アプサラ・カンファレンス２０２５
最新情報

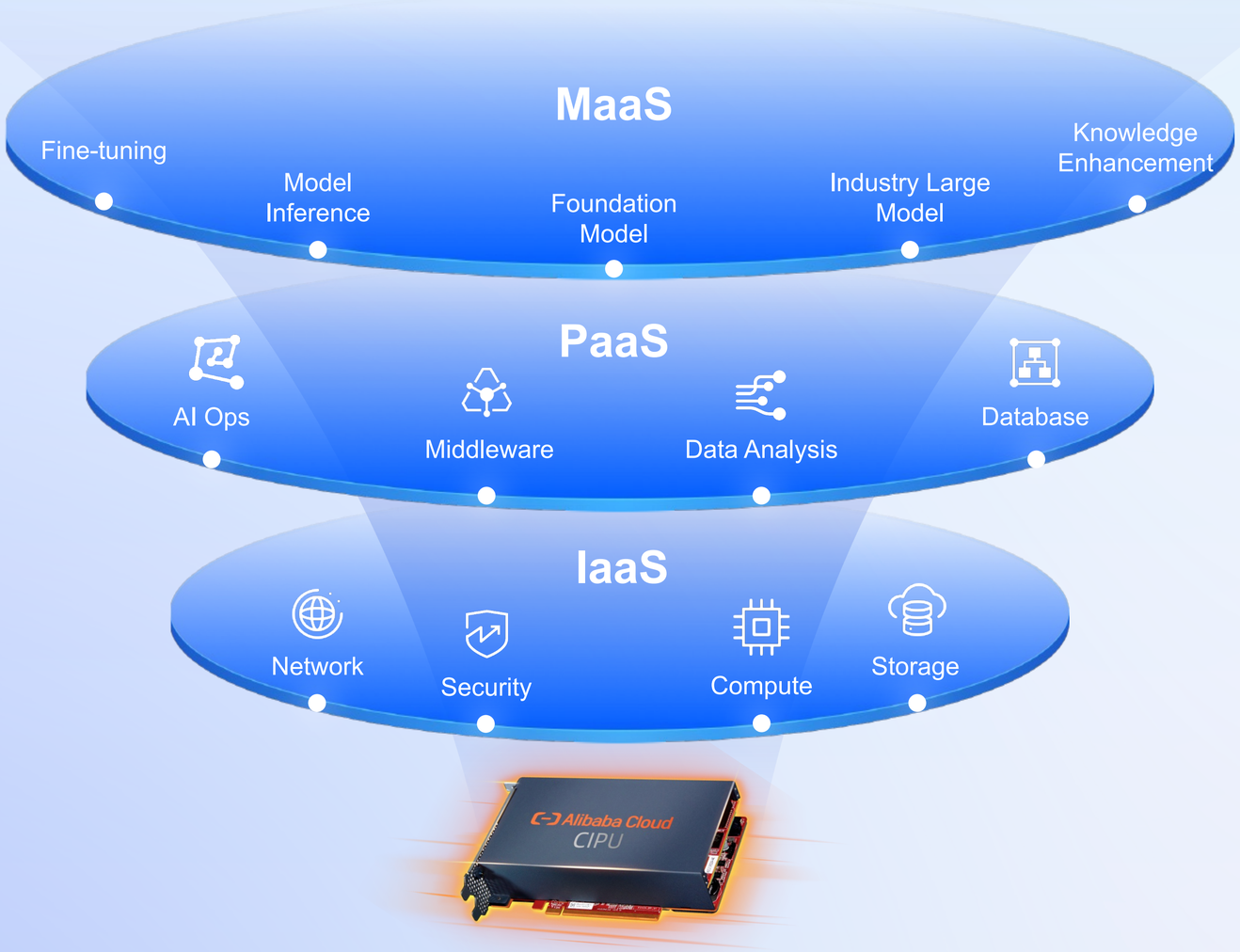
Jack Wang
Solution Architect
Alibaba Cloud Intelligence Group

王 志正 (Jack Wang)
Wang Zhizheng

キャリア経歴：

- 2024年4月アリババクラウドジャパンに入社
- 前職　Datadog Japan、SBCloud、EMC
- 現職　アリババクラウド　ソリューションアーキテクト

Full-stack AI としてのクラウドサービスはグローバルリーダーの位置付けになっています



費用対効果が高く、信頼性の高いエンドツーエンドのソリューション

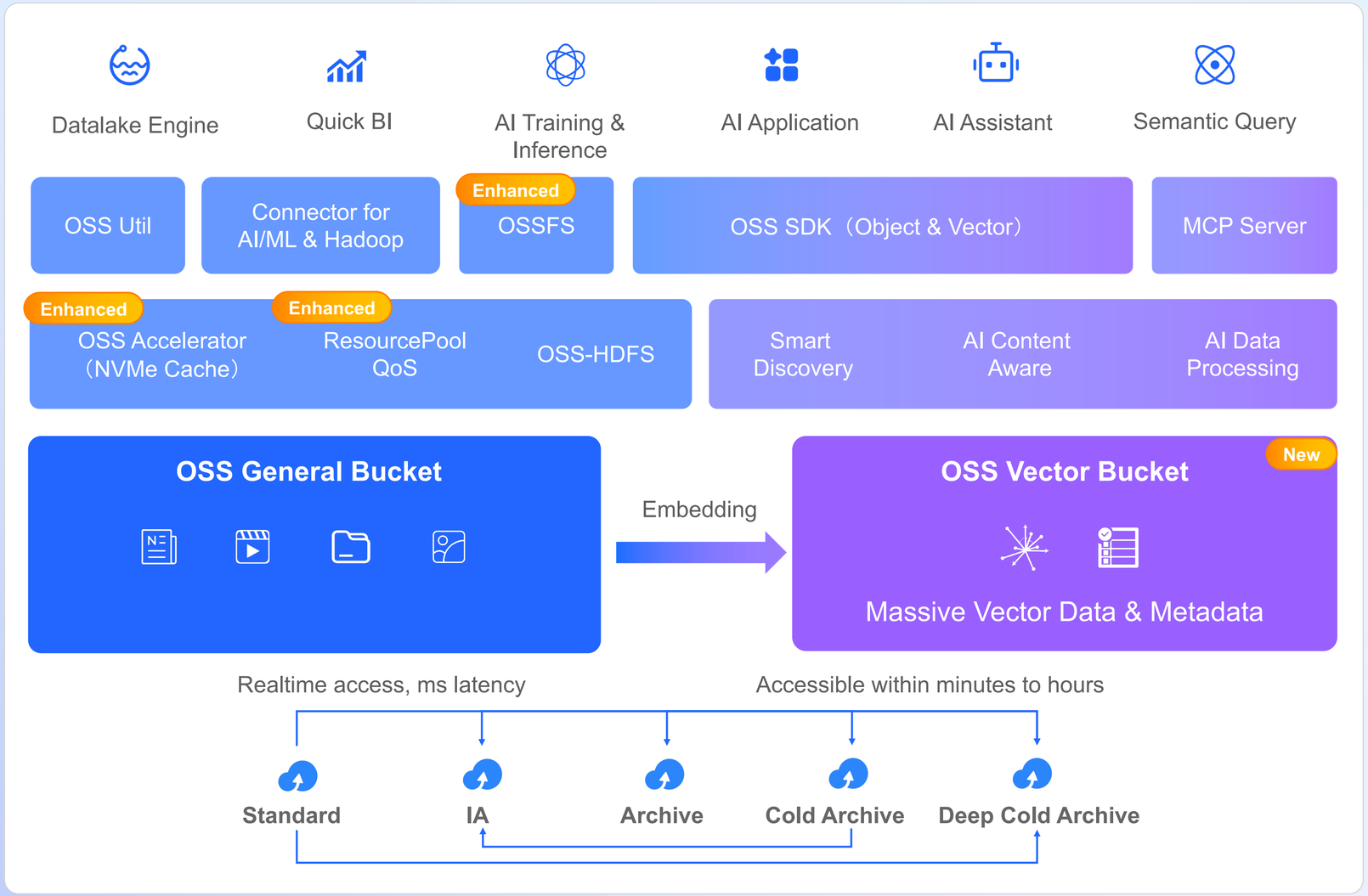
水平および垂直製品ポートフォリオの統合

自社開発のソフトウェアとハードウェアによる技術リーダーシップ

APSARA 云栖大会
APSARA CONFERENCE

AIイノベーションを大幅に加速

AIワークロード向けの大規模オブジェクトストレージ



組み込みのVector Bucketにより従来のVector DBのコストを大幅に削減します。

パフォーマンスQoS – 優先制御と最小スループットの保証

OSSアクセラレータ – 最大100GB/秒
OSSFS 2.0 – 新しいシンプルなFUSEクライアント

7.65x ↑
Single Thread Reading(*1)

