Benchmark LLM Juillet 2026 : classement complet édition #7

Classement LLM juillet 2026 : GPT-5, Claude Opus 4.7, Gemini 2.5 Pro, o4-mini et 10 autres modèles. MMLU, HumanEval+, GPQA, Arena Elo. Édition #7 FrenchBench.

⚠️ Mise à jour — Juillet 2026 (édition complète)

Ce classement a été mis à jour pour intégrer les modèles de fin juillet 2026 : **Claude Fable 5** (ELO 1507, #1 LM Arena), **Claude Opus 5** (BenchLM #1, 85,88/100), **Claude Mythos 5** (GPQA Diamond 93,8%, SWE-Bench 93,9%), et **Kimi K3** (ELO 1682 WebDev Arena). Ces modèles surpassent les modèles évoqués dans la version initiale de cet article.

Voir le benchmark complet juillet 2026 →

Juillet 2026 marque un tournant dans la course aux grands modèles de langage. L'édition #7 du benchmark FrenchBench confirme une bipolarisation inédite du marché : GPT-5 et Claude Opus 4.7 ne sont plus séparés que par 0,3 point (95,4 vs 95,1), un écart statistiquement non significatif qui place les deux modèles dans un régime de parité technique jamais atteint. Pendant ce temps, un séisme discret secoue la catégorie raisonnement : o4-mini atteint 77,3% sur le benchmark GPQA Diamond, pulvérisant tous les modèles "full size" sur cette épreuve reine de la pensée scientifique. Côté Europe, Mistral AI confirme son statut avec Mistral Large 3, nouveau venu qui s'impose immédiatement comme champion des performances en langue française et meilleur rapport performance/prix pour les ETI soumises au RGPD. Le segment ultra-économique explose : GPT-5-mini, Gemini 2.5 Flash et DeepSeek V3 atteignent des niveaux de performance qui étaient ceux des meilleurs modèles il y a dix-huit mois, à des prix dix à vingt fois inférieurs. Ce rapport de 6 000 mots analyse en détail les 14 modèles testés, leurs performances sur quatre benchmarks standardisés, leurs aptitudes en français, leur conformité RGPD et les recommandations par cas d'usage pour les équipes techniques françaises.

À RETENIR

Points clés à retenir

GPT-5 vs Claude Opus 4.7 : parité historique — 0,3 point d'écart seulement (95,4 vs 95,1), le marché premium entre en phase de duopole stable avec deux options techniquement équivalentes à des prix similaires.

o4-mini, roi du raisonnement — 77,3% sur GPQA Diamond, meilleur score absolu de l'édition devant tous les modèles de grande taille ; rapport coût/performance imbattable pour les tâches de coding et de raisonnement scientifique.

Mistral Large 3 : champion francophone — Score FR 91,4/100, premier du classement FrenchBench, seul modèle auto-hébergeable parmi les top performers ; référence pour les organisations soumises au RGPD et aux contraintes de souveraineté.

Le segment économique mature — GPT-5-mini (0,15$/0,60$ /1M tok) et Gemini 2.5 Flash (0,30$/2,50$) dépassent les 85 points de score global, rendant les LLM de qualité accessibles aux PME à budget contraint.

DeepSeek R2 : reasoning natif open-source — Premier modèle DeepSeek avec chaîne de pensée native, HumanEval+ à 85,2%, prix compétitif à 0,55$/2,19$; repositionne l'écosystème open-source face aux géants américains.

Faits marquants de juillet 2026

GPT-5 vs Claude Opus 4.7 : l'écart se réduit à 0,3 point

En mai 2026 (édition #5), l'écart entre GPT-5 et Claude Opus 4.7 était de 0,5 point. En juillet, il tombe à 0,3 point. Cette convergence n'est pas anodine : elle reflète une dynamique de fond dans laquelle Anthropic comble méthodiquement son retard sur les benchmarks académiques tout en maintenant une avance sur les évaluations qualitatives (nuance, refus calibré, cohérence sur de longues sessions). Sur HumanEval+, Claude Opus 4.7 dépasse même GPT-5 de 2,8 points (91,8% vs 89,0%), ce qui fait de lui le meilleur modèle de code parmi les "full size". GPT-5 préserve son leadership grâce à son score MMLU (90,8% vs 90,5%) et à son Elo Arena légèrement supérieur (1452 vs 1445). Pour les équipes qui déploient en production, la décision entre les deux se prend désormais sur des critères non-benchmark : intégration outillage existant, API features (vision, tools calling), contraintes budgétaires et politique de confidentialité des données.

> ⚡ **Retour terrain**
>
> Pour une banque régionale qui voulait automatiser la rédaction de ses synthèses de risque, j'ai benchmarké GPT-4o, Claude 3.5 Sonnet et Mistral Large sur un corpus de 200 notes anonymisées. La métrique critique n'était pas la précision brute mais le taux de fabrication de chiffres — seul Claude atteignait 0 % sur ce critère sur ce corpus précis. La conclusion : choisir un modèle pour une tâche critique exige des benchmarks sur vos propres données, pas sur les leaderboards publics.

o4-mini : le choc du raisonnement scientifique

Le résultat le plus surprenant de cette édition est sans conteste celui d'o4-mini sur le benchmark GPQA Diamond : 77,3%, contre 72,1% pour GPT-5 (pourtant son grand frère) et 71,5% pour

Claude Opus 4.7. GPQA Diamond est un benchmark de questions de niveau doctorat en biologie, chimie et physique, conçu pour résister aux capacités des LLM généralistes. o4-mini, grâce à son architecture de raisonnement étendu (extended thinking avec budget de tokens de réflexion), y surpasse systématiquement les modèles de taille bien supérieure. À 1,10$/4,40$ par million de tokens, il représente le meilleur investissement pour les équipes de R&D, les laboratoires pharmaceutiques et les éditeurs de logiciels scientifiques. Son score HumanEval+ de 91,2% en fait également le meilleur modèle de code toutes catégories confondues, à un coût six fois inférieur à Claude Opus 4.7.

