Computer Science and Application 计算机科学与应用, 2026, 16(1), 44-55
Published Online January 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.161005

Hans汉斯

一种面向资源受限环境的高吞吐量、置信度门控的细粒度文本分类框架

董致男[1]，潘昱暄[2]，黎阳诗[3]

[1]纽约大学坦登工程学院，美国 纽约
[2]辽宁工程技术大学软件学院，辽宁 葫芦岛
[3]Oracle，美国 奥斯汀

收稿日期：2025年12月3日；录用日期：2026年1月2日；发布日期：2026年1月12日

摘　要

针对通用大语言模型在资源受限环境下处理大规模细粒度文本分类任务时面临的效率瓶颈与幻觉问题，本文提出了一种名为DeepConf-Verify (DCV)的高性能框架。该框架首先通过领域微调将小参数量模型的准确率基线从不足30%显著提升至90%以上；进而引入双阈值动态置信度门控机制，利用词元级置信度轨迹实时监控生成过程，实现对“困惑”样本的立即熔断和对高确信样本的快速通行；最后，对处于临界置信度区间的样本执行双模型一致性验证以消除尾部风险。实验结果表明，在单张NVIDIA A100 GPU受限条件下，DCV框架在保持95.2%企业级准确率的同时，相比原有的通用大模型系统实现了超过1200%的吞吐量提升(达60.2条/秒)，相比同参数量的单一微调模型亦有24%的效率优化。系统成功支持日处理超过500万条评论数据，并将人工审核率控制在4.5%以内。本研究为在低资源环境下构建高吞吐、高可靠的垂直领域AI系统提供了有效的理论与实践范式。

关键词

细粒度文本分类，资源受限环境，置信度门控，高吞吐量，双模型验证

A Confidence-Gated Framework for High-Throughput Fine-Grained Text Classification in Resource-Constrained Environments

Zhinan Dong[1], Yutong Pan[2], Yangshi Li[3]

[1]Tandon School of Engineering, New York University, New York, USA
[2]School of Software, Liaoning Technical University, Huludao Liaoning

文章引用: 董致男, 潘昱暄, 黎阳诗. 一种面向资源受限环境的高吞吐量、置信度门控的细粒度文本分类框架[J]. 计算机科学与应用, 2026, 16(1): 44-55. DOI: 10.12677/csa.2026.161005

[3]Oracle, Austin, USA

Received: December 3, 2025; accepted: January 2, 2026; published: January 12, 2026

Abstract

To address the efficiency bottlenecks and hallucination issues faced by general-purpose Large Language Models (LLMs) in handling large-scale, fine-grained text classification tasks within resource-constrained environments, this paper proposes a high-performance framework named DeepConf-Verify (DCV). Building upon domain-specific fine-tuning, which elevates the accuracy baseline of small-parameter models from under 30% to over 90%, the framework integrates a Dual-Threshold Dynamic Confidence Gating mechanism. This mechanism utilizes token-level confidence trajectories to monitor the generation process in real-time, executing an immediate "Panic Exit" for "confused" samples and a "Fast Pass" for high-confidence samples. Furthermore, a Dual-Model Verification protocol is employed to enforce consensus on samples within critical confidence intervals, thereby mitigating tail risks. Experimental results on a single NVIDIA A100 GPU demonstrate that DCV achieves an enterprise-grade accuracy of 95.2%. Notably, it boosts throughput by over 1200% (reaching 60.2 comments/sec) compared to the original general-purpose LLM system, and achieves a 24% efficiency optimization compared to a single fine-tuned model of equivalent parameter size. The system successfully scales to process over 5 million comments daily while keeping the manual audit rate within 4.5%. This study provides a robust theoretical and practical paradigm for constructing high-throughput and reliable vertical-domain AI systems in low-resource settings.

Keywords

Fine-Grained Text Classification, Resource-Constrained Environments, Confidence-Gated, High-Throughput, Dual-Model Verification

1. 引言

在如今这个数据驱动、竞争激烈的市场环境中，白色家电制造商面临着快速适应日益变化的数据需求的挑战[1] [2]。我们发现，通过从社交媒体平台(如抖音、小红书等)收集大量非结构化用户反馈并深入分析，已成为提升企业竞争力的重要途径[3]。通过这样的方式，企业不仅能够主动保障产品质量，及早发现潜在的产品问题并避免演变为危机，还能精准地评估公众情绪，进而优化市场战略[4]。不过，处理这些数据的量是非常庞大的，例如我们的客户每天需要分析超过 500 万条评论。在如此巨大的数据量和处理速度面前，传统的人工分析或低效的计算方法已经远远无法满足需求，因此迫切需要一种自动化、高效且精确的解决方案[5]。

然而，目前许多家电制造商依旧依赖未经过优化的通用大语言模型(参数量高达 140 亿)来处理这些高度专业化的任务，比如将用户评论准确地归类到具体的行业类别(如“冰箱压缩机异响”、“洗衣机脱水程序报错 E4”等) [6]。在调研中，我们发现这种做法存在两个主要问题。首先，庞大且计算密集的模型导致它们在有限的硬件资源下，每天只能处理约 40 万条评论，而客户每日的评论量却高达 500 万条，这就造成了数据积压，影响了企业响应市场变化的速度，错失了决策的最佳时机。其次，由于缺乏对家

电行业知识的深入理解，通用模型在细致分类任务中的准确率不到 30%，这不仅使得分析结果不准确，甚至可能误导企业决策，带来不小的商业风险。

为了解决这些问题，我们提出了一个名为 DeepConf-Verify (DCV)的解决方案。这个框架结合了三项关键技术：首先，通过行业专有的高质量数据对小型模型进行微调，取代了庞大的通用模型，使得模型能更好地理解行业术语和用户的表达习惯[7]；其次，采用动态置信度门控机制，实时监控模型的置信度，提前终止那些低质量的计算，从而提高处理效率[8]；最后，采用双模型验证机制，通过同时使用两个独立的模型并要求其结果一致，确保分类结果的可靠性。这些技术的结合有效提高了分类任务的效率、准确性，并提升了系统的整体稳定性[9]。

