



Benchmarks IA : LLMs, Image, Vidéo et Agents

Comment évaluer et comparer les modèles de fondation entre
2024 et 2026

LLMs

Vision

Vidéo

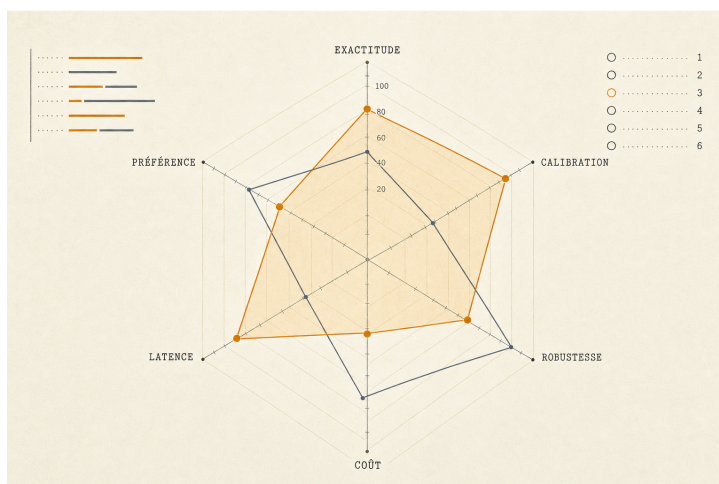
Agents

Tool use

Long contexte

Contamination

Usage réel



Entre 2024 et 2026, l'évaluation des modèles de fondation a profondément changé. Les benchmarks historiques (MMLU, HumanEval, BIG-bench) restent indispensables comme repères, mais ils sont saturés, contaminés ou supplantés par des suites plus difficiles, plus dynamiques et plus spécialisées : MMLU-Pro, MMLU-CF, LiveBench, BFCL, BrowseComp, LiveCodeBench, SWE-bench. La comparaison moderne combine plusieurs axes : exactitude, calibration, robustesse, résistance à la contamination, préférence humaine, coût, latence et reproductibilité.

Pourquoi les benchmarks ont changé entre 2024 et 2026

L'évaluation des modèles de fondation a nettement changé de nature en deux ans. Les benchmarks historiques (MMLU, HumanEval, BIG-bench...) restent la langue commune, mais ils sont saturés, contaminés, parfois entachés d'erreurs de labels — et surtout ils ne disent plus grand-chose des modèles de pointe.

Une nouvelle génération les a complétés ou remplacés : MMLU-Pro et BBEH durcissent les anciens, MMLU-CF et LiveBench protègent contre la contamination, SuperGPQA élargit les disciplines, BFCL et SWE-bench mesurent l'usage réel d'outils et la résolution d'issues GitHub, BrowseComp et ~~AgentBench~~ évaluent les agents web.

Le message méthodologique est clair : **un seul score ne suffit plus**. Une comparaison crédible combine exactitude, calibration, robustesse, préférence humaine, coût, latence et reproductibilité — et, pour les agents, réussite de tâche et respect de contraintes.

Ce qu'il faut retenir

- Les benchmarks historiques (GLUE, SuperGLUE, MMLU, HumanEval, TruthfulQA, BIG-bench) restent la langue commune des papiers et des system cards, mais sont largement saturés pour les modèles de pointe.
- Les nouveaux benchmarks visent explicitement à corriger les défauts des anciens : saturation (MMLU-Pro, BBEH), erreurs de labels (MMLU-Redux), contamination (MMLU-CF, LiveBench), biais disciplinaire (SuperGPQA).
- Les leaderboards se sont professionnalisés : Open LLM Leaderboard v2, HELM, lm-evaluation-harness, OpenCompass, BFCL, LiveCodeBench, SWE-bench, et les arènes de préférences humaines (Chatbot Arena / LMArena).
- Des hubs indépendants agrègent qualité, prix et vitesse à travers les fournisseurs (Artificial Analysis, Vellum, Epoch AI), tandis qu'OpenRouter Rankings expose la préférence révélée — quels modèles sont réellement consommés en production.
- Les évaluations privées anti-contamination (SEAL de Scale AI) et les arènes spécialisées (WebDev Arena, Copilot Arena, Aider Polyglot, Search Arena) prennent de l'importance face à la saturation des benchmarks publics.
- Pour l'image, la vidéo et les agents, il n'existe pas d'équivalent unique de MMLU : on compare des familles de tâches (perception, OCR, raisonnement visuel, génération vidéo, agents web).
- Une comparaison crédible doit combiner exactitude, calibration, robustesse, préférence humaine, coût, latence, reproductibilité et — pour les agents — réussite de tâche et respect de contraintes.

Vocabulaire : ne pas tout appeler « benchmark »

Dans les conversations sur « les benchmarks », quatre objets distincts sont régulièrement confondus. Les distinguer évite la plupart des malentendus. Un benchmark est un **jeu de tâches** ; une suite est un **outillage logiciel** qui exécute ces tâches ; un leaderboard **publie** des comparaisons ; les model-card evals sont des batteries **internes** au fournisseur.

