FAST GROWTH Advisors

LIVRE BLANC

AOÛT 2026

LE MESSAGE RÉCUPÉRABLE

Ce que les moteurs de réponse lisent vraiment de votre site

> *Vingt cas limites, trois extracteurs, 94 audits. Une mesure, pas une opinion.*

Mesuré

Observé dans une étude publiée à grande échelle, ou testé par nous et reproductible.

Déduit

Conséquence logique nécessaire d'un fait mesuré, signalée comme telle.

Non vérifié

Circule sans source primaire. Nous le disons plutôt que de le reprendre.

Ce qui n'y est pas

Aucune tactique, aucune promesse de visibilité. Les limites sont écrites dans le corps du texte.

Ce document ne vend rien. Il mesure ce qu'un moteur retient d'un site, et dit comment le vérifier soi-même.

SECTION 01

EN RÉSUMÉ

Une entreprise B2B soigne deux lectures de son site : celle de l'acheteur et celle de Google. Une troisième s'est installée sans prévenir, celle des moteurs de réponse, et elle n'obéit aux règles ni de l'une ni de l'autre.

Ce document établit trois choses.

D'abord que cette lecture se mesure. Nous avons soumis une page de test de vingt cas limites aux trois extracteurs qui servent à préparer les corpus des grands modèles, et plusieurs résultats contredisent des réflexes hérités du référencement classique.

Ensuite que la French Tech y est massivement défaillante. Sur 94 audits complets de startups post-levée, la lisibilité par les moteurs IA est le plus faible des quinze sous-critères mesurés, à 0,91 sur 5.

Enfin que la perte n'est pas uniforme, et c'est le résultat qui compte. Le discours survit. Les preuves meurent. Un moteur sait alors dire ce que fait l'entreprise, pas ce qui la distingue, et cite comme référence de la catégorie le concurrent qui a mis sa preuve en HTML.

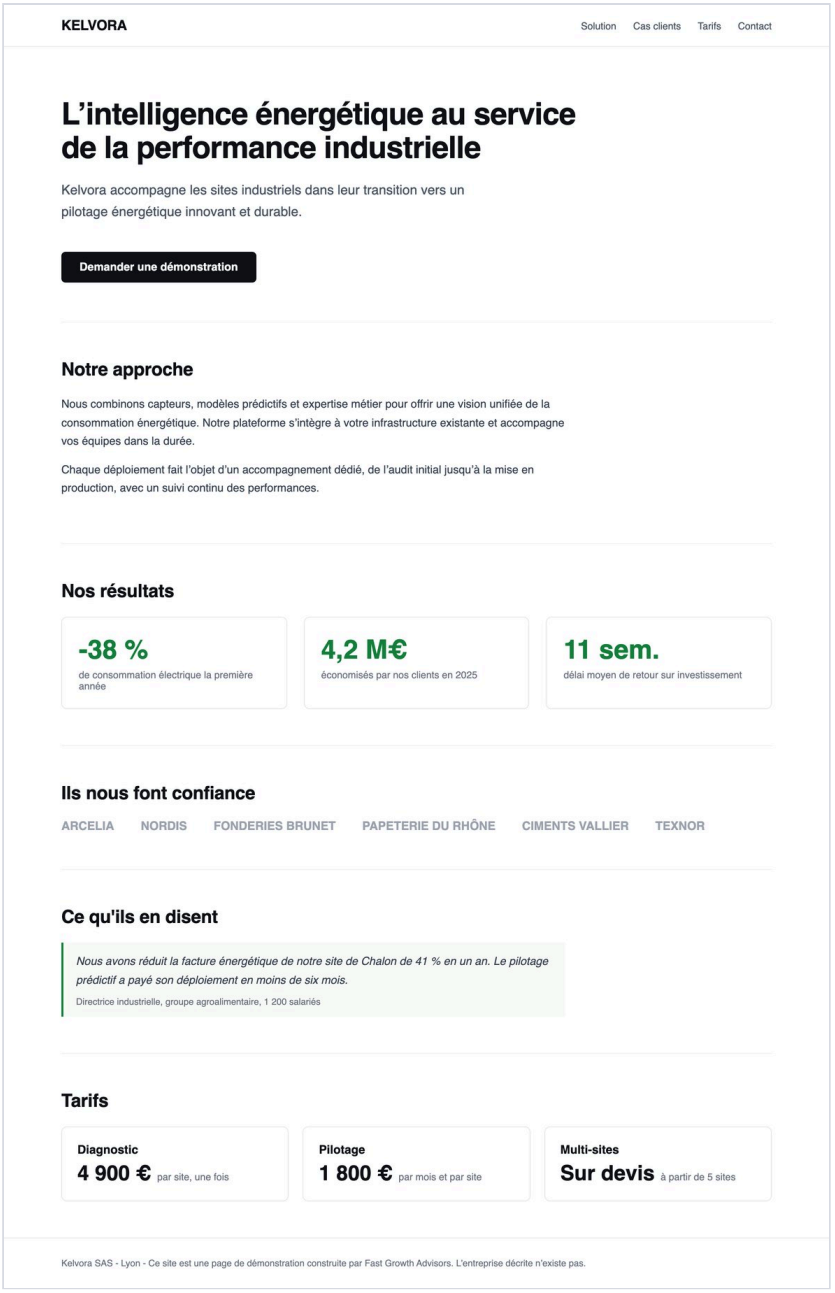
Nous en tirons une grandeur d'audit nouvelle, le message récupérable : la part du positionnement qui survit à une lecture sans JavaScript, extraction comprise. Et un plan d'action qui relève de l'intervention ciblée, pas de la refonte.