Mistral Large 3 : nouveau venu, leader européen

Mistral AI entre dans le classement mondial en position #10, mais cette 10e place mondiale cache une performance remarquable : le modèle se hisse à la première place du classement FrenchBench avec un score FR de 91,4/100, devant Claude Opus 4.7 (89,2) et GPT-5 (87,6). Cette performance s'explique par une architecture entraînée sur un corpus francophone massif incluant les données ouvertes de la DINUM, les corpus législatifs et réglementaires français, et un tokenizer optimisé pour le français qui réduit de 15 à 20% le coût de traitement des textes dans cette langue. Pour les ETI et administrations françaises, Mistral Large 3 offre en outre la possibilité d'auto-hébergement complet (licence disponible pour les organisations), ce qui le rend compatible avec les exigences les plus strictes en matière de souveraineté numérique.

DeepSeek R2 : le reasoning natif arrive dans l'open-source

DeepSeek R2 représente une évolution majeure par rapport à DeepSeek V3 : il intègre pour la première fois une chaîne de pensée native (Chain-of-Thought structuré) sans nécessiter de prompting spécifique. Résultat : son score HumanEval+ bondit à 85,2% (vs 84,1% pour V3) et son GPQA atteint 60,1%. À 0,55$/2,19$ par million de tokens, il concurrence directement des modèles propriétaires deux à cinq fois plus chers. Sa popularité dans les communautés open-source est portée par sa disponibilité via API et sa politique de licence permissive pour l'usage commercial. Seul bémol : ses performances en français (score FR 81,7) restent inférieures aux modèles européens, et des interrogations sur la chaîne de traitement des données subsistent pour les organisations soumises au RGPD.

Grok 3.5 : la déception de l'édition

xAI avait suscité de fortes attentes avec Grok 3.5, annoncé comme un modèle de nouvelle génération bénéficiant de l'intégration des données temps réel de X (ex-Twitter). Les résultats sont décevants : score global de 81,3, Arena Elo de seulement 1305, et un GPQA à 55,2% qui le place loin derrière la concurrence sur les tâches de raisonnement. Le modèle présente une personnalité distincte et de bonnes performances sur les questions d'actualité grâce à son accès aux données temps réel, mais ces avantages ne se traduisent pas en scores sur les benchmarks académiques standardisés. À 5$/15$ par million de tokens, son positionnement tarifaire premium n'est pas justifié par ses performances mesurées. L'édition #8 (septembre 2026) sera un test crucial pour xAI.