Quatre objets à ne pas confondre

Benchmark / dataset d'évaluation

Ensemble de tâches et de réponses de référence, comme MMLU, HumanEval, EgoSchema ou VQAv2. C'est l'unité de base : un jeu fixe de questions/cas et une métrique associée.

Reproductible, comparable dans le temps, public

Peut se saturer, être contaminé, ou contenir des erreurs de labels

Suite / framework d'évaluation

Logiciel qui exécute et normalise des benchmarks, comme HELM (Stanford), lm-evaluation-harness (EleutherAI) ou OpenCompass. Ce n'est pas un benchmark mais un outillage.

Standardise prompts, métriques, harness ; rend les résultats reproductibles

Le choix du bundle de tâches reste subjectif

Leaderboard

Publication continue de résultats comparables : Open LLM Leaderboard v2, BFCL v4, LiveCodeBench, SWE-bench, Chatbot Arena. Souvent dynamique, parfois rafraîchi.

Vue d'ensemble du paysage, mise à jour régulière

Score agrégé selon un bundle implicite ; sensible au gaming

Model-card evals

Batteries d'évaluations publiées par un fournisseur (system cards GPT-4o, GPT-4.1, o3/o4-mini d'OpenAI). Utiles pour comparer dans une même série.

Détaillés, alignés sur les capacités annoncées

Pas équivalents à un benchmark indépendant ; choix de protocole opaque

Les benchmarks LLM historiques

Ces benchmarks ne suffisent plus seuls à départager les meilleurs modèles, mais ils restent **indispensables comme repères** : ils forment la langue commune des papiers, des leaderboards et des system cards. Les ignorer, c'est se priver de toute comparaison historique avec les générations précédentes.

Benchmarks LLM historiques

Toujours cités, mais largement insuffisants seuls pour départager les meilleurs modèles en 2026.

GLUE / SuperGLUE

Saturé

9 tâches NLU (entailment, sentiment, similarité). Largement saturés avant 2024. Repère historique et point d'entrée pédagogique.

'[GLUE \(ACL 2018\)](#) · [SuperGLUE](#)

MMLU

Historique

57 matières académiques en QCM. Toujours reporté dans les model cards 2024-2025, mais peu discriminant pour les meilleurs modèles. ~6,5% d'erreurs de labels selon MMLU-Redux.

'[Paper MMLU](#)

HumanEval

Historique

Fonctions Python à partir de docstrings, métrique pass@k. Standard historique du code, mais peu représentatif du repo-level coding moderne.

'[Codex paper](#) · [repo](#)

TruthfulQA

Utile

817 questions conçues pour piéger les fausses croyances humaines. Toujours utile pour la factualité courte, rarement seul.

'[Paper officiel](#)

BIG-bench / BBH

Saturé

204 tâches diverses (BIG-bench) ; 23 tâches dures (BBH). Omniprésent en 2024, proche de la saturation en 2025.

'[BIG-bench](#) · [BBH](#)

MT-Bench

Historique

Questions multi-tours ouvertes, jugées par un LLM. Important pour l'histoire du LLM-as-a-judge ; biais de position et de verbosité connus.

'[Paper](#)

La nouvelle génération anti-saturation

Chacun de ces benchmarks vise **explicitement à corriger un défaut** des précédents. MMLU-Pro contre la saturation, MMLU-Redux contre les erreurs de labels, MMLU-CF contre la contamination, SuperGPQA contre le biais STEM, LiveBench contre le statisme. C'est désormais ce socle qui sert à classer les modèles de pointe.

Benchmarks LLM devenus centraux en 2024-2026

Ils corrigent explicitement un défaut des anciens : saturation, contamination, erreurs de labels, biais disciplinaire ou statisme.

MMLU-Pro Standard 2024+ Version durcie de MMLU (10 options, plus de raisonnement). Chute de 16 à 33 points vs MMLU. Meilleure stabilité au prompt. Nouveau standard depuis 2024. Paper arXiv	MMLU-CF Anti-contamination Benchmark fermé, décontaminé, plus difficile. GPT-4o n'atteint que 71,9% en 0-shot dans le papier. Excellent antidote à la contamination. Paper arXiv
GPQA / SuperGPQA Expert GPQA : 448 QCM 'Google-proof' rédigés par des PhD (humains ~65%). SuperGPQA (2025) couvre 285 disciplines graduate, contre le biais STEM. GPQA · SuperGPQA	LiveBench Dynamique Questions fréquemment mises à jour, vérité vérifiable, contamination limitée. Coûteux à maintenir mais bien plus crédible dans le temps. livebench.ai
BIG-Bench Extra Hard Très difficile Successeur plus dur de BBH. Le meilleur généraliste plafonne à 9,8% ; le meilleur modèle de raisonnement à 44,8%. Paper BBEH	SimpleQA Factualité 4 326 questions factuelles courtes avec distinction réponse/refus. Très utile pour mesurer les hallucinations courtes. OpenAI blog

Là où se joue la vraie différenciation

Au-delà des grands QCM académiques, ce sont les benchmarks spécialisés qui apportent le plus de signal pratique : qualité de code (LiveCodeBench, SWE-bench), usage d'outils (BFCL), longs contextes (LongBench v2, RULER), et préférence conversationnelle (Arena, AlpacaEval). C'est ici qu'un modèle gagne ou perd un déploiement produit.

