Qwen3 Tech Report解读

2025年5月13日　3 分钟阅读

#AI #Qwen3 #大模型 #Qwen #Paper

> *全方位解读Qwen3的论文技术报告*

模型介绍

Qwen3，这是Qwen系列大型语言模型的最新版本，该系列模型涵盖从0.6B到235B的多种参数规模，包含稠密模型和混合专家模型（MoE）。 Qwen3的关键创新在于集成了思考模式和非思考模式，允许模型根据需要动态切换推理深度，并引入了思考预算机制以优化计算资源。文档详细阐述了模型的架构设计、三阶段预训练过程（通用、推理、长文本），以及四阶段后训练方法（包括长CoT冷启动、推理RL、思维模式融合和通用RL）。此外，还介绍了强大的弱模型蒸馏技术，用于优化轻量级模型的性能。全面的评估结果显示，无论在预训练还是后训练阶段，Qwen3在多项基准测试中均表现出色，尤其在编码、数学、推理和Agent任务上具有竞争力，并显著提升了多语言支持能力，涵盖119种语言和方言。

技术指标

根据来源资料，通义千问3 (Qwen3) 相较于前代模型在几个核心技术方面取得了显著进展，从而提升了其性能：

> **思维模式与非思维模式的整合**：通义千问3将两种不同的操作模式整合到单个模型中。这包括用于复杂、多步推理的思维模式和用于快速、上下文驱动响应的非思维模式。这种统一的框架消除了在不同模型之间切换的需要，例如从 Qwen2.5 切换到 QwQ (Qwen Team, 2024) 或其他专门的推理模型。模型可以根据用户查询或聊天模板动态切换模式。
>
> **思维预算机制**：通义千问3引入了思维预算机制，允许用户在推理期间自适应地分配计算资源。这有助于根据任务复杂性平衡延迟和性能。评估表明，增加思维代币的思维预算可以持续改善模型在各种任务上的表现。

目录

文章信息

字数

阅读时间

发布时间

更新时间

标签

#AI #Qwen3 #大模型 # #Paper

多语言能力显著扩展：通义千问3将多语言支持从 Qwen2.5 的 29 种语言和方言扩展到了 119 种。这通过改进的跨语言理解和生成能力增强了全球可访问性。所有 Qwen3 模型都在包含 119 种语言和方言的庞大且多样化的数据集上进行了训练。多语言数据被特别增加，以增强低资源语言任务的表现。在 Belebele 基准测试中，Qwen3 的表现与类似规模的 Gemma 模型相当，同时显著优于 Qwen2.5。

更大规模、更多样化的预训练数据：通义千问3 的预训练过程使用了约 **36 万亿**代币的庞大数据集，旨在确保语言和领域的 다양성。这比 Qwen2.5 的预训练代币量翻了一番，语言数量增加了三倍。数据收集采用了多模态方法，例如使用 Qwen2.5-VL 从 PDF 文档中提取文本，并使用领域专用模型（如 Qwen2.5-Math 和 Qwen2.5-Coder）生成合成数据。此外，开发了一个多语言数据标注系统，对超过 30 万亿代币进行了标注，以支持更有效的数据过滤和组合。与之前在数据源或领域级别优化数据混合的研究不同，Qwen3 的方法在实例级别优化数据混合。

架构改进：通义千问3 密集模型的架构与 Qwen2.5 类似，使用了分组查询注意力 (GQA)、SwiGLU、旋转位置嵌入 (RoPE) 和 RMSNorm。此外，移除了 Qwen2 中使用的 QKV-bias，并引入了 QK-Norm 到注意力机制中，以确保训练的稳定性。

上下文长度增加：在预训练的第三阶段，模型在数千亿代币上进行了训练，将上下文长度从 4,096 增加到 **32,768 代币。**通过使用 ABF (Xiong et al., 2023)、YARN (Peng et al., 2023) 和 Dual Chunk Attention (DCA, An et al., 2024) 等技术，在推理过程中实现了序列长度能力的四倍提升。在非思维模式下，Qwen3 在长上下文处理任务中优于类似大小的 Qwen2.5 模型。

