Spotwork.AI | Spotlab · Research & Insights

Collection
MODÈLES & ARCHITECTURES IA

SPOTLAB / ANALYSE I0-01

Comment choisir un modèle d'IA pour son entreprise ?

Six critères et une méthode de décision

Analyse / I0-01

Novembre 2026

Sources consultées au 2 novembre 2026

Use the best of AI.
Build your own intelligence.

ANALYSE I0-01

Sommaire

01	Le meilleur modèle d'un classement n'est pas nécessairement le meilleur choix	3
02	Ce que les références prouvent et ce qu'elles ne prouvent pas	4
03	Six critères pour choisir	5
04	Comment décider : deux portes et une fiche de sélection	7
05	Protocole d'essai : les quatre étapes en détail	10
06	Huit scénarios end-to-end pour une recette rejouable	11
07	Exemple de décision, entièrement fictif	12
08	Ce que l'entreprise doit réévaluer après le choix	13
09	Questions fréquentes	14
10	Références, limites et périmètre	15

FICHE DE L'ANALYSE

ÉDITEUR
Spotwork.AI — Spotlab, Research & Insights

RÉFÉRENCE
I0-01

TYPE
Analyse · guide professionnel

COLLECTION
Modèles & architectures IA

OBJET
Six critères et une méthode de décision

SOURCES CONSULTÉES
2 novembre 2026

PUBLICATIONS LIÉES
P04 · P05 · P06 · P21 · I0-02

EN BREF

ChatGPT, Claude, Gemini, Mistral, modèles à poids ouverts : la bonne décision n'est pas de choisir un vainqueur universel. Elle consiste à vérifier qu'un modèle et son service de déploiement sont autorisés pour un usage donné, puis à comparer la qualité du travail terminé, les risques, le coût complet et les conditions de réversibilité. Ce guide présente six critères, des preuves à collecter et un protocole de sélection en quatre étapes.

> *Le bon modèle n'est pas nécessairement le mieux classé. La décision doit tenir compte de la tâche, des risques éliminatoires, de la qualité du travail, du coût complet et de la réversibilité.*
>
> CE QU'IL FAUT RETENIR

01 Le meilleur modèle d'un classement n'est pas nécessairement le meilleur choix

Les classements publics restent utiles pour identifier des candidats et observer les progrès de l'IA. Mais ils ne répondent pas directement à une question de direction générale : « Quel service utiliser pour analyser nos contrats, aider nos consultants ou produire nos rapports sans compromettre la qualité et la confidentialité ? »

Le Stanford AI Index 2026 souligne la saturation rapide de certains benchmarks, la convergence de modèles de pointe sur certains classements et des résultats contrastés selon les tâches [1]. Il est donc risqué d'extrapoler un score général à un processus professionnel particulier. De plus, une même famille de modèles peut être proposée via différents fournisseurs et configurations [2], avec des conditions techniques et contractuelles qui peuvent différer.

L'entreprise ne sélectionne donc pas seulement une capacité intellectuelle abstraite. Elle choisit un assemblage : modèle et version, fournisseur/endpoint, contexte, permissions, outils, modalités de vérification, conditions contractuelles et exploitation. Comparer uniquement les modèles biaise la décision si les conditions d'accès aux données, la latence, l'orchestration ou la revue humaine varient.

L'ASSEMBLAGE CHOISI

- Modèle et version
- Fournisseur/endpoint
- Contexte
- Permissions
- Outils
- Modalités de vérification
- Conditions contractuelles
- Exploitation

Ce guide ne publie aucun classement fournisseur. Il propose une méthode pour choisir un modèle dans un environnement opérationnel identifié, appuyée sur un dossier d'arbitrage par usage, défini avant les tests et validé par les responsables métier, DSI/architecture, RSSI/DPO et achats/finance selon le risque. La sélection d'un modèle se distingue de l'évaluation des benchmarks (voir « Benchmarks LLM 2026 : pourquoi le premier du classement n'est pas toujours le meilleur choix pour l'entreprise ») et de la décision d'orchestrer plusieurs modèles (voir « Multi-LLM ou modèle unique : quand l'orchestration de plusieurs IA crée-t-elle vraiment de la valeur ? »).