2. 背景与相关工作

2.1. 文本分类技术的演进：从统计方法到专业化模型

我们首先简要回顾文本分类技术的发展历程，为我们的方法选择提供历史和理论背景。众所周知，早期的统计方法，如词袋模型(Bag-of-Words)和 TF-IDF，为该领域奠定了基础，但它们在捕捉复杂语义关系方面存在固有局限性[10]。我们很快将讨论的重点转向了深度学习模型，特别是基于循环神经网络(RNN)和卷积神经网络(CNN)的架构，它们在特征自动提取方面展现了巨大的优势[11]。我们认为，随着 Transformer 架构的出现和大规模预训练模型(如 BERT)的兴起，文本分类的性能被推向了新的高度。然而，需要指出的是，尽管这些大模型代表了技术的巨大飞跃，但它们在企业级应用中的真正潜力，需要通过专业化才能完全释放。我们引用了关于模型微调和小型语言模型(SLM)兴起的研究，将此视为在资源受限环境中实现高性能的关键趋势。我们指出，这一趋势直接为我们放弃客户原有的庞大通用模型、转向小型专业化模型的决策提供了坚实的理论依据。

2.2. 推理加速：自适应计算与提前退出的兴起

我们接下来将探讨推理延迟这一在实际部署中至关重要的问题。我们引入了自适应计算的概念，即根据输入样本的内在复杂度动态调整模型的计算资源消耗[12]。由此引出对“提前退出”机制的深入讨论，该机制允许模型在处理较简单的输入时，从其中间层直接产生输出，从而避免了不必要的深层计算，显著加速了推理过程[13]。我们观察到，现有的大多数提前退出机制主要通过在模型的不同层次设置退出点来操作，即在模型的不同深度设置退出点。我们认为，虽然这种方法有效，但对于文本分类这类需要快速决策的任务，退出点的粒度可能过于粗糙，无法及时终止计算[14]。我们的工作在此基础上提出了一项关键创新：将自适应计算的思想应用到更细致的词元级别，通过监控分类决策形成过程中的内部状态，实现了一种更为动态和灵敏的提前终止策略。据我们所知，这是首次将此类细粒度的实时监控机制应用于加速文本分类任务。

2.3. 可靠性保障：集成方法与 NLP 中的验证机制

我们为双模型架构建立了坚实的学术基础。我们回顾了集成学习的基本原理，如装袋法和提升法，这些方法通过组合多个弱学习器来构建一个强大的、鲁棒性更强的学习器。具体来说，集成方法的核心优势在于通过模型多样性来降低预测的方差。随后我们将这一概念与模型验证机制联系起来，并引用了相关工作，在这些工作中，我们使用多个模型进行交叉验证，从而识别和分流那些模型间存在分歧的疑难输入样本[15]。我们认为，这一系列研究为双模型验证系统提供了强有力的理论支持，证明了它并非一种冗余的计算开销，而是一种最大化准确性、并为系统内置不确定性处理能力的原则性方法[16]。

我们认识到，领域专业化微调、自适应推理加速和基于集成的可靠性保障这三个独立的研究趋势正

在趋于融合，共同催生了企业级人工智能的一个新范式[17]。过去，企业在部署 AI 时常常被迫在几个选项中做出妥协：选择速度快但准确率低的简单模型，选择准确率高但速度慢的大型模型，或者选择性能强大但开发复杂的定制化集成系统[18]。我们的研究则展示了将这三种趋势统一到一个内聚框架中的可能性。通过这种方式，构建了一个能够同时提供速度、准确性和效率的系统，而无需在任何一方面做出重大妥协。我们首先通过微调技术构建了一个高性能的基座模型，然后利用置信度门控机制使其运行得更快，最后通过双模型验证协议确保其结果的可靠性。我们相信，这种综合性的设计理念本身就是一项重要的创新。

3. DeepConf-Verify (DCV)分类框架

3.1. 基础：领域特定的模型专业化

首先将阐述模型选择与微调的过程，这是 DCV 框架的基石。我们选择了一个紧凑的、参数量为 30 亿的预训练语言模型作为我们的基础架构。我们做出这一选择，是经过对模型表达能力、客户现有 GPU 的显存容量及处理能力限制之间进行综合权衡后得出的最优解。

我们遵循了模型微调的最佳实践来设计我们的微调流程。具体步骤如下：

1. 数据策管与预处理：我们使用了客户提供的高质量、包含超过 100 万条已标注的历史评论数据集。我们认为，这是模型学习领域知识不可替代的关键资源。同时，我们构建了一条严谨的数据清洗与预处理管道，并参考了标准技术，该管道包括：(1) 分词；(2) 对网络俚语和平台特有黑话(例如“YYDS”、“绝绝子”)进行归一化处理；(3) 移除表情符号、URL、HTML 标签等对分类任务无益的无关噪声。

2. 监督微调：我们在文本分类任务上对模型进行了监督微调。采用了 AdamW 优化器，并结合了倾斜三角形学习率策略，以实现高效且稳定的收敛[19] [20]。我们通过调整模型权重，使其能够深入理解与白色家电相关的 400 个细分领域的语义。需要特别指出的是，我们从同一个基础架构出发，但使用不同的随机种子独立训练了两个模型(模型 A 和模型 B)。因为，这种做法能够在保持两个模型高性能的同时，引入足够的随机性，使它们在面对模糊或边缘案例时可能产生不同的“观点”，从而为后续的双模型验证环节提供有益的、必要的模型差异性。

