# 泛在计算环境下模型推理输出长度预测

牛昊一[1,2],林 露[1],裘向明[1],张北北[1],邹 涛[1],高 丰[1*]
(1. 之江实验室,高效能计算设施研究中心,浙江杭州 311100;2. 温州肯恩大学国际前沿交叉研究院,浙江温州 325060)

**摘 要:** 在泛在计算场景中,系统涵盖大、中、小等多种规模的模型与端侧、边侧、云侧等多种算力形态,基于服务质量需求,需对模型推理服务进行精细化调度。在此背景下,精准预测模型推理的输出长度,对实现高效资源管理至关重要。然而,由于生成结果存在高度可变性且受语义影响显著,输出长度预测始终面临挑战。本研究提出一种轻量级的预测模型训练框架(Output Length Prediction Model, OLPM),仅依据输入提示即可预测请求的输出长度区间。框架采用基于DistilBERT的架构以保障计算效率,并引入两项关键优化手段:一是基于正态分布的过滤策略,通过降低异常值偏差提升训练稳定性;二是语义提示增强方法,提高模型对控制长度语言线索的敏感度。评估结果表明,OLPM的预测精度达到当前最优水平,最高准确率可达99.6%。与现有模型相比,OLPM不仅显著提升了预测精度,还大幅缩短了模型收敛时间,实现了精度与效率的协同优化。实验结果进一步说明,该方法收敛迅速、性能稳定,能够实现精准且高效的输出长度预测。OLPM能够在保持高预测精度同时,具备较低的模型复杂度与计算开销,显著优于现有方法,可有效满足泛在计算场景下对资源调度精细化的需求,展现出卓越的实用性、可扩展性与泛化能力,为实现泛在计算环境下高效、智能的模型推理服务提供了先进的技术路径。

**关键词:** 泛在计算;模型推理;输出长度预测;过滤策略;语义提示增强
**基金项目:** 国家重点研发计划(No.2022YFB4501804)
**中图分类号:** TP39 **文献标识码:** A **文章编号:** 0372-2112(2026)04-1614-12
**电子学报 URL:**http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250897

# Prediction on Model Inference Output Length in Ubiquitous Computing Environment

NIU Haoyi[1,2], LIN Lu[1], XI Xiangming[1], ZHANG Beibei[1], ZOU Tao[1], GAO Feng[1*]
(1. *Research Center for High Efficiency Computing Infrastructure*, *Zhejiang Lab*, *Hangzhou*, *Zhejiang* 311100, *China*;
2. *International Frontier Interdisciplinary Research Institute*, *Wenzhou-Kean University*, *Wenzhou*, *Zhejiang* 325060, *China*)

**Abstract:** In ubiquitous computing scenarios, systems encompass large, medium, and small-scale models, as well as diverse computing resources spanning edge-side, device-side, and cloud-side infrastructures. Efficient scheduling of model inference services, driven by quality-of-service requirements, is critical to optimizing system performance. Accurate prediction of output length during model inference plays a pivotal role in enabling fine-grained resource management. However, this task remains challenging due to the high variability and strong semantic dependence of generated outputs. To address this, we propose a lightweight training framework, which is output length prediction model (OLPM) to predict the output length range of requests based solely on the input prompt. Our framework employs a DistilBERT architecture to ensure computational efficiency, enhanced with two key innovations: (1) a normal-distribution-based filtering strategy that improves training stability by mitigating the impact of outliers, and (2) a semantic prompt augmentation method that enhances the model's sensitivity to linguistic cues indicative of output length. The evaluation results show that OLPM achieves state-of-the-art prediction accuracy, with a maximum accuracy of up to 96.8%. Compared to existing models, OLPM not only significantly improves prediction accuracy but also greatly reduces model convergence time, achieving a synergistic optimization of precision and efficiency. Experimental results further show that the method converges rapidly and performs stably, enabling accurate and efficient output length prediction. By maintaining high prediction accuracy while featuring low model complexity and computational overhead, OLPM significantly outperforms existing approaches. It effectively meets the demands for fine-grained resource scheduling in ubiquitous computing environments, demonstrating outstanding practicality, scalability, and generalization capability. OLPM provides an advanced technical pathway for realizing efficient and intelligent model inference services in ubiquitous computing scenarios.

收稿日期:2026-01-14;录用日期:2026-01-23;责任编委:彭鑫
*通信作者:高丰

Keywords: ubiquitous computing; model inference; output length prediction; filtering strategy; semantic prompt augmentation

Foundation Item(s): National Key Research and Development Program of China (No.2022YFB4501804)

## 0 引言

在泛在计算环境中,推理模型已成为人工智能领域的核心组成部分,其应用场景的广泛性与多样性,催生了对高效、可扩展推理服务系统的迫切需求[1-2]。随着应用场景的多样化与服务需求的激增,构建高效、可扩展且服务质量可保障的模型推理系统已成为关键挑战。泛在计算场景呈现出多模型共存(涵盖大、中、小规模模型)、多算力形态协同(融合终端、边缘与云端资源)的复杂架构特征,要求推理服务具备高度动态化与精细化的调度能力[3-4]。然而,推理模型的自回归生成机制导致其解码延迟与输出序列长度呈近似线性增长关系,而不同请求的生成长度差异巨大,从数十至数千 token 不等,造成执行时间的高度异构性[5]。与传统可预测计算负载的模型不同,模型推理请求的计算需求难以预先估计,若缺乏对输出长度的先验认知,资源分配过程中极易引发内存碎片化[6]、批处理效率下降[7]、尾部延迟增加等问题[8],严重制约系统吞吐与响应性能。

为应对上述挑战,研究者提出语义路由[9]、动态批处理[10]和 KV 缓存优化[11]等技术提升计算效率,但这些方法的有效性高度依赖于对推理过程的精准预判,尤其是对输出长度的准确估计[12]。因此,在推理开始前,基于输入提示预测生成序列 token 数量成为实现高效资源管理的关键前提[13]。现有研究表明,对推理请求计算负载的精准预估,能够显著降低尾部延迟并提升集群吞吐量[14]。然而,该任务在建模上面临显著挑战:模型的推理输出长度受任务类型、指令明确性、上下文复杂度及用户意图等多重因素耦合影响,这些因素呈现非线性、任务相关且语义敏感的特性[13],导致基于输入长度或简单启发式规则的方法在多样化工作负载下性能急剧下降。

尽管已有研究尝试通过浅层模型[15-16]或微调预训练编码器[17]预测输出长度,但其在泛在计算场景下面临显著局限:浅层模型依赖人工特征,难以捕捉提示词中的细微语义差异,泛化能力弱;基于模型微调的方法虽然语义理解能力强,但因高参数量导致推理延迟大、资源消耗高,难以满足端侧与边侧设备的低延迟部署需求。此外,泛在环境下的推理任务负载具有高度动态性与异构性,导致输出长度分布差异显著。现有方法易受数据分布变化影响,性能不稳定,且通常孤立优化精度或效率,难以在预测准确性、计算开销与部署可行性之间取得平衡[15-17]。因此,亟需兼具高精度、低开销与强泛化能力的轻量级预测框架,以支持泛在计算环境下的高效推理。

