arXiv:2602.00952v1 [cs.LG] 1 Feb 2026

大语言模型的最优预算适配

Jing Wang [1] **Jie Shen** [2] **Dean Foster** [3] **Zohar Karnin** [4] **Jeremy C Weiss** [1]

摘要

带标签数据的可用性与下游准确率之间的权衡，仍然是大语言模型（LLM）微调中的核心挑战。我们提出了一个理论严谨的框架，通过将大语言模型适配建模为上下文斯塔克尔伯格博弈，实现预算感知监督微调。在我们的建模中，学习器（领导者）承诺采用评分策略和标签查询策略，而自适应环境（跟随者）则对此做出响应，选择具有挑战性的监督备选。为明确解决标签效率问题，我们将有限监督预算直接纳入学习目标。我们的算法在全反馈机制下运行，并在标准线性上下文假设下实现 $\tilde{O}(d\sqrt{T})$ 遗憾。我们进一步引入最大延迟优先（LLF）置信门控来选择性地查询标签，在 $B = \beta T$ 条件下实现 $\tilde{O}(\sqrt{dB} + c\sqrt{B})$ 的预算感知遗憾界。

1. 引言

微调大语言模型（LLM）已成为将通用模型适配到专业下游任务的主流范式。在医疗、金融和科学推理等高风险领域，领域偏移以及获取标注数据的渠道有限，往往使现成模型难以胜任（Jiang 等人，2023）。虽然监督微调能够显著提升性能，但获取高质量标签的成本高昂，因此推动了标签高效的适配方法的发展。

一个关键挑战在于，并非所有训练样本都同样具有信息量。对所有可用数据盲目进行微调，可能会将标签资源浪费在影响较小的更新上；而选择性地查询标签，则需要有原则的方法来平衡性能与成本。这个问题自然引出了在线决策的视角：学习器必须确定何时需要监督，以及如何在不确定性下更新模型。

监督微调的斯塔克尔伯格博弈视角。我们将监督式大语言模型适配建模为学习器（领导者）与自适应环境（跟随者）之间的斯塔克尔伯格博弈（Balcan 等人，2025；Harris 等人，2024）。每一轮中，学习器针对一组候选回复提交评分策略，而跟随者则根据学习器当前的策略选择具有挑战性的监督备选方案。关键在于，监督信号虽可获得但成本高昂：学习器受固定标注预算限制，必须决定何时请求真实标签。

该表述有别于经典的对抗式或零和训练。跟随者并不操纵标签或奖励，而是通过呈现相互竞争的带标签候选来控制监督决策的难度。学习器的目标是在遵守监督预算的同时最小化累积预测误差。

通过置信度门控实现预算感知监督。 为了高效分配标签，我们引入一种受最大延迟优先（LLF）调度启发的基于置信度的查询机制（Roughgarden，2001）。学习器维护候选得分的上下置信度界，仅当预测不确定性超过阈值时才请求监督。这使得模型能够跳过低信息轮次，同时将监督集中于信息丰富的样本。

综上所述，我们的主要贡献如下：

- 我们将有限标签条件下的监督式大语言模型微调表述为具有显式预算约束的上下文斯塔克尔伯格博弈。
- 我们提出STACKSL，一种可实现全反馈下的 $\tilde{O}(d\sqrt{T})$ 遗憾的监督学习器。
- 我们提出了 LLF-STACKSL，这是一种预算变体，仅使用 $O(B)$ 个带标签轮次即可实现次线性遗憾。
- 我们在真实世界的大语言模型微调任务和合成老虎机问题上验证了我们的框架，展现出很高的标签效率。

[1]美国国家医学图书馆 [2]史蒂文斯理工学院 [3]亚马逊 [4]技术创新研究所。联系邮箱：Jing Wang <jing.wang20@nih.gov>。

预印本。*2026*年*2*月*3*日。

表 1. 比　　较各斯塔克尔伯格老虎机设置下的遗憾保证。此处，d是特征空间的维度。

方法	遗憾	#标签	反馈	模型规模
Harris 等人（完整）（Harris 等人，2024）	$\tilde{O}(d\sqrt{T})$	T	Full	小
Harris 等人（多臂老虎机）（Harris 等人，2024）	$\tilde{O}(T^{2/3})$	none	多臂老虎机	小型
Balcan 等人 (Balcan 等人，2025)	$\tilde{O}(d\sqrt{T})$	none	多臂老虎机	中位数
StackSL（无预算）	$\tilde{O}(d\sqrt{T})$	T	预算	大
LLF-StackSL（预算, $0 < \beta < 1$）	$\tilde{O}\left(d\sqrt{B} + c\sqrt{B}\right)$	$B = \beta T$	预算	大型

我们还对遗憾和标签预算进行了理论分析，相关内容在表中突出展示1。具体而言，（Harris 等人，2024）分析了上下文斯塔克尔伯格博弈，取得了 $\tilde{O}$（$\sqrt{T}$）完全反馈下的遗憾，而 $\tilde{O}$（$T^{2/3}$）则是老虎机机制下的最佳界。随后，（Balcan 等人，2025）通过效用空间归约，在老虎机设定下实现了近乎最优的$\tilde{O}$（$d\sqrt{T}$）遗憾。我们实现了通常只有完全反馈场景才能达到的近乎最优遗憾，同时仅使用了部分标签（$B = \eta T$）。这证明，在斯塔克尔伯格设定下，要实现收敛，并不需要为每个词元或提示词都获取标签。同时，我们的工作有效地证明了“监督学习”和“老虎机学习”处于一个连续谱的两端，而最高效的大语言模型训练发生在中间，此时预算 B 决定了信息流。