3.2. 吞吐量提升：动态置信度门控

动态置信度门控是我们研究的核心机制。我们提出了一个核心假设：即使是对于看似“一步到位”的分类任务，模型在输出最终类别之前，其内部也会生成一系列的中间状态和对后续 token 的概率分布[21]。我们认为，这个内部的“思考”过程的置信度轨迹，蕴含着关于输入样本难度的宝贵信息。我们的机制正是为了实时监控并利用这一内部过程而设计的。

为此，我们定义了组置信度 C_{G_i} 这一关键指标，用于量化模型在处理文本时局部片段的确定性。其数学表达式如下方公式(1)所示：

$$C_{G_i} = \frac{1}{|G_i|}\sum_{t\in G_i} C_t \tag{1}$$

其中，C_t 是单个 token 的置信度，而 G_i 是一个包含多个 token 的滑动窗口。我们进一步定义了最低组置信度 C_{least}，它代表了整个处理过程中最不确定的那个片段的置信度，其数学表达式如下方公式(2)所示：

$$C_{least}(t) = \min_{G_j\in\mathcal{G}} C_{G_j} \tag{2}$$

该机制的实现方式为：当模型处理一条输入评论时，系统会实时监控一个大小为 W (设为 64 token)的滑动窗口内的组置信度。对于长度不足 W 的短文本输入，则计算当前已生成的所有 token 的平均置信度作为替代指标。为了平衡效率与处理能力，我们引入了动态置信度阈值策略。设定统一置信度阈值 s。

当模型 A 生成过程中，实时计算的最低组置信度 C_{least} 出现以下情况时执行不同操作：

1. 立即熔断：若 $C_{least} < s$ (设定为 0.75)，判定模型完全‘困惑’，继续计算毫无意义。此时立即终止当前模型的生成过程，并将该样本标记为‘待验证’，无缝路由至双模型验证阶段(即激活模型 B)，以避免错误输出。这一机制避免了模型 A 在低置信度样本上浪费计算资源，构成了吞吐量提升的主要收益。

