Qwen3.7 Max Thinking d'Alibaba : le champion open-source Apache 2.0 — benchmark juillet 2026

Qwen3.7 Max Thinking d'Alibaba Cloud est en juillet 2026 la meilleure réponse à la question que se posent de plus en plus d'organisations : "Comment obtenir des performances proches du top 5 mondial, en Apache 2.0, à moins de $1,25/M tokens ?" Avec un ELO LM Arena de 1475 (11ème mondial sur 360+ modèles), un GPQA Diamond de 92,4%, un SWE-Bench de 80,4% et une vitesse de génération de 197 tokens/seconde, Qwen3.7 Max Thinking est le modèle le plus puissant disponible sous licence Apache 2.0 en juillet 2026 — permettant un usage commercial libre, un déploiement on-premise et un fine-tuning sans restriction. Ce guide analyse l'architecture MoE, les benchmarks complets, la performance en français, et les cas d'usage où Qwen3.7 Max Thinking est le meilleur choix.

La famille Qwen3 — positionnement et gamme

Alibaba Cloud a structuré sa gamme Qwen3 en plusieurs tiers pour couvrir tous les usages, de l'inférence ultra-rapide aux tâches de raisonnement complexe :

Modèle	Params (actifs)	LM Arena ELO	GPQA	Vitesse	Prix in /M	Licence
Qwen3.7 Max Thinking	235B (22B actifs)	1475	92,4%	197 tok/s	$1,25	Apache 2.0
Qwen3.6 Max Preview	—	1446	—	—	$7,80	Apache 2.0
Qwen3.7 Plus	72B (18B actifs)	~1420	~88%	~280 tok/s	$0,42	Apache 2.0
Qwen3.7 Turbo	—	~1380	—	~450 tok/s	$0,12	Apache 2.0
Qwen2.5 Max (2025)	—	~1380	~82%	—	—	

Le nom "Thinking" indique la variante avec raisonnement étendu activé par défaut (Chain-of-Thought interne masqué, similaire aux variantes "Thinking" de Claude ou au mode "reasoning" de GPT-5.x). Cette variante obtient des scores significativement supérieurs sur les benchmarks de raisonnement au prix d'une latence plus élevée.

Architecture MoE — 235 milliards de paramètres, 22 milliards actifs

Qwen3.7 Max Thinking repose sur une architecture **Mixture of Experts (MoE)**, ce qui lui confère un avantage décisif en matière d'efficience computationnelle :

Paramètres totaux : 235 milliards

Paramètres actifs par inférence : 22 milliards (seulement 9,4% des paramètres sont activés pour chaque token)

Experts par couche : 64 experts spécialisés, 4 actifs par token

Architecture de base : Transformer avec attention multi-head groupée (GQA), RoPE positional encodings, SwiGLU activation

Fenêtre de contexte : 256 000 tokens (256K) — inférieure à Claude/GPT-5/Kimi K3 (1M), limitation à prendre en compte pour les cas d'usage nécessitant de très longs contextes

Entraînement : plus de 20 trillions de tokens incluant une proportion importante de textes en chinois, anglais, arabe, français, coréen et japonais

L'architecture MoE permet à Qwen3.7 Max Thinking d'atteindre des performances de modèle dense à 100B+ paramètres tout en maintenant le coût d'inférence au niveau d'un modèle de 22B — d'où l'excellent rapport qualité/prix.

Benchmarks officiels — performances complètes

LM Arena

Catégorie	ELO Qwen3.7 Max	Rang	Leader
Global	1475	11ème	Claude Fable 5 (1507)
Coding / WebDev	~1450	~10ème	Kimi K3 (1682)
Raisonnement STEM	~1490	~7ème	Claude Opus 4.7 Thinking

Benchmarks académiques — comparatif détaillé

Benchmark	Qwen3.7 Max	Claude Fable 5	GPT-5.6 Sol	Kimi K3	Muse Spark 1.1
GPQA Diamond	92,4%	~92%	94,6%	93,5%	89,5%
SWE-Bench Verified	80,4%	~85%	—	—	77,4%
AIME 2024	~88%	—	—	—	—
MATH (olympiades)	~90%	~87%	—	—	~83%
HumanEval (code)	~91%	~92%	—	—	~88%
MMLU Pro	~85%	~88%	—	—	~83%
LiveCodeBench	~75%	—	—	—	—

Résultat surprenant : Qwen3.7 Max Thinking **surpasse Claude Fable 5 sur les mathématiques** (MATH ~90% vs ~87%). Alibaba a clairement priorisé les compétences STEM dans son data mixture, ce qui se confirme dans les scores AIME et MATH. Pour les applications de raisonnement scientifique et mathématique en budget contraint, Qwen3.7 Max Thinking est un concurrent sérieux des modèles propriétaires deux à quatre fois plus coûteux.

