Qwen3.8-Flash-Next 技術・アーキテクチャ包括分析報告：次世代Qwen4アーキテクチャの先行実装と超疎MoEの新潮流

Gemini 3.7 Flash

概要と開発の背景

Alibabaの基盤モデル研究開発チームであるQwenは、2026年8月26日、次世代基盤モデルの設計思想を具現化したオープンウェイトモデル「Qwen3.8-Flash-Next」を公開した[1]。本モデルは、将来的なメジャーアップデートとなる「Qwen4」シリーズに採用予定の新規アーキテクチャをコミュニティおよびエンタープライズ向けに先行開示し、実環境での検証を行うための技術プレビューとして位置付けられている[1]。これは、過去にQwen3.5世代の基盤を実証した「Qwen3-Next」と同様の開発戦略を踏襲したものである[5]。

近年の基盤モデル開発においては、単純なパラメータ規模の拡大や全結合層の拡張に伴う計算資源およびメモリ帯域のボトルネックが深刻化している[8]。Qwenチームは本モデルにおいて、「演算の実行領域」「長期知識の保持領域」「長大コンテキストの検索領域」を構造的に分離する設計方針を採用した[9]。テキストのみならず画像や動画の高度な理解をネイティブで備えるマルチモーダル構成を維持しつつ、主骨格の総パラメータ数1,250億（125B）、外部記憶表510億（51B）、マルチトークン予測モジュール40億（4B）からなる大規模構造を構築している[2]。その一方で、1トークンあたりの処理時に稼働するアクティブパラメータ数をわずか60億（6B）に抑え込む超疎（Ultra-Sparse）Mixture-of-Experts（MoE）設計を確立した[2]。

この抜本的な構造刷新により、前世代の「Qwen3.7-Plus」と比較して約9分の1という極めて低い計算コストで事前学習を完了させながら、ソフトウェアエンジニアリングやオフィス自動化エージェントタスクにおいて最高峰のフロンティアモデルを凌駕する実行性能を達成している[3]。

4大コアアーキテクチャの技術的刷新

Qwen3.8-Flash-Nextの技術的進歩は、「アテンション機構」「残差接続」「埋め込み空間」「最適化手法」という4つの主要領域における体系的な再構築によって支えられている[3]。

GDN + QSA ハイブリッド・アテンション

長大コンテキスト処理時の計算量増大とKey-Value（KV）キャッシュによるメモリ逼迫を根本から解決するため、全48層のトランスフォーマーブロックは「Gated DeltaNet（GDN）」と「Qwen Sparse Attention（QSA）」を組み合わせたハイブリッド構成を採用している[3]。モデル全体は4層を1ブロックとする12周期の構造からなり、各ブロックの前半3層（計36層）にGDN、後半1層（計12層）にQSAが配置されている[3]。

GDNは線形再帰型の注意機構であり、入力されるコンテキスト履歴を固定サイズの内部状態へと逐次的に圧縮し続けることで、メモリ消費量をコンテキスト長に依存しない一定値に保つ役割を果たす[3]。GDN層における線形アテンションヘッドはValue（V）が48ヘッド、Query-Key（QK）が16ヘッド、ヘッド次元は128で構成される[14]。一方、4層ごとに挿入されるQSAは、微小ブロック（micro-block）粒度で

動作する高精度なスパース検索機構である[12]。QSA層はQuery（Q）が24ヘッド、Key-Value（KV）が2ヘッド（ヘッド次元256、RoPE次元64）で構成され、検索専用のインデクサーにはMulti-Query Attention（4 Queryヘッド / 1 Shared Keyヘッド、ヘッド次元128）が割り振られている[8]。検索予算は512ブロック（2,048トークン）に制限されている[14]。
この2系統の機構が相互補完することにより、オンライン推論環境（プレフィックスキャッシュ命中率90%想定）の100万トークン入力時において、Qwen3.7-Plus比で約8.6倍のPrefillスループット向上を記録している[3]。

