Qu'est-ce que llms.txt ? Un guide SEO sur la norme web de l'IA

By rankstudio.net Publié le 17 octobre 2025

Résumé

Le fichier `/llms.txt` est un nouveau standard web proposé, destiné à aider les *grands modèles linguistiques* (LLM) et les outils d'IA à mieux découvrir, analyser et interpréter le contenu des sites web. D'esprit analogue au `robots.txt` de longue date pour les robots d'exploration web, `llms.txt` agit comme une **carte structurée et organisée des pages clés et des informations d'un site** pour les agents IA. Les partisans soutiennent que, parce que les LLM ont des fenêtres de contexte limitées et ont souvent du mal à extraire le contenu textuel pertinent de pages web complexes, un `llms.txt` rédigé par un humain peut améliorer considérablement la précision de l'IA en orientant les modèles directement vers les ressources en texte brut les plus importantes (Source: searchengineland.com) (Source: www.released.so). Les premiers adoptants — y compris des plateformes de développement et certaines entreprises technologiques — ont commencé à créer des fichiers `llms.txt`, et des outils/générateurs ont émergé pour faciliter l'implémentation (Source: www.released.so) (Source: github.com).

Cependant, **le débat est loin d'être clos**. Certaines voix de l'industrie mettent en garde contre le fait que `llms.txt` pourrait être une solution prématurée ou inutile, arguant que l' optimisation pour les moteurs de recherche (SEO) traditionnelle suffit déjà pour les cas d'utilisation de l'IA. Des représentants de Google ont explicitement déclaré que les Aperçus IA de Google s'appuient sur le SEO standard et *n'utiliseront pas* `llms.txt` (Source: searchengineland.com). De même, des praticiens SEO respectés notent que les mécanismes existants (par exemple, les sitemaps XML ou les licences Creative Commons) peuvent répondre à de nombreux besoins sans un nouveau format de fichier (Source: searchengineland.com) (Source: searchengineland.com). Une analyse empirique montre une adoption négligeable parmi les 1 000 premiers sites web (effectivement 0%) (Source: www.rankability.com) (Source: www.rankability.com), bien que des communautés plus petites signalent des politiques "autoriser l'IA" relativement élevées sur les sites qui l'implémentent *effectivement* (Source: llmscentral.com). En pesant les perspectives des développeurs d'IA, des experts SEO, des opérateurs de sites web et des défenseurs de la vie privée, ce rapport conclut que **`/llms.txt` est une innovation théoriquement convaincante mais dont l'impact pratique est incertain**. Sa valeur dépendra probablement de la question de savoir si les mainteneurs de plateformes d'IA en tiendront réellement compte, et de la manière dont les éditeurs web équilibreront les coûts de rédaction des métadonnées LLM par rapport aux avantages potentiels en termes de portée de l'IA.

Introduction et Contexte

Alors que l'IA générative et les grands modèles linguistiques (LLM) comme GPT d'OpenAI et Gemini de Google deviennent des interfaces omniprésentes pour l'information, il y a un intérêt croissant à rendre le Web existant plus **compatible avec les LLM**. Actuellement, les sites web sont principalement conçus pour les lecteurs humains et les moteurs de recherche traditionnels ; les humains naviguent facilement dans des interfaces complexes, et Googlebot indexe les pages via des liens et des sitemaps. Mais les LLM sont confrontés à un handicap critique : des *fenêtres de contexte limitées*. Ils ne peuvent pas ingérer des pages web complexes entières et sont souvent distraits ou confus par les barres de navigation, les publicités, les scripts et autres éléments non textuels (Source: searchengineland.com) (Source: llms-txt.io). Comme le note Jeremy Howard, le technologue à l'origine de la proposition `llms.txt` :

> *« Les grands modèles linguistiques s'appuient de plus en plus sur les informations des sites web, mais sont confrontés à une limitation critique : les fenêtres de contexte sont trop petites pour gérer la plupart des sites web dans leur intégralité. La conversion de pages HTML complexes avec navigation, publicités et JavaScript en texte brut convivial pour les LLM est à la fois difficile et imprécise. » (Source: searchengineland.com)*

Cette limitation fondamentale signifie qu'un agent IA essayant de répondre à la question d'un utilisateur en explorant un site peut manquer l'information clé ou l'interpréter de manière erronée. Les techniques traditionnelles de SEO et de conception web mettent l'accent sur la convivialité humaine et la visibilité pour les moteurs de recherche, mais elles ne répondent pas directement aux besoins des agents IA au moment de l'inférence (Source: llms-txt.io). En pratique, un LLM doit passer au crible le désordre de la page et ne peut conserver qu'un extrait limité. Par exemple, un développeur a rapporté avoir dû aplatir un site de documentation entier en un seul fichier texte de 115 378 mots (966 Ko) pour le fournir à un LLM avec un contexte complet (Source: searchengineland.com).