2. 相关工作

大语言模型微调。 领域特定的大语言模型适配的最新进展凸显了高效监督和鲁棒性方面的挑战。大量研究致力于将预训练 Transformer 语言模型中的模型行为定位到具体位置（Dai 等人，2022；Olsson 等人，2022）。例如，研究者运用因果中介分析探索 GPT 中隐状态激活的因果效应（Pearl，2022；Vig 等人，2020）。研究发现，在处理主体名称的最后一个词元时，多个中间层中的前馈 MLP 起着决定性作用（Meng 等人，2022）。与此同时，另一系列研究证实了参数高效微调（PEFT）的有效性。LoRA 冻结预训练模型权重，并将可训练的秩分解矩阵注入 Transformer 的每一层，从而使 GPT-3 的可训练参数数量减少 10,000 倍（Hu 等人，2022）。（He 等人，2021）。我们的工作与这些研究密切相关，但重点转向了查询预算有限时的自适应监督。以往研究要么侧重于全局参数高效微调，要么侧重于在特定层和注意力头进行局部干预；与之不同，我们提出了一种动态微调框架，将置信度驱动的查询与局部模型更新相结合。

上下文老虎机与主动学习 我们的方法与主动学习密切相关。在主动学习中，学习器会选择性地查询标签，以尽量减少遗憾或提高准确率（Settles，2009年；Shen 等人，2022；Wang 等人，2022；Shen，2021）。近期研究推进了上下文老虎机中的选择性采样：模型根据不确定性阈值决定是否查询标签（Abbasi-Yadkori 等人，2011）。协同主动学习是一种学习器与人类用户之间的交互模型（Tucker 等人，2024）。不同于典型的基于人类反馈的强化学习（RLHF）训练——在这种训练中，待比较的两个项目是固定的，或像对决老虎机那样从策略中采样（Yue 和 Joachims，2009年），协同主动学习会定义反馈质量的衡量标准，并允许用户引导探索。

我们基于这些概念，引入一种最大延迟优先（LLF）置信门控机制。该机制根据模型的内部不确定性进行选择性查询，从而在遵守资源约束的同时，实现预算感知遗憾界$\tilde{O}$（$\sqrt{dB} + c\sqrt{B}$），其理论效率与标准线性上下文假设相当。协同训练使用带标签样本训练一个弱假设，并利用该假设为另一组样本打标签（Diakonikolas 等人，2024）。

在线多臂老虎机与斯塔克尔伯格博弈。在线学习优化技术广泛应用于多臂老虎机领域（Harris 等人，2024；Shen 等人，2014年）。采用 UCB 分析的线性上下文多臂老虎机是强大的工具。（Abbasi-Yadkori 等人，2011）结合斯塔克尔伯格博弈，在提供跟随者及随时间变化的上下文信息时，可应用在线学习算法来学习领导者。这些算法在完全反馈下实现了遗憾 O（$\sqrt{T}$），并在多臂老虎机反馈设定下实现了遗憾 O（$T^{2/3}$）。在（Balcan 等人，2025）中，采用线性上下文多臂老虎机算法作为领导者的效用。领导者采取能够产生该效用向量的策略，并将奖励作为反馈提供。该算法实现的遗憾为O（$T^{2/3}$）。

斯塔克尔伯格博弈问题已得到广泛研究（Von Stackelberg，2010年；Zhao et al.，2023）。研究表明，经典的多臂老虎机上置信界算法可同时用于

领导者和代理，以实现最优性能（Kaoet al.，2022）。已有大量研究探讨带侧信息的多臂老虎机（Langford &Zhang，2007；Foster et al.，2021）。本文也研究带辅助信息的斯塔克尔伯格博弈，并通过上下文多臂老虎机来解决这一问题。

预算约束下的标签查询。在预算约束下进行选择性查询的策略已在主动学习和在线学习中得到研究（Shen，2021；Shen & Zhang，2021）。例如，（Foster etal.，2020）计算得分最低的动作，并为其他每个动作分配一个与该动作得分和最低得分之间的差距成反比的概率。（Roughgarden，2001）为领导者提供用于集中任务分配的预算，随后跟随者可以自由选择最佳动作。在主动学习中，预算约束下的标签查询长期以来被视为至关重要，早期工作包括预算约束下的特征获取（Lizotteet al.，2003）和成本敏感型标签查询（Kapoor et al.，2007）。这些方法强调在成本约束下仅查询信息量最大的标签。选择性分类允许学习器在不确定性较高时放弃查询，从而在风险与覆盖率之间取得平衡（Chow，1970；El-Yaniv & Wiener，2010年；Bartlett& Wegkamp，2008）。

3. 问题设定

我们在在线学习框架下，研究标注预算有限的大语言模型监督微调。在每一轮 $t \in \{1, \dots, T\}$，学习器观察到一个输入查询 $z_t \in \mathcal{Z}$（例如自然语言提示），并必须从有限的候选集中选择一个响应。

候选集。 对于每个查询 z_t，一个候选集

$$C_t = \{x_{t,1}, \dots, x_{t,|C_t|}\} \subset \mathcal{X}$$

将被构建，其中每个元素对应一个带标签的备选项（例如多项选择选项或排序后的响应）。环境（跟随者）可以根据学习器当前的策略自适应地选择候选集，以体现决策难度的变化。

评分模型。 我们假设在源自大语言模型的特征表示之上使用线性评分函数：

$$s_\theta(z, x) = \theta^\top \phi(z, x),$$

其中 $\phi(z,\ x) \in \mathbb{R}^d$ 是固定或缓慢更新的特征映射，而 $\theta \in \mathbb{R}^d$ 是可训练参数。学习器预测

$$\hat{x}_t = \arg\max_{x \in C_t} s_\theta(z_t, x).$$

监督与损失。需要监督时，学习器会观察真实标签 $x_t^\star \in C_t$，并产生一个监督式 softmax 交叉熵损失：

$$\ell_t(\theta) = -\log \frac{\exp(s_\theta(z_t, x_t^\star))}{\sum_{x \in C_t} \exp(s_\theta(z_t, x))}.$$

预算约束。 标签获取需要付出代价。学习器最多只能在 $B = \beta T$轮中请求监督信号，其中 $\beta \in (0, 1]$。每一轮，学习器可以请求标签并更新其参数，也可以跳过监督信号并执行保持稳定性的更新。

