SPOTLAB / ANALYSE P06

# Multi-LLM ou modèle unique : quand l'orchestration de plusieurs IA crée-t-elle vraiment de la valeur ?

Modèle unique, spécialisation, routeur ou cascade : comment prouver que la complexité multi-LLM apporte réellement de la valeur ?

**Analyse / P06**

Novembre 2026

Sources consultées au 2 novembre 2026

ANALYSE P06

# Sommaire


| | | |
|---|---|---|
| 01 | Trois enseignements pour les décideurs | 3 |
| 02 | Ce que les recherches permettent d'affirmer — et leurs limites | 3 |
| 03 | Quatre stratégies différentes, souvent confondues | 4 |
| 04 | La règle de décision : deux portes avant de regarder le prix | 5 |
| 05 | Exemple : trois familles de travail dans une direction des risques | 6 |
| 06 | Le protocole en six décisions | 7 |
| 07 | L'indépendance ne vient pas du nombre de modèles | 8 |
| 08 | À quelles conditions le multi-LLM est-il une mauvaise idée ? | 8 |
| 09 | La décision à retenir | 9 |
| A–F | Annexes méthodologiques | 10 |
| 10 | Sources et méthode | 14 |



FICHE DE L'ANALYSE

ÉDITEUR
Spotwork.AI — Spotlab, Research & Insights

RÉFÉRENCE
P06

TYPE
Analyse stratégique et méthode de décision

COLLECTION
Modèles & architectures IA

SOURCES CONSULTÉES
2 novembre 2026

PUBLICATIONS LIÉES
I0-01 · P04 · P05 · P21



EN BREF

Faut-il confier tout le travail à un seul modèle d'IA, sélectionner un modèle par métier ou confier cette décision à un routeur ? La stratégie multi-LLM peut améliorer le rapport entre qualité, coût et rapidité, mais elle peut aussi multiplier les appels, les erreurs et les dépendances. La bonne question n'est pas combien de modèles l'entreprise possède : quels résultats supplémentaires justifient la complexité introduite ?


> *Une approche multi-modèles doit démontrer son avantage net, en tenant compte du routage, de la supervision, de la qualité et de la gouvernance.*
>
> CE QU'IL FAUT RETENIR

## 01 Trois enseignements pour les décideurs

1. Un **modèle unique bien évalué** est une vraie option stratégique, notamment lorsque les usages sont homogènes, les volumes faibles ou les contraintes strictes.

2. Le multi-LLM peut créer de la valeur sur des **portefeuilles de tâches hétérogènes**, mais un bénéfice sur benchmark public n'est pas automatiquement transférable à l'activité métier.

3. Le bon niveau de décision est le **travail terminé et accepté**, en intégrant qualité, reprises, temps, risque, coût de routage et coût de gouvernance.

## 02 Ce que les recherches permettent d'affirmer — et leurs limites

La performance des modèles reste contrastée selon les tâches. Le Stanford AI Index 2026 observe que les principaux systèmes se rapprochent en classement général, tout en conservant des écarts selon les domaines et les épreuves. L'étude souligne aussi les limites et défauts de certains benchmarks : une première place ne garantit pas la meilleure performance sur les documents, les contraintes ou le secteur d'une entreprise [1].

Les travaux RouteLLM, publiés à ICLR 2025, apprennent à orienter les demandes vers un modèle plus puissant ou plus économique à partir de données de préférences humaines. Les auteurs rapportent, dans certains cas, une réduction de coût supérieure à un facteur deux sans perte de qualité sur les évaluations étudiées [2]. C'est une preuve de possibilité technique dans un dispositif expérimental, pas une économie garantie, ni un benchmark de Spotwork.AI. Les travaux antérieurs FrugalGPT étudient également les cascades de modèles, avec des résultats expérimentaux importants, mais dépendants des prix, tâches, modèles et métriques de leur époque [3].

