

Kwaku Agents

Autant d'agents IA que vous avez de métiers, définis avec vous, qui travaillent sur vos documents et vos outils internes. Le modèle tourne sur le matériel de l'hébergeur et les données de chaque entreprise restent cloisonnées.

Agents sur mesure

Multi-entreprises

Modèle exécuté localement

Validation humaine

Kwaku Agents en une page

LE BESOIN

Utiliser l'IA sur ses propres documents

Sans envoyer les dossiers clients chez un prestataire externe, et sans accorder à un modèle un droit d'écriture non encadré.

LA RÉPONSE

Des agents taillés pour vos métiers

Autant d'agents que nécessaire, chacun chacun avec son nom, son objectif et son son prompt, sur une base de connaissances connaissances alimentée par vos outils. outils.

LA DIFFÉRENCE

Le contrôle reste chez vous

Le modèle tourne sur votre matériel. Aucune écriture n'est autorisée par défaut. défaut. Chaque action validée est tracée au tracée au nom de son auteur.

5 agents en standard

+ agents sur mesure

8 types de sources

0 action sans autorisation

Pourquoi un assistant générique ne suffit pas en entreprise



Il ne connaît pas votre entreprise

Il ignore vos procédures, vos contrats et vos échanges internes. Joindre un document à chaque question reste un contournement, que personne ne tient sur la durée.



Il envoie vos données ailleurs

Chaque question transmet son contenu à un prestataire externe. Sur un dossier client, une fiche de paie ou un devis en cours, c'est le plus souvent rédhibitoire.



Il n'a aucune limite d'action

Dès qu'un assistant peut écrire dans Slack, Teams ou Drive, il faut pouvoir dire ce qu'il a le droit de publier et sur quelle cible. La plupart des outils ne répondent pas à cette question.

Des agents définis pour vos métiers

Agent	Son terrain	Une demande réelle
Iris · Marketing	Contenus et argumentaires	Rédige la fiche produit du nouveau contrat
Marc · RH	Procédures internes	Que dit notre accord télétravail ?
Yasmine · LinkedIn	Publications professionnelles	Un post à partir du communiqué de presse
Zoé · Instagram	Formats courts et visuels	Trois légendes pour la photo du chantier
Tom · Développeur	Code et documentation	Explique ce déploiement à un arrivant
Votre agent métier	Défini avec vous	Toute activité qui mérite un interlocuteur

Les cinq premiers sont livrés en standard. Le nombre d'agents n'est pas plafonné : nom, rôle, objectif, prompt, ton, modèle et couleur sont définis avec vous, puis ajoutés à votre instance. Les agents standard se modifient de la même façon, ou se retirent.

De la mise en service au quotidien

01

L'administrateur connecte connecte les outils

Slack, Microsoft 365, Notion, Notion, Google Drive : une autorisation OAuth, puis des ressources cochées une par une.

02

La base de connaissances se remplit

Les sources se synchronisent sur synchronisent sur planning, avec avec élagage de ce qui a disparu disparu chez le fournisseur.

03

L'équipe pose ses questions

Chaque réponse cite ses sources. Un document qu'une qu'une personne n'a pas le droit droit de lire ne remonte pas dans ses résultats.

04

L'agent propose, l'humain décide

Publier un message ou créer une une page : l'agent propose, l'utilisateur valide d'un clic, et et l'opération est tracée à son son nom.

Les quatre capacités de la plateforme



Conversation

Autant d'agents que de métiers, mémoire par conversation, pièces jointes image et PDF, réponses formatées avec leurs citations.



Connaissance

Huit types de sources, recherche hybride, synchronisation planifiée, droits d'accès document par document.



Connecteurs

Slack, Microsoft 365, Notion, Notion, Google Workspace et tout serveur MCP déclaré par par le client.



Surfaces

Une API par clé, un serveur MCP en lecture seule, un robot qui répond dans un fil Slack.

Tout se règle par entreprise/services

Chaque entreprise a ses agents, ses sources, ses connexions et ses droits. Un agent porte son nom, son objectif, son prompt, son modèle et sa couleur, et la liste des agents s'étend à la demande.