Gated Residual（ゲート付き残差接続）

モデルの深層化に伴う表現力の低下や勾配消失・爆発を防ぐため、従来の単一バイパス経路による残差接続を拡張した「Hyper-Connections」技術が導入された[3]。残差ストリームの経路数を4本に並列化し、各経路の情報の読み出しと書き込みを動的ゲートによって制御する[3]。
具体的には、入力データに依存して要素単位で調整されるリードゲートと、ブランチ単位のスカラーライトゲートが機能することで、初期層で抽出された局所的な文脈情報が深層に至る過程で希釈されるのを防ぐ[3]。これにより、推論時のオーバーヘッドを最小限に抑えつつ、深層にわたる情報伝達効率と学習時の数値的安定性を飛躍的に高めている[3]。

N-gram Embedding（外部条件付き記憶機構）

MoEが担う動的な条件付き計算を補完し、静的な知識パターンをオフロードするための機構として、510億パラメータ（約2,000万エントリ）に及ぶ「N-gram Embedding」テーブルが第2層に統合されている[12]。自然言語における頻出熟語やプログラミング言語における定型構文（構文宣言やライブラリの定型ブロック等）は、直前の数トークン（bigramおよびtrigram）から参照表を直接ルックアップすることでベクトル表現を取得する[15]。
この設計により、トランスフォーマー本体の初期層が表層的な構文再構成の負荷から解放され、より高次な論理推論や文脈追跡にニューラルネットワークの容量を割り当てることが可能となった[17]。さらに、N-gram埋め込みはトークンごとに極めて疎なアクセスしか発生しないため、高価なGPU VRAMではなくホストメインメモリ（DDR RAM）やNVMe SSD上にテーブルを配置できる[3]。Layer 1の演算中にLayer 2で必要となる埋め込みベクトルを非同期にプリフェッチするパイプラインがモデルレベルで実装されており、推論レイテンシを劣化させることなく大幅なメモリ節約を実現している[12]。

Muon オプティマイザと学習プロセスの最適化

事前学習の効率を極限まで高めるため、行列更新において直交化制約を活用する新規オプティマイザ「Muon」と、埋め込み層やバイアス項の更新に適した「AdamW」をパラメータ種別ごとに併用するハイブリッド学習手法が採用された[7]。新アーキテクチャに合わせてスケーリング則を再適合させた結果、従来の事前学習で必須とされていたバッチサイズの段階的ウォームアップ期間を撤廃し、学習開始時点から目標バッチサイズおよび大きな学習率を適用することが可能となった[12]。これにより総最適化ステップ数が大幅に削減され、Qwen3.7-Plusと同等以上の能力を獲得するために要した計算資源は約9分の1に抑制されている[3]。

モデル仕様とハードウェア実行要件

Qwen3.8-Flash-Nextの構造パラメトリクス、コンテキスト仕様、量子化ファイル構成、および動作環境の仕様を以下に示す。

基本構造パラメータ仕様

項目	詳細仕様
総パラメータ数	1,250億（125B Backbone）＋ 510億（51B N-gram表）＋ 40億（4B MTP）[7]
トークンあたりアクティブパラメータ数	60億（6B）[2]
レイヤー数および構造	全48層（12 × [3×GDN-MoE ＋ 1×QSA-MoE]）[14]
隠れ層次元（Hidden Dimension）	2,560[16]
語彙数（Token Vocabulary）	248,320（Padded）[8]
コンテキスト長	ネイティブ：262,144トークン（256K） / YaRN拡張時：最大1,000,000トークン（1M）[3]
対応モダリティ	テキスト、静止画像、動画（ネイティブ・マルチモーダル）[2]
推論・推論制御機構	1層MTP（Multi-Token Prediction）、Thinking / Non-thinking 動的制御[8]

フォーマット構成とメモリ所要量

公式およびオープンソースコミュニティ（Unsloth等）から提供されている主要な重みフォーマットと動作環境要件は以下の通りである[3]。

形式・量子化レベル	ストレージ容量 / メモリ要件	主なターゲット環境
公式 **BF16**	約355～360 GB[3]	データセンター向けクラスタ（NVIDIA GB300 NVL72、8×H100/H200等）[24]
公式 **FP8**	約185 GB[3]	ワークステーション環境（4×RTX PRO 6000 Blackwell、DGX Station）[24]
Unsloth UD-Q4_K_XL	約111.3 GB[3]	128GB統合メモリ搭載PC（Apple Silicon M-Max/Ultra

		等)[17]
Unsloth UD-IQ1_S	約72.5 GB[3]	96GB統合メモリ機、またはホストRAMオフロード構成PC[22]