3x ↑
Model Loading Speed(*2)

*1 ecs.g7.32xlarge, FIO, reading 100 GB file
*2 ecs.8gi.48xlarge, llama, loading DeepSeek-R1-GGUF model

コンテンツ認識ストレージ

OSS:データのコンテキストを理解するAIストレージ

RAG アプリ | セマンティック検索 | AIエージェント | AIデータ管理

既存コンテンツの検出

セマンティッククエリ

Query “cat”

キーワードを生成する

“cat”, “lion”, “feline”

新しいコンテンツの発見

新規データに自動タグを付ける

New “kitten” uploaded

OSS Raw Data

One-click activation
"AI Content Aware"

OSS組み込みメタデータとベクターインデックス

- OSSMeta, Multimedia Meta
- UserMeta, Vector Data,
- Content Description…

使いやすい

複雑な操作なしでワンクリックで機能を起動

グローバル展開

迅速なグローバル展開を可能にする包括的なSDK ツール

マルチディスカバリーモード

既存のコンテンツを検出してクエリし、新しいデータに自動的にタグを付ける

信頼できる結果

85%の精度を誇る成熟したアルゴリズム技術

コスト効率が高い

初期設定コストが低く、従量課金制

APSARA 云栖大会
APSARA CONFERENCE

AIワークロード向けALB：フルシナリオロードバランサーの構築

GA サポート エニーキャスト IP

Enhanced

- 高可用性ノード
- 最適なルーティング
- 最寄りのPoP経由の低遅延アクセス

近接アクセスのためのエニーキャストを備えた GA

セキュリティが組み込まれた統合エントリポイント

Enhanced

- 統合アクセスアーキテクチャ
- 集中管理
- WAFは悪意のあるトラフィックをブロックします

GA WAF

LLMインテリジェントプロキシとスケジューリング

- 異なるAIバックエンド統合
- 最適なモデル選択のためのスケジューリング
- 認証+レート制限によるコスト管理
- 完全な可観測性

ALB for AI workloads

New

ACK Kubernetes: AIコンピューティングの効率的で安定した基盤

Open Framework

Spark | Ray | Argo Workflow | Kubeflow / KServe | Tensorflow / Pytorch | Deepspeed | vLLM | SGLang

ACK Kubernetes Cluster

Scheduling
Batch | Queues | Topology-aware

Data & Model Acceleration
Fluid

AI Inference Suite (New)
RBG | Inference Gateway | LLM Performance Observation

Elastic
Resources | Tasks

Heterogeneous Resource Management
CPU | GPU | RDMA | eRDMA | CPFS | OSS | NAS

Stability (Enhanced)
Fault self-healing | inspection | diagnosis

Observability (Enhanced)
Logs | Monitoring | AI Profiling

ECS/EGS node pool

General Computing Instance

Accelerated Computing Instance

Lingjun Node Pool (New)

High-performance GPU Bare Metal cluster

Virtual Node

ACS Pod | ACS Pod | ACS Pod

AIフレームワークのオープンエコシステム

最適化されたスケジューリング | 柔軟性 | 可観測性 | セキュリティ

AI Suite: 効率的なトレーニングと推論

- LLM推論スループットが2倍に向上
- モデルのコールドスタート時間が90%短縮

インテリジェントなO&M

自己修復：GPU効率が85%向上

ACS コンテナ: Smarter, Simpler, Stronger

Enhanced

コンテナの性能向上

Intel Processor Performance :

General Purpose **20%** ↑ Performance type **24%** ↑

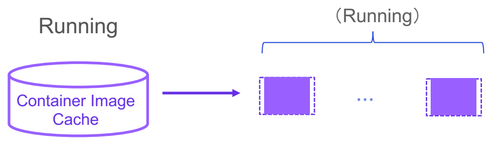
AMD processors Performance:

General Purpose **15%** ↑ Performance type **16%** ↑

Note: specCpu 2017 benchmark

New

コンテナイメージのキャッシュ加速

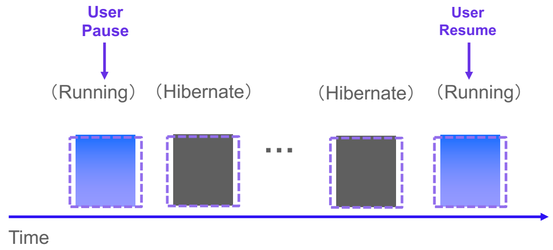


Thousands of pods millisecond pull container images

End-to-end start in seconds

Coming soon

状態保持：数秒で回復、コスト削減



During suspend: vCPU free,

memory/storage remain billed

Container Computing Service ACS

汎用 - CPU & GPU		性能 - CPU power		容量予約 - GPU Power
Default GPU power support for RDMA	BestEffort	Default	BestEffort	Default

高性能

Container performance ahead of competition

極めて高い弾力性

15,000 pods/min startup per region

利便性

12-month GA: 10M+ Pods/day on average

PAI-EAS:推論のためのコスト効率の高いサービス

Enhanced

より迅速な展開	89.8% Cold Start 97.6% Scale out	↓
低レイテンシー	20.3% TTFT 70.6% TPOT	↓
より高いスループット	71.0% TPS	↑

モデル展開レイヤー	負荷分散	スケジュールの最適化	モデル展開の加速
推論エンジンレイヤー	大規模EP	コンピューティング通信の重複	演算子の量子化
	PD分離からAF分離へ	KVスマートキャッシュ	MTP

APSARA 云栖大会
APSARA CONFERENCE

AI アプリケーション向けのフルスタック Observable Cloud Monitor 2.0

モデルアプリケーション

アプリケーション

- RAG, MCP call monitoring
- AI application monitoring
- Model call monitoring
- Safety diagnosis

モデル

- Model monitoring
- Model Log
- model link

AIプラットフォーム (PaaS)

モデル推論

- Service and Component Monitoring
- Inference Engine Monitoring

サービスコンポーネント

- Vector database monitoring
- AI Gateway Monitoring

モデルのトレーニング

- Training task monitoring
- Data Flow Monitoring

Container Kubernetes

コンテナのスケジュール

ワークロードの監視、コントロールプレーンの監視、GPU モニタリング、イングレス モニタリング、イベントおよびログ監査、AI スイートの監視

インテリジェントコンピューティングサーバー（IaaS）

AIインフラストラクチャ

- Lingjun コンピューティングノードの監視
- ネットワーク RDMA 監視
- 高速ストレージ CPFS 監視
- GPU クラウドサーバー監視

フルスタック監視
Intelligent IaaS, PaaS, Application Full Linインテリジェントな IaaS、PaaS、アプリケーションのフルリンクk

