Computer Science and Application 计算机科学与应用, 2025, 15(8), 11-20
Published Online August 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.158193

Hans汉斯

相近尺寸下生成式模型和判别式模型在文本多分类任务上的性能比较研究

黄　浩，李　崭，李雨航

中国电子科技集团公司第十研究所，四川 成都

收稿日期：2025年6月30日；录用日期：2025年7月28日；发布日期：2025年8月5日

摘　要

针对自然语言处理中应用广泛的文本多分类任务，为了探究其在实际落地过程中最合适的模型选型。本文从预测准确率和输出响应时延两个维度，对尺寸大小相近的生成式模型Qwen3-0.6b和判别式模型Bert-base在文本多分类任务上的性能表现进行了对比研究。作者使用了THUCNews新闻数据集，设置了Bert微调、Qwen3-0.6b零样本提示、Qwen3-0.6b全参微调和Qwen3-0.6bLoRA微调四组对比实验。实验表明，在相同训练样本的条件下，判别式模型在宏观平均F1值和预测速度上都优于最好的生成式模型方案，分别高出了0.8%和356.8%。

关键词

生成式模型，判别式模型，多文本分类，监督微调，宏观平均F1，响应时延

A Comparative Study on the Performance of Generative and Discriminative Models in Text Multi-Classification Tasks under Similar Model Sizes

Hao Huang, Zhan Li, Yuhang Li

The 10th Research Institute of China Electronics Technology Group Corporation, Chengdu Sichuan

Received: Jun. 30th, 2025; accepted: Jul. 28th, 2025; published: Aug. 5th, 2025

Abstract

In the widely applied text multi-classification task of natural language processing, this paper aims

文章引用: 黄浩, 李崭, 李雨航. 相近尺寸下生成式模型和判别式模型在文本多分类任务上的性能比较研究[J]. 计算机科学与应用, 2025, 15(8): 11-20. DOI: 10.12677/csa.2025.158193

to explore the most suitable model selection for practical implementation. From the perspectives of prediction accuracy and response latency, we conducted a comparative study on the performance of the generative model Qwen3-0.6b and the discriminative model Bert-base in text multi-classification tasks. The experiments were based on the THUCNews news dataset, with four comparison groups set up: Bert fine-tuning, Qwen3-0.6b zero-shot prompting, Qwen3-0.6b full-parameter fine-tuning, and Qwen3-0.6b LoRA fine-tuning. The results demonstrated that, under the same training data conditions, the discriminative model outperformed the best-performing generative model approach in both macro-average F1 score and prediction speed, achieving improvements of 0.8% and 356.8%, respectively.

Keywords

Generative Model, Discriminative Model, Multi-Text Classification, Supervised Fine-Tuning, Macro Average F1, Response Latency

1. 引言

文本分类任务是传统的自然语言处理任务之一，应用于情感分析、垃圾邮件过滤、新闻分类、病历文本处理等基础功能，并延伸至舆情监控、个性化推荐、法律案件管理等专业化需求。随着生成式大语言模型的问世，因为其强大的自然语言理解和文本生成能力，几乎所有自然语言处理(NLP)的任务都可以被一个模型所解决。并且随着大语言模型的权重参数量向着更小更强的方向发展，更多时敏性高的场景也看到了大模型的身影，比如手机等边端设备上大量植入小尺寸模型。另一方面，大模型虽然基线较高，但是存在资源需求高、预测时延高、输出幻觉等问题。在实际应用中，对准确率和响应时延都有较高的要求多文本分类任务仍然采用基于小参数量的判别式模型 Bert 优化得到。鉴于以上情况，本文对模型尺寸相近的生成式模型和判别式模型在文本多分类任务上的性能表现进行对比研究，以支撑实际应用落地时的模型选择。作者使用了 THUCNews 新闻数据集。实验表明，在相同训练样本的条件下，判别式模型在分类 F1 值和预测速度上都优于生成式模型。

