

Architecture · Sémantique · Contexte · Exécution

Ce que doit avoir une plateforme IA aujourd'hui

Un référentiel d'architecture pour comprendre,
concevoir et évaluer une plateforme IA professionnelle.

Des mécanismes techniques aux preuves de maturité : organiser les connaissances,
maîtriser le contexte, exposer les capacités et conduire un travail vérifiable.

Structurer

Ontologie métier, catalogue de
capacités et cycle de vie des
données.

Orchestrer

Contexte, mémoire, RAG,
modèles et exécution durable.

Gouverner

MCP, identité, effets,
observabilité et évaluations.

La profondeur technique se démontre par les contrats, les frontières de confiance et les résultats mesurés.

Architecture de référence indépendante d'un produit.

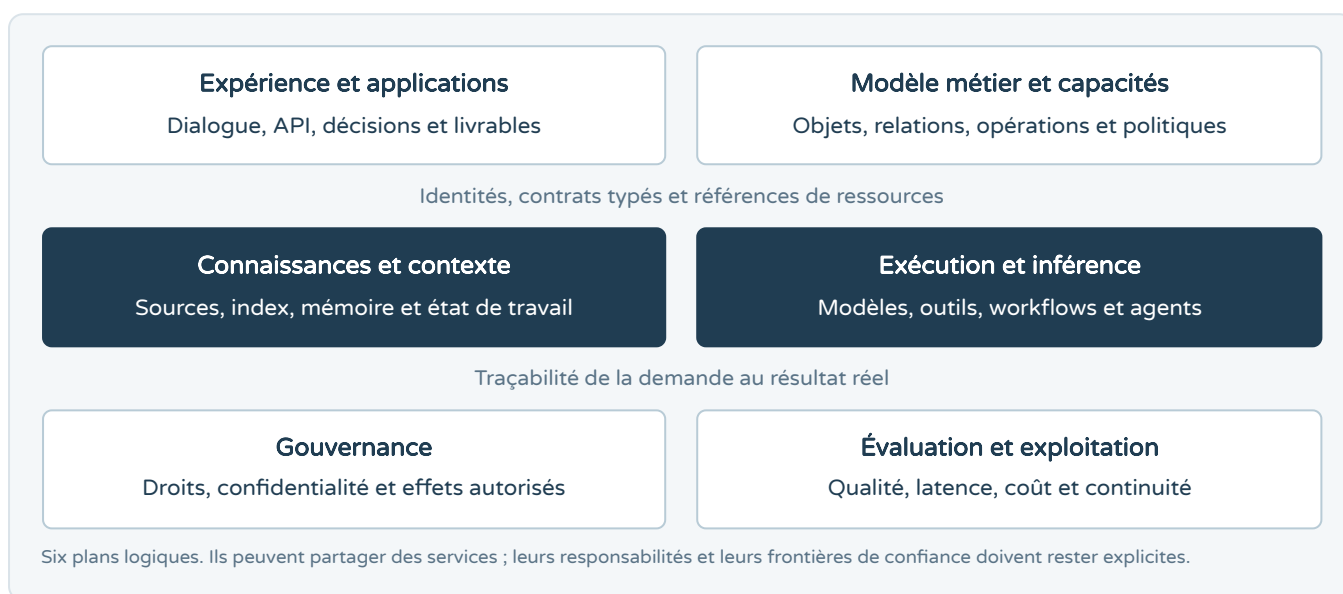
Illustrations ponctuelles issues de WorkProba.

Édition révisée · Improba · Octobre 2026

Ce qu'un responsable IA doit pouvoir évaluer

Une plateforme IA mature organise les connaissances, le contexte et les actions autour de contrats explicites, avec des résultats mesurables.

Ce document s'appuie sur le projet WorkProba pour illustrer les concepts présentés.



Trois niveaux de lecture

Le socle attendu comprend l'identité, les sources versionnées, les interfaces d'outils, les contrôles d'exécution et les évaluations. **Les choix d'architecture** portent sur le routage, la recherche, la mémoire et la persistance. **Les options avancées**, telles qu'un graphe de connaissances ou plusieurs agents spécialisés, se justifient par un besoin et un gain vérifié.

Le fil conducteur du rapport

Un scénario fictif de revue contractuelle sert d'exemple : retrouver deux versions, identifier des clauses, comparer leurs effets et produire une note vérifiable. Les sections expliquent les contrats techniques qui permettent de faire évoluer ce scénario vers un travail métier durable.

Critère de maturité. Chaque brique doit avoir une responsabilité, une source d'autorité, des limites et une preuve de bon fonctionnement. L'accumulation de composants ne suffit pas à établir la qualité de l'ensemble.

Parcours : sémantique p. 3 ; capacités p. 4 ; contexte p. 5-6 ; données et RAG p. 7-8 ; preuves p. 9 ; MCP p. 10-11 ; exécution p. 12 ; modèles p. 13 ; exploitation p. 14 ; évaluation p. 15 ; grille de maturité p. 16.