优化的后训练流程：后训练流程旨在实现思维控制和通用强化学习等核心目标。这包括 Long-CoT 冷启动、推理 RL、思维模式融合、通用 RL 和强到弱蒸馏等阶段。强到弱蒸馏使得小型模型能够以显著减少的成本和努力实现强大的推理能力。思维模式融合阶段构建了包含思维和非思维数据的数据集，并设计了聊天模板以动态切换模式。评估结果表明，这个流程增强了模型的各种能力，包括模式切换和事实准确性。

卓越的性能表现：通义千问3 在各种基准测试中取得了最先进的结果，与更大的 MoE 模型和专有模型相比具有竞争力。通义千问3 密集模型在更高参数规模上与 Qwen2.5 密集模型表现相当，特别是在 STEM、编码和推理基准测试中表现甚至超越了 Qwen2.5 模型。值得注意的是，参数量小于 Qwen2.5-72B-Base 一半的 Qwen3-32B-Base 在 15 个评估基准测试中的 10 个上表现优于 Qwen2.5-72B-Base。通义千问3-235B-A22B (非思维模式) 在多数基准测试中超越了其他领先的开源模型，包括 DeepSeek-V3 和 LLaMA-4-Maverick，以及之前的旗舰模型 Qwen2.5-72B-Instruct，甚至在 23 个基准测试中的 18 个上超越了闭源的 GPT-4o-2024-11-20。通义千问3-32B (思维模式) 在 23 个基准测试中的 17 个上优于 QwQ-32B，成为 32B 规模新的

最先进推理模型，并与闭源的 OpenAI-o3-mini (medium) 竞争。通义千问3-32B (非思维模式) 在几乎所有基准测试上表现出优于所有基线的性能。即使是边缘侧模型 (如 8B, 4B, 1.7B, 0.6B) 也表现出色，在某些情况下甚至超越了参数更多的基线模型，包括先前的 Qwen2.5 模型。

训练方法

通义千问3 (Qwen3) 的训练过程主要分为**预训练 (Pre-training)** 和**训练后处理 (Post-training)** 两个主要阶段。

1. 预训练阶段 (Pre-training)

预训练为模型构建了强大的基础能力，包括通用知识、语言理解和生成能力，以及多语言能力。

大规模且多样化的数据: Qwen3 模型在约 **36万亿 (trillion) token** 的数据集上进行了预训练。

数据集规模相比 Qwen2.5 **增加了一倍**.

数据涵盖了高达 **119种语言和方言**，比 Qwen2.5 增加了三倍。这极大地增强了模型的多语言能力。

数据来源广泛，包括编程、STEM (科学、技术、工程和数学)、推理任务、书籍、多语言文本和合成数据等高质量内容。

高效的数据扩展方法: 为了扩大训练数据语料库：

使用 **Qwen2.5-VL** 模型从大量的 PDF 文档中提取文本，然后使用 Qwen2.5 模型进行提炼以提高质量。

使用领域特定模型生成合成数据，例如使用 **Qwen2.5-Math** 生成数学相关内容，使用 **Qwen2.5-Coder** 生成代码相关数据。

三阶段预训练策略: Qwen3 模型通过一个三阶段过程进行预训练：

第一阶段 (通用阶段 - S1): 所有模型使用 4,096 token 的序列长度，在超过 30万亿 token 的数据上进行训练。这一阶段主要建立语言能力和通用世界知识的基础，涵盖 119种语言。

第二阶段 (推理阶段 - S2): 通过增加 STEM、编程、推理和合成数据的比例来优化语料库，进一步提升推理能力。模型使用 4,096 token 的序列长度，在约 5万亿更高质量的 token 上进行额外预训练。此阶段加速了学习率衰减。

第三阶段 (长上下文阶段): 收集高质量的长上下文语料库，将模型的上下文长度从 4,096 token 扩展到 **32,768 token**。模型在数千亿 token 上进行预训练，其中包含 75% 长度在 16,384到 32,768 token 之间的文本和 25% 长度在 4,096到 16,384 token 之间的文本。使用了 ABF 技术将旋转位置嵌入 (RoPE) 的基础频

率从 10,000 提高到 1,000,000。引入了 **YARN** 和 **Dual Chunk Attention (DCA)** 技术，以在推理时实现序列长度容量的四倍提升。

2. 训练后处理阶段 (Post-training)

