

Version en ligne : aipm-standard.com/aipm/cas/rakuten-claude-code-vllm · édition du 30/09/2026. Le statut de preuve peut évoluer : la page en ligne fait foi.

Cas approfondi

Sept heures d'autonomie, une référence pour vérifier : la Directive bien écrite

Un ingénieur de Rakuten confie à Claude Code une tâche dans vLLM, bibliothèque open source de 12,5 millions de lignes. L'agent travaille sept heures d'affilée ; le résultat atteint 99,9 % de précision numérique face à une méthode de référence. Ce qui rend la démonstration instructive n'est pas la durée : c'est qu'une référence existait pour la juger.

Toute la banque de cas (aipm-standard.com/aipm/cas) · La règle de preuve (aipm-standard.com/aipm/methode-de-preuve)

Statut de preuve : Annoncé/projeté. Ce statut n'est pas une opinion sur l'entreprise ou sur la technologie : il décrit ce que les sources consultées permettent d'affirmer, et rien de plus. La règle qui l'attribue est publiée sur la page Méthode de preuve (aipm-standard.com/aipm/methode-de-preuve). Il porte sur le constat central de la fiche ; les chiffres cités à l'intérieur conservent le statut de leur propre source, et ceux qui proviennent du seul éditeur restent « annoncés », y compris dans une fiche confirmée.

Le contexte

Qui, quand, quel système

Rakuten, groupe japonais de plus de 70 activités, publie le 26 juin 2025 sur son blog *Rakuten Today* un retour d'expérience sur Claude Code, l'agent de programmation d'Anthropic. Le même texte figure, non daté, parmi les études de cas clients d'Anthropic. Il s'inscrit dans la stratégie que le groupe appelle « AI-nization ».

Le texte réunit deux choses de nature différente : une démonstration technique, conduite par un ingénieur en apprentissage automatique, Kenta Naruse ; et un résultat d'exploitation, la baisse du délai de mise sur le marché des nouvelles fonctionnalités. Les confondre, c'est se tromper sur ce que la source prouve.

La décision, et qui l'a prise

La Directive de fait

La décision initiale est un test : Kenta Naruse confie à Claude Code l'implémentation d'une méthode d'extraction de vecteurs d'activation dans vLLM, une bibliothèque open source de 12,5 millions de lignes en plusieurs langages. La tâche a une propriété décisive : il existe une méthode de référence contre laquelle comparer le résultat.

Selon le texte, ce test, combiné à l'alignement de valeurs avec Anthropic, a « convaincu la direction » que Claude Code pouvait transformer son développement. La décision suivante — redessiner les flux de travail autour de l'outil plutôt que de l'ajouter aux processus existants — est présentée comme acquise ; c'est « à la suite de pilotes réussis » que la direction s'engage à étendre l'accès.

Ce que le système a réellement fait

Claude Code a réalisé la tâche en sept heures de travail autonome, en une seule exécution. « Je n'ai écrit aucune ligne de code pendant ces sept heures », rapporte Naruse, « j'ai juste donné quelques indications de temps en temps. » L'implémentation atteint 99,9 % de précision numérique par rapport à la méthode de référence.

En exploitation, Rakuten annonce que le délai moyen de mise sur le marché des nouvelles fonctionnalités est passé de 24 jours ouvrés à 5 jours, soit une réduction de 79 %. Les usages décrits vont des tests unitaires à la documentation, avec des sessions parallèles : « cinq tâches en parallèle en déléguant quatre », selon Yusuke Kaji, directeur de l'IA pour les métiers.

Un projet plus ambitieux est en cours : un « agent ambiant » répartissant la mise à jour du monorepo du groupe sur 24 sessions parallèles, pour un travail qui prendrait manuellement plus d'un mois. C'est une intention, pas un résultat.