Performance en langue française

Qwen3.7 Max Thinking obtient un score de **8,5/10 en moyenne** pour le français — position intermédiaire entre Muse Spark 1.1 (8,4/10) et Kimi K3 (8,9/10), mais significativement en deçà de Claude Fable 5 (9,8/10) ou Gemini 3.1 Pro (9,4/10).

Critère	Score / 10	Points forts	Points faibles
Grammaire & orthographe	8,7	Règles de base solides, bons accords	Subjonctifs plus-que-parfaits rares
Richesse lexicale	8,8	Vocabulaire technique excellent (IT, science)	Registre familier limité
Maîtrise des registres	8,3	Bon en registre formel/professionnel	Registre littéraire avec "patine traduit"
Nuances culturelles	8,1	Références culturelles générales	Humour français, sous-entendus culturels
Fluidité et naturel	8,5	Prose technique fluide	Légère lourdeur dans la prose littéraire
Moyenne	**8,5/10**	Fort sur le français technique, moyen sur le littéraire	

Pour les organisations francophones ayant besoin d'un LLM en Apache 2.0, Qwen3.7 Max Thinking offre le meilleur niveau de français parmi les options entièrement libres. En comparaison, Qwen2.5 Max (2025) atteignait seulement 7,4/10 — progression notable due à l'augmentation des données françaises dans le corpus d'entraînement Qwen3.

Artificial Analysis — vitesse, coût et intelligence

Qwen3.7 Max Thinking se distingue par un excellent équilibre vitesse/coût/intelligence selon les mesures d'Artificial Analysis :

Indicateur	Qwen3.7 Max Thinking	Comparatif
Score d'intelligence	46/100	Claude Fable 5 : 61 · GPT-5.6 Sol : 59
Vitesse de génération	197 tok/s	Claude Fable 5 : ~55 · Gemini 3.6 Flash : 237
Coût par tâche standard	$1,03	Claude Fable 5 : $2,03 · GPT-5.6 Sol : $1,04
Latence TTFT	~0,8s	Bonne pour un modèle de cette taille
Prix in / out /M	$1,25 / $5	Claude Fable 5 : $10 / $50 (8x plus cher)

À 197 tok/s et $1,03/tâche, Qwen3.7 Max Thinking offre le meilleur rapport intelligence/coût parmi les modèles avec GPQA >90%. Pour les pipelines qui nécessitent du raisonnement avancé à grand volume (>100 000 requêtes/jour), l'économie vs Claude Fable 5 peut atteindre 95% ($1,03 vs $2,03/tâche × volume).

Apache 2.0 — pourquoi la licence change tout

Qwen3.7 Max Thinking est distribué sous **licence Apache 2.0**, la licence open-source la plus permissive du marché pour un modèle de ce niveau. Concrètement, Apache 2.0 signifie :

Usage commercial libre : pas de restriction sur la commercialisation des applications construites avec le modèle.

Modification libre : fine-tuning, distillation, création de modèles dérivés sans obligation de publier les modifications.

Redistribution libre : inclure les poids dans votre produit, votre image Docker, votre package npm.

Pas de royalties : contrairement à certaines licences "open" avec clause de partage des bénéfices.

Pas de limite d'utilisateurs : contrairement à la licence Llama 5 (700M MAU limit) ou Llama 4 (600M MAU limit).

Cette liberté totale est particulièrement précieuse pour :

Les éditeurs de logiciels qui veulent intégrer un LLM dans un produit packagé sans risque légal.

Les entreprises qui souhaitent fine-tuner sur leurs données propriétaires et garder le modèle résultant confidentiel.

Les startups qui veulent démarrer avec l'API publique (Alibaba Cloud, Together AI, Fireworks AI) et migrer vers un déploiement propre si les coûts augmentent.

Les institutions académiques et de recherche qui ont besoin de la liberté totale de modification et publication.