针对上述问题,本文提出一种轻量级、高精度的输出长度预测模型(Output Length Prediction Model, OLPM),面向泛在计算场景下的推理模型工作负载,实现基于输入提示的输出长度区间预测。OLPM 结合“正态分布过滤-语义增强”双重优化策略,以较小的计算开销实现了高精度预测。具体而言,将长度预测建模为基于分位数的区间分类任务,提升模型对噪声与异常值的鲁棒性;双重优化策略通过基于统计的自适应过滤机制,剔除极端生成样本以稳定训练过程,同时引入保留语义的提示增强技术,引导模型学习控制长度的关键语言线索,增强其对语义敏感度的感知能力,该方法融合符号化指令设计与数据驱动学习,实现可控且可解释的训练流程。

相较于基于输入长度、词频等手工特征的浅层回归模型,OLPM 通过语义保留提示增强机制,显式建模语言指令与输出长度间的因果关联,提升对复杂语义的敏感度。其次,区别于直接微调的方法,OLPM 以 DistilBERT 为骨干编码架构,在保持强大语义理解能力的同时降低参数量与推理延迟,满足端侧和边侧设备的轻量化部署需求。实验结果表明,OLPM 在多类典型自然语言处理任务上达到当前最优预测精度,预测结果可集成至各类调度框架,有效支持静态内存预留、最优批处理规划等优化策略,为泛在计算场景下模型推理服务的精细化调度提供有力支撑。

## 1 问题建模

本研究聚焦于泛在计算场景下模型推理输出长度的预测问题,旨在构建一个轻量级、高精度且具备良好泛化能力的预测模型。预测模型的核心任务是:给定一个输入提示 $r_i$,预测其在目标模型上推理生成的输出序列长度所归属的预定义长度区间。与直接回归精确 token 数量不同,本研究将问题建模为多分类任务,以增强对噪声和异常值的鲁棒性,并为下游调度系统提供更具操作性的决策粒度。预测模型可形式化表示为一个映射函数 $f(\cdot)$:

$$\hat{y}_i = f\left(r_i;\ \theta\right) \tag{1}$$

其中,$\hat{y}_i \in \{1,2,\cdots,K\}$ 表示预测的长度区间类别,$K$ 为预设的区间总数,$\theta$ 为模型可学习参数。

为量化预测结果在调度系统中的潜在价值,本文进一步建立数学模型,分析推理输出长度对负载均衡

与响应延迟的影响。考虑一个由若干个工作节点 $W=\{w_1,w_2,\cdots,w_j\}$ 构成的模型推理系统，持续接收若干个推理请求 $R=\{r_1,r_2,\cdots,r_i\}$，其中每个请求 $r_i$ 在时刻 $t$ 到达。设 $C_j(t)$ 为该队列中最后一个请求的完成时间，若解码时间与输出长度成正比，则将请求 $r_i$ 分配至节点 $w_j$ 的预计完成时间可表示为

$$C_j(t;\ r_i)=C_j(t)+\gamma\cdot\hat{L}_i \tag{2}$$

其中，$\gamma$ 为单位 token 的平均解码时间，$\hat{L}_i$ 为模型针对请求 $r_i$ 的预测长度。高效的推理请求调度策略应最小化请求延迟与考虑系统负载均衡。基于这个目标，可以构建如下目标函数：

$$w^*=\arg\min_{w_j\in W}\left[C_j(t;\ r_i)+\lambda\cdot\left|\frac{N_j(t)\cdot\mu(t)+\hat{L}_i}{N_j(t)+1}-\mu(t)\right|\right] \tag{3}$$

其中，$w^*$ 为最优分配节点，$N_j(t)$ 为当前已分配的请求数量，$\mu(t)$ 为系统平均负载，$\lambda$ 为权重因子。该模型表明，预测误差会直接、成比例地放大为调度决策中的偏差。高精度的预测结果能确保目标函数真实反映不同分配方案的优劣，从而使调度器能够选择最优节点，而低精度的预测则会使成本函数失真，导致调度决策偏离最优解，最终降低系统整体性能。

## 2 方法架构

为解决泛在计算环境下模型推理输出长度预测问题，本文提出一种轻量级、高精度且具备强泛化能力的输出长度预测模型 OLPM。OLPM 的设计核心在于将复杂的长度预测问题转化为基于语义理解的分类任务，并通过协同优化数据质量与语义敏感度，实现精度与效率的平衡。算法框架如图 1 所示，OLPM 的整体架构遵循“数据预处理-特征编码-分类预测”的流程，系统性地整合了语义增强与统计过滤的双重优势。首先，在多任务的工作负载数据集上构建训练样本；随后，通过正态分布分析对各任务的输出长度分布进行自适应过滤，有效剔除极端异常值，保留典型的类别模式；在此基础上，设计语义保留提示增强方法，通过可控地修改输入提示中与输出长度相关的关键语言线索，显式地向模型注入长度控制的因果信号，增强其对细微语义差异的感知能力；最后，采用轻量化的 DistilBERT 作为骨干模型对输入提示进行编码，并通过分类头输出其所属的长度区间。该架构不仅计算开销低、易于部署，而且通过数据层面的双重优化策略，能够提升模型对多样化任务的适应能力与预测鲁棒性，为后续的调度优化提供高可靠依据。

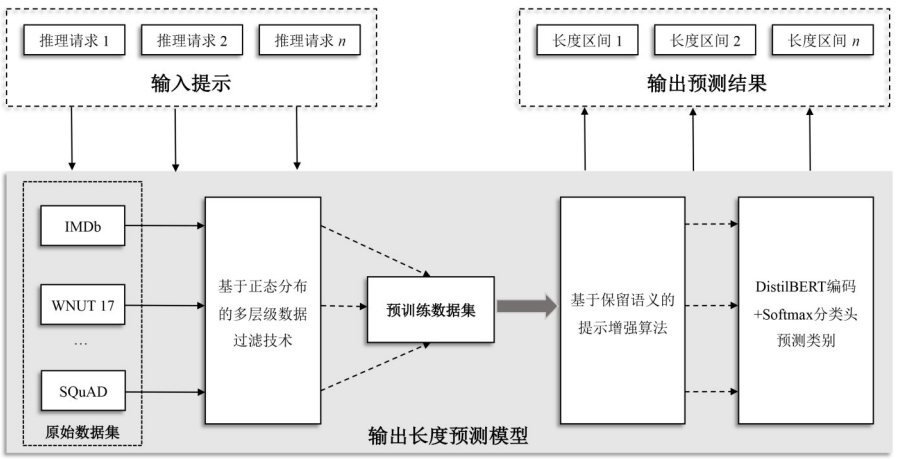



图 1 OLPM 整体技术架构

Figure 1 Framework of OLPM

### 2.1 预测模型架构

OLPM 采用 DistilBERT 作为骨干编码架构，旨在构建一个计算高效且语义表征能力强的轻量化预测模型。DistilBERT 通过知识蒸馏技术从 BERT 中学习，保留了其大部分语义理解能力，同时显著减少了模型参数量和推理延迟，使其更适合在资源受限的泛在计算场景中部署。给定输入提示 $r_i$，OLPM 首先通过 DistilBERT 编码器生成其上下文感知的词元级嵌入序列 $\boldsymbol{h}=\text{DistilBERT}(r_i)$。进一步地，为获得固定维度的句子级表征，OLPM 采用平均池化策略[17]对所有词元进行聚合：

$$\boldsymbol{h}_{\text{avg}}=\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{h}_j \tag{4}$$

