XXXX

基于LLM语义指导与主动推理的云边协同资源调度优化

诸葛斌，洪仕玉，许云汉，张子天，董黎刚，蒋献
（浙江工商大学，浙江 杭州 310018）

摘　要：针对云边协同环境下LLM推理面临的实时性与动态不确定性挑战，本文提出一种LLM与主动推理协同的决策框架AIF-LLM。该框架利用Sentence Transformer将宏观语义指导量化为策略偏好向量，并将其融入预期自由能计算中，实现高层指导下的精确决策；设计元学习模块，根据环境不确定性动态调整语义偏好权重λ，以平衡宏观指导与主动推理精确成本模型；同时利用LLM的语义理解能力生成初始信念先验，显著优化冷启动性能和样本效率。实验表明，AIF-LLM的QoS满意度达92.57%，相较于主流SAC、PPO、DQN和A2C算法分别实现2.90、4.90、8.00和10.00个百分点的绝对提升。在系统极限负载区间，其将QoS违约风险和长尾失败率分别削减了28.07%和47.67%，验证了框架在复杂环境下卓越的鲁棒性与自适应性。

关键词：LLM推理卸载；资源分配；云边计算；主动推理；多模态感知
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.

An Optimization Framework for Cloud-Edge Collaborative Resource Scheduling Based on LLM Semantic Guidance and Active Inference

ZHUGE Bin, HONG Shiyu, Xu Yunhan, Zhang Zitian, Dong Ligang, Jiang Xian
Zhejiang Gong shang University, Hangzhou 310018, China

Abstract: To address the challenges of real-time performance and dynamic uncertainty in Large Language Model (LLM) inference within cloud-edge collaborative environments, this paper proposes a synergistic decision-making framework combining LLMs and Active Inference (AIF-LLM). The framework utilizes Sentence Transformer to quantify macroscopic semantic guidance into policy preference vectors, which are then integrated into the calculation of Expected Free Energy to achieve precise decision-making under high-level guidance. A meta-learning module is designed to dynamically adjust the semantic preference weight λ based on environmental uncertainty, thereby balancing macroscopic guidance with the precise cost modeling of Active Inference. Simultaneously, the framework leverages the semantic understanding capabilities of the LLM to generate initial belief priors, significantly optimizing cold-start performance and sample efficiency. Simulation results indicate that AIF-LLM reaches a Quality of Service (QoS) satisfaction rate of 92.57%, achieving absolute improvements of 2.90, 4.90, 8.00, and 10.00 percentage points

收稿日期：XXXX−XX−XX；修回日期：XXXX−XX−XX

compared to mainstream SAC, PPO, DQN, and A2C algorithms, respectively. In system limit load scenarios, the framework successfully reduces the QoS violation risk and long-tail failure rate by 28.07% and 47.67%, respectively, validating the framework's superior robustness and adaptability in complex environments.

Key words: LLM inference offloading, resource allocation, cloud-edge computing, active inference, multimodal perception

1 引言

近年来，大型语言模型（LLM）凭借卓越的自然语言处理能力，在智能问答与代码辅助等领域取得突破。然而，其动辄数千亿的参数量导致推理阶段对计算资源需求极高，在实时数据分析等延迟敏感场景中，如何在受限资源下实现高效推理成为亟待解决的难题。传统的集中式云计算虽算力强大，但受限于网络延迟；边缘计算虽能降低延迟，但单设备资源匮乏。云边协同计算通过任务灵活卸载与资源分配，为LLM推理提供了理想平台[1]。

然而，将LLM部署于云边环境面临重重挑战：首先是计算密集性与边缘资源受限的矛盾；其次是动态不确定性（如节点故障、负载突发）使传统静态或纯强化学习方法易陷入“奖励黑客”困境，鲁棒性不足。

但是LLM的推理延迟和计算开销可能抵消其优化带来的性能增益，甚至威胁到边缘系统的实时性。针对这一核心挑战，本框架并未采取‘全时在线’的LLM决策模式，而是设计了一种‘轻量级语义协同’机制。我们利用轻量级编码器（如Sentence Transformer）将LLM的宏观指导快速转化为低维向量，并将昂贵的LLM推理过程异步化，仅在冷启动或环境剧变时按需触发。这种设计确保了系统在获得高级语义指导的同时，将决策开销维持在毫秒级的主动推理（AIF）层面，从而有效规避了引入大模型带来的额外延迟风险。

针对“实时性”这一核心痛点，必须关注“决策开销悖论”：即引入LLM辅助决策所带来的额外推理延迟，可能抵消其优化带来的性能增益，甚至威胁系统实时性。为此，本框架摒弃了“全时在线”的LLM决策模式，设计了“轻量级语义协同”机制。通过轻量级编码器（Sentence Transformer）将宏观指导向量化，并采取异步触发机制，仅在环境剧变或冷启动时调用昂贵的LLM推理。这确保了日常决策开销维持在毫秒级的主动推理层面，从根本上规避了决策延迟风险。

主动推理基于自由能原理，通过维持和更新关于世界状态的“信念”，在观测有噪声的环境中展现出极强的鲁棒性[2]。本文提出的AIF-LLM框架核心逻辑在于：利用LLM的“宏观语义理解”应对未知任务，结合AIF的“信念驱动决策”处理微观不确定性。