エンドツーエンドの診断
10 種類以上の主流エージェント フレームワークをサポート

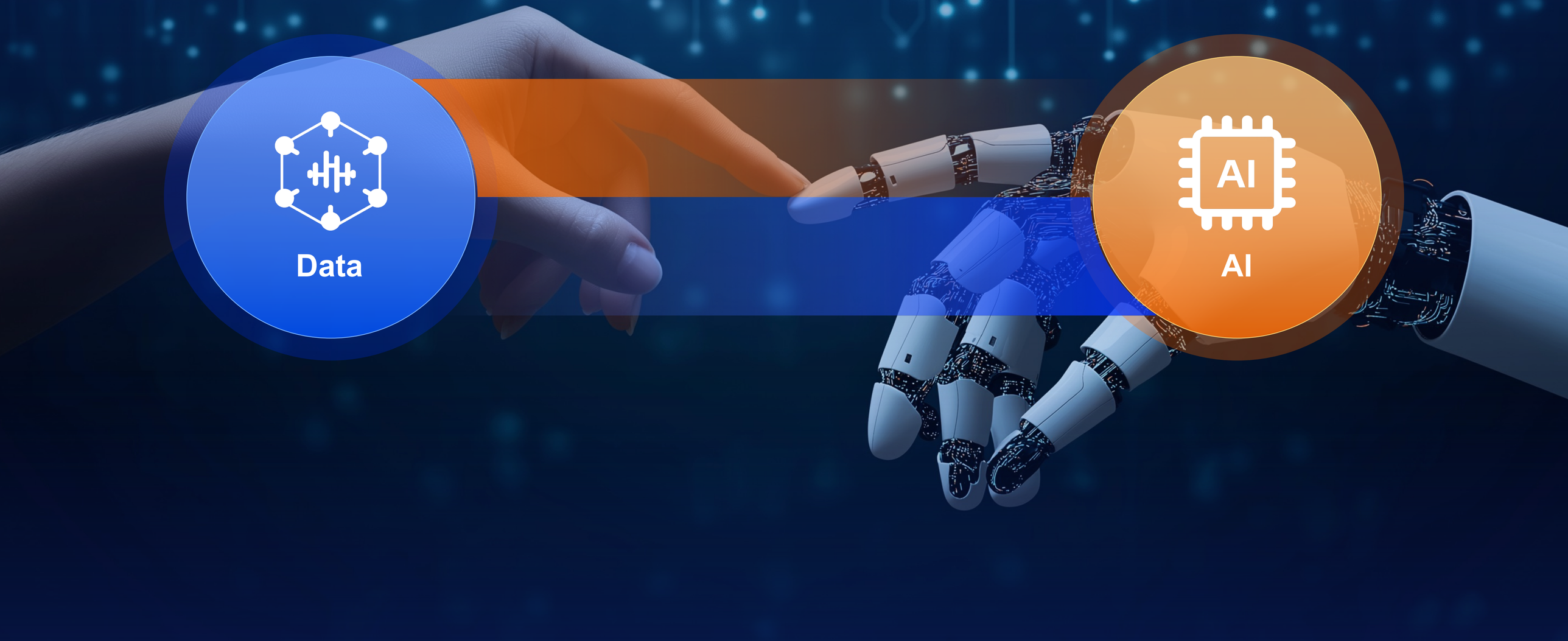
重要なアクセスデータ収集
コンテナ、PAI、その他のプラットフォームとデフォルトで統合

コスト管理とモデル品質評価
トークン、GPU使用コスト監視

モデルの入力と出力の品質評価
セマンティクス、出力品質、セキュリティリスクをカバーする

APSARA 云栖大会
APSARA CONFERENCE

Revolutionize Data Intelligence with AI



データ管理とデータ準備のための単一のプラットフォーム

統合データ管理

Enhanced

すぐに使用できるデータ エージェントと開発プラットフォーム、並列クエリ + SIMD と自動インデックスによる高速化によるパフォーマンスの向上

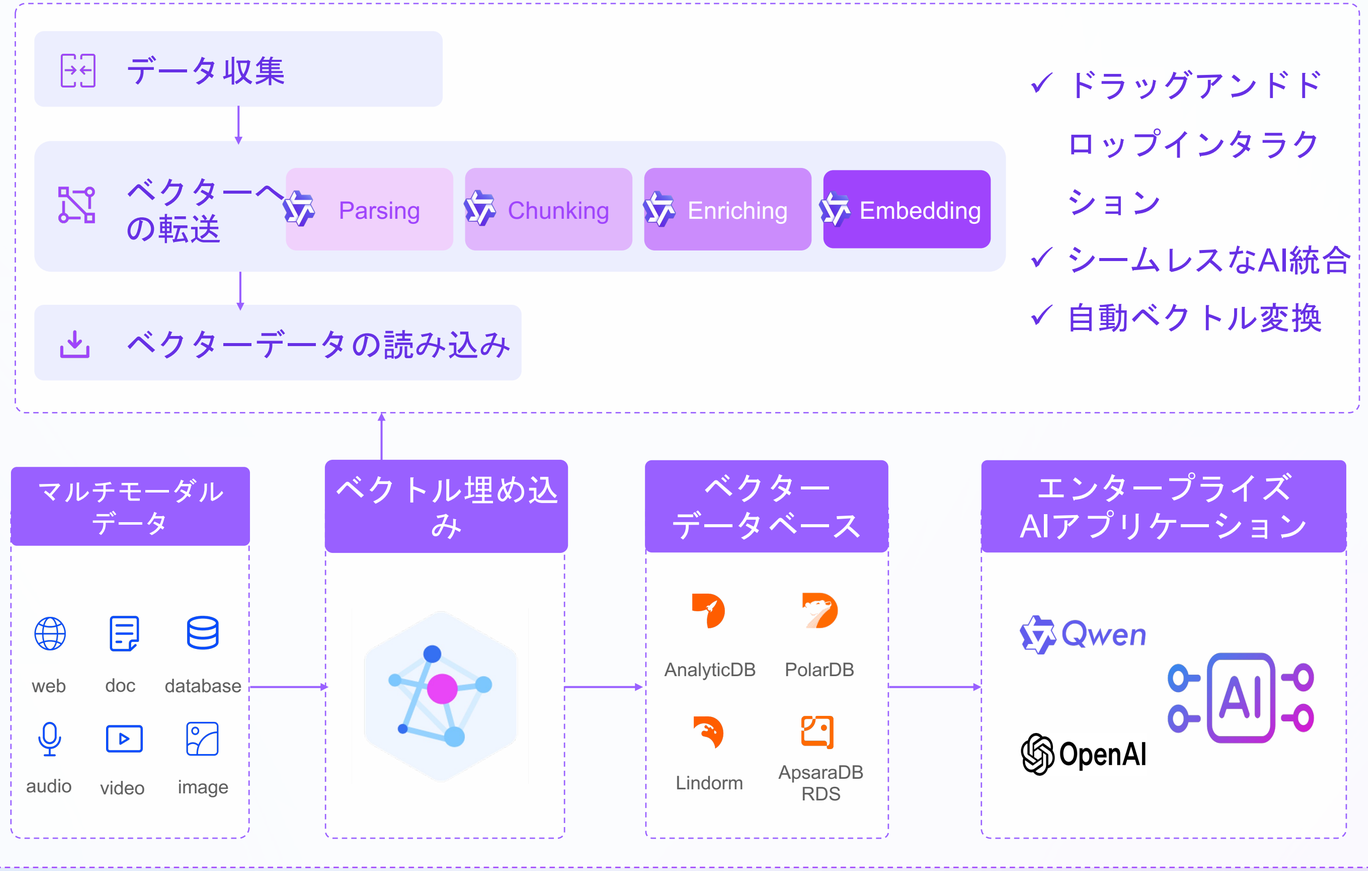
- 統合されたメタデータ
- 統合データガバナンス
- 統合データ運用
- データエージェント

データ管理

Alibaba Cloud DMS
(Data Management)