La base de connaissances

Dépôt manuel

Base SQL en lecture seule

API HTTP paginée

Pages web et sitemap

Notion

Google Drive

Slack

Microsoft 365

RECHERCHE

Vectorielle et plein texte

Les deux classements sont fusionnés par rang réciproque. Une distance cosinus et un score textuel ne sont pas additionnables.

MISE À JOUR

Incrémentale et planifiable

Ce qui n'a pas changé n'est pas recalculé. Ce qui a été supprimé chez le fournisseur est retiré de l'index au passage suivant.

LECTURE

Des droits par document

Refus par défaut. Une liste d'autorisés vide ne vide ne donne accès à personne, et des droits droits que le connecteur n'a pas su lire non non plus.

Lecture et écriture, par outil connecté

Outil connecté	Lecture	Écriture, sur cible autorisée
Slack	Fils des canaux publics ouverts	Publier un message
Microsoft 365	SharePoint, OneDrive, canaux Teams	Publier dans un canal, déposer un fichier
Notion	Pages et bases partagées	Créer une page ou une entrée
Google Workspace	Dossiers et Drive partagés	Créer un document
Serveur MCP du client	Outils déclarés en lecture seule	Validation humaine toujours imposée

Condition de la colonne Écriture

Une autorisation d'écriture est créée par un administrateur, pour un agent précis et sur une cible nommée. La liste des cibles est saisie à la main. Une liste vide interdit l'action au lieu de l'ouvrir à tout.

Trois façons d'atteindre les agents sans ouvrir le navigateur



API agent

Une clé par entreprise, avec son périmètre : agents autorisés, débit, modèle. Un système tiers ouvre une conversation et reçoit une réponse accompagnée de ses sources.



Serveur MCP exposé

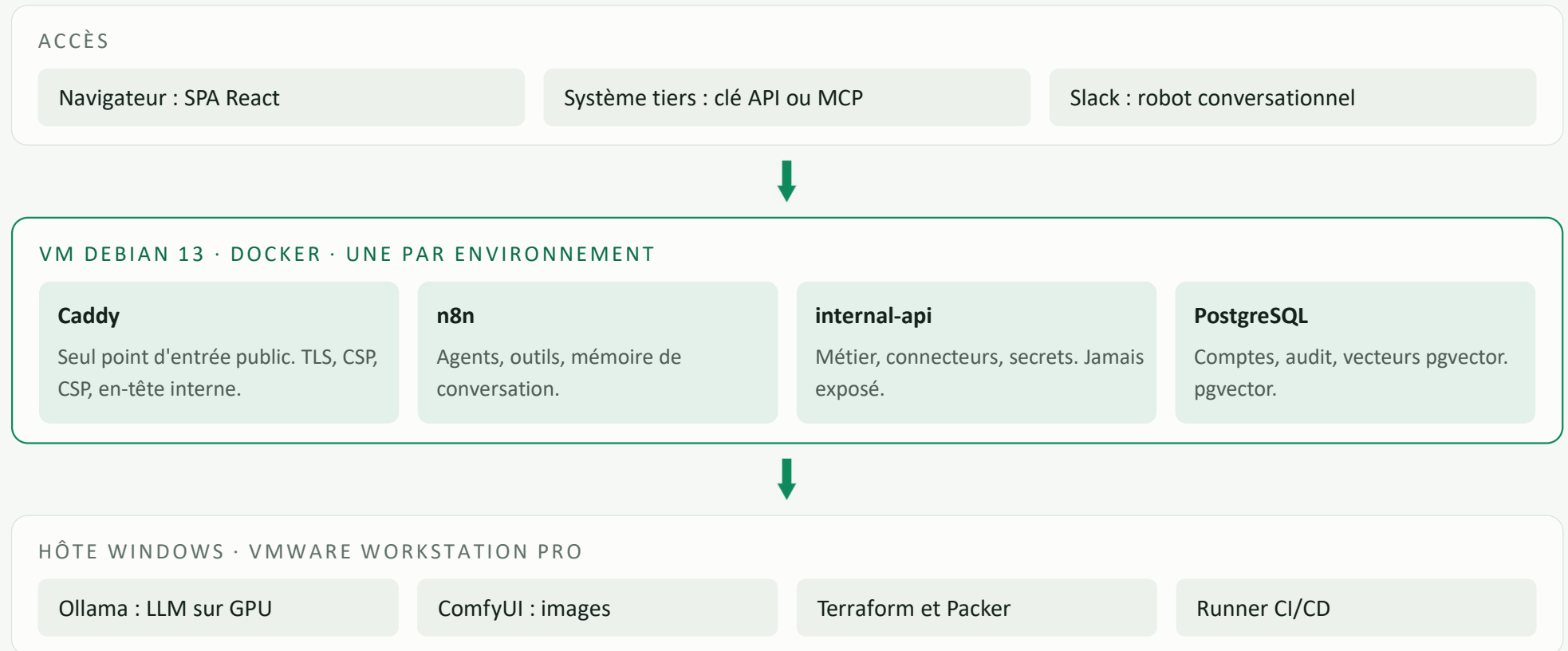
Un assistant tiers interroge votre base de connaissances depuis son propre outil. L'accès est limité à la lecture, pour ne pas chaîner deux agents dont aucun ne connaît les règles de l'autre.