训练后处理管线经过策略性设计，旨在实现两个核心目标：**思维控制 (Thinking Control)** 和 **通用强化学习 (General RL)**。

思维控制: 这是 Qwen3 的一项关键创新，将"非思维"模式和"思维"模式整合到一个模型中。

通过在用户查询或系统消息中引入 `/think` 和 `/no think` 标记，实现模式的**动态切换**。默认情况下模型运行在思维模式。

引入**思维预算机制**，允许用户在推理过程中**自适应地分配计算资源**，从而平衡延迟和性能。增加思维 token 预算可以持续提升模型在各类任务上的表现。

思维模式融合 (Thinking Mode Fusion)：训练后管线中的一个阶段。在此阶段构建的监督微调 (SFT) 数据集结合了"思维"和"非思维"数据。数据集设计包括过滤掉无需 CoT (Chain-of-Thought) 推理的简单查询和推理过程不足的响应。目标是让模型获得**根据标记正确切换思维模式**的能力。

通用强化学习 (General RL): 训练后管线包括推理 RL (Reasoning RL) 和通用 RL 阶段。

Long-CoT Cold Start: 训练后管线的初始阶段，旨在向模型灌输基础的推理模式，不强调立即的推理表现。数据集经过过滤，只包含需要更深层推理的复杂问题。

推理 RL: 训练后管线的一个阶段，专门通过强化学习提升推理能力。

通用 RL: 训练后管线的最后一个阶段。目标包括**遵循格式 (Format Following)**，例如根据 `/think` 和 `/no think` 标记正确切换模式，并在输出中使用 `<think>` 标记分隔内容。还包括**偏好对齐 (Preference Alignment)**，旨在提升模型在开放性查询上的有用性、吸引力和风格，提供更自然的用户体验。

由强到弱蒸馏 (Strong-to-Weak Distillation): 这是训练后处理方法的一部分，特别针对**轻量级模型**设计（包括 0.6B, 1.7B, 4B, 8B, 14B 密集模型和 30B-A3B MoE 模型）。这种方法被证明能够有效地赋予轻量级模型深刻的推理能力。例如，Qwen3-30B-A3B 在思维模式下，使用更小的模型规模和显著少于十分之一的激活参数量，实现了与 QwQ-32B 相当的性能。这使得构建轻量级模型所需的成本和精力显著降低，同时性能令人印象深刻。

提升复杂任务和多语言场景表现的流程和机制:

复杂任务: 主要通过预训练阶段增加 STEM、编程、推理和合成数据的比例，以及训练后阶段的思维控制机制（思维模式、思维预算） 和推理相关的强化学习 (Reasoning RL) 来提升。思维模式和思维预算允许模型在需要时进行更深入的推理。

多语言场景: 主要通过在预训练阶段使用**大规模且多样化、涵盖119种语言和方言的数据集** 构建基础能力。训练后处理阶段虽然没有专门针对多语言训练的流程描述，但通用强化学习中的格式遵循和偏好对齐等能力会应用于多语言输出。对模型的多语言能力评估是在训练后进行的，结果表明 Qwen3 在多种多语言基准测试中取得了有竞争力的表现。

总之，Qwen3 通过结合大规模多阶段预训练、创新的思维控制训练后机制、通用强化学习以及面向轻量级模型的蒸馏策略，在复杂任务处理和多语言支持方面实现了显著提升。

模型评估

根据来源资料和我们之间的对话历史，通义千问3 (Qwen3) 的评估是一个**全面且多维度**的过程，涵盖了模型的**预训练基础能力**和**训练后处理后的表现**，特别关注其在复杂任务处理、多语言场景以及独特的思维控制机制下的性能。评估旨在展示Qwen3在广泛任务和领域上的**领先性能。**

评估过程主要涵盖以下几个方面：

对预训练基础模型的评估：

评估重点在于基础模型在**通用知识、推理、数学、科学知识、编码和多语言能力**等方面的表现。

使用了15个标准基准数据集进行评估，其中包括：

通用任务：MMLU、MMLU-Pro、MMLU-redux、BBH (Few-shot, CoT)、SuperGPQA (Few-shot, CoT)。

其他类别如数学、科学知识、编码和多语言能力也使用了相应的基准。

评估中将Qwen3系列基础模型与Qwen2.5基础模型以及其他领先的开源基础模型进行了比较，包括DeepSeek-V3 Base、Gemma-3、Llama-3 和 Llama-4 系列基础模型。