目标与遗憾。 令 $x_t^\star = \arg\max_{x \in C_t} s_{\theta^\star}(z_t, x)$ 表示真实参数 $\theta^\star$ 下的最优响应。我们通过累积评分遗憾来衡量性能：

$$R(T) = \sum_{t=1}^{T} \left[s(z_t, x_t^\star) - s(z_t, \hat{x}_t) \right].$$

目标是设计能够实现次线性遗憾$R(T) = o(T)$并遵守标注预算 B的算法。

4. 方法

我们提出了一种预算感知监督微调框架，用于大语言模型（LLM）。与对比表征学习不同，我们的方法完全在监督设定下运行：假设标签可用但获取成本高昂，学习器必须在固定的标注预算下决定何时请求监督信号。学习过程被建模为学习器（领导者）与自适应环境（跟随者）之间的斯塔克尔伯格博弈，其中环境控制监督决策备选项的难度。

4.1. 监督评分模型

令 $z_t \in \mathcal{Z}$ 表示第 t 轮的输入查询，并令 $C_t = \{x_{t,1}, \dots, x_{t,|C_t|}\}$ 为一个有限的已标注响应候选集（例如，多项选择题的选项）。我们假设在冻结或缓慢更新的大语言模型编码器之上有一个线性评分头。每个查询–候选项对 $(z,\ x)$ 都映射为一个特征向量 $\phi(z,\ x) \in \mathbb{R}^d$，模型会为其分配一个分数

$$s_\theta(z, x) = \theta^\top \phi(z, x), \tag{4.1}$$

其中 $\theta \in \mathbb{R}^d$ 是可训练参数。

学习器预测

$$\hat{x}_t = \arg\max_{x \in C_t} s_\theta(z_t, x). \tag{4.2}$$

4.2. 监督学习目标

请求监督时，学习器会观察正确标签 $x_t^\star \in C_t$ 并执行一次监督更新

使用候选集上的 Softmax 交叉熵损失：

$$\ell_t(\theta) = -\log \frac{\exp(s_\theta(z_t, x_t^\star))}{\sum_{x \in C_t} \exp(s_\theta(z_t, x))}. \tag{4.3}$$

该目标对应于标准的多类别分类或排序。重要的是，所有候选项都是针对同一查询的带标签备选项；其中不使用对比表示学习或基于视图的数据增强。

为了在自适应候选选择和有限监督下稳定学习，我们加入 KL正则化项，以约束相对于先前策略的偏离：

$$\ell_t^{\text{reg}}(\theta) = \ell_t(\theta) + \lambda \, \mathrm{KL}\big(\pi_\theta(\cdot \mid z_t) \,\|\, \pi_{\theta_{t-1}}(\cdot \mid z_t)\big), \tag{4.4}$$

其中， π_θ 表示由 s_θ 诱导的softmax策略，而$\lambda > 0$控制更新稳定性。

4.3. 斯塔克尔伯格交互

我们将学习过程建模为斯塔克尔伯格博弈。学习器（领导者）首先确定评分函数 s_θ，而跟随者则根据学习器当前的策略，自适应地选择具有挑战性的候选集 C_t 作为响应。这些候选集对应于相互竞争的带标签选项，体现了多项选择问答和回答排序等现实中的监督决策问题。

4.4. 通过 LLF 门控进行预算监督

标签获取受预算 $B = \beta T$ 限制，其中T为总轮数，$\beta \in (0, 1]$。为高效分配监督，我们采用了一种受最大延迟优先（LLF）原则启发的基于置信度的门控机制。

在每一轮中，学习器都会为每个 $x \in C_t$ 维护置信区间上界 $\mathrm{UCB}_t(x)$ 和下界 $\mathrm{LCB}_t(x)$，并计算

$$\Delta_t = \max_{x \in C_t} \mathrm{UCB}_t(x) - \min_{x \in C_t} \mathrm{LCB}_t(x). \tag{4.5}$$

如果 Δ_t 超过阈值 ε_t，学习器就会请求监督并执行监督更新。否则，将跳过该标签，仅应用 KL 正则化步骤。这样可以确保有限的预算用于预期信息增益较高的轮次。

4.5. 算法

我们介绍两种算法。**算法1（Stackelberg监督学习器）**对应全反馈机制，在该机制下，每一轮都会查询标签。它采用上置信界（UCB）选择策略来识别响应，并使用 Softmax 交叉熵损失进行监督更新。

算法1 Stackelberg监督学习器（StackSL）

初始化： 策略 π_{θ_0}，格拉姆矩阵 $V_0 = \lambda I$，裁剪级别 ρ，探索 $\{\beta_t\}_{t=1}^T$
对于 $t = 1$到 T **执行**
  观察查询 $z_t \sim \mathcal{D}$
  跟随者最佳响应：
    候选集 $C_t \subseteq \mathcal{X}$ 在 θ_{t-1}条件下
  UCB选择： $\arg\max_{x \in C_t} \Big(\theta_{t-1}^\top \phi(z_t, x) + \beta_t \|\phi(z_t, x)\|_{V_{t-1}^{-1}}\Big)$
    $\hat{x}_t \leftarrow$
    真实标签 $x_t^\star \in C_t$
    $\ell_t \leftarrow -\log \frac{\exp(\theta_{t-1}^\top \phi(z_t, x_t^\star))}{\sum_{x \in C_t} \exp(\theta_{t-1}^\top \phi(z_t, x))}$
  Lossc pp ng
    $\ell_t \leftarrow$最小化{最大值(ℓ_t, 0) ρ}
    $\theta_t \leftarrow \theta_{t-1} - \eta \nabla_\theta \big[\ell_t + \lambda \, \mathrm{KL}\big(\pi_\theta(\cdot|z_t) \,\|\, \pi_{\theta_{t-1}}(\cdot|z_t)\big)\big]_{\theta=\theta_{t-1}}$
  置信度跟踪：
    $V_t \leftarrow V_{t-1} + \phi(z_t, \hat{x}_t)\phi(z_t, \hat{x}_t)^\top$
结束循环

算法 2（LLF预算监督学习器）通过根据置信边际 Δ_t 选择性地查询监督信号，来确保不超出标注预算。该算法仅当置信边际 Δ_t（即候选池中最高上置信界与最低下置信界（LCB）之差）超过阈值 ϵ_t 时，才会选择性地查询标签。

