Chapitre gratuit



Les workflows avant les agents

Chapitre gratuit d'Agents en production : le guide pratique des frameworks agentiques, de MCP, des skills et de l'automatisation par l'IA

Écrit par des agents IA (Claude), sous la direction de Yann

Édition 2026 à 2027

Ce qu'est ce livre

Agents en production est un guide pratique pour faire passer des agents IA d'un pilote prometteur à un système qui tourne tous les jours et se rentabilise. Il est indépendant des éditeurs et construit autour d'un seul modèle mental : un agent, c'est un modèle, des outils, une boucle et du contexte, enveloppés dans un *harness* (le code qui encadre la boucle du modèle) qui applique budgets, garde-fous, état et télémétrie.

Il s'adresse à quatre lecteurs, plus un parcours Claude pour les équipes qui construisent sur Claude : Tech (développeurs et tech leads), Manager, Solo (fondateurs solo) et No-code (pros de l'automatisation sans code). Chaque chapitre indique les sections que votre parcours peut sauter.

Vous saurez :

- Décider quand un *workflow* (enchaînement d'étapes prédéfinies dans le code) fixe vaut mieux qu'un agent, et concevoir des outils, des serveurs MCP, des *skills* (dossiers avec un `SKILL.md`, chargés à la demande) et un contexte qu'un modèle peut exploiter.
- Choisir un framework, ou aucun, et envelopper la boucle dans un harness avec budgets, garde-fous, état et télémétrie.
- Exploiter des agents avec des *evals* (évaluations automatisées : jeux de tâches avec résultat attendu et notation), du traçage, le coût par succès, la sécurité et la gouvernance, y compris le règlement européen sur l'IA (AI Act).
- Mettre des agents au travail pour construire et tester du logiciel, avec un point de contrôle humain à chaque poste qui en a besoin.
- Déployer un agent, ou l'arrêter, sur la base de preuves qu'un comité de pilotage peut vérifier.

-

Écrit par des agents IA, à dessein

Ce livre a été écrit par une équipe d'agents IA. Claude, un modèle conçu par Anthropic, a orchestré plus de trente agents spécialisés qui ont mené les recherches, rédigé, vérifié auprès des sources primaires, dessiné les figures et construit les pages que vous lisez. Un humain, Yann, a dirigé le travail par de courtes décisions à choix multiples et répond de ce qui est publié. Cette édition française a elle aussi été écrite par des agents IA, sous la direction de Yann : des agents traducteurs l'ont produite à partir de l'édition anglaise. Un livre sur les agents en production est aussi un test des agents en production : si vous trouvez une erreur, nous voulons le savoir. La postface, « Comment ce livre a été fait », montre la chaîne de production, les chiffres, et les erreurs rattrapées ou peut-être manquées.

Comment il traite les faits et les histoires

Chaque affirmation précise (un chiffre, une date, une version, une échéance légale) renvoie à sa source, et le site web liste toutes les sources par chapitre ([SITE URL]). Le cahier d'exercices compagnon contient les exercices, les corrigés, le projet fil rouge et l'annexe Claude Kit. Les chiffres illustratifs sont signalés comme tels. Un conflit d'intérêts : les agents qui ont écrit ce livre tournent sur Claude, et le livre traite des produits, de la recherche et des données d'Anthropic au même titre que de ceux de tous les autres. Le chapitre 19, le seul chapitre propre à un éditeur, a été ajouté à la demande du directeur et obéit à la même règle de sourçage. Pour tous les éditeurs : des sources primaires, des chiffres rapportés par les éditeurs signalés comme tels, et aucune recommandation de produit sans règle de décision derrière. Les agents n'ont pas d'expérience vécue, et ce livre ne prétend donc pas en avoir : les histoires sont soit des incidents publics cités, soit des cas composites clairement signalés. Quand le directeur humain ajoute sa propre expérience de *product builder* (développeur polyvalent responsable d'un produit) travaillant sur une plateforme d'environ 100 microservices et un assistant RAG interne, elle apparaît dans un encadré « Note de Yann ». Cette édition en contient un, au chapitre 13, rédigé à partir de ses réponses à un entretien et approuvé par lui ; d'autres pourront venir dans les éditions suivantes.

Compléments en ligne

Gratuits sur le site web ([SITE URL]) :

- Notes de terrain 2026 : le recueil daté complet derrière le chapitre 20, avec une source et un niveau de preuve pour chaque entrée.
- L'annexe B, la checklist de conception des outils et serveurs MCP, et l'annexe C, le modèle de modélisation des menaces, chacune avec un PDF.
- Pour chaque chapitre : les questions des lecteurs, un quiz d'autoévaluation et toutes les sources.
- Le glossaire.

