面向边缘异构大语言模型可靠协作推理的自适应重复查询机制

王腾胜[①] 余 涛[②] 李济宏[①] 郑谷寒[①] 张舜卿*[①]

[①](上海大学通信与信息工程学院 上海 200444)

[②](香港大学电机与电子工程系 香港)

摘 要：在网络边缘部署大语言模型(Large Language Models, LLMs)对泛在人工智能(Artificial Intelligence, AI)至关重要。与单一LLM相比，多LLM协同推理可以显著提高系统的稳健性，但严格的资源约束对可靠推理提出了挑战。本文提出了门控自适应重复查询(Gating Adaptive Repeat Query, G-ARQ)框架，将协同推理转化为语义驱动的闭环过程，增强了系统的健壮性。在G-ARQ中，语义空间对齐和错误引导重传(Semantic-Space Alignment and Error-Guided Retransmission, SEMAR)机制将不同类型的LLM输出对齐到一个统一的语义空间，并使用推理错误信息来指导下一次重复查询LLM的选择，而轨迹优化器在不需要显式系统动态的情况下联合优化LLM门控和资源分配。在SQuAD和TriviaQA混合数据集上的实验表明，在严格的时延和能耗约束下，G-ARQ相比基线方案的准确率最大提高2.2%。G-ARQ框架建立了一个帕累托边界来表征准确率和时延之间的权衡，为可靠的第六代(Sixth-Generation, 6G)边缘智能提供了一条实用的路径。

关键词：6G边缘智能；大语言模型；协同推理；门控自适应重复查询

中图分类号：TN929.5 **文献标识码**：A **文章编号**：1009-5896(2026)00-0001-11

DOI: 10.11999/JEIT260218 **CSTR**: 32379.14.JEIT260218

1 引言

在过去的几十年里，无线网络技术在提高数据速率、连接规模和减少延迟方面不断创新。同时，人工智能(Artificial Intelligence, AI)的强大能力正在推动无线网络的智能化转型[1]。在此背景下，在网络边缘部署大语言模型(Large Language Models, LLMs)以支持资源受限的物联网(Internet of Things, IoT)用户设备(User Equipment, UE)的智能服务已成为实现环境智能和普适AI服务的关键途径[2]。随着LLM边缘部署的普及，为许多新兴应用打开了大门[3]，在未来的电子认证中，系统可利用UE提供的自然语言指令和生物信息，使用LLM生成特定的认证信息，进而通过电子验证系统完成验证，这可以提供更加灵活的体验来提升用户满意度。然而，当面对UE复杂的任务请求时，由于知识限制和固有的偏差，单一的LLM推理往往被证明是不可靠的[4]。协作地利用多个异构LLM的智慧来获得更准确的推理结果已成为一种有前途的解决方案[5]。

然而，实现可靠的协同边缘推理面临三个核心挑战。首先，IoT UE在严格的延迟和能量限制下运行[6]，这使得查询所有可用LLM的策略是不现实的，需要复杂的顺序动态决策。这个过程包括智能地选择LLM，分配发射功率，并在资源耗尽之前对结果进行集成输出以最大化任务成功概率，这对静态资源分配方法提出了挑战。其次，存在一个基本的范式冲突：像混合自动重复请求(Hybrid Automatic Repeat Request, HARQ)[7]这样的协议专注于可靠地重传比特，而智能系统必须基于语义错误来决定下一步查询哪个LLM。此外，缺乏对异构LLM输出的有效融合和动态适应机制阻碍了将不成功的尝试转化为有价值的指导，严重限制了在有限资源下有效的模型空间探索。

为了应对这些挑战，现有研究提供了部分解决方案，但仍存在重大缺陷。在通信可靠性方面，HARQ及其演进方案[8]有效地确保了比特传输，但完全不知道语义内容。在边缘智能领域，模型集成[9]和提前退出机制[10]等方法提高了推理效率，但它们大多采用静态方案，缺乏基于实时反馈的动态调整能力。新兴的语义通信[11]和面向任务的通信[12]框架开始关注语义传输，但它们主要集中在单个数据流或具有已知效用函数的预定义任务上，难以处理多个LLM输出和顺序决策的异质融合。对于复杂的动态顺序决策问题，传统的轨迹优化方法由于严重依赖于精确的系统动力学模型，往往不适用于非均匀、时变的边缘环境。虽然强化学习提供了一些模

收稿日期：2026-02-28；改回日期：2026-06-24；网络出版：2026-07-04

*通信作者： 张舜卿 shunqing@shu.edu.cn

基金项目：国家自然科学基金(62571307)，上海市科委基金(24DP1500703, 24DP1500500)，国家重点研发计划(2022YFB2902304)

Foundation Items: The National Natural Science Foundation of China(62571307), The Science and Technology Commission Foundation of Shanghai(24DP1500703, 24DP1500500), The National Key Research and Development Program of China (2022YFB2902304)

型不可知的特性[13]，但它存在样本效率低和约束满足能力差的问题，导致在实际系统中收敛缓慢。

在此背景下，本文提出了门控自适应重复查询(Gating Adaptive Repeat Query, G-ARQ)，这是一种创新框架，将协作推理重新构建为语义驱动的闭环过程，从根本上将范式从传统的比特级重传转向语义级重传。表1总结了HARQ与GARQ的区别。具体来说，HARQ主要关注比特级或传输块级可靠性，其返回确认(Acknowledgement, ACK)/否定确认(Negative Acknowledgement, NACK)反馈由链路层解码结果产生，重传对象通常为同一信息块的冗余编码比特[14]。相比之下，G-ARQ面向任务级语义可靠性，其ACK/NACK由LLM生成结果的任务完成情况决定。G-ARQ并非重复发送同一比特流，而是根据前序语义反馈重新选择合适的LLM进行查询，以获得更具互补性的语义信息。同时，G-ARQ的优化对象也从传统 HARQ中的编码率和重传次数等扩展为LLM查询序列、通信资源分配以及语义融合权重。

该框架的一个关键创新是语义空间对齐与误差引导重传(Semantic-Space Alignment and Error-Guided Retransmission, SEMAR)，这是一种通过语义空间对齐和误差导向来解决异构LLM输出的统一表示和动态引导的技术。此外，引入了一种基于零阶扩散采样的黑箱轨迹优化器，可在无需明确系统动力学模型的情况下联合优化LLM选择和资源分配。实验结果表明，在通信受限的情况下，所提出的G-ARQ框架在准确率上可提高最多2.2%，并为G-ARQ框架建立了帕累托边界分析，为资源受限的边缘环境中的可靠推理提供了有效的解决方案。

2 系统模型

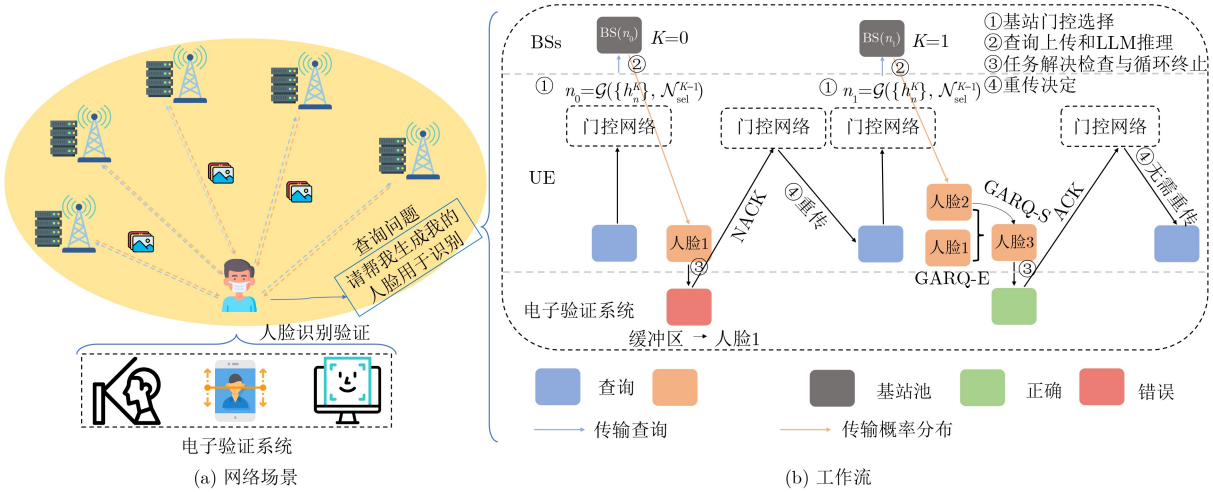
在这部分中，我们将介绍G-ARQ工作流，然后对通信资源和门控策略进行联合优化，以提高资源受限的边缘LLM协作系统的健壮性。