两种算法优化相同的监督目标，区别仅在于标签获取策略。

时间复杂度。 在预算约束设定下，学习器仅在 $B = \beta T$ 轮中查询监督信号。这些轮次的成本与 STACKSL 相同，即通过 $O(d \cdot |C_t|)$ 对候选项进行评分并计算监督损失。其余 $T - B$ 轮不查询标签，只进行 KL正则化稳定性更新，成本为 $O(d)$。因此，总时间复杂度为

$$\text{Total time (LLF-StackSL)} = O(B \cdot d \cdot |C_t| + (T - B) \cdot d),$$

其中第二项体现了跳过轮次所降低的成本。

内存复杂度。预算版本维护与 STACKSL 相同的参数向量、特征表示和格拉姆矩阵。预算监督减少了带标签更新的次数，但不会影响存储需求。因此，渐近内存用量仍然

$$\text{Memory (LLF-StackSL)} = O(d^2 + d \cdot |C_t|),$$

受置信度矩阵 V_t支配。

算法 2 大语言模型斯塔克尔伯格监督学习器，采用 LLF 对冲（LLF-StackSL）

输入： 预算 $B = \beta T$、置信度尺度 c、KL尺度 λ、学习率 η
初始化： 策略 π_θ、格拉姆矩阵 $V \leftarrow \lambda I$，查询计数器 $q \leftarrow 0$
对于 $t = 1$至 T **执行**
  观察查询 $z_t \sim \mathcal{D}$
  跟随者最佳响应：
    候选集 $C_t \subseteq \mathcal{X}$ 在 θ条件下
  UCB选择：
    $\mathrm{UCB}_t(x) = \theta^\top \phi(z_t, x) + \beta_t \|\phi(z_t, x)\|_{V^{-1}}$
    $\mathrm{LCB}_t(x) = \theta^\top \phi(z_t, x) - \beta_t \|\phi(z_t, x)\|_{V^{-1}}$
    $\Delta_t \leftarrow$ 最大值$_{x \in C_t}\mathrm{UCB}_t(x) -$ 最小值$_{x \in C_t}\mathrm{LCB}_t(x)$
    $\varepsilon_t \leftarrow \dfrac{c}{\sqrt{1+q}}$
    $\hat{x}_t \leftarrow \mathrm{UCB}_t(x)$ 的最大参数$_{x \in C_t}$
  如果 $\Delta_t \le \varepsilon_t$ **则**
    $\theta \leftarrow \theta - \eta \nabla_\theta \Big[\lambda\,\mathrm{KL}\,\text{散度}\big(\pi_\theta(\cdot|z_t) \,\|\, \pi_{\theta_{\text{ref}}}(\cdot|z_t)\big)\Big]$
    如果 $q < B$ **那么**
      真实标签 $x_t^\star \in C_t$
      $\ell_t \leftarrow -\text{对数}\,\frac{\text{指数函数}(\theta^\top \phi(z_t,\ x_t^\star))}{\sum_{x \in C_t} \text{指数函数}(\theta^\top \phi(z_t, x))}$
      $\theta \leftarrow \theta - \eta \nabla_\theta \Big[\ell_t + \lambda\,\mathrm{KL}\big(\pi_\theta(\cdot|z_t) \,\|\, \pi_{\theta_{\text{ref}}}(\cdot|z_t)\big)\Big]$
      $q \leftarrow q + 1$
    结束 if
  结束 if
  置信度跟踪：
    $V \leftarrow V + \phi(z_t,\ \hat{x}_t)\phi(z_t,\ \hat{x}_t)^\top$
结束 for

比较。 STACKSL 以 $O(T)$ 次标签查询为代价实现更强的收敛。相比之下，LLF-STACKSL 仅使用 $O(B)$ 个监督轮次来确保标签效率，同时维持无遗憾保证，从而在性能与监督成本之间实现了有原则的权衡。

4.6. 参数分析

我们总结如何设置算法 1和算法 2中的关键参数，并说明它们对查询率、稳定性和计算量的影响。

标签预算 $B = \beta T$。比例 $\beta \in (0, 1]$ 由可用监督资源决定。较大的 β 允许进行更多监督更新，通常能提高准确率，但也会增加标注成本。我们的遗憾界会随 B 增大而改善（参见 $O(\sqrt{dB} + c\sqrt{B})$），因此实践中应采用可行的最大预算。

置信度尺度 c **（LLF 阈值）**。LLF 门控使用阈值 $\varepsilon_t = c/\sqrt{1+q}$，其中 q 是迄今为止的查询轮次。较大的 c 会导致更少的标签查询（门控更严格，更多更新仅使用 KL 散度），而较小的 c 会带来更多的监督。一个实用的规则是根据典型的初始 UCB 宽度校准 c：

$$c \approx \kappa \cdot \mathrm{median}\left\{\max_{x \in C_t} \mathrm{UCB}_t(x) - \min_{x \in C_t} \mathrm{LCB}_t(x)\right\}_{t \le T_0},$$

结合 $\kappa \in [0.5, 1.5]$ 和较短的预热阶段 T_0（例如，500–1,000 ˆ 步）。若要达到目标查询预算 B，可通过二分法调整 c，使实际的 $q \approx B$达到目标。

KL尺度 λ。参数 λ 用于惩罚与先前策略 $\pi_{\theta_{t-1}}$之间的偏差，起到信赖域约束的作用。它可以理解为强制施加每步 KL 上限 τ：

$$\min_{\Delta\theta} \ell_t(\theta + \Delta\theta) + \lambda\,\mathrm{KL}(\pi_{\theta+\Delta\theta} \| \pi_\theta),$$
$$\text{s.t.} \quad \mathrm{KL}(\pi_{\theta+\Delta\theta} \| \pi_\theta) \le \tau.$$

在实践中，我们设置 $\tau \in [10^{-4}, 10^{-2}]$ 并调整 λ，使观测到的 KL 接近 τ。较小的预算通常适合使用更大的 λ，以避免对少数监督轮次过拟合。