本文的主要贡献包括：

● 提出LLM与AIF协同的决策框架：该框架实现了LLM宏观语义指导与AIF微观精准执行的闭环协同，提升了动态不确定环境下云边资源调度的智能管理能力与适应性；

● 语义向量化与预期自由能的精确集成：利用Sentence Transformer将LLM的宏观指导（如资源偏好）量化为384维策略偏好向量V_{pref}，并将其无缝集成至AIF的预期自由能计算中，实现可量化的精确指导；

● 基于元学习的动态EFE权重λ自适应优化：引入MAML思想设计元学习模块，根据环境不确定性指标（如故障率、负载波动）动态调节语义偏好权重λ。该机制有效平衡了LLM的宏

观逻辑与 AIF 的成本模型，增强了系统的泛化能力；

● LLM 驱动的信念先验初始化：利用 LLM 的语义理解能力，根据任务描述为 AIF 提供初始信念先验 D_{LLM}，取代传统的随机初始化，显著提升了新任务场景下的样本效率与冷启动性能；

● 全面的实验验证：仿真结果显示，AIF-LLM 在 QoS 满意度上较主流 DRL 算法（A2C、DQN、PPO）分别提升 10.00、8.00 和 4.90 个百分点；相较于 SOTA 基线 SAC 仍保持 2.90 个百分点优势。尤其在高负载区间，本框架将 QoS 违约风险和长尾任务失败率分别削减了 28.07% 和 47.67%，充分验证了其卓越的鲁棒性。

2　研究现状

在云边协同环境中，LLM 推理卸载与资源分配是当前的研究热点。由于 LLM 模型规模持续扩大，边缘设备在计算能力与内存方面面临严峻挑战，促使研究者提出了多种推理卸载策略[3]。这些策略通常涵盖模型分割、任务调度与结果聚合等核心技术。例如，通过将模型按层分割，使部分层在边缘执行而其余层在云端处理，能够有效平衡延迟与计算负载[4]；此外，针对网络状况与设备负载的动态决策机制可显著优化整体系统性能[5]。在资源分配方面，传统的调度算法往往难以适应云边环境中异构资源和动态工作负载的特点。近年来，基于强化学习的方法被广泛应用于云边资源分配，通过智能体与环境的交互学习最优的资源调度策略[6]。然而，这些方法通常需要大量的训练数据和复杂的奖励机制设计，且在面对快速变化的环境时，其适应性仍有提升空间。

主动推理作为一种统一的感知、行动与学习理论，为不确定环境下的决策提供了生物学启发的框架。其核心逻辑在于通过最小化变分自由能来降低环境模型的预测误差，进而持续更新系统对世界状态的“信念”[7]。在任务调度中，AIF 能够比传统方法更有效地处理资源波动与网络延迟等动态不确定性[8][9]。部分研究已利用 AIF 指导信息增益最大化的资源探测，或通过对未来状态的预测优化负载均衡决策。然而，传统 AIF 框架在处理高维感知信息与高级语义推理方面仍面临瓶颈，难以直接从系统原始数据中提取出深层次的高级特征。

将 LLM 与系统管理技术结合是当前新兴的研究方向。LLM 强大的语言理解和生成能力可以为系统提供高级指令理解、任务规划、情境感知和人机交互等功能[10]。例如，LLM 可以帮助系统理解复杂的自然语言指令，将其转化为可执行的调度策略[11]。在资源分配任务中，LLM 可以根据历史经验和语义信息，为系统提供调度策略的建议，甚至生成新的优化目标[12]。然而，如何有效地将 LLM 的符号推理能力与系统底层的感知-行动循环相结合，并解决 LLM 推理的实时性问题，仍然是一个开放性问题。本研究正是基于此背景，旨在通过云边协同的 LLM 推理卸载和资源分配机制，赋能系统更智能、更高效地在动态环境中进行资源管理。

3　问题分析与框架概述

3.1　问题定义

在高度动态的云边协同环境中，实现高效的 LLM 推理卸载与资源分配面临多重挑战，本研究旨在解决以下核心问题：

智能卸载与异构资源管理：云边环境由计算能力、存储及带宽各异的异构节点组成。系统需实时感知任务需求与节点状态，智能决策任务应在本地执行、邻近卸载还是上传至云端，以实现延迟、能耗与资源利用率的多目标优化[13]。针对突发负载和网络波动等动态不确定性，传统静态策略已失效，亟需构建能够实时调整的自适应机制。传统静态策略已失效，亟需构建能够实时调

整的自适应机制[14]。

决策优化与环境感知：实现智能调度的关键在于从设备指标、流量等原始多源数据中提炼出高层次的语义信息，并将其转化为AIF模型可理解的信念状态[15]。在有限资源下，系统必须平衡多任务优先级与QoS约束，利用LLM的先验知识指导资源探测，减少无效尝试，并确保决策在毫秒级内完成以响应突发请求[16]。