Comment cette édition reste à jour

C'est une édition vivante, avec des mises à jour gratuites jusqu'à fin 2027. Les affirmations qui évoluent vite sont marquées « (fin 2026) » pour que vous sachiez quoi revérifier.

Droits d'auteur et mentions

Agents en production : Le guide pratique des frameworks agentiques, de MCP, des skills et de l'automatisation par l'IA. Édition 2026 à 2027. Édition numérique, 2026.

Copyright © 2026 Yann. Tous droits réservés. Aucune partie de ce livre ne peut être reproduite, stockée ou transmise sous quelque forme que ce soit sans l'autorisation écrite de l'éditeur, à l'exception de brèves citations dans des critiques, ainsi que des checklists et des modèles, que les lecteurs peuvent copier pour un usage interne à leur propre organisation.

Paternité IA. Anthropic n'a ni commandé, ni relu, ni approuvé ce livre. La protection des textes générés par IA varie selon les juridictions ; les droits sont revendiqués sur la sélection, l'agencement et les contributions d'auteur humain.

Avertissement. Ce livre a une vocation d'information et de formation. Il ne constitue pas un conseil juridique, financier, fiscal, de sécurité ou autre conseil professionnel. Les faits, prix, versions et réglementations ont été vérifiés en novembre 2026 et évoluent vite ; vérifiez-les avant d'agir. L'auteur et l'éditeur déclinent toute responsabilité pour les décisions prises sur la base de ce livre.

Marques. Les noms de produits et d'entreprises (Claude, ChatGPT, Gemini, Copilot, n8n, Zapier, Make et tous les autres cités dans ce livre) sont des marques de leurs propriétaires. Ils ne sont utilisés que pour identifier des produits, et aucune affiliation ni approbation n'est sous-entendue.

ISBN (édition numérique) : à attribuer. Publié par [Publisher name].

Composé en Bricolage Grotesque, Source Serif 4 et JetBrains Mono, tous sous licence SIL Open Font License 1.1. Figures dessinées par l'équipe d'agents.

Vous avez trouvé une erreur ? Chaque correction améliore l'édition suivante : [errata contact to be added].

Les citations de sources anglophones sont traduites par nos soins ; les liens renvoient aux originaux.

2

Les workflows avant les agents

L'agent le moins cher et le plus fiable est souvent celui que vous n'avez pas construit

Tech Manager Solo No-code Claude

Cas composite (cas réaliste reconstitué, non observé) : un éditeur de logiciels passe deux semaines à construire quatre agents coopérants pour trier les tickets de support, à quarante secondes par ticket, jusqu'à ce que quelqu'un dessine le vrai processus sur un tableau blanc : cinq boîtes, deux branches. Reconstruit en un après-midi sous la forme d'un *workflow* (enchaînement d'étapes prédéfinies dans le code) avec deux appels au modèle, le système est plus rapide, moins cher et plus facile à tester.

Un **workflow** est un système où votre code décide de la séquence des étapes, même si certaines étapes appellent un modèle ; dans un agent (chapitre 1), c'est le modèle qui décide. La différence tient à qui contrôle le flux.

La règle de base : si vous pouvez dessiner le logigramme, construisez le workflow

Si vous pouvez dessiner le logigramme avant le début de la tâche, construisez le workflow. N'utilisez un agent que pour les parties où l'étape suivante dépend de ce que vous découvrez en chemin.

Comment nous en sommes arrivés là

Les workflows sont exécutables depuis cinquante ans

Composer de petites étapes prévisibles est une vieille idée. Les pipes Unix, attribués à Doug McIlroy et implémentés par Ken Thompson (Bell Labs, Unix history), apparaissent pour la première fois dans le manuel d'Unix Version 3 de février 1973 (Oral history of Unix). BPMN 2.0 a ensuite rendu les diagrammes de processus exécutables (prologue). Chaque génération a appris qu'un flux explicite est plus facile à tester, à auditer et à modifier qu'un comportement caché dans un composant astucieux.

Leçon pour les agents : un appel au modèle n'est qu'une étape de plus dans un pipeline ; gardez le pipeline explicite partout où vous pouvez le dessiner.

Dans un workflow, chaque chemin est du code que vous avez écrit : testable, journalisé, chiffré, explicable. Dans un agent, le modèle choisit le chemin à l'exécution : vous pouvez le borner, pas l'énumérer. Ce compromis n'est rentable que si vous avez besoin de cette flexibilité. L'article d'Anthropic « Building effective agents » trace la même ligne et conseille de commencer par la solution la plus simple, en n'ajoutant de la complexité qu'en cas de besoin (Anthropic, 19 décembre 2024). La plupart des « agents » en production sont des workflows avec une courte section agentique, et c'est la forme saine.