À l'inverse, les guides de production Microsoft recommandent de comparer le routeur à un déploiement direct de référence, de mesurer séparément les familles d'usage et d'examiner les latences de queue, notamment le p95, ainsi que les contraintes de régions et de politique [4]. L'orientation vers plusieurs modèles est donc une hypothèse à tester, et non une bonne pratique universelle.

ROUTELLM · ICLR 2025

**> ×2**

réduction de coût rapportée dans certains cas, sans perte de qualité sur les évaluations étudiées

Preuve de possibilité, pas économie garantie

FRUGALGPT · 2023

### Cascades de modèles

résultats expérimentaux importants, dépendants des prix, tâches, modèles et métriques de leur époque

MICROSOFT LEARN · 2026

### Évaluer le routeur

comparer à un déploiement direct, mesurer par famille d'usage, examiner le p95, les régions et la politique

## 03 Quatre stratégies différentes, souvent confondues

| Stratégie | Principe | À envisager si... | Risque principal |
|---|---|---|---|
| **Modèle unique** | Un modèle sert les usages dans un dispositif gouverné | Besoins homogènes, petit périmètre, simplicité prioritaire | Dépendance et inadéquation sur certains cas |
| **Sélection statique** | Un modèle par tâche, domaine ou niveau de risque | Familles d'usages stables et identifiables | Règles vieillissantes, exceptions mal traitées |
| **Routage dynamique** | Un mécanisme choisit le modèle autorisé pour chaque demande | Variabilité de difficulté, volumes suffisants pour apprendre et mesurer | Erreurs de classification, coûts et latence supplémentaires |
| **Cascade / coopération** | Plusieurs appels successifs ou parallèles pour produire, vérifier ou améliorer | Les gains de qualité compensent les coûts et le délai | Reprises, désaccords, contamination des erreurs, surconsommation |

**À DISTINGUER**

Il faut distinguer multi-fournisseur et multi-appel : disposer de deux fournisseurs de secours ne signifie pas envoyer systématiquement chaque document à deux modèles. De même, un routeur n'est pas nécessairement une équipe d'agents ni une boucle de révision automatique.

# 04 La règle de décision : deux portes avant de regarder le prix

Cette règle applique au choix d'une architecture les deux portes du cadre de sélection Spotwork.AI, présentées dans notre publication « Comment choisir un modèle d'IA pour son entreprise ? Six critères et une méthode de décision» : on écarte d'abord, on compare ensuite.

PORTE 1 / ADMISSIBILITÉ

**Et risques bloquants**

- ☐ Données, région, contrats
- ☐ Sorties et outils nécessaires
- ☐ Permissions, rétention, transfert
- ☐ Connecteurs et journalisation
- ☐ Fuite d'information, erreur réglementaire

Un modèle non admissible est exclu, même s'il obtient le meilleur score ou le prix le plus bas.

→

PORTE 2 / COMPARAISON

**Valeur nette et stabilité**

- Résultat accepté défini au préalable
- Coût complet par livrable accepté
- Délai p50/p95 et reprises
- Stabilité par famille de demandes
- Frais de routage et de gouvernance

Si la différence est faible ou incertaine, conserver l'architecture la plus simple.

### Porte 1 — Admissibilité et risques bloquants

Le modèle doit être autorisé pour les données, la région, les contrats, les sorties et les outils nécessaires. Un modèle non admissible est exclu, même s'il obtient le meilleur score ou le prix le plus bas. Cela inclut permissions, rétention, transfert de données, connecteurs et journalisation. Les modèles disponibles dans le routeur doivent être limités à ceux qu'autorise l'entreprise. Les risques bloquants relèvent de la même porte : une moyenne élevée ne compense pas une fuite d'information ou une erreur réglementaire bloquante. Les risques liés aux injections, aux données sensibles et à la consommation non bornée sont notamment documentés par l'OWASP Top 10 for LLM Applications 2026 [5].

