benchmark-models

benchmark-models

> *Catalogue généré le 2026-05-11*

En une phrase

Compare plusieurs modèles d'IA (Claude, GPT, Gemini) sur la même tâche pour voir lequel est le plus rapide, le moins cher et le plus performant, avec des chiffres au lieu d'impressions.

Quand l'utiliser

- Tu veux savoir quel modèle utiliser pour un type de tâche précis (Claude vs GPT vs Gemini).
- Tu suspectes qu'un de tes skills gstack tournerait mieux avec un autre modèle.
- Tu veux mesurer l'impact d'une nouvelle version de modèle sur tes coûts et tes temps de réponse.
- Tu hésites entre plusieurs modèles pour une nouvelle automatisation et tu veux des données.
- Tu veux établir une référence pour détecter les régressions de qualité quand les modèles évoluent.

Comment l'invoquer

- **Slash command** : `/benchmark-models`
- **Voice triggers** : « compare models » · « model shootout » · « which model is best »
- **Phrases déclencheurs (texte)** : "benchmark models", "compare models", "which model is best for X", "cross-model comparison", "model shootout"

- **Auto-invocation** : Sur demande explicite.

Description détaillée

Le skill `benchmark-models` est un comparatif côte à côte. Tu lui donnes un prompt (ou un skill gstack), il l'envoie au même moment à Claude (via Claude Code), GPT (via le CLI Codex d'OpenAI), et Gemini, puis te ressort un tableau comparatif : temps de réponse, nombre de tokens consommés, coût, et optionnellement une note de qualité décidée par un juge LLM (un autre modèle qui note les sorties sur 10).

À ne pas confondre avec `/benchmark` qui mesure la performance d'une page web (Core Web Vitals). Celui-ci compare des modèles d'IA. Le workflow est interactif : Claude te demande d'abord quel prompt utiliser (un skill gstack, un prompt en ligne, ou un fichier sur disque), puis quels providers inclure. Une commande "dry-run" préalable te montre lesquels sont authentifiés sur ta machine — pas de surprise au moment de payer les appels API.

Le juge qualité ajoute environ 0,05 $ par run. Tu peux le désactiver si tu veux juste comparer vitesse et coût. À la fin du test, Claude résume : le plus rapide, le moins cher, le plus qualitatif (si le juge a tourné), et te propose de sauvegarder les résultats en JSON dans `~/.gstack/benchmarks/` pour pouvoir comparer plus tard. Chaque ligne de tableau montre le coût réel — tu sais ce que tu dépenses avant le prochain run.

Source

- **Plugin** : `gstack`
- **Nom interne** : `benchmark-models`
- **Fichier** : `/home/thymon/.claude/skills/gstack/benchmark-models/SKILL.md`

Revision #2
Created 2026-05-11 21:19:03 UTC by thymon
Updated 2026-05-11 21:36:43 UTC by thymon