其中，$n$ 为输入序列长度。随后，将池化后的向量 $\boldsymbol{h}_{\text{avg}}$ 输入一个全连接分类头，通过 Softmax 函数输出其属于各个预定义长度区间的概率：

$$p_i=\text{Softmax}(\boldsymbol{W}\boldsymbol{h}_{\text{avg}}+\boldsymbol{b}) \tag{5}$$

基于 $p_i$ 最大值得到的类别 $\hat{y}_i$ 为预测长度区间。

该架构设计简洁高效,生成的高质量语义表征不仅为下游分类任务提供坚实基础,也为本文提出的正态分布过滤和语义增强方法提供了可靠的语义相似性度量依据,确保了数据处理过程的安全性与有效性。

### 2.2　基于正态分布的多层级数据过滤技术

训练数据中的异常值是影响预测模型稳定性的关键因素,在模型推理过程中,异常值可能源于模型幻觉、重复循环或罕见的边缘案例,这些异常的输出长度训练数据会影响预测模型对典型推理行为的学习过程,导致学习偏差。为了解决这个问题,本研究提出一种基于正态分布的多层级数据过滤技术,自适应地过滤并构建分布合理的训练数据集。对于构建训练数据集的自然语言处理任务,每一类任务因其内在语义目标与输出形式不同,天然形成独立的长度分布模式。因此,我们在每个任务子集 $D_k$ 上独立估计均值 $\omega_k$ 与标准差 $\sigma_k$,实现任务感知的自适应过滤,避免跨任务分布混杂导致的误剔除。假设在特定任务类别 $t_k \in T$ 下,符合预期的输出长度 $\hat{L}_i$ 近似服从正态分布 $\hat{L}_i \sim N(\omega_k, \sigma_k^2)$,$\omega_k$ 和 $\sigma_k$ 的计算公式如下:

$$\omega_k = \frac{1}{\left|D_k\right|} \sum_{(r_i, L_i) \in D_k} L_i \tag{6}$$

$$\sigma_k = \sqrt{\frac{1}{\left|D_k - 1\right|} \sum_{(r_i, L_i) \in D_k} \left(L_i - \omega_k\right)^2} \tag{7}$$

其中,$D_k$ 表示属于任务 $t_k$ 的样本子集。基于此统计模型,设定一个基于 $2\sigma$ 原则的过滤阈值,剔除输出长度偏离均值超过两个标准差的样本,以保留高置信度区间内的典型推理数据,过滤后的数据集 $D_{\text{filtered}}$ 定义为

$$D_{\text{filtered}} = \bigcup_{t_k \in T} \left\{ \left(r_i, L_i\right) \in D_k \middle| \left| L_i - \omega_k \right| \leqslant 2\sigma \right\} \tag{8}$$

基于正态分布的多层级数据过滤技术在保留典型推理样本的同时,有效滤除了长尾中的极端异常值,相较于固定阈值或全局百分位数过滤,该方法具有任务自适应性,能够根据不同任务的固有长度分布特性进行动态调整,避免滤除语义合理但长度较长的响应数据。通过这一多层级的过滤过程,能够显著提升训练数据的质量,为模型学习稳健的统计规律提供可靠保障。

### 2.3　基于保留语义的提示增强算法

提示词中的细微语言线索与输出长度密切相关,但这些特征在原始数据集中出现频率较低且分布稀疏,模型通常难以捕捉。尽管同义词替换等传统数据增强技术已被广泛证实能有效提升模型在小样本场景下的泛化边界与鲁棒性[18],但简单的随机替换容易破坏特定语境下的长度控制语义。为显式增强模型对这类语义信号的敏感度,本研究提出一种基于保留语义的提示增强算法,使模型能够学习到特定语言变化与输出长度之间的因果关联。该方法的核心逻辑如下:首先,基于训练数据开展触发词挖掘,通过统计分析高频长度相关指令、人工校验语义倾向性的方式,筛选出具备明确长度引导作用的核心词汇(定义为触发词);其次,为确保替换过程不改变提示词核心语义,基于主流大模型和中英文同义词库构建触发词语义关联网络,挖掘同义变体形成触发词候选集,每个触发词均关联明确的长度倾向(短/中/长),能在严格遵循“保留核心语义”约束的前提下,生成多样化且语境适配的表达变体,有效避免传统同义词替换可能导致的语义偏移问题;最后,对于训练集中包含触发词的提示词,算法通过语义相似性匹配,寻找一个在语义上相近但长度倾向相反的替代词。对于原提示词中出现的触发词 $s$,设 $S$ 为与 $s$ 语义相关且长度倾向相反的候选替换词集合,$s''$ 表示集合 $S$ 中的任一候选词。通过计算 $s$ 与各候选词之间的语义相似度,从 $S$ 中选择语义最接近的替代词 $s'$:

$$s' = \arg\max_{s'' \in S} \text{sim}(s, s'') \tag{9}$$

其中,替换后生成新提示 $r_i' = r_i, s'$,并记录其对应的模型实际生成长度 $L_i'$,该增强过程可形式化为如下映射关系:

$$\Phi: (r_i, s, s') \rightarrow (r_i', L_i') \tag{10}$$

增强样本 $(r_i', L_i')$ 与原始样本 $(r_i, L_i)$ 构成一对组对比样本数据。在训练过程中,模型通过最小化这两者预测分布的差异,被显式引导学习 $s$ 与 $s'$ 的语义差异如何导致输出长度发生变化。算法将符号化的指令设计思想融入数据驱动的学习过程,实现了可控、可解释的训练增强,有效提升模型对语义指令的解析能力与预测精度。

## 3　实验分析

### 3.1　实验环境

为了全面评估 OLPM 在真实泛在计算场景下的性能表现,本研究构建了一个包含模型训练与在线推理调度的端到端实验环境。所有实验均在高性能计算集群上完成,计算节点配备 NVIDIA A100 GPU,显存带宽约为 1.5 TB/s,软件环境基于 SGLang 推理引擎构建。在推理系统架构方面,扩展了 SGLang 的核心路由组件 sgl-router,构建了包含 4 个推理工作节点 Workers 的集群拓扑。如图 2 所示,OLPM 被作为可插拔的预处理插件集成至路由层,前端通过 sglang_router.bench_serving 工具模拟真实的高并发流式请求,OLPM 能够实时拦截推理请求并预测输出长度区间,随后系统基于自定义的负载均衡策略将请求分发至后端。后端推理实例统一加载 DeepSeek-R1-Distill-Llama-8B 模型,该模型基于 Llama-3.1 架构,是当前端

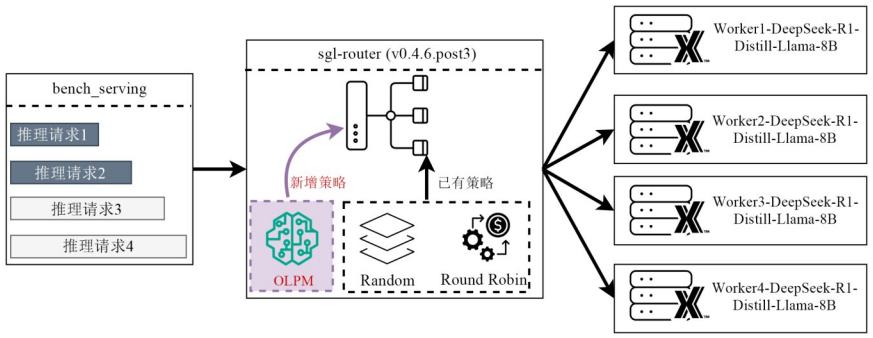



图2 OLPM与SGLang集成方案示意图

Figure 2 Schematic diagram of OLPM and SGLang integration

侧与边缘侧广泛应用的代表性轻量化大语言模型。

针对预测模型OLPM的训练配置，在训练阶段，将所有数据集按8∶1∶1的比例严格划分为训练集、验证集与测试集。为防止模型过拟合并确保泛化能力，训练中引入早停机制，最大训练轮数设定为200，若验证集在连续5轮内无明显下降则自动终止训练。此外，文中报告的所有实验结果均基于五折交叉验证取平均值，以消除数据划分随机性对实验结论的影响。