2.1 G-ARQ工作流

如图1(a)所示，我们考虑这样一个网络，其中有一个戴着口罩且需要进行人脸验证的UE接入多个N_{BS}基站(Base Stations, BSs)，并且在UE上部署用于选择BS的门控网络$\mathcal{G}$，每个BS与托管着人脸生成的LLM的边缘服务器共存。其核心概念是基于重复查询的协作推理过程，通过逐步查询更多的LLM来交换通信资源以获得更高的可靠性。K是当前查询尝试索引，$\mathcal{N}_{\mathrm{sel}}^{K-1}=\{n^{0:\,K-1}\}$是在先前尝试中选择的BS索引集合，其中$K_{\max}$是允许的最大尝试次数。在BS$n$，$g_n(\boldsymbol{X})=[\mathcal{P}_n^1(y^1),\mathcal{P}_n^2(y^2),\cdots,\mathcal{P}_n^T(y^T)]$，其中$\mathcal{P}_n^t(y^t)$是令牌$y^t$的令牌级概率分

表 1 HARQ与GARQ的区别

对比维度	HARQ	GARQ
可靠性目标	关注比特级可靠性，通过冗余比特重传保证数据正确解码。	关注任务级语义可靠性，通过重新查询合适的LLM提升最终任务完成质量。
错误判定单元	以传输块为判定单元，ACK/NACK通常由CRC校验产生，NACK表示解码失败。	以LLM语义输出为判定单元，ACK/NACK表示输出是否满足任务需求，NACK表示语义失败。
资源优化对象	单次链路或有限重传中的编码率、功率、重传次数等链路资源。	联合优化多轮查询中的LLM选择、通信资源分配和语义融合权重。



(a) 网络场景 (b) 工作流

图 1 G-ARQ系统架构

布，T是最大序列长度，输出令牌集合$\boldsymbol{Y}=\{y^1,y^2,\cdots,y^T\}$。这在图1(b)中详细描述的迭代循环G-ARQ工作流中实现，其中第K次查询尝试的过程如下：

(1)基站门控选择。在第K次尝试开始时，UE调用门控网络$\mathcal{G}$从未查询的BS集合$\mathcal{N}_{\mathrm{BS}}-\mathcal{N}_{\mathrm{sel}}^{K-1}$中选择BS$n^K$。门控网络$\mathcal{G}$可以利用来自$\mathcal{N}_{\mathrm{sel}}^{K-1}$的失败尝试历史，选择更加互补的LLM。此外，门控网络$\mathcal{G}$应考虑无线信道的影响，以避免超过最大时延和能耗限制。

(2)查询上传和LLM推理。UE将输入的查询$\boldsymbol{X}$上传到BSn^K，BSn^K向UE返回概率分布$g_{n^K}(\boldsymbol{X})$。

(3)任务解决检查与循环终止。在从BSn^K新查询的LLM接收到推理结果后，所选LLM集合会更新为$\mathcal{N}_{\mathrm{sel}}^{K}=\mathcal{N}_{\mathrm{sel}}^{K-1}\cup\{n^K\}$。然后UE尝试使用最新输出解决任务。我们提出了两种生成该输出的配置，稍后将进行介绍。循环的终止由电子验证系统来自动验证输出。如果正确，它将返回ACK；如果不正确，它将返回NACK。

(4)重传决定。如果收到NACK，路由器检查当前查询次数K是否达到$K_{\max}$。如果不是，则该过程返回到步骤(1)以进行下一次重传；否则，它以失败而终止。

所描述的G-ARQ工作流为可靠的边缘推理建立了一个系统框架，为了满足不同的应用需求，GARQ 支持两种生成每次尝试最终输出的替代配置：(a)GARQ集成输出(GARQ with Ensemble, GARQ-E)：最终输出是从由$\mathcal{N}_{\mathrm{sel}}^{K}$中的所有LLM输出的概率分布中以动态的权重进行融合；(b)GARQ单一输出(GARQ with Single LLM, GARQ-S)：输出完全依赖于新查询的LLMn^K，而之前尝试的LLM输出$\mathcal{N}_{\mathrm{sel}}^{K-1}$仅用于门控网络$\mathcal{G}$。

2.2 通信模型

我们考虑无线上行链路场景，其中UE向一个或多个BS发送查询。对于第K次重传尝试BSn^K，上行链路等待时间是$D^K=1/R^K(n^K)$，其中$R^K(n^K)=B\log_2\left(1+p^K/\sigma^2 h_n^K\right)$是可实现的信道容量。这里，$B$是每个BS分配的带宽，$p^K\geq 0$是发射功率，$\sigma^2$是噪声功率谱密度，$h_n^K$是UE和BS$n$之间的有效信道增益。假设该信道增益在单次查询尝试期间保持准静态，但在重传期间可能变化。

端到端时延包括传输时延、传播时延和BS处理时延。我们假设BS和边缘服务器有足够的计算和下行资源，而UE的上行链路功率是严格有限的。因此，K次尝试后的累计时延和能耗为：

$$D_{\mathrm{total}}^K=\sum_{k=0}^{K}D^k \tag{1}$$

$$E_{\mathrm{total}}^K=\sum_{k=0}^{K}E^k(\{n^k,p^k\})=\sum_{k=0}^{K}\left(\eta\cdot p^k\cdot D^k+p_C\cdot D^k\right) \tag{2}$$

其中η是功率放大器的效率，p_C表示传输过程中的固定电路功耗。

3 问题建模

设K^*表示任务成功所需的随机传输次数。通过联合优化重传策略，建立了通信约束下最小化失败概率任务$\Pr(K^*>K_{\max})$的核心优化问题。可靠推理问题被表示为：

$$\begin{aligned}\mathcal{P}_0:\ &\underset{\{n^K,p^K\}_{K=0}^{K_{\max}}}{\text{minimize}}\Pr\left(K^*>K_{\max}\right),\\ \text{s.t. C1:}\ &n^K=\mathcal{G}\left(\{h_n^K\},\mathcal{N}_{\mathrm{sel}}^{K-1}\right),\\ &\text{C2:}\ D_{\mathrm{total}}^{K_{\max}}\leq D_{\max},\\ &\text{C3:}\ E_{\mathrm{total}}^{K_{\max}}\leq E_{\max},\\ &\text{C4:}\ 0\leq p^K\leq p_{\max}.\end{aligned} \tag{3}$$

其中，$\mathcal{G}$是基于h_n^K和$\mathcal{N}_{\mathrm{sel}}^{K-1}$选择下一次BS$n^K$的门控函数，$D_{\max}$和$E_{\max}$表示最大等待时延和能量消耗，$p_{\max}$是最大发射功率。

问题$\mathcal{P}_0$旨在最大化$K_{\max}$重传下由最坏情况时延和能量约束所定义的严格资源上限内的任务成功率。然而，它面临着以下挑战：对其黑盒混合动力学建模的不可行性，其非凸、受限的顺序决策的联合复杂性，以及从比特级到语义级可靠性所必需的范式转换。