Donner un sens commun aux objets et aux actions

Un modèle sémantique explicite les concepts et leurs relations. Une ontologie formelle peut le compléter lorsque le domaine exige des règles d'inférence.

Un modèle exploitable identifie les entités, leurs propriétés, les relations autorisées et les sources qui font foi. Il distingue une entreprise, un contrat, une version, une clause, une décision et une action. Les identifiants stables relient ces objets aux systèmes d'origine.

Couche sémantique	Rôle technique
Vocabulaire / taxonomie	Termes préférés, synonymes, langues, catégories et hiérarchies. SKOS fournit un repère pour les vocabulaires contrôlés.
Modèle métier	Types d'objets, relations, attributs, unités, états et transitions. Implémentation possible en schémas et services de domaine.
Ontologie formelle	Axiomes et relations interprétables par machine ; OWL est une option lorsque des besoins d'inférence le justifient.
Graphe de connaissances	Instances et relations concrètes entre objets, avec provenance et droits. Il peut servir aux parcours de relations et à certaines recherches.

Résoudre les ambiguïtés et le temps

« ABC », une raison sociale et un identifiant CRM peuvent désigner le même acteur. La résolution d'entités doit conserver les correspondances, leur origine et les ambiguïtés restantes. Une similarité de nom ne suffit pas à fusionner deux identités.

Une clause possède une version et une période de validité. La date de collecte et la date d'effet sont différentes. Les faits dérivés gardent leur source, leur statut et les transformations subies. PROV-O constitue un repère pour représenter cette provenance.

Client → est partie à → Contrat → possède → Version
Version → contient → Clause → fait l'objet de → Décision

Les droits s'appliquent aux objets et aux relations parcourues. Une relation entre un dossier et un document ne doit pas publier le document à tous les lecteurs du dossier.

ILLUSTRATION · WORKPROBA

Le modèle distingue Conversation, Espace, Dossier, processus, Mission et livrable. Cela illustre une sémantique applicative explicite ; les sources examinées n'établissent pas une ontologie OWL ni un graphe de connaissances généralisé. [W1]

Repères : [R1], [R2], [R3]. Un stockage graphe est un choix d'implémentation ; sa présence ne suffit pas à qualifier la sémantique du domaine.

Construire un catalogue de capacités composables

Une capacité décrit un résultat accessible, les données nécessaires et les effets possibles. Le catalogue relie cette intention aux contrats d'exécution.

Connecteur	Accès administré à une application ou une source, avec identité, périmètre et configuration.
Outil / action	Opération typée : lire une version, comparer des clauses, préparer un brouillon ou publier un résultat.
Procédure / workflow	Enchaînement dont les transitions et les exceptions sont explicites ; certaines étapes peuvent appeler un modèle probabiliste.
Agent	Sélection dynamique d'étapes et d'outils dans un objectif, un budget et des permissions bornés.
Processus métier	Coordination durable de participants, livrables, validations et dépendances.

Le contrat d'une capacité

Le catalogue doit préciser identifiant et version, description d'usage, schémas d'entrée et de sortie, préconditions, permissions, classe d'effet, délais, quotas, modalités de reprise et éléments de preuve attendus. L'unité d'un montant, les états d'une facture et les erreurs récupérables appartiennent à ce contrat.

La disponibilité résulte de plusieurs filtres : installation, configuration, droits de l'utilisateur, politique de l'organisation et restrictions de l'exécution. La découverte retourne seulement les définitions accessibles, puis l'appel vérifie de nouveau l'autorisation.

Découverte progressive et gestion du changement

Un catalogue volumineux gagne à proposer une recherche de capacités, une fiche détaillée et le chargement des seuls schémas utiles au tour. Cette sélection réduit le contexte d'outils et les choix ambigus. Le nom, la version et le sens d'une opération doivent rester stables entre découverte et exécution.

Une mise à jour qui ajoute une permission ou change un effet exige une nouvelle revue. Les appels et approbations en attente restent liés à une définition précise ; ils ne migrent pas silencieusement vers une opération différente.

ILLUSTRATION · WORKPROBA

Les définitions de plugins, leurs versions publiées, l'acceptation par une organisation et la connexion sont séparées. La découverte métier expose des agents, procédures et processus, avec recherche puis consultation d'une définition. [W4, W6]

Preuve à examiner. Retirer une permission après découverte, puis appeler l'ancienne définition : le serveur doit refuser l'opération concernée.

Les contrats ci-dessus forment une proposition d'architecture. MCP peut transporter une partie des interfaces ; la sémantique et les politiques métier demeurent applicatives.

Piloter ce qui est transmis à chaque appel modèle

Le contexte est une projection calculée de l'état de travail. Sa composition doit être explicable, bornée et adaptée au modèle choisi.

Avant chaque appel, un assembleur sélectionne les instructions applicables, l'objectif courant, les décisions humaines, l'historique pertinent, les souvenirs autorisés, les passages documentaires et les résultats d'outils. Les éléments retenus gardent leur provenance et leur rôle : instruction, déclaration, observation ou inférence. Un score éventuel de confiance doit être calibré.