2. 相关工作

文本分类任务是传统的自然语言处理任务之一，应用于情感分析、垃圾邮件过滤、新闻分类、病历文本处理等基础功能，并延伸至舆情监控、个性化推荐、法律案件管理等专业化需求。进入深度学习时代以来，文本分类任务主要的技术解决方案有两大类型，一类是以 Bert [1]判别式模型为基础衍生的改进算法，另一类是以 LlaMA [2]、Qwen [3]生成式大语言模型为基础的解决方式。对于 Bert 类解决方案，大多采用微调方法调整 BERT 模型权重，基于任务特定数据集提升目标任务(如文本多分类)性能。Hey 等人[4]提出需求分类的迁移学习方法“NoRBERT”，使用两个预训练 BERT 模型(BERTbase 和 BERTlarge)作为分类模型。这些模型针对二元和多类分类任务进行微调，使用 PROMISE NFR 数据集及其类别标签训练。Sainani 等人[5]研究从软件工程合同文档数据集中提取和分类需求的不同机器学习模型。数据集包含 5472 个需求，标注为 14 个类别(包括项目交付、法律流程、筛选/入职、供应商企业、HR 客户政策、HR 法律和人员分配)。研究应用了 NB、随机森林(RF)、支持向量机(SVM)、BiLSTM 和 BERT5。结果显示，微调 BERT 模型优于其他四个模型，在 9 个需求类别上 F1 分数超过 80%。Chatterjee 等人[6]报告了涉及

2122 个 NFR 的大规模多类分类任务研究，采用三个基于 BiLSTM 的模型和一个预训练 BERTbase 模型。研究表明，微调 BERT 优于其他所有模型，证实了 BERT 在该任务中的优势。该研究使用的数据集也未公开。

另一方面，近年来以 LLaMA 和 Qwen 为代表的大语言模型(LLMs)展现出卓越的语言理解能力涌现现象，开创了无需微调即可完成分类任务的新范式。这些模型既能进行零样本选择，也可采用少样本提示[7]和思维链(CoT) [8]等技术。[9]则系统地回顾了针对大型语言模型(LLM)的参数高效微调方法。文章将这些方法分为五类：自适应微调、加性微调、选择性微调、重构参数微调和混合微调。每种方法下又细分为多个子类。[10]将文献资源分类视为分类号生成任务，利用图书馆编目数据构造训练集和测试集，基于 ChatGLM 3、Llama2 等大语言模型在训练集上进行模型的高效微调，并在中英文测试集上分析模型的分类效果。

最后，在资源消耗方面，[11]表示生成式模型计算需求高，需要大量训练数据，且易受维度诅咒(Curse of Dimensionality)影响。例如，生成模型需建模联合分布 $p(x|y)$，计算复杂度显著高于判别模型。在可解释性方面，生成式模型通过建模联合分布，可解释特征与标签的因果关系。同时，通过输出后验概率分布，提供分类置信度，避免在低置信度时预测结果。而判别式模型存在黑盒特性，其直接学习输入到输出的映射，缺乏对数据生成机制的描述，可解释性较差。例如，支持向量机(SVM)依赖决策边界，但无法解释特征间依赖关系。

本文分别从两个类别中选取有代表性、权重参数量级相近的模型，面向文本多分类任务，基于全参监督微调、高效监督微调、零样本提示工程等技术，对比研究各种方法在 F1 值和预测速度上的性能表现，支撑实际业务应用中的解决方案选择。

3. 本文方法

3.1. 多分类任务定义

在机器学习中，多分类(Multiclass Classification)任务数学定义为：将样本划分到 K 个互斥类别，每个样本仅属于一个类别。

数学描述：

输出空间 $=\{1,2,\cdots,K\}$。

模型输出概率分布 $P(y=k|x)$，满足公式(1)：

$$\sum_{k=1}^{K} P(y=k|x)=1 \tag{1}$$

预测时取最大概率对应的类别为公式(2)。

$$y=\arg\max\left(P(y=k|x)\right) \tag{2}$$

示例：手写数字识别(类别为 0~9)。

3.2. 基于 Bert 模型微调的文本多分类技术

基于 Bert 模型微调的文本多分类技术原理主要依托 BERT 模型的预训练－微调范式，其核心是通过在预训练模型基础上添加线性分类层，并针对下游任务调整参数。整体模型结构由 Bert 层和线性分类层组成。如图 1 所示，BERT 层作为特征提取器，通过多层 Transformer 编码器捕获文本的上下文语义信息。其双向注意力机制能同时考虑词语的左右语境。