推論実行フレームワークとシステム統合

本モデルのN-gram非同期プリフェッチおよび超疎MoEルーティングは、主要なサービングフレームワークであるSGLang、vLLM、TokenSpeed、KTransformers等にDay Zeroから統合されている[3]。エンタープライズ向けのスケールアウト検証においては、NVIDIA GB300 NVL72システム上でGPUあたり16,000 tokens/sec以上、ユーザーあたり200 tokens/sec超の高スループット・低遅延サービングが実証されている[24]。また、ローカル推論においてはllama.cppおよびMLXエンジンを通じてCPU-GPU混在環境での効率的な実行が可能となっている[7]。

ベンチマーク評価と実務性能の分析

Qwen3.8-Flash-Nextは、アクティブパラメータが6Bという軽量な計算プロファイルでありながら、実務的な自律エージェント、コーディング、推論、マルチモーダルタスクにおいて最高峰のフロンティアモデルに匹敵あるいは凌駕するスコアを記録している[3]。

包括的性能比較

評価領域 / ベンチマーク指標	**Qwen3.8-Flash-Next (125B-A6B)**	**Claude Opus 4.6 (Max) /** 最上位水準	**Qwen3.8-27B (Dense)**	**Qwen3.7-Plus (397B**相当**)**
ソフトウェア開発 **(SWE-bench Pro)**	**62.5%** [cite: 8, 12, 28]	53.4%[3] (Claude Mythos 5: 80.3%[28])	61.7%[3]	53.5% (Qwen3.6)[30]
多言語ソフトウェア開発 **(SWE-bench ML)**	**81.0%** [cite: 8, 12, 28]	77.5%[12] (Claude Opus 5: 89.5%[28])	75.8%[12]	73.8%[12]
リポジトリ規模コード生成 **(NL2Repo)**	**48.1%** [cite: 8, 12, 28]	47.6%[12] (DeepSeek-V4 Pro: 61.5%[28])	41.1%[12]	42.3%[12]
自律型コーディング **(DeepSWE**	**58.7%** [cite: 8, 28, 31]	-- (GPT-5.6 Sol: 72.7%[28])	42.2%[31]	--

1.1)				
競技プログラミング **(LiveCodeBench v6)**	**91.9%** [cite: 12, 28, 32]	-- (Sakana Fugu-Ultra: 93.2%[28])	91.9%[29]	--
長時間オフィス実務 **(CoWorkBench)**	**73.9%** [cite: 8, 12, 28]	68.2%[3]	65.1%[12]	45.1%[12]
実務職務遂行 **(JobBench)**	**55.7%** [cite: 8, 12, 28]	36.6%[3]	33.4%[3]	41.3%[12]
長時間ツール実行 **(Toolathlon-Verified)**	**73.5%** [cite: 28, 31]	-- (Claude Opus 5: 80.6%[28])	67.1%[31]	--
モバイルUI自律操作 **(AndroidWorld)**	**84.5%** [cite: 12, 28]	-- (Qwen3.8 Max: 85.3%[28])	81.0%[12]	81.9%[12]
デスクトップOS操作 **(OSWorld 2.0)**	**19.4% / 48.0%** [cite: 28, 34]	2.8% / 21.5%[34]	--	--
最難関学術知識 **(HLE w/o tools)**	**35.9%** [cite: 28]	-- (Claude Mythos 5: 59.0%[28])	--	--
視覚数学 **(MathVision w/o / w/ CI)**	**90.6 / 95.7** [cite: 34, 35]	65.5 / --[34]	90.3 / 88.7[34]	90.0 / 94.6[34]
Webアプリ再構築 **(RecreationBe**	**49.9%** [cite: 8, 12]	--	30.2%[12]	47.1%[12]

nch)				
UIフロントエンド生成 (Vision2Web)	**64.0%** [cite: 8]	--	42.1%[8]	62.9%[8]

実務遂行能力と推論制御メカニズム

ベンチマークの分析において特筆すべきは、単発の知識検索よりも、ツール実行や自律的な検証ループを伴う長大タスク（SWE-bench Proの62.5%やCoWorkBenchの73.9%）での際立った強さである[3]。本モデルは単体での一発正答率に頼るのではなく、テストの生成、構文エラーの検知、ログ解析に基づく自律修正といった一連のワークフローを粘り強く反復する特性を持つ[36]。GDNによる履歴圧縮とQSAによる精密検索がコンテキストの破綻を防ぐため、エージェントハーネス（Claude Code、mini-SWE-agent、Qwen Code等）を介した推論時スケーリングにおいて卓越した完遂力を発揮する[3]。