在负载构建与数据集设置方面，实验数据涵盖八个具有代表性的自然语言处理任务：IMDb情感分析、ELI5复杂问答、WNUT-17实体识别、SQuAD问答、Books摘要、MMLU-PRO多任务语言理解、GPQA-Diamond专家级理解和推理以及AIME2025数学推理。数据集按照8∶1∶1划分为训练集、验证集与测试集，原始输入提示长度范围在109~12 791 tokens之间，原始输出长度范围在18~51 634 tokens之间。实验中将输出长度离散化为短（1~333 tokens）、中（334~512 tokens）、长（512~∞ tokens）3个区间，匹配泛在场景算力节点资源承载能力和调度决策需求，具体构成与统计特性如表1所示。并据此构建了四类测试场景，分别是短输入短输出的Chat场景、长输入短输出的Summary场景、短输入长输出的Creative场景，以及包含三类场景各1 000条请求的Mixed场景。Chat场景输入长度范围为0~100，输出为短序列；Summary场景输入长度范围为3 000~3 300，输出为短序列；Creative场景输入长度范围为1 000~2 800，输出为长序列。

**表1 不同数据集结构与统计特性**

Table 1 Different dataset structures and statistical properties

| 数据集 | 任务类型 | 样本数量 | 原始输出长度(均值±标准差)/tokens | 长度范围/tokens |
|---|---|---|---|---|
| MMLU-PRO | 多任务语言理解 | 6 231 | 513.6±1 370.9 | 99~51 634 |
| AIME2025 | 数学推理 | 30 | 990.4±1 289.5 | 328~7 285 |
| ELI5 | 复杂问答 | 5 000 | 782.1±198.5 | 420~1 024 |
| GPQA-Diamond | 专家级理解和推理 | 198 | 607.6±203.4 | 185~1 407 |
| Books | 书籍摘要 | 5 000 | 624.8±156.3 | 312~986 |
| WNUT-17 | 实体识别 | 4 525 | 385.6±94.2 | 198~528 |
| SQuAD | 阅读理解问答 | 5 200 | 210.3±68.7 | 22~312 |
| IMDb | 情感分析 | 4 800 | 98.4±42.1 | 18~254 |

### 3.2 基于保留语义的提示增强算法的长度触发词库构建

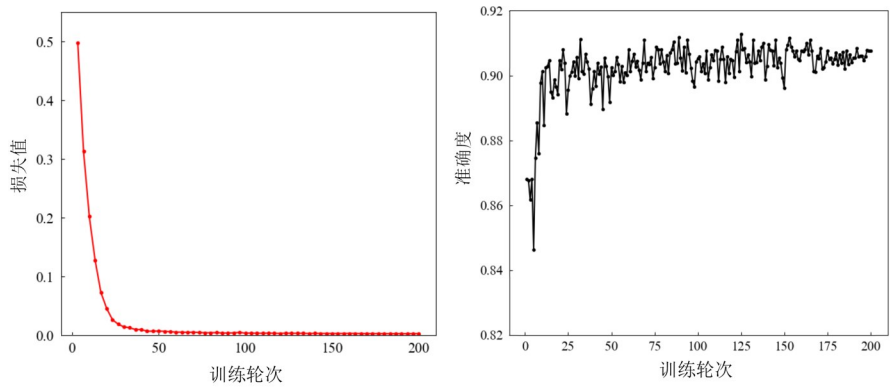
基于2.3节中的通用方法，本文针对实验所用数据集的特点，引入当前主流的Doubao 1.5 Pro大模型，结合WordNet和知网同义词词林，生成了一套适配当前场景的长度控制触发词库，共包含24个核心触发词，短/中/长3类各8个，详见附录A。该词库是方法的具体落地实例，而非通用唯一解：其核心触发词及同义变体均源自实验数据集的高频表达，长度倾向经过人工交叉校验，一致率≥95%，适配本文所涉及的任务类型与数据分布。在其他数据集或任务场景中，可替换为DeepSeek-V3、通义千问Qwen3等具备强推理或长文本处理能力的主流大模型，通过本文提出的“触发词挖掘-语义保留替换”流程，重新生成适配该场景的触发词库，无需修改方法核心逻辑。

此外，为验证提示增强过程中的语义一致性，本研究采用人工标注方式，通过随机抽取1 000个增强样本，由标注员独立判断两者核心语义是否一致。结果显示，语义完全一致率达92.3%，基本一致率达7.1%，语义偏差率仅0.6%，证明本方法在增强长度关联线索的同时，能有效保留输入提示的核心语义。词

库在各数据集的适用比例分别为:IMDb(89.2%)、ELI5(91.7%)、WNUT-17(85.3%)、SQuAD(93.5%)、Books(87.6%)、MMLU-PRO(91.6%)、GPQA-Diamond(88.3%)和AIME2025(90.5%),表明词库对不同类型自然语言处理任务具有良好的适配性。针对多语言和多任务场景,词库的中英文双语设计可直接支持跨语言迁移,并可通过语义相似性匹配机制进行替换,具备跨任务推广能力。

### 3.3　预测精度与训练效率分析

图3所示为OLPM训练过程的收敛曲线,如图3(a)所示,得益于"正态分布过滤-语义增强"策略的协同作用,预测模型的损失值在初始阶段迅速下降,并在前50轮内收敛至稳定值,表明本研究提出的方法能够有效优化效率与训练速度。与之对比,图4中未采用优化策略的基线模型收敛缓慢且波动较大。此外,验证集上的预测精度对比结果表明,在训练初期,OLPM的预测精度能够快速提升,在约100轮后趋于稳定,最终达到超过90%的高精度水平。基线模型的精度提升较为平缓,经过200轮训练后仅达到约48%,性能差距显著,验证了本文所提方法在加速模型收敛和提升预测性能方面的有效性。