2. 完整推理：若 $C_{least} \geq s$，模型 A 将完成完整的生成过程，产生初步结果 R_A，随后进入第 3.3 节的验证判断流程。

图 1 展示了模型 A 内部的动态门控逻辑(注：为简化展示，图 1 仅描述了早停触发后的逻辑终点，完整的双模型流转见图 2)。

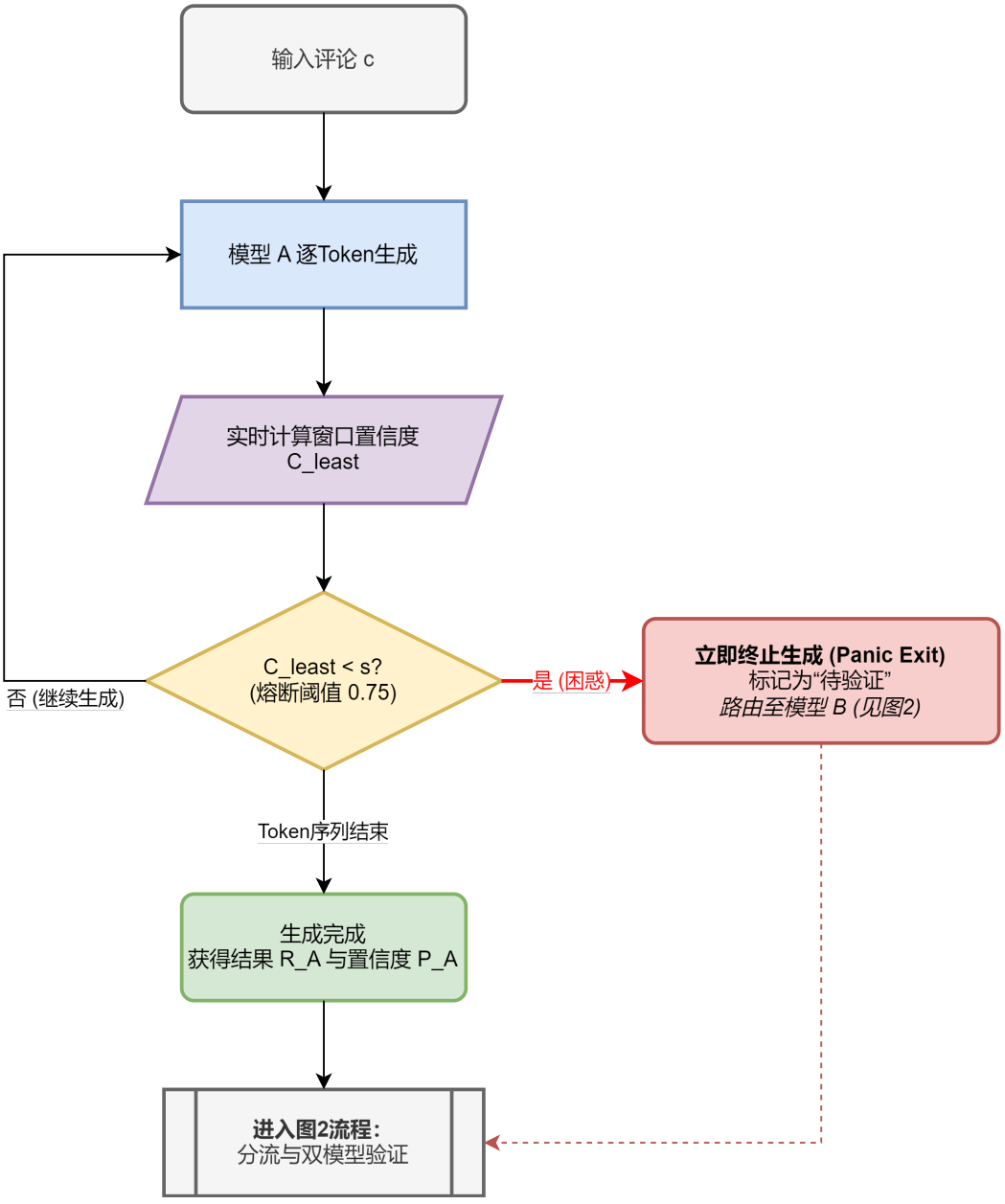

Figure 1. Dynamic confidence gating implementation
图 1. 动态置信度门控实现方式

我们将这种方法命名为“动态置信度门控”，以强调其在推理过程中的实时性和自适应性。这一应用表明，即使是对于一个快速的分类决策，模型的内部“思考过程”也存在一个可被测量的置信度轨迹，并且我们可以通过拦截这一轨迹来优化系统性能。这种方法比仅仅对模型的最终输出概率设置阈值要精妙得多。其背后的逻辑是，当模型在“阅读”和“构思”答案的过程中置信度出现骤降，这往往预示着它难以理解输入文本的语义，这种情况远早于它最终给出一个(很可能是低概率的)答案。因此，我们认为这是一个预示歧义的“领先指标”，使其成为实现提前终止的强大工具。