决策延迟约束下的实时性平衡：不同于传统瞬时调度假设，LLM辅助决策本身产生的推理开销不可忽略。系统必须在“获取高质量语义指导”与“降低额外决策延迟”之间寻找帕累托最优解，防止决策过程过长导致原本可达标的任务因超时而发生QoS违约。

LLM与主动推理框架的有效融合：需将LLM的符号推理能力转化为AIF概率框架下的操作先验或策略建议[17]。构建后验结果反馈链路，通过任务执行状态动态更新元学习参数或优化Prompt上下文，驱动系统持续进化。通过智能卸载与分配机制，确保集成LLM后的调度框架仍能维持实用化的响应速度[18]。

3.2 新框架概述

AIF-LLM框架通过五个核心模块的深度协同，构建了一个具备“宏观认知”与“微观自适应”双重特性的智能系统，旨在攻克云边协同中LLM推理的实时性与动态不确定性难题。图1展示了本文提出的框架的整体架构。

该框架主要由以下几个关键模块组成：

（1） 多模态感知与特征提取融合模块

该模块是框架实现环境感知的核心组件，其设计目标是高效整合云边环境中多源异构的数据流，并将其转化为统一、高层次的感知表示。

多维数据整合：实时监测资源状态（CPU/GPU利用率、功耗、温度）、网络状态（带宽、延迟波动）以及任务特征（复杂度、QoS约束）。

三阶段工作流程：首先进行数据预处理，对原始数据进行清洗、归一化与特征编码；其次是特征提取，利用时序分析与异常检测等深度学习模型，从低级统计特征中提炼出能够捕捉“设备过载”或“网络拥堵”情境的高级语义特征；最后通过特征融合，采用注意力机制与图神经网络捕捉异构数据间的内在关联，最终输出融合状态向量。

决策支撑：该融合状态向量为AIF模型提供了语义丰富的观测输入，是其进行信念更新与最小化预期自由能策略选择的必要前提[19]。

（2） LLM辅助决策模块

该模块作为系统的高级认知大脑，利用LLM强大的推理能力为底层AIF模型提供宏观指导。

语义信号转化：LLM接收来自感知模块的高级特征，进行风险评估与规划。它生成关于最优状态的初始信念先验D_{LLM}以解决AIF的冷启动问题，并将宏观策略建议通过Sentence Transformer编码为384维的策略偏好向量v_{pref}，作为语义偏好项集成至AIF的预期自由能计算中。

异步协同设计（化解延迟悖论）：为避免LLM自身高昂的推理开销威胁边缘实时性，框架设计了“轻量级语义协同”机制。系统仅在冷启动或环境剧增（如Ut超过阈值）时按需触发LLM推理，而在常规周期内仅依赖AIF模型和存量v_{pref}。同时，将LLM任务异步卸载至云端或高算力节点，确保关键决策路径的开销维持在毫秒级[20]。