Code, contexte long, tool use et préférence

C'est ici que se joue la vraie différenciation entre modèles de pointe.

LiveCodeBench

Code dynamique

Problèmes de programmation publiés dans le temps, contamination limitée. Bien plus représentatif du progrès réel en code que HumanEval seul.

'livecodebench.github.io

SWE-bench (et variantes)

Agent logiciel

Issues GitHub réelles dans des dépôts Python ; SWE-bench Verified (500 cas validés), Multilingual (9 langages, 42 repos), Multimodal (517 issues avec captures). En 2026, OpenAI juge SWE-bench Verified trop contaminé pour le frontier coding.

'swebench.com · [OpenAI 2026](#)

BFCL v4 (Berkeley Function Calling)

Tool use

Appels de fonctions réels, scénarios single/multi-turn, métriques AST + exécution + coût/latence. Référence majeure du tool use depuis 2024.

'[Leaderboard Berkeley](#)

LongBench v2 / RULER / InfiniteBench

Long contexte

Benchmarks de contexte long. LongBench v2 : 503 QCM très difficiles, 8k à 2M mots ; humains experts à 53,7%, o1-preview à 57,7%. RULER isole la 'vraie' taille de contexte ; InfiniteBench cible >100k tokens.

'[LongBench v2](#) · [RULER](#) · [InfiniteBench](#)

Chatbot Arena / LMArena

Préférence

Préférences humaines pairwise sur dialogues réels (Elo / Bradley-Terry). Toujours central : excellent proxy de préférence utilisateur, moins contrôlé qu'un benchmark fermé.

'lmarena.ai

AlpacaEval 2.0 / Arena-Hard-Auto

Auto-jugé

Évaluation par juge LLM avec contrôle de longueur (AlpacaEval) ou 500 prompts difficiles (Arena-Hard-Auto). Bons corrélats d'Arena, mais auto-jugeables donc gameables.

'[AlpacaEval](#) · [Arena-Hard-Auto](#)

Frameworks et leaderboards : l'écosystème d'évaluation

Les frameworks (HELM, lm-evaluation-harness, OpenCompass) **standardisent** l'exécution des benchmarks ; les leaderboards les **publient en continu**. Sans cet outillage, deux scores du même benchmark peuvent ne pas être comparables — c'est la condition de reproductibilité de toute évaluation sérieuse.

Frameworks et leaderboards de référence

L'outillage qui rend les évaluations comparables et reproductibles

Frameworks d'évaluation

HELM
Stanford CRFM
Living benchmark multi-dimension : accuracy, calibration, robustesse, biais, toxicité, efficacité.
[Open source](#)

lm-evaluation-harness
EleutherAI
Standard de réplication pour un très grand catalogue de tâches.
[Open source](#)

OpenCompass
Shanghai AI Lab
100+ datasets, modèles open et API. Très actif en 2025-2026.
[Open source](#)

Leaderboards LLM

Open LLM Leaderboard v2
Hugging Face
Bundle IFEval, BBH, MATH, GPQA, MuSR, MMLU-Pro. Référence open models depuis juin 2024.
[Public](#)

Chatbot Arena (LMArena)
LMSYS
Préférences humaines à grande échelle, Elo.
[Public](#)

LiveBench

Indépendant

Benchmark dynamique anti-contamination.

Public

Leaderboards spécialisés

BFCL

UC Berkeley

Function calling, tool use.

Public

LiveCodeBench

Indépendant

Programmation continue, contamination limitée.

Public

SWE-bench

Princeton NLP

Résolution d'issues GitHub réelles.

Public

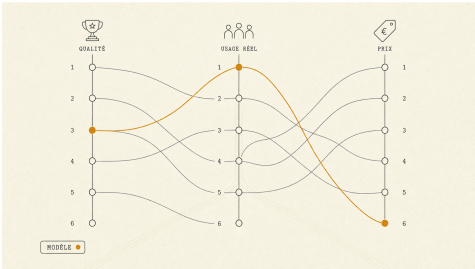
Sites de comparaison indépendants : où regarder en 2026

À côté des frameworks académiques (HELM, lm-eval-harness, OpenCompass) et des leaderboards officiels par benchmark, un écosystème de **hubs de comparaison indépendants** s'est imposé entre 2024 et 2026. Ils agrègent des dizaines de modèles à travers les fournisseurs et exposent, en un coup d'œil, qualité, prix, latence, voire popularité réelle.

Point important : **OpenRouter Rankings** mesure une chose qu'aucun benchmark ne capture — quels modèles les développeurs paient effectivement pour utiliser en production. C'est un signal de préférence révélée orthogonal aux scores.