4 算法设计

第2节和第3节我们建立了G-ARQ框架和问题建模，本节我们将介绍针对G-ARQ框架设计的求解算法，算法框架如图2所示。首先先讨论语义对齐并实现差错重传，在此基础上，使用处理黑箱动力学的轨迹优化器，通过联合优化通信约束下的基站门控和功率分配来解决资源分配问题。

4.1 SEMAR：语义空间对齐和错误引导重传

在这一部分中，我们提出了G-ARQ框架的概率重构，引入了统一的语义表示来对齐异构LLM输出，并引入了基于潜在误差的自适应重传策略优化的期望最大化(Expectation-Maximization, EM)方法。

4.1.1 异构LLMs的语义空间对齐

我们框架中的一个关键挑战是独立训练的LLM之间固有的词汇空间$\mathcal{V}_n$和表示空间不匹配，这阻碍了直接输出集成。我们通过[15]提出的相对表示方法来解决这个问题。具体地，每个LLM的输出令牌级概率向量$\mathcal{P}_n^t$经由投影矩阵$\boldsymbol{\Lambda}_n\in\mathbb{R}^{\mathcal{V}_n\times\mathcal{A}}$映

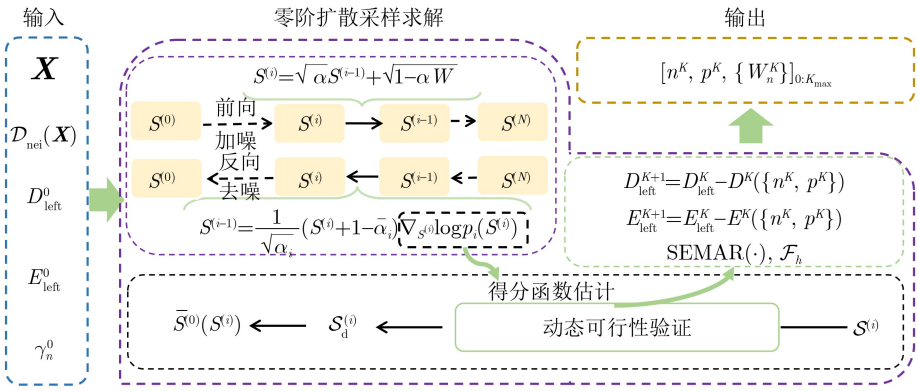


图 2 所提算法框图

射到可比向量，其中$\mathcal{A}$表示基于锚的共享词汇空间，由所有LLM词汇表的交集得到。

$$\boldsymbol{\lambda}_n^t = \mathcal{P}_n^t \cdot \boldsymbol{\Lambda}_n \tag{4}$$

$\boldsymbol{\Lambda}_n$的每个元素都封装了$\mathcal{V}_n$中的一个词元与$\mathcal{A}$中每个词元之间的语义亲和力，该亲和力通过余弦相似度来衡量，具体来说，$\boldsymbol{\Lambda}_n$的第i行$\Lambda[i]$由式(5)得到。该投影将模型特定的置信度分数转换为统一的、可比较的表示$\boldsymbol{\lambda}_n^t$。

$$\Lambda[i] = (\cos(e_{w^{(i)}}, e_{a^{(1)}}), \cdots, \cos(e_{w^{(i)}}, e_{a^{(|\mathcal{A}|)}})), \tag{5}$$

其中，w是固有的词汇空间$\mathcal{V}_n$的token，而a是共享锚点词汇集$\mathcal{A}$的token。

4.1.2 基于潜在误差引导的概率重构

为了解决原始问题的挑战，我们将路由重新表述为概率推理。我们的目标是推断潜在的校正变量$\boldsymbol{z}^K$，引导后续的LLM选择朝向与当前误差正交的输出，并避免重复错误。

初始步骤涉及将用户输入$\boldsymbol{X}$和LLM输出两者映射到共享锚令牌集合$\mathcal{A}$。$\boldsymbol{X}$被表示为$\boldsymbol{\psi} = \mathcal{P}_{\boldsymbol{X}} \cdot \Lambda_{\text{ref}}$，其中$\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{v}) = \mathbb{I}(\boldsymbol{X}^t = \boldsymbol{v})$定义映射函数。相应地，如公式(4)所示，每个LLM的输出被表示为$\boldsymbol{\lambda}_n = \mathcal{P}_{\boldsymbol{Y}_n} \cdot \Lambda_n$。

然后，我们建立了一个概率模型来刻画K次尝试后的系统状态。概率图模型由所有LLM输出的相对表示$\{\boldsymbol{\lambda}_n\}$和第$K$次尝试的理想输出$\boldsymbol{z}^K$定义，其联合概率分布由下式给出，

$$\begin{aligned}&\Pr(\boldsymbol{z}^K, \{\boldsymbol{\lambda}_i\} | \boldsymbol{\psi}, \boldsymbol{z}^{K-1}) \\ &= \frac{1}{\Gamma} \exp\left(-\sum_{n=1}^{N_{\text{BS}}} \theta_n^K \|\boldsymbol{\lambda}_n - \boldsymbol{z}^K\|^2 \right. \\ &\quad \left. - \sum_{n=1}^{N_{\text{BS}}} \gamma_n^K \left\langle \boldsymbol{\lambda}_n - \boldsymbol{z}^K, \boldsymbol{z}^{K-1} - \boldsymbol{z}^K \right\rangle \right)\end{aligned} \tag{6}$$

其中，Γ是对数配分函数。θ_n^K表示第n个LLM的输出与理想输出$\boldsymbol{z}^K$的差异，γ_n^K捕获其误差方向相对于前一个错误输出方向$\boldsymbol{z}^{K-1}$的相关性。

作为一种计算工具，该联合概率模型能够推断出最优的门控策略。它与原问题$\mathcal{P}_0$的联系在于任务成功概率$\mathbb{P}(K^* \le K_{\max})$取决于路由策略$\mathcal{F}$，其有效性由参数$\{\theta_n^K\}$和$\{\gamma_n^K\}$决定。

4.1.3 基于期望最大化的参数学习

最优基站门控决策是通过最大化联合分布的可能性来获得的，提供了对原始目标的直接优化的替代方案。公式(6)中的联合似然的最优参数是通过求解以下问题来确定的。

$$\begin{aligned}\mathcal{P}_1\colon &\underset{\{\boldsymbol{z}^K, \{\theta_n^K\}, \{\gamma_n^K\}\}}{\text{minimize}} L(\boldsymbol{z}^K, \{\theta_n^K\}, \{\gamma_n^K\}), \\ \text{s.t. } &\text{C5}\colon \gamma_n^0 = 0, \\ &\text{C6}\colon \theta_n^K > 0, \\ &\text{C7}\colon 0 < \gamma_n^K \le \theta_n^K.\end{aligned} \tag{7}$$

其中$L(\boldsymbol{z}^K, \{\theta_n^K\}, \{\gamma_n^K\})$表示负对数似然，定义为，

$$\begin{aligned}L(\boldsymbol{z}^K, \{\theta_n^K\}, \{\gamma_n^K\}) = &\sum_{n=1}^{N_{\text{BS}}} \theta_n^K \|\boldsymbol{\lambda}_n - \boldsymbol{z}^K\|^2 \\ &+ \sum_{n=1}^{N_{\text{BS}}} \gamma_n^K \left\langle \boldsymbol{\lambda}_n - \boldsymbol{z}^K, \boldsymbol{z}^{K-1} - \boldsymbol{z}^K \right\rangle\end{aligned} \tag{8}$$

在问题$\mathcal{P}_1$中，C5指示在缺乏先验方向知识的情况下，为初始查询选择具有最小几何距离的LLM。C6确保概率分布是有效的。同时，C7用于抑制表现出类似于已经失败的那些错误模式的LLM。我们使用EM算法来解决问题$\mathcal{P}_1$：