3.3. 可靠性保障：双模型验证

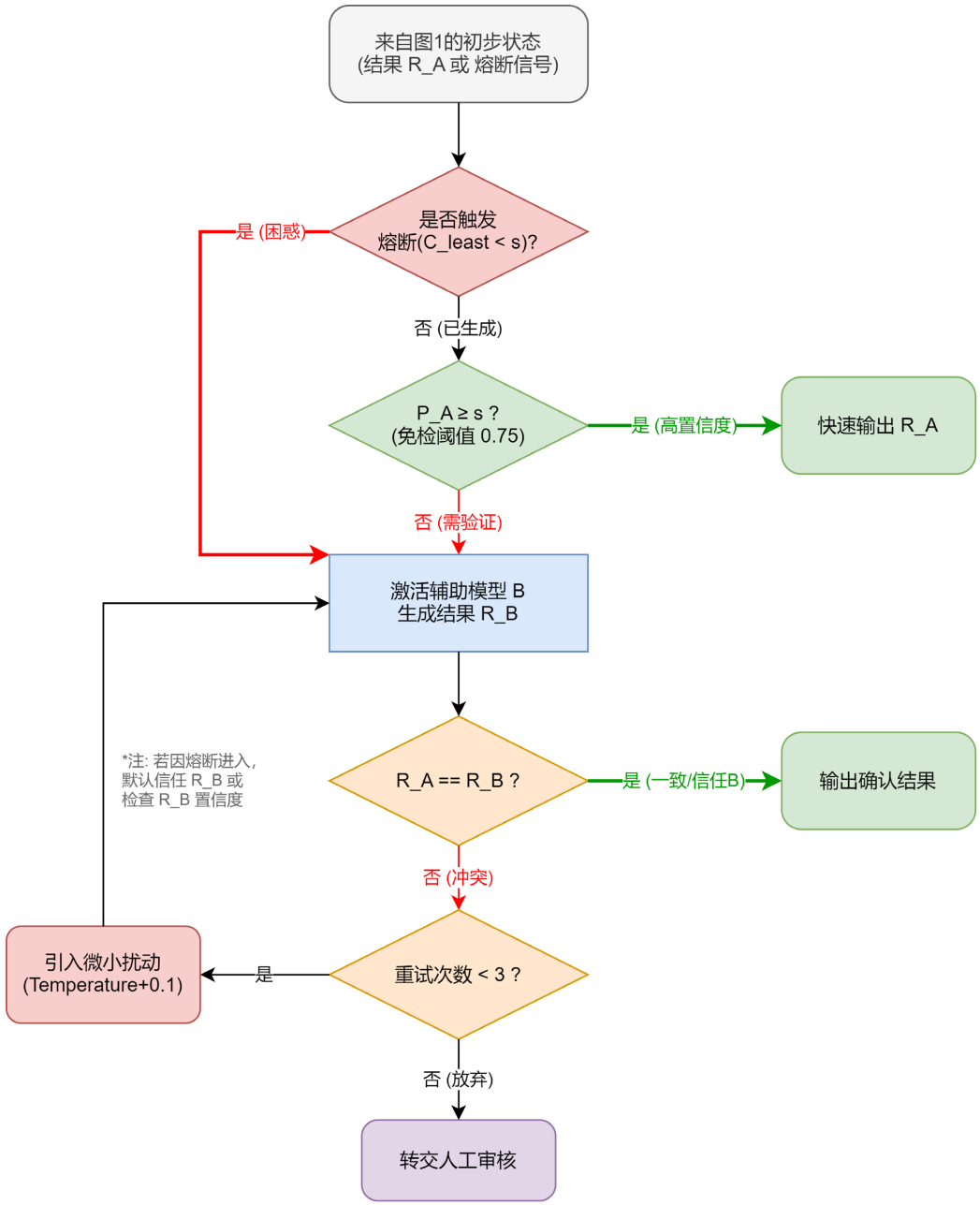

Figure 2. DCV adaptive cascading verification process
图 2. DCV 自适应级联验证流程

接下来描述我们系统的架构，这是保障最终结果高度可靠的关键。我们设计让每一条成功通过了置信度门控的输入评论，我们采用级联验证策略：

1. 快速路径：如果模型 A 顺利完成了生成，且全程 $C_{\text{least}} \geq s$，系统直接输出结果。这覆盖了约 85%的样本。

2. 补救路径：如果模型 A 因置信度过低触发了“早停”，或者最终结果置信度存疑，该样本将被无缝路由至模型 B 进行二次推理。

我们通过下面这个算法来阐述验证逻辑：

1. 接收输入：接收输入评论 c。

2. 主模型推理：令主模型 A 对 c 进行分类，得到初步结果 R_A 和置信度 P_A。

置信度门控判断：

1. 如果 $P_A \geq s$ (验证阈值，设为 0.75)，判定模型 A 结果可信，直接输出 R_A，流程结束。

2. 如果 $P_A < s$，判定样本存在歧义，进入级联验证阶段。

3. 辅助模型介入：激活辅助模型 B 对 c 进行分类，得到结果 R_B。

4. 一致性校验：比较 R_A 与 R_B。如果 $R_A = R_B$，确认该分类结果并输出。

5. 冲突消解(重试机制)：如果结果不匹配，启动重试循环(上限调整为 3 次)。在重试中对解码参数引入微小扰动。

6. 若在重试内达成一致，输出确认结果。

7. 若 3 次重试后仍无法达成一致，判定为难例，转交人工审核。

图 2 来展示完整的 DCV 框架工作流程。

这套双模型验证系统在实际运行中，隐式地构建了一个强大的主动学习循环[22]。那些因置信度不足或模型分歧而被持续标记并送往人工审核的样本，根据其定义，正是数据集中最困难、最模棱两可、信息量最大的案例。我们发现，这个经过系统自动筛选出的高质量数据流，对于后续的模型迭代微调而言，其价值远高于随机抽样的评论。这使得我们的系统能够进入一个“自我进化”的良性循环：系统从自身处理过程中的不确定性中学习，从而随着时间的推移变得更加鲁棒和智能。我们将验证机制与主动学习原则联系起来，因为我们系统的人工审核队列实际上构成了一个高质量的主动学习样本池，为模型的持续改进提供了宝贵的数据资源。

4. 实验设计

4.1. 数据集与评估指标

关于实验数据集的构建，为了在资源受限环境下进行可复现的评估，我们构建了一个包含 150 万条评论的合成增强数据集。这一规模既足以覆盖 400 个细分领域的长尾特征，又符合低成本训练的约束。我们特意将这 400 个类别的分布设计为长尾分布，以模拟真实世界中某些问题(如“噪音大”)常见而另一些问题(如“特定型号的罕见故障代码”)罕见的现象。我们还在数据集中注入了不同比例的噪声，如拼写错误、网络俚语和不相关的文本，以测试模型在噪声和错误输入下的鲁棒性。

我们定义了三个主要的、相互关联的评估指标来全面衡量系统性能：

1. 分类准确率(%)：我们将模型预测结果与一个包含 10 万条评论的、经过专家交叉验证的标准化黄金标签测试集进行比对，计算正确分类的评论所占的百分比。这是衡量系统效果的核心指标。

2. 吞吐量(评论/秒)：我们在一个标准化的硬件配置上(单张 NVIDIA A100 80GB GPU)，测量系统每秒能够稳定处理的评论总数。这是衡量系统效率的关键指标。

3. 人工审核率(%)：我们计算被置信度门控或双模型验证模块转交给人工审核队列的评论，占总评论

数的百分比。我们认为这是一个关键的运营成本指标，直接关系到方案的经济可行性。

4.2. 基于语境感知的合成数据生成协议

为了解决细粒度分类中长尾样本稀缺的问题，并模拟真实社交媒体环境中高噪声的输入特征，我们构建了一个包含 150 万样本的合成数据集。不同于传统的基于规则的同义词替换，我们采用了一种“教师–学生”(Teacher-Student)生成范式。

4.3. 基准系统与实现细节

1. 基准系统：我们明确定义了用于比较的基准，即客户现有的系统。该系统使用一个通用的、参数量为 140 亿的模型，未经过任何领域微调或推理加速优化，并在与我们系统完全相同的标准化硬件上运行。

2. DCV 系统：我们使用了两个经过微调的 30 亿参数模型。为置信度门控机制设置了 64 个 token 的滑动窗口大小，并通过在验证集上进行网格搜索，校准出一个能够在吞吐量和准确率之间取得最佳平衡的停止阈值 s，其值为 0.75。双模型验证的重试次数上限设为 3 次。

5. 性能分析与结果

5.1. 综合性能对比评估

表 1 展示了使用真实的客户数据的基准系统与 DCV 框架在各项关键指标上的表现，表明后者在所有维度上都显著优于前者。

Table 1. Performance comparison between the baseline system and the DCV framework
表 1. 基准系统与 DCV 框架的性能对比

指标	基准系统(14 B 通用模型)	DCV 框架(2x3 B 微调模型)	提升幅度
分类准确率(%)	28.0%	95.2%	+67.2 百分点
吞吐量(评论/秒)	4.6	60.2	+1208%
日处理能力(万条评论)	约 40	约 520	+1200%
人工审核率(%)	100% (因结果不可用)	4.5%	N/A

从这张表格中得出结论，我们的 DCV 框架在所有维度上都显著优于基准系统，实现了压倒性的性能提升。我们将准确率从不可用的 28%提升到了企业级的 95.2%，这意味着分析结果从不可信变为高度可信。同时，将吞吐量提升了超过 12 倍，使得日处理能力从远低于需求的 40 万条跃升至超过 500 万条，不仅完全满足了当前的业务需求，还为未来的数据增长预留了充足的处理能力。同时，我们将需要人工干预的比例从实际上(由于低准确率)的 100%降低到了仅 4.5%的可管理水平内，极大地节约了运营成本。

5.2. 消融研究：解析各组件的贡献

为了证明我们框架中每个组件的独立价值和协同效应，我们进行了一项严谨的消融研究。我们认为，这项研究能够增加我们工作的学术严谨性，并为我们相对复杂的多组件设计提供合理的解释。

Table 2. Ablation study of the DCV framework components
表 2. DCV 框架组件的消融研究

系统配置	准确率 (%)	平均生成长度 (Tokens)	吞吐量 (评论/秒)	人工审核率(%)
1. 基准系统(14 B 模型)	28.0%	128.4	4.6	100%
2. 单个微调模型(3 B) [无门控]	92.5%	85.2	48.5	7.5%
3. 单个微调模型(3 B) + Token 级早停	85.6%	12.5	65.8	NA (机器强行输出结果)
4. DCV 完整框架(早停 + 级联验证)	95.2%	24.6	60.2	4.5%