SECTION 02

À QUOI RESSEMBLE LA PERTE, SUR UN CAS CONCRET ?

Avant d'expliquer le mécanisme, montrons-le.

Nous avons construit une page d'accueil de startup B2B, fictive, avec les défauts que nous rencontrons le plus souvent en audit : les chiffres de résultat, les logos clients, le témoignage et la grille tarifaire sont injectés par JavaScript, et une ancienne version du positionnement est restée dans le code, masquée en CSS. L'entreprise décrite n'existe pas. La page et le protocole sont reproductibles.



Ce que voit un visiteur. La page affiche une promesse, trois chiffres de résultat, six logos clients, un témoignage chiffré et trois tarifs.

Voici maintenant ce que trois extracteurs retiennent de cette même page, sans exécuter le JavaScript, comme le font les robots qui construisent les corpus. Ces sorties ne sont pas reconstituées : elles proviennent de l'exécution réelle de trafilatura, jusText et Resiliparse sur les octets servis.

Ce que retient trafilatura

```
Kelvora accompagne les sites industriels dans leur transition vers un pilotage énergétique innovant et durable.
Demander une démonstrationNous combinons capteurs, modèles prédictifs et expertise métier pour offrir une vision unifiée de la consommation énergétique. Notre plateforme s’intègre à votre infrastructure existante et accompagne vos équipes dans la durée.
Chaque déploiement fait l’objet d’un accompagnement dédié, de l’audit initial jusqu’à la mise en production, avec un suivi continu des performances.
Kelvora divise par deux le coût énergétique d’une ligne de production en quatre-vingt-dix jours, sans arrêt d’exploitation et sans investissement matériel.
```

Ce que retient jusText

```
L'intelligence énergétique au service de la performance industrielle
Notre approche
Nous combinons capteurs, modèles prédictifs et expertise métier pour offrir une vision unifiée de la consommation énergétique. Notre plateforme s'intègre à votre infrastructure existante et accompagne vos équipes dans la durée.
Chaque déploiement fait l'objet d'un accompagnement dédié, de l'audit initial jusqu'à la mise en production, avec un suivi continu des performances.
Nos résultats
Ils nous font confiance
Ce qu'ils en disent
Tarifs
Ancienne version du positionnement, laissée dans le code lors de la refonte
Kelvora divise par deux le coût énergétique d'une ligne de production en quatre-vingt-dix jours, sans arrêt d'exploitation et sans investissement matériel.
```