E步：给定当前参数估计$\{\theta_n^{K,(i)}\}$和$\{\gamma_n^{K,(i)}\}$，更新$\boldsymbol{z}^K$的估计，

$$\boldsymbol{z}^{K,(i)} = \frac{\sum_{n=1}^{N_{\text{BS}}} \left[(2\theta_n^{K,(i)} + \gamma_n^{K,(i)}) \boldsymbol{\lambda}_n + \gamma_n^{K,(i)} \boldsymbol{z}^{K-1} \right]}{2 \sum_{n=1}^{N_{\text{BS}}} \left(\theta_n^{K,(i)} + \gamma_n^{K,(i)} \right)} \tag{9}$$

M步：保持$\boldsymbol{z}^{K,(i)}$固定，通过以下方式更新参数$\{\theta_n^K\}$和$\{\gamma_n^K\}$，

$$\begin{aligned}&\underset{\{\theta_n^K\},\{\gamma_n^K\}}{\text{minimize}} L(\boldsymbol{z}^{K,(i)},\{\theta_n^K\},\{\gamma_n^K\}),\\&\text{s.t.(C5)}-\text{(C7)}\end{aligned}\tag{10}$$

EM算法迭代直到收敛，此时参数在约束下最大化联合似然。

为了估计查询前参数$\boldsymbol{\lambda}_n$，我们通过基于相似性的聚集来估计质量度量$\{\theta_n^K\}$和$\{\gamma_n^K\}$。对于输入查询$\boldsymbol{X}$对应表示的$\boldsymbol{\psi}$，通过欧几里德距离从本地数据库检索相似查询$\mathcal{D}_{\text{nei}}(\boldsymbol{X})$对应表示的$\mathcal{D}_{\text{nei}}(\boldsymbol{\psi})$。然后将这些参数计算为其历史值的邻域平均值：

$$\theta_n^K(\boldsymbol{\psi})=\frac{1}{N_0}\sum_{\boldsymbol{\psi}'\in\mathcal{D}_{\text{nei}}(\boldsymbol{\psi})}\overline{\theta}_n^K(\boldsymbol{\psi}')\tag{11}$$

$$\gamma_n^K(\boldsymbol{\psi})=\frac{1}{N_0}\sum_{\boldsymbol{\psi}'\in\mathcal{D}_{\text{nei}}(\boldsymbol{\psi})}\overline{\gamma}_n^K(\boldsymbol{\psi}')\tag{12}$$

在第K次重传时选择BSn^K之后，我们更新$\mathcal{N}_{\text{sel}}^K=\mathcal{N}_{\text{sel}}^{K-1}\cup\{n^K\}$并计算，

$$\boldsymbol{z}^K=\sum_{n\in\mathcal{N}_{\text{sel}}^K}w_n^K\boldsymbol{\lambda}_n+w_0^K\boldsymbol{z}^{K-1},\tag{13}$$

$$w_n^K=\frac{\overline{w}_n^K}{\sum_{n'\in\mathcal{N}_{\text{sel}}}\overline{w}_{n'}^K+\overline{w}_0^K}\tag{14}$$

其中$\{\overline{w}_n^K\}$表示在N_{BS}中的所有BS都被激活的条件下的最优集成权重，

$$\overline{w}_n^K=\begin{cases}\dfrac{2\theta_n^{K,*}+\gamma_n^{K,*}}{2\sum_{n=1}^{N_{\text{BS}}}\left(\theta_n^{K,*}+\gamma_n^{K,*}\right)},n\neq 0,\\[2ex]\dfrac{\sum_{n=1}^{N_{\text{BS}}}\gamma_n^{K,*}}{2\sum_{n=1}^{N_{\text{BS}}}\left(\theta_n^{K,*}+\gamma_n^{K,*}\right)},n=0.\end{cases}\tag{15}$$

4.2 基于零阶扩散采样的联合基站选择与资源分配

4.1节中所提出的SEMAR机制实现了最佳基站门控。然而，在实际通信约束下，基站门控和功率分配必须联合优化。 本部分介绍了一个基于模型的扩散框架，该框架将顺序决策建模为轨迹采样来求解。

4.2.1 作为轨迹规划的联合优化

在G-ARQ框架中，联合优化基站门控和功率分配是一个顺序决策问题，自然地被描述为一个以最大化时延和能量约束下的任务成功概率为目标的轨迹优化问题。

我们定义决策变量$\boldsymbol{S}=\{D_{\text{left}}^K,E_{\text{left}}^K,\{h_n^K\},\{\overline{w}_n^K\},\boldsymbol{z}^K,n^K,p^K\}_{K=0}^{K_{\max}}$，其中$D_{\text{left}}^K$和$E_{\text{left}}^K$分别是第$K$次尝试后的剩余时延和能量预算。然后将G-ARQ框架中的联合门控和功率分配问题表示为黑盒轨迹优化问题：

$$\begin{aligned}\mathcal{P}_2\colon&\underset{\boldsymbol{S}}{\text{minimize}}J(\boldsymbol{S})=\mathcal{F}(D_{\text{left}}^{K_{\max}},E_{\text{left}}^{K_{\max}})\\&\quad+\sum_{K=0}^{K_{\max}}\sum_{n\in\mathcal{N}_{\text{sel}}^K}\overline{w}_n^K,\\\text{s.t. C8}\colon&\ D_{\text{left}}^0=D_{\max},E_{\text{left}}^0=E_{\max},\gamma_n^0=0,\\\text{C9}\colon&\ n^{K+1}\in\mathcal{N}_{\text{BS}}-\mathcal{N}_{\text{sel}}^K,p^K\leq p_{\max},\\\text{C10}\colon&\ D_{\text{left}}^{K+1}=D_{\text{left}}^K-D^K(\{n^K,p^K\}),\\\text{C11}\colon&\ E_{\text{left}}^{K+1}=E_{\text{left}}^K-E^K(\{n^K,p^K\}),\\\text{C12}\colon&\ \mathcal{N}_{\text{sel}}^{K+1}=\mathcal{N}_{\text{sel}}^K+n^{K+1}\\\text{C12}\colon&\ \{h_n^{K+1}\}=\mathcal{F}_h(\{h_n^K\}),\\\text{C13}\colon&\ \{\{\overline{w}_n^{K+1}\},\boldsymbol{z}^{K+1}\}=\text{SEMAR}(\boldsymbol{z}^K)\end{aligned}\tag{16}$$

这里，$\mathcal{F}(D_{\text{left}}^{K_{\max}},E_{\text{left}}^{K_{\max}})=\rho[\min\{D_{\text{left}}^{K_{\max}},0\}+\min\{E_{\text{left}}^{K_{\max}},0\}]$是惩罚函数，$\rho>0$，$\mathcal{F}_h$是黑盒信道动态函数，SEMAR$(\cdot)$是4.1节所提出的SEMAR方案。

问题$\mathcal{P}_2$的一个重要挑战源于信道动态$\mathcal{F}_h(\cdot)$和语义动态SEMAR$(\cdot)$的黑箱性质，即这些动态在分析上是未知的。传统的轨迹优化方法在这种环境下难以适用，因为缺乏明确的状态转移函数。另一方面，在基于强化学习的方法中，信道动态和语义动态通常被视为状态转移概率的一部分，该概率通过神经网络进行逼近。然而，即便是对于简单任务，这通常也需要数百万次的数据交互和学习步骤。

