FELIXIA · AGENCE IA · MONTPELLIER

Tier list IA 2026 : les outils et les modèles qu'on utilise vraiment

Pas un annuaire de 200 outils. La liste courte de ce qui tourne tous les jours chez Felixia, ce qu'on a testé puis abandonné, et ce qu'on refuse d'installer chez un client — avec la raison à chaque fois.

1. Comment lire cette tier list
2. Tier S — le socle
3. Tier A — solides au quotidien
4. Tier B — à tester en 2026
5. Tier C — utiles mais surestimés
6. Tier D — ce qu'on ne conseille pas
7. Modèles : prix réels et arbitrages
8. Trois stacks types et leur coût
9. Les 7 erreurs qu'on voit le plus
10. Parler à un humain

Édition septembre 2026. Prix API et scores relevés le 17/09/2026 — ce marché bouge tous les mois, la méthode compte plus que le classement.

Felixia — Romain Sost · 202 rue des Mégissiers, 34070 Montpellier · contact@felixia.io · www.felixia.io

01

Comment lire cette tier list

Le classement ne mesure pas la puissance brute d'un outil. Il mesure ce qu'il rapporte à une PME ou une agence de moins de 20 personnes, une fois payé, installé, et maintenu.

CRITÈRE	CE QU'ON REGARDE
Valeur sans IA	L'outil sert-il encore si on coupe les modèles ? Un outil qui ne vaut rien sans LLM est un outil dont le coût grimpe avec l'usage.
Coût réel	Abonnement + consommation de tokens + temps d'installation et de maintenance. Le « gratuit open source » se paie en heures.
Surface de risque	Où partent les données, quelles permissions l'outil réclame, ce qu'il peut exécuter tout seul sur votre machine.
Sortie de secours	Peut-on récupérer ses workflows, ses documents, ses prompts et partir ? Sinon c'est une dépendance, pas un outil.
Rendement par euro	Un modèle 15× moins cher qui échoue une fois sur trois coûte plus cher qu'un modèle premium qui passe du premier coup.

RÈGLE DE TRI EN UNE PHRASE

Un outil mérite le tier S s'il tient les trois : il fait gagner du temps *mesurable*, il reste utile quand l'IA est coupée, et son coût ne dérape pas quand l'usage double.

CE QUE CE DOCUMENT N'EST PAS

Aucun lien affilié, aucun partenariat commercial, aucun outil placé contre rémunération. C'est un retour d'expérience daté, pas un benchmark de laboratoire — nos verdicts « à éviter » sont des avis fondés sur nos tests, pas des vérités universelles.

02

Tier S — le socle

Trois outils. Si on devait tout couper sauf trois lignes, ce sont celles-là.

S n8n — automatisation

obligatoire chez nous

Le seul outil de la liste qui produit de la valeur **sans consommer un seul token**. On construit le workflow avec un assistant IA (via son serveur MCP), mais une fois publié, le workflow tourne en pur code : webhook, condition, appel API, écriture Sheets. Zéro coût variable.

- Auto-hébergeable, donc les données clients ne sortent pas de votre infra.
- L'IA devient un nœud parmi d'autres, appelé seulement là où elle est nécessaire.
- Les branches d'erreur sont explicites — un workflow qui casse se voit, contrairement à un agent qui « improvise ».

Piège : un nœud Code qui dépasse 60 s fait tomber le task runner et vous perdez les exécutions silencieusement. Prévoyez un workflow d'erreur + une alerte, dès le premier jour.

S Claude — travail en équipe

must-have dès 2 personnes

Ce qui le rend indispensable n'est pas le modèle seul, c'est le **partage** : des Projects communs, des fichiers de contexte que toute l'équipe voit, et désormais la génération de slides et de documents directement exploitables. On arrête de se renvoyer des prompts par mail.

- Un Project = un client : ses contraintes, son vocabulaire, ses documents. Le contexte ne se perd plus entre deux personnes.
- Les livrables sortent en slides et docs, pas seulement en texte dans un chat.
- C'est aussi la référence de qualité : les modèles concurrents se comparent à lui, rarement l'inverse.

Arbitrage : le plus cher au token de la liste. On l'utilise là où l'erreur coûte cher (code, analyse client, rédaction livrée) et on descend en gamme pour le volume.

S Oh My Pi — agent de code

notre poste de travail

Rapide, économe en tokens, et surtout il **capitalise** : chaque procédure résolue une fois devient une « skill » réutilisable à la session suivante. C'est la différence entre un assistant qui redécouvre votre projet chaque matin et un collègue qui s'en souvient.

