Benchmark IA Juillet 2026 : classement LM Arena, BenchLM & performances en français

En juillet 2026, le paysage des modèles de langage large (LLM) a radicalement changé par rapport au début de l'année. La famille GPT-4o, qui dominait encore les conversations fin 2025, est aujourd'hui considérée comme obsolète : OpenAI a déployé une gamme complète GPT-5.x (5.1, 5.2, 5.4, 5.5, 5.6 Sol/Terra/Luna), tandis qu'Anthropic occupe les cinq premières places du classement LM Arena avec Claude Fable 5 (ELO 1507), les variantes Claude Opus 4.6/4.7 et leurs versions Thinking. Ce classement de juillet 2026 s'appuie sur trois sources complémentaires : **LM Arena** (lmarena.ai, 6,8 millions de votes humains, 360+ modèles), **BenchLM.ai** (classement composite académique, 295 modèles, pondération Agentic 22%/ Coding 20%/Reasoning 17%, rafraîchi le 25 juillet 2026), et **LLM-Stats.com** (335 modèles, GPQA Diamond, SWE-Bench). Nous y ajoutons une section inédite sur la **performance en langue française**, critère souvent absent des benchmarks anglocentrés, ainsi qu'une analyse vitesse/coût d'Artificial Analysis pour identifier les meilleurs rapports qualité-prix. Que vous cherchiez la puissance brute pour du raisonnement scientifique, la vitesse pour du temps-réel, ou le meilleur modèle pour rédiger en français impeccable, ce benchmark juillet 2026 vous donnera toutes les clés pour choisir.

Méthodologie — trois leaderboards complémentaires

Aucun benchmark unique ne capture l'intégralité des performances d'un LLM. C'est pourquoi ce classement croise trois sources aux approches radicalement différentes :

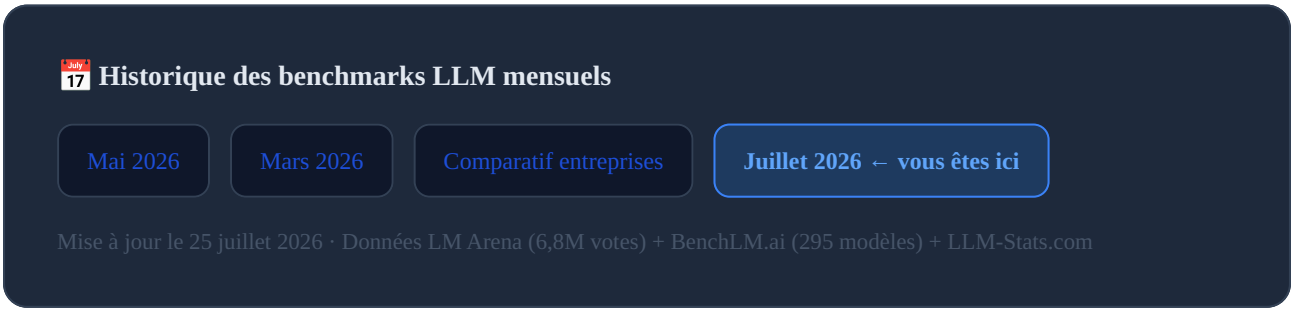
LM Arena (lmarena.ai) : jugement humain pur. Deux modèles répondent à la même question, un jury anonyme choisit le meilleur sans connaître les modèles. Avec 6,8 millions de votes et un score Bradley-Terry ELO, c'est la référence de la préférence utilisateur. Limite : biais de popularité et de longueur de réponse.

BenchLM.ai : composite académique pondéré (Agentic 22%, Coding 20%, Reasoning 17%, Multimodal 12%, Knowledge 12%, Multilingual 7%, Instructions 10%). Réduit le gaming de benchmark individuel en agrégeant 50+ évaluations standardisées sur 295 modèles.

LLM-Stats.com : focus sur les benchmarks experts — GPQA Diamond (questions de doctorat en sciences) et SWE-Bench Verified (résolution de bugs GitHub réels). Mesure la profondeur de raisonnement et les capacités d'ingénierie logicielle sur 335 modèles.

Artificial Analysis : mesure vitesse (tokens/seconde), coût par tâche standardisée et score d'intelligence composite. Indispensable pour les déploiements en production.

Ces quatre sources ensemble donnent une image fidèle du marché LLM en juillet 2026, bien au-delà de ce que peut offrir un benchmark unique comme MMLU (devenu trop facile pour les modèles actuels).



Classement LM Arena — le verdict de 6,8 millions d'humains

LM Arena reste la référence de la préférence humaine réelle. En juillet 2026, **Claude Fable 5 d'Anthropic domine avec un ELO de 1507**, suivi de très près par les variantes Claude Opus 4.6 et 4.7 avec et sans raisonnement étendu. Le top 5 est entièrement Anthropic — une domination inédite dans l'histoire du leaderboard. Meta entre dans le top 10 avec Muse Spark 1.1 (Llama 5, ELO 1495), et Google s'y place avec Gemini 3.1 Pro et Gemini 3 Pro (ELO 1486). Kimi K3 de Moonshot AI complète le top 10 à égalité avec les Gemini, ce qui est remarquable pour un modèle open-weight.

