试阶段通过最小化输入困惑度更新参数，从而实现对目标测试分布的自适应 (Hu et al., 2025)。从生成式建模的直觉出发，该目标直接鼓励模型提高对” 当前测试域文本分布” 的似然拟合。因此，困惑度可被理解为一种可计算的” 域不匹配程度” 代理指标：当模型在某些样本上的困惑度显著更高时，往往意味着该样本更偏离训练分布或超出当前模型的语言知识覆盖范围，也更可能为适配提供有效的梯度信号。

然而，将困惑度作为无标签适配信号也天然伴随目标错配（*objective mismatch*）风险：困惑度衡量的是生成式对数似然，而下游任务损失可能是判别式、结构化或推理型指标；因此，“更善于生成/更善于解释输入”并不必然等价于“更善于完成任务”。主文通过理论近似与消融实验表明，困惑度更新对下游改进需要一定条件（例如更新方向与任务目标之间的对齐性），且整体收益对更新频率、样本选择与更新规模等策略较为敏感 (Hu et al., 2025)。从课程角度看，这提示在设计测试时目标时应显式讨论：(*i*) 无标签代理损失与任务损失之间的关联假设；(*ii*) 当关联不成立时的失败模式与检测信号；(*iii*) 在预算受限下如何通过策略性数据组织提高“有效梯度”的比例。

6.3. 参数高效方法的稳定性含义

在部署约束下，测试时学习不仅追求有效性，还必须强调可部署性，包括更新开销、显存/带宽预算以及对模型行为的可控性。主文选择将 LoRA 作为测试时可训练参数子集：通过在若干权重矩阵上引入低秩增量（$\Delta W \approx BA$）近似全参数更新，从而显著减少可训练参数量与优化噪声，并降低测试时更新的资源成本 (Hu et al., 2022; 2025)。与训练期参数高效微调类似，LoRA 的核心价值在于：以受限自由度的结构化更新换取更高的工程可行性与稳定性，并允许将适配增量与基座权重解耦管理（例如按场景加载/卸载）(Hu et al., 2022)。主文进一步将这种受限更新用于测试阶段，以缓解在线更新可能带来的不稳定与性能回退。

从更一般的稳定性—可塑性（stability–plasticity）视角看，测试时学习与持续学习/连续学习共享同一核心张力：模型需要对新分布迅速塑形，又要避免破坏既有能力（可理解为遗忘或漂移）。EWC 等持续学习方法通过对重要参数施加约束来缓解灾难性遗忘，凸显” 受限更新” 在稳定性上的意义 (Kirkpatrick et al., 2017)。尽管 LoRA 与 EWC 的机制不同，但在课程框架下，两者可统一理解为通过降低有效参数自由度或限制更新空间提高在线更新的可控性。这既解释了主文为何倾向使用 LoRA 而非全参数更新，也提示了其潜在限制：当领域偏移较大或任务需要更深层的能力重构时，过强的参数约束可能导致适配容量不足 (Hu et al., 2022)。

6.4. 开放问题与未来方向

第一，非平稳测试分布与概念漂移（**concept drift**）。部署环境中的测试分布往往随时间演化，可能呈现渐变、突变或周期性漂移；概念漂移综述系统讨论了漂移类型与在线适应的评估挑战 (Gama et al., 2014)。主文提出的在线协议与样本优先策略为持续适配提供了起点，但如何在非平稳情形下设计漂移检测、更新触发与遗忘控制，仍需要更系统的机制与基准 (Quinonero-Candela et al., 2008)。

第二，更新策略的安全鲁棒性。测试时自适应引入了” 用测试数据反向影响模型参数” 的通道，因此可能暴露新的攻击面或数据投毒风险。已有研究表明，即便不改变训练阶段，测试时自适应过程本身也可能带来可被利用的对抗风险，这提示需要将安全性纳入测试时

学习的标准评估维度 (Wu et al., 2023)。对主文范式而言，一个直接问题是：当高困惑度样本被优先用于更新时，是否会放大异常样本对参数轨迹的影响；相应地，如何在保持样本效率的同时引入鲁棒过滤或保守更新机制，值得进一步研究。