- Mises à jour quotidiennes — les corrections arrivent en jours, pas en trimestres.
- Mémoire longue durée + skills : le coût d'un même travail baisse à chaque répétition.
- Serveurs MCP branchés (hébergeur, n8n, navigateur) : il agit sur l'infra, il ne se contente pas de proposer.

Condition d'usage : un agent qui exécute doit travailler dans un dépôt versionné. Sans Git, pas de retour arrière — c'est la seule vraie ceinture de sécurité.

03

Tier A — solides au quotidien

Pas indispensables à tout le monde, mais chacun règle un problème précis mieux que l'alternative payante.

A Google Apps Script

gratuit, déjà dans votre Workspace

Le plus sous-estimé du lot. Un Sheet + 150 lignes de script remplacent un SaaS de reporting à 50 €/mois : consolidation Google Ads, envoi de mails, synchronisation Drive, tâches planifiées. Aucun serveur à payer, aucune donnée qui sort de votre Workspace.

Limite : quotas d'exécution et pas de vrai debug. Au-delà de ~500 lignes, passez sur n8n.

A OpenRouter — routeur de modèles

indispensable si vous testez

Une clé, un format d'appel, des dizaines de modèles derrière. Permet de changer de modèle dans un workflow sans réécrire le code, et de comparer coût/qualité sur vos vrais prompts. C'est aussi la porte d'entrée vers les modèles chinois sans ouvrir un compte chez chacun.

Alternative auto-hébergée : LiteLLM (Apache 2.0) pour le même rôle — mais il faut Redis + Postgres et de la maintenance. Ne vaut le coup qu'à fort volume.

A Chatwoot — support client

alternative Intercom/Zendesk

Boîte de réception omnicanale open source (MIT sur le cœur), auto-hébergeable, avec widget de chat et centre d'aide. Son agent « Captain » apprend de votre FAQ pour absorber les premières réponses.

Honnêteté : auto-héberger un support critique, c'est accepter serveurs, sauvegardes et mises à jour. À partir d'un certain volume, le SaaS revient moins cher que vos heures.

A Docling — documents vers IA

Apache 2.0, local

Convertit PDF, scans et tableaux en Markdown/JSON propre, en local. C'est la brique qui rend un RAG exploitable : la qualité d'une recherche documentaire dépend d'abord du parsing, pas du modèle. Le modèle Granite-Docling (258M) tourne sur une machine ordinaire.

Gain typique : quelques dizaines de centimes d'API pour traiter une centaine de pages, contre plusieurs dizaines d'euros par mois pour un parseur commercial équivalent.

A Langfuse — observabilité LLM

MIT, à partir de 3 workflows IA

Tant que vous avez deux prompts, un fichier suffit. Dès que l'IA est en production chez un client, il faut tracer : quel prompt, quel modèle, quel coût, quelle latence, quelle sortie. Sinon vous débuggez à l'aveugle et vous facturez à l'aveugle.

Coût caché : compter 4 cœurs / 16 Go pour une instance sérieuse. Commencez sur leur cloud, migrez après.

04

Tier B — à tester en 2026

Prometteurs, open source, avec une vraie raison d'exister. On les évalue en interne avant de les poser chez un client — un outil de moins de six mois n'a pas encore prouvé qu'il tiendra.

B Buzz (Block) — chat humains + agents

Apache 2.0, lancé 21/07/2026

Le pari : mettre les agents IA dans la conversation d'équipe comme des membres à part entière, avec leur propre identité cryptographique. Canaux, fils, DM, voix, dépôts de code et automatisations, bâtis sur le protocole Nostr — les données vivent sur un « relay » ouvert, pas sur les serveurs d'un éditeur.

- Auto-hébergeable, ou hébergé par Block sur buzz.xyz.
- Agnostique quant aux modèles — vous branchez ce que vous voulez.
- Vrai intérêt : le contexte de l'agent est la conversation d'équipe, pas un prompt isolé.

À vérifier avant adoption : maturité des permissions par agent, et ce qu'implique un relay pour la confidentialité client. À tester sur un projet interne, pas sur un compte client.

B Papermark — envoi de documents tracé

alternative DocSend

Partage de PDF et data rooms avec analytics page par page : vous savez si votre proposition a été lue jusqu'au prix. Auto-hébergeable, API + CLI + serveur MCP pour l'automatiser. Pour une agence, c'est du signal commercial gratuit sur chaque devis envoyé.