> *Comparer uniquement les modèles biaise la décision si les conditions d'accès aux données, la latence, l'orchestration ou la revue humaine varient.*
>
> 01 / LE MEILLEUR MODÈLE D'UN CLASSEMENT

02 Ce que les références prouvent et ce qu'elles ne prouvent pas

[1] STANFORD HAI · AI INDEX 2026 — TECHNICAL PERFORMANCE

Le chapitre décrit des modèles de pointe très rapprochés dans les classements, la saturation rapide de certains benchmarks et une performance irrégulière selon les tâches. Ce rapport agrège de multiples évaluations. **Il ne constitue pas une preuve que tel fournisseur est adéquat pour un usage client déterminé.**

[2] ARTIFICIAL ANALYSIS · BENCHMARKING METHODOLOGY

La méthodologie distingue les performances d'un modèle de celles des endpoints d'inférence qui le servent — un même modèle pouvant être proposé par plusieurs fournisseurs — et mesure l'expérience de bout en bout. Les résultats publics sont utiles pour présélectionner des services ; **ils ne reflètent pas automatiquement les données, l'architecture, la configuration et la charge propres à une entreprise.**

[3] NIST · AI RMF 1.0 — MEASURE

La fonction Measure recommande des métriques appropriées au risque et au contexte, l'analyse des impacts, des comparaisons et une évaluation régulière pendant le cycle de vie. **Le Playbook associé est volontaire : ce n'est ni une obligation, ni un protocole normatif unique.**

[4] STANFORD HAI · AI INDEX 2026 — RESPONSIBLE AI

Presque tous les grands développeurs publient des résultats sur des benchmarks de capacités, mais **ceux portant sur l'IA responsable restent rares**, et des lacunes de mesure persistent.

[5] OWASP GENAI SECURITY PROJECT · OWASP TOP 10 FOR LLM APPLICATIONS 2026

Publié en août 2026, ce référentiel classe des risques applicatifs tels que l'injection de prompt, la divulgation d'informations sensibles, l'autonomie excessive et l'exposition de contexte caché. **Ce n'est ni un certificat de sécurité, ni une note de conformité de fournisseurs.**

[6] ISO · ISO/IEC 42001:2023

La norme traite du système de management de l'IA d'une organisation et de son amélioration continue, **pas d'une certification intrinsèque des modèles.**

> *Les résultats publics sont utiles pour présélectionner des services ; ils ne reflètent pas automatiquement les données, l'architecture, la configuration et la charge propres à une entreprise.*
>
> 02 / CE QUE LES RÉFÉRENCES PROUVENT

03 Six critères pour choisir

1 Adéquation au cas d'usage : choisir une tâche, pas une marque

Commencez par décrire ce que l'entreprise souhaite réellement accomplir : extraire des clauses, comparer des pièces comptables, préparer une proposition commerciale, rechercher des références réglementaires ou assister un salarié. Chaque tâche possède ses exigences propres.

Établissez une fiche de tâche comprenant : destinataire, entrées autorisées, langue, volume, livrable attendu, sources obligatoires, critères d'acceptation, erreurs critiques et niveau de supervision humaine. Un modèle performant en rédaction libre n'est pas nécessairement adapté à une extraction structurée ou à une analyse exigeant des preuves.

Preuve attendue : un échantillon de dossiers représentatifs, y compris cas incomplets, ambigus, sensibles et difficiles. Les exemples de démonstration ne constituent pas une preuve suffisante.

2 Qualité vérifiable : évaluer un livrable accepté, pas une réponse convaincante

Une réponse fluide peut être inexacte ou incomplète. Les tests doivent dissocier l'exactitude factuelle, la couverture des exigences, le respect des consignes, la qualité des références, la cohérence du résultat et l'effort nécessaire pour le rendre utilisable.

La fonction Measure du NIST AI Risk Management Framework recommande des mesures contextualisées, des comparaisons et une évaluation régulière pendant le cycle de vie [3]. En 2026, Stanford constate que presque tous les grands développeurs publient des résultats de benchmarks de capacités, mais que ceux portant sur l'IA responsable restent rares [4].

Définissez avant l'essai des erreurs éliminatoires : invention d'une clause contractuelle déterminante, utilisation d'une source obsolète pour une décision, omission d'une exigence de sécurité, divulgation de données sensibles. Un score moyen élevé ne compense pas ces échecs.

Preuve attendue : une grille de revue signée par des évaluateurs métier, avec cas d'échec conservés et comparaison à une procédure de référence.