Pour combler cette lacune, **le fichier `/llms.txt` a été proposé fin 2024** par Jeremy Howard (co-fondateur de Answer.AI et fast.ai) comme une *extension bienveillante* des standards de métadonnées web. L'idée est simple : à la racine d'un site web (tout comme avec `robots.txt`), le webmaster peut placer un fichier Markdown en texte brut nommé `llms.txt` qui contient :

- Un titre H1 avec le nom du site ou du projet (un élément requis).
- Une courte introduction ou un "résumé" sous forme de citation en bloc, donnant un contexte clé.
- Une ou plusieurs sections narratives pour expliquer le site ou son utilisation à une IA.
- Des listes à puces sous des titres H2, chacune listant les pages importantes sous forme de liens Markdown `[Titre](URL)` avec des descriptions facultatives.
- (Facultatif) Une section "Optionnel" distincte pour les liens de moindre priorité que le LLM peut ignorer s'il est contraint.

Un tel fichier vise à fonctionner comme **"une carte au trésor pour l'IA"** (Source: www.linkedin.com). Au lieu de forcer l'IA à analyser le HTML du site web, le `llms.txt` sert de **table des matières organisée** pointant vers tout le contenu pertinent. Le fichier lui-même est écrit en Markdown clair, supprimant les scripts et la navigation afin que le LLM ne voie que du texte brut. En pratique, un agent ou un outil IA peut récupérer "/llms.txt" et voir, par exemple, un titre, un résumé de l'entreprise, puis des sections comme "Produits" ou "Docs" avec des liens à puces. Cela donne au modèle un accès immédiat aux pages et au contexte que ses créateurs considèrent comme les plus importants.

Cette notion fait écho aux efforts historiques pour rendre le web "compréhensible pour les machines". En fait, les critiques l'ont comparée à l'initiative longtemps dormante du **Web sémantique**, qui tentait d'annoter le contenu web pour l'interprétation machine (Source: news.ycombinator.com). La vision de Tim Berners-Lee, vieille de plusieurs décennies, d'agents "analysant toutes les données sur le Web" dans un "Web sémantique" de machine à machine n'a jamais été pleinement réalisée (Source: news.ycombinator.com). L'approche `llms.txt` évite les ontologies lourdes ou les schémas RDF, s'appuyant plutôt sur du texte brut. Comme l'a observé un partisan, elle évite la complexité qui a écrasé l'effort du Web sémantique et utilise des "formats sans état" (Markdown, XML) pour communiquer avec l'IA (Source: news.ycombinator.com).

De manière cruciale, `llms.txt` ne concerne pas le *blocage* ou le contrôle légal, mais le *guidage* de l'IA. Contrairement à `robots.txt` (qui utilise des règles "Disallow: URL" pour interdire l'indexation), `llms.txt` n'a **aucune directive de blocage**. Il est entièrement facultatif et instructif – le propriétaire du site choisit les pages à mettre en évidence. Les implémenteurs soulignent qu'il s'agit "plutôt d'un choix concernant le contenu qui doit être montré contextuellement ou entièrement à une plateforme d'IA" (Source: searchengineland.com). En fait, il dit à un LLM "si vous voulez apprendre notre site, voici exactement où chercher". Par exemple, Howard et ses collaborateurs décrivent l'utilisation d'un petit `llms.txt` pour alimenter des outils comme Cursor ou Claude avec une documentation précisément organisée, évitant à chaque utilisateur de devoir rassembler manuellement le contexte (Source: news.ycombinator.com).