（3） AIF模型

AIF模型是框架的核心决策引擎，它基于自由能原理，将资源调度视为一个持续的闭环感知-行动过程[21]。

变分自由能（VFE）最小化：系统内部维护一个关于环境状态的概率图模型，在接收到观测数据时，通过最小化VFE更新后验信念。引入的D_{LLM}显著加速了这一过程。

预期自由能（EFE）引导行动：系统选择能

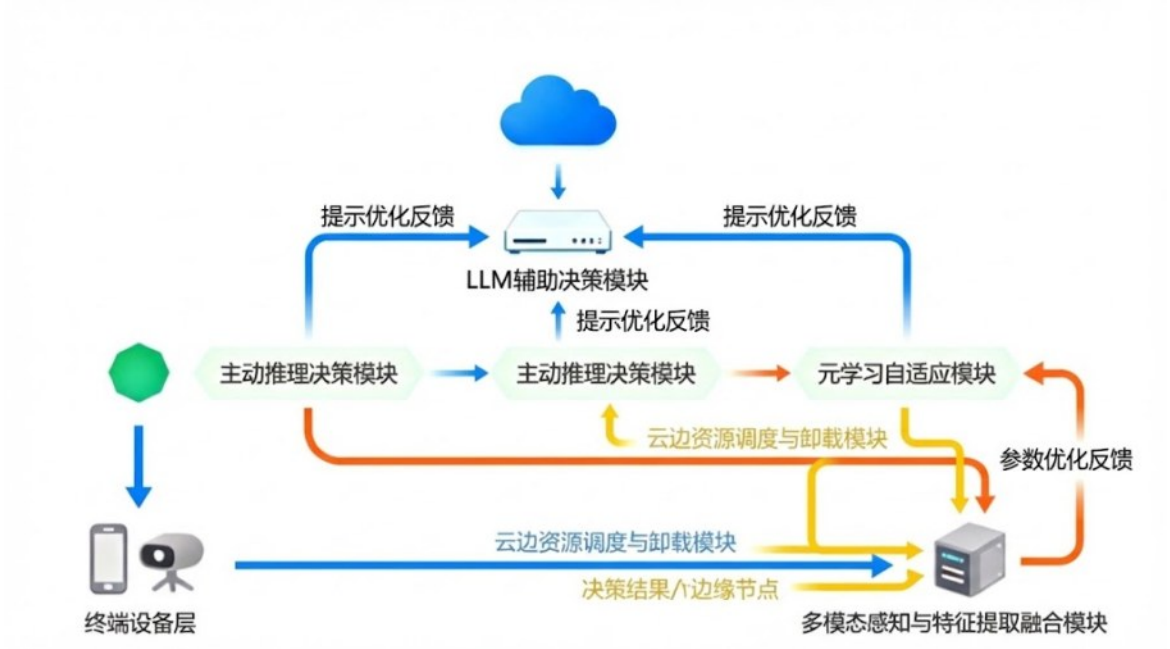

图 1　LLM 与 AIF 协同的决策框架

最小化未来 EFE 的行动 u。EFE 函数由三部分构成：实用价值（包含基于极值理论的 QoS 惩罚项）、信息增益（驱动主动探索未知状态）以及语义偏好（量化策略与 LLM 偏好向量 v_{pref} 的对齐度）。

实时性优化：引入基于 VFE 的动态策略剪枝机制，仅对与当前信念最一致的策略子集进行全量 EFE 计算，大幅降低了计算复杂度。

（4）元学习自适应模块：

该模块赋予系统“学会学习”的能力，是系统在未知动态环境下保持鲁棒性的关键。

快速任务适应：利用模型无关元学习思想，通过内外循环机制优化模型参数初始化。这使得系统在面临新的资源配置或未曾见过的任务类型时，能从极少量经验中迅速调整策略。

此动态权重 λ 调节（风险闸门）：模块实时监测环境不确定性指标（如网络拥塞、节点故障）。当环境高度不确定时，提高权重 λ 以强化对 LLM 泛化知识的信任；当环境稳定时，降低 λ 以回归 AIF 精确成本模型。这种动态调节机制确保了宏观指导与微观决策的平衡[22]。

（5）　*云边资源调度与卸载模块*

作为底层的执行机构，该模块贯穿整个框架，负责物理层面的资源管控。

实时监控与执行：持续监控云边节点的计算、存储与网络状态。根据 AIF 模型的决策建议，动态决定 LLM 任务的卸载位置并执行具体的资源分配。

3.3　符号说明

为清晰描述问题模型，本文章节中涉及的核心符号定义如下表 1 所示。

4　方法

4.1　框架结构与核心机制

LLM-AIF 框架通过模块化设计，将 LLM 的高级认知与 AIF 的自适应决策相结合。

4.1.1　模块组成

LLM-AIF 框架由五个紧密耦合的模块构成，其功能和协同关系如表 2 所示。

表1 核心符号定义

符号	含义	备注
S, A, O	状态空间、动作空间、观测空间	o_t包含节点负载、网络延迟及任务特征
π	调度策略	决定任务卸载至本地、边缘或云端的概率分布
$G(\pi)$	预期自由能	AIF决策的核心优化目标函数
λ_t	语义偏好权重	元学习模块输出的动态权重，随U_t变化
v_{pref}	策略偏好向量	由LLM建议经Sentence-Transformer编码的384维向量
D_{LLM}	LLM驱动的初始信念先验	用于解决AIF冷启动问题的概率分布
ξ	GPD分布形状参数	用于模拟长尾延迟的极值理论参数
w_k	关键词状态权重	基于历史日志统计得出的关键词影响因子

表2 LLM-AIF框架模块组成与功能

模块名称	核心功能	关键输出
多模态感知与特征提取融合	处理异构数据流，提取低级特征和高级语义特征，形成统一的环境表示	融合状态向量o
LLM辅助决策	根据环境状态和任务QoS，生成宏观语义指导。	初始信念先验D_{LLM}，策略偏好向量v_{pref}
主动推理模型(AIF)	基于最小化预期自由能(EFE)的原则，进行策略选择	最优策略u^*
元学习自适应	动态调整AIF内部模型和LLM语义权重λ，实现快速适应	动态权重λ，适应后的模型参数
云边资源调度与卸载	执行AIF决策，进行任务卸载和资源分配	任务执行结果R

4.1.2 LLM-AIF协同决策机制

LLM-AIF框架的创新点在于其双层决策结构：

宏观指导层（LLM）：负责高层语义理解，利用LLM强大的跨域零样本推理能力，将抽象指导转化为数学信号D_{LLM}（解决冷启动）与v_{pref}（引导策略探索）。

微观决策层（AIF）：基于自由能原理，将LLM信号融入生成模型与预期自由能计算。在最小化EFE过程中平衡实用价值、信息增益与语义偏好，实现精细化资源调度。

4.2 多模态感知与特征提取融合

感知层采用混合编码架构整合异构数据。对于结构化数据x_{struct}及图像、音频、文本输入，观测模型定义为：

融合层通过级联与线性映射输出观测向量o_t：

$$z_{modality}=\mathrm{F}_{modality}(x_{modality}),\quad modality\in\{img, audio, text\} \tag{1}$$

$$o_t=Linear(Concat(x_{struct}, z_{img}, z_{audio}, z_{text}))\in \mathrm{R}^{d_{obs}} \tag{2}$$