Sites indépendants pour comparer les modèles

Hubs agrégateurs, classements d'usage réel et arènes spécialisées — complémentaires des benchmarks académiques



Hubs de comparaison qualité / prix / vitesse

Artificial Analysis

Indépendant

100+ modèles comparés sur qualité, prix, tokens/s, TTFT. Intelligence Index composite (MMLU-Pro, GPQA Diamond, HLE, LiveCodeBench, SciCode, MATH-500, AIME, IFBench). Couvre aussi image, vidéo, speech.

Public

Vellum LLM Leaderboard

Vellum

Comparaison côte-à-côte orientée praticiens : raisonnement, code, math, multimodal, agentique, fenêtre de contexte, coût par 1M tokens, date de sortie.

Public

Epoch AI Benchmarking Hub

Epoch AI

Suivi frontier models : FrontierMath, GPQA Diamond, MATH niveau 5, OTIS Mock AIME. Perspective tendances/scaling et base Notable AI Models (compute, paramètres, coût d'entraînement).

Public

Usage réel et préférence révélée

OpenRouter Rankings

OpenRouter

Classement par volume de tokens consommés via la passerelle. Catégories : Programming, Roleplay, Marketing, Technology, Science, Translation, Academia, Finance, Health... Signal de préférence révélée — ce que les développeurs paient vraiment pour utiliser.

🔴Public

Chatbot Arena (LMArena)

LMArena

Préférences humaines pairwise (Elo). Sous-arènes : WebDev Arena, Copilot Arena (in-IDE via VS Code), Search Arena, Vision Arena, RepoChat Arena, Text-to-Image / Video Arena.

🔴Public

Aider Polyglot Leaderboard

Aider

Coding agent : C++, Go, Java, JavaScript, Python, Rust (exercices Exercism). Mesure aussi la fiabilité du format d'édition (diffs applicables), critique pour l'intégration agent.

🔴Public

Évaluations privées / anti-contamination

SEAL Leaderboards

Scale AI

Évaluations expertes à prompts non publics : Humanity's Last Exam, MultiChallenge, VISTA (vision), Tool Use, Agentic Tool Use, Adversarial Robustness, Coding, Instruction Following. Meilleure défense actuelle contre la contamination.

🔴Public (prompts privés)

AGI-Eval

Aggregateur CN

Hub bilingue (chinois/anglais) couvrant agents et tests de domaine (droit, médecine, finance). Meilleure couverture des modèles chinois (Qwen, DeepSeek, GLM, Doubao, Kimi, MiniMax) sous-représentés ailleurs.

🔴Public

Vectara Hallucination Leaderboard

Vectara

Étroit mais largement cité : taux d'hallucination sur tâches de résumé. Utile en complément factualité quand SimpleQA / TruthfulQA ne suffisent pas.

🔴Public

Image et vision-langage : pas d'équivalent unique de MMLU

Pour l'image, on ne compare jamais un modèle sur un seul benchmark. Les évaluations sérieuses combinent **trois familles** : la perception classique (ImageNet, COCO, LVIS), le raisonnement visuo-linguistique (VQA, MMMU, MathVista, MMStar) et la lecture documentaire (DocVQA, OCRBench). Deux pièges fréquents : ImageBind n'est pas un benchmark mais un modèle d'embedding ; et un score MME agrégé masque souvent de fortes asymétries entre sous-capacités.

Benchmarks image et vision-langage

Pour l'image, pas d'équivalent unique de MMLU : on combine perception classique, lecture documentaire et raisonnement visuo-linguistique.

ImageNet / ILSVRC

Perception

Classification à grande échelle (top-1, top-5). Toujours un repère, surtout pour les encodeurs visuels.

image-net.org

COCO

Standard

Détection, segmentation, keypoints, captions. Standard polyvalent (AP/AR, BLEU/CIDEr) ; complété par LVIS pour le long-tail.

cocodataset.org

LVIS

Long-tail

>1200 catégories, focus long-tail. Évite le biais vers les classes fréquentes. Complète COCO plutôt que le remplace.

lvisdataset.org

VQA / VQAv2

VQA

Questions ouvertes sur images COCO, 10 réponses annotées. Fondamental pour vision-langage, désormais complété par DocVQA / OCRBench / MMMU.

visualqa.org

DocVQA / OCRBench

Documents

QA sur documents (DocVQA, ANLS) et OCR + compréhension textuelle visuelle (OCRBench, OCRBench v2). Clé pour les assistants documentaires.

[DocVQA](#) · [OCRBench](#)

MMMU / MathVista / CharXiv

Raisonnement visuel

Triade fréquente dans les system cards VLM : compréhension multimodale niveau universitaire, math visuelle, lecture de figures scientifiques.

[MMMU](#)

MMStar

VLM

Mesure perception et raisonnement visuels fins. Souvent combiné à MMMU, MathVista, OCRBench dans les comparaisons VLM.

[Repo MMStar](#)

MME / MME-Unify

Multi-tâches

Évaluation multimodale large/unifiée par sous-capacités. Photo rapide, mais score agrégé parfois opaque.