候选集规模 $|C_t|$。计算量随 $|C_t|$线性增长。对于大型语料库，可通过近似最近邻检索（按编码器相似度选取前-M 项）构建 C_t，然后进行轻量级重排序。较大的 $|C_t|$能提高对不同选项的区分能力，但会增加成本；我们建议将 $M \in [64, 512]$ 作为一种良好的权衡。

损失裁剪 ρ。对监督损失进行裁剪，即 $\ell_t \leftarrow \min\{\max(\ell_t,\ 0),\ \rho\}$，可避免候选集高度不平衡时出现罕见但极端的梯度。稳健的做法是将 ρ 设为滑动窗口内损失的较高分位数（例如第 90–95 百分位），或设为中位数的一个较小倍数。

学习率 η **和优化器。** 我们使用 AdamW，并采用线性预热和余弦衰减。随着 β 减小，执行的监督步骤会减少，因此每次更新的影响更大；相应地，应降低 η。对于采用 LoRA/PEFT 的 7B–14B 大语言模型，典型的初始值为$\eta \in [10^{-5}, 5 \times 10^{-5}]$。

置信界参数。 对于自归一化 UCB 界，我们采用岭参数 $\lambda_{\text{ridge}} \in [10^{-3}, 1]$和置信水平 $\delta \in \{0.05, 0.1\}$。特征向量经过归一化，以保持 $\|\phi(z,\ x)\| \le 1$，从而稳定 UCB 宽度。如果 LLF 门控很少开启，增大 λ_{ridge} 或加强特征归一化通常会有效。

参考策略 π_{ref}。我们使用前一次迭代 $\pi_{\theta_{t-1}}$作为参考策略。为提高稳定性，可以使用指数移动平均：$\pi_{\text{ref}} = \pi_{\bar{\theta}}$，其中$\bar{\theta} \leftarrow \alpha\theta + (1-\alpha)\bar{\theta}, \alpha \in [0.9, 0.999]$。

实用方案。小预算($\beta \le 0.1$)采用更大的置信度尺度 c（更严格的 LLF 门控）、更大的 KL 权重λ，以避免在少数监督轮次上过拟合；并采用更小的 学习率 η，因为每次查询更新的影响更大。中等预算（$\beta \in (0.1, 0.3]$）采用适中的 c 和 λ，以平衡标签效率与适配速度。大预算（$\beta > 0.3$）采用更小的 c（门控更宽松、监督更频繁）、较弱的 KL 正则化，以及标准学习率 η，逐步接近全反馈训练。

5. 理论分析

回顾一下，我们维护着岭正则化格拉姆矩阵

$$V_t = \lambda I + \sum_{s<t} \phi(z_s, x_s)\phi(z_s, x_s)^\top,$$

ˆ以及岭估计 θ_t。定义置信半径

$$\beta_t = \sigma\sqrt{d\log\Big(1 + \tfrac{tL^2}{\lambda d}\Big) + 2\log\tfrac{1}{\delta}} + \sqrt{\lambda}\,S,$$

以及 UCB/LCB 指标

$$\text{UCB}_t(x) = \hat{\theta}_t^\top \phi(z_t, x) + \beta_t \|\phi(z_t, x)\|_{V_t^{-1}}, \tag{5.1}$$

$$\text{LCB}_t(x) = \hat{\theta}_t^\top \phi(z_t, x) - \beta_t \|\phi(z_t, x)\|_{V_t^{-1}}. \tag{5.2}$$

令 $x_t^\star = \arg\max_{x \in C_t} s(z_t, x)$，并令 x_t为算法选取的动作。我们将得分遗憾定义为 $R(T) = \sum_{t=1}^{T} [s(z_t, x_t^\star) - s(z_t, x_t)]$。使用 LLF 门控时，定义候选池上的宽度（间隔）$\Delta_t = \max_{x \in C_t} \text{UCB}_t(x) - \min_{x \in C_t} \text{LCB}_t(x)$，以及阈值 $\varepsilon_t = c/\sqrt{1+q}$，其中 q 表示截至目前的查询轮次。标签/查询预算为 $B = \beta T$。

假设1（线性得分模型）。存在 $\theta^\star \in \mathbb{R}^d$，使得对于所有 (z, x)，$s(z, x) = \theta^{\star\top}\phi(z, x)$，其中$\|\phi(z, x)\| \le L$ 且 $\|\theta^\star\| \le S$。

假设2（次高斯噪声）。观测到的反馈围绕真实得分呈 σ-次高斯分布。

定理 1（全反馈下的遗憾）。在假设 *1–2*下，以至少 $1-\delta$的概率，全反馈乐观学习器的累积遗憾满足

$$R(T) \le \tilde{O}(d\sqrt{T}).$$

证明。根据线性老虎机的集中不等式，以至少 $1-\delta$ 的概率，真实得分$s(z_t, x)$ 位于 $[\text{LCB}_t(x), \text{UCB}_t(x)]$ 之间，对所有 t 和 $x \in C_t$。在此置信事件下，由乐观性可知，瞬时遗憾有界于所选动作处置信区间的宽度：$r_t \le 2\beta_t \|\phi(z_t, x_t)\|_{V_t^{-1}}$。对 t 求和，并应用椭圆势引理（Abbasi-Yadkori 等人，2011），得到$\sum_{t=1}^{T} \|\phi(z_t, x_t)\|^2_{V_t^{-1}} \le 2d\log\big(1 + \frac{TL^2}{\lambda d}\big)$。综合这些步骤，并将 β_t 上界为 $\tilde{O}(\sqrt{d})$，即可得到所述的 $\tilde{O}(d\sqrt{T})$遗憾界。

□

定理 1 中的速率是典型的 线性上下文多臂老虎机 速率，并且在 T 中对于此设定是极小极大最优的。尽管我们的遗憾分析依赖于$s(z, x)$的线性结构，但这对于利用置信区间推导非渐近界是标准做法。在实验中，我们使用非线性大语言模型表示来评估相同的框架。

正文中的算法1采用贪心更新以提高计算效率。尽管定理 1 对 UCB 式乐观变体进行了形式化分析，但我们观察到，贪心版本在实践中表现出无遗憾行为，这表明大规模大语言模型交互可能提供了充分的“隐式探索”，正如近期深度多臂老虎机研究中观察到的那样。