第三，性能—成本曲线与可复现报告规范。测试时学习的单位样本增益应与推断延迟、反向传播频率、显存占用等部署指标一并呈现。在线优化文献强调在资源约束下分析算法过程与反馈结构的重要性 (Hazan, 2015; Shalev-Shwartz, 2011)；主文也通过降低反向传播频率与采用 LoRA 平衡收益与成本。未来更理想的报告方式是给出明确的性能—成本曲线（而非单点最优结果），并在不同漂移强度与不同预算约束下进行统一对比。

第四，与 **in-context learning** 的关系。大模型在测试阶段既可通过参数更新（TTL/TTT 类）实现适应，也可通过上下文条件化（in-context learning, ICL）在不更新参数的情况下完成任务迁移 (Brown et al., 2020)。主文与 few-shot 场景的测试时训练工作共同提示：参数更新与上下文学习可能在不同任务类型与偏移形态下互补，但二者的统一理论刻画、协同策略与安全边界仍有待建立 (Hu et al., 2025; Akyürek et al., 2025)。

7. 结论

本文围绕 ICML 2025 的 *Test-Time Learning for Large Language Models*，对" 大语言模型在分布偏移下如何在无标签测试数据条件下实现可部署的动态适配" 这一问题进行了精读与比较式综述 (Hu et al., 2025)。与将部署推理视为静态映射不同，该工作将测试阶段自适应形式化为一个预算受限的优化过程：在测试阶段仅依赖输入数据构造自监督信号，通过少步梯度更新改变模型状态，从而缓解目标域分布与训练分布不一致导致的性能退化。

在方法层面，主文的核心贡献可概括为三点。第一，提出面向 LLM 的 TTL（TLM）范式，并将测试时学习目标具体化为输入困惑度最小化：利用自回归语言建模的序列似然信号在无标签输入上构造可微目标，为测试时更新提供直接驱动；同时给出理论近似与经验证据（梯度内积统计）支持" 输入困惑度下降可促进输出困惑度下降" 的关键假设。第二，提出高困惑度样本优先的样本高效学习策略，通过对测试流进行筛选或加权，将有限的反向传播预算集中到信息量更高的样本上，以提升适配效率并减少无效更新。第三，为增强稳定性与可部署性，采用 LoRA 将测试时可更新集合约束为低秩增量子空间，以减少更新参数量并缓解灾难性遗忘风险，使测试时学习更贴近在线部署需求 (Hu et al., 2025; 2022)。在评测方面，主文构建 AdaptEval 基准，系统覆盖领域知识、指令跟随与推理类场景，并报告其在领域知识适配上相对原始 LLM 至少提升 20% 的总体收益，从而为 TTL 的可行性提供了较完整的实证支撑。

在比较式分析框架下，本文将主文与四类代表性顶会工作进行对照。视觉侧的 TTT（ICML 2020）通过自监督预任务在测试前更新表征，并可扩展到在线数据流 (Sun et al., 2020)；Tent（ICLR 2021）在 fully test-time adaptation 设定下以预测熵最小化驱动在线更新，并将更新限制在归一化相关参数以提升稳定性 (Wang et al., 2021)；MAE-TTT（NeurIPS 2022）以 masked autoencoding 的重建目标强化单样本测试时训练信号，并从偏差—方差视角解释其增益来源 (Gandelsman et al., 2022)；LLM few-shot 场景的 TTT（ICML 2025）则表明，在结构性推理任务上，通过基于 in-context 演示构造损失并进行临时参数更新，可显著提升 few-shot 学习与推理能力 (Akyürek et al., 2025)。上述对比揭示：测试时自适应的成败高度依赖于（i）测试时目标函数是否与模型生成/判别机制匹配，（ii）更新子空间是否受控且具备部署可行性，以及（iii）数据组织能否在有限预算下提供高质量梯度信号 (Hu et al., 2025; Wang et al., 2021; Sun et al., 2020; Gandelsman et al., 2022; Akyürek et al., 2025)。