Déploiement on-premise — guide pratique

Les poids de Qwen3.7 Max Thinking sont disponibles sur Hugging Face sous le nom `Qwen/Qwen3.7-Max-Thinking`. Options de déploiement :

Méthode	Configuration	Vitesse	Usage
vLLM (recommandé prod)	4x H100 80GB	~200 tok/s	Production haute charge
vLLM + AWQ	4x A100 80GB	~150 tok/s	Production standard
Ollama (locale)	4x A100 40GB + INT4	~50 tok/s	Équipe dev, usage interne
HuggingFace TGI	8x A100 80GB FP16	~250 tok/s	Très haute charge
Together AI (cloud)	API tierce	~197 tok/s	Pas d'infra à gérer

La quantification AWQ (Activation-aware Weight Quantization) permet de réduire les besoins VRAM de 40% avec moins de 1% de dégradation sur les benchmarks — recommandée pour les déploiements sur A100 40GB.

Cas d'usage où Qwen3.7 Max Thinking excelle

Applications STEM et mathématiques : avec MATH ~90% (supérieur à Claude Fable 5), Qwen3.7 est le meilleur modèle open-source Apache 2.0 pour les tuteurs IA en mathématiques, les assistants de calcul scientifique et les outils d'aide à la démonstration.

Génération de code à haute fréquence : 197 tok/s + HumanEval 91% = idéal pour les assistants IDE comme continuation de Copilot sur infrastructure propre.

Pipelines RAG à grand volume : rapport coût/performance optimal pour les architectures RAG traitant des millions de documents.

Applications multilingues en Asie : performances supérieures en chinois, japonais, coréen et arabe par rapport aux modèles occidentaux de même niveau.

Recherche et académique : Apache 2.0 + haute performance = première option pour les labs de recherche souhaitant publier des résultats sans contraintes légales.

Limitations de Qwen3.7 Max Thinking

Fenêtre de contexte 256K : limitation significative par rapport aux 1M tokens de Claude, GPT-5.6, Kimi K3. Pour les cas d'usage nécessitant d'ingérer de très longs documents ou codebases, Kimi K3 ou DeepSeek-V4-Pro-Max sont préférables.

Alignement de sécurité : certains rapports indiquent que Qwen3.7 est légèrement moins conservateur sur les contenus sensibles que Claude ou GPT-5.6 — avantage pour la recherche en sécurité, inconvénient pour les applications grand public non moderées.

Inférence locale coûteuse : malgré la MoE, déployer Qwen3.7 Max Thinking on-premise requiert au minimum 4 GPU A100/H100 — inaccessible pour les très petites équipes sans cloud GPU.

Registres français informels : score 8,5/10 correct mais inférieur aux modèles Anthropic et Google pour du contenu créatif en français soutenu.

À RETENIR

À retenir — Qwen3.7 Max Thinking

🏅 11ème LM Arena mondial (ELO 1475) — meilleur Apache 2.0 du classement

🧮 GPQA Diamond 92,4% — au niveau de Claude Fable 5, loin devant tous les autres open-source

💻 SWE-Bench 80,4% — résolution de bugs GitHub réels, excellent pour un modèle open

⚡ 197 tok/s — vitesse de génération 3,5x supérieure à Claude Fable 5

💰 $1,25/M tokens — 8x moins cher que Claude Fable 5 pour des performances proches

📜 Apache 2.0 — la licence la plus libre du marché pour ce niveau de performance

✅ Recommandé pour : STEM, code haute fréquence, déploiement souverain sans restriction, recherche

⚠️ Limite : contexte 256K (vs 1M pour Claude/GPT-5), français légèrement inférieur aux leaders

Multilingue — les performances de Qwen3.7 au-delà du français

Qwen3.7 Max Thinking a été entraîné avec un corpus multilingue exceptionnellement riche, héritage direct d'Alibaba Cloud qui opère des services dans des dizaines de langues à travers l'Asie. Voici les performances estimées par langue :

Langue	Score Qwen3.7	Claude Fable 5	GPT-5.6 Sol	Avantage Qwen
Anglais	9,7/10	9,8/10	9,8/10	Parité
Chinois (simplifié)	9,9/10	8,5/10	8,8/10	🏆 Large avantage
Japonais	9,4/10	8,8/10	9,0/10	Avantage Qwen
Coréen	9,3/10	8,7/10	8,9/10	Avantage Qwen
Arabe	9,2/10	8,6/10	8,7/10	Avantage Qwen
Français	8,5/10	9,8/10	9,2/10	Désavantage Qwen
Allemand	8,8/10	9,4/10	9,3/10	Désavantage Qwen
Espagnol	8,9/10	9,5/10	9,4/10	Désavantage Qwen

Pour les applications multilingues couvrant l'Asie-Pacifique, Qwen3.7 Max Thinking est clairement supérieur aux modèles occidentaux. Pour les applications franco-européennes uniquement, Claude Fable 5 reste le meilleur choix.