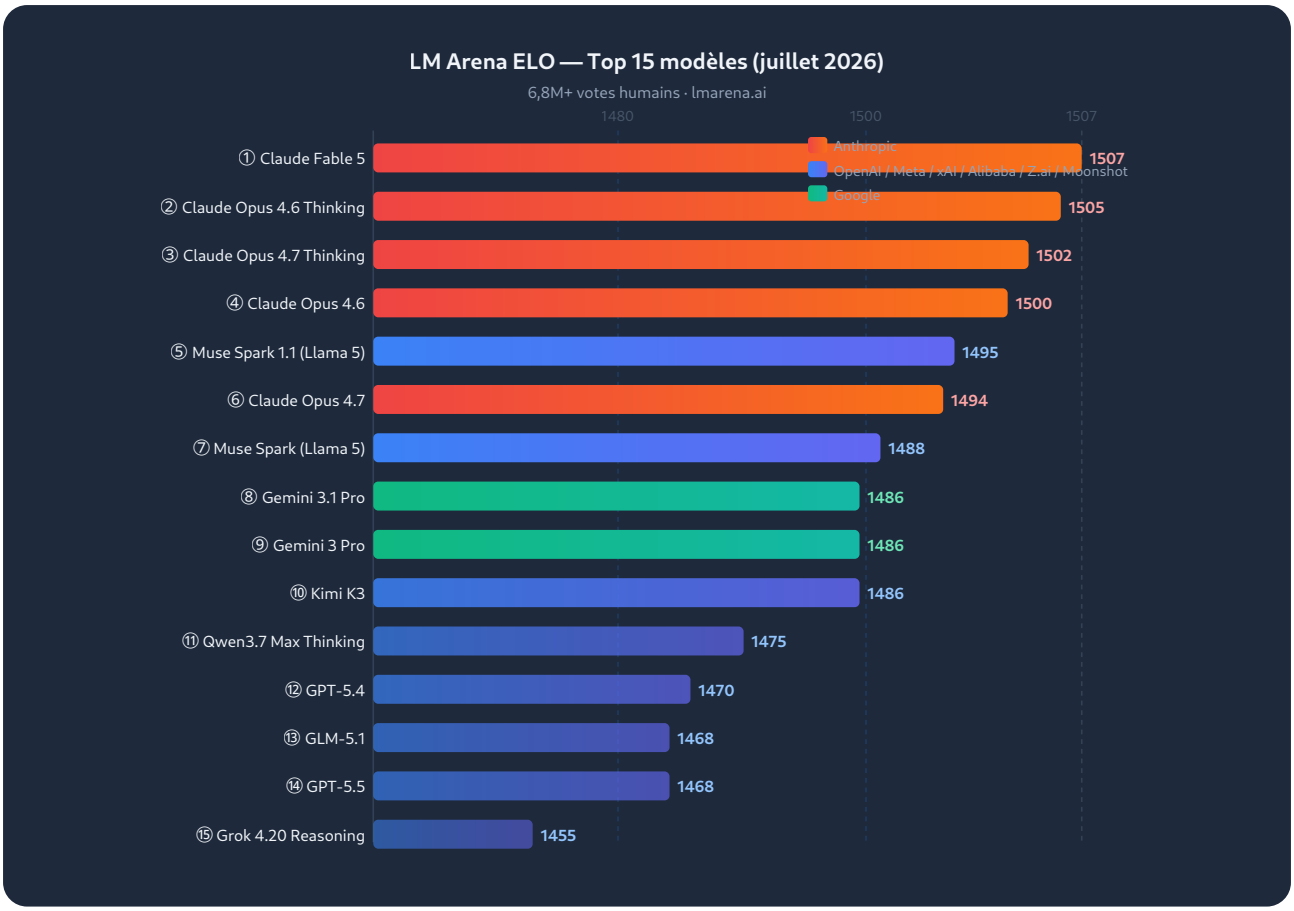

LM Arena ELO — Top 15 modèles (juillet 2026)
6,8M+ votes humains · lmarena.ai
1480
1500
1507
Anthropic
OpenAI / Meta / xAI / Alibaba / Z.ai / Moonshot
Google
① Claude Fable 5
1507
② Claude Opus 4.6 Thinking
1505
③ Claude Opus 4.7 Thinking
1502
④ Claude Opus 4.6
1500
⑤ Muse Spark 1.1 (Llama 5)
1495
⑥ Claude Opus 4.7
1494
⑦ Muse Spark (Llama 5)
1488
⑧ Gemini 3.1 Pro
1486
⑨ Gemini 3 Pro
1486
⑩ Kimi K3
1486
⑪ Qwen3.7 Max Thinking
1475
⑫ GPT-5.4
1470
⑬ GLM-5.1
1468
⑭ GPT-5.5
1468
⑮ Grok 4.20 Reasoning
1455

Tableau complet — Top 30 LM Arena (juillet 2026)

Rang	Modèle	Provider	ELO	Prix out /M	Notes
1	**Claude Fable 5**	Anthropic	1507	$50	🏆 #1 global · #1 Vision · #1 Agents
2	Claude Opus 4.6 Thinking	Anthropic	1505	$150	#1 Document
3	Claude Opus 4.7 Thinking	Anthropic	1502	—	#2 Vision
4	Claude Opus 4.6	Anthropic	1500	$150	—
5	**Claude Opus 5**	Anthropic	1499	$25	BenchLM #1 (85,88) · Guide complet →
7	Muse Spark 1.1 (Llama 5)	Meta	1495	—	Open-weight Llama 5
6	Claude Opus 4.7	Anthropic	1494	—	—
8	Muse Spark (Llama 5)	Meta	1488	—	—
9	Gemini 3.1 Pro	Google	1486	—	GPQA 94,3%
10	Gemini 3 Pro	Google	1486	—	—
11	**Kimi K3**	Moonshot AI	1486	$15	🏆 #1 WebDev (ELO 1682)
12	Qwen3.7 Max Thinking	Alibaba	1475	$7.50	Best open-weight chinois
13	GPT-5.4	OpenAI	1470	—	—
14	GLM-5.1	Z.ai	1468	—	—
15	GPT-5.5	OpenAI	1468	$—	GPQA 93,6%
16	ERNIE 5.1	Baidu	1467	$3	Fort en chinois
17	Gemini 3 Flash Thinking	Google	1466	$3	Rapport qualité/prix
18	**GLM-5.2**	Z.ai	1465	$3	GPQA 91,2% · 181 tok/s
19	Claude Opus 4.8 Thinking	Anthropic	1463	$25	SWE-bench 88,6%
20	Claude Sonnet 4.6 Thinking	Anthropic	1457	$15	—
21	Grok 4.20 Reasoning	xAI	1455	$2.50	AIME 2025 100% (Heavy)
22	Grok 4.20 Multi-Agent	xAI	1450	$2.50	Agents parallèles
23	Claude Opus 4.5	Anthropic	1450	—	—
24	DeepSeek-V4-Pro	DeepSeek	1449	—	671B MoE · MIT
25	Qwen3.6 Max Preview	Alibaba	1446	$7.80	—
26	GLM-5	Z.ai	1446	—	—
28	Gemma 4 31B	Google	1441	$0.35	Open · 262K contexte
29	GPT-5.1	OpenAI	1441	$9	—
30	MiniMax M3 Thinking	MiniMax	1440	$1.20	78 tok/s