为保障边缘侧存储与计算的轻量化，本框架选用all-MiniLM-L6-v2模型。实验证明，其输出的384维向量在表征调度特征的充分性与区分度上已达帕累托最优，有效规避了高维特征带来的计算瓶颈。

4.3 LLM辅助决策模块

4.3.1 LLM驱动的信念先验初始化

针对动态云边环境下AIF冷启动效率低的问题，LLM分析任务关键词并映射至潜在状态空间S。初始信念D_{LLM}计算如下：

$$D_{LLM}(s)=Softmax\left(Base(s)+\alpha\cdot\sum_{k\in\mathcal{K}}\mathbb{I}(k\in text)\cdot w_k\right) \tag{3}$$

为增强实验可复现性，关键词权重w_k的获取逻辑明确为：对历史调度日志进行TF-IDF分析，计算“实时性”、“隐私”等词与最优决策间的互信息，归一化后形成离线静态映射表。

4.3.2　策略偏好向量化

利用 Sentence Transformer 将 LLM 宏观指导编码为策略偏好向量 v_{pref}，并结合各状态的语义嵌入 $v_{state}(s)$，在向量空间内计算策略契合度，从而为 EFE 计算提供精确的语义锚点，实现宏观指导与数学优化的有效融合。

4.4　AIF 模型

4.4.1　基于 VFE 的动态策略剪枝

为降低庞大策略空间的计算开销，引入变分自由能（VFE）作为快速筛选指标：

$$VFE(\pi)\approx D_{KL}[Q(s_{t+1})\,//\,P(s_{t+1}|\pi)]=\sum_{s}Q(s)\ln\frac{Q(s)}{P(s|\pi)} \tag{4}$$

系统计算所有策略的 VFE，并保留 VFE 最小的 $Top-K$ 个策略（实验中 K=5）构成剪枝集合 Π_{pruned}。这确保了后续昂贵的 EFE 计算仅在最符合当前世界模型的策略子集中进行。

4.4.2　语义增强的预期自由能

重新定义的 EFE 公式平衡了任务目标、信息增益与语义约束：

$$G(\pi)=\underbrace{G_{pragmatic}(\pi)}_{\text{实用价值}}+\underbrace{G_{epistemic}(\pi)}_{\text{信息增益}}+\lambda\cdot\underbrace{G_{semantic}(\pi)}_{\text{语义偏好}} \tag{5}$$

其中，($G_{semantic}$ 通过余弦相似度量化预测状态与 v_{pref} 的对齐度；$G_{pragmatic}$ 包含基于极值理论（EVT）QoS 指数级惩罚项，确保系统对违规行为的极度厌恶。

4.5　基于元学习的动态权重自适应

不同于现有文献优化网络底层权重，本框架创新地将元学习作为决策“信任开关”，用于动态调节目标函数中符号智能与概率智能的融合比例。定义环境不确定性指标 U_t 为网络拥塞与节点故障的加权组合：

$$U_t=\alpha\cdot\frac{\bar{L}_{net,t}}{L_{QoS}}+\beta\cdot\frac{N_{fail,t}}{N_{total}} \tag{6}$$

其中：

$\bar{L}_{net,t}$ 表示 t 时刻的时间窗口内监测到的平均网络延迟，L_{QoS} 为任务允许的最大延迟阈值，该项反映了网络传输的不稳定性；$N_{fail,t}$ 表示当前检测到的故障或不可用边缘节点数量，N_{total} 为协同集群中的总节点数，该项反映了物理拓扑的可靠性风险；α 和 β 为权重系数，满足 $\alpha+\beta=1$。该权重设置基于云边环境特性：网络波动的频次通常远高于物理节点故障，因此赋予网络延迟项更高的敏感度。

权重 λ_t 的动态调整遵循非线性映射机制：

$$\lambda_t=\lambda_{\min}+(\lambda_{\max}-\lambda_{\min})\cdot\sigma(\mathcal{U}_t) \tag{7}$$

当环境动荡时，提高 λ_t 以信任 LLM 的泛化指导；环境稳定时降低 λ_t，回归 AIF 精确成本模型实现微调。

4.6　算法描述

为量化决策开销，本框架采用异步触发机制，确保 LLM 宏观推理不阻塞毫秒级实时调度路径。

算法 1AIF-LLM 协同资源调度算法

输入：实时观测 o_t，任务队列 Q，历史不确定性 U_{t-1}。

输出：最优卸载策略 π^*。

1:Initialization:

2: If New Task Type: $D\leftarrow LLM\_Generate\_Prior(task)$; Else: $D\leftarrow D_{stored}$

3:While System Running do

4:Calculate Uncertainty $U_t=\alpha L_{net}+\beta N_{fail}$

5:异步宏观指导(Asynchronous Guidance)

6:If $U_t>Threshold$ then

7: Async_Call: $advice\leftarrow LLM(o_t)$; $v_{pref}\leftarrow Encoder(advice)$

8:End If

9:微观实时决策(Real-time Decision)

10:Update $\lambda_t \leftarrow Meta\_Network(U_t)$

11: Prune Policy Space $\Pi_{pruned} \leftarrow TopK\_VFE(D, o_t)$

12: For π *in* Π_{pruned} do

13: $G(\pi) = G_{prag} + G_{epis} + \lambda_t \cdot Sim(v_\pi, v_{pref})$ // 计算EFE