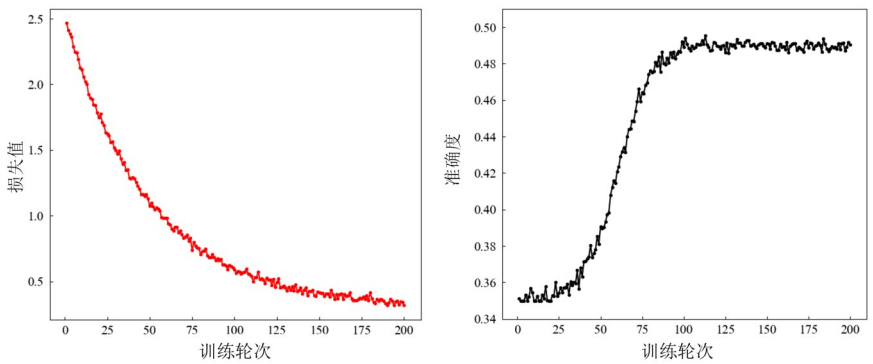



(a) 损失曲线　　(b) 预测准确度曲线

(a) Loss curve　　(b) Prediction accuracy curve

图3　OLPM训练过程损失曲线及预测准确度曲线

Figure 3　Training loss curve and prediction accuracy curve of OLPM




(a) 损失曲线　　(b) 预测准确度曲线

(a) Loss curve　　(b) Prediction accuracy curve

图4　基线模型训练过程损失曲线及预测准确度曲线

Figure 4　Baseline model training loss curve and prediction accuracy curve

为量化OLPM的综合性能,图5展示了OLPM与多种模型在7类数据集上的分类精度对比结果。基于数据量的考虑,我们将GPQA-Diamond数据集和AIME2025数据集合并在一起。实验结果表明,OLPM在所有数据集上均取得了最高的预测精度。具体地,OLPM针对MMLU-PRO数据集达到了99.6%的准确性,并在IMDb、ELI5、WNUT-17、SQuAD、Books和GPQA-Diamond+AIME2025上分别取得了96.8%、79.1%、77.3%、78.9%、72.6%和70.2%的结果,超过了BERT-base、RoBERTa-large、DeBERTa-v3和T5-small等对比模型。

为进一步验证OLPM在轻量化性能方面的竞争

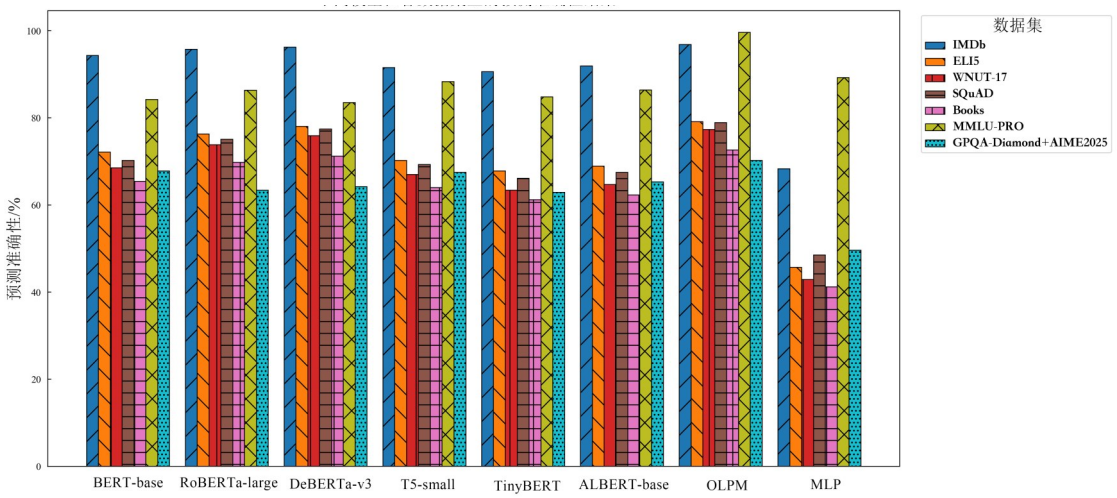



图5 不同模型在数据集上的预测对比结果

Figure 5 Comparison of prediction results of different models on the dataset

力，引入TinyBERT与ALBERT-base两类高效预训练语言模型及MLP浅层模型作为基线。如图5所示，尽管TinyBERT与ALBERT-base在参数规模上具备优势，其预测准确率仍低于OLPM。在IMDb上，TinyBERT和ALBERT-base的预测准确性分别为90.6%和91.9%，OLPM达到96.8%；在长文本生成任务Books和MMLU-PRO上，两者准确率不足63%和90%，OLPM提升至72.6%和99.6%，主要原因是高度压缩的架构削弱了语义建模能力，难以捕捉长度控制相关的细微语言线索。而MLP模型虽实现极致轻量化，但由于仅依赖词频等浅层统计特征，无法建模语义关联，导致精度大幅下降，在复杂语义任务ELI5、WNUT-17和GPQA-Diamond+AIME2025中，MLP精度不足50%。

此外，在ELI5等需要模型具有深度语义理解能力的复杂数据集上，OLPM具有显著优势，说明其具有较强的精准捕捉复杂语言线索的能力，证明了OLPM对于泛在场景多样化任务的适应性，能够提供精准的长度预测结果。上述结果充分证明，OLPM实现了“轻量化-语义理解-预测精度”的协同优化，既避免了传统预训练模型的高开销问题，又克服了浅层模型语义建模能力薄弱的缺陷，能够适配泛在场景多样化任务的精准长度预测需求。

### 3.4 计算开销分析

为系统评估OLPM的轻量化特性，从参数量、计算复杂度、推理延迟、相对开销占比及预测准确率五个核心维度展开量化分析，结果如表2所示。所有测试均在统一的NVIDIA A100 GPU硬件配置下完成，测试条件统一设定：单样本推理延迟测试采用batch size=1、并发请求数=1，以精准量化模型基础推理效率。OLPM基于DistilBERT构建，参数量维持$66\times10^6$，计算复杂度仅为3.2 GFLOPs，显著低于RoBERTa-large、BERT-base等模型；单样本推理平均延迟4.8 ms，表明其对硬件资源的占用更低。尽管TinyBERT和ALBERT-base的参数量和计算复杂度更低，但由于其高度压缩的架构削弱了语义建模能力，预测精度显著低于OLPM，难以满足泛在场景的精度需求。相比RoBERTa-large，OLPM参数减少81.4%、计算复杂度降低82.9%、推理提速5.9倍、吞吐性能提升3.9倍，同时保持最高预测精度，实现了高精度与低开销的协同优化。

**表2 不同模型计算开销对比结果**

Table 2 Comparison of computational costs of different models

| 模型 | 参数量×$10^6$ | 计算复杂度/GFLOPs | 推理延迟/ms | 相对开销占比/% | 预测准确率/% |
|---|---|---|---|---|---|
| BERT-base | 110 | 10.8 | 11.5 | 1.10 | 74.64 |
| RoBERTa-large | 355 | 18.7 | 28.7 | 2.73 | 77.20 |
| DeBERTa-v3 | 184 | 14.3 | 19.3 | 1.84 | 78.06 |
| T5-small | 60 | 4.1 | 8.3 | 0.79 | 73.97 |
| TinyBERT | 14.5 | 0.8 | 2.1 | 0.20 | 70.97 |
| ALBERT-base | 11.7 | 0.7 | 2.3 | 0.22 | 72.43 |
| MLP | 8.0 | 0.3 | 1.8 | 0.17 | 55.06 |
| OLPM | 66 | 3.2 | 4.8 | 0.46 | 82.07 |

为了更直观地评估模型引入的计算开销，本节引入“相对开销占比”指标，选取本研究中最短任务数据集IMDb作为基准。根据表1统计，IMDb的平均输出长度为98.4 tokens，基于NVIDIA A100硬件测试，DeepSeek-R1-Distill-8B模型的单token生成耗时约为10.67 ms，忽略预处理与系统调度的边际开销前提下，以解码耗时为核心估算，完成一个最短请求的耗时通常在1 049.6 ms量级。在此基准下，OLPM的4.8 ms推理延迟仅占整个任务耗时的0.46%。相比之下，虽然MLP的相对开销占比降低至0.17%，但两者在绝对

