Patchie 6 Nano

Le modèle le plus rapide de la gamme Patchie 6.

Patchie Lab a annoncé Patchie 6 Nano le même jour que Patchie 6. Il remplace Patchie 4.8 Nano comme modèle léger de la gamme : celui qu'on appelle quand la réponse doit arriver tout de suite, à grand volume et à faible coût. Assistants en temps réel, agents de recherche, support client, complétion de code : c'est le palier de première ligne, Patchie 6 restant le modèle des tâches les plus lourdes.

Par rapport à 4.8 Nano, le saut est net sur tous les axes mesurés : la qualité des réponses, la taille du contexte et la longueur des sorties.

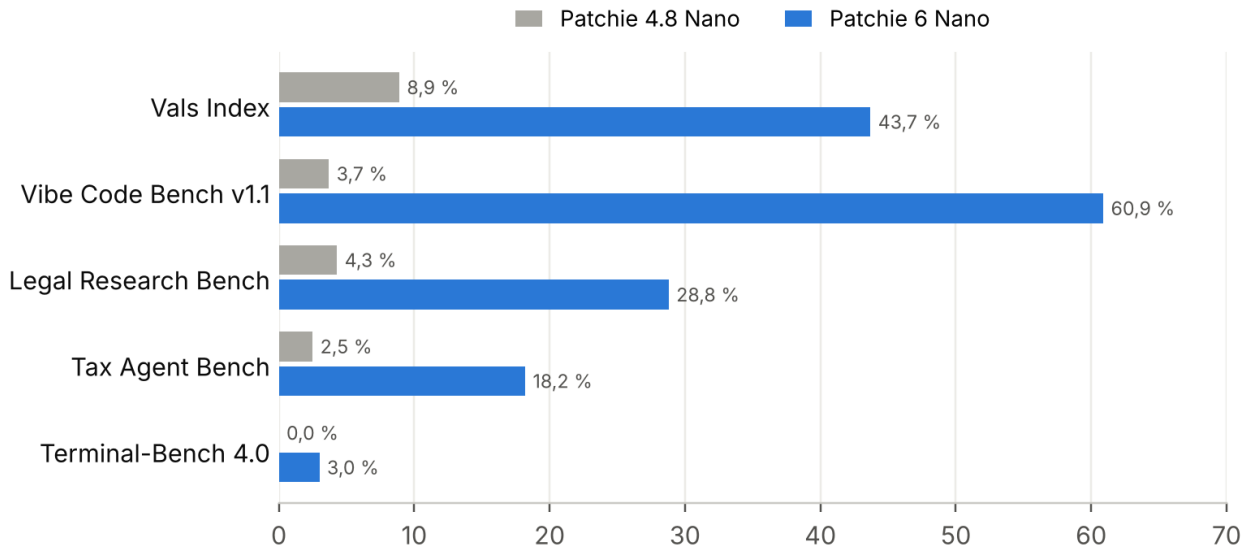
En bref

	Patchie 4.8 Nano	Patchie 6 Nano	Évolution
Vals Index (score global)	8,9 %	43,7 %	× 4,9
Vibe Code Bench v1.1	3,7 %	60,9 %	+ 57,2 pts
Fenêtre de contexte	260 K tokens	1 M tokens	× 3,8
Sortie maximale	65 536 tokens	131 072 tokens	× 2
Débit annoncé	≈ 1 100 tokens/s	≈ 1 070 tokens/s	stable

Face à ses concurrents directs, GPT-6 Luna et Claude Haiku 5.5, Patchie 6 Nano obtient le meilleur score sur 8 des 12 benchmarks publiés, et bat chacun des deux sur 10 benchmarks sur 12.

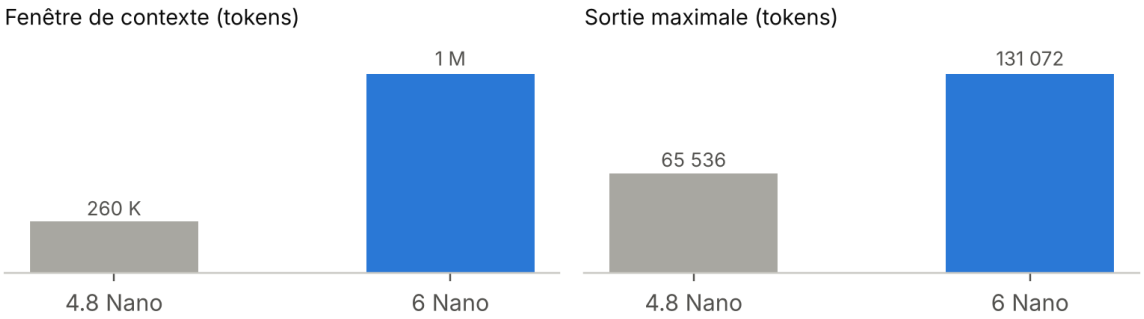
1. Ce qui change par rapport à Patchie 4.8 Nano

Les deux modèles ont été évalués par le même organisme indépendant (Vals AI), sur les mêmes versions de benchmarks. Seuls les benchmarks disponibles à l'identique pour les deux générations figurent ici.