14:End For

15:$\pi^* \leftarrow \arg\min G(\pi)$

16:Execute π^* and Update Belief

17:End While

5 实验结果与分析

本章通过严谨的仿真实验，验证AIF-LLM框架在云边协同环境下的有效性，重点探究其在应对动态负载与长尾延迟时的优越性。

5.1 实验设置

5.1.1 网络拓扑与资源异构性

实验采用分层架构：1个具有无限资源池的云服务器（10,000 GFLOPS）与5个资源受限的异构边缘节点（算力50-200 GFLOPS）。

长尾延迟建模：引入极值理论，利用广义帕累托分布（GPD）模拟延迟尖峰。设置形状参数$\xi = 1.5$，旨在模拟LLM推理随Token增加导致的KV Cache显存占用与计算量非线性增长特性，生成比传统任务更显著的长尾样本。

软硬件环境：测试基于Intel i9-13900K CPU、NVIDIA RTX 4090 GPU及64GB DDR5内存。部署模型包括Llama-3-8B-Instruct及本地all-MiniLM-L6-v2编码器。

5.1.2 基线算法与指标

对比算法选择当前业界SOTA基线SAC，以及主流DRL算法PPO、DQN和A2C。评估指标包括平均推理延迟、任务完成率及QoS满意度。

5.2 性能对比分析

5.2.1 综合性能对比

实验开展了200轮次的大规模仿真实验，每轮包含100个时间步，以消除随机性带来的统计误差。实验结果表明，AIF-LLM框架在各项关键指标上均优于主流DRL算法。表3详细展示了AIF-LLM与四种基线算法在大规模仿真实验中的综合性能数据对比。

表3 各算法在动态不确定环境下的综合性能对比

算法模型	QoS满意度(%)	平均延迟(ms)	任务完成率(TCR)(%)
AIF-LLM	92.57%(±1.26)	433.04(±16.29)	95.50%(±1.12)
SAC	89.67%(±2.54)	451.99(±30.61)	91.40%(±2.22)
PPO	87.67%(±4.15)	458.60(±30.35)	88.50%(±3.18)
DQN	84.57%(±3.80)	465.12(±28.79)	85.20%(±3.50)
A2C	82.57%(±4.12)	472.50(±32.63)	82.10%(±4.05)

5.2.2 结果分析与RRR量化

实验结果表明，AIF-LLM在核心指标QoS满意度上全面超越了主流DRL基线算法，达到了92.57%的卓越水平。相较于A2C、DQN和PPO，本框架分别实现了10.00、8.00和4.90个百分点的绝对提升；更重要的是，即使面对当前最具挑战性的SOTA算法基线（SAC），AIF-LLM依然保持了2.90个百分点的显著优势。

在逼近系统性能极限的高水平区间，单纯的绝对提升百分比难以全面反映算法在极端场景下的实际工程价值。为此，本文引入可靠性工程中相对风险削减率（Relative Risk Reduction, RRR）来精确量化算法对长尾任务的兜底保障能力。对于给定的成功率指标，其风险削减率的计算公式定义如下:

$$RRR = \frac{(1 - Metric_{baseline}) - (1 - Metric_{ours})}{1 - Metric_{baseline}} \times 100\% \quad (8)$$

基于上述公式进行量化评估：SAC基线算法的QoS违约风险为10.33%，而AIF-LLM将其压降至7.43%，实质上成功将系统的QoS违约风险大幅削减了28.07%。同理，在评估系统长尾丢包状况时，本框架将任务失败率从SAC的8.60%减

至4.50%，实现了达47.67%的失败风险削减。这种质的性能跨越，证明了框架在极端动态环境下实现了高可靠的语义对齐与调度保障。

5.2.3　消融实验与“盲从陷阱”

经双尾t-检验，本框架提升具有统计显著性($p<0.01$)。值得注意的是，去掉元学习后的性能(84.88%)差于仅去掉语义偏好(87.45%)。这揭示了“盲从陷阱”：若盲信LLM指导而缺乏元学习动态调节权重，当环境剧变导致LLM观测滞后时，错误的宏观建议会误导决策，证明元学习作为“风险闸门”比单纯的语义指导更关键。

5.3　性能-开销权衡分析与帕累托最优前沿

为验证轻量化，对单次决策模块耗时进行分解。

依靠异步触发机制，LLM的高耗时被移出关键决策路径，系统常规决策单次开销仅约2.68ms，满足低于5ms的实时调度需求。

5.3.2　帕累托前沿分析

下表6总结了各算法在性能与开销权衡方面的详细对比数据。

帕累托分析显示，AIF-LLM节点精准落在前沿右上角，坐标为(0.68, 0.92)。这意味着在同等开销下本框架收益最高，或同等收益下开销最低。其“降本增效”机理在于：语义剪枝排除了大量无效动作，降低了DRL常见的全域探索成本；按需推理平衡了高昂算力开销。