面向未来，本文认为优先推进的研究方向包括：**(1)** 非平稳测试分布下的稳定在线适配，即在概念漂移与域切换情形下的更新触发、遗忘控制与评测协议统一；**(2)** 测试时更新的安全与鲁棒性，在存在异常样本、对抗扰动或投毒风险时避免适配过程成为新的攻击面；**(3)** 性能—成本曲线的标准化报告，将准确率/质量增益与反向传播频率、延迟、显存占用等部署指标共同纳入评价；以及 **(4)** 与 **in-context learning** 的关系刻画与协同机制，系统理解" 上下文条件化" 与" 参数更新" 两种适配路径在不同任务偏移形态下的互补与边界 (Hu et al., 2025; Akyürek et al., 2025)。我最终得到的" 可复用原则" 是三条：目标函数要尊重生成机制（避免错配），更新子空间要可控（避免漂移/遗忘），数据组织要提高梯度信噪比（预算下的有效学习）；附录的 Quantile-SEL 则是我沿第三条原则做的最小改动验证。

致谢

本结课报告在尚凡华老师的指导下完成。尚老师在课程内容、方案设计与实验分析等方面给予了重要的帮助和宝贵的建议，在此谨致以衷心的感谢。

References

Akyürek, E., Damani, M., Zweiger, A., Qiu, L., Guo, H., Pari, J., Kim, Y., and Andreas, J. The surprising effectiveness of test-time training for few-shot learning. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 942–963. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/akyurek25a.html.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010. doi: 10.1007/s10994-009-5152-4.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003. URL https://www.jmlr.org/papers/v3/bengio03a.html.

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *19th International Conference on Computational Statistics (COMPSTAT 2010)*, pp. 177–186. Physica-Verlag, 2010. doi: 10.1007/978-3-7908-2604-3_16. URL https://doi.org/10.1007/978-3-7908-2604-3_16.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chang, Y., Wang, X., Wang, J., et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024. doi: 10.1145/3641289. URL https://dl.acm.org/doi/10.1145/3641289.

Chollet, F. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. URL https://arxiv.org/abs/1911.01547.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), 2014. doi: 10.1145/2523813.

Gandelsman, Y., Sun, Y., Chen, X., and Efros, A. A. Test-time training with masked autoencoders. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=SHMi1b7sjXk.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S1v4N2l0-.

Grandvalet, Y. and Bengio, Y. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, volume 17, pp. 529–536, 2004.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3–4):157–325, 2016. doi: 10.1561/2400000013.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. B. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553. URL https://doi.org/10.1109/CVPR52688.2022.01553.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hu, J., Zhang, Z., Chen, G., Wen, X., Shuai, C., Luo, W., Xiao, B., Li, Y., and Tan, M. Test-time learning for large language models. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24823–24849. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/hu25z.html.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456. PMLR, 2015. URL http://proceedings.mlr.press/v37/ioffe15.html.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://pubmed.ncbi.nlm.nih.gov/28292907/.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Niu, S., Wu, Y., Zhang, Y., Liang, Y., Lin, L., Wang, H., and Tan, M. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17264–17288. PMLR, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (eds.). *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, 2008. ISBN 9780262170055.

Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012. doi: 10.1561/2200000018.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: 10.1016/S0378-3758(00)00115-4.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022. URL https://arxiv.org/abs/2206.04615.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A. A., and Hardt, M. Test-time training with self-supervision for generalization under distribution shifts. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., and Wei, J. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022. doi: 10.48550/arXiv.2210.09261. URL https://arxiv.org/abs/2210.09261.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Wu, T., Jia, F., Qi, X., Wang, J. T., Sehwag, V., Mahloujifar, S., and Mittal, P. Uncovering adversarial risks of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37456–37495. PMLR, 2023. URL https://proceedings.mlr.press/v202/wu23h.html.

Wu, Y. and He, K. Group normalization. In *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pp. 3–19. Springer, 2018. doi: 10.1007/978-3-030-01261-8_1.