[MME-Unify \(2025\)](#)

ImageBind

Pas un benchmark

Attention : ce n'est pas un benchmark mais un modèle de représentation joint (image, texte, audio, profondeur, IMU). Souvent cité à tort.

[Paper](#)

Vidéo : compréhension et génération sont deux mondes

Le monde vidéo est encore plus fragmenté que l'image. Une comparaison crédible distingue au minimum **classification d'actions** (Kinetics, Something-Something), **localisation spatio-temporelle** (AVA, ActivityNet), **vidéo-QA** (NExT-QA, EgoSchema, Video-MME) et **génération vidéo** (VBench, VBench-2.0).

Mélanger un score Video-MME avec un score VBench dans une moyenne globale n'a strictement aucun sens : on n'évalue pas la même chose.

Benchmarks vidéo : compréhension et génération

Le monde vidéo est fragmenté : ne jamais fusionner compréhension et génération dans une moyenne globale.

Kinetics-400/600/700

Classification

Classification d'actions humaines courtes (clips ~10s). Top-1/top-5 accuracy. Toujours un benchmark de base, faible couverture du raisonnement.

[Google Research](#)

Something-Something V2

Temporalité

Compréhension d'actions fines et relations d'objets. Plus sensible à la dynamique temporelle que Kinetics.

[Qualcomm](#)

AVA / ActivityNet

Localisation

AVA : actions atomiques localisées en espace/temps (mAP). ActivityNet : challenge multi-tâches reconnaissance/détection/captioning.

[AVA](#) · [ActivityNet](#)

NExT-QA

Vidéo-QA

Vidéo-QA causal et temporel, compréhension de scène. Très utile pour diagnostiquer le raisonnement temporel.

[Paper](#)

Ego4D / EgoSchema

Egocentrique

Référence pour le first-person video : episodic memory, forecasting, hands/objects, social. EgoSchema : QA sur longues vidéos egocentriques.

[Ego4D](#) · [EgoSchema](#)

Video-MME / Video-MME v2

VLM vidéo

Évaluation multimodale vidéo générale. v2 (2026) : 900 vidéos, 7 dimensions, 25 tâches. Devenu standard pour les VLM vidéo.

[Site officiel](#)

VBench / Vbench-2.0

Génération

Génération vidéo : 16 dimensions hiérarchiques de qualité. Vbench-2.0 / Vbench++ étendent vers human fidelity, controllability, physics, commonsense.

[Project page](#)

WebVid

Pas un benchmark

Attention : grand corpus de pré-entraînement vidéo-texte, pas un benchmark public standardisé. Souvent confondu.

[Repo](#)

Agents : trois mondes qui coexistent

Les benchmarks d'agents ont muté plus rapidement que tous les autres. **Trois mondes** coexistent désormais : agents web/assistants (BrowseComp, AssistantBench), agents logiciels (SWE-bench), agents incarnés et robotique (Meta-World, ManiSkill-HAB).

Leur particularité : la performance dépend autant du modèle que du **scaffold** — outils, mémoire, navigateur, budget de pas. À lire avec encore plus de prudence qu'un leaderboard LLM classique.

Benchmarks d'agents

Trois mondes coexistent : agents web/assistants, agents logiciels, agents incarnés/robotique. Le scaffold compte autant que le modèle.

BrowseComp

Web

1 266 questions courtes, vérifiables, très dures à trouver sur le web. Benchmark clé du browsing en 2025-2026.
[OpenAI](#) · [PDF](#)

AssistantBench

Web

214 tâches réalistes, 525+ pages, 258 sites. Très bon test de navigation multi-site, proche de besoins réels.
[Site officiel](#)

Äbench

Tool use

Interaction agent-outils-utilisateur, respect de règles (success rate / conformité). Variance non négligeable entre runs.
[taubench.com](#)

Äbench

Coordination

Dual-control : utilisateur et agent modifient un monde partagé. Ajoute la coordination, absente des benchmarks classiques.
[Sierra Research](#)

SWE-bench (variantes)

Code

Coding agent : résolution d'issues réelles. Verified, Multilingual, Multimodal et SWE-bench Pro recommandé en 2026 pour le frontier coding.
[swebench.com](#)

BFCL

Function calling

Tool-using : validité et exécution des appels de fonctions, sous-scores par scénario. Souvent le meilleur test public de function calling pur.
[Leaderboard](#)

MiniGrid / Atari ALE / CALE**RL**

RL classique : exploration et planification (MiniGrid), contrôle séquentiel (Atari ALE/CALE). Référence historique, éloignée des assistants modernes.

[MiniGrid](#) · [ALE](#)

Meta-World / ManiSkill-HAB**Robotique**

Robotique : 50 tâches de manipulation (Meta-World) ; manipulation basse couche en réarrangement domestique (ManiSkill-HAB, 2025).

[Meta-World](#) · [ManiSkill-HAB](#)

Mini-BEHAVIOR**Embodied**

Embodied / household : version procédurale 2D simplifiée de BEHAVIOR. Sert surtout à l'accessibilité expérimentale.