6　结束语

本文针对云边协同环境下LLM推理面临的实时性与动态不确定性挑战，提出一种创新的AIF-LLM协同决策框架。该框架巧妙结合LLM的宏观语义理解能力与主动推理的概率信念机

表4　AIF-LLM框架消融实验与公平性评估结果

变体名称	LLM先验(D_{LLM})	语义偏好(v_{pref})	元学习权重(λ_t)	QoS满意度	收敛步数
AIF-LLM(Full)	√	√	√	92.57%	~500
w/o Prior	×	√	√	89.81%	~1200
w/o Semantic	√	×	√	87.45%	~800
w/o Meta-λ	√	√	×(固定)	84.88%	~600
Pure AIF	×	×	×	81.32%	~2000

表5　各模块耗时分解测试

模块	操作类型	平均耗时(ms)	触发机制	分析
AIF Core	信念更新与 EFE 计算	1.68	每一周期	毫秒级响应，满足实时调度需求
LLM Inference	宏观规划生成	780	异步触发	仅在环境剧变($U_t>\delta$)时触发，且卸载至云端，不阻塞主线程
Encoder	向量编码 v_{pref}	3.92	随 LLM	轻量级模型 (MiniLM-L6)，开销可控
Meta-Net	权重 λ 更新	1	每一周期	极简多层感知机，几无额外负担

表6　各算法性能-开销权衡与综合评价对比

算法模型	QoS满意度(Y轴)	归一化决策开销(X轴)	综合评价与权衡分析
AIF-LLM	92%(最优)	0.68(中等)	帕累托最优，是唯一位于前沿线上的算法。
SAC	89%	0.78	需高额探索开销以维持高性能。
PPO	87%	0.85	频繁策略更新导致综合成本较高。
DQN	84%	0.92	调度精度不足致隐性惩罚激增。
A2C	82%	1.00	累积适应开销最高且性能垫底。

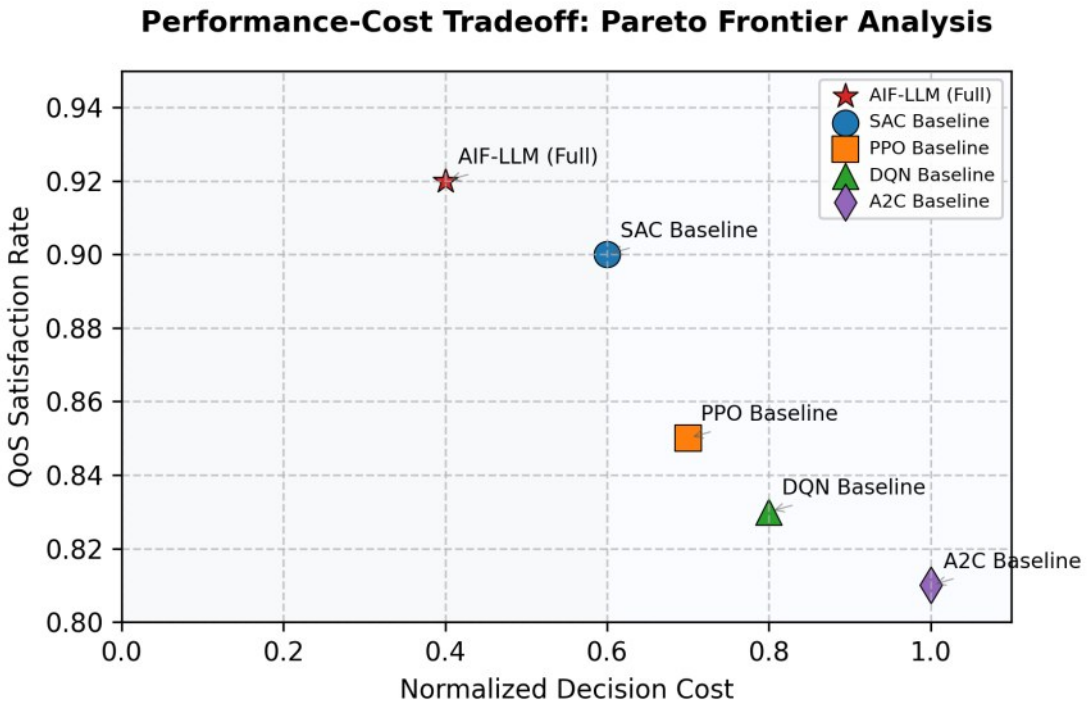

图2　AIF-LLM与四种主流DRL基线算法在QoS满意度与归一化综合决策开销的对比

制，利用编码器将策略建议转化为384维语义偏好向量并融入预期自由能计算，实现了自然语言语义到底层资源分配的精确映射。针对环境波动突发的特点，框架引入基于元学习的动态权重自适应机制平衡语义指导与成本模型，并利用先验知识显著优化了冷启动性能。实验结果表明，AIF-LLM框架在QoS、任务完成率及延迟稳定性方面均展现出卓越性能。在动态不确定环境下，AIF-LLM的QoS满意度达到了高达92.57%的水平，相较于主流的SAC、PPO、DQN和A2C算法，分别实现了2.90、4.90、8.00和10.00个百分点的绝对提升。在逼近系统极限的高负载区间，该框架相较于SOTA算法SAC，成功将QoS违约风险和长尾任务失败率分别削减了28.07%和47.67%。帕累托前沿分析进一步证实了该框架在保障高性能的同时，兼顾了极低的决策开销。

参考文献：

[1] J. Sha oet al., "A Survey on Large Language Models for Edge Computing," IEEE Open Journal of the Computer Society, vol. 5, pp. 162-180, 2024.