### Porte 2 — Comparaison : qualité acceptée, valeur nette et stabilité

Définir au préalable ce qu'est un résultat accepté : exactitude vérifiée, couverture des exigences, sources, format, revue humaine et absence d'incident critique. Comparer seulement les stratégies qui ont passé la première porte. Mesurer le coût complet par livrable accepté, le délai p50/p95, les reprises et la stabilité des résultats sur plusieurs familles de demandes. Ajouter les frais de routage, évaluateurs, normalisation des formats, observabilité, intégrations, maintenance et incidents. Si la différence est faible ou incertaine, conserver l'architecture la plus simple.

# 05 Exemple : trois familles de travail dans une direction des risques

SIMULATION PÉDAGOGIQUE FICTIVE aucun test réel ni performance produit revendiqués

Une direction utilise l'IA pour (A) classer des demandes simples, (B) synthétiser des dossiers documentaires et (C) préparer des analyses de risque à forte conséquence.

## A

**Classer des demandes simples**

Un modèle économique peut suffire.

## B

**Synthétiser des dossiers documentaires**

Un modèle performant sur la synthèse et la citation de sources peut être préférable.

## C

**Analyses de risque à forte conséquence**

Le système pourrait retenir un modèle avancé et une validation humaine renforcée — mais aucune réponse du modèle ne vaut à elle seule décision de conformité.

Trois variantes sont comparées. Le test doit inclure des demandes ambiguës, afin de mesurer les erreurs de routage, pas seulement les cas faciles.

**S1** · modèle unique de référence

**S2** · modèle attribué par famille

**S3** · routeur qui classe chaque demande, avec règle de repli vers un modèle autorisé

LECTURE DE LA DÉCISION

Imaginons que S2 coûte moins cher et soit plus rapide en moyenne. Si quelques demandes critiques sont mal orientées, **la porte d'admissibilité peut invalider S2** malgré son économie. Inversement, si S1 atteint les seuils sur tout le portefeuille avec une gouvernance plus simple, déployer S3 peut être un surinvestissement. Ces scénarios sont des cas de décision, pas des résultats mesurés.

# 06 Le protocole en six décisions

Ces six décisions s'inscrivent dans les quatre étapes présentées dans notre publication « Comment choisir un modèle d'IA pour son entreprise ? Six critères et une méthode de décision» : cadrer (1–2), écarter (3), comparer (4–5), décider et réévaluer (6).

ÉTAPE 1

## Cadrer

**1. Segmenter le portefeuille.**

Distinguer tâches, volumes, sensibilité, conséquences d'erreur, complexité et variations temporelles.

**2. Établir un référentiel.**

Tester d'abord un modèle unique correctement configuré sur des cas représentatifs ; conserver les versions, paramètres et prompts.

ÉTAPE 2 · PORTE 1

## Écarter

**3. Définir les seuils bloquants.**

Autorisations, taux d'erreurs critiques, qualité minimale, latence maximale et budget complet.

ÉTAPE 3 · PORTE 2

## Comparer

**4. Comparer progressivement.**

Tester une sélection statique, puis seulement si nécessaire un routage dynamique ou une cascade ; garder constants les autres composants.

**5. Rejouer et surveiller.**

Répéter les cas incertains selon criticité, borner essais et coûts, vérifier les régressions sur les cas rares, journaliser les changements de modèle et de version.

ÉTAPE 4

## Décider et réévaluer

**6. Arbitrer et maintenir la sortie.**

Décider « conserver le modèle unique », « spécialiser », « router », « corriger » ou « arrêter ». Préparer un repli en cas d'indisponibilité et réévaluer lorsque prix, modèles ou tâches changent.

CONSERVER LE MODÈLE UNIQUE | SPÉCIALISER | ROUTER | CORRECTION PUIS RETEST | NO-GO — ARRÊT

Une comparaison contrôlée ne dispense pas d'un essai progressif en conditions d'utilisation. Les recommandations d'évaluation OpenAI rappellent de calibrer les scores automatisés sur des évaluations humaines et de garder un jeu de tests représentatif, avec une surveillance continue [6].