Pour vérifier que vous pouvez le dessiner, prenez cinq entrées réelles et notez les étapes qu'un humain compétent suit pour chacune. Mêmes étapes, ou deux ou trois variantes connues : un workflow. Un chemin différent à travers des systèmes différents à chaque fois : peut-être un agent.

Anti-pattern

Construire l'agent pour découvrir ce qu'est le workflow

Les équipes qui ne connaissent pas encore le processus construisent un agent en espérant que le modèle le trouvera. Il le trouvera, différemment à chaque fois, et vous passerez des semaines à rétro-analyser des traces. Interrogez l'humain qui fait la tâche aujourd'hui, dessinez le logigramme, construisez-le. Gardez l'agent pour les boîtes marquées « ça dépend ».

Le catalogue des patterns

Dans les quatre premiers patterns, votre code maîtrise le flux. Orchestrateur-exécutants se situe à la frontière. Les deux derniers sont des agents. Les cinq premières définitions suivent les recommandations d'Anthropic.

Pattern	Forme	À utiliser quand	À éviter quand	Coût relatif	Défaillance typique
Chaînage de prompts (*prompt chaining*)	Séquence fixe d'appels avec des contrôles en code entre eux	Étapes claires (extraire, rédiger, vérifier)	Les étapes dépendent les unes des autres de façon imprévisible	Faible	Une erreur précoce se propage aux étapes suivantes
Routage (*routing*)	Classer, puis envoyer vers un chemin spécialisé	Catégories d'entrées distinctes	Les catégories se recoupent ou changent sans cesse	Faible	Mauvais routage avec assurance
Parallélisation	Appels indépendants simultanés (découpage) ou même appel N fois (vote)	Sous-tâches indépendantes, jugements à fort enjeu	État partagé	Moyen, N fois les appels	Votes concordants sur la même mauvaise réponse
Évaluateur-optimiseur (*evaluator-optimizer*)	Générateur plus critique, en boucle	Critères de qualité vérifiables	Le premier jet suffit	Moyen à élevé	Peaufinage sans fin
Orchestrateur-exécutants (*orchestrator-workers*)	Un modèle décompose et délègue	Sous-tâches inconnues à l'avance	Vous pouvez lister les sous-tâches	Élevé	Mauvaise décomposition
Boucle d'agent (ReAct)	Raisonner, appeler un outil, lire, recommencer	Le chemin dépend des résultats d'outils	Vous pouvez dessiner le logigramme	Borné seulement par les plafonds	Boucles, mauvais outils, paramètres inventés
Planifier puis exécuter (*plan-and-execute*)	Écrire un plan, l'exécuter, le réviser	Tâches longues où un plan visible aide la relecture	Tâches courtes	Moyen à élevé	Suivre fidèlement un mauvais plan

Ce que le tableau ne dit pas :

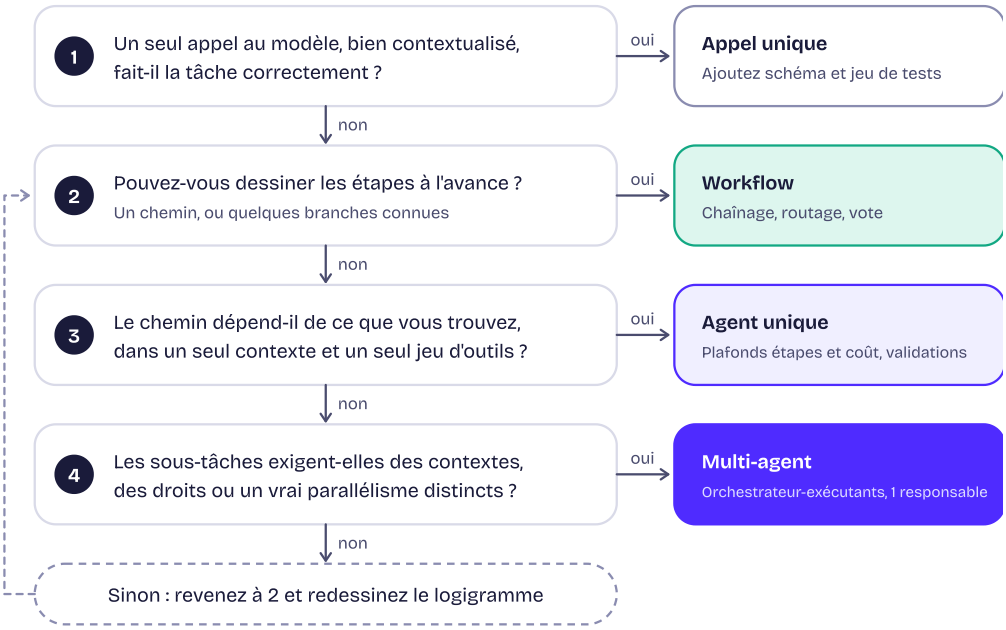
- **Le chaînage** n'est fiable que grâce aux points de contrôle : du code simple, entre les appels, vérifie le schéma, les champs obligatoires et la longueur.
- **Le routage** est votre principal levier de coût (petits modèles pour les cas faciles). Donnez toujours au routeur une catégorie « incertain » qui part vers un humain ou un chemin plus puissant ; un routeur obligé de choisir choisira avec assurance, et à tort.
- **Le vote** détecte l'incertitude à peu de frais : un désaccord entre exécutions est un signal.
- **L'évaluateur-optimiseur** a besoin d'une condition d'arrêt et d'un plafond d'itérations. Le critique trouve toujours quelque chose.
- **La boucle d'agent** (nommée d'après l'article ReAct) exige un *harness* (le code qui encadre la boucle du modèle : budget, garde-fous, état, télémétrie) complet. Arize décrit des agents qui interrogent sans cesse l'état d'une tâche au lieu d'attendre, transformant une tâche en centaines d'appels d'API alors que chaque ligne de log a l'air saine (Arize).
- **Planifier puis exécuter** donne aux humains un plan à relire avant que quoi que ce soit ne s'exécute : un point d'approbation naturel.

