Jumeau numérique · Sous-section

Choisir son modèle IA

Comparer les modèles via Artificial Analysis, tester via OpenRouter ZDR, et trancher : DeepSeek V4 Flash, GLM 5.2, Kimi 2.6, Qwen pour l'open source ; GPT 5.6 (ChatGPT) et Fable / Claude Opus 4.8 (Claude Code) pour le closed source.

Source : https://prisme.one/plateforme/ia-juillet-2026/semaine-1-choisir-modele-ia

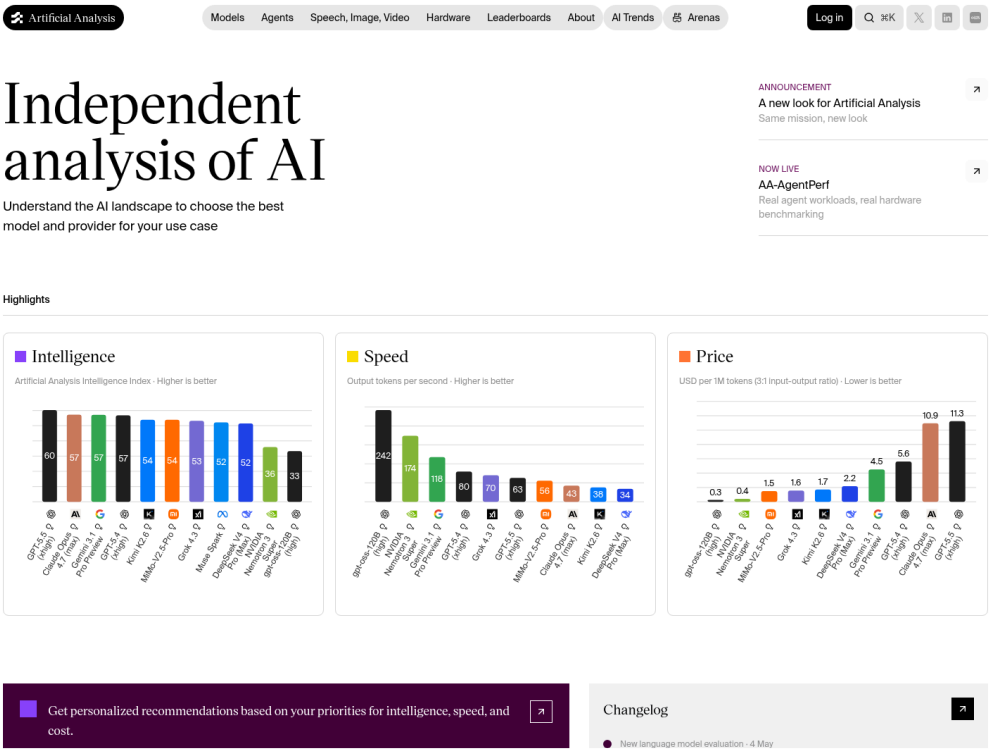
---

Choisir son modèle IA

La recommandation officielle du bootcamp (OpenCode Desktop + ChatGPT + Synthetic AI pour GLM 5.2 + DeepSeek V4 Flash sur OpenRouter) est détaillée dans Comprendre l'anatomie de ton système. Cette leçon va plus loin : comment comparer les modèles, lesquels tester, et comment trancher.

D'abord, regarder où on en est

Le paysage des modèles bouge tous les mois. Avant de te fixer sur un choix, **va voir où on en est réellement** sur Artificial Analysis. C'est le benchmark indépendant le plus utile : il compare intelligence, vitesse, prix sur tous les modèles principaux.



Source : artificialanalysis.ai. À consulter régulièrement.

Le constat que tu fais en regardant : **les modèles dits « high-end » se valent à peu près** sur les tâches du quotidien. À 10 points près sur l'index d'intelligence, ils font globalement le même travail. La vraie différence se joue ailleurs : prix, vitesse, confidentialité, accès via abonnement ou API.

La meilleure façon de choisir : tester

Aucun benchmark ne remplace l'expérience. Tu vas trouver ton modèle préféré en testant plusieurs sur tes propres tâches.