时间上的差异仅为3 ms,对于1 000 ms量级的生成过程中,这种微弱差异对于用户感知的响应延迟几乎可以忽略不计;对于长文本生成任务,OLPM的相对开销占比将被进一步稀释至0.1%以下。

在“计算开销”与“预测精度”的权衡分析中,表2显示MLP的平均预测准确率仅为55.06%,这表明其虽然实现了极致的参数轻量化,却无法提供有效的调度决策支持,极易因错误的长度预估导致后端显存碎片化与算力浪费。而OLPM在保持0.46%极低计算开销的同时,提供了高达82.07%的预测可靠性。综上所述,OLPM的轻量化架构能够在不影响用户体验的前提下,实现输出长度预测收益的最大化。

### 3.5　消融实验分析

为系统评估OLPM框架中各技术方法的贡献,本研究针对任务复杂度高、长度分布广的ELI5数据集开展消融实验,通过逐一移除关键方法并对比性能变化,结果如表3所示。

**表3　消融实验对比结果**

Table 3　Comparison of ablation experiments

| 消融策略 | 预测准确率/% | 收敛轮次 |
|---|---|---|
| OLPM | 79.1 | 38 |
| 移除正态分布过滤策略 | 64.8 | 61 |
| 移除语义保留提示增强算法 | 36.3 | 52 |
| 对tokens进行回归 | 52.5 | 78 |

#### 3.5.1　正态分布过滤策略分析

完整的OLPM模型能够达到79.1%的最高准确率、最低验证损失以及最快的收敛速度,验证了整体架构设计的有效性与高效性。移除正态分布过滤策略后,准确率下降至64.8%,收敛速度明显变慢,收敛轮次延长至61轮。这主要是由于原始ELI5数据集中存在约5%的极端异常样本,这些样本的长度分布超过了正常范围,导致模型对正常样本的预测出现偏差。而正态分布过滤策略通过统计每个任务的长度分布特征,能够自适应地剔除极端样本,并保留符合任务典型规律的数据,使模型学习到更稳健的长度关联规律,最终提升训练数据质量与预测稳定性。

#### 3.5.2　语义保留提示增强算法分析

当移除语义保留提示增强模块后,模型精度从79.1%骤降至36.3%,收敛轮次从38轮延长至52轮。这一结果表明,该模块是OLPM捕捉长度控制语义线索的核心组件,模型高度依赖此类显式语言信号进行长度判断。在原始训练数据中,长度相关指令出现频率较低且分布稀疏,导致模型难以建立语言风格与输出长度通过语义增强生成的训练样本,为模型提供了大量明确的“语义-长度”配对特征,使其能够精准识别影响输出长度的关键短语。

值得注意的是,该模块的有效性在很大程度上依赖于所用触发词库的质量与代表性:若缺乏语义清晰、倾向一致的长度引导词,模型将难以从原始稀疏信号中有效学习语义-长度映射关系,转而依赖输入长度等浅层统计特征,从而在面对复杂或多样化的提示时性能明显下降。

#### 3.5.3　预测区间分类策略分析

将预测任务替换为对token数的直接回归后,模型精度下降至52.5%,收敛轮次大幅延长至78轮。这一结果验证了区间分类策略的合理性:模型的推理输出受随机因素影响存在波动,对精确token数进行回归会放大这种波动的负面影响。此外,回归任务对异常样本更敏感,极端值会显著拉低整体预测精度,而预测区间分类能够通过离散化的手段降低异常值的干扰。因此,预测区间分类策略是OLPM实现高鲁棒性预测的关键,更符合泛在场景下的实际资源管理与调度需求。

### 3.6　模型预测置信度分析

在实际部署中,预测结果的可靠性与可信度对于调度决策至关重要,本节对OLPM模型的预测置信度进行分析。模型输出的类别概率可直接作为预测置信度的度量。实验统计发现,OLPM在测试集上的平均预测置信度高达0.91,表明模型对绝大多数样本的分类决策具有高度信心。图6所示为置信度与预测准确度统计结果,当模型预测置信度大于0.85时,其对应样本的预测准确率超过98%。在置信度为0.7~0.85的区间内,准确率仍保持在85%以上,表明模型的高置信度输出对应着高准确性预测结果。

此外,本研究分析了不同长度区间的置信度分布。如图6所示,对于“短”和“中”区间,模型表现出更高的置信度,分别为0.93和0.90,对于“长”区间内的预测置信度为0.85,主要是由于长输出的生成过程更易受模型内部随机性和复杂上下文影响,模型对其预测的不确定性也相应增加。基于上述分析可知,OLPM不仅能提供高精度的预测,还能输出可靠的置信度评估,能够在泛在计算环境中提供有效的决策支持。

### 3.7　OLPM驱动的推理调度性能评估

基于SGLang集成环境,对比OLPM与Random、RoundRobin、CacheAware 3种主流调度策略的端到端性能,在低并发场景下,由于系统计算资源与显存相对充裕,各类调度策略的吞吐量性能差异较小,随着并发量增加,资源竞争加剧,OLPM的调度优势开始显著体现。因此,本文重点分析高并发极端场景下OLPM的性能表现,核心测试结果如表4所示。

实验结果表明,OLPM在各类场景下均展现出了优越的调度性能,尤其在负载构成复杂的Mixed场景中优势最为明显,在这类包含Chat、Summary和Cre-

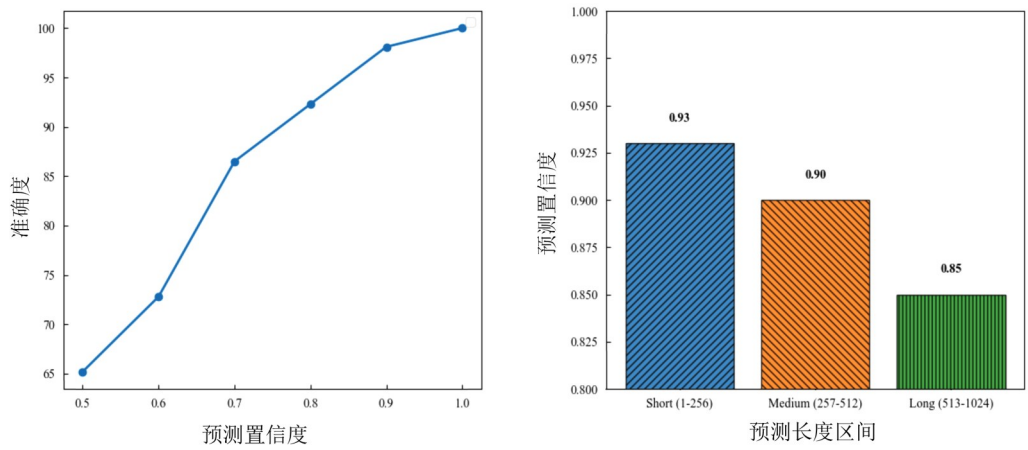



图6 OLPM预测置信度与准确度统计结果

Figure 6 Statistical results of prediction confidence and accuracy of OLPM

**表4 融合不同调度策略的推理性能对比结果**

Table 4 Comparison of inference performance using different scheduling strategies

| 场景 | 指标 | Random | RoundRobin | CacheAware | OLPM | 提升幅度*/% |
|---|---|---|---|---|---|---|
| Chat | 请求吞吐量/(req·s$^{-1}$) | 136.82 | **160.81** | 132.28 | 175.40 | +9.1 |
| | 端到端延迟/ms | 6 537 | **6 045** | 6 332 | 5 210 | −13.8 |
| Summary | 请求吞吐量/(req·s$^{-1}$) | 29.03 | 33.30 | **76.50** | 84.20 | +10.1 |
| | 端到端延迟/ms | 28 535 | 28 106 | **11 762** | 10 540 | −10.4 |
| Creative | 请求吞吐量/(req·s$^{-1}$) | 0.23 | **0.25** | 0.24 | 0.31 | +24.0 |
| | 端到端延迟/ms | 2 764 053 | **2 662 278** | 2 741 067 | 2 150 300 | −19.2 |
| Mixed | 请求吞吐量/(req·s$^{-1}$) | **0.61** | 0.58 | 0.58 | 0.78 | +27.9 |
| | 端到端延迟/ms | 435 767 | **395 320** | 395 925 | 308 420 | −22.0 |

