Qwen3.8-Flash-Next 調査レポート

― 2026 年 8 月 26 日公開の事実確認、技術的特徴、性能主張の検証と実務への示唆 ―

2026 年 8 月 27 日

Claude Opus 5

---

1. 要旨

- 前提は正しい。2026 年 8 月 26 日、Alibaba（阿里巴巴）の Qwen（通義千問）チームは「Qwen3.8-Flash-Next」をオープンウェイトで公開した。モデル名の表記も正確である。同日、API 提供の量産版「Qwen3.8-Flash」も併せてローンチされている。[1,2,5]
- その正体は「Qwen4 アーキテクチャの早期プレビュー」である。総パラメータ 125B・アクティブ 6B のマルチモーダル MoE で、GDN＋QSA ハイブリッドアテンション、N-gram 埋め込み、Gated Residual、MTP を搭載する。Qwen3.7-Plus 比で学習コスト約 9 分の 1 を主張し、自社ベンチでは Claude Opus 4.6 を多くの項目で上回るとする。[1,2,7]
- ただし公表数値はすべて自社ベンチであり、執筆時点で第三者による独立評価は公開されていない。またライセンスは Apache 2.0 ではない独自の「Qwen Community License 1.0」である。実務投入時は、性能値の未検証性とライセンスの MaaS 条項の双方に留意を要する。[12,17,18]

2. 前提の真偽検証

ご提示の前提「2026 年 8 月 26 日に Qwen が『Qwen3.8-Flash-Next』を公開した」は、一次情報源で確認でき、正確である。以下、確認事項を整理する。

公開主体は Alibaba Group の Qwen（通義千問）チームであり、正式名称は Qwen3.8-Flash-Next（Hugging Face リポジトリ ID：Qwen/Qwen3.8-Flash-Next）である。[1,2,3]

公開日については、オープンウェイトが 2026 年 8 月 26 日 23 時（北京時間、UTC+8）に Hugging Face および ModelScope で公開された。Reuters は同日、北京発で、Alibaba 傘下の Qwen が新たなマルチモーダルモデルを公開し、コーディングおよびオフィス業務の性能向上と学習コスト削減を同社が声明で表明した旨を報じている。[4] 中国メディア（IT 之家、界面新闻、扬子晚报等）も同日付で報道した。[5,20,32] なお OrcaRouter は、Hugging Face のリポジトリ自体は 8 月 24 日から露出していたと報告しているが、これは一次情報での裏付けが取れていないため、参考情報として扱う。[8]

2.1　混同しやすい近縁モデル

名称の近い別モデルが複数存在するため、実務上の取り違えに注意を要する。

- Qwen3.8-Flash：同日ローンチされた API 提供の量産版。オープンウェイト版とは別物である。[15,20]
- Qwen3-Next（Qwen3-Next-80B-A3B）：2025 年 9 月 11 日公開の前世代プレビュー。総 80B／アクティブ 3.9B、Gated DeltaNet＋Gated Attention のハイブリッド構成。[7]
- 表記のゆれとしてしばしば見られる「Qwen3.8-Next-Flash」は誤りである。

3. モデル規模・構成

以下は Hugging Face 公式モデルカードおよび公式ブログに基づく確認事実である。[1,2]