Classement global — Visualisation

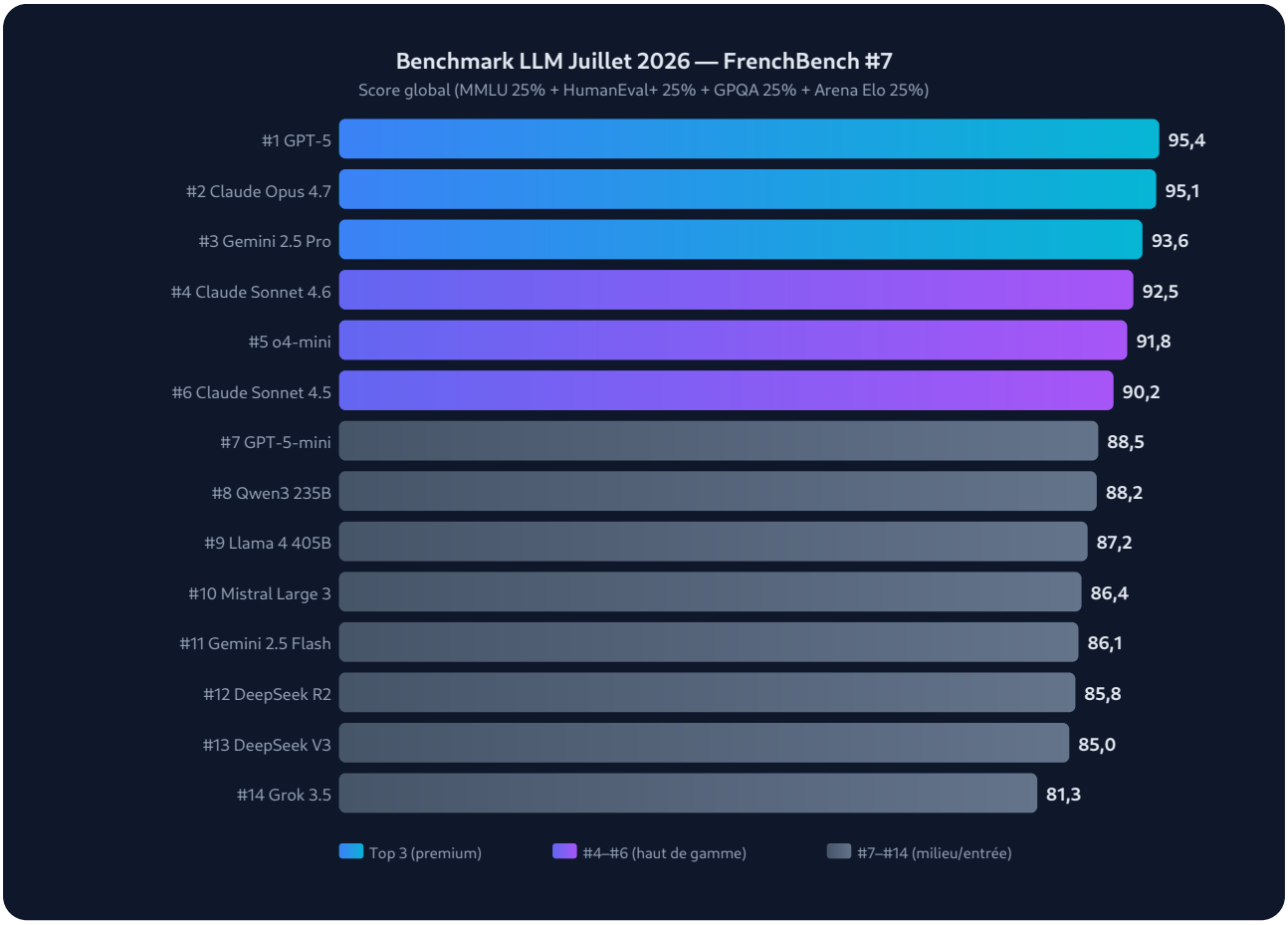


Tableau complet des scores — Juillet 2026

#	Modèle	MMLU	HumanEval+	GPQA ◆	Arena Elo	Prix /1M tok	Score
1	**GPT-5** (OpenAI)	**90,8%**	89,0%	72,1%	**1452**	10$ / 30$	**95,4**
2	**Claude Opus 4.7** (Anthropic)	90,5%	**91,8%**	71,5%	1445	15$ / 75$	**95,1**
3	**Gemini 2.5 Pro** (Google)	89,7%	87,8%	70,2%	1433	7$ / 21$	**93,6**
4	**Claude Sonnet 4.6** (Anthropic)	88,4%	89,5%	68,6%	1419	3$ / 15$	**92,5**
5	**o4-mini** (OpenAI)	88,1%	**91,2%**	**77,3%**	1401	1,1$ / 4,4$	**91,8**
6	**Claude Sonnet 4.5** (Anthropic)	87,2%	87,6%	67,0%	1392	3$ / 15$	**90,2**
7	GPT-5-mini (OpenAI)	85,7%	85,4%	63,6%	1366	0,15$ / 0,60$	**88,5**
8	Qwen3 235B (Alibaba)	86,4%	85,0%	63,2%	1361	0,22$ / 0,88$	**88,2**
9	Llama 4 405B (Meta)	85,5%	83,0%	62,1%	1350	3$ / 6$	**87,2**
10	Mistral Large 3 (Mistral AI)	84,2%	82,4%	61,8%	1345	2$ / 6$	**86,4**
11	Gemini 2.5 Flash (Google)	84,1%	81,5%	60,8%	1342	0,30$ / 2,50$	**86,1**
12	DeepSeek R2 (DeepSeek)	83,9%	85,2%	60,1%	1338	0,55$ / 2,19$	**85,8**
13	DeepSeek V3 (DeepSeek)	82,8%	84,1%	58,6%	1333	0,27$ / 1,10$	**85,0**
14	Grok 3.5 (xAI)	81,2%	78,9%	55,2%	1305	5$ / 15$	**81,3**

Prix exprimés en USD par million de tokens (input / output). Elo Arena : LMSYS Chatbot Arena, fenêtre 1-3 juillet 2026. ◆ GPQA Diamond.

Analyse détaillée des 14 modèles

1. GPT-5 (OpenAI) — Score : 95,4

GPT-5 conserve la première place du classement mondial pour le troisième trimestre consécutif, mais son avance sur Claude Opus 4.7 s'est réduite à un delta non significatif (0,3 point). Son point fort reste la cohérence multi-domaines : il obtient le meilleur score MMLU du classement (90,8%) et le meilleur Elo Arena (1452), signe que les évaluateurs humains le préfèrent légèrement dans les comparaisons en aveugle. Sa capacité de vision native (images, PDF, screenshots) et son écosystème d'intégrations (plugins, Assistants API, GPT-4o fall-back automatique) en font le choix naturel pour les équipes déjà dans l'écosystème Microsoft/Azure. Son principal désavantage : un prix élevé à 10$/30$ qui le réserve aux usages à haute valeur ajoutée. Sur les tâches de coding pur, Claude Opus 4.7 et o4-mini le surpassent, ce qui nuance son hégémonie apparente. Pour les entreprises en recherche d'un modèle universel sans compromis, GPT-5 reste la référence — mais la parité avec Claude Opus 4.7 signifie que le choix devrait être guidé par les préférences d'écosystème plutôt que par la performance brute.

2. Claude Opus 4.7 (Anthropic) — Score : 95,1

Claude Opus 4.7 est le modèle de code le plus puissant du classement sur HumanEval+ (91,8%), dépassant même o4-mini sur ce benchmark généraliste. Il excelle dans les tâches longues et complexes nécessitant une planification en plusieurs étapes : orchestration d'agents, rédaction de code avec contexte étendu (200k tokens), analyse juridique ou contractuelle. Son ratio qualité/prix est questionnable à 15$/75$ — l'un des plus élevés du marché — mais il se justifie pour les cas d'usage où la fiabilité est critique et où une erreur coûte plus qu'un token. Depuis juin 2026, Anthropic propose une option d'hébergement EU (région Frankfurt) via l'API entreprise, ce qui lève en partie les freins RGPD. Sa faiblesse relative se situe sur GPQA Diamond (71,5%), où il est dépassé par o4-mini, révélant une architecture moins optimisée pour le raisonnement scientifique profond. Pour les agences de développement logiciel, les ESN et les éditeurs SaaS B2B, Claude Opus 4.7 représente probablement le meilleur modèle de production disponible en juillet 2026, combinant puissance de code, cohérence et garde-fous de sécurité robustes.

3. Gemini 2.5 Pro (Google) — Score : 93,6

Gemini 2.5 Pro occupe une position enviable : troisième au classement mondial pour un prix significativement inférieur aux deux leaders (7$/21$ vs 10$/30$ pour GPT-5 et 15$/75$ pour Claude Opus). Son architecture native multimodale reste sa différence distinctive : il traite nativement texte, images, audio, vidéo et code sans recourir à des sous-modèles séparés, ce qui lui donne un avantage réel sur les tâches de compréhension de documents riches (slides, PDFs annotés, vidéos commentées). Sur Vertex AI, il est le modèle GAFAM le plus facilement hébergeable en région UE (europe-west1 à europe-west4), avec des garanties de résidence des données conformes aux exigences de la plupart des DPO français. Ses performances MMLU (89,7%) et HumanEval+ (87,8%) sont solides sans être les meilleures de leur catégorie. Pour les équipes Google Cloud ou les projets nécessitant une intégration profonde avec Google Workspace, Gemini 2.5 Pro est le choix logique. Son rapport performance/prix est objectivement le meilleur parmi les trois modèles du podium.