评估使用了**相同的评估流程和广泛采用的评估设置**，以确保公平比较。

评估结果显示，Qwen3密集型基础模型在较高参数规模下与Qwen2.5基础模型表现相当，尤其在STEM、编码和推理基准上甚至超越了参数规模更高的Qwen2.5模型。旗舰基础模型Qwen3-235B-A22B-Base在大多数评估基准上取得了最高分。

对训练后处理模型的评估：

训练后处理后的Qwen3模型在**思维模式 (Thinking Mode)** 和**非思维模式 (Non-thinking Mode)** 下都进行了评估。

评估任务分类更加多样化，包括：

通用任务：MMLU-Redux、GPQA-Diamond、C-Eval、LiveBench。

对齐任务：IFEval (严格提示准确率)、Arena-Hard、AlignBench v1.1、Creative Writing V3 和 WritingBench。这些评估了模型遵循指令、与人类偏好对齐的能力。

数学与文本推理：MATH-500、AIME'24 和 AIME'25 (对每个问题采样多次并取平均准确率)、ZebraLogic、AutoLogi。

Agent 与编码：BFCL v3 (评估使用FC格式，并在长上下文设置下评估多轮对话)、LiveCodeBench v5 (针对思维模式调整了提示模板)、Codeforces Ratings (计算Elo等级分)。

多语言任务：Multi-IF、INCLUDE、MMMLU (14种语言)、MT-AIME2024、PolyMath、MLogiQA。特别使用了Belebele基准，评估了模型在**80种受支持语言**上的自然语言理解能力。

评估时使用了特定的采样超参数。对于思维模式，温度设置为0.6，top-p为0.95，top-k为20。非思维模式的设置略有不同。最大输出长度通常设置为32,768 tokens，但在AIME等任务中会延长以提供足够的思维空间。

训练后模型与多种领先的开源和闭源模型进行比较，包括DeepSeek-R1/V3、Grok-3-Beta (Thinking)、Gemini2.5-Pro、OpenAI-o1/o3-mini/o4、QwQ-32B (作为之前的强大推理模型)、各种Qwen2.5-Instruct模型、Llama-4/3.1、Gemma-3-IT模型 和 Phi-4/mini。

结果显示，Qwen3旗舰模型Qwen3-235B-A22B在两种模式下均达到了开源模型的SOTA水平，并与闭源领先模型具有高度竞争力。Qwen3-32B在多种基准上超越了之前的推理模型QwQ-32B，并在非思维模式下表现优异，超越了Qwen2.5-72B-Instruct。

对特定能力的评估：

思维预算的有效性：通过调整思维tokens的预算，评估模型在数学、编码和STEM领域基准上的性能。结果表明，增加思维预算可以持续提升模型性能。

长上下文能力：使用RULER基准评估模型处理长上下文的能力。评估在思维模式和非思维模式下进行。结果显示非思维模式下Qwen3优于同等大小的Qwen2.5模型，而思维模式下的性能略有下降。

模式切换能力：使用ThinkFollow基准评估模型是否能根据用户查询中的`/think`和`/no think`标记正确切换思维模式。

评估轻量级模型的性能：

Qwen3系列中的轻量级模型（如0.6B, 1.7B, 4B, 8B, 14B和30B-A3B）也进行了广泛评估。

评估结果证明了**由强到弱蒸馏 (Strong-to-Weak Distillation)**方法的有效性。例如，Qwen3-30B-A3B在思维模式下使用远少于十分之一的激活参数量就实现了与QwQ-32B相当的推理性能。Qwen3-14B/30B-A3B在非思维模式下也以显著更少的参数超越了Qwen2.5-32B-Instruct。Qwen3-8B/4B/1.7B/0.6B等边缘端模型也在多种基准上表现出色，甚至超越了参数量更大的Qwen2.5模型。

综上所述，Qwen3的评估是一个全面而严谨的过程，涵盖了从基础能力到高级特性的多方面测试，通过与现有领先模型的对比，充分展示了其在性能、效率、复杂任务处理和多语言支持方面的进步。

> **DeepSeek R1 Paper...**
>
> A comprehensive review of t...