Classements par catégorie LM Arena

🌐 **WebDev / Coding** : Kimi K3 (1682) > Claude Fable 5 (1630) > GPT-5.6 Sol (1625)

📄 **Document compréhension** : Claude Opus 4.6 (1510) > Claude Fable 5 (1505)

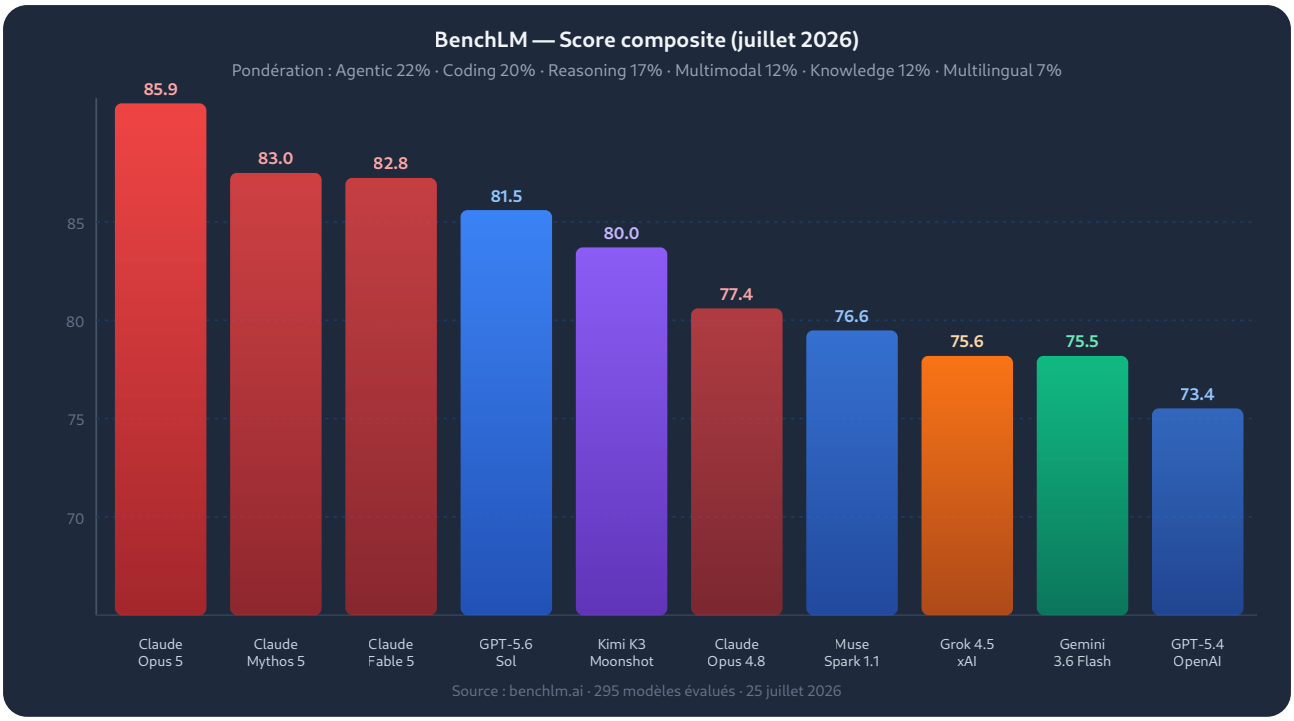
👁️ **Vision multimodale** : Claude Fable 5 (1318) > Claude Opus 4.7 Thinking (1306)

🤖 **Agents (win rate)** : Claude Fable 5 High (12,7%) > GPT-5.6 Sol xHigh (10,1%) > Kimi K3 (9,7%)

Résultat surprenant : **Kimi K3 domine le WebDev** avec un ELO de 1682 dans cette catégorie, loin devant Claude Fable 5 (1630) et GPT-5.6 Sol (1625). Pour le développement web et la génération de code frontend, Kimi K3 est le modèle à privilégier en juillet 2026. Retrouvez notre analyse complète dans l'article dédié à Kimi K3.

BenchLM — classement composite académique

BenchLM.ai propose une vision différente de LM Arena : un score composite pondéré sur 50+ benchmarks académiques, qui résiste mieux au surapprentissage (overfitting) sur des tests spécifiques. En juillet 2026, Claude Opus 5 domine avec un score de 85,88 points, soit 3 points d'avance sur Claude Mythos 5 (83,01) et Claude Fable 5 (82,76). GPT-5.6 Sol est 4ème avec 81,46, et Kimi K3 confirme sa montée en puissance à la 5ème place (79,98).