La méthode simple :

1. **Crée un compte OpenRouter** (déjà fait à l'étape précédente si tu as suivi l'installation)
2. **Active le ZDR** dans les paramètres : OpenRouter ne route alors que vers des providers qui ne stockent pas tes prompts
3. **Mets quelques euros de crédit** (10 € suffisent pour des semaines de test)
4. **Dans OpenCode Desktop**, sélectionne le modèle, lance des requêtes sur tes vraies tâches du quotidien
5. **Compare** : qualité de la réponse, vitesse, ressenti général

OpenRouter te donne accès à tous les modèles open source avec une seule clé API et une seule facture. Tu peux basculer de DeepSeek à GLM à Qwen en une seconde.

Les modèles à essayer (mi-2026)

Voici la liste courte des modèles qui valent le détour aujourd'hui. Le paysage évolue vite, mais c'est ce qui est solide à l'été 2026.

Mes tops, à mettre dans ta stack quotidienne :

Open source · privé · Synthetic AI ou OpenRouter ZDR
GLM 5.2 (Zhipu AI)
Mon modèle privé recommandé pour les tâches complexes. Vraiment intelligent, raisonnement solide, gère bien les contextes longs. Dispo en abonnement via Synthetic AI (30 $/mois, ZDR), ou au token via OpenRouter ZDR.
Open source, via OpenRouter
DeepSeek V4 Flash (DeepSeek)
Le modèle par défaut du bootcamp. Très bon, très rapide, ridiculement bon marché (environ 100 fois moins cher que GPT 5.6). Couvre 80 % des tâches quotidiennes.
Closed source, via ChatGPT
GPT 5.6 (OpenAI)
GPT 5.6 est aujourd'hui en tête des benchmarks, devant Fable / Claude Opus 4.8 d'Anthropic. Excellent, surtout sur les tâches créatives, le code, et la génération de contenu long. S'intègre dans OpenCode Desktop via ton abonnement ChatGPT Plus (20 €/mois).
Closed source, via Claude Code uniquement
Fable / Claude Opus 4.8 (Anthropic)
Le meilleur du meilleur sur les tâches qui demandent du raisonnement long, de l'analyse profonde, de la rédaction nuancée. Disponible uniquement via Claude Code (Anthropic a fermé l'accès via interfaces ouvertes). Cf. tutoriel Claude Code.

Aussi solides, à essayer pour varier ou se faire un deuxième avis :

Open source, via OpenRouter
Kimi 2.6 (Moonshot AI)
Excellent sur les tâches agentic et le code long, avec un contexte 256k. Une alternative crédible à GLM 5.2 et DeepSeek, parfois plus original sur les tournures. Bonus : la version Turbo est dispo en illimité via Fireworks FirePass (voir plus bas).
Open source, via OpenRouter
Qwen 3.6 (Alibaba)
Très fort sur le multilingue (notamment langues asiatiques), bonne option de secours. La version 235B est ce qui se fait de mieux sur la branche Qwen aujourd'hui.

Un mot sur les providers

Un modèle est un cerveau. Mais ce cerveau tourne sur des serveurs, et c'est ce qu'on appelle un **provider**. Le même modèle (DeepSeek V4 Flash par exemple) peut être hébergé par plusieurs providers différents : Fireworks, DeepInfra, Parasail, Novita, Together AI, etc. Chacun a sa politique de prix, sa vitesse, sa politique de confidentialité.

Ça peut vite devenir complexe. La bonne nouvelle : tu n'as pas à choisir un provider à la main.

La solution : OpenRouter. OpenRouter est un proxy qui route automatiquement ta requête vers le bon provider en fonction de la disponibilité du moment. Tu paies OpenRouter, OpenRouter paie le provider. Une seule facture. Un seul ZDR à activer. Et si un provider est down, OpenRouter bascule vers un autre sans que tu t'en aperçoives.

Pour les modèles closed source (Claude, GPT), il n'y a pas de provider tiers : Anthropic et OpenAI hébergent eux-mêmes. Pour eux, tu prends l'abonnement direct (ChatGPT Plus 20 €/mois, ou Claude Pro/Max via Claude Code).