定理 2（LLF 预算门控下的遗憾）。在假设 *1-2*下，令 $B = \beta T$为监督预算，$\varepsilon_t = c/\sqrt{1+q}$ 为 *LLF* 阈值，其中 q 为迄今为止的查询轮次数。考虑算法 *2*，该算法仅当 $\Delta_t > \varepsilon_t$ 且 $q < B$时查询监督。于是，至少以概率 $1-\delta$，累积遗憾满足

$$R(T) = O\left(\sqrt{dB\log\Big(1 + \tfrac{TL^2}{\lambda d}\Big)} + c\sqrt{B}\right) = o(T).$$

证明。遗憾分为两部分：在查询轮次上的遗憾，以及在跳过的轮次上的遗憾。在查询轮次上，与定理 1 中相同的乐观性论证给出 $r_t \le 2\beta_t \|\phi(z_t, x_t)\|_{V_t^{-1}}$，对至多 $q \le B$ 个查询轮次求和，并结合椭圆势引理，得到 $\tilde{O}(\sqrt{dB})$ 项。在跳过的轮次上，LLF门控条件 $\Delta_t \le \varepsilon_t$ 直接将瞬时遗憾界定为 ε_t，并且由于 $\varepsilon_t = c/\sqrt{1+q}$ 仅在出现新查询时减小，对跳过的轮次求和得到 $\sum \varepsilon_t \le 2c\sqrt{q} \le 2c\sqrt{B}$，从而产生 $c\sqrt{B}$ 项。将这两部分贡献相加，即得到所述的预算感知遗憾界。

□

定理 2 将定理 1 恢复为一种特殊情形：当$B = T$且 c 足够小时，LLF门控在所有轮次上打开，遗憾降至完整反馈 $\tilde{O}(d\sqrt{T})$ 速率。

最优性。</style}截至对数因子，$\sqrt{dB}$ 依赖关系与线性模型的极小极大下界相匹配

对于 B 条信息性观测，这表明预算约束下的遗憾界在信息论意义上是紧的。

6. 实验

本节总结关键实证发现，以验证斯塔克尔伯格监督微调框架在真实世界大语言模型微调和合成线性赌博机上的表现。所用大语言模型为 Qwen3-4B-base（Team, 2025）。实验在一台配备两块 GPU 的服务器上进行，分别为 NVIDIA RTX PRO 6000 和 NVIDIA RTX 6000。我们算法中的 β 设为 0.5。

数据集 实验使用单项选择问答数据集 MedQA 的英语版本（题目采用美国医学执照考试的风格）（Jin 等人, 2021）。该数据集包含 10,178 道训练题和 1,273 道测试题。我们还纳入了大规模多任务语言理解（MMLU）数据集（Hendrycks 等人, 2020）中的临床、大学生物学和医学遗传学等主题。临床主题包含 29 道训练题和 265 道测试题。大学生物学主题包含 16 道训练题和 144 道测试题。医学遗传学主题包含 11 道训练题和 100 道测试题。

6.1. 参数探索与敏感性分析

我们通过受控模拟，研究 STACKSL 和 LLF-STACKSL 中关键算法参数的影响。

LLF-StackSL 的标签效率为验证预算算法 LLF-STACKSL（定理 2），我们固定超参数，并改变标签预算比例 $\beta \in \{1.0, 0.5, 0.25, 0.1\}$。根据表 2，仅使用 10% 的标签即可达到全信息性能的 ≈94%。这从大语言模型层面展示了我们的标签效率保证：LLF-STACKSL 在性能仅有小幅下降的情况下，大幅节省了资源。

表 2. 使用 Qwen3-4B 在 MedQA 上的标签效率。

方法	β	使用比例	测试准确率	占完整模型的百分比
完整 STACKSL	1.0	1.00	0.5918	100.0%
LLF-STACKSL	0.5	0.50	0.5826	98.45%
LLF-STACKSL	0.25	0.25	0.5721	96.68%
LLF-STACKSL	0.10	≈0.10	0.5585	94.38%

合成线性老虎机：验证定理 2 我们运行受控模拟（$T = 20,000$，$d = 20$）来检验预算感知的 $\tilde{O}(\sqrt{dB} + c\sqrt{B})$ 遗憾行为。如表 3所示，LLF 的每步遗憾接近全信息基线（0.0020 对比 0.0024），而相同预算下的随机查询产生的遗憾高出一个数量级。

表 3. 合成线性赌博机遗憾与预算的关系。

方法	β	遗憾/T	查询/T
LLF	0.10	0.00244	0.10
随机	0.10	0.02284	0.10
LLF	0.25	0.00243	0.25
LLF	0.50	0.00209	0.50
完整 ($\beta = 1$)	1.00	0.00200	1.00

表4. 裁剪 ρ对损失分布的影响。

方法	平均损失	最大损失
具有挑战性的候选项（No ρ）	1.37	23.5
STACKSL（完整）	1.36	5.0

通过裁剪实现稳定性 我们考察裁剪ρ 对损失尾部的影响。如表 4所示，在 $\rho = 5$处进行裁剪会显著截断罕见的对抗性损失尖峰（最大值： 24.5 →5.0），同时基本不改变平均损失，这支持了裁剪在稳定斯塔克尔伯格动态中的作用。

KL 参数我们探索了表5 和表6所示的参数设置。参数 k 控制监督候选集中包含的具有挑战性的带标签备选项数量 C_t。根据结果，我们选择 KL 参数0。7 和 $k = 4$。

6.2. 现实世界中的应用

我们首先在 MedQA 多项选择任务上，将 StackSL 应用于 Qwen3-4B-Base，并使用冻结的 Qwen3- 4B 副本作为 KL 项的参考模型。我们在表 7 中将我们的方法与标准监督微调（SFT）及基线方法进行比较。