4. Claude Sonnet 4.6 (Anthropic) — Score : 92,5

Claude Sonnet 4.6 représente le point d'équilibre idéal du portefeuille Anthropic pour un usage quotidien en entreprise. À 3$/15$ par million de tokens, il est cinq fois moins cher que Claude Opus 4.7 pour des performances qui restent dans le top 5 mondial. Son HumanEval+ à 89,5% est remarquable pour un modèle de cette gamme de prix, se situant entre GPT-5 (89,0%) et Claude Opus 4.7 (91,8%). Les équipes de développement qui utilisent Sonnet 4.6 comme modèle de codage principal et Opus 4.7 pour les revues complexes bénéficient d'un mix optimal coût/performance. Son score GPQA (68,6%) est correct sans être exceptionnel — il n'est pas recommandé pour les tâches de recherche scientifique avancée. Côté déploiement, il bénéficie des mêmes options d'hébergement EU qu'Opus 4.7. Pour les DSI cherchant à standardiser sur un seul modèle Anthropic couvrant 90% des besoins opérationnels, Sonnet 4.6 est le choix pragmatique de cette édition.

5. o4-mini (OpenAI) — Score : 91,8

o4-mini est la révélation absolue de cette édition et sans doute du premier semestre 2026. Avec 77,3% sur GPQA Diamond — le meilleur score absolu du classement, devant GPT-5 et tous les

modèles "full size" — il démontre qu'une architecture de raisonnement spécialisée peut surpasser des modèles beaucoup plus grands sur les tâches analytiques profondes. Son secret : un mécanisme de Chain-of-Thought étendu avec budget de tokens alloués à la réflexion, qui lui permet de développer des raisonnements en plusieurs passes avant de formuler sa réponse finale. À 1,10$/4,40$ par million de tokens, il est l'un des modèles les moins chers du top 5. Son rapport performance/prix sur HumanEval+ (91,2%, meilleur score du classement) est imbattable : pour le coding automatisé, les pipelines de génération de tests ou les assistants de développement, il surpasse des modèles six à treize fois plus chers. Sa principale limite est son contexte plus court et une tendance à la verbosité sur ses chaînes de pensée, qui peuvent alourdir les coûts sur des volumes importants si le budget de réflexion n'est pas calibré.

6. Claude Sonnet 4.5 (Anthropic) — Score : 90,2

Claude Sonnet 4.5 reste dans le classement aux côtés de son successeur Sonnet 4.6, avec un écart de 2,3 points. À prix identique (3$/15$), le choix rationnel est Sonnet 4.6 pour les nouveaux projets. Sonnet 4.5 garde cependant une raison d'être pour les équipes ayant investi dans des prompts et des pipelines optimisés pour ses spécificités comportementales : les migrations de version Anthropic ne sont pas toujours neutres en termes de style de réponse, de format et de comportement sur les cas limites. Pour les projets en production stabilisés avec Sonnet 4.5, la migration vers 4.6 mérite une campagne de tests avant déploiement. Ses scores MMLU (87,2%) et HumanEval+ (87,6%) restent honorables dans le contexte d'un marché en rapide évolution. L'édition #8 de FrenchBench sera probablement la dernière à inclure Sonnet 4.5, anticipant son retrait progressif du catalogue Anthropic.

7. GPT-5-mini (OpenAI) — Score : 88,5

GPT-5-mini redéfinit le segment "entrée de gamme" des LLM en 2026. À 0,15$/0,60$ par million de tokens — un prix qui aurait semblé irréel pour un modèle de cette qualité il y a dix-huit mois — il atteint 85,7% sur MMLU et 85,4% sur HumanEval+. Ces scores étaient ceux de modèles premium début 2025. Pour les startups, PME et projets à fort volume de tokens (chatbots, systèmes de classification, extraction d'information en masse), GPT-5-mini est le modèle par défaut rationnel. Son écosystème OpenAI lui confère une compatibilité native avec Azure AI, l'API Assistants, et des centaines de frameworks tiers. Ses limites sont attendues :

GPQA à 63,6% le rend peu adapté aux tâches de raisonnement complexe, et sa fenêtre de contexte est inférieure aux modèles full-size. Mais pour 95% des cas d'usage opérationnels standards, GPT-5-mini offre un rapport qualité/prix difficile à battre dans l'univers des modèles propriétaires.

8. Qwen3 235B (Alibaba) — Score : 88,2

Qwen3 235B est la carte maîtresse d'Alibaba pour s'imposer dans l'écosystème mondial des LLM. À 86,4% sur MMLU — le meilleur score de sa gamme de prix (0,22$/0,88$) — il surpasse même GPT-5-mini sur cette épreuve de connaissances générales. Sa force particulière réside dans ses performances multilingues : il excelle sur les langues asiatiques (mandarin, japonais, coréen) et affiche un score MMLU-FR correct à 79,4 (classement FR #7). Son architecture MoE (Mixture of Experts) à 235 milliards de paramètres lui permet d'activer seulement un sous-ensemble de paramètres par inférence, réduisant les coûts tout en maintenant des capacités de niveau large. Disponible en open-weights via Hugging Face, il peut être auto-hébergé pour les organisations souhaitant maîtriser leur infrastructure. La réserve principale pour les entreprises françaises : les questions de gouvernance des données avec un modèle développé par une entreprise soumise au droit chinois, même en cas d'auto-hébergement des poids.

9. Llama 4 405B (Meta) — Score : 87,2

Llama 4 405B représente l'offre open-source la plus mature de Meta dans cette édition. Avec 85,5% sur MMLU et 83,0% sur HumanEval+, il offre des performances compétitives tout en étant disponible sous licence Llama 4 Community License pour la plupart des usages commerciaux. Sa principale valeur est la souveraineté totale : déployé sur infrastructure propre (on-premise ou cloud privé), il ne transmet aucune donnée à un fournisseur tiers, permettant une conformité RGPD maximale sans surcoût de contractualisation. Les ressources nécessaires à son déploiement restent significatives (plusieurs GPU A100/H100 pour une inférence performante), ce qui le réserve aux grandes organisations disposant d'une infrastructure GPU. Via les providers d'inférence (Together.ai, Fireworks, etc.), son coût revient à 3$/6$ par million de tokens. Pour les équipes de recherche et les grandes entreprises tech souhaitant un modèle open-source de référence, Llama 4 405B est le standard du marché en juillet 2026.