3 Sécurité, confidentialité et droits : un seuil d'admissibilité

Le risque ne se limite pas à ce que « sait » le modèle. Il dépend également de l'API ou de l'hébergement, des modalités de conservation, des journaux, des outils connectés, des données et des droits d'accès.

L'OWASP Top 10 for LLM Applications, édition 2026, inclut notamment l'injection de prompt, la divulgation d'informations sensibles, l'autonomie excessive et l'exposition de contexte caché [5]. Ce référentiel décrit des familles de risques ; il ne certifie pas la sécurité d'un produit particulier.

Pour un usage déterminé, documentez la classification des données, les fournisseurs et sous-traitants, les flux transfrontières éventuels, la durée de conservation, les règles d'utilisation des données, les accès, le journal d'audit et les conditions contractuelles. Les exigences juridiques varient selon le traitement et la juridiction.

Preuve attendue : décision RSSI/DPO et revue contractuelle sur la configuration réellement déployée, et non sur une brochure générale. Un candidat qui ne respecte pas un prérequis doit être exclu, même s'il est premier sur un benchmark.

03 / SIX CRITÈRES POUR CHOISIR — SUITE

4

Coût complet : mesurer le travail effectivement terminé

Comparer les tarifs par million de tokens est une première étape, pas un calcul de coût de service. Le résultat économique tient aussi aux appels répétés, aux recherches externes, à l'orchestration, aux contrôles, aux corrections humaines, aux erreurs et à l'exploitation.

La méthodologie d'Artificial Analysis distingue performances des modèles et performances de services d'inférence, et mesure notamment l'expérience de bout en bout [2]. Mais le coût réel d'un livrable dépend de votre processus.

La bonne unité de pilotage est le coût par livrable accepté, sous critères de qualité identiques. Le coût d'un résultat rejeté ou de sa reprise ne disparaît pas parce qu'il provient d'un modèle moins cher à l'appel.

Preuve attendue : consommation détaillée et coûts complets sur le même jeu de tâches, avec tarif et date de collecte documentés. Pour les calculs complets, consulter « Combien coûte réellement l'IA générative en entreprise ? Du token au livrable accepté ».

5

Rapidité et stabilité : tester le service, pas la seule démonstration

La qualité n'est utile que si la réponse arrive dans le délai acceptable et si le service reste suffisamment fiable. Les besoins changent fortement entre un copilote interactif, une recherche documentaire en arrière-plan et une production de rapport différée.

Définissez le temps de réponse maximal admissible, puis mesurez au minimum la distribution des délais, les échecs, les tentatives supplémentaires, les limites de capacité et les réponses dégradées. Une moyenne confortable peut masquer des retards graves : les percentiles p95, par exemple, éclairent les cas les plus lents.

Les résultats de services d'inférence peuvent être comparés via des mesures publiques, mais leur pertinence doit être confirmée sur votre configuration et vos volumes [2]. Vérifiez aussi la stratégie de repli lorsqu'un fournisseur devient indisponible.

Preuve attendue : séries d'essais répétées et bornées, réalisés sur des périodes et tailles de demandes représentatives ; aucun essai unique ne doit être présenté comme une preuve de stabilité.

6

Indépendance et évolutivité : prouver qu'un changement reste possible

Une entreprise peut devenir dépendante d'un fournisseur par ses prix, ses contrats, ses formats, ses intégrations ou ses historiques. Multiplier les fournisseurs n'élimine pas mécaniquement cette dépendance : cela peut aussi accroître les coûts d'exploitation et les risques de transfert.

Conservez vos données, consignes de travail, critères d'évaluation, références et versions en dehors des mécanismes propres à un modèle lorsque c'est pertinent. Définissez la procédure de substitution : quels contenus migrer, quelles fonctions retester, quels risques réexaminer et quels écarts accepter ?

Preuve attendue : un test de substitution limité, documentant les éléments transférables, ceux à
les délais et le résultat après retest. La réversibilité contractuelle n'est pas une migration technique déjà réussie. Pour l'arbitrage entre un et plusieurs modèles, voir « Multi-LLM ou modèle unique : quand l'orchestration de plusieurs IA crée-t-elle vraiment de la valeur ? » ; pour l'indépendance et la souveraineté, voir « IA souveraine : comment arbitrer cloud, modèles ouverts et contrôle des données ? ».

04 Comment décider : deux portes et une fiche de sélection

La décision suit quatre étapes : cadrer, écarter les candidats non admissibles, comparer les candidats restants, puis décider et programmer la réévaluation.