4.2.2 基于扩散采样的求解方法

为了解决上述问题，我们使用基于扩散的规划器从目标分布$p_0(\boldsymbol{S})\propto p_{\text{d}}(\boldsymbol{S})p_{\text{J}}(\boldsymbol{S})p_{\text{c}}(\boldsymbol{S})$中采样轨迹，该目标分布编码了动态可行性$p_{\text{d}}(\boldsymbol{S})$，最优性$p_{\text{J}}(\boldsymbol{S})$和约束$p_{\text{c}}(\boldsymbol{S})$，$\mathbb{I}[\cdot]$是狄拉克增量函数：

$$\begin{aligned}p_0(\boldsymbol{S})\propto p_{\text{d}}(\boldsymbol{S})\cdot p_{\text{J}}(\boldsymbol{S})\cdot p_{\text{c}}(\boldsymbol{S})\propto\prod_{K=0}^{K_{\max}}\mathbb{I}\left[(\text{C10})\sim(\text{C13})\right]\\\cdot\exp\left(-\frac{J(\boldsymbol{S})}{\tau}\right)\cdot\prod_{K=0}^{K_{\max}}\mathbb{I}\left[(\text{C8})\sim(\text{C9})\right]\end{aligned}\tag{17}$$

现有理论[16]证明，当$\tau\to 0$时，从$p_0(\boldsymbol{S})$采样等价于求解$\mathcal{P}_2$。然而，由于$p_0(\boldsymbol{S})$的高维性和稀疏性，直接采样是不可行的。因此，我们采用一个扩散过程，将样本从简单分布逐步去噪回目标分布。

扩散过程包括一个正向过程和一个反向过程。前向过程在N个步长上用高斯噪声逐步地加噪初始分布$p_0(\cdot)$，将其变换为各向同性的高斯分布$p_N(\cdot)$，其中每个步长由$p_{i|i-1}(\cdot|\boldsymbol{S}^{(i-1)})=\mathcal{N}(\sqrt{\alpha_i}\boldsymbol{S}^{(i-1)},(1-\alpha_i)\boldsymbol{I})$定义，其中$\mathcal{N}(\cdot)$表示高斯分布，$\alpha_i\in(0,1)$是比例因子。反向过程通过迭代地将$p_N(\cdot)$去噪回目标分布。每个反向步长由分数函数上的梯度上升步长来近似：

$$
\begin{aligned}
S^{(i-1)} &= \frac{1}{\sqrt{\alpha_i}}\left(S^{(i)} + (1-\overline{\alpha}_i)\nabla_{\boldsymbol{S}^{(i)}} \ln p_i(\boldsymbol{S}^{(i)})\right) \\
&= \frac{1}{\sqrt{\alpha_i}}\left(S^{(i)} + (1-\prod_{k=1}^{i}\alpha_k)\nabla_{\boldsymbol{S}^{(i)}} \ln p_i(\boldsymbol{S}^{(i)})\right)
\end{aligned} \tag{18}
$$

我们通过使用模型指导的先验$p_0(\boldsymbol{S}^{(0)})$和贝叶斯法则来估计得分函数：

$$
\begin{aligned}
&\nabla_{\boldsymbol{S}^{(i)}} \ln p_i(\boldsymbol{S}^{(i)}) \\
&= \frac{\nabla_{\boldsymbol{S}^{(i)}} p_i(\boldsymbol{S}^{(i)})}{p_i(\boldsymbol{S}^{(i)})} \\
&= -\frac{\boldsymbol{S}^{(i)}}{1-\overline{\alpha}_i} + \frac{\sqrt{\overline{\alpha}_i}}{1-\overline{\alpha}_i} \\
&\quad \cdot \frac{\displaystyle\int \boldsymbol{S}^{(0)} p_{i|0}(\boldsymbol{S}^{(i)} \mid \boldsymbol{S}^{(0)}) p_0(\boldsymbol{S}^{(0)}) d\boldsymbol{S}^{(0)}}{\displaystyle\int p_{i|0}(\boldsymbol{S}^{(i)} \mid \boldsymbol{S}^{(0)}) p_0(\boldsymbol{S}^{(0)}) d\boldsymbol{S}^{(0)}}
\end{aligned} \tag{19}
$$

然后采用蒙特卡洛方法来进一步估计得分函数：

$$
\begin{aligned}
\nabla_{\boldsymbol{S}^{(i)}} \ln p_i(\boldsymbol{S}^{(i)}) &\approx -\frac{\boldsymbol{S}^{(i)}}{1-\overline{\alpha}_i} + \frac{\sqrt{\overline{\alpha}_i}}{1-\overline{\alpha}_i} \underbrace{\frac{\displaystyle\sum_{\mathrm{S}^{(0)}\in\mathcal{S}^{(i)}} \boldsymbol{S}^{(0)} p_0(\boldsymbol{S}^{(0)})}{\displaystyle\sum_{\mathrm{S}^{(0)}\in\mathcal{S}^{(i)}} p_0(\boldsymbol{S}^{(0)})}}_{\text{蒙特卡洛估计}} \\
&\triangleq -\frac{\boldsymbol{S}^{(i)}}{1-\overline{\alpha i}} + \frac{\sqrt{\overline{\alpha i}}}{1-\overline{\alpha i}} \overline{\mathrm{S}}^{(0)}(\mathcal{S}^{(i)})
\end{aligned} \tag{20}
$$

式中，$\mathcal{S}^{(i)}$是$\phi_i(\boldsymbol{S}^{(0)}) = \mathbb{N}\left(\boldsymbol{S}^{(i)}/\sqrt{\overline{\alpha}_i}, \boldsymbol{I}/\overline{\alpha} - \boldsymbol{I}\right)$的一批样本。

为了处理$\phi_i(\cdot)$中由于动态可行性约束而产生的低似然样本，我们实施了一种策略。首先从$\phi_i(\cdot)$中提取候选样本，然后使用它们对应的控制序列$\boldsymbol{U} = \{n^K, p^K\}_{K=0}^{K_{\max}}$通过系统动力学C10～C13传播候选样本，以获得动态可行子集$\mathcal{S}_{\mathrm{d}}^{(i)}$。为每个验证样本$\boldsymbol{S}^{(0)}$分配权重$w(\boldsymbol{S}^{(0)}) = p_{\mathrm{J}}(\boldsymbol{S}^{(0)}) p_{\mathrm{c}}(\boldsymbol{S}^{(0)})$，然后对定义在公式(20)中使用的样本集进行加权计算：

$$
\overline{\boldsymbol{S}}^{(0)}(\mathcal{S}^{(i)}) = \frac{\displaystyle\sum_{\boldsymbol{S}^{(0)}\in\mathcal{S}_{\mathrm{d}}^{(i)}} \boldsymbol{S}^{(0)} w(\boldsymbol{S}^{(0)})}{\displaystyle\sum_{\boldsymbol{S}^{(0)}\in\mathcal{S}_{\mathrm{d}}^{(i)}} w(\boldsymbol{S}^{(0)})} \tag{21}
$$

4.2.3 所提扩散采样轨迹优化器的时间复杂度分析

设共享锚点语义空间维度为A，EM迭代次数为I_{EM}。G-ARQ的复杂度主要由SEMAR更新和扩散轨迹优化两部分构成。SEMAR中的投影矩阵可离线预计算，在线阶段主要进行语义表示更新、误差方向估计和权重更新，其复杂度约为$O(I_{\mathrm{EM}}K_{\max}A + K_{\max}N_{\mathrm{BS}}A)$。扩散轨迹优化器是主要计算开销来源。每个采样轨迹包含轮基站选择与功率分配，单轮状态传播、约束检查和效用评估的复杂度为N_{BS}，则经过N个扩散步和M个蒙特卡洛样本后，其复杂度为$O(NMK_{\max}N_{\mathrm{BS}})$，因此，G-ARQ的主导复杂度可表示为$O(I_{\mathrm{EM}}K_{\max}A + K_{\max}N_{\mathrm{BS}}A + NMK_{\max}N_{\mathrm{BS}}) \approx O(NMK_{\max}N_{\mathrm{BS}})$。