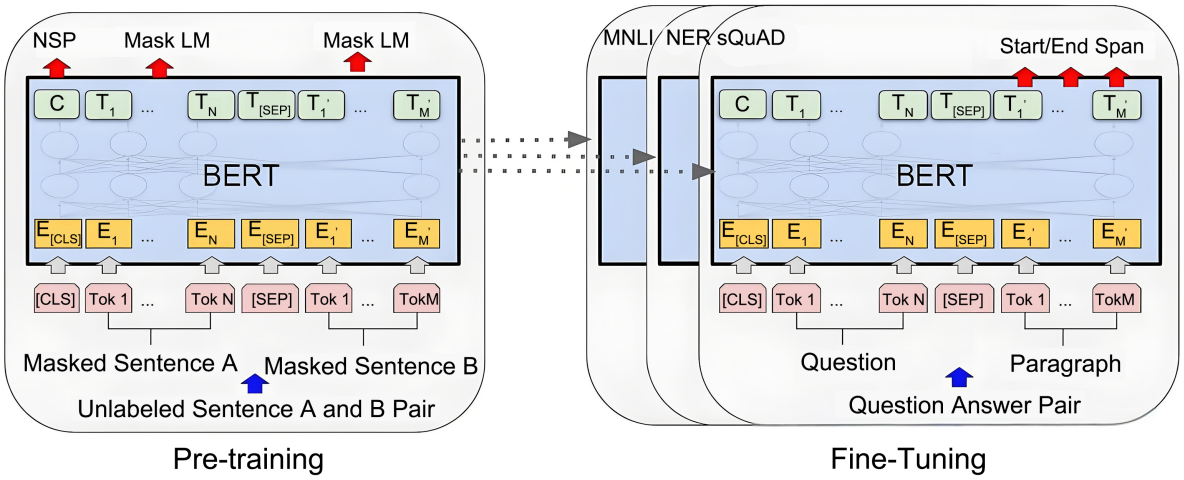

Figure 1. Structure of Bert model
图 1. Bert 层模型结构

如图 2 所示，线性分类层在 BERT 输出的[CLS]标记向量(代表整句语义)后接一个全连接层(Linear Layer)，通过 Softmax 函数输出多分类概率。例如，对于句子“这电影真棒”，BERT 提取全局特征后，线性层将其映射到“positive”类别。