Ce que retient Resiliparse

```
L'intelligence énergétique au service de la performance industrielle

Kelvora accompagne les sites industriels dans leur transition vers un pilotage énergétique innovant et durable.

Demander une démonstration

Notre approche

Nous combinons capteurs, modèles prédictifs et expertise métier pour offrir une vision unifiée de la consommation énergétique. Notre plateforme s'intègre à votre infrastructure existante et accompagne vos équipes dans la durée.

Chaque déploiement fait l'objet d'un accompagnement dédié, de l'audit initial jusqu'à la mise en production, avec un suivi continu des performances.

Nos résultats

Ils nous font confiance

Ce qu'ils en disent

Tarifs

Ancienne version du positionnement, laissée dans le code lors de la refonte

Kelvora divise par deux le coût énergétique d'une ligne de production en quatre-vingt-dix jours, sans arrêt d'exploitation et sans investissement matériel.
```

LECTURE	OCTETS DE TEXTE	PART	ÉLÉMENTS SUIVIS RETENUS
Vue par un humain, JavaScript exécuté	1450	100 %	8 / 8
trafilatura	706	49 %	1 / 8
jusText	796	55 %	2 / 8
Resiliparse	949	65 %	2 / 8

Volume de texte retenu, et survie des éléments de preuve, sur la page de démonstration.

Le volume seul serait rassurant : entre 49 % et 65 % du texte survit. Il masque l'essentiel. Ce qui disparaît n'est pas réparti au hasard, c'est la part qui prouve. Sur les huit éléments suivis, les extracteurs en retiennent un ou deux, et ce ne sont jamais les mêmes que ceux qui comptent : la promesse passe chez deux d'entre eux, l'ancien positionnement

masqué passe chez les trois. Les six éléments de preuve proprement dits, les trois chiffres de résultat, les logos clients, le témoignage et les tarifs, ne survivent nulle part.

Deux détails méritent qu'on s'y arrête.

Le premier : jusText et Resiliparse conservent les *titres* des sections, « Nos résultats », « Ils nous font confiance », « Tarifs », mais vides. Le moteur sait donc qu'il existe des résultats, des clients et des prix. Il n'en connaît aucun. C'est la situation décrite plus loin dans ce document : il peut dire ce que fait l'entreprise, pas ce qui la distingue.

Le second est plus gênant. L'ancienne promesse laissée en `display:none`, celle que l'équipe croyait avoir supprimée, ressort dans les trois extracteurs. C'est le seul chiffre commercial que cette page transmet réellement, et c'est celui qu'elle ne voulait plus dire.

Reste le canal des métadonnées, qui obéit à d'autres règles. Trafilatura ne retient pas le titre affiché mais l'`og:title`, ici « Kelvora, pilotage énergétique industriel ». Une balise que beaucoup traitent comme un détail social décide donc du titre sous lequel la page entre dans le corpus.

SECTION 03

POURQUOI CE SUJET DEVIENT-IL URGENT MAINTENANT ?

Trois faits datés, et aucun n'est de nous.

Les visites référées par ChatGPT vers les sites B2B ont été multipliées par quatre en un an : environ 645 000 visites mensuelles en juin 2025, 2,6 millions en juin 2026, soit une hausse de 303 %, sur un ensemble de plus de 11 milliards de visites analysées par Demandbase et publié le 12 août 2026. Le cabinet IDC (International Data Corporation) projette que 70 % des acheteurs B2B américains s'appuieront sur l'IA générative pour découvrir, évaluer et sélectionner leurs fournisseurs d'ici 2028. Et depuis le 13 août 2026, Microsoft Clarity affiche un ratio entre pages aspirées et visiteurs renvoyés qui installe le sujet dans les comités de direction : environ 6 000 pages aspirées pour un visiteur, dans l'exemple publié par Microsoft.

Ce dernier chiffre alimente un débat de rémunération qui n'est pas notre sujet. La question opérationnelle est ailleurs. Puisque ces moteurs lisent vos pages en volume, que retiennent-ils ?