DTS One Channel For AI：マルチモーダルデータをAI対応ベクトルに変換

Enhanced

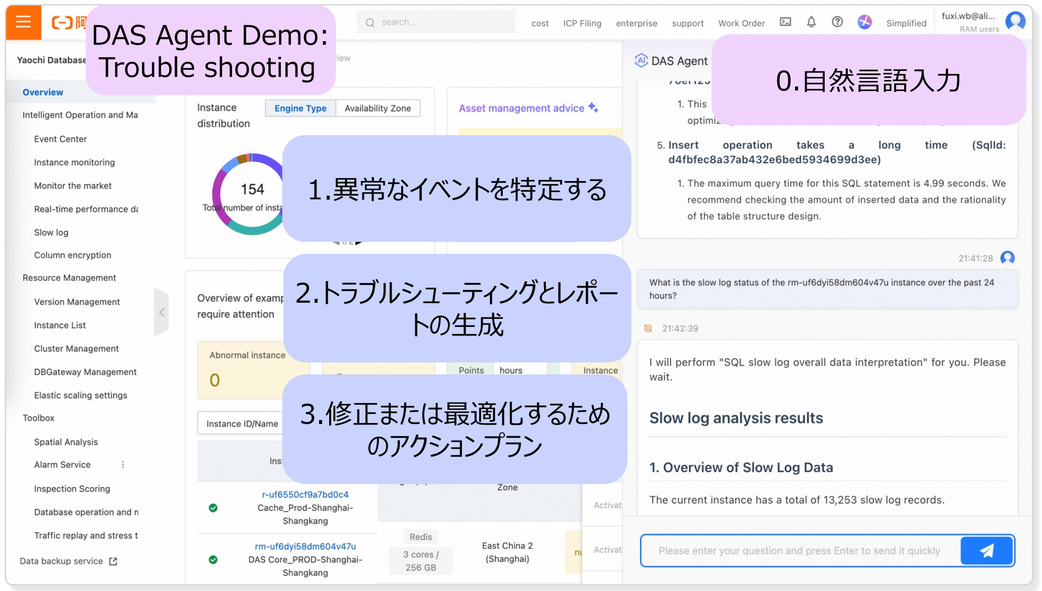


APSARA 云栖大会
APSARA CONFERENCE

ApsaraDB エージェント - NEW

DAS エージェント
データベース運用エージェント

New

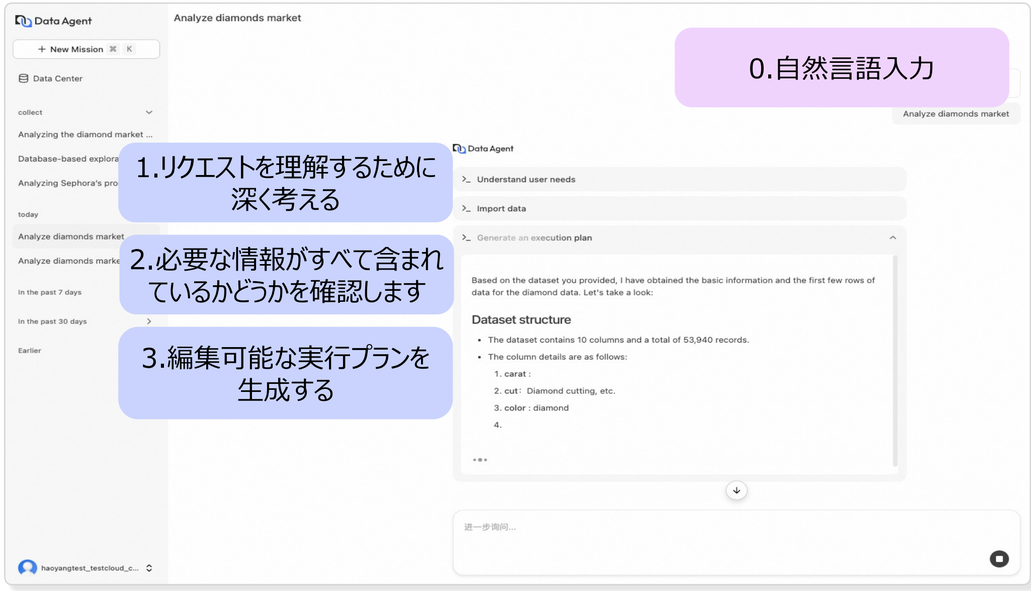


自動トラブルシューティング
30s
レポートと提案を生成する

正確率
92%+
専門家（85%）よりも優れており、通常のLLMよりも33%優れています

DMS データエージェント
データインサイトエージェント

New



データ統合
マルチモーダルデータ

データ保護
エンタープライズ専用環境

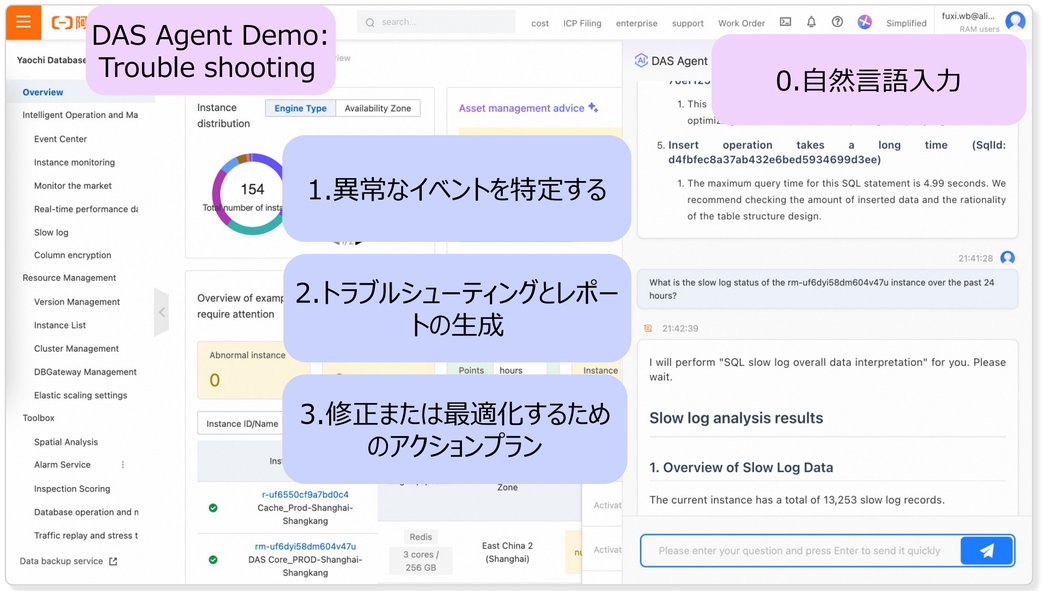
クエリの精度
複数ターンの会話
深い探求と思考

APSARA 云栖大会
APSARA CONFERENCE

ApsaraDB エージェント - NEW

DAS エージェント
データベース運用エージェント

New

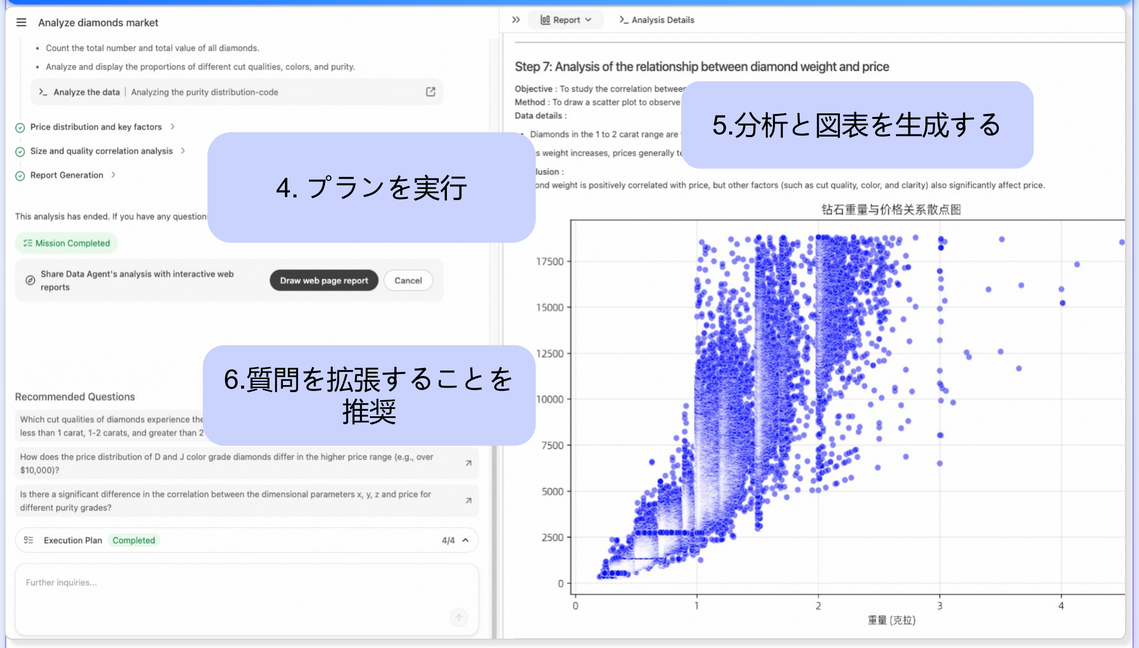


自動トラブルシューティング
30s
レポートと提案を生成する

正確率
92%+
専門家（85%）よりも優れており、通常のLLMよりも33%優れています

DMS データエージェント
データインサイトエージェント

New



データ統合
マルチモーダルデータ

データ保護
エンタープライズ専用環境

クエリの精度
複数ターンの会話深い探求と思考

APSARA 云栖大会
APSARA CONFERENCE

Simplify Cloud Operations with AI

APSARA 云栖大会
APSARA CONFERENCE

クラウド製品 AI アシスタント: オープンで柔軟なクラウド エクスペリエンス

ECS AIアシスタント：Q&Aからインテリジェントな運用まで

Q&A — Enhanced

- 製品ガイダンス
- トラブルシューティングのアドバイス

インスタンスの選択 — New

- シナリオベースのインスタンス推奨
- リアルタイムの価格と在庫の統合

オペレーション — New

- リソース操作: 開始/停止、パスワードのリセットなど。
- システムメンテナンス

診断 — New

- 150以上の診断と自動修復プラン

監視と分析 — New

- 視覚的な監視 + 異常の洞察

OSS AIアシスタント: AIワークロードのためのよりスマートなストレージ管理

問い合わせサポート、エラー診断、購入ガイダンス

会話形式のQ&A、コード例、ツールの最適化

エラー診断 — Coming soon