Si OpenRouter ne te suffit pas

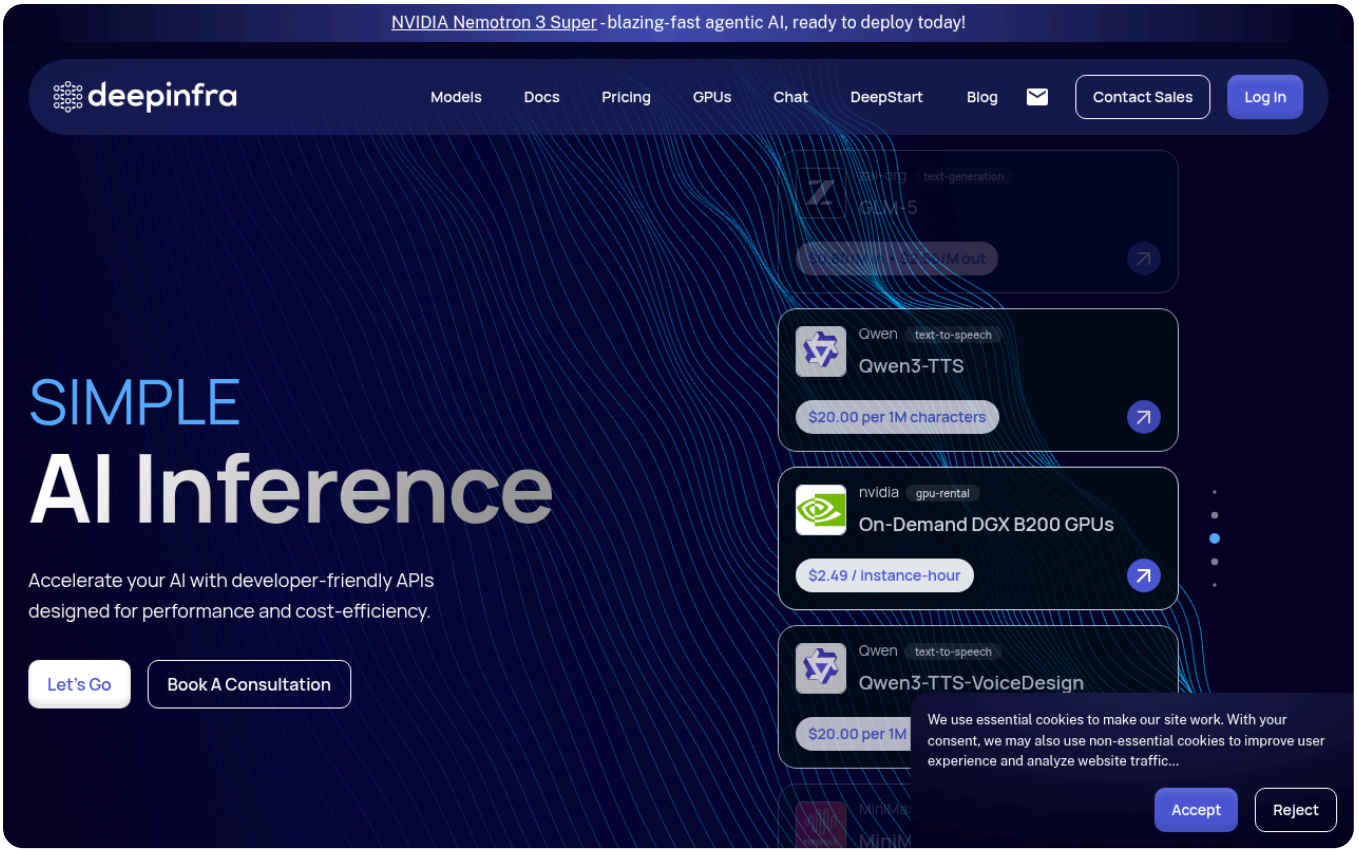
OpenRouter couvre 95 % des cas. Mais parfois tu veux un setup plus pointu : aller chez le provider directement pour bénéficier d'un plan illimité, d'un ZDR contractuel renforcé, d'une infra européenne, ou de la vitesse maximale possible. Voici les providers que je recommande, par ordre de préférence.