Intégration dans l'écosystème Alibaba Cloud

Qwen3.7 Max Thinking bénéficie d'une intégration native dans l'écosystème Alibaba Cloud (Aliyun), qui peut être attractif pour les entreprises déjà clientes :

DashScope API : accès unifié à toute la gamme Qwen, facturation consolidée avec les autres services Alibaba Cloud.

PAI (Platform for AI) : fine-tuning managé de Qwen3.7 sans gestion d'infrastructure, avec des templates préconfigurés pour les cas d'usage courants.

Model Studio : interface no-code pour construire des applications IA basées sur Qwen3.7, avec workflow visuel.

MaxCompute / DataWorks : intégration avec les pipelines de données Alibaba pour des scénarios analytics augmentés par LLM.

Quantification et optimisation pour déploiement économique

L'architecture MoE de Qwen3.7 Max Thinking offre des possibilités de quantification particulièrement avantageuses. Voici les configurations testées par la communauté :

Précision	VRAM requise	Vitesse	Dégradation GPQA	Recommandation
FP16 (référence)	~460 GB	197 tok/s	—	Cloud haute performance
BF16	~460 GB	200 tok/s	<0,1%	Équivalent FP16
INT8 (AWQ)	~230 GB	280 tok/s	~0,5%	✅ Production standard
INT4 (AWQ/GPTQ)	~115 GB	380 tok/s	~1,5%	Budget / déploiement edge
INT4 (experts actifs)	~30 GB	~150 tok/s	~3%	4x GPU 40GB consumer

La configuration INT8 (AWQ) sur 3-4 GPU H100 80GB est le meilleur compromis pour la production : 280 tok/s avec une dégradation GPQA négligeable (<0,5%). La configuration INT4 experts actifs (~30GB VRAM) ouvre la possibilité d'un déploiement sur 4 GPU gaming haute gamme ou 2 GPU A100 40GB — réduisant massivement le coût d'infrastructure on-premise.

Comparatif détaillé — Qwen3.7 vs DeepSeek-V4-Pro-Max

Les deux modèles sont souvent comparés car ils visent le même segment : performance top 10 mondiale, open-source, déploiement souverain. Voici une comparaison approfondie :

Critère	Qwen3.7 Max Thinking	DeepSeek-V4-Pro-Max	Verdict
LM Arena ELO	**1475**	1449	Qwen +26 points
GPQA Diamond	**92,4%**	90,1%	Qwen +2,3%
SWE-Bench	80,4%	**80,6%**	Quasi-parité
MATH olympiades	~90%	~88%	Qwen légèrement supérieur
Contexte	256K	**1M**	DeepSeek 4x plus long
Vitesse	**~197 tok/s**	~150 tok/s	Qwen +31%
Prix API in /M	$1,25	$1,60	Qwen moins cher
Licence	Apache 2.0	**MIT**	MIT légèrement plus libre
Chinois / Asiatique	**9,9/10**	9,5/10	Qwen supérieur
Français	8,5/10	7,8/10	Qwen supérieur

Verdict : Qwen3.7 Max Thinking est supérieur sur la plupart des critères. DeepSeek-V4-Pro-Max est préférable si vous avez besoin d'un contexte de 1 million de tokens ou d'une licence MIT strictement sans restriction. Pour tous les autres cas, Qwen3.7 Max Thinking est le meilleur choix Apache 2.0 en juillet 2026.

Perspectives — Qwen 4 et la roadmap Alibaba

Alibaba n'a pas encore communiqué de date pour la prochaine génération Qwen. Sur la base de la cadence historique (Qwen1→2 : 8 mois, Qwen2→2.5 : 6 mois, Qwen2.5→Qwen3 : 8 mois), Qwen4 serait attendu au 1er trimestre 2027. Les améliorations attendues selon les annonces partielles d'Alibaba Research :

Extension du contexte à 1 million de tokens (rattrapage des concurrents).