5 数值结果

我们考虑一个400 m×400 m的区域，有五个基站。UE以1.1 m/s的速度移动，最大发射功率为27 dBm。系统带宽为20 MHz，噪声功率谱密度为–174 dBm/Hz，信道遵循35+35lg(d)路径损耗模型，其中d表示UE和BS之间的距离。对于UE的人脸生成任务，由于数据集缺失，为了方便起见，我们使用SQuAD(阅读理解)和TriviaQA(知识密集型问答)的混合数据集来进行验证，共2000个样本，其中SquAD和TriviaQA分别为1 000个样本，通过精确匹配准确率来评估任务成功率。我们使用五个性能相当的LLM：Mistral-7B, Vicuna-7B, Nous-Capybala-7B, Gemma-7B和Llama-2-7B，详细实验参数见表2所示。

5.1 组件级评估

(1)SEMAR的有效性

图3(a)说明了SEMAR的校正机制。当来自第$(K-1)$次重传的语义方向$\boldsymbol{z}^{K-1}$偏离真实方向$\boldsymbol{z}^*$时，SEMAR使用误差正交化来指导用于第K次重传的基站选择。最终结合来自BS0的语义方向$\boldsymbol{\lambda}_0$产生新的方向$\boldsymbol{z}^K$，该方向与$\boldsymbol{z}^*$更接近。

通过比较两种方案：(a)Top-k门控+平均权重，依次选择初始权重$\{\theta_n^0\}$最高的前K个基站，并使用平均权重进行集成输出；(b)SEMAR门控+平均权重，该方案使用所提出的SEMAR进行门控，但保留了平均权重；如图3(b)所示，当$K=0$时，准确率达到65.60%，表明很大部分简单请求无需重传即可满足用户需求。SEMAR门控+平均权重方案比Top-k门控+平均权重方案的平均精度提高了

表 2 仿真参数设置

参数名称	参数值
基站个数(N_{BS})	5
带宽(B)	20 MHz
最大发射功率($p_{\max}$)	27 dBm
路径损耗	35+35lg(d)
噪声功率谱密度(σ^2)	–174 dBm/Hz
扩散步数(N)	1000
蒙特卡洛采样样本数(M)	8192
最大重查询次数($K_{\max}$)	2

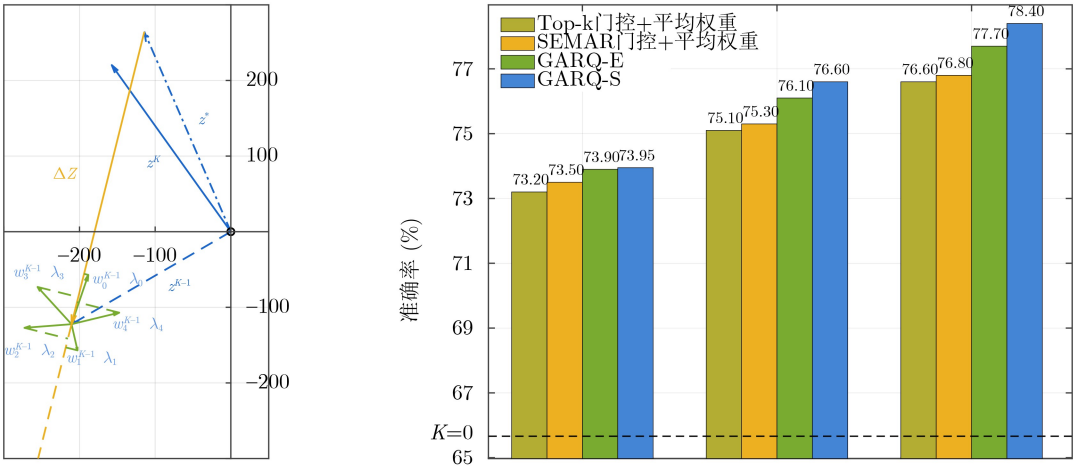
0.23%，证明了SEMAR误差引导门控的有效性。GARQ-E在SEMAR门控+平均权重方案的基础上进一步提高了0.7%的平均精度，突出了SEMAR动态调整集成权重的显著增益。此外，顺序查询中的信息互补增益与错误语义传播之间也存在权衡。增加查询次数能够引入不同LLM的互补知识，但若前序错误输出被静态融合到最终结果中，错误语义也可能随之传播，从而削弱甚至抵消多模型协同收益。GARQ-S仅利用历史输出进行门控决策，而不直接融合历史答案，因此能够避免部分错误语义污染最终输出。相比之下，GARQ-E虽然仍采用融合输出，但SEMAR会根据前序失败反馈动态调整模型选择和融合权重来抑制错误语义传播，因此相比平均融合等静态策略仍具有性能增益。

(2)黑箱轨迹优化器的有效性

为了在G-ARQ框架内评估黑盒轨迹优化器的有效性，我们与两种方案进行了比较，均采用注水法分配功率来满足能量约束：(a)信道贪婪：其在每次重传时选择具有最佳信道条件的基站；(b)模型贪婪：其在每次重传时选择具有最高权重$\max\{w_n^K\}$条件的基站。如图4所示，由于信道贪婪方案每次都选择信道最好的基站，所以有着最低的时延；而模型贪婪由于选择权重最高的基站可能信道条件很差导致其不满足时延约束；由于GARQ-E和GARQ-S只在最终输出上有区别，所以它们有着一致的时延消耗。GARQ-E和GARQ-S在严格遵守时延约束的同时实现了模型优先策略的接近最优的准确率，验证了黑盒轨迹优化器的有效性。图5通过跟踪在第K次重传时倾向于选BS0的概率偏好$P(\boldsymbol{S}^{(i)}(n^K=0))$和$J(\boldsymbol{S})$随着扩散步骤的演变来验证所提出的优化器的收敛，可以看出二者均呈现收敛趋势。

5.2 系统级性能

我们在不同的时延和能耗约束下对G-ARQ进行了评估，比较了两种方案来说明系统级的有效性：(1)基线1：模型贪婪选择+平均权重+近端策略优化(Proximal Policy Optimization, PPO)功率优化，首先在固定基站选择序列下通过PPO优化功率，然后在此基础上通过模型贪婪策略选择能满足系统约束的权重累加最高的基站序列，最后通过平均权重进行集成输出；(2)基线2：模型贪婪选



(a) SEMAR的工作机制　　(b) 不同方案下的准确率随K的变化

图 3 评估SEMAR有效性的实验结果

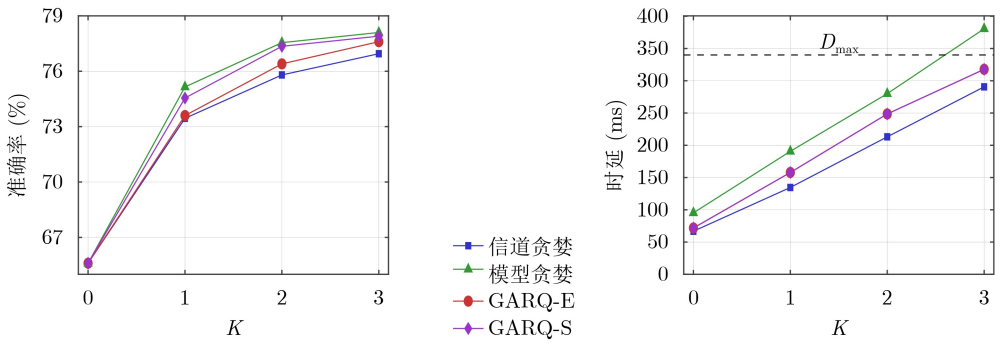


图 4 不同方案下的准确率、时延随K的变化

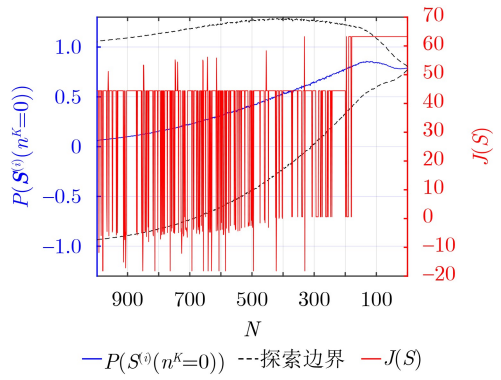


图 5 所提优化算法的收敛性

择+平均权重+模拟退火功率优化，首先在固定基站选择序列下通过模拟退火算法优化功率，然后在此基础上通过模型贪婪策略选择能满足系统约束的权重累加最高的基站序列，最后通过平均权重进行集成输出。