## 07 L'indépendance ne vient pas du nombre de modèles

Une entreprise peut utiliser trois fournisseurs et rester dépendante d'un format de sortie, d'un outil de recherche, d'une mémoire conversationnelle ou d'une interface spécifique. La réversibilité exige des données et artefacts conservés sous son contrôle, des tests portables, des spécifications de sorties, des permissions et une preuve de remplacement sur un cas réel. Ce point complète notre publication « IA souveraine : comment arbitrer cloud, modèles ouverts et contrôle des données ?».

L'architecture doit aussi permettre de constater pourquoi une requête a été orientée vers un modèle : contexte de la décision, contraintes appliquées, version, résultat, coût et éventuel repli. En environnement réglementé, la responsabilité et la validation restent organisées au niveau du travail, pas déléguées au routeur.

## 08 À quelles conditions le multi-LLM est-il une mauvaise idée ?

Il l'est :

- si les cas d'usage sont rares ou identiques
- si le coût de maintenance dépasse les gains
- si les jeux d'évaluation sont insuffisants
- si le routage allonge excessivement les délais
- si la politique de données devient difficile à faire respecter
- si personne ne sait expliquer une décision de routage

Il peut aussi être inutile lorsque les besoins sont couverts par un modèle unique accompagné d'outils, de bonnes sources, d'un processus de contrôle et de versions maîtrisées. L'évaluation de l'ensemble du système prime alors sur la multiplication des fournisseurs.

> *Plus de modèles ne signifie ni plus d'intelligence, ni plus de souveraineté.*
>
> 08 / QUAND LE MULTI-LLM EST UNE MAUVAISE IDÉE

# 09 La décision à retenir

## Une stratégie multi-LLM doit être gagnée par les preuves.

Évaluer un modèle unique comme référence, poser des conditions éliminatoires, mesurer le travail accepté, puis n'activer une spécialisation ou un routeur que si les gains restent visibles après coût et gouvernance. L'entreprise préserve ainsi sa capacité de choisir demain, sans payer aujourd'hui pour une complexité non justifiée.

1. **Modèle unique comme référence**
2. **Conditions éliminatoires**
3. **Mesurer le travail accepté**
4. **Spécialiser ou router si gain net visible**

ANNEXES MÉTHODOLOGIQUES

Les annexes qui suivent approfondissent la mise en œuvre : comparaison loyale des configurations (A), modèle d'économie en cohortes (B), gestion des erreurs de routage (C), sécurité, souveraineté et continuité (D), protocole de replay et décision (E), éléments à valider par les responsables (F).

# A Comment définir une comparaison loyale ?

La bonne unité d'évaluation est la tâche terminée et acceptée, pas la simple réponse textuelle. Comparer au minimum le modèle unique de référence (S1) et une alternative justifiée (S2, modèle attribué statiquement selon la famille d'usage), en laissant l'option de ne pas adopter le multi-LLM. N'ajouter S3 que si un routeur dynamique ou une cascade répond à une hypothèse distincte et démontrable ; préciser alors laquelle de ces deux architectures est testée, afin d'attribuer son surcoût et son effet. Une équipe peut conclure après S1 et S2.

Les entrées doivent être comparables : mêmes documents autorisés, mêmes exigences, mêmes critères d'acceptation, même budget maximum, mêmes niveaux de supervision. Les prompts ne sont pas nécessairement syntaxiquement identiques entre modèles, mais toute adaptation doit être documentée pour ne pas attribuer au routage un gain dû à un meilleur contexte ou à une instruction différente.