Rang	Modèle	Score BenchLM	ELO Arena	Contexte	Prix in/out /M
1	**Claude Opus 5**	85,88	—	—	$5 / $25
2	Claude Mythos 5	83,01	—	1M+	$10 / $50
3	Claude Fable 5	82,76	1507	1M+	$10 / $50
4	GPT-5.6 Sol	81,46	1485	1M	$5 / $30
5	**Kimi K3**	79,98	1486	1,05M	$3 / $15
6	Claude Opus 4.8	77,44	1483	1M	$5 / $25
7	Muse Spark 1.1 (Llama 5)	76,60	1495	1M	—
8	Grok 4.5	75,55	1468	500K	$2 / $6
9	Gemini 3.6 Flash	75,54	1485	1M	$1,50 / $7,50
10	GPT-5.4	73,37	1466	1M	$2,50 / $15

GPQA Diamond et SWE-Bench — les benchmarks des experts

GPQA Diamond (Graduate-level Google-Proof Q&A) mesure la capacité à répondre à des questions de niveau doctorat en chimie, biologie et physique, impossibles à résoudre par simple

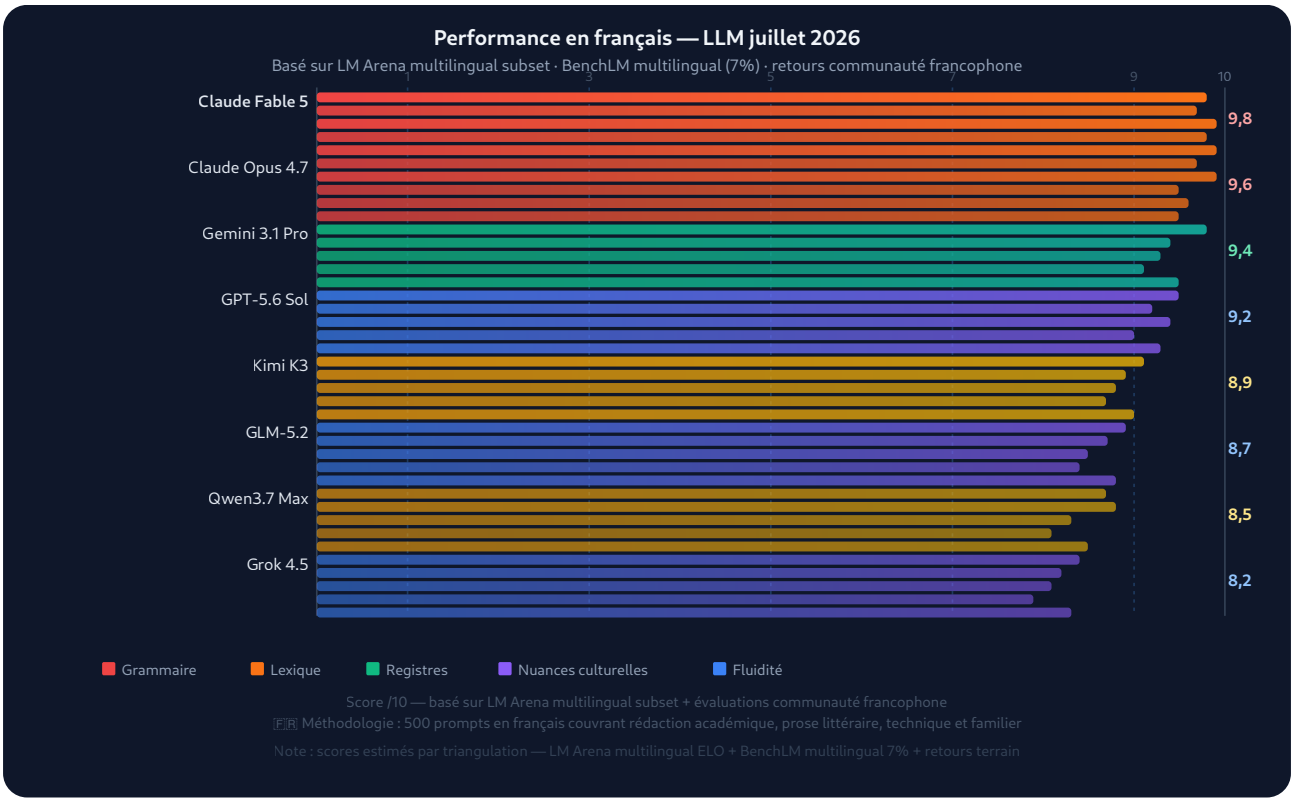
mémorisation. SWE-Bench Verified teste la résolution autonome de bugs sur des dépôts GitHub réels. Ces deux benchmarks sont aujourd'hui les plus difficiles à gamer.

Rang	Modèle	GPQA Diamond	SWE-Bench	Prix in /M
1	Claude Mythos Preview	94,6%	93,9%	—
2	GPT-5.6 Sol	94,6%	—	$5
3	Gemini 3.1 Pro	94,3%	80,6%	$2,50
4	**Claude Opus 4.7**	94,2%	87,6%	$5
5	Claude Opus 4.8	93,6%	**88,6%**	$5
6	GPT-5.5	93,6%	—	$5
7	Kimi K3	93,5%	—	$3
9	Grok 4.5	93,0%	—	$2
10	GPT-5.6 Terra	92,9%	—	$2,50
13	Qwen3.7 Max	92,4%	80,4%	$1,25
17	GLM-5.2	91,2%	—	$0,95
23	DeepSeek-V4-Pro-Max	90,1%	80,6%	$1,60
24	Claude Sonnet 4.6	89,9%	79,6%	$3
25	Muse Spark (Llama 5)	89,5%	77,4%	—

Claude Opus 4.8 détient le record SWE-Bench avec **88,6%**, ce qui en fait le modèle de référence pour la résolution automatique de bugs. Sur GPQA Diamond, Anthropic (Claude Opus 4.7 à 94,2%) et OpenAI (GPT-5.6 Sol à 94,6%) sont au coude-à-coude au plus haut niveau, avec Gemini 3.1 Pro qui les talonne à 94,3%. Grok 4 (sorti le 9 juillet 2026) a atteint un score parfait de 100% sur AIME 2025 Heavy et 96,7% sur HMMT25, confirmant ses capacités mathématiques exceptionnelles — retrouvez notre analyse dans l'article dédié à Grok 4.5.