Une convention, avant d'entrer dans le dur. Chaque affirmation de ce document est étiquetée Mesuré (observé dans une étude publiée à grande échelle, ou testé par nous et reproductible), Déduit (conséquence logique nécessaire d'un fait mesuré) ou Non vérifié (circule sans source primaire). Un document qui ne fait pas cette distinction finit par faire circuler des inférences comme des faits. Nous en avons fait l'expérience en interne. Nous préférons l'éviter en public.

SECTION 04

POURQUOI FAUT-IL RAISONNER EN TROIS ÉTAGES, ET PAS EN UN ?

L'erreur la plus fréquente consiste à raisonner en une seule étape : « le robot voit la page ».

Il y en a trois, et une information peut survivre aux deux premières puis mourir à la troisième.

Le premier étage est la récupération. L'agent demande l'URL. Il peut échouer sur un pare-feu qui bloque, une page absente, une limitation de débit, un bandeau de consentement qui masque le corps. La correction relève de l'infrastructure.

Le deuxième est la réponse. Le serveur renvoie des octets, et aucun JavaScript n'est exécuté à ce stade. Si le texte n'est pas dans ces octets, il n'existe pas pour la suite. La correction relève de l'architecture de rendu.

Reste le troisième, l'extraction. Un outil y transforme le HTML en texte, et jette une partie du document au passage. C'est l'étage presque toujours oublié, et celui où nous avons mesuré les résultats les moins intuitifs.

Trois étages, trois causes d'invisibilité, trois corrections qui n'ont rien à voir entre elles. Un audit qui confond ces trois niveaux produit un diagnostic faux et envoie le budget au mauvais endroit, typiquement vers une refonte technique quand le défaut se situait dans le balisage, ou l'inverse.

SECTION 05

QUI EXÉCUTE LE JAVASCRIPT, ET QUI NE L'EXÉCUTE PAS ?

La réponse est documentée, et elle est nette.

Mesuré, par Vercel et MERJ (cabinet d'analyse d'audience) en décembre 2024, sur le réseau Vercel et deux infrastructures tierces : les robots d'OpenAI (GPTBot, OAI-SearchBot qui est son robot de recherche, ChatGPT-User), d'Anthropic (ClaudeBot), de Perplexity, de Meta, de ByteDance et de Common Crawl n'exécutent pas le JavaScript. Googlebot et Applebot, eux, rendent les pages dans un navigateur sans interface.

Le détail qui tranche : ces robots téléchargent les fichiers JavaScript sans les exécuter. 11,5 % des requêtes de ChatGPT et 23,8 % de celles de Claude portent sur des scripts, traités comme du texte, jamais comme des programmes.

Deux réserves de fraîcheur accompagnent ce tableau. La mesure date de décembre 2024 et reste la dernière publiée à grande échelle. Aucun éditeur de modèle ne documente sa capacité de rendu, donc tout repose sur de l'observation tierce, à reconfirmer périodiquement.

Et le rendu existe désormais ailleurs. Les navigateurs agentiques exécutent bien le JavaScript, mais sur le poste de l'utilisateur, une page à la fois, sans rien écrire de durable.

D'où le paradoxe qui structure ce document. Le rendu existe, et il se produit exactement là où il n'alimente aucun index. La couche qui décide des citations futures, celle qui construit les corpus et les index de réponse, ne rend rien.

SECTION 06

QUE SURVIT-IL RÉELLEMENT À L'EXTRACTION ?

Le 29 juillet 2026, nous avons soumis une page de test aux trois extracteurs documentés dans les chaînes de préparation de corpus : jusText, Resiliparse et trafilatura. Vingt marqueurs uniques, un par cas limite, dans un corps de texte de 5 389 caractères. La longueur n'est pas un détail : sur un document trop court, le classifieur de jusText range tout en habillage et le test devient faux. Les versions exactes sont consignées dans nos fichiers de test, et la mesure est à rejouer à chaque montée de version.