Budget d'entrée = fenêtre utilisable – réserve de sortie – marge technique
Entrée = instructions + schémas + historique + état + sources + résultats

Allouer la capacité selon la tâche

Poste	Traitement	Règle d'allocation
Instructions et outils	Sélection	Politiques et schémas effectivement nécessaires.
Historique récent	Adaptation	Dialogue utile et échanges en cours.
État / mémoire	Priorisation	Objectif, contraintes et faits sourcés.
Sources / outils	Récupération ciblée	Passages et observations sélectionnés.
Sortie réservée	Réservation	Réponse et raisonnement selon leur comptabilisation.
Marge	Dimensionnement	Variations et sûreté du dimensionnement.

Aucune proportion fixe ne convient à tous les usages. La somme entrée, sortie réservée et marge respecte la fenêtre applicable, ainsi que les limites propres d'entrée et de sortie. Le tokenizer et la comptabilisation du raisonnement dépendent du fournisseur.

Arbitrer sans perdre les contraintes essentielles

L'ordre de traitement peut combiner déduplication, sélection de passages, réduction des résultats volumineux, récupération à la demande et compactage. Le choix doit préserver les interdictions, décisions et éléments nécessaires à la vérification des actions.

Lorsqu'un résultat d'outil dépasse le budget, la vue envoyée au modèle peut fournir un extrait, une limite de couverture et une référence de continuation. Le journal conserve le résultat canonique autorisé. Une omission doit rester détectable et récupérable.

Preuve à examiner. Faire croître le contexte jusqu'à la limite et vérifier que les contraintes initiales restent accessibles. Si le contexte essentiel ne tient plus, l'exécution doit demander un autre périmètre ou une capacité adaptée.

Synthèse d'architecture : [R4], [R5], [R6], [R19]. Les estimations de tokens doivent être confrontées aux usages réels de l'API.

Séparer journal, contexte actif et état durable

Quatre couches distinctes permettent de reprendre une tâche sans transformer un résumé du modèle en vérité métier.

Couche	Contenu	Statut
Journal canonique	Messages, appels, résultats et décisions, avec politique de rétention.	Trace consultable ; accès contrôlé.
Contexte actif	Projection envoyée pour l'appel courant.	Partiel et reconstituable.
Mémoire durable	Préférences et faits sélectionnés, avec source, propriétaire et date.	Révisable, oubliable et périssable.
État d'exécution	Étape, objets visés, autorisations, reçus et actions en attente.	Autorité des services métier.

Compacter avec des références réouvrables

Le compactage peut retenir objectif, contraintes, décisions, questions ouvertes et prochaine étape. Les éléments importants référencent les messages ou résultats d'origine. Une correction ultérieure doit remplacer explicitement l'ancienne orientation, avec une révision identifiable.

La récupération ciblée du dialogue écarté complète ce résumé. Le système conserve les liens entre appels d'outils et résultats, les décisions humaines et les limites connues des sources. Les procédures d'effacement doivent couvrir journal, mémoire, index et caches selon leurs politiques.

Contrôler les conflits et la fraîcheur

Une mémoire indique si le fait vient d'une déclaration, d'une observation de connecteur ou d'une inférence. Des faits contradictoires demandent une règle de priorité et une vérification ; un fait ancien ne doit pas être présenté comme une observation récente.

Le cache de préfixe d'inférence peut réduire le retraitement de contenu stable selon le fournisseur. Il ne remplace ni la mémoire métier ni les contrôles d'accès. Sa durée de vie, son isolation et ses règles d'invalidation sont des paramètres distincts.

ILLUSTRATION · WORKPROBA

Le code sépare le journal du contexte projeté. Un état de travail versionné référence des extraits d'origine ; les résumés ne deviennent pas des décisions humaines. Les passages écartés peuvent être relus et l'état métier doit être récupéré dans ses outils autorisés. [W2]

Repères : [R4], [R5], [R6], [R7]. Les mécanismes de compactage exigent des tests de conservation des contraintes, de corrections tardives et de reprise.

Industrialiser l'ingestion et la qualité documentaire

Un index fiable décrit ce qu'il contient, la version dont il provient et les limites de sa préparation.

01

Acquérir

Source, identité, droits et version.

02

Extraire

Structure, OCR et visuels utiles.

03

Indexer

Passages, métadonnées et vecteurs.

04

Maintenir

Révisions, révocations et effacement.

Des transformations traçables

L'ingestion enregistre empreinte, type MIME, langue, dates, origine et état de traitement. Les transformations sont versionnées : parseur, OCR, règles de découpage et modèle d'embedding. Cette chaîne permet d'expliquer un résultat et de reconstruire un index lors d'une évolution.

L'extraction conserve les titres, listes, tableaux et repères de pages, cellules ou diapositives. Le découpage peut utiliser des relations parent/enfant et des en-têtes répétés. Une représentation multimodale exige des liens entre texte, figure, légende et page d'origine.

Rendre explicites les limites et les coûts

Une feuille de calcul peut contenir formules et valeurs mises en cache périmées. Un PDF peut mêler texte et scans. Une extraction partielle doit exposer sa couverture ; une page OCRisée ne démontre pas l'interprétation d'un graphique.