- 種別：Causal Language Model ＋ Vision Encoder。テキスト・画像・動画入力に対応するマルチモーダルモデル。[2,10]
- パラメータ：主モデル 125B（うちアクティブ 6B／トークン）、N-gram 埋め込み 51B、MTP 4B。ディスク上の総計は約 180B（BF16 表記）。[2]
- MoE 構成：512 エキスパート、アクティブ 10 routed ＋ 1 shared、エキスパート中間次元 640。[2]
- レイヤ：48 層。レイアウトは 12×（3×（Gated DeltaNet→MoE）→ 1×（Qwen Sparse Attention→MoE））で、GDN 3 層に対し QSA 1 層の 4 対 1 構成。[1,2]
- コンテキスト長：ネイティブ 262,144 トークン、YaRN により最大 1,000,000 トークンまで拡張可能。[2,13]
- 派生モデル：FP8 量子化版（Qwen3.8-Flash-Next-FP8）を同時公開。[11] Instruct／Thinking のリポジトリ分割はなく、単一チェックポイントがハイブリッド思考モデルとして機能する（enable_thinking／reasoning_effort で制御、既定は思考モード ON・reasoning_effort=xhigh）。[2]
- ライセンス：qwen-community-1.0（Qwen Community License 1.0）。Apache 2.0 ではない点に注意。同シリーズの Qwen3.8-27B が Apache 2.0 であったのと対照的である。[12,24]

【留意】総パラメータの表記は情報源によって「125B」「176B」「180B」と揺れる。[9] 本レポートは公式モデルカードの内訳（125B＋51B＋4B）と Hugging Face の「180B params」表記を採用した。「176B」は二次情報における計上差または丸めと推定される（分析）。

4. 技術的特徴

4.1 「Next」の位置づけ

公式は本モデルを「Qwen4 を支えるアーキテクチャの実験的プレビュー」と明言している。[1] これは Qwen3-Next が Qwen3.5 に対して果たした役割と同一であり、完全な Qwen4 ファミリーの構築前に、アーキテクチャ変更をコミュニティに早期検証させる狙いと位置づけられる。[7] 系譜としては、Qwen3-Next で導入された Gated DeltaNet（GDN）＋Gated Attention のハイブリッドが Qwen3.5 以降の各世代で標準採用され、今回それが GDN＋QSA（Qwen Sparse Attention）へと刷新された。

4.2 4 つの刷新軸

- Hybrid Attention with QSA：QSA は軽量 Indexer により micro-block 粒度で重要コンテキストを選択する。公式主張では、1M トークン時の Attention Kernel が prefill で最大 7.6 倍・decode で 4.9 倍高速化し、Qwen3.7-Plus 比で prefill スループットは 8.6 倍とされる。[1]
- Gated Residual：4 分岐の残差ストリームをデータ依存ゲートで制御し、学習安定性と表現力を両立させる。[1]
- N-gram Embedding：51B のルックアップ表（bigram／trigram、2,000 万エントリ）を第 2 層に配置する。演算量をほぼ増やさずに容量を拡張でき、ホスト RAM への非同期オフロードが可能である（現状オフロード対応は NVIDIA デバイスのみ）。[1,14]
- Tailored Training Recipe：Muon と AdamW を重み種別ごとに適用し、batch-size warmup を廃して目標バッチサイズから直接開始する。[1]

効率面では、Qwen3.7-Plus（397B／アクティブ 17B）比で学習コスト約 9 分の 1 を主張する。中国語プレスは「学習コスト約 90%減、推論コストも大幅減」と表現している。[5,32]

5. ベンチマーク性能

以下はすべて Alibaba 自社公表値である（公式モデルカード掲載の言語系ベンチマーク）。[2]

ベンチマーク	Qwen3.8-Flash-Next	DeepSeek-V4-Flash	Claude Opus 4.6
DeepSWE 1.1（Agentic coding）	58.7	54.4	—
SWE-bench Pro	62.5	56.0	53.4

ベンチマーク	Qwen3.8-Flash-Next	DeepSeek-V4-Flash	Claude Opus 4.6
SWE-bench Multilingual	81.0	—	77.5
NL2Repo-Bench	48.1	54.2	—
CoWorkBench	73.9	45.1	—
JobBench	55.7	—	36.6
IFBench	81.3	—	62.5
GPQA Diamond	91.7	—	91.3
HLE（Humanity's Last Exam）	35.9	—	40.0
LiveCodeBench v6	91.9	—	88.8