Les quatre découvertes

Le contenu masqué par CSS est lu. Un bloc en `display:none`, servi par le serveur, ressort dans le texte extrait par jusText et Resiliparse. La raison est mécanique : ne pas exécuter le JavaScript implique ne pas appliquer le CSS, donc l'extracteur n'a aucun moyen de savoir qu'un bloc est masqué. C'est l'inversion complète du référencement classique. Conséquence heureuse, une FAQ repliée dans un élément `details` rendu par le serveur est lue en entier. Conséquence gênante, les résidus de refonte, les variantes de tests A/B et les anciens contenus laissés en `display:none` polluent la lecture que le moteur fait de votre positionnement. Personne n'audite ce que son HTML cache.

Le titre de niveau 1 n'est pas le titre qui gagne. Trafilatura n'inclut pas le `h1` dans le corps extrait. Il prend son titre dans les métadonnées, et dans notre test c'est l'`og:title` qui l'emporte. Cette balise que tout le monde traite comme décorative, destinée aux partages sur les réseaux, est dans au moins une chaîne d'extraction courante le titre qui atteint le corpus. Un `og:title` négligé, ou en désaccord de vocabulaire avec le `h1`, crée deux positionnements concurrents : l'un pour le visiteur, l'autre pour le corpus.

Le test recommandé partout produit un faux positif exactement dans le cas qu'il doit détecter. Une commande `curl` suivie d'un `grep` sur la réponse brute trouve la phrase cherchée même quand elle est injectée par JavaScript : elle est présente comme chaîne littérale dans le bloc `script`, portion que tout extracteur jette et qu'aucun moteur ne lira comme du contenu. Mesuré chez nous : recherche naïve, trouvé ; après retrait des scripts, absent ; dans le texte extrait, absent. Ce faux positif est silencieux et systématique. Toute mesure sérieuse retire les blocs `script`, `style` et les commentaires avant de chercher, ou travaille sur la sortie d'un extracteur réel.

La coquille vide, quantifiée. Une application en rendu client pur sert 128 octets de HTML. Texte utile extrait : zéro octet. Le cas est binaire, et il devient minoritaire, plusieurs plateformes pré-rendant désormais leurs sites par défaut. Il existe encore, et il ne pardonne rien.

À quoi s'ajoutent les pertes silencieuses. Les attributs `data-*`, les titres d'iframe et le JSON embarqué dans la page sont perdus par les trois extracteurs, même présents dans la réponse. L'attribut `alt` des images est peu fiable. Tout contenu en iframe ou porté par une image est à considérer comme non acquis.

Ce que nous ne savons pas

Par symétrie, voici ce qui reste non vérifié et ne doit pas être vendu comme un fait. Le lien causal entre rendu serveur et taux de citation : aucune étude avec groupe témoin n'existe, et les gains de visibilité spectaculaires qui circulent viennent d'études de cas d'éditeurs. La part du web en rendu client pur : aucune source publique ne la mesure. L'extracteur réellement utilisé par chaque éditeur de modèle : aucun ne le publie, d'où notre règle de ne considérer fiable que ce qui passe les trois.

SECTION 07

POURQUOI LES PREUVES PARTENT-ELLES EN PREMIER ?

Voici le résultat qui relie la mécanique d'extraction au chiffre d'affaires.

Sur une startup typique, la couverture du texte est élevée. Le corps de page passe, le discours passe, la promesse passe. Ce qui disparaît est concentré sur une famille précise d'éléments : les logos clients en carrousel JavaScript, les avis en widget tiers, la grille tarifaire à bascule, les tableaux comparatifs dynamiques, les chiffres clés animés au défilement, la liste d'intégrations chargée depuis une interface de programmation, les études de cas en fenêtre modale.

Tout ce qui prouve.

La restitution que produit alors un moteur de réponse est prévisible. Il sait dire ce que fait l'entreprise et décrit correctement la catégorie. Il ne peut pas dire ce qui la distingue, alors il cite comme référence de la catégorie le concurrent qui a mis sa preuve en HTML.