如图6和图7所示，在不同的时延和能耗约束下，所提方案GARQ-E和GARQ-S均优于基线1和基线2，显著提高了准确性。在开头区域，也就是时延和能耗限制比较严格的区域，增益是由于联合优化功率和基站选择以及SEMAR的动态权重共同带来的；而在结尾区域，也就是资源限制比较宽松的区域，是由于SEMAR的动态权重所带来的性能增益。另外，在更为严格的时延和能耗约束下，所提出算法的性能提升最为显著，在$K_{\max}=1$时最大增益达到2.2%，而在$K_{\max}=2$时最大增益为1.9%。这表明G-ARQ框架具有很强的适用性，并为未来的低时延、高可靠性系统建立了一个帕累托边界。

如表3所示，我们对比了在不同数据集下的泛化性，可以看出，GARQ-E和GARQ-S在所有四个基准测试上都取得了最佳的准确率，包括 GSM8K (数学推理)、TriviaQA、ARC-C(开放知识)和SquAD。值得注意的是，在GSM8K数据集上，GARQ-E方案明显优于GARQ-S方案。我们的分析表明，数学推理任务对逻辑一致性要求非常高。GARQ-E通过 SEMAR 机制整合了多个LLM的共识，有效纠正了单个模型的逻辑偏差或计算错误，从而获得更可靠的结果。相反，在TriviaQA、SquAD以及ARC-C等需要广泛覆盖事实性和开放性知识的任务中，GARQ-S的表现更接近甚至在某些方面超过GARQ-E。这表明此类任务可能要求 LLM 具有独特的知识多样性，而简单的共识融合有时可能会抹平关键的非共识正确答案。如表4所示，我们对比了不同方案在NVIDIA RTX A6000上的实际运行时间，可以看出，所提出方案的计算复杂度相对于基线方案有着很好的优势。

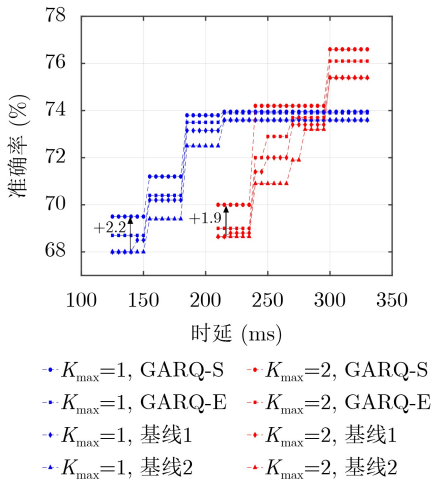


图 6 不同方案和$K_{\max}$下的准确率随最大时延约束的变化

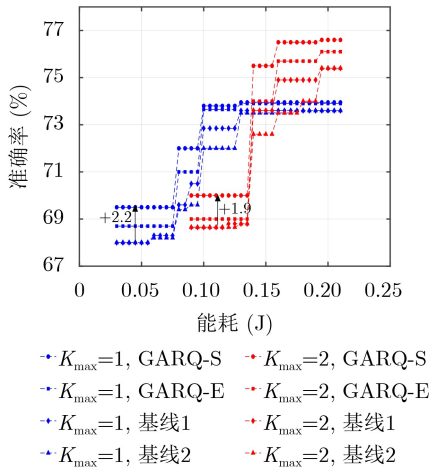


图 7 不同方案和$K_{\max}$下的准确率随最大能耗约束的变化

表 3 在不同数据集下的准确率比较

方案	数据集			
	GSM8K	TriviaQA	ARC-C	SquAD
基线1	48.65	70.20	72.10	80.70
基线2	48.05	69.80	71.90	80.30
GARQ-E	**51.20**	71.30	72.35	80.90
GARQ-S	50.30	**71.40**	**74.32**	**81.80**

表 4 不同方案的运行时间比较

方案	GARQ	基线1	基线2
运行时间/秒	4.287	16.415	55.621

6 结论

在本文中，我们提出了G-ARQ，这是一种新型的将协作边缘推理转化为语义引导的重传的闭环框架。通过引入用于错误驱动模型选择的SEMAR和一个黑箱轨迹优化器，该框架在严格的资源约束下实现了显著的准确率提升。实验评估验证了G-ARQ在可靠的6G边缘智能应用中的有效性。

参 考 文 献

[1] CUI Yuanhao, CAO Xiaowen, ZHU Guangxu, *et al*. Edge perception: Intelligent wireless sensing at network edge[J]. *IEEE Communications Magazine*, 2025, 63(3): 166–173. doi: 10.1109/MCOM.001.2300660.

[2] LIU Chang and ZHAO Jun. Resource allocation in large language model integrated 6G vehicular networks[C]. 2024 IEEE 99th Vehicular Technology Conference, Singapore, Singapore, 2024: 1–6. doi: 10.1109/VTC2024-Spring62846.2024.10683673.

[3] LIN Zheng, QU Guanqiao, CHEN Qiyuan, *et al*. Pushing large language models to the 6G edge: Vision, challenges, and opportunities[J]. *IEEE Communications Magazine*, 2025, 63(9): 52–59. doi: 10.1109/MCOM.001.2400764.

[4] LUO Haoxiang, LIU Yingqiu, ZHANG Ruichen, *et al*. Toward edge general intelligence with multiple-large language model (Multi-LLM): Architecture, trust, and orchestration[J]. *IEEE Transactions on Cognitive Communications and Networking*, 2025, 11(6): 3563–3585. doi: 10.1109/TCCN.2025.3612760.

[5] WANG Lijing, GHOSH D, GONZALEZ DIAZ M T, *et al*. Wisdom of the ensemble: Improving consistency of deep learning models[C]. Proceedings of the 34th Conference on Neural Information Processing Systems, Vancouver, Canada, 2020: 19750–19761.

[6] MAZHAR N, ULLAH S A, CHAUHDARY S H, *et al*. Optimizing age of information in energy-constrained IIoT networks: A reinforcement learning framework[J]. *IEEE Internet of Things Journal*, 2025, 12(20): 42813–42828. doi: 10.1109/JIOT.2025.3594665.

[7] AHMED A, AL-DWEIK A, IRAQI Y, *et al*. Hybrid automatic repeat request (HARQ) in wireless communications systems and standards: A contemporary survey[J]. *IEEE Communications Surveys & Tutorials*, 2021, 23(4): 2711–2752. doi: 10.1109/COMST.2021.3094401.

[8] LONG Hang, XIANG Wei, SHEN Shanshan, *et al*. Analysis of conditional error rate and combining schemes in HARQ[J]. *IEEE Transactions on Signal Processing*, 2012, 60(5): 2677–2682. doi: 10.1109/TSP.2012.2184100.

[9] SHLEZINGER N, FARHAN E, MORGENSTERN H, *et al*. Collaborative inference via ensembles on the edge[C]. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Canada, 2021: 8478–8482. doi: 10.1109/ICASSP39728.2021.9414740.