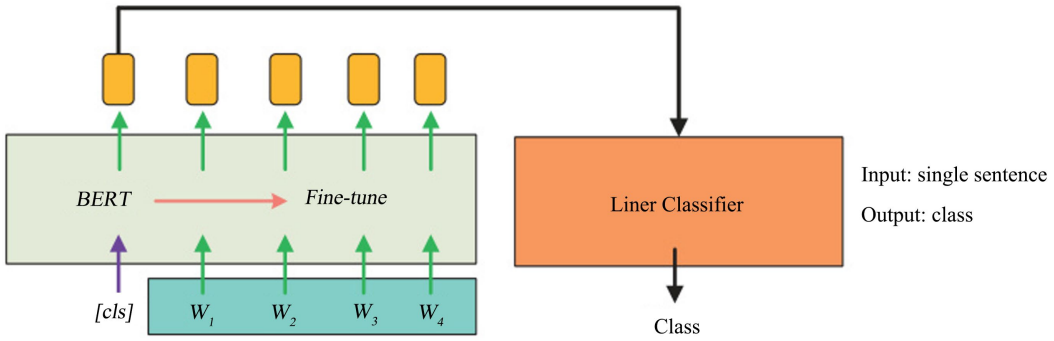

Figure 2. Linear classification layer
图 2. 线性分类层结构

模型微调过程包含参数调整和学习率策略设置。

- **参数调整**：微调时，整个 BERT 模型(包括 Embedding 层、Transformer 层)与分类层一同训练，利用任务标注数据通过反向传播更新参数。这种端到端训练使模型适应特定任务的语言模式。
- **学习率策略**：通常采用分层衰减学习率(Layer-wise Learning Rate Decay)，深层参数使用更小的学习率以缓解灾难性遗忘问题，浅层参数更新幅度较大以捕捉任务特性。

3.3. 基于 Qwen3 和 LoRA 微调的文本多分类技术原理

基于 Qwen3 和 LoRA (Low-Rank Adaptation)微调的文本多分类技术原理，主要结合了预训练语言模型的能力与参数高效微调策略，其核心原理包括 LoRA 原来和多分类适配。

1) LoRA 基本原理

LoRA 通过冻结预训练模型的原始权重，仅对低秩分解矩阵进行微调。具体而言，对于预训练权重矩阵 $W_0 \in R^{d\times k}$，LoRA 引入两个低秩矩阵 $B \in R^{d\times r}$ 和 $A \in R^{r\times k}$ $(r \ll \min(d, k))$，将权重更新量表示为 $\Delta W =$

BA，最终输出为公式(3)：

$$h_{\text{LoRA}} = W_0 x + BAx \tag{3}$$

其中，B 和 A 为可训练参数，初始时 B 为随机高斯分布，A 为零矩阵，确保初始更新量为零。这种设计将可训练参数数量从 $d \times k$ 减少到 $r \times (d + k)$，显著降低内存需求和计算成本。

2) 文本多分类中任务适配

面向多文本分类任务，基于 LoRA 微调的 Qwen3 模型需要做模型适配和超参数的选择。模型适配方面：1) 进行注意力层调整。LoRA 通常被注入到 Transformer 模型的注意力机制中(如 Q、K、V 矩阵)，因为这些层对任务适应敏感。2) 分类头扩展。在预训练模型的顶部添加多分类层(如全连接层)，结合 LoRA 调整后的特征表示，实现类别区分。在超参数配置方面，需要保证效率与性能的平衡。1) 低秩维度选择：通过调整秩 r (如实验常用的 $r = 64$ 或 $r = 8$)，平衡模型容量与过拟合风险。较大的 r 增强表达能力，但可能增加计算量；较小的 r 则更高效。2) 缩放因子 α：引入超参数 α (如 $\alpha = 16$)对 BA 的输出进行缩放，以控制低秩更新的强度，优化训练稳定性。

4. 实验

4.1. 实验方法

为了更加充分对比生成式模型和判别式模型在多分类任务上性能表现的差异，本文设置了四组实验，分别为基于 Bert 模型微调的文本多分类(以下简称 Bert 微调)、基于 Qwen3-0.6b 的零样本文本多分类(以下简称 Qwen3-0.6b 零样本)、基于 Qwen3-0.6b 全参数微调多分类(以下简称 Qwen3-0.6b 全参微调)、基于 Qwen3-0.6b 的 LoRA 微调多分类(以下简称 Qwen3-0.6bLoRA 微调).为了进一步研究解决方案在不同数量类别下的泛化性，每组实验都分别在 4 个类别、9 个类别和 14 个类别上进行性能测算。

4.2. 实验数据

实验采用 THUCNews 数据集[11]，它是根据新浪新闻 RSS 订阅频道 2005~2011 年间的历史数据筛选过滤生成，包含 74 万篇新闻文档(2.19 GB)，均为 UTF-8 纯文本格式。数据划分为 14 个候选分类类别：财经、彩票、房产、股票、家居、教育、科技、社会、时尚、时政、体育、星座、游戏、娱乐。