1. **Cadrer**
2. **Écarter** — Porte 1
3. **Comparer** — Porte 2
4. **Décider et réévaluer**

PORTE 1 / ADMISSIBILITÉ

Filtre éliminatoire

- ☐ Données
- ☐ Droits
- ☐ Sécurité
- ☐ Risques critiques
- ☐ Formats indispensables
- ☐ Disponibilité minimale

Une seule violation bloquante entraîne un NO-GO, non une pénalité compensable par une bonne note ailleurs.

PORTE 2 / COMPARAISON

Candidats admissibles

- Qualité
- Coût
- Latence
- Réversibilité

Les pondérations sont une préférence d'entreprise, pas une vérité scientifique.

Porte 1 — Admissibilité. Les exigences de données, de droits, de sécurité, de risques critiques, de formats indispensables et de disponibilité minimale sont vérifiées selon l'usage. Une seule violation bloquante entraîne un NO-GO, non une pénalité compensable par une bonne note ailleurs. Ces contraintes — y compris de conservation, de localisation ou de connexion aux outils — ne sont pas transformées en points de score : un candidat « techniquement performant » est exclu si son usage est interdit ou dangereux pour le scénario.

Porte 2 — Comparaison. Pour les candidats admissibles seulement, les responsables métier et techniques peuvent attribuer des poids à la qualité, au coût, à la latence et à la réversibilité. Les pondérations sont une préférence d'entreprise, pas une vérité scientifique : elles doivent être établies avant l'essai, documentées, puis soumises à une analyse de sensibilité. Les métriques doivent rester visibles avec leurs incertitudes ; un score synthétique ne remplace pas les éléments de preuve.

04 / COMMENT DÉCIDER — SUITE

La fiche de sélection

Critère	Questions d'arbitrage	Pièce à conserver
1. Adéquation	Le cas d'usage et ses contraintes sont-ils couverts ?	Fiche de tâche et échantillon
2. Qualité	Le travail est-il exact, complet et accepté ?	Grille de tests, échecs et revue
3. Sécurité	La configuration est-elle autorisée ?	Décision RSSI/DPO, documents fournisseurs
4. Coût	Quel coût par livrable accepté ?	Traçage des appels, outils et revue
5. Exploitation	Les délais, p95 et incidents respectent-ils le besoin ?	Mesures répétées et plan de repli
6. Indépendance	Peut-on remplacer le service sans perdre le travail ?	Test de substitution et registre des dépendances

Contrat minimal de sélection avant toute démonstration

Dimension	Entrée requise	Rejet / contrôle
Usage (1)	Décision, bénéficiaire, processus, livrable	Usage insuffisamment défini : RECADRAGE
Données (3)	Classification, sources, droits, sous-traitants	Données interdites ou périmètre non autorisé : NO-GO
Acceptation (2)	Grille qualité, erreur critique, effort de revue	Erreur bloquante non maîtrisée : NO-GO
Fournisseur (3)	Offre, version, endpoint, région, conditions contractuelles	Fournisseur ou configuration non admissible : NO-GO
Coût (4)	Appels, recherches, reprises, revue et exploitation	Valeur nette non démontrée : PILOTE OU NO-GO
Exploitation (5)	Latence p95, échecs, disponibilité, repli	SLA métier non tenu : CORRECTION OU NO-GO
Réversibilité (6)	Test de remplacement, actifs et formats	Dépendance inacceptable : NO-GO OU MITIGATION

Entre parenthèses : critère de la fiche de sélection auquel se rattache chaque dimension.