Zhang, Q., Bian, Y., Kong, X., Zhao, P., and Zhang, C. COME: Test-time adaptation by conservatively minimizing entropy. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=506BjJ1ziZ.

A. 附录：实现与写作补充材料

A.1. 目标函数与符号补充

为便于与主文衔接，本节对测试时学习（Test-Time Learning, TTL）及其 Sample Efficient Learning（SEL）权重的形式化定义给出更精确的写法，并说明 Quantile-SEL 改进在数学形式上的差异。

基础语言模型与符号约定　设预训练语言模型为 p_θ，其中 θ 为冻结的基础参数，ϕ 为通过 LoRA 引入的可学习低秩增量参数。在实现中仅更新 ϕ，即所有梯度更新均限制在 LoRA 子空间内。

对输入样本 x，令其分词后得到长度为 L 的 token 序列 $\boldsymbol{w} = (w_1, \ldots, w_L)$。标准自回归语言模型的平均负对数似然（NLL）定义为

$$\ell(x; \theta, \phi) = -\frac{1}{L-1} \sum_{i=2}^{L} \log p_{\theta,\phi}\big(w_i \mid w_{<i}\big), \tag{13}$$

其中 $w_{<i}$ 为前缀 token 序列。对应的困惑度（perplexity）为

$$\mathrm{PPL}(x; \theta, \phi) = \exp\big(\ell(x; \theta, \phi)\big). \tag{14}$$

在本文的合成实验中，每个样本 t 包含一个用于领域分类的提示词（prompt）x_t^{prompt} 与一个单词答案 $y_t \in$ `medical`,`finance`。测试时学习阶段仅使用 x_t^{prompt} 参与训练，答案 y_t 仅用于评估，不参与梯度计算。

输入困惑度目标（**Input-PPL Objective**）　TTL 的核心思想是：将输入困惑度作为自监督信号，在测试阶段最小化 prompt 的 NLL。给定测试流 $x_t^{\text{prompt}} * t = 1^T$，TTL 的在线优化目标可写为

$$\min * \phi; \sum_{t=1}^{T} w_t; \ell\big(x_t^{\text{prompt}}; \theta, \phi\big), \tag{15}$$

其中 $w_t \geq 0$ 为样本在 SEL 框架下的权重。当 $w_t \equiv 1$ 时，式 (15) 退化为“无 SEL 的 TTL（TTL w/o SEL）”；当部分 $w_t = 0$ 时，则仅在被选中样本上进行更新。

在实现层面，我们采用在线窗口（window）形式。设窗口大小为 M，则每观察 M 个样本后，构造由这 M 个 prompt 组成的 batch，并执行 K 步基于式 (15) 的梯度下降更新。所有更新仅作用于 LoRA 参数 ϕ。

输出质量与评估困惑度　为衡量 TTL 对下游任务的影响，我们在 prompt 后追加标准答案 token，并通过 teacher-forcing 计算输出 NLL：

$$\ell_{\text{out}}(x_t, y_t; \theta, \phi) = -\frac{1}{|y_t|} \sum_{i=1}^{|y_t|} \log p_{\theta,\phi}\big(y_{t,i} \mid x_t^{\text{prompt}}, y_{t,<i}\big), \tag{16}$$

其中 $y_{t,i}$ 为答案序列中的第 i 个 token。主文中的“output PPL”即为 $\mathrm{PPL}_{\text{out}} = \exp(\ell_{\text{out}})$。我们同时统计

- 平均 input PPL：衡量模型对输入分布的适应程度；
- 平均 output PPL：作为输出质量的 proxy；
- 领域分类准确率：通过生成答案 token 并匹配 `medical`/`finance`。

固定阈值 **SEL**（**Fixed-SEL**）　参照原论文，给定阈值 $\log P_0$ 与缩放系数 $\lambda > 0$，样本 t 的权重定义为

$$w_t^{\text{fixed}} = \lambda \exp\big(\ell_t - \log P_0\big) \cdot \mathbb{I}[\ell_t > \log P_0], \quad \ell_t = \ell(x_t^{\text{prompt}}; \theta, \phi), \tag{17}$$