HLE は主要項目のうち唯一 Claude が上回った項目である。マルチモーダル系では AndroidWorld 84.5（Claude 62.0）、MathVision 90.6、ERQA 72.3 等で優位を主張している。[2,16]

5.1　独立検証の状況

執筆時点（2026 年 8 月末）において、第三者による独立評価は公開されていない。Artificial Analysis、LMArena、LiveBench、Aider、SWE-bench 公式リーダーボードのいずれも、Flash-Next および API 版 Qwen3.8-Flash のスコアを未掲載である。近縁の Qwen3.8-Max（Intelligence Index 58）[18] や Qwen3.8-27B（同 52）[19] のスコアは存在するが、いずれも別モデルである。集計サイト BenchLM は「67.54／100・226 中 25 位」と表示するものの、同サイト自身が検証済みポジションに足る出典が不足している旨を明記しており、独立実測ではなくベンダー申告ベースの推計と解される。[17]

したがって、公表値は現時点では「主張」として扱うのが妥当である（評価）。

6. 提供形態と価格

- オープンウェイト：Hugging Face および ModelScope で BF16 と FP8 を公開。[2,11]
- API 量産版「Qwen3.8-Flash」（QwenCloud）：入力 0.16 米ドル／100 万トークン、出力 0.47 米ドル／100 万トークン（中国国内表記では入力 1 元・出力 3 元／百万トークン）。既定 1M コンテキスト、公式内蔵ツール付き。[15,20,21]
- 中国メディアは、DeepSeek-V4-Flash の最安価格帯の 3 分の 1 と報じている。[21]

- 上位の Qwen3.8-Max（入力 2.00 米ドル／出力 6.00 米ドル）と比較して約 12 分の 1 の価格水準にあたる。[18]

7. Alibaba のモデル戦略における位置づけ（分析）

2026 年 8 月の Alibaba は異例の高頻度リリースを展開した。8 月 3 日にフラグシップ Qwen3.8-Max（2.4 兆パラメータ、アクティブ 95B）を API ローンチし[23,26]、8 月 12〜13 日にそのオープンウェイト版 Qwen3.8-2.4T-A95B、8 月 14 日に軽量 Qwen3.8-27B（Apache 2.0）、そして 8 月 26 日に Flash-Next を投入している。[24,25]

普及規模について Alibaba は、2026 年 8 月 15 日付のメール声明で、過去 6 か月の全世界ダウンロードが 30 億回超、オープンソース化したモデルが 460 以上、派生モデルが 30 万件超に達したと主張している。[27] ただしこの自社発表値は第三者プラットフォームのデータと乖離がある。Hugging Face の「State of Open Models」レポート（2026 年 8 月公表）は、同 Hub 上の Qwen 累計ダウンロードを約 20.45 億回、派生モデルを 151,448 件と集計しており、企業公表値はプラットフォーム実測の約 1.5 倍規模となる。[28] 実務上は、後者を実勢値の下限として参照するのが安全である（評価）。

戦略構造は明確に階層化されている。すなわち Max 系＝最上位（クローズド起点からオープン化）、Flash／Next 系＝高効率・低価格の量産およびアーキテクチャ実験レイヤーという配置である。[22] 特筆すべきは、Max 級モデルまでオープンウェイト化するという 2024 年までの方針からの転換であり、これは DeepSeek が実証した「オープン＋低価格で市場をコモディティ化する」プレイブックを、大企業スケールで実行しているものと解釈できる（分析）。[26]

Flash-Next において「Qwen4 アーキテクチャを完成品より先に公開する」という動きは、①ランタイム・量子化・アプリのエコシステムを Qwen4 本番投入前に整備させる、②開発者を早期に囲い込む、③新アーキテクチャをコミュニティ環境で大規模にグレーテストする、という複合的な狙いによるものと分析される。[8,34]