4.3. 实验参数设置

4.3.1. Bert 微调方法参数设置

Bert 微调方法参数设置如表 1 所示：

Table 1. Bert finetune parameter setup
表 1. Bert 微调方法参数设置

参数名称	设置值
lr_scheduler_type	cosine
learning_rate	1.0e−5
train_batch_size	64
eval_batch_size	256
num_train_epochs	3
weight_decay	1e−6
eval_steps	0.05

4.3.2. Qwen3-0.6b 零样本提示词设置

提示词为：你是一个分类器，你的任务是判断下面文本所属的类别(候选类别如下)，只给出最合适的类别，以 json 格式返回，键为类别。

**返回值示例

{“类别”：“xx”}

候选类别

财经、彩票、房产、股票、家居、教育、科技、社会、时尚、时政、体育、星座、游戏、娱乐

文本

Xxx

4.3.3. Qwen3-0.6b 全参微调方法参数设置

Qwen3-0.6b 全参微调方法参数设置如表 2 所示：

Table 2. Qwen3-0.6b full parameter SFT setup
表 2. Qwen3-0.6b 全参微调方法参数设置

参数名称	设置值
per_device_train_batch_size	12
gradient_accumulation_steps	8
learning_rate	1.2e−5
warmup_ratio	0.01
num_train_epochs	3
lr_scheduler_type	cosine
bf16	true

4.3.4. Qwen3-0.6b Lora 微调方法参数设置

Qwen3-0.6b Lora 微调方法参数设置如表 3 所示：

Table 3. Qwen3-0.6b Lora finetune parameter setup
表 3. Qwen3-0.6b Lora 微调方法参数设置

参数名称	设置值
rank	16
lora_alpha	32
lora_dropout	0.05
bias	none
per_device_train_batch_size	8
gradient_accumulation_steps	4
learning_rate	4
num_train_epochs	3

4.4. 评价指标

本文选择宏观平均精确率 P_{macro}、宏观平均召回率 R_{macro} 和宏观平均 $F1_{\text{macro}}$ 值作为多分类任务的评价指标，其计算公式如公式(4)~(6)：

$$P_{\text{macro}} = \frac{1}{C}\sum_{C=1}^{C}\frac{TP_i}{TP_i + FP_i} \tag{4}$$

$$R_{\text{macro}} = \frac{1}{C}\sum_{C=1}^{C}\frac{TP_i}{TP_i + FP_i} \tag{5}$$

$$F1_{\text{macro}} = \frac{2 * P_{\text{macro}} * R_{\text{macro}}}{P_{\text{macro}} + R_{\text{macro}}} \tag{6}$$

其中，TP_i 为真正例(True Positive)，指的是模型预测为第 i 类且实际为第 i 类的样本数量。FP_i 为假正例(False Positive)，指的是模型预测为第 i 类但实际为其它类的样本数量。FN 为假负例(False Negative)，指的是模型预测为其它类但实际为第 i 类的样本数量。

4.5. 实验结果与分析

4.5.1. 总体实验结果分析

从表 4 和图 3 可以看出，尽管 Bert 模型参数量最小，但是经过监督微调后，其宏观平均 F1 值最高，分别高出后续方法 0.8%、3%、18.9%，后续依次是 Qwen3-0.6b 全参微调、Qwen3-0.6bLoRA 微调和 Qwen3-0.6b 零样本方案。Qwen3-0.6b 零样本方式仅利用了预训练模型的通用能力，没有经过训练优化，准确率低于 0.8。Qwen3-0.6bLoRA 微调方式针对新闻分类场景做了针对性优化，采用高效微调方式，面向训练样本数据集特征，更新了部分参数，使得准确率提升至 0.911。Qwen3-0.6b 全参微调方式在相同的训练样本条件下，通过后训练，更新了整个模型权重，使之在相同类型的文本分类场景中表现更优。Bert 微调方式基于参数量最小、基线能力最差的基座模型，在相同训练条件下取得最佳结果(排除测试误差等原因，基本与 Qwen3-0.6b 全参微调表现一致)，可能的原因有二。第一，判别式模型采用编码器结构(Encoder-only)，其任务的输出空间是预先定义好且有限的(一组标签)或者严格限定在输入文本范围内(如抽取式答案)，不存在生成式模型的“幻觉问题”，输出的标签不会产生“漂移”。第二，Bert 权重大小是 0.1 B，仅为 Qwen3-0.6b 权重的 1/6，同等规模的训练样本使模型优化的更充分，更符合训练样本的特征分布。