[OpenReview](#)

Comparer correctement : ce qu'il faut inclure, ce qu'il faut éviter

Une comparaison moderne ne se résume pas à un score. Elle ajoute au minimum six couches au-dessus du chiffre principal : calibration, robustesse, préférence humaine, coût, latence, reproductibilité. Et elle évite quelques pièges récurrents — au premier rang desquels confondre score de benchmark et performance produit.

Ce qu'une comparaison moderne doit inclure

Multi-benchmark

Au moins un benchmark fermé/stable + un benchmark difficile/récent + un test réaliste pour la capacité visée.

Calibration

Distinguer 'réponse correcte', 'réponse fausse' et 'refus justifié' ; rapporter la calibration probabiliste quand disponible.

Robustesse au prompt

Vérifier que les scores ne dépendent pas trop du wording (MMLU-Pro est conçu pour cela).

Préférence humaine

Compléter les benchmarks fermés par Arena ou un panel humain — surtout pour la qualité conversationnelle.

Coût et latence

Tracer une courbe qualité/coût et qualité/latence, pas seulement un classement par score brut.

Reproductibilité

Documenter modèle, version, prompt, température, juge, harness, seed, tokenization.

Variance

Pour agents, tool use et juges LLM, faire plusieurs runs et reporter intervalles de confiance.

Sous-scores

Publier par capacité et par slice, pas seulement la moyenne agrégée.

Les pièges fréquents

Confondre score et performance produit

Un excellent MMLU ne dit rien sur le tool use, la navigation web ou la lecture de documents.

Agréger abusivement

Une moyenne sur 6 benchmarks dépend des poids implicites — les scores agrégés cachent les vraies forces et faiblesses.

Protocoles hétérogènes

0-shot vs few-shot, sans outil vs avec outil, direct vs CoT, juge humain vs juge LLM, 1 run vs 5 runs : pas comparable.

Mélanger benchmark, framework et leaderboard

OpenCompass et lm-eval-harness sont des outillages ; ImageBind est un modèle ; BFCL et BrowseComp sont des benchmarks.

Ignorer la contamination

Vérifier la date, la fermeture du benchmark, et l'existence de versions Redux/CF/Verified/Pro.

Benchmark gaming

Les benchmarks auto-jugés sont manipulables (cf. 'Cheating Automatic LLM Benchmarks'). Croiser avec un échantillon humain.

Protocole pratique selon la modalité

À chaque type de modèle, sa combinaison de benchmarks. Voici la recommandation minimale pour comparer honnêtement deux modèles dans chacune des grandes familles : LLM généraliste, VLM image-document, modèle vidéo et agent.

Protocole recommandé par modalité

LLM généraliste

Combiner un benchmark de connaissance/raisonnement moins saturé (MMLU-Pro ou SuperGPQA), un benchmark de factualité (SimpleQA ou TruthfulQA), un benchmark de préférence conversationnelle (Arena ou AlpacaEval), un benchmark de code ou tool use si pertinent (LiveCodeBench, BFCL), et un benchmark de long contexte (LongBench v2 ou RULER) si le modèle est vendu comme '128k+' ou '1M tokens'.

Couvre les vraies capacités produit, résiste à la contamination, expose la robustesse

VLM image-document

Séparer perception, lecture et raisonnement : ImageNet / COCO / LVIS pour la perception de base, VQAv2 / MMStar / MMMU / MathVista pour le raisonnement visuo-linguistique, DocVQA / OCRBench pour la lecture documentaire. Pour un usage entreprise, les benchmarks documentaires sont souvent plus prédictifs.

Diagnostic précis des sous-capacités, alignement avec les usages métier

Modèle vidéo

Annoncer clairement compréhension vs génération. En compréhension : Kinetics, AVA, NExT-QA, Ego4D, EgoSchema, Video-MME (v2). En génération : VBench puis VBench-2.0 / VBench++. Ne jamais fusionner les deux dans une moyenne.

Évite les comparaisons absurdes entre tâches incompatibles

Agent

Triangle : un benchmark web réaliste (BrowseComp ou AssistantBench), un benchmark de tool-use/règles (AIBench ou BFCL), et un benchmark logiciel si le produit code (SWE-bench ou CoreCodeBench). Reporting du scaffold, du nombre d'étapes et de la variance non négociable.

Couvre les trois axes critiques : navigation, outils, code

Conclusion : du score unique à la batterie diagnostique

L'évolution méthodologique entre 2024 et 2026

Si l'on devait résumer la période 2024-2026 en une phrase : **les benchmarks utiles sont devenus moins des « examens finaux » que des batteries diagnostiques**. Les meilleurs rapports d'évaluation ne montrent plus seulement « qui gagne », mais **sur quoi** le modèle gagne, **à quel coût**, **dans quelles conditions**, **avec quelle robustesse**, et **avec quel risque de tromperie méthodologique**.