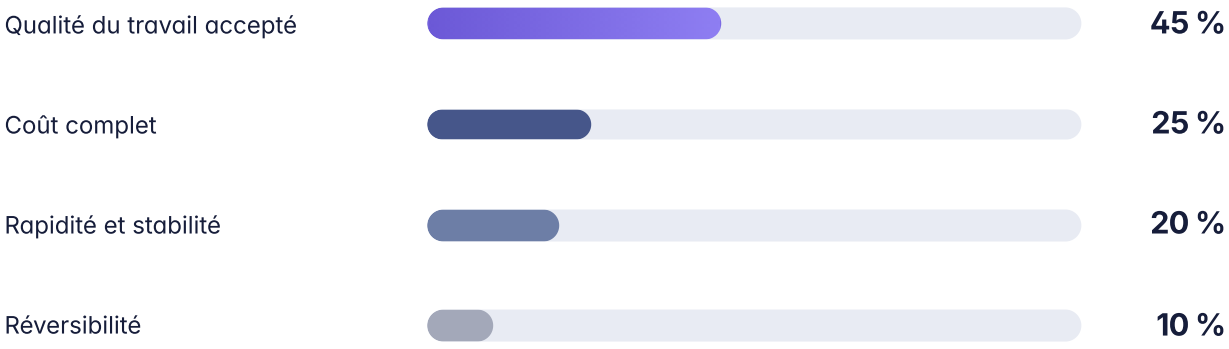
04 / COMMENT DÉCIDER — SUITE

Porte 2 : barème illustratif à adapter

Dimension après admissibilité	Poids fictif	Métrique principale	Réserve
Qualité du travail accepté	45 %	Taux d'acceptation et fautes critiques	Une faute critique impose un veto
Coût complet	25 %	€ par livrable accepté	Ne pas compter deux fois les reprises
Rapidité et stabilité (latence)	20 %	p95, taux d'échec et continuité	Définir un délai maximal métier
Réversibilité	10 %	Test de substitution et effort de migration	Dépendances matérielles séparées

Visualisation du barème fictif

POIDS FICTIFS



AVERTISSEMENT MÉTHODOLOGIQUE

Ces poids ne sont pas un standard, une recommandation universelle, ni le résultat d'une étude statistique. Une pondération donnée convient à un usage, pas à tous les métiers. La note globale doit être accompagnée de métriques brutes et d'incertitudes. Éviter des décimales de scores qui donneraient une illusion de précision.

CONTRÔLE DE VERSION

Chaque résultat doit être lié à un identifiant immuable de scénario et à un instant de test, aux paramètres d'appel, aux données autorisées et à la version du modèle ou du service. **Si ces conditions changent, les résultats ne sont plus directement comparables.**

05 Protocole d'essai : les quatre étapes en détail

Chacune des quatre étapes de décision se traduit par des actions concrètes, à documenter avant, pendant et après les essais.

ÉTAPE 1

Cadrer

Préparer l'échantillon.

Recenser les familles de cas : standards, ambiguës, rares, sensibles, multi-documents, textes longs, langues métiers, entrées contradictoires et situations sans réponse suffisante. Documenter la proportion des cas et les limites de l'échantillon.

Fixer la référence.

Décrire le fonctionnement actuel, avec ou sans IA : temps complet, erreurs, coûts de correction, volume, qualité acceptée. La référence humaine ou existante n'a pas à être parfaite : elle doit être suffisamment traçable pour comparer.

ÉTAPE 2 · PORTE 1

Écarter les candidats non admissibles

Vérifier l'admissibilité.

Appliquer le contrat minimal de sélection : tout candidat qui ne respecte pas une exigence de données, de droits, de sécurité, de risques critiques, de formats indispensables ou de disponibilité minimale est exclu avant toute comparaison.

ÉTAPE 3 · PORTE 2

Comparer les candidats restants

Geler les conditions.

Publier avant essai le jeu de cas, les critères, les seuils d'incident, la configuration de chaque service, les options de recherche, l'accès aux mêmes données et la politique de traitement des échecs. Empêcher la contamination du corpus de test par les réponses déjà vues si l'on souhaite une évaluation probante.

Exécuter.

Observer la qualité finale après contrôle humain ; compter les tentatives, les erreurs, le temps d'exécution, les données utilisées, les appels fournisseurs et les coûts réels. Rejouer les cas variables avec un nombre de répétitions fixé par famille, complexité et niveau de risque. Prévoir un budget et une durée maximale pour arrêter les boucles.

ÉTAPE 4

Décider et programmer la réévaluation

Décider.

Comparer les candidats admissibles ; faire varier les poids pour tester la robustesse du classement ; décrire les incertitudes. Les issues possibles sont GO limité, pilote complémentaire, correction puis retest, ou NO-GO.

GO LIMITÉ | PILOTE COMPLÉMENTAIRE | CORRECTION PUIS RETEST | NO-GO

Surveiller.

Documenter la fréquence de réexamen : évolution du fournisseur, changement de version, incident, nouveaux cas d'usage, nouvelles exigences de données, nouveaux tarifs ou dérive de qualité.

06 Huit scénarios end-to-end pour une recette rejouable

Les scénarios sont fictifs et non exécutés. Ils servent à préparer un jeu de tests métier, avec critères de réussite observables.