Scores Vals AI, en %. Plus haut = meilleur.

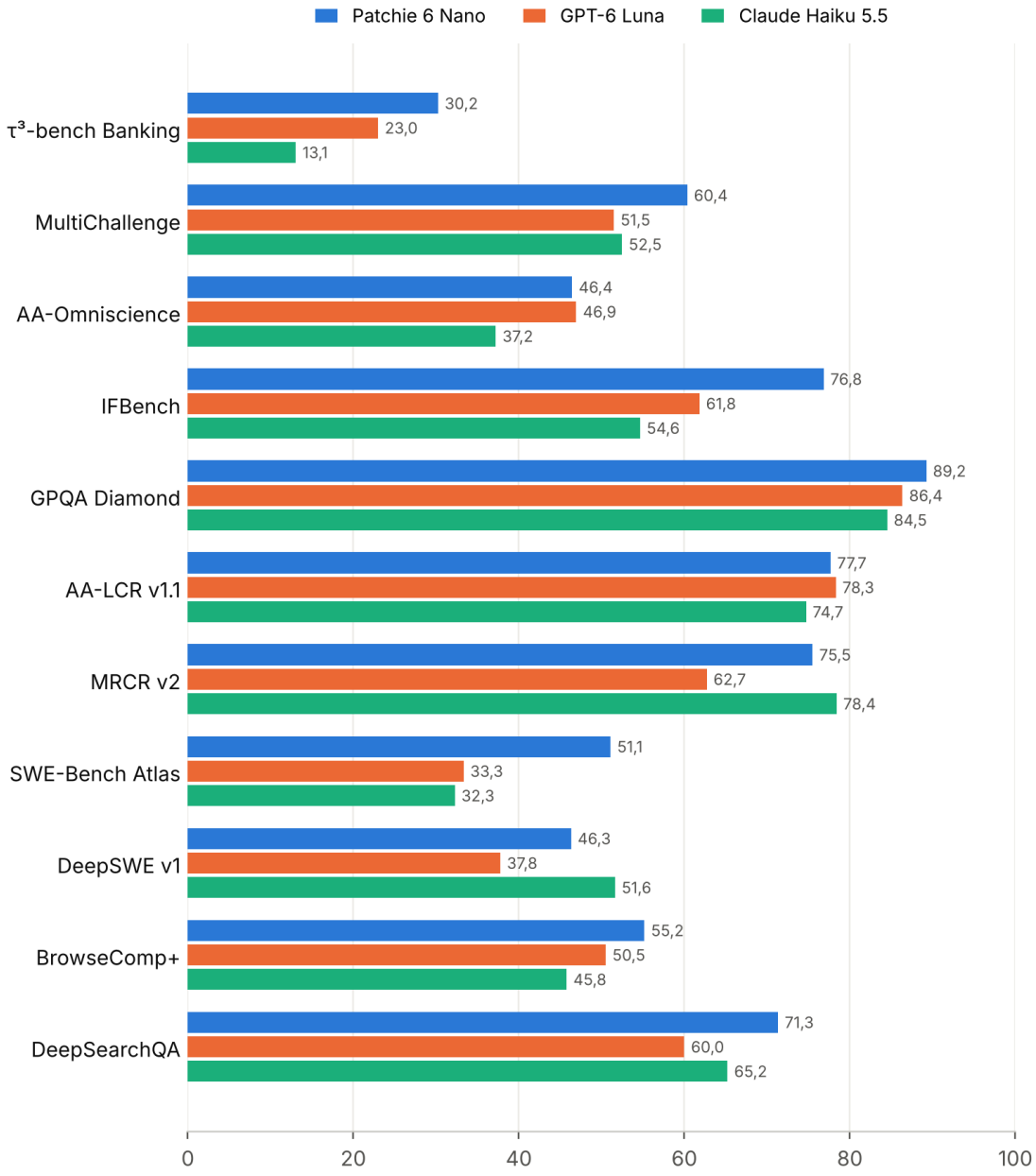
Le Vals Index, qui agrège l'ensemble des tâches professionnelles de Vals AI, est multiplié par près de cinq. Le gain le plus spectaculaire concerne la génération d'applications complètes (Vibe Code Bench) : 4.8 Nano n'y parvenait presque jamais, 6 Nano réussit six tâches sur dix.



Le contexte passe de 260 000 à 1 million de tokens : un dépôt de code moyen ou plusieurs centaines de pages tiennent désormais en une seule requête.

2. Face à la concurrence

Patchie 6 Nano se mesure aux deux modèles qui visent le même usage : rapides, peu chers, déployés à grande échelle. Les scores ci-dessous sont obtenus avec un effort de raisonnement moyen.



Scores en %. Plus haut = meilleur.