C'est précisément la direction prise par HELM, LiveBench, BFCL, les suites de long contexte, les benchmarks agents récents et les analyses de contamination/leaderboards. Les system cards des fournisseurs reflètent ce changement : la page GPT-4.1 d'OpenAI cite explicitement OpenAI-MRCR, Graphwalks, MMMU, MathVista, CharXiv, ComplexFuncBench et ~~A~~bench, avec des chiffres précis par sous-capacité plutôt qu'un score unique.

Un benchmark qui servait encore de référence en 2023-2024 peut être en 2026 soit **saturé** (BBH), soit **corrigé** (MMLU-Redux), soit **décontaminé** (MMLU-CF), soit **déclassé** par un successeur plus difficile (BBEH, SWE-bench Pro). Construire une comparaison crédible, c'est désormais combiner plusieurs angles — et accepter qu'aucun score unique ne raconte toute l'histoire.

Le principe pratique à retenir : choisir d'abord l'**usage visé** (connaissance, code, doc, image, vidéo, agent), puis combiner un benchmark stable, un benchmark récent et difficile, et un test réaliste — en documentant scrupuleusement le protocole. C'est la seule manière d'éviter à la fois la saturation, la contamination et le benchmark gaming.

Sources et ressources

- MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark (arXiv 2406.01574)
<https://arxiv.org/abs/2406.01574>
- MMLU-Redux: Are We Done with MMLU? (arXiv 2406.04127)
<https://arxiv.org/html/2406.04127v1>
- MMLU-CF: A Contamination-Free Multi-Task Language Understanding Benchmark (arXiv 2412.15194)
<https://arxiv.org/abs/2412.15194>
- AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models (arXiv 2304.06364)
<https://arxiv.org/abs/2304.06364>
- SuperGPQA: Scaling LLM Evaluation across 285 Graduate Disciplines (arXiv 2502.14739)
<https://arxiv.org/abs/2502.14739>
- BIG-Bench Extra Hard / BBEH (arXiv 2502.19187)
<https://arxiv.org/abs/2502.19187>
- BIG-bench: Beyond the Imitation Game (arXiv 2206.04615)
<https://arxiv.org/abs/2206.04615>
- Challenging BIG-Bench tasks and chain-of-thought / BBH (arXiv 2210.09261)
<https://arxiv.org/abs/2210.09261>
- SuperGLUE (arXiv 1905.00537)
<https://arxiv.org/abs/1905.00537>
- GLUE — A Multi-Task Benchmark for NLU (BlackboxNLP 2018)
<https://aclanthology.org/W18-5446/>
- TruthfulQA — official paper
<https://truthful.ai/papers/truthfulqa/>
- SimpleQA — OpenAI (introduction)
<https://openai.com/index/introducing-simpleqa/>
- MT-Bench / Judging LLM-as-a-Judge (arXiv 2306.05685)
<https://arxiv.org/abs/2306.05685>
- MT-Bench (PDF)
<https://arxiv.org/pdf/2306.05685>
- AlpacaEval — repo & leaderboard
https://github.com/tatsu-lab/alpaca_eval
- Arena-Hard-Auto — repo
<https://github.com/lmarena/arena-hard-auto>
- LiveBench — official
<https://livebench.ai/>
- LMentry — Elementary Language Tasks (arXiv 2211.02069 / ACL Findings 2023)
<https://aclanthology.org/2023.findings-acl.666/>

- HumanEval — Evaluating LLMs Trained on Code (Codex paper, arXiv 2107.03374)
<https://arxiv.org/abs/2107.03374>
- HumanEval — repo
<https://github.com/openai/human-eval>
- LiveCodeBench — official
<https://livecodebench.github.io/>
- SWE-bench — Can Language Models Resolve Real-World GitHub Issues? (arXiv 2310.06770)
<https://arxiv.org/abs/2310.06770>
- SWE-bench — official site
<https://www.swebench.com/>
- SWE-bench Verified
<https://www.swebench.com/verified.html>
- SWE-bench Multilingual
<https://www.swebench.com/multilingual-leaderboard.html>
- SWE-bench Multimodal
<https://www.swebench.com/multimodal>
- OpenAI — Why we no longer evaluate SWE-bench Verified (2026)
<https://openai.com/index/why-we-no-longer-evaluate-swe-bench-verified/>
- CoreCodeBench — repo
<https://github.com/AGI-Eval-Official/CoreCodeBench>
- Berkeley Function-Calling Leaderboard (BFCL) — Gorilla
<https://gorilla.cs.berkeley.edu/leaderboard.html>
- Gorilla / BFCL — repo
<https://github.com/shishirpatil/gorilla>
- LongBench (arXiv 2308.14508)
<https://arxiv.org/abs/2308.14508>
- LongBench v2 (arXiv 2412.15204)
<https://arxiv.org/abs/2412.15204>
- RULER: What's the Real Context Size of Your Long-Context Models? (arXiv 2404.06654)
<https://arxiv.org/abs/2404.06654>
- InfiniteBench — Extending Long Context Evaluation Beyond 100K Tokens (arXiv 2402.13718)
<https://arxiv.org/abs/2402.13718>
- L-Eval — Standardizing Long-Context Evaluation (ACL 2024)
<https://aclanthology.org/2024.acl-long.776/>
- HELM — Stanford CRFM (Holistic Evaluation of Language Models)
<https://crfm.stanford.edu/helm/>
- lm-evaluation-harness — EleutherAI
<https://github.com/EleutherAI/lm-evaluation-harness>
- OpenCompass — repo
<https://github.com/open-compass/opencompass>