通过分析表 2 的数据，我们可以清晰地量化每个组件的具体贡献。

首先，对比配置 1 与配置 2 可以看出，仅通过将模型从通用的 14 B 参数量缩减并微调为专业的 3 B 参数量，吞吐量即从 4.6 条/秒大幅提升至 48.5 条/秒。这验证了在受限环境下，使用小参数量专用模型是提升效率的基础。

在配置 3 中，我们引入了 Token 级置信度门控。实验结果显示，这一机制将吞吐量进一步提升至 65.8 条/秒(相比配置 2 提升约 35%)。通过分析表 2 的‘平均生成长度’数据，我们揭示了双模型比单模型更快的根本原因。在配置 2 (单模型无门控)中，模型在处理非分布内(OOD)或模糊样本时，倾向于产生‘幻觉循环’，生成大量重复或无意义的文本，导致平均生成长度高达 85.2 tokens，严重拖累了推理速度。

相比之下，配置 4 (DCV 完整框架)虽然引入了模型 B 的额外开销，但置信度门控机制在模型 A 刚开始陷入困惑(平均仅生成 12~15 tokens)时便立即通过‘熔断’终止计算，并转交给模型 B。由于模型 B 仅需处理约 15%的疑难样本，从全局来看，系统处理每个样本的平均 Token 生成量由 85.2 大幅下降至 24.6。因此，尽管加载了双模型，整体吞吐量(60.2 条/秒)反而比单模型全量跑完(48.5 条/秒)提升了 24%。这证明了‘快速失败(Fail-fast)’策略在长尾分类任务中的计算经济性。”更重要的是，双模型验证成功修正了单模型中的“盲目自信”错误，将准确率推向了 95.2%的新高。

值得注意的是，配置 4 的最终人工审核率为 4.5%。这一比例是由系统自动筛选出的“高分歧”或“双重低置信度”样本，代表了数据集中最难处理的长尾部分。与配置 3 被迫强制输出错误结果不同，配置 4 选择将这些难例交由人工处理，从而在保障自动处理准确率的同时，构建了可靠的人机协作闭环。

5.3. 置信度阈值 s 的影响分析

接下来将深入探讨停止阈值 s 这一关键超参数所控制的系统性能权衡。我们通过一张图表来可视化地展示 s 值从激进(低 s 值)到保守(高 s 值)变化时，对系统关键性能指标的影响。

我们对该图表(图 3)进行分析后发现，验证阈值 s 的选择直接决定了“快速通道”的通过率，从而控制系统的运行模式。当 s 设定较低(例如 0.60)时，系统策略更为激进：判定条件 $P_A \geq s$ 容易满足，约 95%的样本通过快速通道直接输出，且较少触发重试机制。这带来了极高的吞吐量(78.5 条/秒)，但由于部分边缘样本未经过双模型验证，准确率略有下降(93.1%)。相反，当 s 设定较高(例如 0.90)时，系统策略更为保守：大部分样本因无法达到极高的置信度要求而触发二级模型验证。这使得吞吐量显著降低至 28.4 条/秒(低于单模型基准)，但换取了最高的准确率(96.8%)。实验表明，在 $s = 0.75$ 处，系统达到了最佳平衡点，在保证 95.2%高准确率的同时，维持了 60.2 条/秒的高吞吐量。

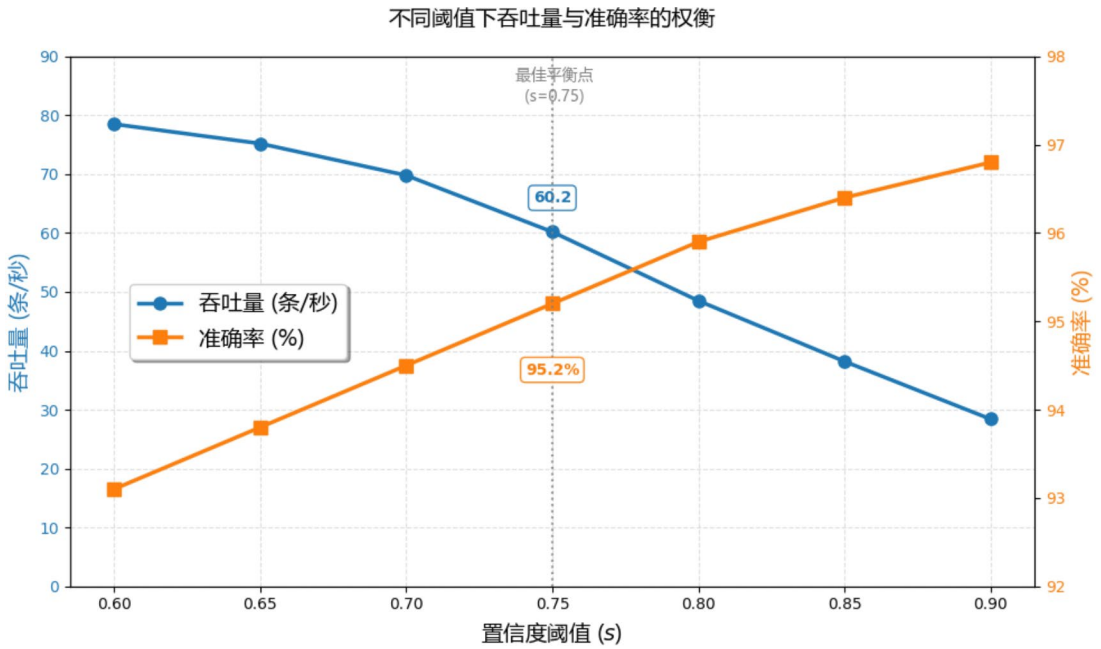

Figure 3. Trade-off impact of the confidence threshold s on system performance
图 3. 置信度阈值 s 对系统性能的权衡影响