また、システム側からモデルの思考プロセスを制御するための専用インターフェースが整備されている[14]。enable_thinking パラメータによる思考モードのオン・オフ切り替えに加え、対話履歴全体の思考トレースを保持して複数ターンにわたる推論精度を高める preserve_thinking 機構を備える[14]。さらに reasoning_effort パラメータによって、難解なタスク向けの xhigh（デフォルト）から、バランス型の medium、高速応答を重視した low、思考を行わない none まで推論深度を調整可能であり、要求されるレイテンシや運用コストに応じた柔軟な制御が実現されている[14]。

ライセンス体系と商用APIエコシステム

Qwen3.8-Flash-Nextの公開形態および商用サービングの条件は、従来のオープンソースモデルと明確に区別されている[3]。

先行して公開された密結合型モデル「Qwen3.8-27B」がApache License 2.0を採用していたのに対し、本モデルは「Qwen Community License 1.0」の下で提供されている[3]。このライセンスに基づき、個人利用、研究開発、および社内利用を目的とした一般的な商用展開は原則として無償で許諾される[3]。しかし、モデル推論APIを外部提供するModel-as-a-Service（MaaS）事業者や、AIコーディング支援およびオフィス自動化ツールの提供を主目的とする事業者が自社製品に組み込む場合は、Alibabaとの個別商用契約の締結が義務付けられている[3]。また、月間アクティブユーザー数（MAU）が1億人を超えるか、または月間売上高が2,000万米ドルを超える大規模サービスで運用される場合には、エンドユーザー向けUI上にモデル名を明記することが条件付けられている[3]。

クラウドAPIとして提供される製品版「Qwen3.8-Flash」は、Qwen Cloudを通じて100万トークンのコンテキストウィンドウと標準ツール群を備えた状態で提供される[8]。その価格設定は、入力100万トークンあたり0.16米ドル（約1人民元）、出力100万トークンあたり0.47米ドル（約3人民元）と極めて廉価に設定されている[12]。これは、同等のエージェント性能を持つ商用フロンティアモデル群（入力5米ドル、出力25〜30米ドル水準）と比較して入力側で30倍以上、出力側で60倍以上のコスト優位性を有しており、エンタープライズにおける大規模バッチ処理や高頻度エージェント実行の経済性を大きく変革する要因となっている[46]。

技術的意義と今後の展望

Qwen3.8-Flash-Nextがもたらした最大の工学的示唆は、基盤モデルにおける「計算と記憶の物理的・論理的分離」を実用レベルで成立させた点にある[9]。従来のアーキテクチャでは、全パラメータを高価なGPU HBM上に配置し、トークンごとに巨大な行列積を駆動させていたが、本モデルはN-gram埋め込みをホストメモリやSSDに逃がし、Layer 1実行中の非同期プリフェッチによって通信オーバーヘッドを隠蔽する手法を実証した[3]。これは、ハードウェアリソースの制限が厳しいオンプレミス環境やエッジワークステーションにおいて、大規模知識モデルを現実的なコストで運用するための新たな指針となる[15]。

さらに、GDNによる線形状態圧縮とQSAによるブロック単位スパース検索のハイブリッド化は、数十万～100万トークンに及ぶ超長文脈処理におけるKVキャッシュ問題に対する実効的な解決策を示した[3]。エージェントが複雑なタスクを完了するまでに発生する大量のコンテキスト読み戻しを、極めて低いPrefillコストと高速なスループットで維持できるため、自律型AIエージェントの基盤としての実用性が極めて高い[3]。

125B総パラメータ・6Bアクティブという極限の疎構造においてフロンティアモデルに匹敵する性能が証明されたことにより、正式な「Qwen4」シリーズではこのハイブリッド・疎結合アーキテクチャがさらに大規模なスケールで展開される見通しである[2]。モデルアーキテクチャの革新と超低コストAPIの供給が両輪となり、オープンウェイトモデルを中心とした生成AIエコシステムの構造転換が今後一段と加速していくと考えられる[12]。