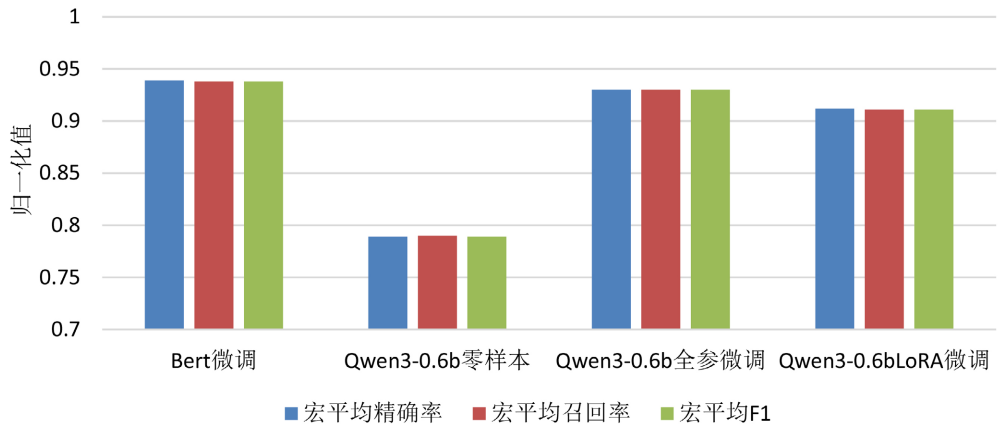

Figure 3. Overall experimental results bar chart
图 3. 总体实验结果柱状图

Table 4. Overall experimental results
表 4. 总体实验结果

实验方法	宏平均精确率	宏平均召回率	宏平均 F1
Bert 微调	0.939	0.938	0.938
Qwen3-0.6b 零样本	0.789	0.790	0.789
Qwen3-0.6b 全参微调	0.930	0.930	0.930
Qwen3-0.6bLoRA 微调	0.912	0.911	0.911