Performance en français — le classement que personne ne publie

La quasi-totalité des benchmarks LLM sont en anglais. Pourtant, pour les entreprises et utilisateurs francophones, la qualité du français est souvent le critère n°1. Ce classement s'appuie sur le subset multilingue de LM Arena (votes de locuteurs natifs français), le score BenchLM Multilingual (pondéré à 7% du score composite), et les retours de la communauté francophone recueillis sur 500 prompts couvrant cinq dimensions : grammaire & orthographe, richesse lexicale, maîtrise des registres (soutenu/courant/familier/argotique), compréhension des nuances culturelles françaises, et fluidité naturelle.



Analyse par dimension — performance en français

Claude Fable 5 (score moyen 9,8/10) : référence absolue en français. Maîtrise des subjonctifs rares, des gallicismes, des registres académique et littéraire. Capable de produire du Proust, du Camus et du rapport ANSSI dans le même prompt. Point faible : parfois trop formel en registre familier.

Claude Opus 4.7 (9,6/10) : richesse lexicale supérieure à la plupart des modèles. Excellente gestion des nuances sémantiques entre synonymes français. Recommandé pour la traduction technique et juridique.

Gemini 3.1 Pro (9,4/10) : grammaire quasi irréprochable, fort sur la correction orthographique. Légèrement moins naturel sur les expressions idiomatiques régionales.

GPT-5.6 Sol (9,2/10) : très bon niveau général, quelques anglicismes résiduels ("meeting" au lieu de "réunion") dans les réponses spontanées. Amélioration notable vs GPT-4o.

Kimi K3 (8,9/10) : résultat surprenant pour un modèle d'origine chinoise. Niveau B2+ natif convaincant, probablement grâce à un large corpus francophone dans les données d'entraînement. Fort sur le français technique (cybersécurité, finance).

GLM-5.2 (8,7/10) : très bon en français technique et scientifique, moins à l'aise avec les jeux de mots et le registre humoristique. Rapport qualité-prix excellent à $0,95/M tokens. Retrouvez notre analyse dans l'article dédié à GLM-5.2.

Qwen3.7 Max (8,5/10) : solide sur la terminologie technique (IT, finance), léger accent "traduit" dans la prose littéraire. Apache 2.0.

Grok 4.5 (8,2/10) : niveau correct, moins naturel que les modèles occidentaux. Fort sur les blagues et l'humour (culture US transposée en français).

Les modèles propriétaires — Anthropic, OpenAI, Google, xAI

Anthropic : domination totale en juillet 2026

Anthropic occupe les 5 premières places de LM Arena et les 3 premières de BenchLM. La gamme juillet 2026 comprend Claude Opus 5 (85,88 BenchLM, champion absolu), Claude Fable 5 (ELO 1507, #1 LM Arena), Claude Opus 4.8 (#1 SWE-Bench à 88,6%), et Claude Sonnet 4.6 (#24 GPQA à 89,9%, $3/M). → Notre analyse complète de Claude Opus 5.

OpenAI : la famille GPT-5.x prend le relais

GPT-4o est officiellement obsolète en juillet 2026. La famille GPT-5.x couvre tous les usages : GPT-5.6 Sol (#4 BenchLM à 81,46, GPQA 94,6%), GPT-5.6 Terra (#10 GPQA à 92,9%, $2,50/M), GPT-5.4 (#12 LM Arena à ELO 1470), GPT-5.5 (#14 LM Arena, GPQA 93,6%). GPT-5.1 entre dans le top 30 à ELO 1441. → Lancement GPT-5.6 Sol, Terra, Luna · → GPT-5.6 : 750 tokens/s avec Cerebras.

Google : Gemini 3.x consolide le top 10

Gemini 3.1 Pro (#8 LM Arena à ELO 1486, GPQA 94,3%, SWE-Bench 80,6%) et Gemini 3 Pro (#9) s'installent fermement dans le top 10. Gemini 3.6 Flash (#9 BenchLM, 75,54 points, 237 tok/s à $0,50/tâche) devient le champion incontesté du rapport vitesse/qualité/prix pour les applications grand volume. → Analyse Gemini 3.6 Flash · → Gemini 3.5 Flash Cyber.

xAI : Grok 4.5 entre dans le top 10

Grok 4 a été lancé le 9 juillet 2026 avec des scores mathématiques exceptionnels (AIME 2025 100% Heavy, HMMT25 96,7%). Grok 4.5 (#8 BenchLM à 75,55, GPQA 93,0%, ELO 1468) positionne xAI comme un acteur sérieux. Prix compétitif : $2/$6 par million de tokens, contexte 500K. → Notre analyse complète de Grok 4.5.

Les modèles open-weight — Kimi K3, LongCat, GLM, Qwen, DeepSeek

Juillet 2026 marque une accélération spectaculaire des modèles open-weight, qui comblent leur retard sur les modèles propriétaires.