L'arbre de décision : appel unique, workflow, agent, multi-agent

Parcourez cet arbre pour chaque nouveau cas d'usage. Arrêtez-vous au premier « oui ».

Parcourez l'arbre, arrêtez-vous au premier oui

Chaque étape vers le bas coûte plus cher et se teste moins bien



> **Arrêtez-vous au premier oui.** Chaque boîte plus bas coûte plus cher et se teste plus difficilement, elle doit donc être justifiée par un cas d'évaluation en échec.

L'étape 4 porte sur les contextes, les permissions et le parallélisme, pas sur des « rôles » : découper ne paie pas parce que la tâche « ressemble » à une équipe. Par défaut, retenez la boîte la plus haute qui passe vos *evals* (évaluations automatisées : jeux de tâches avec résultat attendu et notation), et ne descendez qu'avec un cas d'évaluation en échec que de meilleurs prompts, contextes ou outils ne peuvent pas corriger. Notez ce cas : il justifie le surcoût.

Le budget de déterminisme : où mettre le code et où mettre le modèle

Chaque appel à un LLM ajoute de la variance, et des sorties fausses peuvent avoir l'air justes. Ne dépensez cette variance que là où le modèle apporte quelque chose que le code ne peut pas apporter.

Le modèle propose, le code dispose.

Dans le code (réponse juste calculable)	Dans le modèle (langage et jugement)
Permissions et règles : qui voit quoi, quels remboursements exigent une approbation	De quoi parle ce message, à quel point semble-t-il urgent ?
Arithmétique, dates, échéances de SLA	Lequel de ces documents répond à la question ?
Recherches par clés connues : offre, contrat, responsable	Rédiger une réponse sur le bon ton
Routage sur des champs structurés, validation, nouvelles tentatives, idempotence	Résumer ce fil pour un humain

Le modèle propose, le code dispose. Le modèle renvoie une sortie structurée (une catégorie issue d'une énumération, un brouillon) ; le code la valide, applique les règles métier et décide. Un modèle qui annonce « priorité : P1 » pour un client de l'offre gratuite ne l'emporte pas sur la règle selon laquelle les offres gratuites n'ont pas de file P1. Le test : « si le modèle se trompe là-dessus, un test unitaire l'aurait-il détecté ? » Si oui, cela relève du code. « Les remboursements de plus de 500 euros exigent l'approbation d'un manager » est une règle, pas une ligne de prompt : les prompts ne font pas respecter les règles.

Exemple détaillé : le tri des tickets de support, construit de trois façons

La tâche. Un éditeur SaaS B2B reçoit des tickets de support par e-mail et par chat. Pour chacun, le système attribue une **catégorie** (facturation, technique, accès, fonctionnalité, autre), une **priorité** de P1 (panne pour un client payant) à P4 (question), stockée sous forme d'entier de 1 à 4, une **file** et, éventuellement, un **brouillon de réponse** tiré de la base de connaissances. Les P1 doivent atteindre l'astreinte en quelques minutes, les faux P1 réveillent des gens la nuit, et aucune réponse ne part sans un humain. À disposition : un CRM, une base de connaissances consultable, une page d'état des incidents ouverts.