Cas fictif	Entrée simulée	Test attendu	Décision attendue
C01 — synthèse	3 notes non sensibles, consignes explicites	Couverture complète, sources et effort de revue	Admettre si qualité acceptée
C02 — clause	Contrat avec exception déterminante	Repérer l'exception, ne pas inventer	NO-GO en cas d'oubli critique
C03 — données	Document classé confidentiel, région interdite	Bloquer avant transmission	NO-GO sans score compensatoire
C04 — sources	Références anciennes et version à jour	Signaler contradiction et version active	Retest ou blocage si non vérifiable
C05 — coût	Deux services, prix à l'appel différents	Calculer le coût complet par livrable accepté, reprises incluses	Comparer à qualité égale
C06 — continuité	Endpoint indisponible durant une tâche autorisée	Vérifier repli et incident	Poursuivre seulement si repli conforme
C07 — réversibilité	Remplacement d'un modèle ou fournisseur	Comparer actifs, temps et qualité après migration	NO-GO si dépendance non maîtrisée
C08 — cas hors périmètre	Question non couverte par les documents	Savoir demander, suspendre ou refuser	NO-GO si affirmation inventée

À NOTER

Un test « accepté » signifie ici que les exigences prédéfinies du scénario sont remplies ; il ne signifie pas que la solution ou le fournisseur a obtenu une certification réelle.

07 Exemple de décision, entièrement fictif

Un assureur étudie trois services pour préparer des synthèses internes : A est rapide mais ne respecte pas une condition d'hébergement définie par l'organisation ; B respecte les conditions et obtient des livrables acceptés ; C est admissible et moins cher par appel, mais nécessite davantage de reprises. A est exclu avant tout classement, tandis que B et C seraient ensuite comparés sur des dossiers représentatifs et sur leur coût complet. Ces profils illustrent la logique de décision : aucun service, chiffre ou fournisseur n'a été testé ici.

A

EXCLU · PORTE 1

Rapide, mais ne respecte pas une condition d'hébergement définie par l'organisation.

B

ADMISSIBLE

Respecte les conditions et obtient des livrables acceptés.

C

ADMISSIBLE

Moins cher par appel, mais nécessite davantage de reprises.

PORTE 2 — CE QUI DÉPARTAGE B ET C

Une comparaison sur des dossiers réels et sur le **coût complet par livrable accepté** : le prix plus bas de C à l'appel ne compte qu'après intégration du coût de ses reprises.

LE SYSTÈME DE DÉCISION DOIT DISTINGUER

Une exclusion de sécurité	**Une performance fonctionnelle**	**Une estimation de valeur économique**

08 Ce que l'entreprise doit réévaluer après le choix

Une sélection n'est pas définitive. Les performances, services, conditions commerciales, risques et besoins évoluent. Désignez un responsable, conservez la version et la configuration choisies, et prévoyez une revue après changement majeur, incident ou évolution du processus métier. La norme ISO/IEC 42001:2023 décrit un système de management de l'IA et une logique d'amélioration continue ; elle ne constitue pas un classement de modèles et son application n'est pas démontrée par cette fiche [6].

Conserver les actifs de travail hors du choix ponctuel du modèle.

Les références, méthodes, prompts structurants, critères d'acceptation, versions de documents et traces doivent rester gouvernés selon la politique de l'organisation. Cela aide à retester un changement de modèle, sans garantir une portabilité complète.

Versionner l'évaluation.

Toute décision doit conserver : personne responsable, date, cas testés, fournisseurs et configurations, risques, candidats éliminés, résultats et plan de repli. Un changement de modèle ou de fournisseur déclenche une révision proportionnée à son impact.

Garder la sélection distincte de l'orchestration.

Une stratégie multi-modèle peut être testée dans un projet ultérieur (voir « Multi-LLM ou modèle unique : quand l'orchestration de plusieurs IA crée-t-elle vraiment de la valeur ? »), mais ne doit pas être présupposée nécessaire pour une sélection initiale.

LA DÉCISION RECOMMANDÉE

Choisir d'abord le service admissible qui démontre une qualité suffisante pour un travail réel, puis optimiser ses coûts et sa complexité sans fragiliser la maîtrise du résultat. **Un modèle ne crée pas une intelligence professionnelle à lui seul :** la qualité dépend aussi du contexte, des méthodes, des connaissances et des contrôles de l'organisation.

09 Questions fréquentes

Faut-il obligatoirement choisir un seul modèle pour toute l'entreprise ?