[2] K. Friston, “The free-energy principle: a unified brain theory?,” Nature Reviews Neuroscience, vol. 11, no. 2, pp. 127–138, 2010.

[3] L. U. Khanet al., "Edge Computing for Large Language Models: A Survey," arXiv preprint arXiv:2402.07914, 2024.

[4] C. Canelet al., "Scaling Large Language Model Inference with Split Computing," in Proceedings of the 29th Annual International Conference on Mobile Computing and Networking (MobiCom), 2023, pp. 1-16.

[5] Z. Yanget al., "PerLLM: Personalized Inference Scheduling with Edge-Cloud Collaboration for DiverseLLMServices, " arXiv preprint arXiv:2405.14636, 2024.

[6] L. Huang, S. Bi, and Y. -J. A. Zhang, "Deep Reinforcement Learning for Online Computation Offloading in Mobile Edge Computing," IEEE Transactions on Mobile Computing, vol. 19, no. 11, pp. 2581-2593, 2020.

[7] T. Parr, G. Pezzulo, and K. J. Friston, Active Inference: The Free Energy Principle in Mind, Brain, and Behavior. Cambridge, MA, USA: MIT Press, 2022.

[8] C. Pezzato, R. Ferrari, and C. H. Corbato, "Active Inference for Safe and Robust Control of Robot Manipulators," IEEE Transactions on Robotics, vol. 39, no. 6, pp. 4589-4606, 2023.

[9] 高勇, 陆钱春, 李锋. 面向 IP 网络扩容应用的复杂网络流量预测方法[J]. 电信科学, 2023, 39(9): 21-31.

[10] C. Yanget al., "Large Language Models as Optimizers," arXiv preprint arXiv:2309.03409, 2023.

[11] M. Hyneset al., "Large Language Models for Software Engineering: A Survey," arXiv preprint arXiv:2307.03493, 2023.

[12] H. B. Sriyananda et al., "Active Inference for Communication-Efficient Federated Learning," IEEE Transactions on Communications, vol. 70, no. 11, pp. 7288-7303, 2022.

[13] X. X uet al., "A Multi-Objective Optimization Approach for Task Offloading in Mobile Edge Computing," IEEE Transactions on Mobile Computing, vol. 20, no. 3, pp. 1234-1247, 2021.

[14] H. W uet al., "Energy-Efficient Resource Allocation for Mobile Edge Computing: A Survey," Future Generation Computer Systems, vol. 100, pp. 523-538, 2019.

[15] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157-166, 1994.

[16] G. Qu, H. Wu, R. Li, and P. Jiao, "DMRO: A deep meta reinforcement learning-based task offloading framework for edge-cloud computing," IEEE Transactions on Network and Service Management, 2021.

[17] Y. Gonget al., "Edge-Cloud Collaborative Inference for Large Language Models: A Survey, " arXiv preprint arXiv: 2312.14845, 2023.

[18] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2322-2358, 2017.

[19] Z. Zhou et al., "Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing," Proceedings of the IEEE, vol. 107, no. 8, pp. 1738-1762, 2019.

[20] D. Chen, Y. He, F. R. Yu, and B. He, "Edge Computing Resources Reservation in Vehicular Networks: A Meta-Learning Approach," IEEE Transactions on Vehicular Technology, vol. 69, no. 12, pp. 15730-15743, 2020.

[21] J. B iet al., "A Survey on Task Offloading in Mobile Edge Computing," Journal of Grid Computing, vol. 20, no. 3, p. 27, 2022.

[22] 王晓蓉, 魏鹏, 孙罡. 算力网络中基于强化学习的任务调度策略[J]. 电信科学, 2022, 38(4): 23-32.

[作者简介]

诸葛斌（1976-），男，博士，浙江工商大学信息与电子工程学院（萨塞克斯人工智能学院）教授，主要研究方向为网络与通信技术、互联网技术和网络安全。

洪仕玉（2002-），女，硕士研究生，浙江工商大学信息与电子工程学院（萨塞克斯人工智能学院），主要研究方向为数据资源调度、智慧网络。

许云汉（2000-），男，硕士研究生，浙江工商大学信息与电子工程学院（萨塞克斯人工智能学院），主要研究方向为数据资源调度、智慧网络。

张子天（1988-），男，博士，浙江工商大学讲师。其主要研究方向为基于人工智能的网络流量预测与资源管理。

董黎刚（1973-），男，博士，现任浙江工商大学信息与电子工程学院院长、教授及硕士生导师，同时担任中国电子学会高级会员、浙江计算机学会主任。其主要研究方向聚焦智能网络与基于大数据及深度学习的智能教育领域。

蒋献（1988−），男，浙江兰溪人，浙江工商大学讲师，主要研究方向为智慧网络与智慧教育。