Version A : le workflow

Cinq boîtes : classer, rechercher le client, appliquer les règles de priorité, router, rédiger. Deux ont besoin d'un modèle.

```
CATEGORIES = {"billing", "technical", "access", "feature", "other"}

def workflow_triage(ticket: Ticket) → Triage:
    # Appel au modèle 1 : petit modèle rapide, schéma JSON strict
    # (category, urgency_signal, confidence).
    c = classify(ticket.text, schema=ClassifySchema)
    if c.category not in CATEGORIES or c.confidence < 0.6:
        return to_human(ticket, reason="low_confidence")

    # Code : recherches par clé.
    customer = crm.get_customer(ticket.sender_email)
    if customer is None:
        return to_human(ticket, reason="unknown_sender")
    incident = status.open_incident_for(customer.region)

    # Code : les règles métier renvoient un entier,
    # de 1 (panne) à 4.
    inputs = PriorityInputs(
        category=c.category, urgency=c.urgency_signal,
        tier=customer.tier, open_incident=incident is not None,
    )
    priority = priority_rules(inputs)
    queue = QUEUE_MAP[(c.category, priority)]

    # Appel au modèle 2, seulement là où un brouillon aide.
    draft = None
    if c.category in {"billing", "access"}:
        docs = kb.search(ticket.text, top_k=4)
        draft = draft_reply(ticket.text, docs, customer.name)

    return Triage(c.category, priority, queue, draft,
                  confidence=c.confidence, inputs=inputs,
                  needs_review=True)
```

Deux appels au modèle au plus par ticket. La priorité vient des règles, donc un P1 signifie toujours la même chose, et chaque branche se teste avec des tests unitaires et des tickets passés étiquetés (100 à 200 répartis entre les caté-

gories, avec une surreprésentation des cas rares et coûteux comme les vrais P1). Vous pouvez le chiffrer avant d'écrire le code : nombre de tickets × *tokens* (unités de texte facturées par le modèle) par ticket × prix.

D'où vient le score de confiance. Une confiance que le modèle écrit dans son propre JSON est un indice, pas une probabilité ; le 0,6 est illustratif. Regroupez les tickets étiquetés par confiance déclarée, mesurez l'exactitude de chaque groupe et placez le seuil là où l'exactitude atteint l'exigence métier, probablement par catégorie. Si les scores déclarés ne suivent pas l'exactitude, utilisez l'accord entre deux ou trois exécutions, ou les log-probabilités des tokens quand votre fournisseur les expose.

Ce qu'il ne sait pas faire : enquêter. « L'export est cassé depuis votre mise en production de mardi, voir le log joint » reçoit une catégorie et une priorité, mais personne ne vérifie si la mise en production de mardi a touché aux exports.

Version B : l'agent

Un seul modèle dans une boucle bornée, avec les outils `search_kb`, `get_customer`, `get_open_incidents`, `search_recent_tickets`, `read_attachment` et `submit_triage`. Le prompt contient les catégories, la politique de priorité et la table des files ; le harness plafonne l'exécution à 12 étapes et à un coût fixe ; tous les outils sauf `submit_triage` sont en lecture seule.

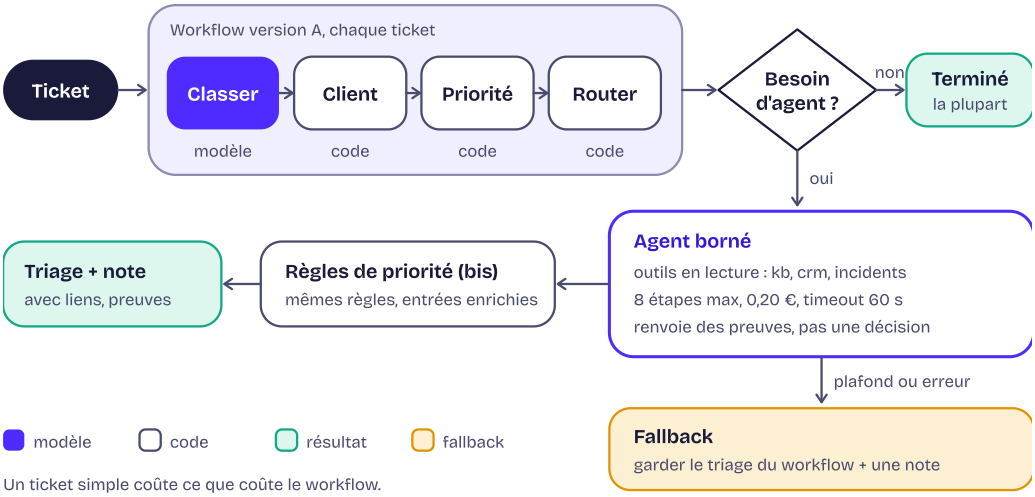
Sur le ticket de l'export de mardi, l'agent lit le log, trouve trois signalements récents similaires, remarque un incident ouvert et propose un P2 avec des liens : ce que ferait un ingénieur support confirmé. Le coût : 3 à 12 appels au modèle par ticket, chacun renvoyant un contexte qui grossit, avec une latence variable. La priorité est désormais la lecture par le modèle d'une politique rédigée en prose, donc des tickets similaires peuvent recevoir des priorités différentes, et les tests doivent couvrir les trajectoires. Sur un simple ticket « comment changer mon adresse de facturation ? », l'agent cherche et vérifie quand même. La plupart des tickets sont simples.