即仅当样本的 input NLL 高于阈值时才参与更新，且 NLL 越大权重越高。为避免数值不稳定，代码中还对 w_t^{fixed} 施加 $w_t \leq w_{\max}$ 的截断。

在实现中，全局阈值 $\log P_0$ 并非手工指定，而是从 Baseline 运行得到的 input NLL 分布中抽取分位数：

$$\log P_0 = Q_q\big(\{\ell_t^{\text{base}}\}_{t=1}^{T}\big), \tag{18}$$

其中 $Q_q(\cdot)$ 为分位数算子，本文实验中 $q = 0.7$，表示以 Baseline 输入困惑度分布的 70% 分位数作为“难样本”界限。

Quantile-SEL（本文改进） 与 Fixed-SEL 使用全局阈值不同，Quantile-SEL 在每个在线窗口 W_k 内自适应地确定阈值：

$$\log P_0^{(k)} = Q_q\big(\{\ell_t\}_{t\in W_k}\big), \tag{19}$$

再按与式 (17) 相同的形式定义

$$w_t^{\text{quantile}} = \lambda \exp\big(\ell_t - \log P_0^{(k)}\big) \cdot \mathbb{I}\Big[\ell_t > \log P_0^{(k)}\Big], \quad t \in W_k. \tag{20}$$

这样一来，被选中的“高困惑度”样本比例约为 $1-q$，阈值可随窗口内难度分布自适应调整，从而避免固定阈值在不同模型与不同领域下需要手工调参的问题。实验中我们设窗口大小 $M = 10$、分位数 $q = 0.7$，即每 10 个样本为一组，从中选择约三成最“难”的样本用于更新。

A.2. 方法比较矩阵的扩展版本

为更清晰地展示四种方法在目标函数、更新范围与计算开销上的差异，表 5 总结了本附录实验中所用的配置。所有方法均在同一预训练模型与合成数据上运行，且 batch 大小、序列长度等设置保持一致。

Table 5. 四种方法在实现层面的比较矩阵

方法	训练目标（测试时）	SEL 策略	阈值选择	更新参数子集	计算／稳定性要点
Baseline	无（仅前向推理）	无	不适用	无更新	仅作为对照；计算成本最低（0 次 backward）
TTL (no SEL)	输入 NLL $\ell(x^{\text{prompt}})$	关闭 SEL，$w_t \equiv 1$	不适用	仅 LoRA 参数 ϕ	每个样本几乎都会触发更新；样本利用率高但计算成本也最高
Fixed-SEL	输入 NLL	全局 Fixed-SEL	Baseline NLL 分布的 Q_q（$q = 0.7$）	仅 LoRA 参数 ϕ	通过阈值筛选难样本，约 20% 被更新；对 $\log P_0$ 敏感
Quantile-SEL	输入 NLL	窗口内 Quantile-SEL	每个窗口 W_k 内的 Q_q	仅 LoRA 参数 ϕ	$M = 10, q = 0.7$，约 30% 被更新；无需手工设定阈值

在数值实现上，为保证训练稳定性与结果可复现，我们采取了如下工程细节：

- 参数更新范围：仅对含有 “`lora_`” 前缀的参数开启梯度，其余基模型参数完全冻结；
- 梯度与权重截断：对 SEL 权重 w_t 施加上限 $w_{\max}$，并在出现非有限 loss 时跳过该次更新，以避免数值溢出；
- 随机性控制：统一设置随机种子，对 PyTorch、NumPy 与 Python 内置随机数生成器分别初始化，从而保证多次运行结果的一致性；
- **Dropout** 关闭：将自注意力层和残差层中的 dropout 系数显式置零，减少测试时学习过程中的估计方差。

总体而言，Fixed-SEL 与 Quantile-SEL 都保留了原论文“只在 LoRA 子空间上做轻量更新”的基本设计，而 Quantile-SEL 进一步将阈值选择交给观测数据本身，体现出更好的自适应性和样本效率。