8. 反響・評価

8.1 技術コミュニティ

Hacker News では公開前の予告段階から首位（272 ポイント）を獲得した。[8] Reddit の r/LocalLLaMA では、4bit 量子化で 80〜90GB VRAM を要するという見積もりが議論され、N-gram 表（約 24GB）をシステム RAM にオフロードできる設計が、高 RAM 構成の個人でも動かせ

る点として評価された。[14,31] Unsloth は公開初日に GGUF を提供している。[29,30]

8.2　メディアの温度差

中国国内メディア（搜狐、新浪、IT 之家、扬子晚报など）は「全新架构、极致性价比（全く新しいアーキテクチャ、究極のコストパフォーマンス）」として好意的に大きく報じ、「千億パラメータ規模でアクティブ 60 億のみで Opus 4.6 を上回る」という効率の物語を前面に置いた。[5,21,32] 一方、英語圏メディア（The Decoder、Reuters、Decrypt、MarkTechPost 等）は、報道自体は同様に扱いつつ、自社ベンチである点、実世界性能は別である点、価格圧力という文脈を併記する傾向が見られる。[7,15]

8.3　懐疑論・批判

- ベンチマークはすべて自社ハーネスによるものであり、一部（CoWorkBench、RecreationBench 等）は自社製ベンチである。独立再現がない。[17]
- 比較対象の Claude Opus 4.6 は 2026 年 2 月時点のモデルであり、最新の Opus 4.8 ではない点がフェアネス上の論点となる。[16]
- ライセンスが Apache 2.0 でない点。[12]

9.　ライセンス精査

Hugging Face 掲載の LICENSE 原文を確認した結果、以下のとおり整理される。[12]

- 地理的制限は存在しない。一部で流布した「米国・EU・英国・韓国でのダウンロード／使用を禁止」との懸念は、実際には 2026 年 8 月初旬の Qwen3.8-Max／27B および MiniMax H3 をめぐる別の議論に由来し、後に発言者自身が表現が強すぎたとして一部撤回したものである。Flash-Next のライセンス本文に該当条項はない。[12,24]
- MAU 閾値は「表示義務のトリガー」であって商用禁止ではない。月間 1 億アクティブユーザーまたは月間売上 2,000 万米ドルを超える製品では、UI にモデル名を明示する必要がある（LICENSE 本文 Condition 1）。[12]
- 実質的な商用ゲートは MaaS 条項である。ライセンシーが「Model as a Service」または「AI Work Assistant」事業（AI 支援コーディング／オフィス生産性の独立製品）を営む場合、商用利用前に Qwen から別途ライセンスを取得する必要がある（Condition 2）。第三者に出力やモデル能力を提供しない内部利用は対象外である。[12]

- OSI 承認のオープンソースではない。訓練データ非開示および利用分野制限を含むため、OSI の Open Source AI Definition を満たさない「オープンウェイト（source-available）」に分類される。[35]

10. 米中フロンティア競争への含意（分析）

Flash-Next は、性能でフロンティアに肉薄しつつ価格を 1 桁下げるという中国オープンウェイト勢の典型であり、クローズドモデル（OpenAI／Anthropic／Google）の API 高価格を持続困難にする圧力として機能する。実際 The Decoder は、OpenAI が新たな GPT-5.6 系において大幅な値下げで応じたと報じている。[15]

オープンウェイトである点は、米国の政策枠組みにおいてオープンウェイトモデルが連邦セキュリティレビューの対象外とされる状況とも相まって、規制摩擦を回避しつつ普及を狙える構造をもたらす。ただし「アーキテクチャ効率でのブレークスルー」という主張が独立検証されるまでは、実力の確定は保留すべきである（評価）。