Les grands documents nécessitent files de traitement, lots, plafonds et points de reprise. La déduplication par empreinte doit intégrer le périmètre de sécurité et les paramètres de transformation. Réutiliser des vecteurs suppose que les passages et le modèle restent compatibles.

Faire vivre les index avec les droits

Une révocation doit empêcher une nouvelle lecture, y compris depuis un cache. La suppression ou l'archivage déclenche une politique cohérente pour les fichiers, index, dérivés et sauvegardes. La réindexation publie une génération identifiée et conserve les références historiques selon les règles de rétention.

ILLUSTRATION · WORKPROBA

La chaîne documentaire conserve la structure de plusieurs formats, des locateurs, une couverture d'extraction et des versions. Elle comporte un OCR sélectif, un cache borné par organisation et des reprises de lots administrées. [W3]

Preuve à examiner. Indexer un document incomplet puis le révoquer : aucune réponse ultérieure ne doit prétendre à une lecture complète ou restituer son contenu via un ancien résultat.

Synthèse d'ingénierie documentaire. La profondeur du traitement visuel doit être choisie et évaluée sur les formats réellement utilisés.

Router la demande vers la bonne méthode de recherche

Le RAG assemble des preuves adaptées à une question. Les stratégies se choisissent selon la nature de la réponse attendue.

Question	Stratégie utile	Point de contrôle
Référence ou code exact	Recherche lexicale, filtres et correspondances d'identifiants.	Rappel sur codes rares.
Formulation différente	Recherche dense combinée au lexical.	Paraphrases et langues.
Comparer des versions	Recherche puis lecture ciblée de chaque version.	Couverture symétrique.
Compter ou calculer	Requête structurée et calcul déterministe sur données autorisées.	Unités et population.
Relations / synthèse globale	Parcours d'entités, graphe ou résumés hiérarchiques si pertinents.	Provenance et coût.

Une chaîne de sélection à plusieurs étages

Le système détermine le périmètre autorisé, prépare la requête et récupère des candidats. Les résultats non autorisés ne sont transmis ni au modèle ni à un reranker externe. Il fusionne ensuite les classements. Le reclassement réduit ce jeu aux passages les plus utiles. La diversification limite les doublons ; l'expansion parent/enfant ou par voisinage rétablit le contexte nécessaire.

Les étapes doivent conserver la requête d'origine et leurs transformations. Une reformulation peut altérer un nom ou élargir la population. Les paramètres sont mesurés à budget comparable : candidats, passages finaux, tokens, latence et coût.

Orchestrer une recherche itérative

Pour une question composite, un agent peut décomposer en sous-questions, comparer les éléments trouvés et effectuer une lecture complémentaire. Il doit disposer d'un critère d'arrêt, de limites d'appels et d'une politique en cas de preuve insuffisante. La synthèse globale peut justifier un GraphRAG ; une recherche locale simple peut suffire pour une clause précise.

ILLUSTRATION · WORKPROBA

Le chemin Cloud combine recherche plein texte et vecteurs pgvector par RRF lorsque les embeddings sont disponibles. Des options de reclassement et de voisinage complètent la sélection. Cet exemple ne démontre pas un GraphRAG ni un planificateur universel de recherche. [W3]

Preuve à examiner. Comparer lexical seul, dense seul, hybride et hybride avec reranking sur le même jeu indépendant, en mesurant aussi les coûts et les cas sans réponse.

Repères : [R8], [R9], [R10]. Le graphe et la recherche agentique sont des options conditionnelles, à justifier par les questions métier.

Relier les affirmations aux sources réellement utilisées

Une réponse fiable permet de retracer une affirmation jusqu'à une source, sa version et les passages effectivement consultés.

Une preuve documentaire structurée

La référence conserve identifiant de source, version, empreinte, locateur et portion consultée. La couverture distingue résultat de recherche, lecture ciblée, lecture d'une plage et consultation de tout le contenu extractible. Un document partiellement extrait reste partiel après parcours de tous ses passages.

Dans une comparaison, les deux versions doivent être identifiées et consultées selon un périmètre cohérent. Une absence de résultat ne prouve pas qu'une clause n'existe pas. Une affirmation négative ou une conclusion exhaustive exige une couverture adaptée à sa portée.

Validation	Ce qu'elle peut établir
Déterministe	Présence de la citation littérale, calcul, unité, date ou identifiant ; cohérence d'un schéma.
Sémantique	Soutien d'une affirmation par les passages associés, avec jugement probabiliste à calibrer.
Métier / humaine	Adéquation au besoin, interprétation spécialisée et décision sur un livrable précis.

Traiter contradiction et incertitude

Le système explicite les versions en conflit, les sources obsolètes et les inférences. Il peut réduire la portée de sa réponse, demander une source complémentaire ou signaler que la conclusion n'est pas établie. La vérification n'accorde ni un droit supplémentaire ni une autorisation d'agir.