エラー入力 → 根本原因診断 + アクション

購入ガイド — Coming soon

請求分析 + リソースプランの推奨事項

ネットワーク AI アシスタント: ツールセットから回答セットへ

Q&A — New

- モデル: ネットワークRAG
- データ: MCP + OpenAPI

パス分析 — New

接続の問題を診断 → 実用的な修正

故障診断 — New

パケットロス解析 → ルートトラフィック + 解決

APSARA 云栖大会
APSARA CONFERENCE

クロスプロダクトクラウドAIOpsエージェント

クラウド監視CMS

アプリケーションの観察

- アプリケーション監視 APM
- ユーザーエクスペリエンスモニタリング
- クラウドダイヤルテスト
- アラームセンター

インテリジェントな運用・保守アシスタント

観察可能なインテリジェントアシスタント

- ログ分析
- インテリジェントな番号チェック
- コールチェーン分析
- トレンド予測

インテリジェントアラーム

根本原因の洞察

AIアプリケーションのフルスタック観察

- AIアプリケーションオブザーバブル
- AIゲートウェイインサイト
- PAIインサイト
- IaaS監視

クラウド製品の洞察 CloudLens

- OSSインサイト
- コンテナインサイト
- データベース観測可能
- PAIインサイト
- ECSインサイト

ログサービス SLS

観測可能 データ処理	データ処理	データストレージ	統合データモデル	
観測可能 データストレージ	SLS ログストア	SLSメトリックストア	SLS イベントストア	SLS エンティティストア

AIOpsエージェント New

自然言語リクエスト：
「なぜネットワークが遅いのですか？」

フルスタックへの高速アクセス

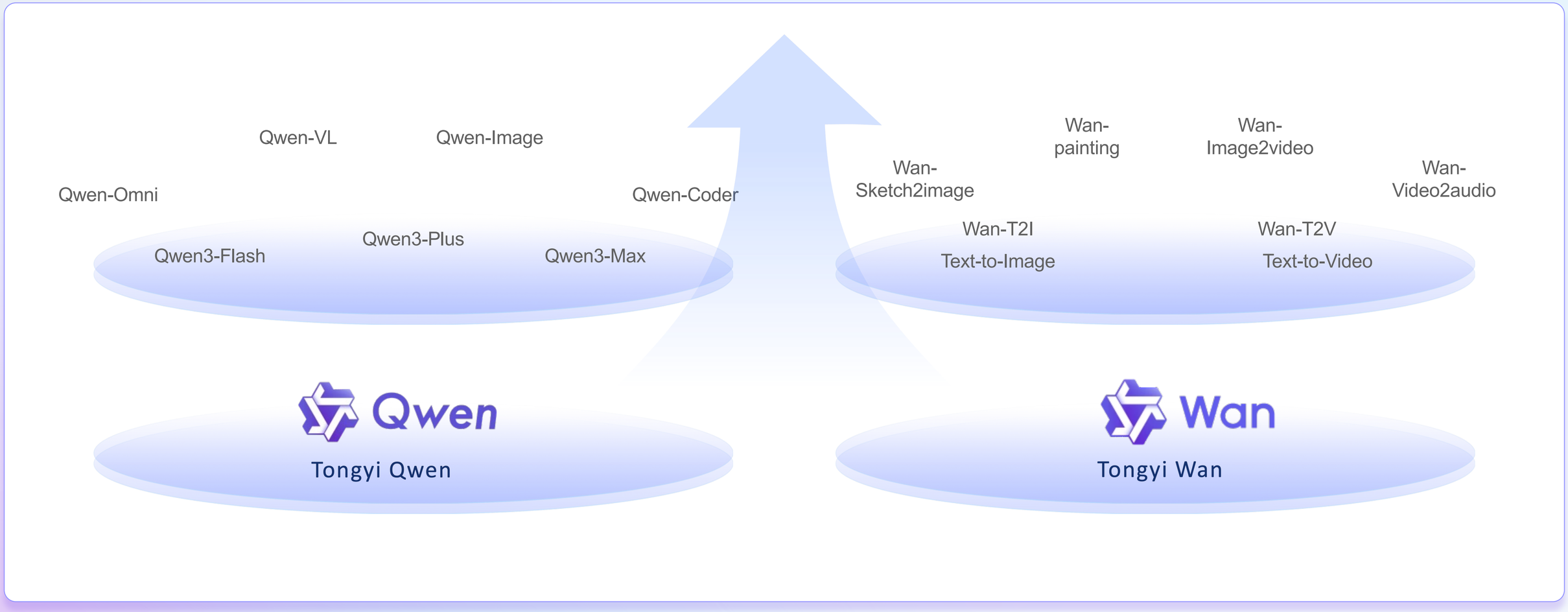
200以上のクラウド製品、
オープンソースコンポーネントをカバー
多言語サポート

統合データストレージおよび処理プラットフォーム

データ統合ストレージがデータサイロを解消｜
効率的なデータ処理がデータ価値を向上

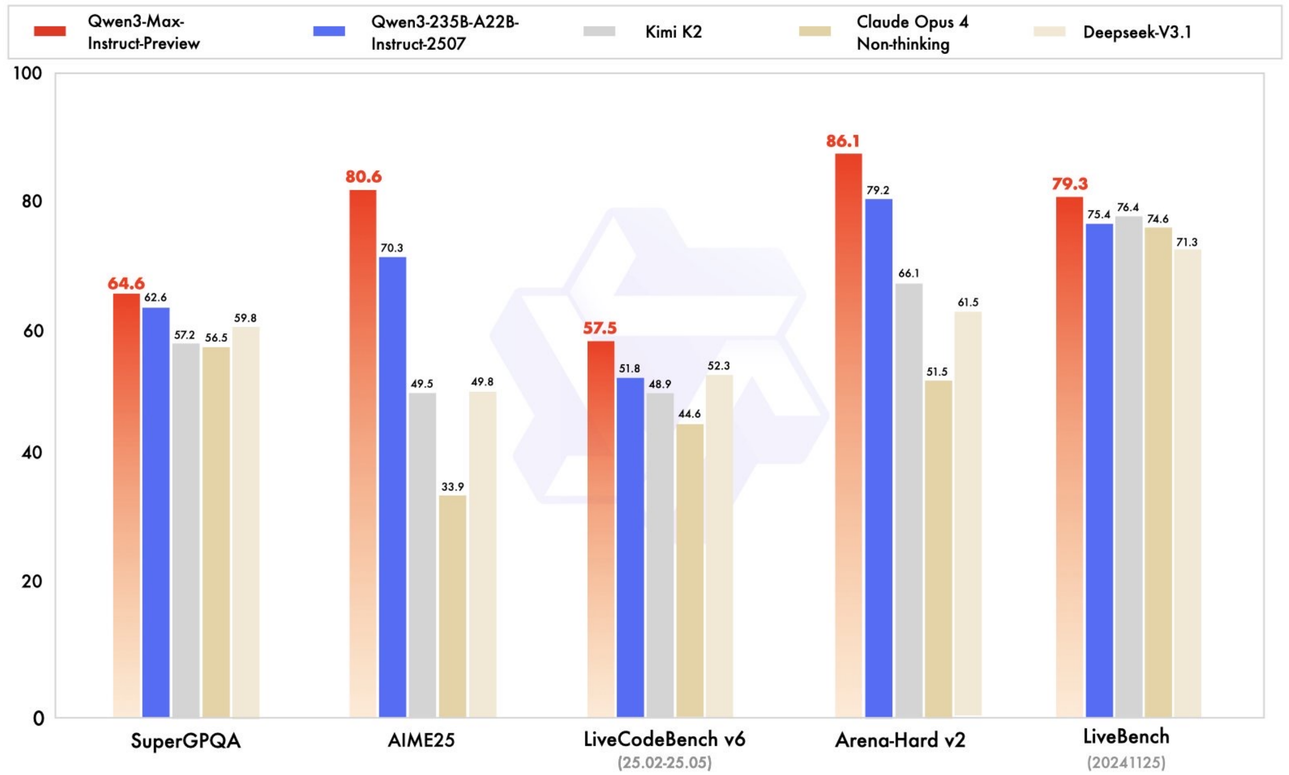
MaaS：TONGYI Model Family

Varied Sizes, Multi-Modalities, Open Source



Qwen3 - Max

Qwen3は、1兆を超えるパラメータを持ち、包括的なパフォーマンス改善を備えた最初のモデルです。
Qwen3-Max, the first model with over one trillion parameters, achieves comprehensive performance improvement.