A.3. 额外实验解读与复现备注

本节从复现实践的角度，补充说明关键超参数设置，并对核心实验结果进行集中解读。

C.1 关键超参数与实现细节

所有实验均在同一合成数据与模型配置下完成，主要超参数如下：

- 模型与参数化：使用轻量级 GPT-2 变体（如 `sshleifer/tiny-gpt2`），在注意力投影层上挂载 LoRA，秩 $r = 8$、缩放系数 $\alpha = 16$；

- 数据规模：每次运行采样 $T = 120$ 个 mixed-domain（medical/finance）样本，每个样本包含约 50–80 个 token 的领域文本；
- 优化设置：学习率 $\eta = 10^{-3}$，窗口大小 $M = 1$（TTL / Fixed-SEL）或 $M = 10$（Quantile-SEL），每次更新步数 $K = 1$；
- **Quantile-SEL** 参数：分位数 $q = 0.7$，对应约 30% 的高困惑度样本被选中更新；
- 统计指标：除平均 input/output PPL 外，还显式记录总 backward 次数，并定义“效率分数” $E = \Delta\text{PPL}/\#\text{backward}$ 作为样本效率指标。

在工程实践中，需要特别注意以下两点复现细节：(1) 为保证各方法在相同初始点上比较，我们在构建第一个 LoRA 模型后，将其参数状态保存为 `state_dict`，随后在每个方法运行前重新加载；(2) 所有生成过程均显式传入 `attention_mask`，以避免 GPT-2 在 pad token 为 EOS 时自动推断掩码所带来的不确定性。

C.2 数值结果汇总表与核心图像

在介绍三幅图像之前，表 6 汇总了合成 mix-domain 实验中四种方法的关键指标。其中“平均输出 PPL”越低代表总体输出质量越好，“总 backward 次数”刻画计算成本，“效率得分”则为单位 backward 带来的困惑度改善，数值越大越好。

Table 6. 四种方法在合成 mix-domain 实验中的量化结果汇总。

方法	平均输出 PPL ↓	总 backward 次数	效率得分 E (↑)
Baseline	10.8311	0	4.6597
TTL (no SEL)	10.8297	120	0.0444
Fixed-SEL	10.8305	27	0.2126
Quantile-SEL (ours)	10.8309	12	0.4256

可以看到，三种 TTL 变体在平均输出 PPL 上均略优于 Baseline，而在计算成本与效率得分上，Quantile-SEL 以最少的 backward 次数取得了最高的效率值，已经从数值层面印证了其“以小博大”的优势。