Pour un livrable, la chaîne conserve les sources, les transformations, les calculs et le résultat enregistré. Une confirmation textuelle du modèle doit être rapprochée d'un reçu de stockage ou d'un état observé dans l'application cible.

ILLUSTRATION · WORKPROBA

Des lectures bornées, des curseurs liés à une version et des reçus de consultation distinguent passages retrouvés et passages transmis pendant le tour. Des contrôles d'affirmations complètent cette chaîne, avec une portée explicitement limitée. [W3]

Preuve à examiner. Demander une conclusion d'absence alors qu'une partie du document manque : la réponse doit conserver cette limite, y compris après un résumé de conversation.

PROV-O [R3] apporte un repère pour la provenance. La fidélité d'une réponse exige une évaluation indépendante des seuls liens de citation.

Exposer des capacités lisibles et composables

MCP standardise l'accès à des capacités de serveurs. Une plateforme peut consommer des serveurs tiers et exposer ses propres services à des assistants externes.

Primitive	Responsabilité	Exemple fictif
Tools	Opérations appelables, décrites par un nom et des schémas.	Rechercher une version ; préparer une note.
Resources	Contenus adressables par URI, lisibles selon les capacités prises en charge.	Une source versionnée ; un résultat de Mission.
Prompts	Gabarits d'interaction paramétrés, selon les fonctions du serveur.	Une trame de revue contractuelle.

Concevoir le contrat d'un outil

Le schéma précise types, champs obligatoires, contraintes et erreurs possibles. La description indique quand utiliser l'opération, son effet métier et ses limites. Les résultats structurés exposent les données utiles, leur provenance, la pagination et l'état de couverture.

Le catalogue doit éviter les ambiguïtés de noms et maîtriser sa taille. Une découverte progressive peut rechercher les capacités pertinentes puis charger leurs détails. Les annotations d'effets sont des indications pour le client ; les contrôles de sécurité restent exécutés par le serveur.

Distinguer protocole et sémantique métier

Une même capacité doit conserver identifiants, unités, états et permissions entre interfaces internes et MCP. L'adaptateur MCP réutilise les services de domaine et traduit leurs résultats. Le protocole ne fournit pas à lui seul l'ontologie métier, les transactions ou la validation des livrables.

La matrice de compatibilité indique révisions du protocole, transports et primitives réellement supportées. Les schémas et versions sont testés avec les clients visés. L'exposition d'outils seule ne démontre pas la prise en charge de toutes les primitives MCP.

ILLUSTRATION · WORKPROBA

Le point d'entrée Cloud transporte des outils et des commandes de Missions sous des contrôles communs. La compatibilité annoncée dans les sources examinées est 2025-11-25 ; elle ne couvre pas la révision 2026-07-28 utilisée comme référence externe dans ce rapport. [W4]

Références : outils [R11], ressources [R12], prompts [R18]. Les trois primitives sont décrites au niveau conceptuel ; leur activation et leurs fonctionnalités dépendent du serveur et de sa révision.

Gouverner ce qui peut être découvert, lu et exécuté

Une interface interopérable doit préserver les droits actuels, le périmètre de délégation et les règles de décision de l'application métier.

Une chaîne d'identité vérifiable

Pour les transports HTTP protégés, le profil d'autorisation MCP s'appuie sur OAuth : découverte des serveurs d'autorisation, cible de ressource explicite, validation du destinataire et protection des flux de connexion. Les scopes sont demandés selon le besoin ; les jetons ne sont pas réutilisés comme justificatifs universels vers d'autres services.

Le serveur combine identité du client et de l'utilisateur, organisation, ressource visée, délégation et politique. L'autorisation est revérifiée à l'appel. Une connexion ancienne ou un outil encore visible dans le client ne conserve pas un droit révoqué.

Rendre les effets sensibles inspectables

Étape	Contrat attendu
Préparation	Cible, arguments, version et identité figés ; classe d'effet identifiée.
Décision	Revue authentifiée par la personne autorisée lorsque la politique l'exige.
Exécution	Nouvelle vérification des droits, des paramètres et des préconditions.
Résultat / reprise	Reçu durable, déduplication et traitement explicite de l'incertitude.

Traiter les risques propres aux intégrations

Les contenus et descriptions d'outils sont des entrées potentiellement hostiles. L'admission d'un serveur, la revue des changements, la validation des sorties et les limites réseau réduisent les risques de manipulation, de fuite et de délégation confuse. Une protection de prompt ne suffit pas à isoler un effet métier.

ILLUSTRATION · WORKPROBA

Les assistants externes disposent d'autorisations OAuth dédiées. Pour certaines écritures, une revue indépendante lie la décision aux arguments exacts. Les droits, révocations et restrictions des connecteurs restent contrôlés par les services Cloud. [W4]

Preuve à examiner. Révoquer un accès après préparation d'une écriture ou modifier ses arguments après approbation : l'effet doit être refusé et le refus traçable.

Références : autorisation [R13], bonnes pratiques de sécurité MCP [R14], autonomie des agents [R15].

Choisir le bon niveau d'autonomie

La boucle d'un agent doit s'inscrire dans un système d'exécution dont les états et les effets restent compréhensibles après une interruption.