Usage chez nous : candidat pour livrer nos propres documents (dont celui-ci) avec un suivi de lecture, à la place d'un simple lien de téléchargement.

B Paperclip — usine à agents auto-hébergée

MIT · attention au malentendu

Souvent cité comme « alternative à Slack » : ce n'en est pas une. Paperclip est une plateforme open source pour faire tourner une **équipe d'agents IA** sur votre infrastructure, avec vos propres clés et vos modèles. Intéressant pour des tâches répétitives massives, hors sujet pour la messagerie interne.

Pour la messagerie d'équipe open source, les vraies alternatives à Slack sont Mattermost (cœur MIT, auto-hébergé) et Rocket.Chat.

B Supabase — base de données des projets IA

Postgres + pgvector

Devenu le back-end par défaut des applications générées par IA. Le vrai argument ici : c'est du Postgres réel, donc l'extension pgvector permet de stocker les embeddings à côté des données métier — un seul système à sauvegarder au lieu de deux.

Auto-hébergement : licence à 0 €, mais vous perdez sauvegardes gérées et restauration à la minute. En dessous de 25 €/mois de cloud, l'auto-hébergement n'est pas rentable.

05

Tier C — utiles, mais surestimés

Ils marchent. Ils ne méritent simplement pas la place qu'on leur donne dans les discours commerciaux.

C Les chatbots « site web » génériques

Un widget branché sur vos pages répond correctement 8 fois sur 10 et invente le reste. Sans base de connaissances tenue à jour et sans garde-fou sur les sujets sensibles (prix, délais, engagement contractuel), c'est un risque juridique déguisé en gain de temps.

Notre position : d'abord la FAQ et le routage des demandes. Le chatbot après, s'il reste un besoin.

C Les plateformes « agent no-code » clés en main

Démo impressionnante, facture imprévisible : vous payez à l'exécution un raisonnement qui, dans 80 % des cas, aurait pu être une condition *si/alors*. Et vos automatisations ne sont pas exportables.

Test décisif : demandez l'export de vos workflows. Si la réponse est « non », vous ne construisez pas un actif, vous louez une dépendance.

Tier D — ce qu'on ne conseille pas

Testés chez nous, écartés. Avis daté et assumé, pas une accusation.

D Hermes

testé puis écarté

- **Sécurité :** périmètre d'exécution et permissions trop larges pour être posés chez un client.
- **Usine à gaz :** beaucoup de configuration pour un résultat qu'un workflow n8n rend en deux nœuds.
- **Gourmand en tokens :** le coût grimpe sans que la qualité suive.

D OpenClaw

testé puis écarté

- Même diagnostic : complexité de mise en œuvre disproportionnée au gain réel.
- Consommation de tokens élevée pour des tâches que des outils plus simples traitent mieux.
- Surface d'exposition difficile à auditer — rédhibitoire dès qu'on manipule des données clients.

À retenir : la question n'est pas « est-ce que ça marche ? » mais « est-ce que je peux expliquer à mon client ce que ça a le droit de faire ? ».

06

Modèles : prix réels et arbitrages

Prix publics par million de tokens, relevés le 17 septembre 2026. Ils baissent vite : GLM a perdu environ 65 % sur son prix d'entrée en trois mois. Vérifiez avant de bâtir un budget.

MODÈLE	ENTRÉE $/M	SORTIE $/M	CE QU'ON LUI CONFIE
Claude Opus 5	5,00	25,00	Architecture, dossiers clients, tout ce qui est relu par un décideur.
Claude Sonnet 5	2,00	10,00	Le vrai cheval de trait : code au quotidien, rédaction livrée. Meilleur rapport qualité/prix de la gamme.
Claude Haiku 4.5	1,00	5,00	Classification, extraction, tâches répétitives à fort volume.
GLM 5.2 (Z.ai)	0,49	1,56	Volume et contexte long (1M tokens). Résumés, tri, première passe. Cache à 0,09 $/M.
MiniMax M2	0,26	1,02	Le moins cher qui reste utilisable. Traitements de masse où une erreur n'a pas de conséquence.
Kimi K2.5	0,45	2,25	Alternative de repli sur le même créneau.