Ce mécanisme se combine avec un motif que l'Observatoire de Fast Growth Advisors relève dans 61 % des audits détaillés : les chiffres qui prouvent la valeur, gains, économies, métriques d'impact, existent bien, mais dans la presse et les communiqués de levée, pas sur le site. Les preuves sont donc doublement absentes, éditorialement d'abord, techniquement ensuite. Une deeptech de série B de notre corpus, 30 millions d'euros levés, un procédé qui réduit les coûts de production de 45 %, présente sur sa page d'accueil des « solutions innovantes de recyclage avancé ». Score d'audit : 24 sur 75.

SECTION 08

OÙ EN EST LA FRENCH TECH, MESURÉE ?

L'Observatoire de Fast Growth Advisors audite le messaging des startups françaises post-levée. Nos mesures portent sur 375 audits simples, 94 audits complets sur quinze sous-critères, et 346 startups croisées avec leur montant de levée.

Le sous-critère le plus faible des quinze est la lisibilité par les moteurs IA : 0,91 sur 5, soit moins de 20 % du potentiel. Aucun des quinze ne dépasse 60 % du potentiel, et les mieux notés, différenciation et clarté, plafonnent à 52 %. C'est le point le plus faible d'un ensemble déjà faible.

Deux autres chiffres éclairent le contexte. 75,9 % des 369 startups françaises post-levée auditées par l'Observatoire se situent sous le seuil critique de clarté, fixé à 37,5 sur 75 (édition Q2 2026). Et la corrélation entre montant levé et score de messaging est quasi nulle, $R^2 = 0{,}036$.

Autrement dit : lever 50 millions plutôt que 5 ne rend ni le message plus clair, ni le site plus lisible. Le problème ne se résout pas par le financement, et il ne se résout pas non plus par l'ajout d'outils, puisqu'il porte sur ce que le site dit et sur la forme sous laquelle il le dit.

SECTION 09

COMMENT ÊTRE TROUVÉ, ET DANS QUELLE LANGUE ?

L'extraction décide de ce qui est retenu d'une page déjà trouvée. Encore faut-il être trouvé. Nos mesures sur les moteurs eux-mêmes ajoutent deux résultats contre-intuitifs.

Le premier tient à la langue.

Reprenons, depuis le début.

Quand un utilisateur pose une question, le moteur ne cherche pas la question. Il la décompose en requêtes qu'il émet lui-même, et ces requêtes se lisent dans les interfaces de programmation officielles. Mesuré sur deux corpus indépendants, 20 questions B2B techniques puis 30 questions réparties dans 15 secteurs : sur les sujets B2B techniques, ChatGPT émet 46 % de ses requêtes en anglais pour des questions posées en français. La littérature de référence du logiciel est anglaise, et le moteur cherche dans la langue où vit l'autorité du sujet.

Conséquence observée sur notre propre site, avec trois signaux convergents mais des volumes modestes, donc à confirmer : la page la plus consultée par les récupérations de ChatGPT, hors accueil, est la version anglaise de notre page sur la visibilité IA, devant sa jumelle française. La version anglaise d'un site B2B français ne sert pas seulement ses clients internationaux. Elle sert ses clients français, par des moteurs qui cherchent en anglais et répondent en français.

Dernier résultat : tous les sujets ne déclenchent pas de recherche. Sur notre panel, les questions de conseil (« lequel choisir », « comment éviter ») sont répondues de mémoire, sans qu'aucune page ne puisse peser, quand les questions nommant des entités déclenchent des requêtes. Avant d'investir sur un sujet, la première mesure est donc : ce moteur va-t-il seulement chercher ?

SECTION 10

QU'EST-CE QUE LE MESSAGE RÉCUPÉRABLE ?

Un audit de messaging mesure classiquement deux grandeurs : le message revendiqué, ce que l'entreprise dit, et le message compris, ce que l'acheteur retient. Ce document en justifie une troisième, indépendante des deux premières.

Le message récupérable est la part du positionnement qui survit à une lecture sans JavaScript, extraction comprise.

Les trois ne se déduisent pas l'une de l'autre.