11. 日本企業の実務への示唆

特に知的財産・研究開発分野におけるローカル／自社環境での LLM 活用を念頭に、段階的な対応を提案する（以下は分析・提案であり、確認事実ではない）。

11.1 評価フェーズ（即時）

新規オープンウェイトは、まず本番から隔離した低リスクの複製ワークロードで既存モデル（社内標準）と出力比較する。Flash-Next は agentic coding および長文脈オフィス業務で強い自社ベンチを示すため、コードレビュー支援、長大な社内文書・特許明細書の要約や横断検索が検証候補となる。判断閾値としては、自社の代表タスクで既存モデル比±5%以内かつコスト 5 分の 1 以下であれば本格 PoC へ進める、といった設定が考えられる。

11.2 ライセンス審査（PoC 前に必須）

qwen-community-1.0 は Apache 2.0 と異なるため、法務レビューを先行させる必要がある。確認すべきは、①自社が「Model as a Service／AI Work Assistant」事業に該当しないか（社外に LLM 機能を提供する SaaS は別途ライセンスを要する）、②1 億 MAU／月商 2,000 万米ドル超であればモデル名の UI 表示義務が生じる点、の 2 点である。[12] 純粋な社内利用（知財調査、研究開発、社内文書処理で出力を第三者に提供しない）であれば、比較的自由に利用できる。MaaS 該当の可能性

がある場合は、Qwen へのライセンス照会を完了するまで商用展開を止めるべきである。

11.3　インフラ（ローカル運用）

BF16 フルは約 355GB 級となり現実的ではない。4bit 量子化で 80〜90GB 級、Unsloth の 1bit 系では約 75GB で RAM／unified memory 上での実行可能性が報告されている。[29,31] N-gram 表の RAM オフロード設計により、GPU VRAM が潤沢でなくとも比較的動作する点は、知財部門の自社サーバ運用にとって有利に働く。[14] 機微な特許・発明情報を扱う場合は、データが社外送信される API 版ではなく、オープンウェイトのローカル／VPC 隔離運用を選択すべきである。

11.4　再評価の条件

Artificial Analysis、LMArena、SWE-bench 公式が Flash-Next のスコアを公開した時点で再評価する。自社ベンチとの乖離が大きい場合（特にコーディング・エージェント系で公表値を大きく下回る場合）は採用を見送り、Apache 2.0 の Qwen3.8-27B や他のオープンモデルを代替候補とするのが妥当である。

12.　留保事項

- 性能数値はすべて Alibaba 自社公表値であり、独立した第三者評価は執筆時点で未確認である。特に CoWorkBench、JobBench、RecreationBench 等は自社製ベンチであり、絶対値の外部比較可能性は限定的である。
- 比較対象の公平性：Claude との比較は Opus 4.6（2026 年 2 月）に対するものであり、最新の Opus 4.8 ではない。DeepSeek-V4-Flash との比較も一部項目で数値が欠落している。
- 普及規模の数値：Alibaba 自社発表（30 億 DL・30 万派生）は Hugging Face 実測（約 20.45 億 DL・151,448 派生）と約 1.5 倍の乖離があり、自社値は複数プラットフォーム合算・自己申告である点に留意を要する。
- パラメータ表記のゆれ：総計を「125B」「176B」「180B」と報じる情報源が混在している。
- 公開日時の先行露出：一部で「Hugging Face リポジトリは 8 月 24 日に先行公開」と報告されているが、一次確認は未了である。正式リリースは 8 月 26 日で確定している。
- ライセンス解釈は法的助言ではない。qwen-community-1.0 の適用可否・MaaS 該当性は個別事情により変わるため、商用展開前に必ず自社法務および専門家の確認を要する。
- Base／Instruct／Thinking 派生：単一のハイブリッド思考チェックポイントとして公開されている。事前学習のみの Base ウェイトの一般公開有無は未確認である。

- 技術レポート：tech_report.pdf が GitHub（QwenLM/Qwen3.8-Flash-Next）に存在するとされるが、本調査では全文を精読していない。学習データ規模・学習 FLOPs の絶対値等は未確認である。[3]
- 参考文献のうち [4]、[27]、[28] は、本調査において一次記事 URL を直接確認できていない。内容の裏付けとしては二次情報経由である点にご留意いただきたい。