L'état de la preuve

Ce que les sources établissent, et ce qu'elles n'établissent pas

Ce que la source établit

- Une tâche réelle, sur une base de 12,5 millions de lignes, a été exécutée en sept heures d'autonomie en une seule exécution
- Le résultat a été jugé contre une méthode de référence, à 99,9 % de précision numérique
- L'ingénieur déclare n'avoir écrit aucune ligne de code pendant la session
- Rakuten annonce un délai moyen de mise sur le marché passé de 24 jours ouvrés à 5 jours

Ce que la source n'établit pas

- Que la démonstration soit représentative : c'est une tâche choisie, sur une base open source, pas un flux de production
- La période, le nombre d'équipes et le nombre de fonctionnalités sur lesquels les 24 et 5 jours sont mesurés
- La part du gain attribuable à l'agent plutôt qu'à la refonte des flux de travail qui l'a accompagné
- Le nombre et la nature des « indications occasionnelles » données pendant les sept heures

- Toute vérification indépendante : le texte est un témoignage client publié par le client et par le fournisseur

Sources consultées : Rakuten Today — Rakuten accelerates development with Claude Code (source primaire 2025-06-26) · Anthropic — étude de cas client Rakuten, Claude Code (page non datée) (source primaire) · Anthropic — entretien Rakuten, Claude Managed Agents (page non datée) (source primaire)

Ce que ce cas dit du référentiel

Ce cas est l'exemple-type du premier champ de la **Fiche Directive** : « résultat vérifiable et méthode de vérification ». Le résultat — implémenter une méthode dans vLLM — est vérifiable parce qu'une méthode de référence existe, et la méthode de vérification est explicite : comparer numériquement les sorties. C'est ce qui permet de laisser l'agent seul sept heures. Sans référence, sept heures d'autonomie ne sont que sept heures sans contrôle.

Il illustre la **preuve exécutable** du Standard : 99,9 % n'est pas le jugement de l'agent sur son propre travail, c'est une mesure externe, rejouable. Il éclaire aussi le **principe 4**, « L'autonomie se mérite, jamais par défaut » : l'élargissement de l'usage est venu après un test sur une tâche bornée, pas avant.

Il sert enfin d'exercice sur le **principe 7**, « L'itération par la preuve mesurée, jamais par la promesse ». La démonstration est solide dans son périmètre ; le gain d'exploitation, lui, est annoncé sans période ni périmètre. Calibrer une gouvernance sur les 79 % comme s'ils étaient confirmés serait exactement l'anti-pattern du risque **7.9.8**.

La consigne

Ce qu'un lecteur fait différemment après avoir lu cette fiche

Ne lancez pas une tâche longue sans oracle. Avant de laisser un agent travailler des heures, écrivez dans la Fiche Directive ce contre quoi le résultat sera comparé : une implémentation de référence, un jeu de tests, des sorties attendues. S'il n'y en a pas, construisez-le d'abord ou raccourcissez la tâche.

Comptez les interventions. « Quelques indications » n'est pas une mesure. Journalisez chaque intervention humaine pendant une session autonome ; c'est la donnée qui justifiera, ou non, d'élargir le périmètre d'autonomie.

Séparez démonstration et exploitation dans vos propres bilans. Un test réussi sur une tâche choisie ouvre un pilote ; il ne mesure pas un gain de production.

Ce qui invaliderait cette lecture. La publication par Rakuten de la période, des équipes et du volume de fonctionnalités derrière le passage de 24 à 5 jours : la fiche passerait alors en « Confirmé opérationnel » pour ce gain. À l'inverse, une information montrant que la comparaison à la référence a été faite par l'agent lui-même, et non par l'ingénieur ou un outil indépendant, retirerait au cas sa valeur d'exemple de méthode de vérification.

Cette fiche suit le gabarit des cas approfondis d'AIPM : huit questions, toujours les mêmes. Voir les autres fiches (aipm-standard.com/aipm/cas-approfondis).

© 2026 DYONYSOS — Tous droits réservés · Bexo® est une marque déposée de DYONYSOS · aipm-standard.com. Reproduction interdite sans autorisation, hors courtes citations avec mention de la source (aipm-standard.com).