Un message peut être clair, différenciant, et malgré tout non récupérable. C'est même le cas le plus fréquent chez les entreprises qui ont investi en même temps dans leur discours et dans un site moderne, puisque ce sont précisément les composants les plus soignés, carrousels et compteurs animés, qui disparaissent à l'extraction.

La sortie utile n'est pas un score.

C'est une localisation : quels blocs précis de la page ne survivent pas, et parmi eux, lesquels portent des preuves. Sur les sites que nous auditons, la réponse tient presque toujours en une liste courte, corrigeable en une intervention. C'est ce qui rend le sujet traitable : le diagnostic est chirurgical, pas architectural.

SECTION 11

QUE FAIRE LUNDI MATIN ?

Le test de trente secondes d'abord. Désactiver JavaScript, recharger la page d'accueil, regarder ce qui reste. C'est l'approximation grossière du message récupérable, et elle suffit à détecter les cas graves. Elle ne dit rien, en revanche, de ce qu'un extracteur jette une fois la page récupérée, qui est la seconde moitié du diagnostic.

Ensuite, dans l'ordre du rendement : sortir les preuves du JavaScript, logos, chiffres, témoignages et tarifs en HTML rendu par le serveur, quitte à garder l'animation par-dessus ; remplacer les accordéons hydratés par des éléments `details` rendus par le serveur ; aligner le vocabulaire de l'`og:title` sur celui du `h1` ; auditer ce que le HTML cache, car les résidus de refonte en `display:none` sont lus ; et vérifier son plan de redirections, puisqu'une part importante du budget de récupération des moteurs génératifs se perd en pages mortes, 34,8 % des récupérations de ChatGPT selon la même étude Vercel.

Sur le fichier `llms.txt`, dont on vous parlera : aucun moteur majeur n'a confirmé le lire. Le poser coûte peu. Le présenter comme un levier est une promesse non mesurée, et ce document n'en fait pas.

SECTION 12

QUELLES SONT NOS LIMITES ?

Les mesures d'extraction datent du 29 juillet 2026, sur trafilatura, jusText et Resiliparse, versions consignées dans nos fichiers de test. Les mesures de décomposition de requêtes et de langue datent du 5 août 2026, par les interfaces de programmation officielles des moteurs, outil de recherche activé, sur deux corpus indépendants, pour un coût total inférieur à 20 dollars. Les chiffres d'audit viennent de l'Observatoire de Fast Growth Advisors 2026, dont le corpus et les seuils sont documentés dans le rapport.

Nos limites, écrites avant qu'on nous les trouve. L'échantillon de l'Observatoire est français et post-levée, il ne décrit pas le web. Les mesures sur notre propre site portent sur des volumes modestes. Et la mesure de référence sur l'exécution du JavaScript date de décembre 2024, ce qui est ancien pour ce sujet.

Chaque chiffre de ce document porte sa base et sa date. Si l'un d'eux ne survit pas à votre propre test, écrivez-nous. C'est le principe.

SECTION 13

QUESTIONS FRÉQUENTES

En quoi le message récupérable diffère-t-il du message compris ?

C'est la part du positionnement d'une entreprise qui survit à une lecture sans JavaScript, extraction comprise. Elle se distingue du message revendiqué, ce que l'entreprise dit, et du message compris, ce que l'acheteur retient. Un message peut être clair, différenciant, et malgré tout non récupérable par un moteur de réponse.

Les moteurs de réponse exécutent-ils le JavaScript de mon site ?

Non, pour les robots qui construisent les corpus et les index. GPTBot, ClaudeBot, PerplexityBot et Common Crawl téléchargent les fichiers JavaScript sans les exécuter, d'après la mesure de Vercel et MERJ de décembre 2024. Googlebot et Applebot rendent les pages. Les navigateurs agentiques aussi, mais sur le poste de l'utilisateur, sans alimenter d'index.

Le test « désactiver JavaScript » suffit-il à savoir ce qu'une IA lit de mon site ?