Ne pas comparer uniquement les cas sur lesquels les modèles divergent fortement. Couvrir aussi la fréquence réelle, les cas simples, les cas critiques, les cas ambigus, les longs documents et les exceptions. Mesurer chaque famille séparément : la performance moyenne peut masquer une baisse grave sur les demandes réglementaires.

| Porte | Mesures d'entrée | Décision permise | Erreur interdite |
|---|---|---|---|
| **Porte 1 · Admissibilité** | Modèle et endpoint autorisés, région, contrat, permissions, outils, conservation, droits d'audit | Exclure les modèles non admissibles | Laisser le routeur contourner une restriction |
| **Risques bloquants (porte 1) et qualité (porte 2)** | Erreurs critiques, livrable accepté, justificatifs, couverture, tests contradictoires, corrections | Rejeter ou corriger les stratégies défaillantes | Compenser un incident critique par un score moyen |
| **Porte 2 · Valeur nette** | Coût complet, relectures, coûts du routeur, p50/p95, maintenance, réversibilité | Choisir la solution admissible la moins complexe si résultats comparables | Confondre économie API et valeur métier |

Chaque porte doit comporter un responsable, un jeu de tests, un seuil défini avant essai, un mécanisme de dérogation exceptionnel approuvé, limité aux règles que l'organisation est habilitée à modifier, et une trace. Un service interdit pour la donnée ou la finalité reste exclu, y compris en repli : une indisponibilité ne le rend jamais admissible. Ce tableau est une proposition méthodologique et non une certification réglementaire ou une norme ISO.

# B Modèle d’économie en cohortes

### Dénominateur

Ne compter que les livrables acceptés selon le barème convenu. Inclure dans le coût les demandes rejetées, les essais supplémentaires, les évaluations et les interventions humaines. Les dossiers non aboutis restent dans le numérateur, sous réserve de la politique d’allocation documentée.

### Coût technique du routage

Ajout possible d’un appel de classification, de jetons et du monitoring. Une cascade peut entraîner deux ou plusieurs appels, parfois sur les mêmes documents. Le nombre d’appels ne détermine pas seul la qualité, et un prix inférieur à la requête ne garantit rien sur le coût d’un travail fini.

### Coût opérationnel de la diversité

Tests d’intégration par fournisseur ; adaptateurs de schéma ; règles de rétention ; journaux ; gestion des versions et des changements de prix ; incidents et secours ; documentation des décisions. Répartir les frais fixes selon une clé d’allocation explicite.

### Économie attribuable

Comparer les configurations sur la même cohorte, puis reconstituer les dépenses réellement évitées ou la capacité utile. Le multi-LLM peut accroître la qualité sans réduire le coût : dans ce cas, l’argument d’investissement doit porter sur le bénéfice métier, avec les preuves adaptées.

## C Gestion des erreurs de routage

Le mode d'échec le plus important n'est pas toujours l'hallucination du modèle final. Il peut survenir avant l'appel : une demande réglementaire est classée « extraction simple », un document sensible est envoyé vers un service non autorisé, ou un résultat de vérification est considéré comme une preuve indépendante alors que deux modèles reposent sur la même source.

Pour chaque routage, conserver l'ID du run, les exigences et permissions applicables, les modèles candidats admissibles, le motif de sélection compréhensible, la version du routeur, le modèle exécutant et sa configuration, le nombre d'appels, les erreurs et replis. Éviter d'exposer une chaîne de pensée interne ou des données sensibles dans la télémétrie.

**Garde-fou déterministe**
Exclure les modèles interdits avant la sélection, y compris pour le repli.

**Garde-fou métier**
Imposer les validations humaines requises après la réponse.

**Garde-fou économique**
Plafonner latence, retries, coût et nombre de cascades.

**Garde-fou de changement**
Exiger des tests de non-régression au changement du routeur ou d'un modèle critique.

## D Sécurité, souveraineté et continuité

Le multi-LLM peut réduire la dépendance vis-à-vis d'un fournisseur, mais aussi multiplier les tiers ayant accès aux données. La réversibilité exige une politique de conservation des documents, des historiques, des métriques et des outputs indépendante des fournisseurs ; des adaptateurs versionnés ; un contrat de sortie testé et une alternative réellement admissible. Un « fallback » qui viole les règles régionales ou la confidentialité n'est pas un secours valide.

Les contrôles de risque s'appliquent aussi au routeur : injections présentes dans des documents externes, traitement de données sensibles, escalades de droits via outils et consommation non bornée. Une architecture peut être simple tout en ayant un bon niveau de contrôle ; une architecture sophistiquée ne prouve pas par elle-même la sécurité.