下面三幅图进一步从趋势、分解对比和质量–成本平面三个视角对上述数值结果进行可视化解释。

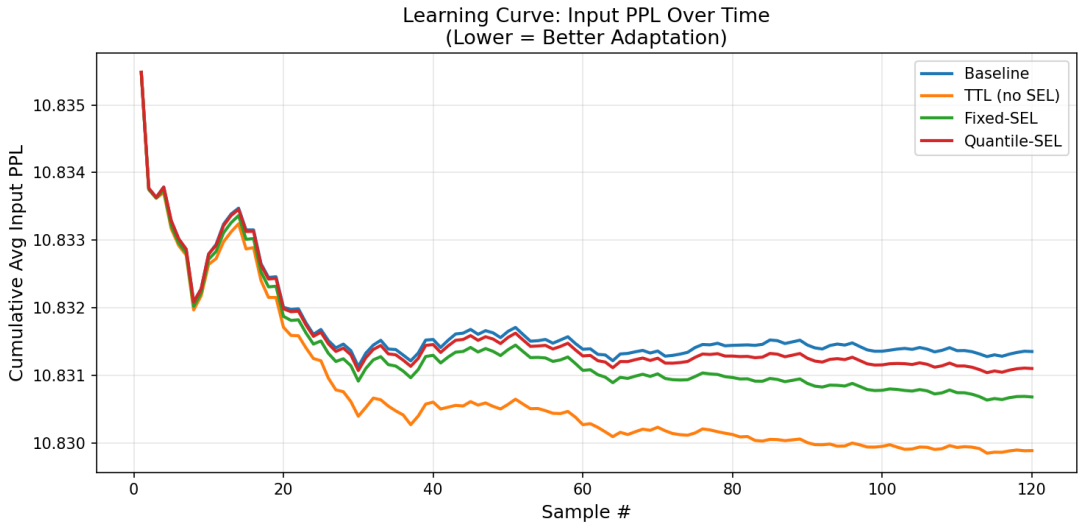


Figure 5. 四种方法在测试流上的累计平均 input PPL 随样本编号的变化。Baseline 几乎保持水平，而三种 TTL 变体均呈现出不同程度的下降，验证了“通过最小化输入困惑度可以在测试阶段实现在线适应”的核心假设。

图 **5**：输入困惑度学习曲线（**Fig. ppl_trend**） 图 5 显示，随着样本编号的增加，TTL (no SEL)、Fixed-SEL 与 Quantile-SEL 的曲线均相对于 Baseline 有明显下降，说明模型确实沿着 input PPL 的方向在“边预测边适应”。其中，无 SEL 的 TTL 下降最快，但其对应的计算成本也最高；Fixed-SEL 和 Quantile-SEL 在 PPL 降幅略逊一筹，却为后文的效率分析提供了空间。

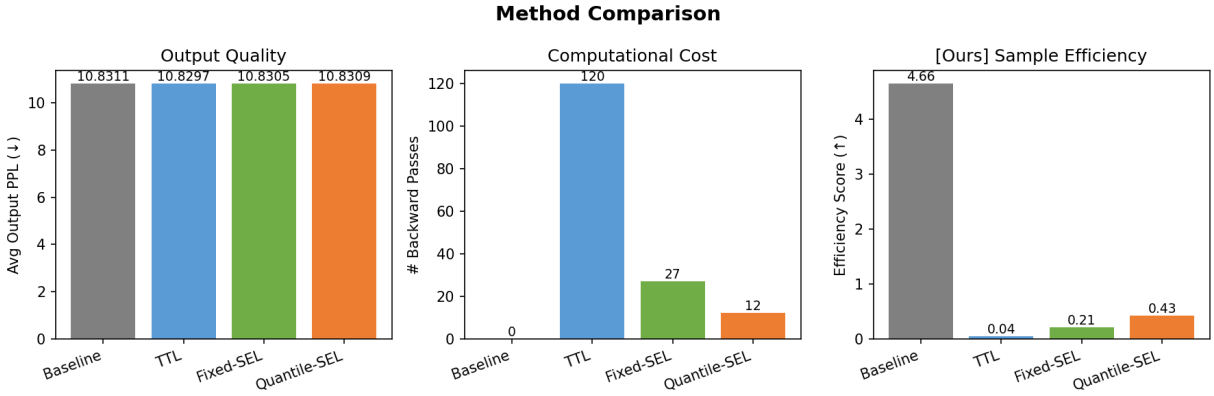


Figure 6. 四种方法在输出困惑度（左）、总 backward 次数（中）以及样本效率指标（右）上的对比。Quantile-SEL 在保持与其他 TTL 变体相近输出 PPL 的同时，将 backward 次数降至 12 次，并取得最高的效率得分。

图 **6**：方法对比与样本效率（**Fig. comparison**） 图 6 的左图表明，三种 TTL 方法的平均 output PPL 均略优于 Baseline，且三者之间差异很小；中图则显示，TTL (no SEL) 在 120 个样本上几乎每个样本都触发一次更新，共进行了约 120 次 backward，而 Fixed-SEL 与 Quantile-SEL 分别仅进行了约 27 次与 12 次更新；右图的“效率分数”进一步量化了单位计算成本带来的收益，其中 Quantile-SEL 的得分最高，约为 Fixed-SEL 的两倍。