Il donne une première approximation fidèle et gratuite, suffisante pour détecter les cas graves, et c'est déjà beaucoup pour un test qui prend trente secondes. Il ne dit rien, en revanche, de l'étage suivant, l'extraction, où un outil transforme le HTML en texte et jette une partie du document. Une mesure complète passe donc par un extracteur réel, et non par une lecture du HTML brut.

Pourquoi mes logos clients et mes chiffres disparaissent-ils alors que mon texte passe ?

Parce que ces éléments sont presque toujours injectés par JavaScript : carrousels, widgets d'avis, grilles tarifaires à bascule, compteurs animés. Le corps de texte, lui, est rendu par le serveur. La perte est donc concentrée sur ce qui prouve, pas sur ce qui décrit, ce qui explique qu'un moteur sache dire ce que fait l'entreprise sans savoir dire ce qui la distingue.

Faut-il refondre son site pour être lisible par les moteurs IA ?

Rarement. Dans la plupart des cas que nous auditons, le discours passe déjà et ce sont les preuves qui manquent, ce qui relève d'une intervention ciblée sur quelques blocs. La refonte ne se justifie que dans le cas binaire du rendu client pur, où le serveur ne renvoie aucun texte exploitable.

SECTION 14

GLOSSAIRE

Message récupérable

Part du positionnement qui survit à une lecture sans JavaScript, extraction comprise. Grandeur d'audit distincte du message revendiqué et du message compris.

Extraction

Troisième étage de la lecture machine, où un outil transforme le HTML en texte et écarte ce qu'il considère comme de l'habillage. C'est l'étage le plus souvent oublié dans les diagnostics.

Rendu serveur

Mode de construction des pages où le HTML est assemblé par le serveur avant envoi. Il s'oppose au rendu client, où le contenu est construit par le navigateur en exécutant du JavaScript.

jusText

Extracteur qui sépare le contenu de l'habillage en mesurant la densité de mots outils. Il classe en habillage les listes de noms propres sans phrase, ce qui fait disparaître les listes de références clients.

Trafilatura

Extracteur de contenu principal. Il n'inclut pas le titre de niveau 1 dans le corps et prend son titre dans les métadonnées de la page, où l'og:title l'emporte.

Resiliparse

Extracteur rapide issu de la chaîne ChatNoir, utilisé sur de très grands corpus. Il conserve le texte masqué par CSS, puisque le CSS n'est pas appliqué.

og:title

Balise de métadonnée destinée à l'origine aux partages sur les réseaux sociaux. Elle atteint le corpus dans au moins une chaîne d'extraction courante, ce qui la rend structurante et non décorative.

SECTION 15

SOURCES

1. Vercel et MERJ. *The rise of the AI crawler*, 17 décembre 2024. Étude sur journaux serveur : aucun des grands robots génératifs n'exécute le JavaScript, et 34,8 % des récupérations de ChatGPT visent des pages inexistantes. vercel.com
2. Microsoft Clarity. *Scrape-to-Referral insights*, 13 août 2026. Introduction du ratio entre pages aspirées et visiteurs renvoyés, avec un exemple public d'environ 6 000 pour 1. clarity.microsoft.com
3. Labs by Demandbase. *ChatGPT referrals to B2B websites*, 12 août 2026. Visites référées par ChatGPT vers les sites B2B : environ 645 000 par mois en juin 2025, 2,6 millions en juin 2026, soit +303 %, sur plus de 11 milliards de visites mesurées.
4. IDC (International Data Corporation), 2024. Prévision : d'ici 2028, 70 % des acheteurs B2B américains s'appuieront sur l'IA générative pour découvrir, évaluer et sélectionner leurs fournisseurs. idc.com
5. trafilatura, documentation officielle. trafilatura.readthedocs.io
6. jusText, dépôt du projet. github.com/miso-belica/jusText
7. Resiliparse, documentation officielle. resiliparse.chatnoir.eu
8. Nvidia NeMo Curator, chaîne de préparation de corpus documentant l'usage de ces extracteurs. github.com
9. Observatoire du Message-Market Fit, Fast Growth Advisors, 2026. 375 audits simples, 94 audits complets sur quinze sous-critères, 346 startups croisées avec leur montant de levée.