Mon coup de cœur du moment · abonnement + ZDR
Synthetic AI

En ce moment, **Synthetic AI** est vraiment excellent. C'est un **abonnement à 30 $/mois** (pas de paiement au token) qui te donne des tokens **ZDR (zéro rétention de données)** sur **GLM 5.2**, servis très rapidement, pour **à peu près le prix d'un abonnement ChatGPT**. C'est ma recommandation pour tout ce qui est privé : difficile de faire mieux aujourd'hui sur le rapport intelligence / vitesse / confidentialité / prix.

synthetic.new
Le moins cher + ZDR
DeepInfra

Le provider **le moins cher du marché** sur la plupart des modèles open source (à partir de 0,02 $/M tokens). **ZDR de série**, SOC 2 et ISO 27001. 190+ modèles open source. C'est aussi un des providers vers lesquels OpenRouter route par défaut en mode Privacy. Bonne option si tu veux passer en direct pour réduire la facture ou simplifier ton stack.



deepinfra.com
100 % européen
Infomaniak AI Tools

Provider **suisse**, datacenters en Suisse, alimentés en énergie renouvelable. GDPR + nLPD-natif, OpenAI-compatible, modèles open source disponibles (Llama 3, Mistral 3, Qwen 3, Gemma 3n, modèles de transcription et d'image). Facturation à l'usage, et **1 million de crédits offerts** pour tester un mois sans engagement. Le meilleur choix si tu veux un acteur français/suisse de confiance pour des données vraiment sensibles (clients EU, santé, juridique).

Setup éclair dans OpenCode

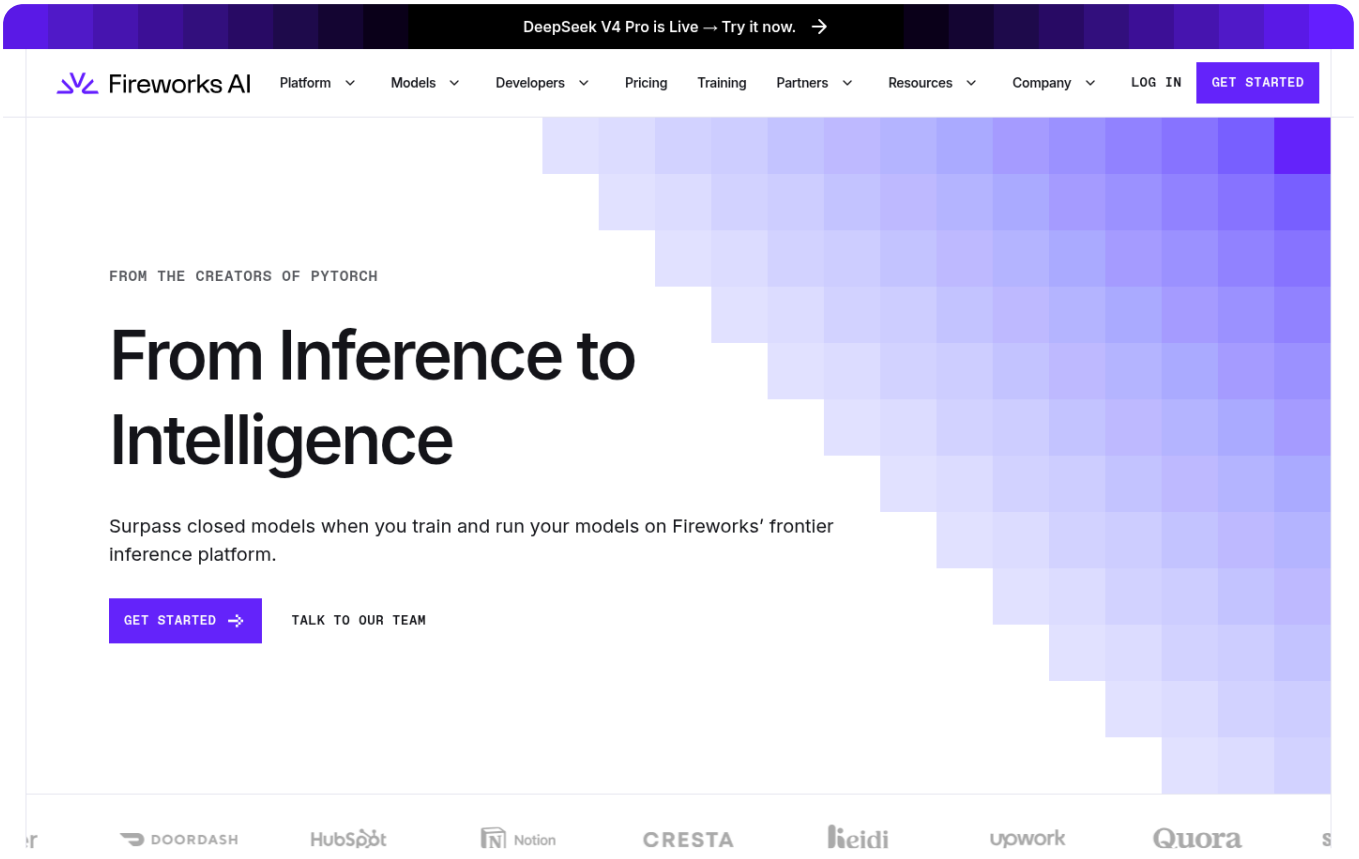
1. Crée un compte sur infomaniak.com et active **AI Tools** (1 M de tokens offerts)
2. Dans ton manager Infomaniak, copie ton **product ID** et génère un **API token**
3. Dans `opencode.json`, ajoute un provider custom de type OpenAI compatible avec `baseURL` = `https://api.infomaniak.com/2/ai/&#123;product_id&#125;/openai/v1` et ton token en header
4. Sélectionne le modèle Mistral 3 ou Qwen 3 dans le sélecteur d'OpenCode et tape ta première requête