Version C : l'hybride

Faites passer chaque ticket par le workflow. N'envoyez à un agent borné que les tickets qui nécessitent une enquête, et laissez l'agent proposer, pas décider.

L'hybride : une ossature déterministe avec une section agentique bornée

L'agent propose des éléments de preuve. Le code applique les règles et a le dernier mot.



Un ticket simple coûte ce que coûte le workflow.

> **La colonne vertébrale d'abord, l'agent là où ça dépend.** Chaque ticket passe par le workflow ; seuls les tickets qui nécessitent une enquête atteignent un agent borné en lecture seule, et c'est toujours le code qui fixe la priorité.

```
from dataclasses import replace

def triage(ticket: Ticket) -> Triage:
    # Version A, inchangée.
    base = workflow_triage(ticket)
    needs_investigation = base.category == "technical" and (
        base.priority <= 2 or ticket.has_attachment
        or base.confidence < 0.8
    )
    # La plupart des tickets s'arrêtent ici.
    if not needs_investigation:
        return base

    result = run_agent(
        task=Task(id=ticket.id,
```

```
                input=investigate_prompt(ticket, base)),
        model=SMALL_MODEL,
        # kb, crm, incidents, tickets
        tools=READ_ONLY_TOOLS, flags=FLAGS,
        budget=Budget(8, 0.20, 60),
    )
    # Un passage de relais garde le tri du workflow,
    # avec la raison.
    if result.status ≠ "done":
        return replace(base, note="enquête incomplète : "
                       + result.reason)
    # Un Finding validé : l'argument de submit_finding.
    finding = result.output

    # Ne garder que les preuves qui pointent vers de vrais
    # enregistrements (ticket, incident, ligne de log).
    evidence = verify_evidence(finding.evidence)

    # L'agent propose ; le code applique les mêmes règles
    # à des entrées plus riches.
    inputs = replace(base.inputs, evidence=evidence)
    p = priority_rules(inputs)
    return replace(base, priority=p, inputs=inputs,
                   queue=QUEUE_MAP[(base.category, p)],
                   note=finding.summary, links=finding.links)
```

`verify_evidence` protège contre la défaillance la plus pernicieuse : une preuve fausse mais plausible. L'agent doit étayer chaque affirmation par un identifiant d'enregistrement ; le code récupère chaque enregistrement et écarte les affirmations dont l'identifiant n'existe pas ou ne correspond pas (un « incident ouvert » qui est clos). Les intuitions non vérifiées peuvent figurer dans le résumé, mais ne font jamais passer un ticket en P1.

Les tickets simples coûtent ce que coûte le workflow, et un P1 signifie toujours ce que disent les règles. Si l'agent échoue ou épuise son budget, le ticket garde le tri du workflow avec une note : une dégradation maîtrisée, pas une panne.

Régler le déclencheur. `needs_investigation` décide à la fois du coût et de la couverture. Journalisez-le sur chaque ticket et suivez deux taux hebdomadaires : les **enquêtes manquées** (pas d'enquête, puis escalade par un hu-

main : déclencheur trop strict) et les **enquêtes inutiles** (le résultat n'a rien changé : vous payez sans valeur). Modifiez une condition à la fois.

La comparaison

Ce sont des estimations de conception sans cache, pas des benchmarks. Le nombre d'appels sous-estime l'écart : une classification lit un ticket, alors que chaque étape d'enquête renvoie un contexte qui grossit. Avec des volumes de tokens illustratifs (environ 1 200 tokens d'entrée par classification et environ 2 800 par ticket pour la version A une fois les brouillons comptés ; une enquête en six étapes qui passe de 5 000 à 11 000 tokens, soit environ 48 000 au total), la version C coûte environ 1,85 fois la version A quand 5 % des tickets font l'objet d'une enquête, et environ 4,4 fois à 20 %. C'est le déclencheur, pas le modèle, qui fixe la facture.