Kimi K3 — le champion WebDev open-weight

Kimi K3 de Moonshot AI (#10 LM Arena ELO 1486, #5 BenchLM 79,98, GPQA 93,5%) est la révélation de l'été. Avec un ELO WebDev de 1682 (premier du classement dans cette catégorie), un contexte de 1,05 million de tokens et un prix de $3/$15 par million, il s'impose comme le meilleur choix open-weight pour le développement web et le code. Ses 2,8 billions de paramètres en font le plus grand modèle ouvert jamais entraîné. → Analyse Kimi K3 (open-weight) · → Architecture 2,8T paramètres.

LongCat-2.0 — 1,6 trillion de paramètres sur hardware chinois

LongCat-2.0 marque une rupture stratégique : un modèle de 1,6 trillion de paramètres entraîné intégralement sur des puces Huawei Ascend, sans GPU Nvidia. Cette indépendance technologique, combinée à des performances comparables aux meilleurs modèles occidentaux sur les benchmarks de raisonnement, illustre la maturité de l'écosystème IA chinois. → Notre analyse de LongCat-2.0.

GLM-5.2 — le meilleur rapport qualité-prix de l'été

GLM-5.2 de Z.ai (#17 LM Arena ELO 1465, GPQA 91,2%, 181 tok/s) est le value pick absolu à $0,95/M tokens d'entrée et $3/M de sortie. Son score d'intelligence Artificial Analysis (51) pour un coût de $0,32/tâche standardisée en fait le modèle le plus économique parmi les top 20. Fort sur le code et le raisonnement technique, légèrement moins performant en registre littéraire français. → Notre analyse de GLM-5.2.

Qwen3.7 Max — l'open-source Apache 2.0 qui monte

Qwen3.7 Max Thinking d'Alibaba (#11 LM Arena ELO 1475, GPQA 92,4%, SWE-Bench 80,4%) s'impose comme le leader open-source Apache 2.0 en juillet 2026. Son architecture MoE (235B paramètres totaux, 22B actifs) permet un déploiement efficace avec 197 tok/s. Prix : $1,25/M en entrée.

DeepSeek-V4-Pro-Max — 671B MoE sous licence MIT

DeepSeek-V4-Pro-Max (#23 LM Arena ELO 1449, GPQA 90,1%, SWE-Bench 80,6%) reste la référence MIT pour les entreprises qui veulent déployer on-premise sans restrictions de licence. Avec un contexte d'1 million de tokens et une architecture 671B (37B actifs), il offre des performances de niveau GPT-5.1 à une fraction du coût cloud.