Highlight

- **Comprehensive leap in general capabilities:** Based on the Qwen3 series, the overall general capabilities of the Qwen3-Max model have been significantly enhanced compared to Qwen2.5-Max, with notably superior performance in Chinese and English general text understanding and other related abilities.
- **総合的な能力の向上：**Qwen3 シリーズをベースにしたこのモデルの全体的な一般機能は Qwen2.5-Max と比較して大幅に向上しており、中国語と英語の一般テキスト理解機能が大幅に向上しています。
- **Significantly improved reasoning ability:** Achieving higher accuracy in tasks requiring reasoning, such as mathematics, coding, logical reasoning, and science subjects.
- **推論能力が大幅に向上：**数学、コーディング、論理的推論、科学など、推論を必要とするタスクにおける高い精度
- **Complex instruction following:** Better meeting users' specific task requirements and response format needs.
- **複雑な指示に従う：**ユーザーの特定のタスク要件、応答形式などをより適切に満たすことができます。
- **Broader Knowledge Coverage:** Better Mastery of Long-tail Knowledge, Reduced Hallucinations, More Reliable and Credible Responses.
- **より広い知識範囲：**ロングテール知識の理解が深まり、錯覚が減り、より信頼性の高い回答が得られます

Qwen3 - Next

基本モデルの将来における2つの大きなトレンド
Two Major Trends in the Future of Foundation Models

コンテキスト拡張 Context-Length Scaling

総パラメータ拡張 Total Parameter Scaling

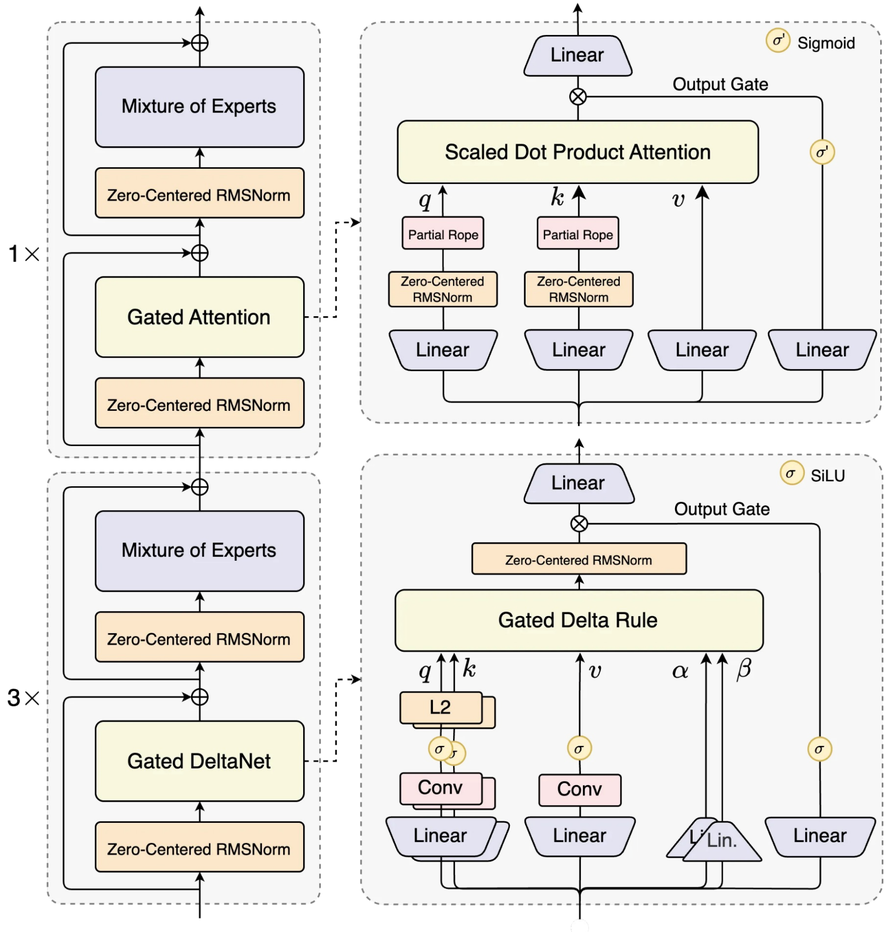
Qwen3-Next

超高スパース比
Ultra-high
sparsity ratio

活性化された専門家の割合1:16（Qwen） → 1:50（Qwen-next）

Activation Rate of Experts（Qwen） 1:50（Qwen-next）

マルチトークン予測メカニズム[3] Multi- Token Prediction
トレーニングの安定性に配慮した設計



[1] Gated Delta Networks: Improving Mamba2 with Delta Rule
[2] Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free
[3] Better & faster large language models via multi-token prediction

Qwen3 VL

Qwen VL 次世代マルチモーダルモデル

Comprehensive scale: From small to large, from Dense to MoE, from Instruct to Thinking, fully covered

包括的なスケール：小規模から大規模、高密度から低密度まで、指示から思考まで

Long sequences: Native 256K support, scalable up to 1M sequence length

長いシーケンス: ネイティブ 256K、1M まで拡張可能な長いシーケンスのサポート

Richer and more comprehensive visual world knowledge, with enhanced object recognition capabilities

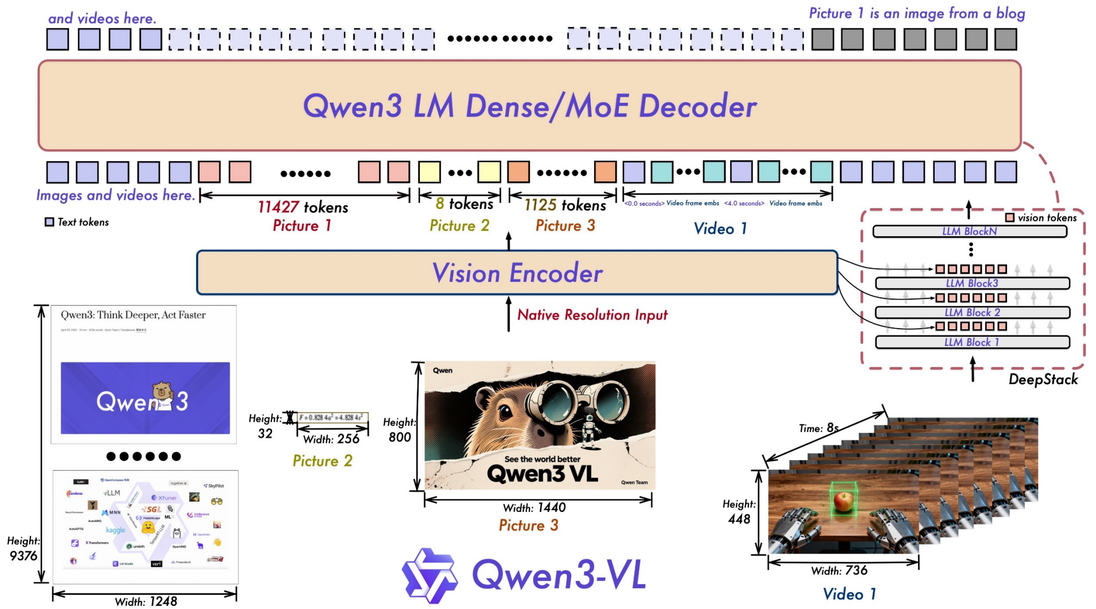
視覚世界に関するより豊かで包括的な知識と、あらゆるものを識別する能力の向上

Enhanced spatial awareness and stronger 2D/3D grounding capabilities

空間認識を改善し、2D/3Dグラウンディング機能を強化します

Superior video understanding capabilities, including long videos and spatiotemporal localization