# E Protocole de replay et décision

Une expérimentation peut suivre quatre phases en 90 jours indicatifs :

**PHASE 1**
Définir baseline et critères d'acceptation

**PHASE 2**
Comparaison hors production incluant cas difficiles et variantes

**PHASE 3**
Essai contrôlé avec journalisation et quotas

**PHASE 4**
Maintenir le modèle unique, adopter la spécialisation, piloter une route dynamique, retester ou arrêter

Le calendrier n'est ni obligatoire ni adapté à tous les risques. Le nombre de cas et les répétitions ne sont pas uniformes : définir une stratégie selon complexité, variabilité et conséquences ; fixer une borne en nombre de tests, dépenses et durée. Mesurer les violations critiques indépendamment des moyennes. Rejouer lorsque changent modèle, outils, version du routeur, règles de protection ou sources structurantes. Aucun test métier n'a été exécuté dans le cadre de cette étude.

# F Éléments à valider par les responsables

| | |
|---|---|
| **DSI / architecture** | Intégrations, latence p95, coût de maintenance, versionnement, mécanismes de repli. |
| **RSSI / DPO** | Politique des fournisseurs et traitements, droits, journaux, sources sensibles, politique de conservation. |
| **Métier** | Pertinence des familles de tâches, barème de qualité, erreurs et relectures obligatoires. |
| **DAF / FinOps** | Coût complet, clefs de ventilation, période d'évaluation, bénéfices réalisables. |

# 10 Sources et méthode

[1] Stanford HAI, AI Index 2026 — Technical Performance.
Synthèse de benchmarks, limites des classements.
hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance

[2] Ong et al., RouteLLM: Learning to Route LLMs from Preference Data, ICLR 2025.
Expérimentation de recherche ; résultats non généralisables sans validation métier.
proceedings.iclr.cc/paper_files/paper/2025/hash/5503a7c69d48a2f86fc00b3dc09de686-Abstract-Conference.html

[3] Chen, Zaharia & Zou, FrugalGPT, 2023.
Prépublication de recherche ; prix et modèles historiques.
arxiv.org/abs/2305.05176

[4] Microsoft Learn, Evaluate model router for your workload, 2026.
Guide fournisseur, non audit indépendant.
learn.microsoft.com/en-us/azure/foundry/openai/how-to/evaluate-model-router

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

[6] OpenAI, Evaluation best practices.
Guide fournisseur, importance des évaluations représentatives.
developers.openai.com/api/docs/guides/evaluation-best-practices

[7] Artificial Analysis, Benchmarking Methodology.
Méthodologie comparative externe ; le fournisseur, le point d'accès et la configuration influent sur les mesures.
artificialanalysis.ai/methodology

**MÉTHODE**

Analyse documentaire et proposition méthodologique de Spotwork.AI ; aucun résultat de test, baisse de coût ou bénéfice de Spotwork.AI n'est revendiqué. Les métriques et seuils proposés doivent être adaptés au portefeuille étudié. L'étude propose un cadre de décision, pas une preuve de performance ou un comparatif contractuel entre fournisseurs. Elle ne constitue pas non plus une évaluation d'impacts réglementaires. Sources consultées au 2 novembre 2026.

## Pour aller plus loin

**I0-01**
Le cadre de sélection commun : six critères, deux portes, quatre étapes

**P04**
Les classements de modèles et leurs limites

**P05**
Le coût complet par livrable accepté

**P21**
Les conditions concrètes de souveraineté et de réversibilité

Spotwork.AI

L'IA professionnelle de confiance

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

SPOTLAB · RESEARCH & INSIGHTS · ANALYSE P06

# Multi-LLM ou modèle unique : quand l'orchestration de plusieurs IA crée-t-elle vraiment de la valeur ?

Modèle unique, spécialisation, routeur ou cascade



© 2026 SPOTWORK.AI