注:*提升幅度计算基于各场景下表现最好的基线指标(加粗项)进行对比。

ative的混合场景中,传统策略由于缺乏对请求输出长度的感知,极易将长输出请求与短请求不合理地打包,导致严重的队头阻塞。OLPM凭借精准的长度预测,能够将吞吐量从基线的0.61 req/s提升至0.78 req/s,并将端到端延迟降低了约22%。这证明了OLPM能够有效识别请求特征,实现“长短分离”的精细化调度。

在Chat场景中,尽管任务简单,但在1 024高并发下,显存碎片化成为瓶颈。OLPM通过预测输出长度,使得调度器能够更紧凑地规划内存页,从而支持更大的批处理大小。实验数据显示,OLPM的吞吐量达到175.40 req/s,优于表现最好的RoundRobin策略,证明了长度感知对提升并发容量的有效性。在Summary和Creative场景中,CacheAware策略虽然通过缓存前缀提升了性能,但OLPM进一步通过预测输出长度,避免了因显存预估不足导致的中间计算结果换入换出。特别是在Creative场景中,OLPM成功将吞吐量提升至0.31 req/s,显著优于其他策略,验证了其在处理长尾负载时的稳定性。

此外,OLPM作为轻量化插件嵌入SGLang后,未显著增加系统开销,且无需额外硬件资源,可直接复用现有节点的空闲算力。与未集成预测模型的调度策略相比,OLPM在保持系统原有部署复杂度的前提下,实现了吞吐量与延迟的协同优化,尤其在高并发、异构负载场景中性能衰减幅度更小,充分证明了OLPM在真实推理系统中的部署可行性与技术优势。

## 4 结论

本研究针对泛在计算场景的精细化调度需求,提出一种轻量级的预测模型训练框架,通过将输出长度分类为离散区间,为多模型、多算力形态、服务质量驱动的模型服务优化提供支撑,提升泛在场景下的计算效率。基于正态分布的多层级数据过滤技术保障训练数据质量、提升训练稳定性,利用基于保留语义的提示增强算法强化模型对长度控制语言线索的敏感度,实现“数据质量-语义感知”双重优化。该模型在多种不同任务集上表现优异,能够为优化模型推理服务提供核心调度依据,满足泛在计算场景下多模型协同、多算力适配、服务质量保障的精细化调度需求。值得指出的是,OLPM的设计遵循“预测-调度解耦”原则,能够作为可插拔模块无缝嵌入主

流推理调度框架,无需修改推理引擎的底层处理逻辑。OLPM可凭借较高的预测精度进一步提升推理系统的吞吐与响应效率,兼具学术创新性与工程落地价值。

未来研究将围绕轻量化深化、泛化能力拓展与多场景适配3大方向展开,进一步提升模型的实用性与部署灵活性:一方面,持续探索轻量化优化路径,通过模型量化与剪枝技术,在保持预测精度的前提下进一步降低参数量与计算复杂度,适配嵌入式设备等极端资源受限场景,并针对短文本等特定任务引入更精简的Transformer结构,增强部署场景的适配性;另一方面,针对泛在场景下更多样的模型类型与算力形态,自适应优化策略参数与模型架构,重点探索OLPM在不同架构模型上的零样本或少样本迁移能力,构建泛化性强的通用输出长度预测模型,同时将当前固定过滤阈值升级为基于在线负载与历史误差反馈的自适应策略,提升系统鲁棒性,此外还将把预测框架延伸至图文生成、语音合成等多模态推理场景,实现对图像分辨率、音频时长等多模态指标的精准预测,为更多泛在计算场景下的智能服务提供技术支撑。

**参考文献**


[1] 章晋睿, 龙婷婷, 张德宇, 等. 端智能推理加速技术综述[J]. 电子学报, 2025, 53(4): 1063-1102.
Zhang Jinrui, Long Tingting, Zhang Deyu, et al. On-device intelligence acceleration technologies: A survey[J]. Acta Electronica Sinica, 2025, 53(4): 1063-1102. (in Chinese)

[2] Wu Haijie, Lin Weiwei, Shen Wangbo, et al. Prediction of heterogeneous device task runtime based on edge server-oriented deep neuro-fuzzy system[J]. IEEE Transactions on Services Computing, 2025, 18(1): 372-384.

[3] Byun J, Choi Y, Lee J, et al. Privacy-preserving inference resistant to model extraction attacks[J]. Expert Systems with Applications, 2024, 256: 124830.

[4] 杨赟辉, 程虎, 魏敬和, 等. 面向Transformer模型边缘端部署的常用激活函数高精度轻量级量化推理方法[J]. 电子学报, 2024, 52(10): 3301-3311.
Yang Yunhui, Cheng Hu, Wei Jinghe, et al. High-precision lightweight quantization inference method for prevalent activation functions in Transformer models in edge device deployment[J]. Acta Electronica Sinica, 2024, 52(10): 3301-3311. (in Chinese)

[5] Chen Huajie, Zhu Tianqing, Ji Shouling, et al. Stand-in model protection: Synthetic defense for membership inference and model inversion attacks[J]. Knowledge-Based Systems, 2025, 316: 113339.

[6] Chen Yuxuan, Li Rongpeng, Yu Xiaoxue, et al. Adaptive layer splitting for wireless large language model inference in edge computing: A model-based reinforcement learning approach[J]. Frontiers of Information Technology & Electronic Engineering, 2025, 26(2): 278-292.

[7] Yang Jin, Wu Qiong, Feng Zhiying, et al. Quality-of-service aware LLM routing for edge computing with multiple experts[J]. IEEE Transactions on Mobile Computing, 2025, 24(12): 13648-13662.

[8] Barkalov A, Lemeshko O, Yeremenko O, et al. Solving load balancing problems in routing and limiting traffic at the network edge[J]. Applied Sciences, 2023, 13(17): 9489.

[9] Seo M, Hyun J, Jeong S, et al. OASIS: Outlier-aware KV cache clustering for scaling LLM inference in CXL memory systems[J]. IEEE Computer Architecture Letters, 2025, 24(1): 165-168.

[10] Cui Lixiao, He Kewen, Li Yusen, et al. SwapKV: A hotness aware in-memory key-value store for hybrid memory systems[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(1): 917-930.

[11] Lu Jinkang, Lv Meng, Li Peixuan, et al. Dhcache: A dual-hash cache for optimizing the read performance in key-value store[J]. The Journal of Supercomputing, 2025, 81(2): 400.

[12] He Ying, Fang Jingcheng, Yu F R, et al. Large language models (LLMs) inference offloading and resource allocation in cloud-edge computing: An active inference approach[J]. IEEE Transactions on Mobile Computing, 2024, 23(12): 11253-11264.