Mode	Usage adapté	Responsabilité
Appel direct	Lecture, recherche ou transformation bien délimitée.	Contrat et résultat.
Workflow	Étapes connues, transitions et validations explicites.	Machine à états.
Agent	Chemin de résolution variable et besoin d'adaptation.	Objectif et limites.
Plusieurs agents	Spécialisations ou travaux indépendants apportant un gain.	Délégation, contexte et consolidation.

Des points de reprise fondés sur l'état

L'exécution persiste l'objectif, la version des définitions, les entrées, l'étape courante, les permissions déléguées, les décisions et les reçus. Les workers utilisent des baux et des barrières de concurrence pour éviter qu'une ancienne instance continue après réattribution.

Les délais, annulations et attentes humaines ont des statuts distincts. La fermeture d'un écran ne vaut pas annulation. Un nouveau worker reprend à partir des résultats confirmés et vérifie l'autorité actuelle avant un nouvel effet.

Concevoir la reprise des écritures

Une clé d'idempotence porte le périmètre et l'empreinte des arguments. Un appel terminé peut restituer son reçu. Après perte de réponse, un effet distant peut être incertain : la réconciliation avec le système cible précède toute répétition. Une compensation, lorsqu'elle existe, est une opération métier autorisée à part entière.

Une délégation entre agents transmet un objectif, un contexte minimal, des permissions bornées et un contrat de résultat. Le parent ne doit pas considérer la seule fin d'un agent comme une preuve de réalisation complète.

ILLUSTRATION · WORKPROBA

Les Missions possèdent un état Cloud indépendant de la présence de la conversation. Des résultats partiels et des effets incertains conservent leur statut, avec des limites de relance pour éviter de rejouer une opération déjà effectuée. [W5]

Preuve à examiner. Interrompre un worker juste après une écriture distante, puis redémarrer : inspecter le nombre d'effets réellement produits et l'état de réconciliation.

L'intérêt de plusieurs agents se mesure avec leur surcoût de contexte, de coordination, de latence et d'erreurs de consolidation. Il n'est pas un prérequis universel.

Piloter capacités, destinations et consommation

Une couche d'inférence commune permet de choisir les modèles selon les tâches et de faire évoluer les fournisseurs sans disperser les politiques.

Un catalogue opérationnel

Chaque entrée décrit fournisseur, identifiant exact, cycle de vie, modalités d'entrée, fenêtre de contexte, limites de sortie, outils, formats structurés et paramètres de raisonnement. Les restrictions d'usage, destinations admises et tarifs applicables sont versionnés.

Chemin	Politique de sélection
Dialogue / raisonnement	Compétence nécessaire, complexité, confidentialité et budget de la tâche.
Vision / OCR	Type de contenu, résolution, couverture et contrôle des pages transmises.
Embeddings / reranking	Cohérence de modèle, dimension, langue et coût des lots.
Utilitaires / vérification	Critères dédiés, budget indépendant et qualité calibrée.

Routage, replis et compatibilité

Une politique peut sélectionner un modèle avant le tour et journaliser le motif. Un repli doit rester compatible avec les modalités, outils, formats, destinations et contraintes de sécurité. Une erreur après un effet métier ne doit pas provoquer le rejeu automatique de cet effet avec un autre modèle.

Le RAG apporte des sources au moment de répondre. Un ajustement de modèle peut cibler un comportement ou un format ; il ne remplace pas l'actualisation des connaissances ni les droits. Les mises à jour exigent des évaluations comparatives, une validation des schémas et une stratégie de retour. Les formats d'historique spécifiques au fournisseur ne sont pas transférés sans conversion contrôlée. La version du modèle et celle du prompt doivent être retrouvables dans la trace.

Une économie par tâche réussie

L'admission peut réserver une enveloppe puis rapprocher estimation et usage réel. Les limites couvrent tokens, appels, concurrence et durée, avec facturation spécifique pour OCR, recherche ou cache. Le tableau de bord rapproche coût, résultat, latence et taux de reprise.

Choix à qualifier. Fenêtre longue, modèle plus puissant, réglage de raisonnement et cache sont des leviers différents. Le choix pertinent réduit le coût d'un résultat acceptable tout en respectant le périmètre de données.

Preuve à examiner. Indisponibiliser le modèle principal et vérifier le repli : compatibilité fonctionnelle, respect des routes autorisées, conservation des contraintes et absence de doublon métier.

Le cache fournisseur est un mécanisme d'inférence documenté notamment dans [R7]. Les politiques de routage et d'admission proposées ici sont des choix d'architecture à mesurer.

Observer le système et contenir les incidents

L'exploitabilité repose sur la corrélation des événements, la maîtrise des frontières de confiance et des objectifs de service vérifiés.

Une trace de bout en bout

Un identifiant de tâche relie dialogue, recherche, assemblage du contexte, inférence, appels d'outils, décisions et livrable. Les événements indiquent versions, latences, coûts, tentatives et résultats. Le journal d'audit porte les décisions et effets ; la télémétrie technique aide au diagnostic.