L'ÉCART DE QUALITÉ, DIT FRANCHEMENT

Les modèles chinois sont excellents *pour leur prix* : GLM et MiniMax coûtent 5 à 20 fois moins que Claude. Mais sur les tâches d'ingénierie mesurées (SWE-bench Verified, Terminal-Bench), la famille Claude reste devant — Sonnet 5 est à 85,2 % là où les meilleurs poids ouverts de MiniMax tournent autour de 80 %. Sur du code livré à un client, ces 5 points sont la différence entre « ça marche » et « je repasse derrière ».

LA RÈGLE DE COÛT QU'ON APPLIQUE

Un modèle 15 fois moins cher qui échoue une fois sur trois vous coûte plus cher : vous payez trois appels, plus votre temps de relecture. Calculez le coût **par tâche réussie**, jamais le coût par million de tokens.

En pratique : modèle économique en première passe (tri, résumé, extraction), modèle premium sur la décision et le livrable. C'est exactement ce que permet un routeur comme OpenRouter.

07

Trois stacks types et leur coût

Ce qu'on installe réellement selon la taille. Les montants sont des ordres de grandeur mensuels, abonnements + consommation.

PROFIL	STACK	ORDRE DE GRANDEUR
Solo / TPE 1 à 3 personnes	Claude (1 siège) + Google Apps Script (déjà payé dans Workspace) + n8n auto-hébergé sur un petit VPS. Aucun autre abonnement avant d'avoir saturé ces trois-là.	25 – 60 €
PME / agence 4 à 20 personnes	Claude en équipe (Projects partagés) + n8n + OpenRouter pour router le volume vers GLM/MiniMax + Langfuse pour tracer + Chatwoot si support client.	150 – 400 €
Équipe technique avec développeurs	Le tout précédent + Oh My Pi sur les postes de dev + Supabase (Postgres + pgvector) + Docling en local pour les documents.	400 – 900 €

DANS QUEL ORDRE, CONCRÈTEMENT

- **Semaine 1 —** un seul processus chronométré avant/après. Pas d'outil nouveau, juste la mesure.
- **Semaine 2 —** ce processus automatisé dans n8n, sans IA si possible.
- **Semaine 3 —** on ajoute un appel modèle là et seulement là où le déterministe ne suffit pas.
- **Semaine 4 —** traçage des coûts et des erreurs, puis on passe au processus suivant.

Une PME qui suit cet ordre dépense moins et garde la main. Celle qui achète d'abord les licences finance des outils que personne n'ouvre.

08

Les 7 erreurs qu'on voit le plus

1. **Mettre de l'IA là où un *si/alors* suffit.** C'est le poste de dépense inutile numéro un. Un workflow déterministe ne coûte rien et ne se trompe jamais.
2. **Acheter les licences avant d'avoir mesuré un processus.** Sans le « avant », vous ne pourrez jamais démontrer le « après ».
3. **Donner à un agent des droits qu'on ne saurait pas expliquer à son client.** Si vous ne pouvez pas décrire son périmètre en une phrase, il est trop large.
4. **Choisir le modèle le moins cher au token.** Le seul chiffre qui compte est le coût par tâche réussie, relecture humaine incluse.
5. **Ne pas tracer.** Sans journal des appels, vous découvrez la dérive de coût sur la facture, un mois trop tard.
6. **Faire travailler un agent de code hors dépôt versionné.** Git est la seule vraie annulation.
7. **Empiler les outils au lieu de les remplacer.** Un outil qui entre doit en faire sortir un autre, sinon vous payez deux fois le même usage.

Parler à un humain

Cette liste dit ce qu'on utilise. Elle ne dit pas ce dont *vous* avez besoin : ça dépend de vos processus, de vos données et de vos contraintes réglementaires.

AUDIT IA, SUR PLACE

On vient chez vous, on chronomètre les processus, on repart avec une liste priorisée de ce qui vaut la peine d'être automatisé — et de ce qui n'en vaut pas la peine.

www.felixia.io/contact

LE RESTE DU LOT

Cette tier list fait partie d'un lot gratuit : plugin WordPress de tracking des leads, 2 scripts Google Ads Performance Max, script Sheets → WordPress + ACF, workflow n8n de visibilité dans les LLM.

www.felixia.io/lot

Felixia — Romain Sost, entrepreneur individuel · SIRET 903 887 818 00016 · 202 rue des Mégissiers, 34070 Montpellier · contact@felixia.io · www.felixia.io

Édition septembre 2026. Prix et scores relevés le 17/09/2026 auprès des pages tarifaires publiques des éditeurs et d'agrégateurs publics. Aucun lien affilié, aucun placement rémunéré.