6. 局限性与未来工作

6.1. 框架的局限性

尽管我们的 DCV 框架取得了显著的成功，但我们必须坦诚地认识到其存在的若干局限性，这为未来研究提供了启示，并指明了发展方向。

1. 对高质量标注数据的依赖：我们框架的性能基石是经过微调的专业化模型。这意味着系统的初始构建严重依赖于一个大规模、高质量、且由领域专家标注的细粒度数据集。对于拥有 400 个类别的复杂任务，创建和维护这样的数据集本身就是一项成本高昂且耗时的工作[23]。

2. 对新发问题的处理能力：我们的系统被训练用于识别和分类已知的 400 个问题类别。然而，对于全新的、前所未见的产品缺陷或用户抱怨(即“未知–未知”问题)，系统可能会将其错误地归类到最相似的现有类别中，而不是识别为一个新问题。这要求我们在系统外围补充一个异常检测模块[24]。

3. 模型“自信地犯错”的风险：我们提出的置信度机制并非完美无缺。在某些情况下，模型可能会对一个错误的分类结果表现出极高的置信度[25]。尽管我们的双模型验证机制通过要求达成共识，极大地缓解了这一问题，但当两个模型碰巧以同样的方式“自信地犯错”时，这种风险依然存在。

4. 双模型部署的计算开销：虽然我们的双 3 B 模型方案远比单个 14 B 模型高效，但与仅部署单个 3 B 模型的方案相比，它在推理阶段仍然需要接近双倍的计算资源。这是为了追求极致准确性而付出的必要成本，但在资源极度受限的场景下，这种权衡需要被仔细评估[26]。

5. 超参数的敏感性：系统的最优性能在一定程度上依赖于对关键超参数(如置信度阈值 s、重试次数上限)的精确校准。如果未来的数据分布发生显著漂移(例如，由于新产品线的引入或社交媒体平台语言风格的变迁)，这些超参数可能需要重新进行调整[27]。

6.2. 未来工作

基于对上述局限性的认识，我们规划了以下几个有前景的未来研究方向：

1. 动态阈值与自适应调整：我们建议探索动态阈值调整技术。未来的系统可以根据实时的系统负载、数据流入的统计特性，甚至从人工审核反馈中学习，来自动调整置信度阈值 s，从而实现一个能够自适应优化的智能系统。

2. 闭环主动学习与自动化再训练：我们建议将当前隐式的主动学习循环正式化、自动化。我们可以建立一个自动化的再训练流水线，定期将经过人工审核确认的困难样本和分歧样本送入训练集，对模型进行增量微调。我们相信，这将创建一个持续自我改进的系统，使其能够与时俱进，不断提升性能。

3. 集成异常检测机制：为了解决对新发问题的处理能力不足的问题，我们计划在 DCV 框架中集成一个无监督的异常检测模块。该模块可以分析那些被双模型判定为高度分歧或所有类别置信度均很低的样本，将它们识别为潜在的新问题类别，并主动推送给产品和质量控制团队进行分析。

7. 结论

7.1. 贡献总结

我们重申了在处理海量、细粒度的社交媒体用户反馈时，企业普遍面临的效率低下和准确性不足的严峻挑战。我们提出的 DeepConf-Verify(DCV)框架为这一问题提供了一个鲁棒的、可投入生产的、端到端的解决方案。我们强调了我们的三项主要贡献：

(1) 我们验证了使用经过领域微调的小型模型，是解决专业化分类任务的正确且高效的基础路径；

(2) 我们原创性地提出并实现了一种名为“动态置信度门控”的机制，通过监控模型内部的词元级置信度轨迹来实现提前终止，将处理速度提升了超过 10 倍；

(3) 我们设计并实现了一套双模型验证系统，通过强制模型达成共识，将分类准确率提升至超过 95%的企业级黄金标准。

我们相信，本系统为构建下一代企业级智能文本分析系统提供了一份兼具理论深度和实践价值的蓝图。

7.2. 更广泛的影响与未来工作

我们最后将讨论我们工作的更广泛影响。我们认为，这种将专业化模型、自适应推理和验证层相结合的混合方法，为未来需要兼顾效率、准确性和可靠性的企业级 AI 系统提供了一份蓝图。对于未来的工作，我们提出了两个方向。首先，我们建议探索动态阈值调整技术，即根据系统负载或数据分布的变化，实时调整置信度阈值 s。其次，我们建议将隐式的主动学习循环正式化，建立一个自动化的再训练流水线，利用经过人工审核的数据来持续地、迭代地改进模型性能。

参考文献

[1] Madhan, S., Monish Raju, T. and S, V. (2025) The Future of Social Media in Marketing with Reference to Electronic Goods. *ASET Journal of Management Science*, **4**, 408-418. https://doi.org/10.47059/ajms/v4i2/40

[2] Pravina, S. and Muthulakshmi, K. (2025) A Study on Customer Attitudes and Purchase Intentions Toward White Goods Through Social Media Marketing. *International Journal of Management*, **16**, 48-59. https://doi.org/10.34218/ijm_16_03_004

[3] Sonawane, A. and Shinde, S. (2025) Sentiment Analysis for Social Media: Using Natural Language Processing to Understand Public Opinion. *International Journal of Scientific Research in Science*, *Engineering and Technology*, **12**, 110-113.