Mesure	Question d'exploitation
Latence p50 / p95	Où le temps est-il consommé : file, source, modèle ou outil ?
Usage du contexte	Quelle part provient des schémas, sources, historiques ou répétitions ?
Coût / succès	Quels usages et versions produisent un résultat acceptable ?
Échecs / incertitudes	Quelles opérations nécessitent reprise, réconciliation ou intervention ?
Qualité / dérive	Une évolution du corpus, du modèle ou d'un outil dégrade-t-elle le service ?

Une isolation appliquée aux données et aux effets

Le parcours des données identifie destinataires, contenus transmis, rétention et effacement pour chaque modèle, OCR, reranker et journal. Les droits s'appliquent aussi aux index, caches et mémoires. Secrets et jetons restent dans les services habilités. Les sorties réseau, parseurs de fichiers et exécutions de code sont confinés selon leur risque ; les politiques de rétention et d'effacement couvrent les dérivés.

Les descriptions d'outils et documents ne sont pas des autorités. Les contrôles de politique et les opérations déterministes doivent rester hors du pouvoir de décision du modèle. Une validation de schéma ou un filtre de texte ne constitue pas une défense complète contre l'injection.

Une résilience qui respecte l'état métier

Files, limitation de débit, annulation propagée et circuits de protection évitent les cascades de pannes. Les objectifs de disponibilité, RPO et RTO sont définis par usage. La restauration teste la cohérence entre fichiers, base, index, clés et opérations en attente.

Preuve à examiner. Tester restauration, révocation avec caches chauds et saturation du service. Mesurer les effets clients, la perte de données et le temps réel de reprise.

Repères : OWASP [R15] ; OpenTelemetry [R17] pour la sémantique de télémétrie. Les conventions GenAI évoluent : leur version et leur statut doivent être suivis, avec minimisation des contenus sensibles dans les traces.

Démontrer la qualité avec des tâches représentatives

Une évaluation utile relie la qualité des composants au résultat métier, en maintenant des jeux indépendants et des conditions reproductibles.

Dimension	Indicateurs possibles	Cas à inclure
Extraction / RAG	Couverture, rappel à k, rang, qualité des tableaux et des locateurs.	Scans, paraphrases, versions et absence de réponse.
Contexte / mémoire	Conservation des contraintes, correction, fraîcheur et rappel utile.	Long dialogue, instruction révisée et fait périmé.
Réponse	Exactitude, fidélité, citations et incertitude bien exprimée.	Sources contradictoires ou partielles.
Outils / agents	Choix de capacité, arguments valides et résultat final observé.	Révocation, attente humaine et interruption.
Production	Réussite, coût, p95, taux de reprise et effets indésirables.	Charge, incidents et changements de versions.

Des expériences interprétables

La campagne fige corpus, versions, configuration, droits et critères d'acceptation. Les données de développement sont séparées du jeu d'évaluation. Les composants sont comparés par ablation : retrait du reranking, du graphe ou d'un mécanisme de contexte pour mesurer leur contribution.

Le rappel à k mesure la part des éléments pertinents attendus retrouvés parmi les k premiers résultats. Le coût par réussite inclut les tentatives et la relecture requise. Les essais sont répétés lorsque la variabilité du modèle le justifie. Le rapport expose effectifs, échecs, intervalles d'incertitude et coûts. Le résultat final dans l'environnement est vérifié indépendamment de la déclaration de l'agent.

Combiner les méthodes de jugement

Les règles déterministes vérifient calculs, schémas et états. Les modèles juges apportent une appréciation souple, mais nécessitent une calibration sur des revues humaines. Les experts métier examinent les cas limites et les erreurs dont l'impact est élevé.

Les tests adversariaux couvrent injections via documents et outils, fuite entre organisations, mémoire empoisonnée et dérive de permissions. Un changement de modèle ou de catalogue déclenche les régressions associées et un déploiement progressif.

Dossier de preuve attendu. Pour chaque usage : objectif, population testée, sources, configuration, seuils métier, résultats reproductibles, coûts, limites et décision de mise en service.

Référence : [R16]. Cette grille est une proposition d'évaluation ; aucun score global, benchmark comparatif ou gain chiffré n'est attribué à un produit dans cette édition.

Les preuves qui rendent une architecture convaincante

Pour apprécier le niveau technique, demander des démonstrations ciblées qui traversent plusieurs couches de la plateforme.

Responsabilité	Preuve attendue	Test révélateur
Modèle sémantique	Objets, relations, versions et unités documentés.	Deux homonymes et une clause révisée.
Catalogue	Contrats typés, effets et versions compatibles.	Permission retirée après découverte.
Contexte	Budget, références réouvrables et état reconstruit.	Correction tardive en dialogue long.
Recherche	Évaluations par question et configuration.	Comparaison sur corpus indépendant.
Preuves	Affirmation reliée à une source et une couverture.	Conclusion d'absence sur source partielle.
MCP	Compatibilité et chaîne d'autorité vérifiées.	Jeton révoqué ou cible modifiée.
Exécution	Reçus, reprise et effet incertain explicites.	Panne après envoi distant.
Exploitation	Traces, budget, disponibilité et restauration.	Saturation puis reprise cohérente.
Qualité	Jeux indépendants, résultats répétés et décision.	Nouvelle version de modèle ou d'outil.