長時間動画や時空間位置の理解を含む、より強力な動画理解機能

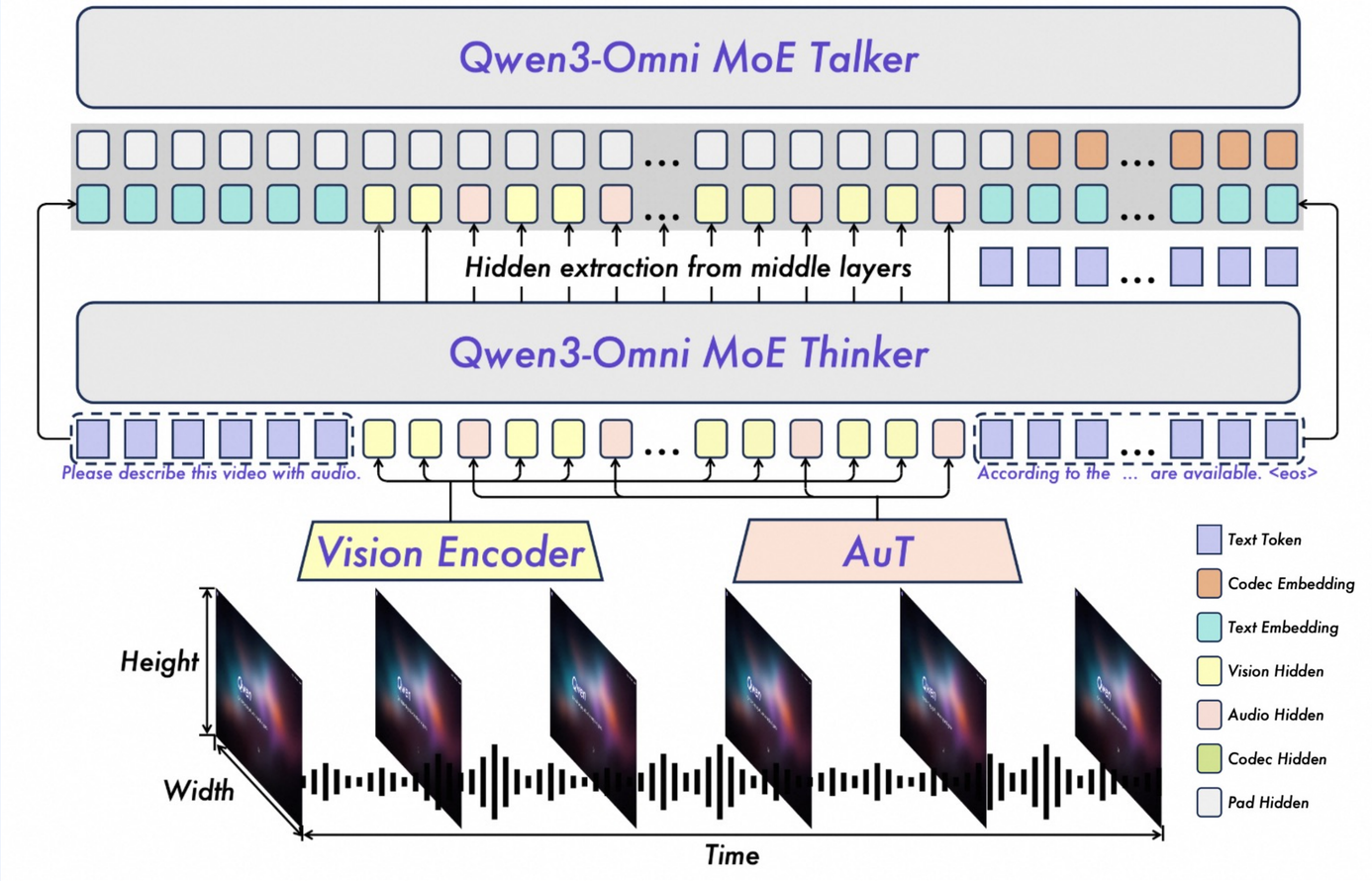


Qwen3 - Omni

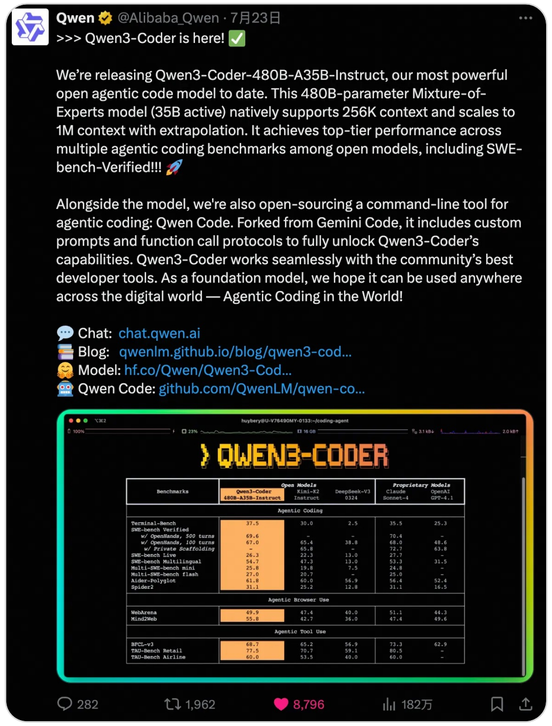
エンドツーエンドのマルチモーダルモデルは、マルチモーダルインタラクションエクスペリエンスを強化します