Tableau détaillé

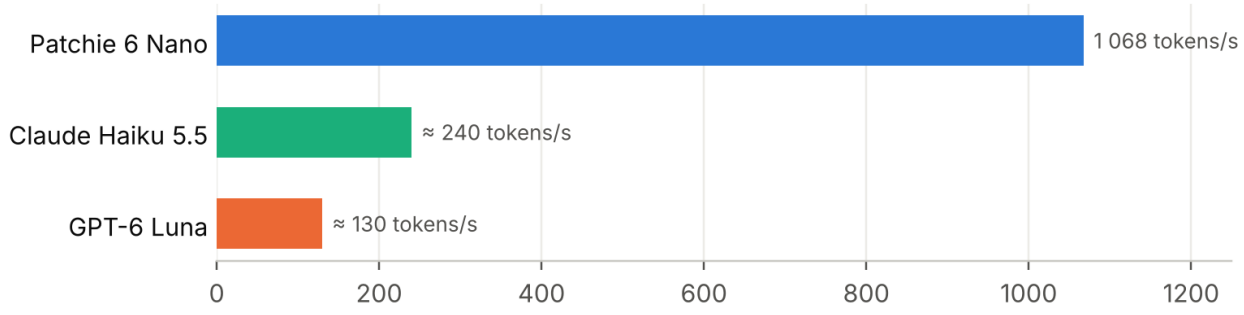
Benchmark	Ce qu'il mesure	Patchie 6 Nano	GPT-6 Luna	Claude Haiku 5.5
GDPval (Elo)	Tâches de travail réelles	**1 237,2**	1 204,0	1 225,0
τ³-bench Banking	Agent de service client	**30,24**	23,00	13,06
MultiChallenge	Conversation en plusieurs tours	**60,35**	51,49	52,48
AA-Omniscience	Exactitude factuelle	46,44	**46,94**	37,22
IFBench	Respect des consignes	**76,84**	61,82	54,65
GPQA Diamond	Sciences, niveau doctorat	**89,23**	86,36	84,51
AA-LCR v1.1	Raisonnement sur long contexte	77,67	**78,30**	74,70
MRCR v2	Rappel sur long contexte	75,50	62,72	**78,40**
SWE-Bench Atlas	Code agentique	**51,08**	33,33	32,26
DeepSWE v1	Code agentique	46,31	37,76	**51,62**
BrowseComp+	Recherche web agentique	**55,18**	50,50	45,78
DeepSearchQA	Recherche agentique	**71,33**	60,00	65,22

En gras : meilleur score de la ligne.

Ce qu'il faut retenir

- **Agents et outils.** Patchie 6 Nano domine nettement sur les tâches d'agent : service client (τ³-bench), recherche (BrowseComp+, DeepSearchQA) et code agentique (SWE-Bench Atlas, + 17,8 points sur GPT-6 Luna).
- **Respect des consignes.** 76,8 sur IFBench, soit 15 points de plus que GPT-6 Luna et 22 de plus que Claude Haiku 5.5. C'est l'écart le plus utile au quotidien.
- **Là où il est battu.** Claude Haiku 5.5 garde l'avantage sur DeepSWE et sur le rappel en long contexte (MRCR v2). GPT-6 Luna est très légèrement devant sur l'exactitude factuelle et le raisonnement long contexte (moins d'un point d'écart).

3. Vitesse



Débit de génération annoncé, sur GPU NVIDIA, en conditions de laboratoire.

En débit brut, Patchie 6 Nano génère environ quatre fois plus vite que Claude Haiku 5.5 et huit fois plus vite que GPT-6 Luna. Les mesures indépendantes en conditions réelles, via une API partagée par de nombreux clients, donnent une médiane plus basse, de l'ordre de 200 à 320 tokens/s, avec environ une seconde de latence. Le temps avant le premier token est le point à surveiller : il a été signalé comme plus lent que le reste de la génération.

4. Quand utiliser Patchie 6 Nano

Usage	Modèle conseillé	Pourquoi
Assistant temps réel, chat grand public	Patchie 6 Nano	Vitesse, respect des consignes
Support client, agents avec outils	Patchie 6 Nano	Meilleur score τ³-bench Banking
Recherche web, synthèse de sources	Patchie 6 Nano	BrowseComp+, DeepSearchQA
Correctifs de code rapides	Patchie 6 Nano	SWE-Bench Atlas
Agents en terminal, tâches longues	Patchie 6	6 Nano à 3 % sur Terminal-Bench 4.0
Retrouver un détail précis dans un très long document	Patchie 6	MRCR v2 en retrait

Méthodologie

Comparaison avec la concurrence : scores publiés au lancement, effort de raisonnement moyen, sans vérification indépendante à ce jour. Comparaison générationnelle : évaluations Vals AI, mêmes versions de benchmarks pour les deux modèles. Les benchmarks dont la version diffère entre 4.8 Nano et 6 Nano (Code Migration, EMB, Finance Agent) ont été écartés. Débits : chiffres annoncés en laboratoire et mesures publiques en conditions réelles, relevées en octobre 2026.