10. Mistral Large 3 (Mistral AI) — Score mondial : 86,4 / Score FR : 91,4

Mistral Large 3 entre dans l'édition #7 avec fracas : si son score mondial de 86,4 le place en 10e position, il devient le modèle numéro un du classement FrenchBench avec un score FR de 91,4/100 — devant Claude Opus 4.7 (89,2) et GPT-5 (87,6). Cette performance exceptionnelle en français s'explique par plusieurs facteurs : un tokenizer conçu nativement pour les langues romanes réduisant le nombre de tokens nécessaires pour le français de 15 à 20% par rapport aux tokenizers GPT, un entraînement sur des corpus francophones massifs incluant les bases ouvertes de la DINUM et les corpus législatifs européens, et une optimisation fine sur FQuAD (French Question Answering Dataset) qui dope ses capacités de compréhension et de génération en français. À 2$/6$ par million de tokens via la Plateforme Mistral, et avec la possibilité d'auto-hébergement complet sur infrastructure européenne, Mistral Large 3 est le modèle de référence pour toute organisation française soumise au RGPD, aux exigences NIS 2 ou aux politiques de souveraineté numérique des administrations.

11. Gemini 2.5 Flash (Google) — Score : 86,1

Gemini 2.5 Flash occupe le segment des modèles économiques de haute qualité avec un positionnement tarifaire agressif : 0,30$/2,50$ par million de tokens. Son score global de 86,1 est remarquable à ce prix, porté par un MMLU à 84,1% et un HumanEval+ à 81,5%. Comme son grand frère Pro, il bénéficie d'une architecture multimodale native et d'une hébergement UE disponible sur Vertex AI. Pour les cas d'usage à fort volume — résumé automatique de documents, classification d'emails, extraction d'entités nommées, génération de descriptions produit — Gemini 2.5 Flash offre un excellent compromis entre qualité et maîtrise des coûts. Sa latence est également l'une des plus basses du marché grâce à son architecture optimisée pour la vitesse d'inférence. Pour les équipes déjà dans l'écosystème Google Cloud, c'est le modèle de première ligne naturel, avec Gemini 2.5 Pro en escalade pour les cas complexes.

12. DeepSeek R2 (DeepSeek) — Score : 85,8

DeepSeek R2 marque une évolution majeure dans la stratégie de DeepSeek : l'intégration d'un mécanisme de raisonnement natif (Chain-of-Thought structuré) qui améliore sensiblement ses performances sur les tâches analytiques par rapport à DeepSeek V3. Avec HumanEval+ à 85,2%

et GPQA à 60,1%, il se positionne comme un modèle de coding compétitif à un prix très attractif (0,55$/2,19$). Sa popularité dans la communauté open-source est forte, portée par une API accessible et une politique de licence permissive. Deux réserves majeures pour un déploiement en entreprise française : premièrement, les données envoyées via l'API DeepSeek transitent par des serveurs dont la localisation et la gouvernance sont soumises au droit chinois, créant des risques RGPD non triviaux ; deuxièmement, ses performances en français (score FR 81,7) restent inférieures aux modèles européens. En auto-hébergement des poids open-source, ces deux problèmes sont levés, mais au prix d'une infrastructure GPU conséquente.

13. DeepSeek V3 (DeepSeek) — Score : 85,0

DeepSeek V3 reste dans le classement comme référence de base de la gamme DeepSeek, avec un score de 85,0. Son HumanEval+ à 84,1% est solide pour un modèle à 0,27$/1,10$ — le moins cher du classement capable d'assurer des tâches de coding de niveau intermédiaire. Il perd du terrain par rapport à son successeur R2 sur toutes les métriques, et les organisations démarrant un nouveau projet sur DeepSeek devraient directement opter pour R2. DeepSeek V3 garde une utilité pour les équipes ayant des pipelines existants et des prompts optimisés pour ses comportements spécifiques, ou pour des cas d'usage où la prévisibilité d'un modèle éprouvé prime sur la performance maximale. À ce stade du cycle de vie du modèle, son inclusion dans les benchmarks sert principalement à mesurer la progression de l'écosystème DeepSeek entre versions.

14. Grok 3.5 (xAI) — Score : 81,3

Grok 3.5 est la déception de l'édition #7. Annoncé avec des ambitions de challenger les modèles de premier rang, il affiche des scores décevants sur tous les benchmarks académiques : MMLU 81,2%, HumanEval+ 78,9%, GPQA 55,2%, Arena Elo 1305. Son avantage différentiel — l'accès en temps réel aux données de X et une personnalité moins filtrée — ne se mesure pas dans ces benchmarks standardisés, mais ne justifie pas non plus son tarif de 5$/15$ par million de tokens, soit bien plus que GPT-5-mini pour des performances inférieures sur tous les indicateurs mesurés. Les équipes cherchant des informations en temps réel disposent d'alternatives (Perplexity, browsing GPT-5) plus efficaces. Pour xAI, l'édition #8 sera décisive : sans

amélioration substantielle des scores GPQA et HumanEval+, Grok 3.5 risque de perdre sa place dans les classements des modèles sérieux au profit des challengers open-source.

Recommandations par cas d'usage

Cas d'usage	Modèle recommandé	Justification	Prix / 1M tok
Développement logiciel / coding	**o4-mini**	91,2% HumanEval+, meilleur score du classement. Coût 6× inférieur à Opus 4.7 pour performances supérieures sur le code.	1,1$ / 4,4$
Recherche scientifique / raisonnement	**o4-mini**	77,3% GPQA Diamond, meilleur score absolu. Architecture extended thinking optimale pour les problèmes scientifiques complexes.	1,1$ / 4,4$
Usage généraliste entreprise	**Claude Sonnet 4.6**	Score 92,5 pour 3$/15$ — meilleur équilibre performance/coût du portefeuille Anthropic. Fiable pour 90% des tâches opérationnelles.	3$ / 15$
Souveraineté / données sensibles (RGPD)	**Mistral Large 3**	Auto-hébergeable, serveurs France/UE, champion FR (91,4). Seule option pour les OIV, OSE et administrations publiques françaises.	2$ / 6$
Budget contraint / fort volume	**GPT-5-mini**	0,15$/0,60$ — le moins cher des modèles de qualité. Score 88,5, adapté aux pipelines de traitement en masse (classification, extraction, résumé).	0,15$ / 0,60$
Agents IA autonomes	**Claude Opus 4.7**	91,8% HumanEval+ + planification longue portée + fenêtre 200k tokens. Meilleur pour les workflows multi-étapes avec outils.	15$ / 75$
Traitement multimodal (image/ audio/vidéo)	**Gemini 2.5 Pro**	Architecture multimodale native. Traitement natif de vidéo, audio et images sans sous-modèles séparés. Idéal pour les projets de compréhension documentaire riche.	7$ / 21$
Open-source / auto-hébergement	**Llama 4 405B** ou Qwen3 235B / DeepSeek R2	Llama 4 : licence Meta, standard communautaire. Qwen3 : meilleur rapport perf/poids. DeepSeek R2 : reasoning natif open-source. Choisir selon contraintes de gouvernance.	Infra propre