[13] Li Yandi, Guo Jianxiong, Tang Zhiqing, et al. Cloud-edge system for scheduling unpredictable LLM requests with combinatorial bandit[J]. IEEE Transactions on Services Computing, 2025, 18(6): 3567-3580.

[14] Hu Yitao, Liu Xiulong, Yang Guotao, et al. Tightllm: Maximizing throughput for llm inference via adaptive offloading policy[J]. IEEE Transactions on Computers, 2025, 74(7): 2195-2209.

[15] Hu Mingzhu, Qin Shengfeng, Wang Shuying, et al. An energy-saving real-time scheduling method based on bi-level multi-agent architecture with bargaining game for flexible job shops[J]. Expert Systems with Applications, 2025, 269: 126527.

[16] Tang Xuhao, Liu Fagui, Xu Dishi, et al. LLM-assisted reinforcement learning: Leveraging lightweight large language model capabilities for efficient task scheduling in multi-cloud environment[J]. IEEE Transactions on Consumer Electronics, 2025, 71(2): 5631-5644.

[17] Zhang Xinyuan, Nie Jiangtian, Huang Yudong, et al. Beyond the cloud: Edge inference for generative large language models in wireless networks[J]. IEEE Transactions

on Wireless Communications, 2025, 24(1): 643-658.
[18] Dai Haixing, Liu Zhengliang, Liao Wenxiong, et al. AugGPT: Leveraging ChatGPT for text data augmentation[J]. IEEE Transactions on Big Data, 2025, 11(3): 907-918.

## 附录A 长度控制触发词库

| 类型 | 中文核心词 | 中文同义变体 | 英文核心词 | 英文同义变体 | 长度倾向 | 核心语义说明 |
|---|---|---|---|---|---|---|
| 短指令词 | 简要 | 简略、简述、简要说明 | Brief | Concise, Brief description | 短 | 要求简洁表述,不展开细节 |
| 短指令词 | 概括 | 归纳、概述、总结 | Summarize | Conclude, Outline, Summarization | 短 | 提炼核心内容,归纳要点 |
| 短指令词 | 一句话 | 一句话概括、简短一句话 | In one sentence | In a single sentence | 短 | 限制输出为单句,极致精简 |
| 短指令词 | 简略 | 简明、简要概括 | Succinct | Terse, Compact | 短 | 简明扼要,省略次要信息 |
| 短指令词 | 简述 | 简要陈述、简要概述 | Outline | Sketch, Brief account | 短 | 简要陈述核心逻辑,不深入阐述 |
| 短指令词 | 精炼 | 凝练、精简表述 | Refine | Condense, Streamline | 短 | 凝练关键信息,去除冗余 |
| 短指令词 | 核心要点 | 关键点、核心内容 | Key points | Core points, Main points | 短 | 仅呈现核心结论或关键信息点 |
| 短指令词 | 快速概述 | 快速总结、简要梳理 | Quick overview | Rapid summary | 短 | 快速呈现整体框架,不展开细节 |
| 中等长度指令词 | 适度说明 | 适中阐述、合理说明、适度表述 | Moderate explanation | Appropriate elaboration, Reasonable description | 中 | 合理展开,兼顾细节与简洁 |
| 中等长度指令词 | 适中回答 | 适度答复、合理回应 | Moderate answer | Appropriate response, Reasonable reply | 中 | 回答全面且不冗长,符合常规需求 |
| 中等长度指令词 | 简要展开 | 适度展开、简要阐述 | Briefly expand | Moderately elaborate, Briefly develop | 中 | 在核心要点基础上适度延伸 |
| 中等长度指令词 | 合理概述 | 适度概述、合理总结 | Reasonable overview | Moderate summary, Appropriate outline | 中 | 全面概述,关键细节补充 |
| 中等长度指令词 | 适中分析 | 适度分析、合理剖析 | Moderate analysis | Appropriate analysis, Reasonable examination | 中 | 分析核心问题,不做深度拓展 |
| 中等长度指令词 | 简要详述 | 适度详述、简要深入 | Briefly detail | Moderately detail, Briefly elaborate | 中 | 表述关键细节,平衡全面性与简洁性 |
| 中等长度指令词 | 适中梳理 | 适度梳理、合理整理 | Moderate sorting | Appropriate sorting, Reasonable organization | 中 | 有条理呈现信息,长度适中 |
| 中等长度指令词 | 合理说明 | 适中说明、适度阐述 | Reasonable explanation | Moderate explanation, Appropriate statement | 中 | 清晰说明逻辑,不冗余不简略 |
| 长指令词 | 详细说明 | 详细阐述、详细描述 | Detailed explanation | Elaborate description | 长 | 全面展开,覆盖细节与逻辑 |
| 长指令词 | 展开论述 | 深入论述、全面论述 | Expand on | Elaborate, Discuss in depth | 长 | 深入阐述,结合论据与分析 |
| 长指令词 | 全面分析 | 深入分析、详细剖析 | Comprehensive analysis | In-depth analysis | 长 | 多角度分析,覆盖利弊与细节 |
| 长指令词 | 尽可能详尽 | 详尽说明、充分阐述 | As detailed as possible | As thorough as possible | 长 | 最大化呈现信息,不遗漏关键细节 |
| 长指令词 | 详细回答 | 详尽答复、全面回应 | Detailed answer | Thorough response | 长 | 全面回应问题,补充背景与延伸 |
| 长指令词 | 深入说明 | 深入阐述、细致说明 | In-depth explanation | Detailed elaboration | 长 | 深度解析,结合原理与实例 |
| 长指令词 | 细致分析 | 细致剖析、详细分析 | Meticulous analysis | Detailed analysis | 长 | 精细化剖析,覆盖微观细节 |
| 长指令词 | 详尽描述 | 详细描述、完整描述 | Detailed description | Complete account, Comprehensive depiction | 长 | 完整呈现过程与细节,无信息省略 |

## 作者简介


**牛昊一**　男,1995年10月出生于河北省张家口市。2023年毕业于浙江大学控制科学与工程学院。现为之江实验室助理研究员。主要研究方向为人工智能和智能优化算法。
E-mail: hyniu95@gmail.com

**张北北**　男,1994年2月出生于黑龙江省五常市。2020年毕业于多伦多大学电子及计算机工程系。现为之江实验室高级研究专员。主要研究方向为智能化软件工程与系统。
E-mail: beibei@zhejianglab.org

**林　露**　男,1986年9月出生于湖北省黄冈市。2011年毕业于武汉大学计算机系。现为之江实验室高级研究专员。主要研究方向为大模型推理技术。
E-mail: limiximil@zhejianglab.org

**邹　涛**　男,1974年11月出生于重庆市。现为之江实验室研究专家、研究员。主要研究方向为智能计算和高性能网络。
E-mail: zout@zhejianglab.org

**裘向明**　男,1987年12月出生于黑龙江省哈尔滨市。2010年毕业于清华大学自动化系。现为之江实验室助理研究员。主要研究方向为非线性优化方法。
E-mail: xixiangming@gmail.com

**高　丰**　男,1974年11月出生于浙江省丽水市。2002年毕业于浙江大学信息与电子工程学院。现为之江实验室高级工程师。主要研究方向为分布式机器学习、云计算、边缘计算。
E-mail: gaof@zhejianglab.org