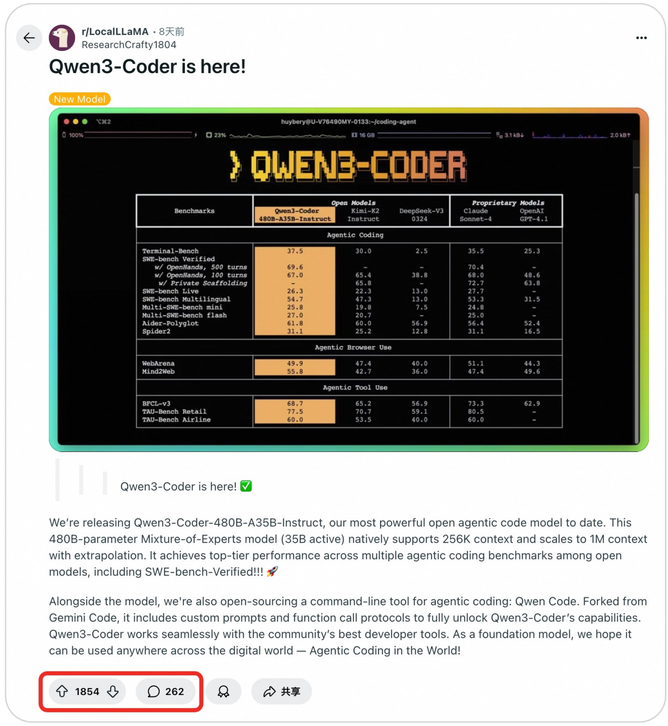
Highlight

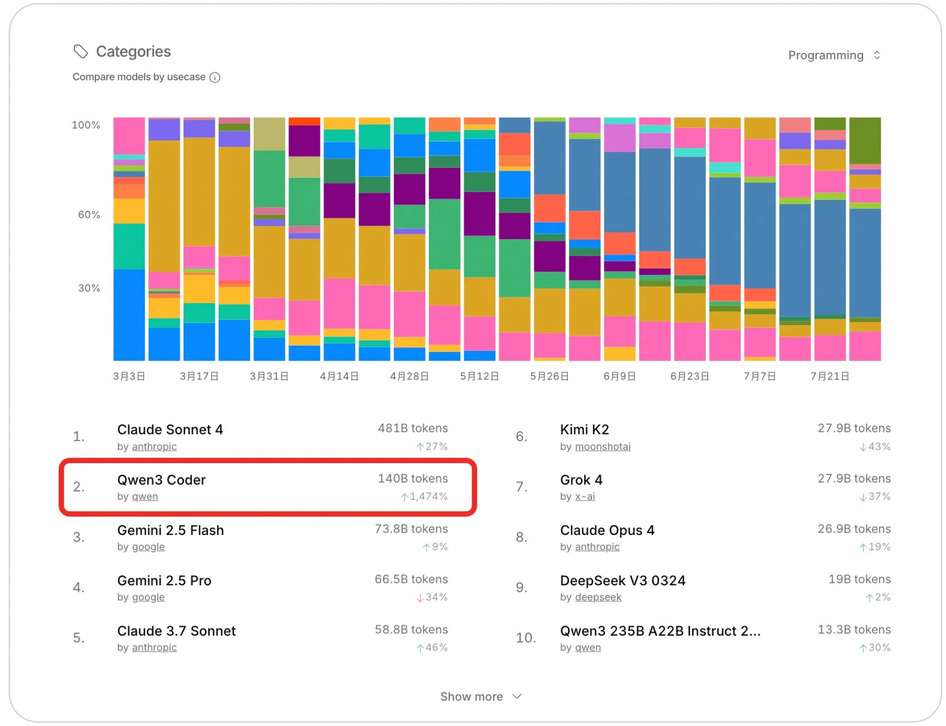
- **モデルアーキテクチャ：**Qwen3-OmniはThinker-Talkerアーキテクチャを採用しています。Thinkerはテキスト生成を担い、TalkerはThinkerから高レベルの意味表現を直接受け取り、ストリーミング音声トークンの生成に集中します。超低遅延ストリーミングを実現するために、Talkerはマルチコードブックシーケンスの自己回帰予測を使用します。各デコードステップにおいて、MTPモジュールは現在のフレームの残差コードブックを出力し、Code2Wavによって対応する波形に合成することで、フレームごとのストリーミングを可能にします。
- **Model Architecture:** Qwen3-Omni adopts the Thinker-Talker architecture. The Thinker is responsible for text generation, while the Talker focuses on generating streaming speech tokens by directly receiving high-level semantic representations from the Thinker.
- **AuT：**オーディオ エンコーダーは、2億時間のオーディオ データでトレーニングされた AuT モデルを使用し、非常に強力な一般的なオーディオ表現機能を備えています。
- **AuT：** The audio encoder adopts the AuT model trained on 200 million hours of audio data。
- **MoE：**Thinker と Talker はどちらも MoE アーキテクチャを採用しており、高い並行性と高速な推論をサポートします。
- **MoE:** Both Thinker and Talker adopt the MoE architecture, supporting high concurrency and fast inference.



Qwen3 - Coder


Qwen @Alibaba_Qwen · 7月23日
>>> Qwen3-Coder is here! ✅
We're releasing Qwen3-Coder-480B-A35B-Instruct, our most powerful open agentic code model to date. This 480B-parameter Mixture-of-Experts model (35B active) natively supports 256K context and scales to 1M context with extrapolation. It achieves top-tier performance across multiple agentic coding benchmarks among open models, including SWE-bench-Verified!!! 🚀
Alongside the model, we're also open-sourcing a command-line tool for agentic coding: Qwen Code. Forked from Gemini Code, it includes custom prompts and function call protocols to fully unlock Qwen3-Coder's capabilities. Qwen3-Coder works seamlessly with the community's best developer tools. As a foundation model, we hope it can be used anywhere across the digital world — Agentic Coding in the World!
💬 Chat: chat.qwen.ai
📝 Blog: qwenlm.github.io/blog/qwen3-cod...
🤗 Model: hf.co/Qwen/Qwen3-Cod...
🤖 Qwen Code: github.com/QwenLM/qwen-co...
QWEN3-CODER
282
1,962
8,796
182万


r/LocalLLaMA · 8天前
ResearchCrafty1804
Qwen3-Coder is here!
New Model
QWEN3-CODER
Qwen3-Coder is here! ✅
We're releasing Qwen3-Coder-480B-A35B-Instruct, our most powerful open agentic code model to date. This 480B-parameter Mixture-of-Experts model (35B active) natively supports 256K context and scales to 1M context with extrapolation. It achieves top-tier performance across multiple agentic coding benchmarks among open models, including SWE-bench-Verified!!! 🚀
Alongside the model, we're also open-sourcing a command-line tool for agentic coding: Qwen Code. Forked from Gemini Code, it includes custom prompts and function call protocols to fully unlock Qwen3-Coder's capabilities. Qwen3-Coder works seamlessly with the community's best developer tools. As a foundation model, we hope it can be used anywhere across the digital world — Agentic Coding in the World!
1854
262
共享


Categories
Programming
Compare models by usecase
100%
60%
30%
3月3日
3月17日
3月31日
4月14日
4月28日
5月12日
5月26日
6月9日
6月23日
7月7日
7月21日
1. Claude Sonnet 4 by anthropic 481B tokens ↑27%
2. Qwen3 Coder by qwen 140B tokens ↑1,474%
3. Gemini 2.5 Flash by google 73.8B tokens ↑9%
4. Gemini 2.5 Pro by google 66.5B tokens ↓34%
5. Claude 3.7 Sonnet by anthropic 58.8B tokens ↑46%
6. Kimi K2 by moonshotai 27.9B tokens ↓43%
7. Grok 4 by x-ai 27.9B tokens ↓37%
8. Claude Opus 4 by anthropic 26.9B tokens ↑19%
9. DeepSeek V3 0324 by deepseek 19B tokens ↑2%
10. Qwen3 235B A22B Instruct 2... by qwen 13.3B tokens ↑30%
Show more

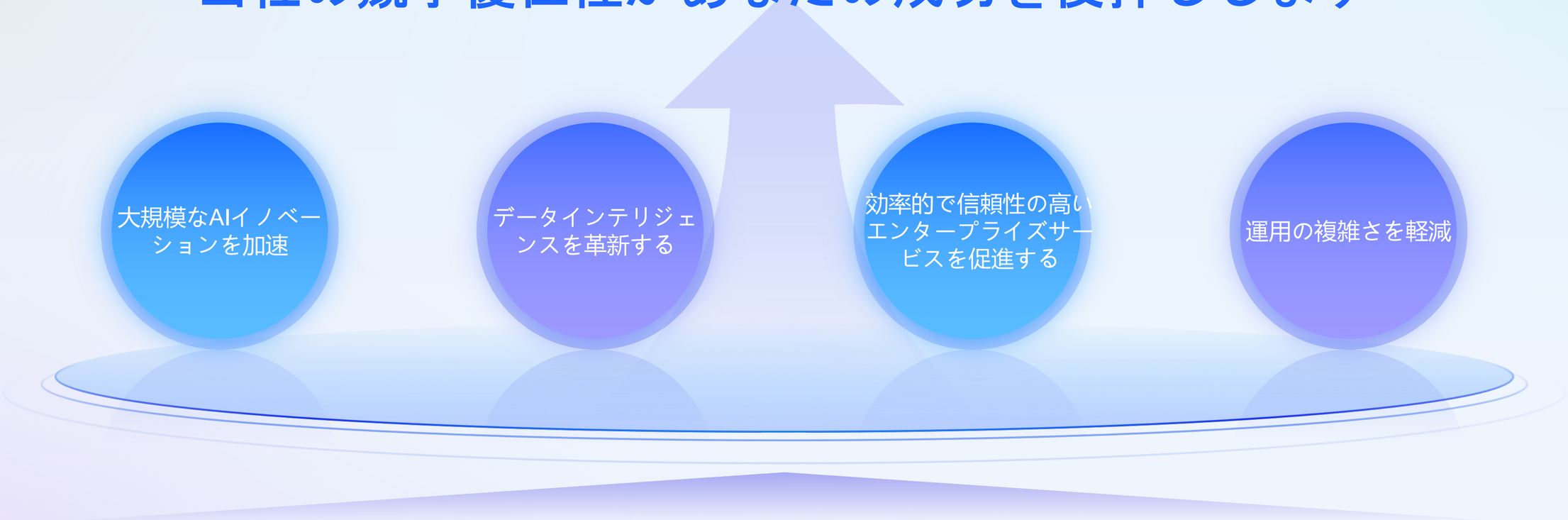
https://chat.qwen.ai/

https://huggingface.co/Qwen

https://modelscope.cn/organization/qwen

https://github.com/QwenLM/Qwen3-VL

当社の競争優位性があなたの成功を後押しします



スタック全体にわたるテクノロジーリーダーシップ

AP最大規模の統合サービス能力

よりコスト効率が高く、より使いやすい

ご清聴頂きありがとうございます