引用文献

1. Qwen3.8-Flash-Next徹底解説:Qwen4を先出しする新モデルの正体と, https://note.com/ai_driven/n/n7a7c2fe2fbad
2. 「Qwen3.8-Flash-Next」27日ダウンロード可能に。125B-A6Bや35B-A3Bか？, https://pc.watch.impress.co.jp/docs/news/2135530.html
3. 「Qwen3.8-Flash-Next」無償公開、Opus 4.6匹敵でQwen3.8-27B超え, https://pc.watch.impress.co.jp/docs/news/2135941.html
4. 「Qwen3.8-Flash-Next」がオープンモデルとして公開される, https://gigazine.net/news/20260827-qwen3-8-flash-next/
5. RTX 5090 + RAM 128GBでQwen3.8-Flash-Nextをllama.cpp ... - Zenn, https://zenn.dev/holy_fox/articles/04887ff8177b87
6. Qwen3.8-Flash-Next: Day-0 Support in SGLang - LMSYS Org, https://www.lmsys.org/blog/2026-08-26-qwen-flash-next/
7. Alibaba’s Qwen Team Releases Qwen3.8-Flash-Next: A 125B Multimodal MoE With 6B Active Parameters Previewing the Qwen4 Architecture, https://www.marktechpost.com/2026/08/26/alibabas-qwen-team-releases-qwen3-8-flash-next-a-125b-multimodal-moe-with-6b-active-parameters-previewing-the-qwen4-architecture/
8. Qwen/Qwen3.8-Flash-Next - Hugging Face, https://huggingface.co/Qwen/Qwen3.8-Flash-Next
9. Alibaba Previews Qwen4 Architecture: 6B Active Params, One-Ninth the Training FLOPs, https://xenospectrum.com/en/qwen-flash-next/
10. Qwen/Qwen3.8-Flash-Next - vLLM Recipes, https://recipes.vllm.ai/Qwen/Qwen3.8-Flash-Next
11. AlibabaがQwen4の仕組みを先行公開 Qwen3.8-Flash-Next,

https://www.ai-jitan-hub.com/news/qwen38-flash-next
12. Qwen3.8-Flash-Next: A New Architecture, Towards Ultimate Cost, https://qwen.ai/blog?id=qwen3.8-flash-next
13. Alibaba releases Qwen3.8-Flash AI model to cut costs, https://www.investing.com/news/stock-market-news/alibaba-releases-qwen38flash-ai-model-to-cut-costs-93CH-4877431
14. Qwen3.8-Flash-Next - Model Details, https://www.modelscope.ai/models/Qwen/Qwen3.8-Flash-Next
15. Qwen 3.8 Flash Next の N-GRAM/PLEを公式資料とDay zero実装, https://zenn.dev/shunmoridev/articles/d276ea75c252cf
16. Qwen3.8-Flash-Next-GGUF - 模型详情页, https://www.modelscope.cn/models/unsloth/Qwen3.8-Flash-Next-GGUF
17. Qwen3.8-Flash-Next の 51B N-gram 埋め込みは SSD から ... - Qiita, https://qiita.com/sukimaengineer/items/222043051726a1ad09c7
18. 见路非道 - 博客园, https://www.cnblogs.com/sing1ee
19. Qwen3.8-Flash-Next: Open Weights, Qwen4 Architecture - LLM Stats, https://llm-stats.com/blog/research/qwen3.8-flash-next-launch
20. Alibaba Cloud、「Qwen3.8-27B」のオープンウェイトを公開, https://gihyo.jp/article/2026/08/qwen-3.8-27b
21. Qwen/Qwen3.8-27B - Hugging Face, https://huggingface.co/Qwen/Qwen3.8-27B
22. Qwen3.8-Flash-Next: ローカルでの実行方法 | Unsloth Documentation, https://unsloth.ai/docs/jp/moderu/qwen3.8-next
23. unsloth/Qwen3.8-Flash-Next-GGUF - Hugging Face, https://huggingface.co/unsloth/Qwen3.8-Flash-Next-GGUF
24. Experiment with Qwen3.8-Flash-Next 176B Model on NVIDIA GB300 NVL72 for Agentic Coding | NVIDIA Technical Blog, https://developer.nvidia.com/blog/experiment-with-qwen3-8-flash-next-176b-model-on-nvidia-gb300-nvl72-for-agentic-coding/
25. unsloth/Qwen3.8-Flash-Next-GGUF · Share your model speed here, https://huggingface.co/unsloth/Qwen3.8-Flash-Next-GGUF/discussions/3
26. Got Qwen3.8-Next-Flash ngram SSD offload working in llama.cpp!, https://www.reddit.com/r/LocalLLM/comments/1vz927j/got_qwen38nextflash_ngram_ssd_offload_working_in/
27. Run Qwen3.8-Flash-Next Locally - Atomic Chat, https://atomic.chat/models/qwen3-8-flash-next
28. Qwen3.8-Flash-Next Benchmarks & Context (August 2026) - BenchLM, https://benchlm.ai/models/qwen3-8-flash-next
29. Qwen3.8-Flash-Next が Qwen4 アーキテクチャを 6B ... - Unite.AI, https://www.unite.ai/ja/qwen3-8-flash-next-previews-qwen4-architecture-with-6b-active-parameters/
30. Qwen3.6-27B vs Qwen3.8-Flash-Next: Benchmarks & Cost - BenchLM, https://benchlm.ai/compare/qwen3-6-27b-vs-qwen3-8-flash-next
31. How to Run Qwen3.8 Flash Next Locally: GGUF, Hardware and, https://atomic.chat/blog/guides/how-to-run-qwen-3-8-flash-next-locally
32. Qwen 3.8 Flash vs DeepSeek 0731 vs GL5 5.3 Flash : r/LocalLLM,