[4] Zavala, A. and Ramirez-Marquez, J.E. (2019) Visual Analytics for Identifying Product Disruptions and Effects via Social Media. *International Journal of Production Economics*, **208**, 544-559. https://doi.org/10.1016/j.ijpe.2018.12.020

[5] Chakraborty, K., Bhattacharyya, S. and Bag, R. (2020) A Survey of Sentiment Analysis from Social Media Data. *IEEE Transactions on Computational Social Systems*, **7**, 450-464. https://doi.org/10.1109/tcss.2019.2956957

[6] Yan, X., Yang, X., Jin, N., Chen, Y. and Li, J. (2025) A General AI Agent Framework for Smart Buildings Based on Large Language Models and React Strategy. *Smart Construction*, **2**, Article 4. https://doi.org/10.55092/sc20250004

[7] Ren, L., Wang, H., Dong, J., Jia, Z., Li, S., Wang, Y., *et al.* (2025) Industrial Foundation Model. *IEEE Transactions on Cybernetics*, **55**, 2286-2301. https://doi.org/10.1109/tcyb.2025.3527632

[8] Gu, A., Gulcehre, C., Paine, T., Hoffman, M. and Pascanu, R. (2020) Improving the Gating Mechanism of Recurrent Neural Networks. *Proceedings of the* 37*th International Conference on Machine Learning*, 13-18 July 2020, 3800-3809.

[9] Zhang, J., Jin, X., Sun, J., Wang, J. and Li, K. (2019) Dual Model Learning Combined with Multiple Feature Selection for Accurate Visual Tracking. *IEEE Access*, **7**, 43956-43969. https://doi.org/10.1109/access.2019.2908668

[10] Said, A.J. and Ismail, A.M. (2025) Trends in Natural Language Processing for Text Classification: A Comprehensive Survey. *International Journal of Science and Research Archive*.

[11] Bansod, D.A. (2025) Enhanced Deep Learning Approaches for Text Classification: A Comprehensive Review. *International Journal for Research in Applied Science and Engineering Technology*, **13**, 2067-2071. https://doi.org/10.22214/ijraset.2025.66731

[12] Ilhan, F., Chow, K., Hu, S., Huang, T., Tekin, S., Wei, W., *et al.* (2024) Adaptive Deep Neural Network Inference Optimization with Eenet. 2024 *IEEE*/*CVF Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 3-8 January 2024, 1362-1371. https://doi.org/10.1109/wacv57701.2024.00140

[13] Scardapane, S., Comminiello, D., Scarpiniti, M., Baccarelli, E. and Uncini, A. (2020) Differentiable Branching in Deep Networks for Fast Inference. *ICASSP* 2020—2020 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 4167-4171. https://doi.org/10.1109/icassp40776.2020.9054209

[14] Zhang, J., Tan, M., Dai, P. and Zhu, W. (2023) LECO: Improving Early Exiting via Learned Exits and Comparison-Based Exiting Mechanism. *Proceedings of the* 61*st Annual Meeting of the Association for Computational Linguistics* (*Volume* 4: *Student Research Workshop*), Toronto, 10-12 July 2023, 298-309. https://doi.org/10.18653/v1/2023.acl-srw.43

[15] Nam, G., Yoon, J., Lee, Y. and Lee, J.Y. (2021) Diversity Matters When Learning from Ensembles. *Advances in Neural Information Processing Systems*, **34**, 35687-35698.

[16] Jung, Y. (2017) Multiple Predicting *k*-Fold Cross-Validation for Model Selection. *Journal of Nonparametric Statistics*, **30**, 197-215. https://doi.org/10.1080/10485252.2017.1404598

[17] Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y.Q., Cui, H., Zhang, X., *et al.* (2023) Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey. arXiv: 2305.18703.

[18] Kang, W. (2024) QOS-Aware Inference Acceleration Using Adaptive Depth Neural Networks. *IEEE Access*, **12**, 49329-49340. https://doi.org/10.1109/access.2024.3384233

[19] Wang, M., Kim, J. and Yan, Y. (2025) Syntactic-aware Text Classification Method Embedding the Weight Vectors of Feature Words. *IEEE Access*, **13**, 37572-37590. https://doi.org/10.1109/access.2025.3545877

[20] Zhuang, Z., Liu, M., Cutkosky, A. and Orabona, F. (2022) Understanding AdamW through Proximal Methods and Scale-Freeness. arXiv: 2202.00089.

[21] Chen, Z., Li, Y., Bengio, S. and Si, S. (2019) You Look Twice: GaterNet for Dynamic Filter Selection in CNNs. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 9164-9172. https://doi.org/10.1109/cvpr.2019.00939

[22] Desreumaux, L. (2019) An Empirical Study of Active Learning Strategies for Supervised Classification. Centrale-Supélec. https://www.semanticscholar.org/paper/7c0ebd3116b8d7bab080351c33c8bc7a3154e01a

[23] Scotta, S. and Messina, A. (2025) Experimenting Task-Specific LLMs. https://www.semanticscholar.org/paper/053529283167c72e61dbc257ee541f9fef27beed

[24] Betta, G., Capriglione, D. and Corvino, M. (2014) A Proposal for the Management of the Measurement Uncertainty in Classification and Recognition Problems. *IEEE Transactions on Instrumentation and Measurement*, **63**, 2056-2064.

[25] Ma, S., Wang, X., Lei, Y., Shi, C., Yin, M. and Ma, X. (2024) "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Honolulu, 11-16 May 2024, 1-20. https://doi.org/10.1145/3613904.3642671

[26] Shi, J., Wang, Z., Zhou, J., Liu, C., Sun, P.Z.H., Zhao, E., *et al.* (2025) MentalQLM: A Lightweight Large Language Model for Mental Healthcare Based on Instruction Tuning and Dual LoRA Modules. *IEEE Journal of Biomedical and Health Informatics*. https://doi.org/10.1109/jbhi.2025.3594133

[27] Taylor, R., Ojha, V. and Martino, I. (2021) Sensitivity Analysis for Deep Learning: Ranking Hyper-Parameter Influence. *IEEE Access*, **9**, 171457-171465.