Robot Slack

L'agent répond dans un fil, aux personnes personnes dont l'adresse correspond à un à un compte actif de l'entreprise. Les canaux sont ouverts un par un par un administrateur.

Vue d'ensemble de l'architecture



Deux environnements indépendants : prod et dev, VM, base et domaine distincts

La stack, couche par couche

Couche	Technologie	Rôle
Infrastructure	Packer, Terraform, VMware	Image Debian 13, clones liés, cycle de vie des VM
Provisioning	Bash idempotent	Docker, pare-feu, TLS, sauvegardes, rejoué à chaque déploiement
Orchestration IA	n8n, Ollama	Agents, outils, mémoire, embeddings
Service métier	TypeScript, Fastify	Authentification, connecteurs, base de connaissances
Données	PostgreSQL 16, pgvector	Relationnel, plein texte, vecteurs
Interface	React, TypeScript, Vite	SPA, trois pages publiques prérendues
Entrée réseau	Caddy	Reverse-proxy, Let's Encrypt, CSP stricte
Supervision	Prometheus, Grafana, Kuma	Métriques, huit tableaux de bord, sondes

Une instance, plusieurs entreprises isolées

PÉRIMÈTRE

Relu en base, pas envoyé par le client

L'entreprise et l'utilisateur sont relus en base depuis le jeton de session, session, à chaque requête. Un identifiant transmis par le navigateur navigateur n'est pas retenu.

REFUS

Une ressource d'ailleurs répond 404

Et non 403. Un refus qui distinguerait les deux cas permettrait de confirmer l'existence d'un identifiant.

SECRETS

Chiffrés par entreprise

AES-256-GCM, avec l'identifiant d'entreprise en donnée authentifiée. authentifiée. Un secret recopié sur une autre entreprise devient devient indéchiffrable.

RESSOURCES

Sièges et débit par entreprise

Quota de comptes, messages par minute, actions par minute. La La consommation d'une entreprise ne dégrade pas le service des des autres.

Du dépôt git à la production, sans intervention manuelle

01

Packer

Une image Debian 13 construite une fois, partagée par les deux environnements.

02

Terraform

Clone lié, réseau, démarrage. Un espace de travail par environnement.

03

Provisioning

Bash idempotent : Docker, Docker, pare-feu, TLS, sauvegardes. Rejoué à chaque fois.

04

Intégration

Un push sur la branche de dev déploie dev, sur la branche principale déploie la prod.

05

Vérification

Le déploiement se termine par un contrôle automatisé de toute la stack.

Sécurité intégrée au processus

Aucun commit direct. Toute modification passe par une branche et une demande de fusion. Une revue de sécurité est obligatoire dès que le changement touche les workflows, le service ou le provisioning. Le passage en production reste une décision humaine.