参考文献

※ URL はいずれも 2026 年 8 月 27 日時点。特記のあるものは一次情報での確認が取れていない。

[1] Qwen Team, Alibaba Group「Qwen3.8-Flash-Next: A New Architecture, Towards Ultimate Cost Efficiency」（公式ブログ） https://qwen.ai/blog?id=qwen3.8-flash-next

[2] Hugging Face モデルカード「Qwen/Qwen3.8-Flash-Next」 https://huggingface.co/Qwen/Qwen3.8-Flash-Next

[3] GitHub「QwenLM/Qwen3.8-Flash-Next」リポジトリ https://github.com/QwenLM/Qwen3.8-Flash-Next/

[4] Reuters（北京発、2026 年 8 月 26 日）Alibaba 傘下 Qwen の新マルチモーダルモデル公開に関する報道 ※本調査では記事 URL を直接確認できていない（要確認）

[5] IT 之家「阿里通义 Qwen3.8-Flash（Next）发布开源：125B 参数 MoE 模型，训练成本仅为前代 1/9」（2026 年 8 月 26 日） https://www.ithome.com/0/994/735.htm

[6] TechNode「Alibaba's Qwen to open-source Qwen3.8-Flash-Next, previewing Qwen4 architecture」（2026 年 8 月 26 日） https://technode.com/2026/08/26/alibabas-qwen-to-open-source-qwen3-8-flash-next-previewing-qwen4-architecture/

[7] MarkTechPost「Alibaba's Qwen Team Releases Qwen3.8-Flash-Next: A 125B Multimodal MoE With 6B Active Parameters Previewing the Qwen4 Architecture」（2026 年 8 月 26 日） 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] OrcaRouter「Qwen3.8-Flash-Next Is Out: Qwen4 Architecture Confirmed」 https://www.orcarouter.ai/blog/qwen-3-8-flash-next-leak

[9] NVIDIA Technical Blog「Experiment with Qwen3.8-Flash-Next 176B Model on NVIDIA GB300 NVL72 for Agentic Coding」 https://developer.nvidia.com/blog/experiment-with-qwen3-8-flash-next-176b-model-on-nvidia-gb300-nvl72-for-agentic-coding/

[10] Unite.AI「Qwen3.8-Flash-Next Previews Qwen4 Architecture With 6B Active Parameters」 https://www.unite.ai/qwen3-8-flash-next-previews-qwen4-architecture-with-6b-active-parameters/

[11] Hugging Face モデルカード「Qwen/Qwen3.8-Flash-Next-FP8」 https://huggingface.co/Qwen/Qwen3.8-Flash-Next-FP8

[12] Qwen Community License Agreement 1.0（Hugging Face 掲載の LICENSE 原文） https://huggingface.co/Qwen/Qwen3.8-Flash-Next/blob/main/LICENSE

[13] vLLM Recipes「Qwen/Qwen3.8-Flash-Next」 https://recipes.vllm.ai/Qwen/Qwen3.8-Flash-Next

[14] BAGUA AI「Deep Dive into Qwen3.8-Flash-Next: How a 24GB n-gram Table Redefines Local LLM Inference」 https://baguaai.com/deep-dive-into-qwen3-8-flash-next-how-a-24gb-n-gram-table-redefines-local-llm-inference/

[15] The Decoder「Alibaba releases Qwen3.8-Flash-Next, targeting "ultimate cost efficiency"」（2026 年 8 月 26 日） https://the-decoder.com/alibaba-releases-qwen3-8-flash-next-targeting-ultimate-cost-efficiency/

[16] OfficeChai「Alibaba Releases Qwen 3.8 Flash-Next, Beats Opus 4.6 On Most Benchmarks」 https://officechai.com/ai/qwen-3-8-flash-next-benchmarks/