4.5.2. 不同方案推理速度分析

为测试四组方案的预测时响应速度，基于英伟达 RTX 3090 (24 G)显卡进行，计算每秒钟模型输出的字符数量，结果如表 5 所示：

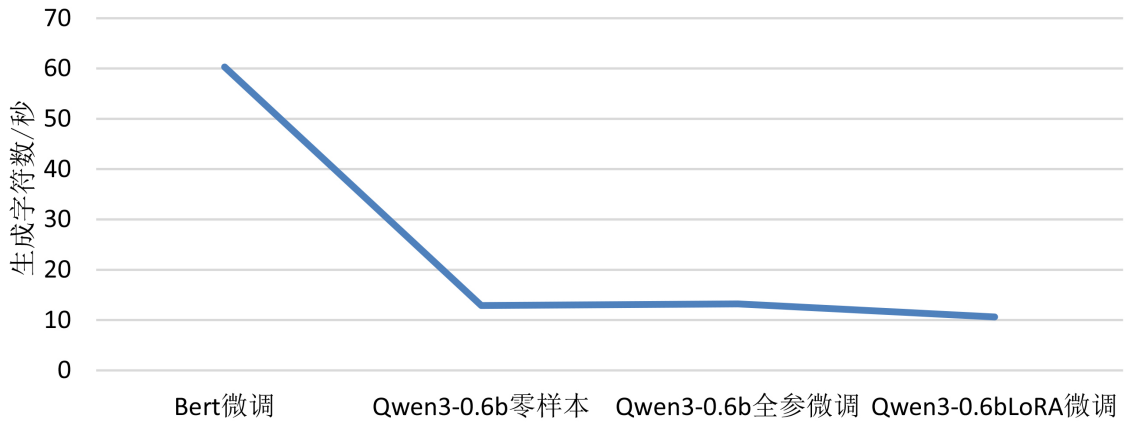

Figure 4. Predict speed results bar chart
图 4. 推理速度测试结果柱状图

Table 5. Predict speed results
表 5. 推理速度测试结果

实验方法	生成字符数/秒	参数量
Bert 微调	60.3	0.1 B
Qwen3-0.6b 零样本	12.9	0.6 B
Qwen3-0.6b 全参微调	13.2	0.6 B
Qwen3-0.6bLoRA 微调	10.65	0.63 B

从表 5 和图 4 可以看出，因 Bert 模型参数量较小，仅 0.1 B，其推理性能最优，且不受输出长度的影响。而基于 Qwen3-0.6b 的方案，全参微调推理性能优于零样本，优于 LoRA 微调方式。具体地，Qwen3-0.6b 全参微调的模型，经过微调后，在推理时无需加载上下文提示词，速度最快。其次，Qwen3-0.6b 零样本方案因需要加入角色扮演、任务信息、分类类别、输出示例等信息，故生成字符数略低于全参微调方案。最后，Qwen3-0.6bLoRA 微调方式将生成一个额外的权重参数，约 0.03 B，模型推理时，需要额外加载计算，故输出性能最低。

4.5.3. 不同分类类别数对实验结果的影响

从表 6 和图 5 可以看出，在 4 类别分类、9 类别分类、14 类别分类任务中，四种方法的整体表现基本一致。随着类别数量的减少，文本分类任务的难度下降，所以四种方法的分类 F1 值均逐步提高，但四种方法的相对排名没有变化。值得一提的是，Qwen3-0.6b 零样本方案在 4 个候选类别的分类任务时，准确率有所提升，分析原因为，低难度任务测试有利于基于模型原始能力的提示词方案。

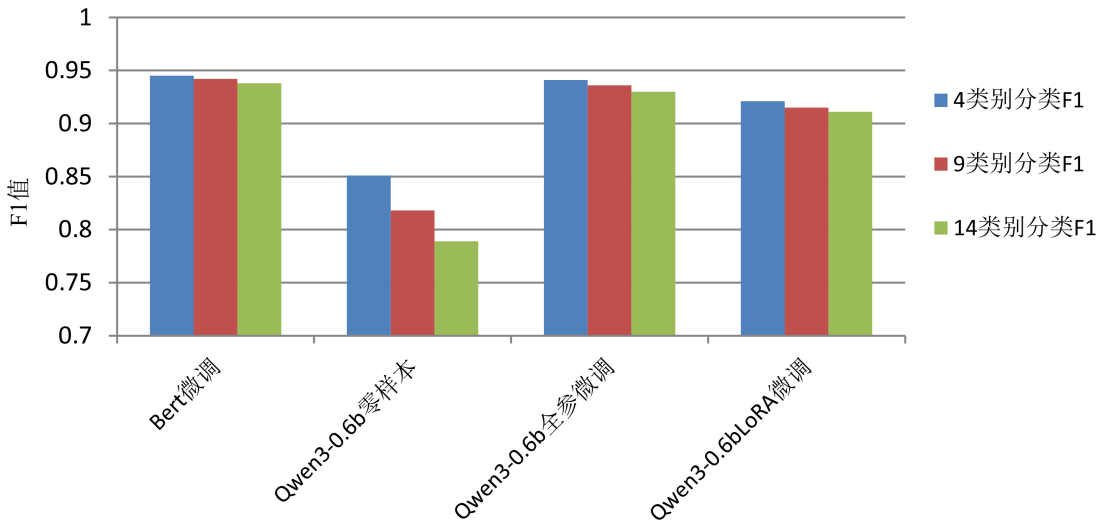

Figure 5. Bar chart of F1 scores for different numbers of classes
图 5. 不同类别数 F1 值测试结果柱状图