Image unavailable in source: Infomaniak AI Tools homepage

https://prisme.one/images/ia-mars-2026/infomaniak-ai-tools.png

infomaniak.com/hosting/ai-services
Alternative intéressante pour usage agent intensif
Fireworks AI (FirePass)

Fireworks est l'un des providers les plus rapides du marché. Leur botte secrète : **FirePass**, un abonnement à **49 $/mois** qui donne accès à **Kimi 2.6 Turbo en illimité**, sans coût par token. Conçu pour les harnesses agentic (OpenCode, Claude Code, Cline). Si tu itères beaucoup avec une IA en mode agent, le ratio est imbattable.



fireworks.ai · docs.fireworks.ai/firepass

> Tu n'as pas besoin de t'inscrire chez tous. **Synthetic AI** pour le meilleur rapport qualité/prix du moment et ton usage privé (abonnement ZDR sur GLM 5.2, 30 $/mois, autour du prix de ChatGPT). **DeepInfra** pour le tarif le plus bas à la consommation. **Infomaniak** si tu as une contrainte juridique européenne et que tu veux un acteur suisse de confiance. **Fireworks FirePass** en alternative si tu utilises l'IA tous les jours en mode agent (illimité à 49 $/mois).

La confidentialité en 3 dimensions

Si tu manipules des données sensibles (clients, projets perso, journaling), trois questions importent :

1. Rétention
Le provider stocke-t-il tes prompts ? Et combien de temps ?
2. Entraînement
Tes prompts servent-ils à entraîner de futurs modèles ?
3. Modération humaine
Tes prompts peuvent-ils être lus par un humain pour des raisons de modération ?

Le label **ZDR (Zero Data Retention)** répond non aux deux premières questions et limite la troisième. Le provider s'engage contractuellement à ne rien stocker, ne rien réutiliser pour l'entraînement, et à supprimer les logs immédiatement.

Le vrai ZDR passe par une offre API/dev, pas par un abonnement grand public

C'est le point que beaucoup ratent, et il faut bien distinguer **deux types d'abonnements** :

- **Abonnement grand public** (ChatGPT Plus, Claude Pro) : conçu pour le chat, **pas de ZDR**. Tes conversations sont retenues pour la modération.
- **Offre API/dev avec ZDR contractuel** : soit au token via une API ZDR-compliante (OpenRouter en mode Privacy), soit en **abonnement dev** comme **Synthetic AI** (30 $/mois, qui sert du GLM 5.2 en ZDR réel). Là, le ZDR est bien contractuel.

Autrement dit, ce n'est pas « abonnement = pas de ZDR ». C'est : **abonnement grand public = pas de ZDR**, mais **offre dev ZDR (Synthetic AI, ou API au token) = ZDR réel**.