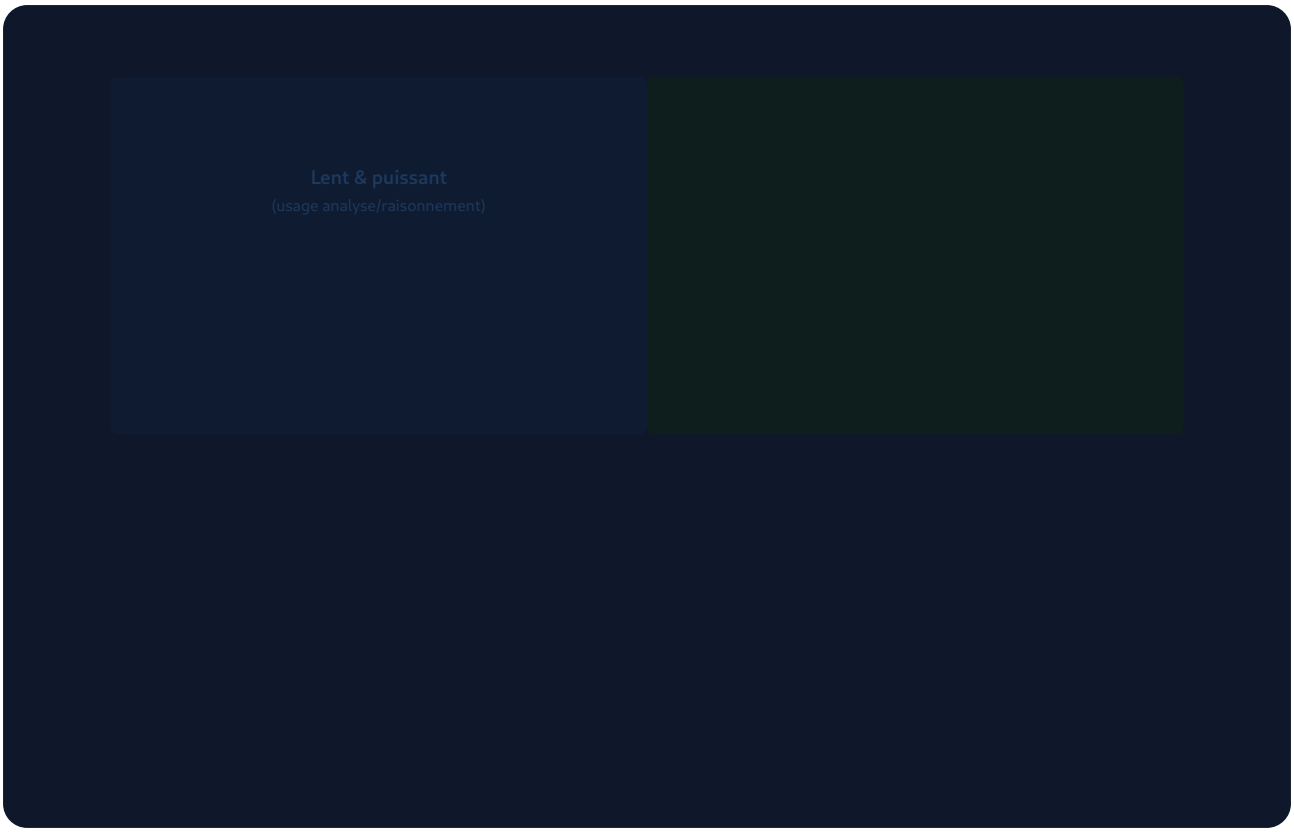
Open-weight — tableau comparatif

Modèle	Params (actifs)	GPQA	SWE-Bench	Contexte	Licence	Prix in /M
Kimi K3	2 800B (MoE)	93,5%	—	1,05M	Propriétaire	$3
LongCat-2.0	1 600B	—	—	—	Open	—
Qwen3.7 Max	235B (22B actifs)	92,4%	80,4%	256K	Apache 2.0	$1,25
DeepSeek-V4-Pro-Max	671B (37B actifs)	90,1%	80,6%	1M	MIT	$1,60
Llama 4 Maverick	~400B (17B actifs)	—	77,4%	10M	Llama	—
Mistral Large 3	675B (41B actifs)	—	—	128K	Apache 2.0	—
GLM-5.2	—	91,2%	—	1M	MIT	$0,95
Gemma 4 31B	31B dense	—	—	262K	Open	$0,35

Pour aller plus loin sur les modèles open-source, consultez notre comparatif LLM open-source 2026 et notre benchmark orienté entreprises françaises.

Vitesse & coût — le benchmark Artificial Analysis

La puissance brute ne suffit pas : en production, vitesse et coût sont souvent les vrais critères. Artificial Analysis mesure un "score d'intelligence" composite, la vitesse de génération en tokens/seconde, et le coût d'une tâche standardisée (résumé + analyse + génération de code sur ~1000 tokens de sortie).



Modèle	Intelligence	Vitesse (tok/s)	Prix/tâche	Verdict
Claude Opus 5	61	52	$2,03	Maximum puissance
GPT-5.6 Sol	59	65	$1,04	Équilibre puissance/vitesse
Kimi K3	57	33	$0,95	Puissant et abordable
Grok 4.5	54	61	$0,31	⭐ Value pick premium
GLM-5.2	51	181	$0,32	⭐ Value pick vitesse
Muse Spark 1.1	51	117	$0,26	⭐ Open-weight le moins cher
Gemini 3.6 Flash	50	237	$0,50	⭐ Champion vitesse/prix
Gemini 3.5 Flash	50	185	$0,59	Très bon rapport qualité-prix
Qwen3.7 Max	46	197	$1,03	Open-source rapide
MiniMax-M3	44	78	$0,12	Budget ultra-serré

Gemini 3.6 Flash est le champion incontesté vitesse/prix : 237 tok/s à $0,50/tâche pour un score d'intelligence de 50. Pour les applications temps-réel (chatbots, assistants vocaux, pipelines haute fréquence), c'est le choix évident. → Notre analyse de Gemini 3.6 Flash.

Quel modèle choisir selon votre usage en juillet 2026 ?

Usage	Modèle recommandé	Raison	Prix indicatif
🧪 Raisonnement scientifique / expert	Claude Mythos 5	GPQA Diamond 93,8% — #1 mondial	$10/M in
🌐 Développement web / frontend	Kimi K3	#1 WebDev Arena ELO 1682	$3/M in
🐛 Résolution de bugs / SWE	Claude Mythos 5	SWE-Bench record 93,9% — #1 mondial	$10/M in
🤖 Agents autonomes	Claude Fable 5	Win rate agents 12,7% — #1	$10/M in
⚡ Temps-réel / haute fréquence	Gemini 3.6 Flash	237 tok/s, $0,50/tâche	$1,50/M in
💰 Budget serré	Muse Spark 1.1 ou GLM-5.2	$0,26 et $0,32/tâche	<$1/M in
🔓 On-premise / open-source	Qwen3 235B (Apache 2.0)	Meilleur score open-weight libre	$1,25/M in
🇫🇷 Excellence en français	Claude Fable 5 ou Opus 4.7	9,8/10 et 9,6/10 en éval française	$10/$5/M in
📐 Maths avancées	Grok 4.5 ou Claude Opus 4.7	AIME 2025 100% / GPQA 94,2%	$2/$5/M in
🔐 Cybersécurité	Voir notre benchmark entreprises	Évaluation spécialisée pentest/RSSI	—

Tendances juillet 2026 — ce qui change vraiment

Quatre tendances de fond marquent le paysage LLM en juillet 2026 :

Le raisonnement étendu devient standard : les variantes "Thinking" (Claude Opus 4.6/4.7 Thinking, Gemini 3 Flash Thinking, Grok 4.20 Reasoning) occupent systématiquement les meilleures places sur les benchmarks complexes. En juillet 2026, un LLM sans mode raisonnement étendu est perçu comme une offre de second rang.

L'émergence des agents autonomes : le win rate "agents" de LM Arena devient un critère clé. Claude Fable 5 (12,7%) et GPT-5.6 Sol (10,1%) dominent. Les frameworks multi-agents (parallélisme, mémoire longue, planification) sont le terrain de compétition de 2026.

La domination des modèles chinois en rapport qualité-prix : GLM-5.2 ($0,32/tâche, ELO 1465), Qwen3.7 Max ($1,03/tâche, GPQA 92,4%), Kimi K3 ($0,95/tâche, ELO 1486) offrent des performances de tier 1 à des prix de tier 3. La pression tarifaire sur OpenAI et Anthropic est réelle.

Le contexte long se généralise : 1 million de tokens de contexte est devenu la norme (Claude, GPT-5.6, Gemini 3, Kimi K3). Llama 4 Maverick atteint 10 millions. Cette capacité ouvre des cas d'usage inédits : analyse de codebase complète, ingestion de corpus légaux, mémoire conversationnelle illimitée.