Table 6. Results of F1 scores for different numbers of classes
表 6. 不同类别数 F1 值测试结果

实验方法	4 类别分类 F1	9 类别分类 F1	14 类别分类 F1
Bert 微调	0.945	0.942	0.938
Qwen3-0.6b 零样本	0.851	0.818	0.789
Qwen3-0.6b 全参微调	0.941	0.936	0.930
Qwen3-0.6bLoRA 微调	0.921	0.915	0.911

5. 结论与展望

本研究通过对生成式模型 Qwen3-0.6b 和判别式模型 Bert-base 在文本多分类任务上的性能对比，得出以下结论：1) 在预测准确率方面，判别式模型 Bert-base 通过监督微调实现了最佳的宏观平均 F1 值，显著优于生成式模型的所有微调方案(包括零样本提示、全参微调和 LoRA 微调)；2) 在预测速度方面，判别式模型同样表现优异，推理速度远高于生成式模型；3) 生成式模型在不同分类类别数下的性能表现较为一致，但在小规模分类任务中(如 4 类别分类)零样本提示方案的性能有所提升，显示出其在简单任务中的潜在优势。综合来看，尽管生成式模型在某些场景下具有灵活性和适应性，但在文本多分类任务中，判别式模型在准确率和效率上更具优势，尤其适合资源受限的实际应用场景。

未来的研究可以从以下几个方面展开：1) 进行更多自然语言任务的对比分析；2) 研究生成式模型思考模式下的性能表现；3) 探索判别式模型和生成式模型压缩加速后的响应速率。

参考文献

[1] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

[2] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., *et al*. (2023) Llama: Open and Efffcient Foundation Language Models. arXiv:2302.13971.

[3] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K. and Deng, X. (2023) Qwen Technical Report. arXiv:2309.16609.

[4] Hey, T., Keim, J., Koziolek, A. and Tichy, W.F. (2020) NoRBERT: Transfer Learning for Requirements Classification. 2020 *IEEE* 28*th International Requirements Engineering Conference (RE*), Zurich, 31 August 2020-4 September 2020, 169-179. https://doi.org/10.1109/re48521.2020.00028

[5] Sainani, A., Anish, P.R., Joshi, V. and Ghaisas, S. (2020) Extracting and Classifying Requirements from Software Engineering Contracts. 2020 *IEEE* 28*th International Requirements Engineering Conference (RE*), Zurich, 31 August 2020-4 September 2020, 147-157. https://doi.org/10.1109/re48521.2020.00026

[6] Chatterjee, R., Ahmed, A., Rose Anish, P., Suman, B., Lawhatre, P. and Ghaisas, S. (2021) A Pipeline for Automating Labeling to Prediction in Classification of NFRs. 2021 *IEEE* 29*th International Requirements Engineering Conference (RE*), Notre Dame, 20-24 September 2021, 323. https://doi.org/10.1109/re51729.2021.00036

[7] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., *et al*. (2020) Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33: *Annual Conference on Neural Information Processing Systems* 2020, *NeurIPS* 2020, Virtual, 6-12 December 2020, 877-901.

[8] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y.-T. and Iwasawa, Y. (2022) Large Language Models Are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, **35**, 22199-22213.

[9] Han, Z., Gao, C., Liu, J., Zhang, J. and Zhang, S.Q. (2024) Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. arXiv:2403.14608.

[10] 罗鹏程, 王继民, 聂磊. 基于生成式大语言模型的文献资源自动分类研究[J]. 情报理论与实践, 2024, 47(12): 174-182.

[11] Awad, M. and Khanna, R. (2015) Support Vector Regression. In: Awad, M. and Khanna, R., Eds., *Efficient Learning Machines*, Apress, 67-80. https://doi.org/10.1007/978-1-4302-5990-9_4