[17] BenchLM「Qwen3.8-Flash-Next Benchmarks & Context (August 2026)」 ※同サイト自身が出典不足を明示（集計値・要注意） https://benchlm.ai/models/qwen3-8-flash-next

[18] Artificial Analysis「Qwen3.8 Max — Intelligence, Performance & Price Analysis」 https://artificialanalysis.ai/models/qwen3-8-max

[19] Artificial Analysis「Qwen3.8 27B (xhigh) — Intelligence, Performance & Price Analysis」 https://artificialanalysis.ai/models/qwen3-8-27b

[20] 界面新闻「阿里千问正式发布 Qwen3.8-Flash」（2026 年 8 月 26 日） https://www.jiemian.com/article/14999311.html

[21] 搜狐「阿里发布 Qwen3.8-Flash 模型并开源」（2026 年 8 月 26 日） https://m.sohu.com/a/1068002885_120988533

[22] Yicai Global「Alibaba Will Make Max AI Model Open Source for First Time」 https://www.yicaiglobal.com/news/alibaba-will-make-max-ai-model-open-source-for-first-time

[23] TechBriefly「Alibaba unveils open-source AI model Qwen3.8-Max」（2026 年 8 月 3 日） https://techbriefly.com/2026/08/03/qwen3-8-max-open-source-ai-model/

[24] Latent.Space / AINews「Qwen 3.8 Max(2.4T) and 27B, new open weights models for Coding and Cowork」（2026 年 8 月 3〜4 日） https://www.latent.space/p/ainews-qwen-38-max24t-and-27b-new

[25] Releasebot「Qwen Release Notes — August 2026 Latest Updates」 https://releasebot.io/updates/qwen

[26] Yahoo Finance「Qwen3.8-Max: The Capability War Begins — Alibaba Matches US Closed-Model Pricing on Eve of Open-Weights Drop」 https://finance.yahoo.com/technology/ai/articles/qwen3-8-max-capability-war-162320284.html

[27] Alibaba のメール声明（2026 年 8 月 15 日付、Fortune／Bloomberg 経由で報じられたオープンモデル普及規模の自社発表値） ※本調査では一次記事 URL を直接確認できていない（要確認）

[28] Hugging Face「State of Open Models」レポート（2026 年 8 月公表、The Next Web 経由） ※本調査では一次レポート URL を直接確認できていない（要確認）

[29] Unsloth Documentation「Qwen3.8-Flash-Next: How to Run Locally」 https://unsloth.ai/docs/models/qwen3.8-next

[30] Hugging Face「unsloth/Qwen3.8-Flash-Next-GGUF」 https://huggingface.co/unsloth/Qwen3.8-Flash-Next-GGUF

[31] Atomic Chat「How to Run Qwen3.8 Flash Next Locally: GGUF, Hardware and Benchmarks」 https://atomic.chat/blog/guides/how-to-run-qwen-3-8-flash-next-locally

[32] 扬子晚报「全新架构，极致性价比！阿里千问 Qwen3.8-Flash 发布并开源」（2026 年 8 月 26 日） https://www.yzwb.net/news/ch/202608/t20260826_386562.html

[33] 80aj「阿里千问明晚开源 Qwen3.8-Flash-Next：基于 Qwen4 架构的多模态 MoE 模型」（2026 年 8 月 25 日） https://www.80aj.com/2026/08/25/qwen3-8-flash-next-moe/

[34] Enterprise DNA「Qwen3.8-Flash-Next ships tomorrow, the first public architecture preview of Qwen4.」（2026 年 8 月 25 日） https://enterprisedna.co/resources/ai-pulse/ai-pulse-2026-08-25-qwen3-8-flash-next-ships-tomorrow-the-first-public-architect/

[35] QubitTool「Open Source AI Licenses [2026]: Apache 2.0 to RAIL Guide」 https://qubittool.com/blog/open-source-ai-license-compliance-guide