与 MC-Ranker 的比较。 作为一种强有力的非保守基线，我们对 Qwen 进行微调，使其根据问题提示词的条件似然为多项选择选项排序。对于每个提示词为 x 、选项为 $\{c_i\}_{i=1}^n$的示例，我们将提示词分别与每个选项拼接，构造五个输入， [提示词 || 选项]，，并对这五个拼接结果进行一次前向传播（批处理）。令 $\mathcal{T}(c_i)$ 表示与选项片段对应的词元索引（通过注意力掩码识别）。我们计算每个选项经长度归一化 后的得分

$$s_i = \frac{1}{|\mathcal{T}(c_i)|} \sum_{t \in \mathcal{T}(c_i)} \log p(y_t \mid x, y_{<t}),$$

也就是说，仅对选项词元计算平均对数似然（屏蔽提示词），从而消除短选项偏差。训练通过最小化五个分数上的标准交叉熵来完成：

$$\mathcal{L}_{\text{CE}} = -\log \frac{\exp(s_{y^\star})}{\sum_{j=1}^{5} \exp(s_j)},$$

表 5。 KL 权重（λ_{KL}）对测试准确率的影响。算法 2，具有挑战性的候选项 k=4，以及 N=1000微调示例。

λ_{KL}	测试准确率 @ 1000
0.3	0.2836
0.5	0.3009
0.7	**0.3166**
0.9	0.3134

表 6。 困难负例数量（k）对测试准确率的影响。算法 2， λ_{KL}=0.7，以及 N=1000 微调示例；模型、数据划分和评估协议均相同。

具有挑战性的候选项 k	测试准确率 @ 1000
1	0.2765
2	0.2687
3	0.3126
4	**0.3166**

其中 $y^\star$ 是正确选项索引。这个基线不使用教师模型、KL正则化、门控/预算机制，也不进行显式的困难负例挖掘。实现细节如下：整理器分别截断提示词和选项（提示词最大长度、选项最大长度）；对选项进行分词时，设置 add special tokens=False，以确保 $|\mathcal{T}(c_i)|$ 仅反映选项词元。将提示词和选项张量沿序列维度拼接，并重塑为 $(B \times n)$ 批次，即可在单次前向计算中为全部五个选项打分。

评估（多项选择准确率）。测试时，我们采用相同的掩码长度归一化评分来获得 $\{s_i\}_{i=1}^{n}$，并预测 $i = \arg\max_i s_i$。准确率是在留出集上， $i = y^\star$ 成立的样例所占的比例。此评估与训练目标一致，并避免了选项长度或词元化差异带来的偏差。

结果表8显示，与零样本 Qwen 基线（0.2521）相比，使用我们的 MC-Ranker 进行监督微调带来了约 0.1155 的显著绝对增益（达到 0.3676），这证实了即使是轻量级的基于似然的训练，也能显著提升多项选择推理能力。我们的 StackSL 方法以**0.3810**的准确率取得最佳表现，进一步优于 MC-Ranker，表明对比式选项感知训练能够更好地区分正确选项与干扰项。预算版本 LLF-StackSL（0.3480）的表现低于完整 StackSL，这可能反映了标签/查询门控所引入的权衡：虽然它减少了标注和计算使用量，但也限制了监督信号，从而可能在固定测试集下限制最高准确率。表9展示了 MMLU 数据集上的结果。“Base+FT”指在训练数据集上进行微调的 Qwen3-4B 大语言模型。StackSL 算法在医学遗传学和大学生物学主题上取得了最佳表现。LLF-StackSL 在临床 主题。

7. 结论

我们从上下文斯塔克尔伯格博弈的视角，提出了一种用于大语言模型监督微调的预算感知框架。该框架将基于源自大语言模型的特征的线性评分函数、置信度驱动的查询策略以及 KL 散度正则项相结合，以便在有限监督下稳定更新。我们提出了两种覆盖不同监督程度的算法：STACKSL（一种全反馈监督学习器）和 LLF-STACKSL（一种置信度门控变体），后者仅在信息丰富的轮次分配固定的标签预算。我们推导了高概率遗憾保证，表明在完全反馈下，STACKSL 达到了经典的 $\tilde{\mathcal{O}}(d\sqrt{T})$ 速率。LLF-STACKSL 达到了预算感知界 $\tilde{\mathcal{O}}(\sqrt{dB}) + c\sqrt{B}$，同时仅在 $B = \beta T$ 轮中查询监督信号。总之，这些结果证明，在冻结或局部线性化的编码器下，我们的框架既能高效利用标签，又能相对于大语言模型表示上的最佳线性评分器实现无遗憾，为满足监督约束的高效大语言模型适配奠定了有理论依据的基础。

影响声明

本文旨在推动机器学习领域的发展。我们的工作可能带来许多社会影响，但我们认为没有哪一项需要在此特别强调。

参考文献

Abbasi-Yadkori, Y., P´al, D., and Szepesv´ari, C. 线性随机多臂老虎机的改进算法。神经信息处理系统进展, 24, 2011.

Balcan, M.-F., Bernasconi, M., Castiglioni, M., Celli, A., Harris, K., and Wu, Z. S. 利用侧信息在斯塔克尔伯格博弈中进行近乎最优的多臂老虎机学习。*arXiv*预印本 *arXiv:2502.00204*，2025。网址 https://arxiv.org/abs/2502.00204。

Bartlett, P. L. 与 Wegkamp, M. H. 使用合页损失的带拒绝选项的分类。载于第*19*届学习理论年会（*COLT*）论文集，2008。Chow, C. K. 论最优识别错误率与拒绝的权衡。*IEEE* 信息论汇刊，16(1): 41–46, 1970。Dai, D.、Dong, L.、Hao, Y.、Sui, Z.、Chang, B. 与 Wei, F. 预训练 Transformer 中的知识神经元。载于第*60*届协会年会论文集

表 7. 使用 Qwen3-4B 的 MedQA 性能。

方法	训练目标		准确率（↑）	F1（↑）	标准差
SFT	标准监督交叉熵	py	0.5981	0.5940	0.0023
监督（交叉熵）	对候选项计算交叉熵	ate 集合 0.6012 0.5972 0.0088			
监督（交叉熵，无正则化）	对困难候选项计算交叉熵	候选集 0.5871 0.5829 0.0069			
STACKSL（完整）	监督交叉熵 +KL 散度正则化	化 +裁剪 **0.5962 0.5922** 0.0045			