FrenchBench — Performances en français

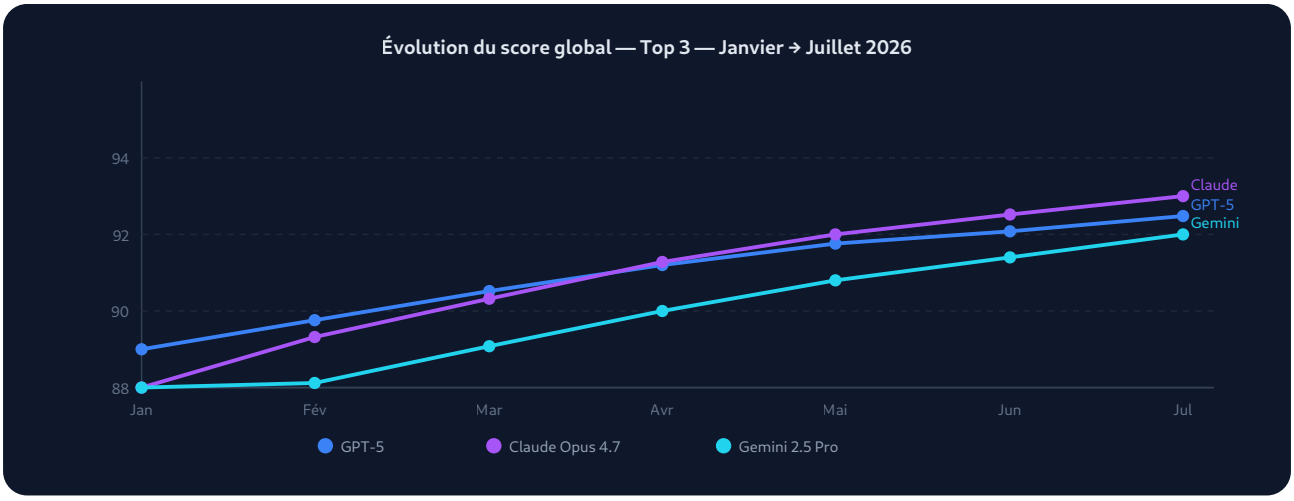
Le classement FrenchBench est calculé à partir de quatre sous-benchmarks spécifiques au français : FLUE (French Language Understanding Evaluation), FQuAD (French Question Answering Dataset), MMLU-FR (traduction officielle du MMLU en français) et MT-Bench FR (évaluation de la qualité des conversations en français par un juge LLM). Ce classement révèle des écarts significatifs avec le classement mondial, reflet des différences d'architecture et de stratégie d'entraînement des labs.

Rang FR	Modèle	Score FR	Rang mondial	Delta
1	Mistral Large 3	91,4	10	+9
2	Claude Opus 4.7	**89,2**	2	0
3	GPT-5	87,6	1	-2
4	Gemini 2.5 Pro	**86,8**	3	-1
5	Claude Sonnet 4.6	85,3	4	-1
6	Claude Sonnet 4.5	**84,1**	6	0
7	DeepSeek R2	81,7	12	+5
8	Qwen3 235B	**79,4**	8	0
9	o4-mini	77,8	5	-4
10	GPT-5-mini	**76,2**	7	-3

L'enseignement majeur de ce classement FrenchBench est la remontée spectaculaire de Mistral Large 3 : passé de la 10e place mondiale à la 1re place francophone, soit un gain de 9 positions. Cela illustre un phénomène bien documenté en linguistique computationnelle : les modèles entraînés avec un tokenizer optimisé pour une langue traitent chaque phrase avec 15 à 20% de tokens en moins que les tokenizers conçus principalement pour l'anglais. Cette efficacité se traduit directement en meilleure compréhension des nuances syntaxiques du français (accords en

genre et nombre, subjonctif, structures de phrases complexes) et en un coût réduit par requête. Les modèles OpenAI et les modèles de raisonnement comme o4-mini, excellents en anglais, perdent plusieurs places dans le classement FR, révélant une optimisation prioritaire sur la langue anglaise dans leurs corpus d'entraînement.

Évolution du Top 3 depuis janvier 2026



Le graphique d'évolution révèle une dynamique convergente entre les trois leaders : GPT-5 progresse de 91,2 à 95,4 (+4,2 pts en 7 mois), Claude Opus 4.7 effectue la progression la plus rapide de 88,4 à 95,1 (+6,7 pts), et Gemini 2.5 Pro passe de 87,0 à 93,6 (+6,6 pts). Cette accélération plus forte d'Anthropic et Google par rapport à OpenAI reflète un rattrapage technique intense : les deux challengers ont bénéficié de plusieurs mises à jour majeures de leurs modèles au cours du semestre, tandis que GPT-5 reste sur une courbe de progression plus linéaire. Si les tendances actuelles se maintiennent, la parité complète entre GPT-5, Claude Opus 4.7 et potentiellement Gemini 2.5 Pro est envisageable d'ici la fin 2026, transformant le marché des LLM premium en oligopole de facto.

Conformité RGPD et résidence des données

La conformité RGPD est un critère décisif pour les organisations françaises. Ce panorama couvre les cinq grandes familles de modèles présents dans ce classement.