L'exemple illustratif de remboursement d'ADK 2.0 publié par Google, sur des API simulées, va dans le même sens : des nœuds déterministes font passer les tokens par exécution de 5 152 à 2 265 et le temps de 7,2 à 5,7 secondes par rapport à un simple agent LLM (Google Developers Blog, juillet 2026, affirmation de l'éditeur).

Dimension	A. Workflow	B. Agent	C. Hybride
Appels au modèle par ticket	1 à 2, fixe	3 à 12, choisis par le modèle	1 à 2 pour la plupart ; jusqu'à 2 + 8 en cas d'enquête
Coût modèle relatif	1 fois	1,5 à 12 fois selon le nombre d'appels, davantage avec la croissance du contexte	Environ 1,85 fois à 5 % d'enquêtes, 4,4 fois à 20 % (illustratif)
Latence	Faible, prévisible	Variable, plafonnée	Faible pour la plupart des tickets
Cohérence des priorités	Déterministe	Varie entre tickets similaires	Déterministe, entrées plus riches

Dimension	A. Workflow	B. Agent	C. Hybride
Tickets inédits, multi-systèmes	Un humain enquête	Oui	Oui, pour les tickets routés
Comportement en cas d'échec	File humaine	Atteint un plafond ; pas de tri sans *fallback* (solution de repli : autre modèle ou chemin si le premier échoue)	Repli sur le tri du workflow
Comment le tester	Tests unitaires plus jeu étiqueté	Evals de trajectoire et de résultat, relecture humaine	Tests du workflow plus un petit jeu de trajectoires

Le verdict

Par défaut, choisissez la version A : un système qui fonctionne, un jeu de données étiqueté et un coût de référence en quelques jours. Livrez la version C quand vous disposez d'assez de tickets mal traités par A faute d'enquête pour constituer un jeu d'évaluation. Ne livrez la version B que si la plupart des tickets exigent une enquête, que le volume est faible et que vous acceptez que les priorités deviennent des jugements.

L'hybride est la forme générale des bons systèmes d'agents : une colonne vertébrale déterministe pour le travail prévisible, des sections agentiques bornées là où la découverte paie, et le code qui a le dernier mot sur les décisions qui comptent. Le chapitre 17 emmène ce système en production.

Règle de décision

Règle de placement de l'hybride

Placez l'agent là où le logigramme dit « ça dépend », donnez-lui des outils en lecture seule et un petit plafond, faites-lui renvoyer des preuves plutôt que des décisions, et laissez le code appliquer les règles. Si sa sortie doit déclencher une action directe (une écriture dans une fiche client), placez une approbation humaine ou une validation déterministe entre l'agent et l'action.

Anti-patterns

- **L'agent pour une tâche cron.** Un rapport nocturne construit comme un agent redécouvre le plan chaque nuit, saute un indicateur jugé « pas pertinent aujourd'hui », et les rapports cessent d'être comparables. Récupérez et comparez en code, puis un seul appel au modèle rédige le résumé.
- **Le théâtre multi-agent.** Un « agent PDG » délègue à des agents « manager », « analyste » et « QA ». Chaque passage de relais perd du contexte, les coûts se multiplient à chaque saut, et personne ne sait dire quel agent a causé une sortie fausse. C'est généralement une chaîne de prompts déguisée : appliquez l'étape 4 de l'arbre.
- **L'agent comme gestionnaire d'erreurs.** Envelopper une intégration fragile dans un agent « pour qu'il comprenne ce qui n'a pas marché ». Corrigez l'intégration : les agents lisent parfois une erreur HTTP 500 sans texte d'échec explicite comme « traité » (Arize).
- **Le même gros modèle partout.** La classification et l'extraction tournent généralement très bien sur un petit modèle. Routez par tâche et évaluez chaque route.

Notes par parcours

- **Tech.** Construisez d'abord la version A, même si vous êtes sûr d'avoir besoin d'un agent. Gardez typée l'interface entre workflow et agent, avec les trois statuts du chapitre 8 : terminé, passage de relais, en attente d'un humain.
- **Manager.** Utilisez l'arbre de décision comme point de contrôle en revue : chaque proposition d'agent nomme l'étape qu'elle ne franchit pas, avec un exemple. Demandez le coût par tâche réussie et ce qui se passe au plafond.
- **Solo.** Les workflows échouent bruyamment et de façon prévisible, vous passerez donc moins de soirées à déboguer. Notez dans un tableur le coût de chaque étape plus votre temps de relecture : cette somme, calculée pour le cas avec nouvelle tentative, est votre prix plancher (chapitre 13).
- **No-code.** Les flux n8n, Zapier ou Make sont déjà des workflows : le routage est un aiguillage après une étape de classification, le chaînage est une suite d'étapes LLM avec des filtres, un nœud IF sur le champ de confiance

envoie les scores faibles vers une file humaine, et l'enquête est un nœud AI agent placé derrière une condition, avec des outils en lecture seule, une limite d'itérations basse et une étape d'approbation avant tout message au client.