https://www.reddit.com/r/LocalLLM/comments/1vz2tae/qwen_38_flash_vs_deepseek_0731_vs_gl5_53_flash/
33. 6B activeでDeepSeek V4 Flash超えを示すQwen公式ベンチ - note, https://note.com/cute_agapan9087/n/n78bf6d12682c
34. qwen3.8-flash-next - Ollama, https://ollama.com/library/qwen3.8-flash-next
35. Qwen3.8-Flash-Next: Alibaba's 125B MoE Model, 1M-Token Context, https://mer.vin/news/qwen3-8-flash-next-alibabas-125b-moe-model-1m-token-context/
36. Qwen3.8-27Bのタスク完遂力がすごい。賢さではなく作業手順の, https://nowokay.hatenablog.com/entry/2026/08/19/152648
37. Qwen3.8-27B におすすめのハーネス｜npaka - note, https://note.com/npaka/n/nf7108e6bd18e
38. Qwen3.8-27B派生版が大量発生、トレンド上位埋め尽くす。10件は「無検閲版」で過半数（生成AIクローズアップ）, https://www.techno-edge.net/article/2026/08/26/5426.html
39. Alibaba、Qwen4設計を先行公開：6B活性・訓練量3分の1でFLOPs9, https://xenospectrum.com/qwen-flash-next/
40. Qwen3.8-Flash-Next: the Qwen4 Architecture Preview Is ... - ModelFit, https://modelfit.io/blog/qwen38-flash-next-open-weights/
41. Qwen3.8-Flash-Next: A Qwen4 Architecture Preview, https://developer.tenten.co/qwen38-flash-next-qwen4-architecture
42. Qwen3.8-Flash-Next API Pricing & Launch Impact, https://www.aipricing.guru/news/qwen3-8-flash-next-api-pricing-august-2026/
43. [Megathread] Qwen3.8-Flash-Next - Release Day : r/LocalLLaMA, https://www.reddit.com/r/LocalLLaMA/comments/1vyq2v4/megathread_qwen38flashnext_release_day/
44. Alibaba releases Qwen3.8-Flash-Next, targeting "ultimate cost, https://the-decoder.com/alibaba-releases-qwen3-8-flash-next-targeting-ultimate-cost-efficiency/
45. Qwen3.8-Flash vs Qwen 3.8: the 6B-active MoE against the 2.4T, https://www.orcarouter.ai/blog/qwen-3-8-flash-vs-qwen-3-8
46. Alibaba’s 125B Qwen Launch Escalates the AI Inference Price War, https://www.ctol.digital/news/alibaba-qwen-3-8-flash-mid-tier-ai-api-pricing/
47. Alibaba's Qwen launches Qwen3.8-Flash AI model with lower training costs, https://m.economictimes.com/tech/artificial-intelligence/alibabas-qwen-launches-qwen3-8-flash-ai-model-with-lower-training-costs/articleshow/133544818.cms