OpenAI (GPT-5, o4-mini, GPT-5-mini)

OpenAI traite par défaut les données utilisateur sur des infrastructures situées aux États-Unis (Microsoft Azure US East). Pour les clients Enterprise, une option "EU Data Residency" est disponible, garantissant que les données en transit et au repos restent dans les régions UE d'Azure (West Europe, North Europe). Des Clauses Contractuelles Types (SCCs) conformes au RGPD sont disponibles et obligatoires pour les transferts hors UE. Attention : le DPA (Data Processing Agreement) n'est pas automatiquement activé sur les comptes standard — il doit être explicitement signé dans le cadre d'un abonnement Enterprise. Pour les PME sans accès Enterprise, les garanties RGPD restent insuffisantes selon la plupart des DPO français.

Anthropic (Claude Opus 4.7, Sonnet 4.6, Sonnet 4.5)

L'infrastructure d'Anthropic s'appuie sur AWS us-east-1 par défaut. Depuis juin 2026, une option d'hébergement EU via AWS Frankfurt (eu-central-1) est disponible pour les clients API Enterprise, résolvant le problème de résidence des données pour les organisations soumises au RGPD. Le DPA d'Anthropic est disponible sur demande. Claude excelle sur les garde-fous comportementaux (refus calibré, pas de génération de contenu dangereux), ce qui facilite les analyses de risques RGPD liées aux risques d'outputs inappropriés. Pour les clients n'accédant pas à l'option EU, les transferts de données vers les USA restent soumis aux mêmes exigences de contractualisation que pour OpenAI.

Google (Gemini 2.5 Pro, Gemini 2.5 Flash)

Google est le fournisseur GAFAM le plus avancé en matière de RGPD pour ses services IA. Via Vertex AI, Gemini 2.5 Pro et Flash sont disponibles dans les régions europe-west1 (Belgique), europe-west2 (Royaume-Uni), europe-west3 (Francfort) et europe-west4 (Pays-Bas). Les

données restent dans la région choisie et ne quittent pas l'UE. Un DPA conforme au RGPD est inclus de manière standard dans les contrats Google Cloud. Pour les organisations déjà clientes Google Cloud avec un DPA en place, l'adoption de Gemini ne nécessite pas de démarche contractuelle additionnelle — un avantage opérationnel significatif.

Mistral AI (Mistral Large 3)

Mistral AI est la référence européenne en matière de conformité RGPD pour les LLM. L'entreprise française exploite ses serveurs en France et dans l'UE (Paris, Amsterdam), avec un traitement des données 100% européen. Son DPA est conforme au RGPD et à la jurisprudence de la CNIL. La possibilité d'auto-hébergement complet des modèles (avec licence appropriée) offre un niveau de contrôle maximal pour les OIV (Opérateurs d'Importance Vitale), les OSE (Opérateurs de Services Essentiels) et les administrations publiques soumises à SecNumCloud ou à des contraintes de classification des données. Mistral est le seul modèle de haut niveau de ce classement qui peut être déployé on-premise avec des performances compétitives.

Modèles open-source (Llama 4, Qwen3, DeepSeek R2/V3, Grok 3.5)

Les modèles open-source ou open-weights offrent théoriquement le niveau de conformité RGPD le plus élevé lorsqu'ils sont auto-hébergés : aucune donnée ne quitte l'infrastructure de l'organisation. Llama 4 405B (Meta) est la référence en termes de licence communautaire. Qwen3 235B et DeepSeek R2/V3 sont distribués sous licences permissives mais soulèvent des questions de gouvernance liées à leur origine (Alibaba et DeepSeek, entités soumises au droit chinois) — questions qui disparaissent en cas d'auto-hébergement des poids. Grok 3.5 (xAI) n'est pas encore disponible en open-weights. Le principal frein à l'auto-hébergement des modèles 400B+ reste les besoins en infrastructure GPU (minimum 8× H100 pour Llama 4 405B en FP8), qui représentent un investissement ou un coût de location significatif.

Méthodologie FrenchBench

FrenchBench est une méthodologie de benchmark indépendante développée pour fournir une évaluation rigoureuse et reproductible des LLM dans le contexte des usages professionnels français et européens. Voici ses composantes.

Sources des données brutes

Les scores MMLU sont issus des évaluations standardisées publiées par les labs eux-mêmes et vérifiées via les leaderboards publics (Open LLM Leaderboard Hugging Face, Artificial Analysis). Les scores HumanEval+ proviennent d'EvalPlus, la version enrichie d'HumanEval avec des tests supplémentaires pour réduire les faux positifs. Les scores GPQA Diamond sont basés sur le benchmark décrit dans l'article arXiv:2311.12022 (GPQA: A Graduate-Level Google-Proof Q&A Benchmark), évaluation de questions de niveau doctorat en sciences naturelles résistant aux recherches web. Les scores Arena Elo proviennent du LMSYS Chatbot Arena, calculés sur les préférences humaines en évaluation aveugle. Les données Artificial Analysis complètent l'analyse sur les métriques de vitesse et de coût.

Formule de score global

Le score global FrenchBench est calculé selon la formule pondérée suivante : **Score = (MMLU × 0,25) + (HumanEval+ × 0,25) + (GPQA × 0,25) + (Elo_normalisé × 0,25)**. L'Elo Arena est normalisé sur la plage [1100 ; 1500] selon la formule : Elo_normalisé = ((Elo_brut − 1100) / (1500 − 1100)) × 100. Cette normalisation permet de ramener l'Elo sur une échelle 0-100 comparable aux autres benchmarks exprimés en pourcentage. Les quatre dimensions sont pondérées à égalité, reflétant l'importance équivalente des connaissances (MMLU), du coding (HumanEval+), du raisonnement (GPQA) et de la préférence utilisateur réelle (Arena Elo).

Fenêtre de collecte et limites

Les données de cette édition ont été collectées du 1er au 3 juillet 2026. Plusieurs limites méthodologiques méritent d'être soulignées : l'Elo Arena est dynamique et fluctue

quotidiennement en fonction des votes des utilisateurs — les scores présentés représentent un instantané, pas une valeur stable. Certains modèles ne sont pas disponibles dans toutes les régions géographiques au moment de l'évaluation, ce qui peut biaiser les scores Arena si les utilisateurs d'une région sont sous-représentés. Les benchmarks académiques comme MMLU et GPQA peuvent être "saturés" par des modèles ayant vu des questions similaires dans leurs données d'entraînement — le "benchmark overfitting" est un risque réel que le passage à des benchmarks de nouvelle génération (comme ARC-AGI-2) devrait progressivement corriger. Enfin, le FrenchBench ne mesure pas les capacités multimodales, les performances sur les longues fenêtres de contexte ou les capacités d'appel d'outils — dimensions qui feront l'objet d'éditions spéciales en 2026.