Architecture hybride MoE-dense améliorée pour réduire encore la VRAM requise.

Capacités vidéo natives (Qwen-VL étendu).

Agents plus autonomes avec planification et mémorisation multi-sessions.

En attendant, Qwen3.7 Max Thinking reste en juillet 2026 le meilleur rapport performance/ licence/prix du marché pour les organisations souhaitant maintenir leur souveraineté numérique. Pour une comparaison avec l'ensemble des modèles disponibles, consultez notre Benchmark IA juillet 2026 — classement LM Arena, BenchLM & performances en français.

FAQ — Qwen3.7 Max Thinking

Quelle est la différence entre Qwen3.7 Max et Qwen3.7 Max Thinking ?

Qwen3.7 Max est la variante de base avec réponses directes. Qwen3.7 Max Thinking active un mode de raisonnement étendu (Chain-of-Thought interne) qui génère une réflexion masquée avant de produire la réponse finale — similaire au mode "extended thinking" de Claude Opus 4.7 ou au "reasoning" de GPT-5.x. La variante Thinking obtient des scores ~3-5 points supérieurs sur GPQA et MATH au prix d'une latence 2-3x plus élevée et d'un coût proportionnellement plus élevé. Pour les tâches complexes (raisonnement scientifique, débogage avancé, mathématiques), Thinking est recommandé. Pour les tâches simples à haute fréquence, Max sans Thinking est plus économique.

Comment comparer Qwen3.7 Max Thinking avec DeepSeek-V4-Pro-Max ?

Les deux sont d'excellents modèles open-source sous licence permissive (Apache 2.0 vs MIT). Qwen3.7 Max Thinking est supérieur sur LM Arena (ELO 1475 vs 1449), GPQA (92,4% vs 90,1%) et vitesse (197 vs ~150 tok/s). DeepSeek-V4-Pro-Max est légèrement supérieur sur SWE-Bench (80,6% vs 80,4%), contexte identique (1M tokens) et licence MIT (encore plus libre qu'Apache 2.0 sur certains points). Pour la plupart des usages, Qwen3.7 Max Thinking est le meilleur choix ; pour les applications de résolution de bugs SWE ou les usages nécessitant un contexte de 1M tokens, DeepSeek-V4-Pro-Max est une alternative sérieuse.

Peut-on fine-tuner Qwen3.7 Max Thinking sur des données propriétaires ?

Oui, c'est l'un des avantages clés de la licence Apache 2.0. Le fine-tuning de Qwen3.7 Max Thinking peut se faire avec les techniques standard : LoRA/QLoRA (paramétrie efficace, recommandé pour les équipes avec ressources limitées), SFT (Supervised Fine-Tuning) complet sur données propriétaires, et DPO (Direct Preference Optimization) pour aligner le comportement du modèle sur vos préférences. Les outils compatibles incluent LLaMA-Factory, Axolotl, et les scripts officiels d'Alibaba sur GitHub. Le modèle fine-tuné résultant est votre propriété exclusive — aucune obligation de partager avec Alibaba ou la communauté.

Qwen3.7 Max Thinking est-il disponible en France ?

Oui, via plusieurs canaux. API Alibaba Cloud (DashScope) avec endpoints européens disponibles depuis début 2026. API partenaires sans dépendance à Alibaba : Together AI, Fireworks AI, OpenRouter. Déploiement on-premise sur votre infrastructure avec les poids Hugging Face — c'est la solution recommandée pour les organisations soumises à des contraintes de localisation des données (RGPD, SecNumCloud). Note : DashScope n'est pas actuellement certifié SecNumCloud, donc pour les données sensibles françaises, le déploiement on-premise ou via un cloud souverain européen est obligatoire.

Qwen3.7 Max Thinking est-il bon pour la cybersécurité ?

Qwen3.7 Max Thinking est pertinent pour la cybersécurité dans plusieurs scénarios : analyse de code source à la recherche de vulnérabilités (HumanEval 91% indique une bonne compréhension du code), génération de scripts de pentest (sous contrôle éthique), analyse de logs et détection d'anomalies, rédaction de rapports de sécurité en français (8,5/10), et recherche en CTF. Son alignement légèrement moins strict que Claude peut être un avantage pour les chercheurs en sécurité qui ont besoin de tester des scénarios d'attaque légitimes. Consultez notre benchmark IA complet de juillet 2026 pour les comparatifs détaillés.