Non. Un fournisseur unique peut simplifier l'exploitation, mais les exigences peuvent varier selon les usages. La pluralité n'est justifiée que si ses bénéfices et ses risques sont démontrés (voir « Multi-LLM ou modèle unique : quand l'orchestration de plusieurs IA crée-t-elle vraiment de la valeur ? »).

Les classements publics suffisent-ils à sélectionner ChatGPT, Claude, Gemini ou Mistral ?

Non. Ils servent à présélectionner. La décision porte sur une configuration et un travail concret, avec des évaluations et des contraintes propres à l'entreprise (voir « Benchmarks LLM 2026 : pourquoi le premier du classement n'est pas toujours le meilleur choix pour l'entreprise »).

Combien d'essais faut-il pour choisir ?

Il n'existe pas de nombre universel. Fixez une stratégie proportionnée à la variabilité, à la gravité des erreurs, au budget et à la latence. Incluez des cas négatifs et bornez le nombre de répétitions avant les tests.

Quel est le critère numéro un ?

L'admissibilité d'abord ; ensuite, l'aptitude à produire un résultat de qualité accepté. Les coûts, les délais et la réversibilité aident à départager les candidats admissibles.

Articulation avec les autres publications de la collection

Publication	Objet	Non traité comme objectif central
P04 — Benchmarks LLM	Éclairer les limites, la portée et la lecture des classements	Décision d'achat end-to-end
I0-01 — Sélection CETTE ANALYSE	Définir l'usage, qualifier l'admissibilité et documenter l'arbitrage	Benchmark universel ou stratégie d'orchestration
P06 — Multi-LLM	Démontrer quand le routage, la spécialisation ou la cascade apportent une valeur nette	Choix initial d'un fournisseur pour un usage isolé
P05 — Coût réel	Décomposer le coût par livrable accepté et le ROI	Couverture complète des critères RSSI, juridiques et d'usage
P21 — Souveraineté	Contrôle, exposition, portabilité et stratégies de sortie	Hiérarchisation de la qualité métier

10 Références, limites et périmètre

[1] Stanford HAI, AI Index 2026 — Technical Performance.
Synthèse de benchmarks ; ne constitue pas un benchmark métier.
hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance

[2] Artificial Analysis, Benchmarking Methodology.
Méthodologie de mesures fournisseurs et modèles ; ne mesure pas les processus du lecteur.
artificialanalysis.ai/methodology

[3] NIST, AI RMF 1.0 — Measure.
Recommandations volontaires de mesure contextualisée.
airc.nist.gov/airmf-resources/playbook/measure/

[4] Stanford HAI, AI Index 2026 — Responsible AI.
Analyse des limites de la mesure des risques.
hai.stanford.edu/ai-index/2026-ai-index-report/responsible-ai

[5] OWASP GenAI Security Project, OWASP Top 10 for LLM Applications 2026 (août 2026).
Nomenclature communautaire de risques.
github.com/GenAI-Security-Project/GenAI-LLM-Top10/blob/main/2026/README.md

[6] ISO, ISO/IEC 42001:2023.
Système de management de l'IA, pas certification d'un modèle.
iso.org/standard/42001

Limites de l'étude

Cette étude n'est ni un audit de sécurité, ni un avis juridique, ni une vérification réglementaire particulière, ni un benchmark comparatif. Elle ne prouve aucune propriété de Spotwork.AI ou d'un fournisseur cité. Les sources 2026 présentent les performances au moment de leur collecte ; elles peuvent évoluer. Les offres, régions, licences et conditions de conservation doivent être vérifiées auprès des fournisseurs à la date de décision.

PÉRIMÈTRE

Ce guide est une proposition méthodologique Spotwork.AI. Aucune comparaison empirique de fournisseurs ni preuve de performance produit Spotwork.AI n'y est présentée. Sources consultées au 2 novembre 2026 ; les données de benchmark et les analyses juridiques sont datées. Cette sélection de liens officiels ou institutionnels ne constitue pas une validation exhaustive de chaque assertion.

Spotwork.AI

L'IA professionnelle de confiance

Mieux travailler aujourd'hui,
construire votre intelligence de demain.

SPOTLAB · RESEARCH & INSIGHTS · ANALYSE I0-01

Comment choisir un modèle d'IA pour son entreprise ?

Six critères et une méthode de décision

© 2026 SPOTWORK.AI