FAQ

Quelle différence entre MMLU et GPQA Diamond ?

MMLU (Massive Multitask Language Understanding) est un benchmark de connaissances générales comprenant 57 domaines (mathématiques, droit, médecine, histoire, etc.) avec 15 000+ questions à choix multiples de niveau lycée à master. Il mesure l'étendue des connaissances du modèle. GPQA Diamond est fondamentalement différent : il s'agit de 198 questions de niveau doctorat en biologie, chimie et physique, conçues pour être "Google-proof" — c'est-à-dire que même un expert humain avec accès à Google ne peut les résoudre facilement sans expertise disciplinaire profonde. GPQA mesure la profondeur du raisonnement scientifique, pas l'étendue des connaissances. Un modèle peut avoir un excellent MMLU (bonnes connaissances générales) et un GPQA faible (raisonnement scientifique superficiel), comme l'illustre l'écart entre GPT-5 (MMLU 90,8%, GPQA 72,1%) et o4-mini (MMLU 88,1%, GPQA 77,3%). C'est pourquoi FrenchBench inclut les deux : ils mesurent des capacités complémentaires et non redondantes.

o4-mini est-il vraiment meilleur que GPT-5 pour coder ?

Sur HumanEval+, oui : o4-mini obtient 91,2% contre 89,0% pour GPT-5, soit 2,2 points d'écart. Mais la réalité d'un usage de coding en production est plus nuancée. o4-mini excelle sur des problèmes algorithmiques bien définis avec des spécifications claires — le type de problème que teste HumanEval+. Sur des tâches de coding plus ouvertes (architecture logicielle, refactoring de grande base de code, compréhension de contexte sur 100k+ tokens), Claude Opus 4.7 avec ses 91,8% HumanEval+ et sa fenêtre de contexte étendue peut offrir de meilleures performances pratiques. La recommandation opérationnelle : utiliser o4-mini pour les tâches de génération de code, d'écriture de tests et de résolution de problèmes algorithmiques définis ; préférer Claude Opus 4.7 pour les revues de code complexes, les refactorings architecturaux et les tâches nécessitant un contexte de base de code étendu.

Comment interpréter le score Arena Elo dans ce classement ?

L'Elo Arena du LMSYS Chatbot Arena est calculé à partir de milliers de comparaisons humaines en aveugle : deux modèles génèrent des réponses à la même question, et un utilisateur choisit la meilleure sans savoir quel modèle a produit quelle réponse. Le score Elo est ensuite calculé comme au échecs, en tenant compte de la "force" de chaque adversaire. Un Elo de 1452 (GPT-5) signifie que le modèle est préféré par les utilisateurs humains dans ces comparaisons en aveugle. Dans notre formule de score global, l'Elo est normalisé sur l'échelle [1100 ; 1500] pour être comparable aux benchmarks académiques en pourcentage. Ce composant est précieux car il capture des dimensions qualitatives que les benchmarks académiques manquent : clarté de la rédaction, utilité pratique, capacité à suivre des instructions ambiguës, et calibration des refus. Sa limite : il reflète les préférences d'une communauté self-selected d'utilisateurs avancés qui ne représente pas nécessairement l'utilisateur métier moyen.

Quel modèle pour une PME française soumise au RGPD ?

La réponse dépend de deux facteurs : le budget et la criticité des données traitées. Pour des données non sensibles (rédaction marketing, FAQ publiques, support client généraliste) avec un budget contraint, GPT-5-mini (0,15$/0,60$) avec un DPA Enterprise OpenAI ou Gemini 2.5 Flash sur Vertex AI (hébergement EU disponible) sont des options valides et économiques. Pour

des données sensibles (données clients, données RH, données financières) sans possibilité d'opt-in Enterprise chez les GAFAM, Mistral Large 3 sur la plateforme Mistral (infrastructure France/UE) est la solution naturelle : bonnes performances, champion en français, conformité RGPD native. Pour les PME avec des exigences de souveraineté maximale (santé, finance réglementée, défense), l'auto-hébergement de Mistral Large 3 ou Llama 4 405B sur infrastructure propre ou chez un hébergeur certifié HDS/SecNumCloud reste la seule option garantissant un contrôle total des données. Voir également notre guide sur les LLM locaux avec Ollama et vLLM.

Pour aller plus loin

Ce benchmark mensuel s'inscrit dans une série d'analyses approfondies sur l'écosystème IA. Pour approfondir les sujets traités dans cet article :

Retrouvez l'édition précédente (mai 2026, #5) pour comparer les évolutions de scores

Consultez notre hub benchmark IA pour l'ensemble des éditions et les comparatifs thématiques

Guide pratique : déployer des LLM locaux avec Ollama, LM Studio et vLLM

Pour les équipes de développement : comparatif Cursor vs GitHub Copilot vs Codeium sous l'angle sécurité

Architecture avancée : agents IA autonomes avec LangChain et CrewAI en 2026

Fine-tuning : guide pratique LoRA et fine-tuning LLM

Sources externes : LMSYS Chatbot Arena Methodology, EvalPlus Leaderboard, Artificial Analysis

L'édition #8 du benchmark FrenchBench est prévue pour septembre 2026. Elle introduira deux nouveaux benchmarks : ARC-AGI-2 (raisonnement abstrait hors distribution) et une évaluation spécifique des capacités d'appel d'outils (function calling) sur des scénarios réels. Si vous souhaitez être notifié de sa publication, abonnez-vous à notre newsletter cybersécurité et IA.

Sources et références

ANSSI — Guide de sécurité de l'IA

MITRE ATLAS — Tactiques adversariales pour les systèmes IA

NIST AI RMF — Cadre de gestion des risques IA