[10] 丁男, 王佳佳, 冀承慧, 等. 边-端协作下基于早期退出机制的深度神经网络动态自适应分区[J]. 电子与信息学报, 2025, 47(10): 4005–4017. doi: 10.11999/JEIT250291.
DING Nan, WANG Jiajia, JI Chenghui, *et al*. Dynamic adaptive partitioning of deep neural networks based on early exit mechanism under edge-end collaboration[J]. *Journal of Electronics & Information Technology*, 2025, 47(10): 4005–4017. doi: 10.11999/JEIT250291.

[11] 陈阳, 马欢, 姬智, 等. 面向图像恢复任务的语义通信网络能耗优化[J]. 电子与信息学报, 2026, 48(1): 183–190. doi: 10.11999/JEIT250915.
CHEN Yang, MA Huan, JI Zhi, *et al*. Optimization of energy consumption in semantic communication networks for image recovery tasks[J]. *Journal of Electronics & Information Technology*, 2026, 48(1): 183–190. doi: 10.11999/JEIT250915.

[12] ZHENG Guhan, NI Qiang, NAVAIE K, *et al*. Semantic communication in satellite-borne edge cloud network for computation offloading[J]. *IEEE Journal on Selected Areas in Communications*, 2024, 42(5): 1145–1158. doi: 10.1109/JSAC.2024.3365879.

[13] 林艳, 夏开元, 张一晋. 基于生成对抗网络辅助多智能体强化学习的边缘计算网络联邦切片资源管理[J]. 电子与信息学报, 2025, 47(3): 666–677. doi: 10.11999/JEIT240773.
LIN Yan, XIA Kaiyuan, and ZHANG Yijin. Federated slicing resource management in edge computing networks based on GAN-assisted multi-agent reinforcement learning[J]. *Journal of Electronics & Information Technology*, 2025, 47(3): 666–677. doi: 10.11999/JEIT240773.

[14] HU Chenbo, YANG Hongjuan, LI Bo, *et al*. HARQ-aided RSMA for integrated satellite-terrestrial networks[J]. *IEEE Transactions on Wireless Communications*, 2026, 25: 14987–15003. doi: 10.1109/TWC.2026.3681284.

[15] HUANG Yichong, FENG Xiaocheng, LI Baohang, *et al*. Ensemble learning for heterogeneous large language models with deep parallel collaboration[C]. 38th Conference on Neural Information Processing Systems, Vancouver, Canada, 2024: 119838–119860. doi: 10.52202/079017-3808.

[16] PAN Chaoyi, YI Zeji, SHI Guanya, *et al*. Model-based diffusion for trajectory optimization[C]. Proceedings of the 38th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2024: 57914–57943.

王腾胜：男，硕士生，研究方向为边缘智能、绿色无线通信、无线资源分配.

余　涛：男，博士后，研究方向为无线网络的节能通信网络、机器学习、深度学习.

李济宏：男，博士生，研究方向为网络切片、无线资源分配、节能通信网络.

郑谷寒：男，副教授，研究方向为非地面网络、车载网络、语义通信.

张舜卿：男，教授，研究方向为节能5G/5G+通信网络、绿色无线网络、异构计算技术.

责任编辑：廖海贝

Gating Adaptive Repeat Query Framework for Reliable Collaborative Inference with Edge Heterogeneous LLMs

WANG Tengsheng[①] YU Tao[②] LI Jihong[①]
ZHENG Guhan[①] ZHANG Shunqing[①]

[①](*School of Communication and Information Engineering, Shanghai University, Shanghai* 200444, *China*)
[②](*Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China*)

Abstract:

Objective Reliable inference at the network edge is indispensable for 6G-enabled ubiquitous AI, yet the deployment of large language models (LLMs) in such environments remains a cornerstone challenge. In resource-constrained edge settings, single-LLM inference often proves unreliable due to knowledge limitations and inherent biases, severely hampering real-world deployment. Collaborative inference leveraging multiple heterogeneous LLMs emerges as a promising remedy to boost robustness, but it introduces nontrivial hurdles under stringent latency and energy budgets, especially when wireless channel conditions and query content vary unpredictably. These challenges include the need for dynamic sequential decision-making for LLM selection and resource allocation, the fundamental paradigm mismatch between bit-level reliability protocols and semantic-level error correction, and the lack of adaptive mechanisms to align and fuse disparate LLM outputs effectively. To fill these critical gaps, this paper presents a novel framework that fundamentally reinterprets collaborative inference as a semantic-driven, closed-loop process, thereby transitioning from conventional bit-retransmission to semantic-retransmission and offering a practical path toward reliable 6G edge intelligence.

Methods In response to these critical challenges, we propose the Gating Adaptive Repeat Query (G-ARQ) framework. Its core innovation is the Semantic-Space Alignment and Error-Guided Retransmission (SEMAR) mechanism. SEMAR first aligns the token-level probability distributions from heterogeneous LLMs into a unified semantic space using relative representation, enabling comparable outputs. It then models the collaborative process probabilistically, explicitly capturing error dependencies among models, and uses an Expectation-Maximization (EM) algorithm to infer a latent error direction, which guides the selection of the next LLM for query retransmission, steering it towards outputs orthogonal to previous errors. To jointly optimize the LLM gating and uplink power allocation under communication constraints without requiring explicit system dynamics—often unavailable in practice—we design a black-box trajectory optimizer. This optimizer formulates the sequential decision problem as sampling from a target distribution that encodes dynamic feasibility, optimality, and constraints. It employs a diffusion-based sampling process with a model-guided prior and Monte Carlo estimation to generate near-optimal policy trajectories that satisfy hard latency and energy limits.

Results and Discussions To evaluate the practical viability of G-ARQ under realistic edge conditions, simulations are conducted in a scenario with five base stations hosting five heterogeneous 7B-parameter LLMs (Mistral-7B, Vicuna-7B, Nous-Capybala-7B, Gemma-7B, and Llama-2-7B). The user equipment (UE) performs a question-answering task evaluated on a mixed SQuAD and TriviaQA dataset. Component-level evaluations, each designed to isolate the contribution of a single innovation, validate the effectiveness of every key element. The error-guided gating of SEMAR, compared to a Top-k gating baseline, improves accuracy by 0.23 % on average, and its dynamic weight ensemble contributes an additional 0.7 % gain (Fig. 3). The black-box trajectory optimizer, which operates without any explicit channel model, achieves accuracy close to that of the unconstrained model-greedy strategy while ensuring strict latency constraints (Fig. 4). The convergence of the optimizer is verified by tracking the evolution of $J(\boldsymbol{S})$ and selection probability over diffusion steps (Fig. 5). System-level performance under varying latency and energy constraints demonstrates that G-ARQ consistently surpasses two baselines: one combining model-greedy selection with Proximal Policy Optimization (PPO) for power optimization, and another combining model-greedy selection with Simulated Annealing for power optimization, both using average output weights. The accuracy improvement is most significant under the most

stringent resource limits, reaching up to 2.2 % for $K_{\max} = 1$ and 1.9 % for $K_{\max} = 2$ (Fig. 6, Fig. 7). The framework successfully establishes a Pareto boundary that characterizes the inherent trade-off between inference accuracy and communication latency, providing a valuable design guideline for resource-constrained edge systems and offering actionable insights for real-world deployment. The GARQ-S variant is noted to outperform GARQ-E by avoiding the integration of outputs from previously erroneous models.

Conclusions This paper proposed G-ARQ framework, an innovative closed-loop framework that transforms collaborative edge inference into a semantics-guided retransmission process. By introducing SEMAR for error-based alignment and selection of heterogeneous LLMs, and employing a black-box trajectory optimizer for joint model selection and power allocation, the framework achieves up to a 2.2% accuracy improvement under strict resource constraints. The results validate G-ARQ as an effective and practical approach toward reliable and efficient 6G edge intelligence.

Key words: 6G edge intelligence; Large language models; Collaborative inference; Gating adaptive repeat query.