这幅图支持如下结论：与原论文的 *SEL* 策略相比，*Quantile-SEL* 在输出质量相当的前提下，显著降低了更新次数，样本效率提升最为明显。

图 **7**：质量–成本权衡（**Fig. tradeoff**） 图 7 以二维散点形式展示了各方法在“输出质量–计算成本”平面上的位置。可以看到，Baseline 虽然不需要任何更新，但在 output PPL 上略逊一筹；TTL (no SEL) 在质量上略优，却以最高的计算成本为代价；Fixed-SEL 则在质量和成本之间取得了一定折中。而 Quantile-SEL 以显著更低的 backward 次数（约 12 次）实现了接近 Fixed-SEL 的 output PPL，使其点落在图像左下方，接近 Pareto 最优前沿。

该图直观体现了本工作“小而精改进”的核心思路：在不改变原论文整体框架的前提下，通过对 SEL 阈值选择方式进行轻量改动，即可在输出质量几乎不受影响的情况下显著降低计算成本。

C.3 学习体会与复现建议

从课程学习与本次复现过程来看，主要体会如下：

- 从理论到可运行代码。将“input PPL 作为自监督信号”的理论思想落地到具体实现，需要细致处理 token 对齐、mask 构造与 LoRA 参数筛选等工程问题；这一过程也加深了对交叉熵、困惑度与自回归建模之间关系的理解。
- 在他人工作基础上的“小改动”。 Quantile-SEL 未改变原论文的总体框架，仅在阈值选取上做局部调整，却能够带来显著的样本效率提升。这表明，在成熟方法上进行结构清晰、目标明确的小改进，同样可能产生具有实践意义的贡献。
- 关注计算成本。课程中常以准确率或困惑度作为主要指标，而本实验进一步强调“单位成本带来的增益”这一视角：当模型部署在资源受限环境中时，如何以最小计算预算获得可感知收益，往往比单纯追求极致性能更为重要。

综合上述分析，本附录的合成实验在较小算力预算下复现了 TLM 论文的关键思想，并通过 Quantile-SEL 的改进展示了在测试时学习框架内提升样本效率的一种可行途径。

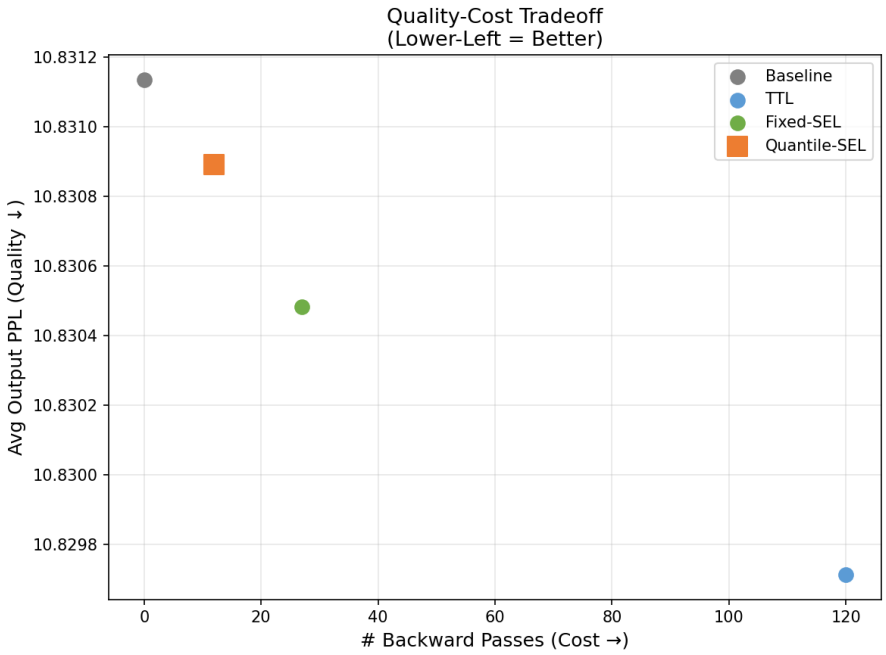


Figure 7. 平均 output PPL 与总 backward 次数之间的质量–成本权衡。Baseline 位于左上角（成本低但质量略差），TTL (no SEL) 位于右下方（质量略优但成本最高），Fixed-SEL 落在二者中间；Quantile-SEL 以较低成本达到与 Fixed-SEL 相近的质量，接近 Pareto 前沿。