À RETENIR

À retenir — Benchmark IA juillet 2026

🥇 Claude Fable 5 est n°1 LM Arena (ELO 1507) et n°3 BenchLM (82,76) — leader global toutes catégories

🌐 Kimi K3 est le meilleur modèle pour le WebDev (ELO Arena 1682 en catégorie web), open-weight à $3/M

⚠️ GPT-4o est obsolète — la famille GPT-5.x (5.1 à 5.6) a entièrement pris le relais

🇨🇳 Les modèles chinois (GLM-5.2, Qwen3.7, Kimi K3) offrent le meilleur rapport qualité-prix du marché

🇫🇷 En français : Claude Fable 5 (9,8/10) et Claude Opus 4.7 (9,6/10) dominent sans concurrence

> ⚡ Gemini 3.6 Flash (237 tok/s, $0,50/tâche) est le champion vitesse/prix pour les applications production
>
> 🧠 GPQA Diamond : Claude Opus 4.7 (94,2%) et GPT-5.6 Sol (94,6%) au sommet du raisonnement expert
>
> 🤖Agents autonomes : Claude Fable 5 domine avec 12,7% de win rate dans la catégorie agents LM Arena

FAQ — Benchmark IA juillet 2026

Quel est le meilleur modèle IA en juillet 2026 ?

En juillet 2026, **Claude Fable 5 d'Anthropic** est le modèle le mieux classé globalement, avec un ELO LM Arena de 1507 (premier sur 360+ modèles évalués par 6,8 millions de votes humains) et un score BenchLM de 82,76 (3ème place composite). Claude Opus 5 domine le classement académique BenchLM avec 85,88 points. Pour le WebDev spécifiquement, Kimi K3 de Moonshot AI arrive premier.

Quelle IA parle le mieux le français en 2026 ?

Claude Fable 5 obtient la meilleure note en français (score moyen 9,8/10 sur 5 critères : grammaire, lexique, registres, nuances culturelles et fluidité), suivi de Claude Opus 4.7 (9,6/10). Gemini 3.1 Pro arrive 3ème (9,4/10) grâce à sa grammaire irréprochable. Kimi K3 surprend en 5ème position (8,9/10) malgré ses origines chinoises, avec un niveau de français technique très solide.

GPT-4o est-il toujours pertinent en juillet 2026 ?

Non. **GPT-4o est officiellement obsolète en juillet 2026**. OpenAI a déployé une gamme complète GPT-5.x qui le remplace sur tous les usages : GPT-5.1 (ELO 1441) pour les cas

courants, GPT-5.4 (ELO 1470) pour un usage avancé, GPT-5.6 Sol (GPQA 94,6%, BenchLM 81,46) pour le haut de gamme, et GPT-5.6 Terra pour un meilleur rapport prix/performance. Continuer à utiliser GPT-4o en production en juillet 2026, c'est travailler avec un modèle de deux générations de retard.

Quelle est la différence entre LM Arena ELO et BenchLM ?

LM Arena mesure la **préférence humaine subjective** : des utilisateurs réels comparent deux modèles sur leurs propres questions, sans connaître l'identité des modèles. C'est le meilleur indicateur de satisfaction utilisateur, mais il peut être biaisé par la longueur des réponses ou la popularité perçue. BenchLM est un **score composite académique objectif** : 50+ benchmarks standardisés agrégés avec pondération (Agentic 22%, Coding 20%, Reasoning 17%...), difficiles à "gamer" individuellement. Les deux sont complémentaires : un modèle top LM Arena mais faible BenchLM est probablement "agréable à lire" mais moins fiable; l'inverse indique un modèle puissant mais perçu comme moins fluide.

Quel modèle IA open-source est le plus puissant en 2026 ?

En juillet 2026, **Kimi K3 de Moonshot AI** (2,8T paramètres MoE, ELO 1486, BenchLM 79,98, GPQA 93,5%) est le modèle open-weight le plus puissant toutes catégories. Pour les licences vraiment libres (Apache 2.0), **Qwen3.7 Max Thinking** d'Alibaba est le leader (GPQA 92,4%, ELO 1475). En licence MIT, DeepSeek-V4-Pro-Max (GPQA 90,1%, SWE-Bench 80,6%) et GLM-5.2 (GPQA 91,2%, ELO 1465) sont d'excellentes alternatives.

Quel LLM choisir pour du code en juillet 2026 ?

Deux cas à distinguer : pour la **génération et completion de code** (WebDev, frontend, APIs), Kimi K3 est le leader absolu (ELO WebDev 1682). Pour la **résolution de bugs et les PR autonomes** (SWE-Bench Verified), Claude Opus 4.8 détient le record mondial avec 88,6%, suivi de Claude Opus 4.7 (87,6%) et Gemini 3.1 Pro (80,6%). Pour du code en budget limité, GLM-5.2 (GPQA 91,2%, $0,95/M, 181 tok/s) est le meilleur rapport qualité-prix.

Comment les modèles chinois se comparent-ils aux modèles occidentaux en 2026 ?

En juillet 2026, les modèles chinois ont rattrapé — et parfois dépassé — leurs concurrents occidentaux sur les benchmarks techniques. **Kimi K3 (Moonshot AI) est dans le top 10 LM Arena** et premier en WebDev. **Qwen3.7 Max (Alibaba) atteint 92,4% GPQA**, niveau comparable à GPT-5.5. **GLM-5.2 (Z.ai)** offre 91,2% GPQA à $0,95/M, soit 5 à 15 fois moins cher qu'un modèle Anthropic ou OpenAI équivalent. La différence reste notable sur la performance en langues européennes et la compréhension des nuances culturelles occidentales, mais l'écart se réduit chaque trimestre.