Checklist

- [ ] Nous avons dessiné le logigramme à partir d'au moins cinq entrées réelles avant de choisir une architecture.
- [ ] Nous avons noté à quelle étape de l'arbre de décision notre conception échoue, avec un exemple.
- [ ] Chaque décision dont la réponse est calculable est dans le code, pas dans un prompt.
- [ ] Chaque sortie du modèle est validée, et chaque classifieur a un chemin « incertain ».
- [ ] Chaque section agentique a des plafonds, un *timeout* (délai d'expiration) et des outils en lecture seule sauf écritures approuvées, et se replie sur un résultat de workflow ou sur un humain.
- [ ] Nous avons une référence de coût et de qualité issue d'un workflow, à laquelle comparer toute conception agentique.

Le chapitre 2 en une page

Les workflows avant les agents

Si vous pouvez dessiner le logigramme, construisez le workflow. N'utilisez un agent que pour les boîtes marquées « ça dépend ».

À retenir

- Parcourez l'arbre de décision depuis le haut et arrêtez-vous au premier oui
- Le modèle propose, le code dispose
- Livrez d'abord le workflow ; passez à l'hybride quand de vrais tickets en montrent le besoin

À faire dès lundi

- Passez une heure avec la personne qui fait la tâche aujourd'hui et notez chaque « ça dépend »
- Cherchez dans vos prompts les seuils, les montants et les « doit » : chacun est une règle à déplacer dans le code
- Chiffrez la version workflow : nombre de tickets × tokens par ticket × prix

1 à 2
appels au modèle par ticket, tri par workflow (estimation de conception)

3 à 12
appels au modèle par ticket, tri par agent (estimation de conception)

Pour aller plus loin : le livre complet

Ceci était un extrait gratuit d'**Agents en production**, un guide pratique pour faire passer des agents du pilote à la production et les y maintenir. Le livre complet comprend un prologue, 21 chapitres en neuf parties, une postface sur la façon dont le livre a été fait et la checklist de mise en production, plus un cahier d'exercices compagnon et des compléments gratuits en ligne.

Le livre complet couvre :

- **Modèles mentaux :** ce qu'est vraiment un agent, et quand un simple *workflow* (enchaînement d'étapes prédéfinies dans le code) vaut mieux.
- **Briques de base :** des outils que les agents peuvent vraiment utiliser, MCP après la révision 2026-07-28, les *skills* (dossiers avec un `SKILL.md`, chargés à la demande), les sous-agents et les plugins, l'ingénierie du contexte et la mémoire.
- **Frameworks et architecture :** une carte indépendante des éditeurs du paysage des frameworks en 2026, un guide de décision et des architectures de référence pour les agents pilotés par les événements et durables.
- **Production :** des *harnesses* (le code qui encadre la boucle du modèle) de fiabilité, des *evals* (évaluations automatisées) en CI, l'observabilité et l'économie unitaire, la sécurité des agents et de MCP, la gouvernance et le règlement européen sur l'IA.
- **Business :** des recettes d'automatisation no-code, des guides pratiques pour indépendants qui font gagner ou économiser de l'argent, et un plan construire, acheter et recruter pour les managers.
- **L'usine logicielle agentique :** des agents qui mènent le logiciel de l'idée à la *pull request* (demande de fusion) fusionnée, le testent de bout en bout et continuent de l'améliorer, avec un projet de synthèse qui fait passer un système par chaque poste.
- **La suite :** un atlas des cas d'usage, les preuves sur la productivité, l'emploi, le coût et l'énergie, et des prévisions sourcées.
- **Annexe :** la checklist de mise en production (59 contrôles, dont 19 obligatoires). En ligne et gratuits : une checklist de conception des outils et serveurs MCP, un modèle de modélisation des menaces, les notes de terrain 2026, un glossaire et les consignes de chaque agent qui a écrit le livre. Le cahier d'exercices compagnon contient les exercices et leurs corrigés.

Quatre parcours de lecture (Tech, Manager, Solo, No-code) vous disent exactement quels chapitres lire et lesquels sauter. Les exercices, le projet fil rouge et les autoévaluations sont fournis avec leurs corrigés complets.

Écrit par des agents IA (Claude, un modèle conçu par Anthropic) sous la direction de Yann.

Disponible sur {{STORE_URL}}