表 8. MedQA 上各方法的测试准确率。

方法	准确率
Qwen（基线）	0.2521
MC-Ranker（基线 + 微调）	0.3676
StackSL（堆叠）	**0.3810**
LLF-StackSL（堆叠预算）	0.3480

表 9. MMLU 各学科的准确率。

科目	Base	基础 + FT	StackSL	LLF-StackSL
医学遗传学	0.3600	0.3700	**0.4000**	0.3509
临床	0.3245	0.3283	0.3547	**0.3660**
大学生物学	0.4306	0.4514	**0.4583**	0.4375

计算语言学（第*1*卷：长篇论文），第8493–8502页，2022。

Diakonikolas, I., Ma, M., Ren, L., 和 Tzamos, C. 通过基于流的主动学习，在弱依赖条件下快速协同训练。发表于国际机器学习大会，2024。

El-Yaniv, R. 和 Wiener, Y. 论选择性分类的基础。《机器学习研究期刊》，11:1605–1641，2010年。

Foster, D. J., Krishnamurthy, A., Kakade, S., 等。超越 UCB：多臂老虎机中的最优且高效的最佳臂识别。发表于*COLT*，2020。

Foster, D. J., Kakade, S. M., Qian, J., 和 Rakhlin, A. 交互式决策制定的统计复杂度。*arXiv* 预印本 *arXiv:2112.13487*，2021。

Harris, K.、Wu, Z.S. 和 Balcan, M.-F. 具有辅助信息的斯塔克尔伯格博弈中的遗憾最小化。载于 *NeurIPS*，2024 年。网址 https://proceedings.neurips.cc/paper_files/paper/2024/file/178b6cd141e003fa5ff808c7d7d8e2cc-Paper-pdf.

He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T. 和 Neubig, G. 迈向参数高效迁移学习的统一视角。*arXiv*预印本 *arXiv:2110.04366*，2021年。Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. 和 Steinhardt, J. 衡量大规模多任务语言理解能力。*arXiv* 预印本 *arXiv:2009.03300*，2020。

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. 等。LoRA：大语言模型的低秩适应。*ICLR*，1(2):3，2022。

Jiang, L. Y.、Liu, X. C.、Nejatian, N. P.、Nasir-Moin, M.、Wang, D.、Abidin, A.、Eaton, K.、Riina, H. A.、Laufer, I.、Punjabi, P. 等。卫生系统规模语言模型是多用途预测引擎。《自然》，619(7969)：357–362，2023 年。

Jin, D.、Pan, E.、Oufattole, N.、Weng, W.-H.、Fang, H. 和 Szolovits, P. 这位患者患有什么疾病？一个基于医学考试的大规模开放域问答数据集。《应用科学》，11(14)：6421，2021 年。

Kao, H.、Wei, C.-Y. 和 Subramanian, V. 基于分层信息结构的去中心化合作强化学习。载于算法学习理论国际会议，第 573–605 页。PMLR，2022。

Kapoor, A.、Graepel, T.、Horvitz, E. 和 Basu, S. 带成本的主动学习。载于人工智能联合国际会议（*IJCAI*）第 *20* 届会议论文集，2007 年。

Langford, J. 和 Zhang, T. 用于上下文多臂赌博机的轮次贪心算法。神经信息处理系统进展，20(1):96–1，2007。

Lizotte, D.、Madani, O. 和 Greiner, R. 朴素贝叶斯分类器的预算学习。载于人工智能中的不确定性第 *19* 届会议论文集（*UAI*），2003 年。

Meng, K.、Bau, D.、Andonian, A. 和 Belinkov, Y. 在 GPT 中定位和编辑事实关联。神经信息处理系统进展，35:17359–17372，2022。

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A. 等。上下文学习与归纳头。*arXiv* 预印本 *arXiv:2209.11895*，2022。Pearl, J. 直接效应与间接效应。载于概率推断与因果推断：*Judea Pearl* 的著作，第 373–392 页。2022。Roughgarden, T. Stackelberg 调度策略。载于第三十三届年度 *ACM* 计算理论研讨会论文集，第 104–113 页，2001。

Settles, B. 主动学习文献综述。威斯康星大学麦迪逊分校计算机科学技术报告 *1648*，2009年。Shen, J. 在恶意噪声下半空间的样本最优 PAC 学习。载于国际机器学习大会，第 9515–9524 页。PMLR，2021。Shen, J. 和 Zhang, C. 在恶意噪声下对半空间进行属性高效学习：近乎最优的标签复杂度与噪声容忍度。载于算法学习理论，第 1072–1113 页。PMLR，2021。Shen, J., Xu, H. 和 Li, P. 最大范数正则化的在线优化。*NeurIPS*，27，2014。Shen, J., Cui, N. 和 Wang, J. 度量公平主动学习。载于国际机器学习大会，第 19809–19826 页。PMLR，2022。Team, Q. Qwen3 技术报告，2025。网址：https://arxiv.org/abs/2505.09388。Tucker, A. D., Brantley, K., Cahall, A. 和 J oachims, T. 利用隐式用户反馈对大语言模型进行协同主动学习。载于国际机器学习大会，2024。Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y. 和 Shieber, S. 使用因果中介分析研究语言模型中的性别偏见。神经信息处理系统进展，33：12388–12401，2020。Von Stackelberg, H. 市场结构与均衡。施普林格科学与商业媒体，2010。Wang, J., Shen, J., Ma, X. 和 Arnold, A. O. 面向阅读理解的基于不确定性的主动学习。机器学习研究汇刊，2022。Yue, Y. 和 Joachims, T. 将信息检索系统的交互式优化建模为双臂老虎机问题。载于国际机器学习大会，第 1201–1208 页，2009年。Zhao, G., Zhu, B., Jiao, J. 和 Jordan, M. 在全知跟随者下斯塔克尔伯格博弈中的在线学习。载于国际机器学习大会，第 42304–42316 页。PMLR，2023。