Six niveaux de vérification

Niveau	Ce qu'il prouve	Quand
Tests du service	Règle métier, cloisonnement, refus, cas limite	Chaque proposition
Tests du client	Fonctions pures de l'interface	Chaque proposition
Tests de construction	Prérendu, métadonnées, référencement	Après la construction
Parcours navigateur	Ce qu'un utilisateur voit et fait réellement	Contre l'environnement dev
Vérification de déploiement	Câblage, réseau, pare-feu, certificats, sauvegardes	Chaque déploiement
Tests manuels documentés	Ce qu'aucun script ne peut faire	Un humain

Deux règles de rédaction

La majorité des cas exercés sont des refus, car un test qui n'emprunte que le chemin passant ne prouve rien du garde-fou. Et une vérification ne prouve que ce qu'elle appelle : ce qui s'affiche dans une page se teste dans une page.

Exploitation et supervision



Métriques

Machine, conteneurs, base de données, données, n8n et hôtes de modèles. Huit Huit tableaux de bord générés depuis le le dépôt, ce qui évite les réglages posés à posés à l'écran et perdus au déploiement déploiement suivant.



Disponibilité

Une sonde par service, y compris le modèle hébergé sur la machine hôte. Un observateur externe complète le dispositif, une supervision auto-hébergée s'arrêtant en même temps que l'hôte.



Sauvegardes

Sauvegarde quotidienne de la base, quatorze jours de rétention, test de restauration hebdomadaire et nettoyage nettoyage disque hebdomadaire. Chaque Chaque tâche écrit son statut.

PRINCIPE DE BASE

Rien n'est autorisé par défaut.

Une action d'écriture n'existe pour un agent qu'après création par un administrateur, sur une cible nommée. La validation humaine est le comportement par défaut, et l'autorisation est revérifiée au moment du clic. Un droit retiré entre la proposition et la décision annule l'action.

Défense en profondeur, couche par couche

Couche	Ce qui est en place
Réseau	Restriction réelle dans la chaîne DOCKER-USER, pas seulement ufw. Debug en LAN.
Système	fail2ban sur SSH, mises à jour automatiques, service interne sans privilège root.
Entrée publique	Caddy seul point d'entrée, TLS Let's Encrypt, CSP stricte, HSTS conditionnel.
Identité	PBKDF2 100 000 itérations, jeton signé 24 h, verrouillage après cinq échecs.
Données	Périmètre relu en base à chaque requête, secrets chiffrés par entreprise.
Sorties réseau	Adresse saisie par un client validée sur l'IP résolue, à chaque redirection.
Traçabilité	Audit de chaque décision, attribué à la personne. Export RGPD en un clic.

Deux invariants

Un mot de passe modifié coupe les deux voies d'accès au compte dans la même opération. Une ressource appartenant à une autre entreprise répond 404 et non 403, pour qu'un refus ne confirme pas l'existence d'un identifiant.

LIMITES ASSUMÉES

Limites connues et documentées

INJECTION INDIRECTE

Réduite, pas résolue

Tout contenu importé est encadré et annoncé comme donnée. Une lecture externe fait retomber une écriture en validation humaine.

GPU PARTAGÉ

Modèle trop gros, réponses réponses lentes

Un abandon côté client n'interrompt pas la génération. Un génération. Un modèle qui déborde de la carte graphique graphique tourne en partie sur le sur le processeur. Le cas est mesuré et alerté.

HÔTE UNIQUE

La supervision s'arrête avec lui

Une panne arrête les deux environnements et leurs superviseurs. Seul un observateur observateur externe fait foi sur sur une indisponibilité.

STREAMING

Expérimental, désactivé

Le mode flux ne persiste rien en en base. Il est désactivé par défaut défaut et documenté comme tel. tel.

Chacune de ces limites dispose de sa page de documentation, et les trois premières d'une règle d'alerte.

À retenir

01

Des agents sur mesure

Nombre, rôles, noms, ton et modèle définis avec vous, puis ajoutés à votre instance.

02

Vos données chez vous

Le modèle tourne sur le matériel matériel de l'hébergeur, chaque chaque entreprise est cloisonnée. cloisonnée.

03

Vos outils, un par un

Slack, Microsoft 365, Notion, Google ou votre propre serveur serveur MCP.

04

Aucune action sans accord

Rien par défaut. Chaque action action validée est tracée au nom nom d'une personne.