Vrai ZDR possible · abonnement dev
Synthetic AI (GLM 5.2)
Abonnement à 30 $/mois qui sert du GLM 5.2 avec un **ZDR contractuel réel**. C'est un abonnement, mais côté dev : rien n'est stocké, rien n'est réutilisé pour l'entraînement. Idéal pour ton usage privé.
Vrai ZDR possible · au token
Modèles open source via OpenRouter API
DeepSeek, GLM, Kimi, Qwen passés via OpenRouter en mode Privacy ne sont routés que vers des providers ZDR-compliants (DeepInfra, certains autres). Rien n'est stocké, rien n'est

utilisé pour l'entraînement.
Pas de ZDR · grand public
Abonnements ChatGPT Plus et Claude Pro
OpenAI et Anthropic gardent tes conversations **au moins 30 jours** pour modération abus, avec lecture humaine possible. Pour Claude, si tu acceptes le partage pour entraînement, ça peut monter à **5 ans**. Le ZDR n'est tout simplement pas une option sur l'abonnement grand public.

Ce que ça veut dire concrètement :

- **Synthetic AI (abonnement dev)** : GLM 5.2 en ZDR contractuel réel pour 30 $/mois. Le moyen le plus simple d'avoir du ZDR sur ton usage privé, sans compter les tokens. C'est l'option qu'on met en avant.
- **OpenRouter Privacy mode** : Settings, Privacy, activer ZDR. Un ZDR réel au token, parfait pour DeepSeek et les tests.
- **ChatGPT Plus / Claude Pro** : tu peux désactiver l'entraînement (« Improve the model for everyone » côté OpenAI, « Help improve Claude » côté Anthropic), mais ça ne supprime pas la rétention 30 jours pour modération. Évite d'y mettre des données vraiment sensibles (RGPD client, médical, juridique).
- **Claude via Claude Code** : politique plus stricte que Claude.ai (rétention 30 jours côté API, pas d'entraînement par défaut), mais ce n'est toujours pas du ZDR formel.

Règle simple : **si la donnée est sensible, route-la via Synthetic AI (GLM 5.2) ou OpenRouter ZDR (DeepSeek, GLM 5.2), pas via ChatGPT.com ou Claude.ai.**

Pour une plongée approfondie, va voir Zoom sur la sécurité.

Recommandation pour démarrer

Si tu hésites encore, voici la stack que je te conseille pour les premières semaines :

1. **ChatGPT Plus** (GPT 5.6, abonnement 20 €/mois) : pour les tâches créatives ou non sensibles (présentations, sites web, brainstorming). Non privé.
2. **Synthetic AI** (GLM 5.2, abonnement 30 $/mois, ZDR) : ton modèle privé quand la tâche est complexe ou touche des données sensibles (clients, raisonnement long).
3. **DeepSeek V4 Flash** (OpenRouter au token, ZDR) : ton défaut ultra bon marché, pour 80 % de tes tâches courantes.
4. **Fable / Claude Opus 4.8** (via Claude Code, en option) : si tu as un abonnement Claude Pro/Max et que tu veux le meilleur sur certaines tâches.

Bascule entre les 4 selon le contexte. C'est ce qui fait la puissance du multi-modèle.

Quand changer ?

Le paysage évolue tous les mois. Une règle simple : **dès qu'une réponse te déçoit, essaye un autre modèle sur la même tâche**. Ça te dit beaucoup sur les forces et faiblesses de chacun. Et ça te garde à jour.

> Pour les détails de connexion (clé API OpenRouter, abonnement ChatGPT dans OpenCode), retourne à Installation technique.

- [ ] Je suis allé voir où on en est sur artificialanalysis.ai
- [ ] Je comprends que tester soi-même reste la meilleure méthode
- [ ] J'ai testé au moins 2 modèles sur mes vraies tâches via OpenRouter
- [ ] Je connais ma stack par défaut : ChatGPT (GPT 5.6) + Synthetic AI (GLM 5.2, privé) + DeepSeek V4 Flash (+ Fable / Claude Opus 4.8 en option)
- [ ] Je sais que ZDR est activé sur OpenRouter dans mes réglages

---

Assets

- Comparaison des modèles IA sur Artificial Analysis
- DeepInfra homepage
- Infomaniak AI Tools homepage
- Fireworks AI homepage