Un socle et des options à justifier

Le socle rend les données accessibles selon les droits, le contexte maîtrisable, les actions contrôlées et les résultats vérifiables. Il doit être démontré pour les usages retenus.

Les options avancées comprennent notamment ontologie formelle, GraphRAG, recherche agentique, traitement multimodal approfondi et coordination de plusieurs agents. Leur pertinence dépend de la complexité métier et de la valeur mesurée.

Ce qui fait la différence. Une plateforme convaincante relie une sémantique explicite, une ingénierie du contexte rigoureuse, des interfaces composables et une exécution dont les preuves résistent aux changements et aux pannes.

Le référentiel décrit un objectif de conception et de qualification. Il ne constitue ni une certification, ni une comparaison commerciale entre fournisseurs.

Sémantique, contexte et recherche

Les références ci-dessous sont des sources primaires. Les choix proposés dans le rapport restent une synthèse d'architecture, à adapter au besoin.

- [R1] [W3C · SKOS Reference](#)
Vocabulaires contrôlés, taxonomies et relations entre concepts.
- [R2] [W3C · OWL 2 Overview](#)
Ontologies formelles et représentation des connaissances.
- [R3] [W3C · PROV-O](#)
Entités, activités et provenance.
- [R4] [Anthropic · Effective context engineering](#)
Sélection du contexte, compactage et récupération à la demande.
- [R5] [Mei et al. · A Survey of Context Engineering](#)
Synthèse de recherche sur l'assemblage et la gestion du contexte.
- [R6] [Anthropic · Context editing](#)
Édition du contexte et gestion des résultats d'outils.
- [R7] [Anthropic · Prompt caching](#)
Réutilisation de préfixes d'inférence et conditions du cache.
- [R8] [Microsoft · Hybrid search](#)
Recherche lexicale, vectorielle et fusion de classements.
- [R9] [Anthropic · Contextual Retrieval](#)
Contexte documentaire et reclassement de candidats.
- [R10] [Microsoft Research · From Local to Global](#)
Approche GraphRAG pour des questions de synthèse globale.

Méthode de lecture

Les sections associent mécanismes, limites et preuves à demander. Les encadrés « Illustration » sont les seuls passages décrivant une implémentation particulière. Les exemples de contrats, de budgets et de scénarios sont fictifs lorsqu'ils sont signalés comme tels.

Le terme « attendu » exprime un objectif de conception pour une plateforme professionnelle. Les technologies conditionnelles ne sont pas des obligations universelles. Une performance d'un article ne peut pas être transférée à un produit sans mesure comparable.

Références consultées le 19 octobre 2026. Liens cliquables dans le PDF. Édition révisée : orientation neutre, ingénierie du contexte, ontologie, catalogue de capacités et exposition MCP approfondis.

Interopérabilité, exploitation et illustrations

Protocoles et évaluation

- [R11] [MCP · Tools, révision 2026-07-28](#)
Contrats d'outils, schémas et résultats.
- [R12] [MCP · Resources, révision 2026-07-28](#)
Accès aux contenus adressables et interfaces de ressources.
- [R13] [MCP · Authorization, révision 2026-07-28](#)
OAuth, découverte, scopes et destinataire des jetons.
- [R14] [MCP · Security Best Practices](#)
Risques de délégation, jetons et frontières de confiance.
- [R15] [OWASP · Excessive Agency](#)
Limitation des fonctions, permissions et décisions autonomes.
- [R16] [Anthropic · Demystifying evals for AI agents](#)
Méthodes de jugement et vérification des résultats finaux.
- [R17] [OpenTelemetry · Gen AI attributes](#)
Repères de télémétrie ; conventions et statuts à versionner.
- [R18] [MCP · Prompts, révision 2025-11-25](#)
Gabarits d'interaction paramétrés ; référence conceptuelle de cette primitive.
- [R19] [Anthropic · Context windows](#)
Limites de contexte, entrée, sortie et raisonnement selon le modèle.

Sources des illustrations WorkProba

Sources figées sur la révision `d19b0503`. Elles illustrent des choix d'implémentation, sans attester l'état d'une installation déployée. L'ouverture des liens peut nécessiter un accès GitHub autorisé.

Repère	Source	Responsabilité illustrée
W1	Modèle produit	Objets, responsabilités et relations.
W2	Contexte · État de travail	Projection et références d'origine.
W3	Chaîne OCR / RAG	Sources, recherche et consultation.
W4	Contrat MCP	Découverte et autorité.
W5	Missions persistantes	Exécution et reprise.
W6	Plugins applicatifs	Versions et activation.

Aucun appel fournisseur payant, benchmark comparatif ni audit d'une installation en production n'a été réalisé pour cette révision. La prise en charge de chaque primitive MCP et de chaque technologie avancée doit être vérifiée dans son propre périmètre.