- Open LLM Leaderboard v2 — Hugging Face (blog)
<https://huggingface.co/spaces/open-llm-leaderboard/blog>
- Chatbot Arena (LMArena) — Arena Leaderboard
<https://lmarena.ai/>
- Artificial Analysis — quality / price / speed comparison
<https://artificialanalysis.ai/>
- OpenRouter Rankings — usage par volume de tokens
<https://openrouter.ai/rankings>
- Vellum LLM Leaderboard
<https://www.vellum.ai/llm-leaderboard>
- Scale SEAL Leaderboards (prompts privés)
<https://scale.com/leaderboard>
- Epoch AI — Benchmarking Hub
<https://epoch.ai/benchmarks>
- Aider Polyglot Coding Leaderboard
<https://aider.chat/docs/leaderboards/>
- AGI-Eval — aggregateur bilingue
<https://agi-eval.cn/>
- Vectara Hallucination Leaderboard
<https://github.com/vectara/hallucination-leaderboard>
- ImageNet — official
<https://www.image-net.org/>
- ILSVRC challenges
<https://www.image-net.org/challenges/LSVRC/>
- COCO dataset
<https://cocodataset.org/>
- LVIS dataset
<https://www.lvisdataset.org/>
- VQA Challenge
<https://visualqa.org/challenge.html>
- DocVQA challenges
<https://www.docvqa.org/challenges>
- OCRBench / MultimodalOCR — repo
<https://github.com/yuliang-liu/multimodalocr>
- OCRBench (challenge)
<https://rrc.cvc.uab.es/?ch=17>
- MMStar — repo
<https://github.com/MMStar-Benchmark/MMStar>
- MME-Unify (arXiv 2503.21755)
<https://arxiv.org/html/2503.21755v1>

- ImageBind — One Embedding Space To Bind Them All (arXiv 2305.05665)
<https://arxiv.org/abs/2305.05665>
- Kinetics — Google Research
<https://research.google/pubs/the-kinetics-human-action-video-dataset/>
- Something-Something V2 — Qualcomm
<https://www.qualcomm.com/developer/software/something-something-v-2-dataset>
- AVA — Google Research
<https://research.google.com/ava/>
- ActivityNet
<https://activity-net.org/>
- Ego4D — official
<https://ego4d-data.org/>
- NExT-QA (arXiv 2105.08276)
<https://arxiv.org/abs/2105.08276>
- Video-MME — CVPR 2025 paper (PDF)
https://openaccess.thecvf.com/content/CVPR2025/papers/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_M
- VBench — project page
<https://vchitect.github.io/VBench-project/>
- VBench — CVPR 2024 paper
https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Model
- WebVid — repo
<https://github.com/m-bain/webvid>
- BrowseComp — OpenAI
<https://openai.com/index/browsecomp/>
- BrowseComp — paper PDF
<https://cdn.openai.com/pdf/5e10f4ab-d6f7-442e-9508-59515c65e35d/browsecomp.pdf>
- AssistantBench — official
<https://assistantbench.github.io/>
- Aibench — official
<https://taubench.com/>
- MiniGrid — publications
<https://minigrid.farama.org/content/publications/>
- Arcade Learning Environment (ALE) — repo
<https://github.com/farama-foundation/arcade-learning-environment>
- CALE — Continuous Atari Learning Environment (NeurIPS 2024 PDF)
https://proceedings.neurips.cc/paper_files/paper/2024/file/f359c6545bea1ffc971f483469b37122-Paper-Datasets_and_Benchmarks_Track.p
- Meta-World — LeRobot docs
<https://github.com/huggingface/lerobot/blob/main/docs/source/metaworld.mdx>
- ManiSkill-HAB — project page
<https://arth-shukla.github.io/mshab/>

- OpenAI — GPT-4o System Card
<https://openai.com/fr-FR/index/gpt-4o-system-card/>
- OpenAI — GPT-4.1 (model page & evals)
<https://openai.com/index/gpt-4-1/>
- Cheating Automatic LLM Benchmarks (arXiv 2410.07137)
<https://arxiv.org/abs/2410.07137>
- Leaderboard analysis & best practices (arXiv 2310.01824)
<https://arxiv.org/html/2310.01824v2>
- LBOps — leaderboard operations & reproducibility (OpenReview)
<https://openreview.net/forum?id=d7KBjml3GmQ>
- Data contamination in foundation models — Balloccu et al., ACL 2024 (PDF)
https://www.cl.uni-heidelberg.de/courses/ss25/the_mystery_of_in-context_learning_of_llms/papers/BalloccuACL2024.pdf