Ainsi, `/llms.txt` incarne une **vision collaborative** : les sites web collaborent explicitement avec les "agents" IA de la même manière qu'ils collaborent avec les moteurs de recherche. Comme le résume un article, *« LLMs.txt est sur le point de changer la façon dont votre contenu est vu, utilisé et protégé dans le monde des grands modèles linguistiques »* (Source: llmsly.com). Dans cette optique, il permet aux créateurs de contenu de "contrôler leur récit" en informant l'IA avec des informations faisant autorité (Source: www.linkedin.com). Les avantages proposés vont de l'amélioration de la précision des réponses de l'IA à un trafic potentiellement mesurable provenant des interfaces de recherche alimentées par l'IA. Les premières expériences des praticiens ont donné des signaux mitigés mais intrigants : les moteurs comme les modèles d'OpenAI explorent apparemment ces fichiers, tandis que Google Search (jusqu'à présent) ne les utilise pas automatiquement (Source: searchengineland.com) (Source: searchengineland.com).

Cependant, la proposition `llms.txt` n'est pas universellement acceptée. Les critiques soulignent les *tensions entre élégance et praticité*. Bien que `llms.txt` puisse simplifier l'exploration par l'IA, il duplique essentiellement ce que le contenu bien conçu devrait déjà faire : être accessible et clair pour *tous* les lecteurs (humains ou IA). Comme l'a noté un commentateur, "Ce n'est pas une bonne UX pour les machines. C'est un correctif pour une mauvaise UX" – un pansement plutôt que de corriger les mises en page imprécises sous-jacentes (Source: news.ycombinator.com). D'autres craignent que sans un processus de normalisation robuste (par exemple, l'enregistrement formel d'un URI bien connu ou de balises meta), le format puisse se fragmenter. Des experts de haut niveau avertissent également que le fait d'exiger des propriétaires de sites qu'ils rédigent manuellement un autre fichier les

alourdit, étant donné qu'*aucun système d'IA ne l'utilise actuellement* (Source: searchengineland.com) (selon Google) ou n'a apparemment demandé un tel fichier. Il existe même un point de vue selon lequel les licences web existantes (Creative Commons, etc.) pourraient régir l'utilisation de l'IA plus proprement qu'un nouveau fichier texte (Source: searchengineland.com).

Dans les sections suivantes, nous approfondirons **ce qu'est `/llms.txt`, comment il est censé fonctionner et pourquoi il peut ou non être important**. Nous examinerons la spécification technique et le format (tel que proposé actuellement), les outils de génération et les différences avec les standards connexes comme `robots.txt` et `sitemap.xml`. Nous passerons en revue l'**état actuel de l'adoption**, y compris des études de cas (par exemple, des entreprises testant `llms.txt` pour la documentation produit) et des données sur le nombre de sites qui l'ont implémenté. Nous résumerons les perspectives des développeurs d'IA, des spécialistes SEO et des défenseurs de la vie privée, en utilisant des entretiens et des déclarations publiées. Nous discuterons également de la réaction des plateformes d'IA – certaines testant activement `llms.txt`, d'autres restant agnostiques (Source: searchengineland.com) (Source: searchengineland.com). Enfin, nous exposerons les implications potentielles pour l'avenir : de la manière dont les entreprises gèrent leur contenu numérique à la manière dont les moteurs de recherche et génératifs évolueront. Grâce à des citations et des analyses exhaustives, le rapport vise à répondre : *`/llms.txt` est-il vraiment révolutionnaire pour la recherche IA, ou juste un autre élément de désordre numérique ?* Les premières preuves suggèrent qu'il pourrait être important pour des cas d'utilisation de niche (comme la documentation pour développeurs et les petits sites), mais son impact global sur la découverte web grand public reste à voir.

Le standard `/llms.txt` : Détails techniques et objectif

La proposition et la spécification de `/llms.txt` sont documentées de manière la plus complète par ses créateurs sur [llmstxt.org] et les dépôts GitHub associés (Source: llmstxt.org) (Source: github.com). En bref, un fichier `llms.txt` est un document **Markdown** en texte brut, situé à la racine d'un site web (par exemple, `https://example.com/llms.txt`). Il utilise la syntaxe Markdown pour présenter un contenu structuré, le rendant à la fois *lisible par l'homme* et analysable par les machines. Le format évite intentionnellement les imbrications arbitraires ou les balises inconnues, au profit d'un agencement bien défini de titres, de paragraphes, de citations en bloc et de listes. L'élément minimal requis est simplement un titre de niveau supérieur (H1) contenant le titre du site ou du projet (Source: llmstxt.org). Au-delà de cela, la spécification définit les composants suivants, dans l'ordre :

- **Titre H1 (requis)** – Le nom du projet ou du site web (par exemple, un nom d'entreprise). Cela ancre l'identité du fichier.
- **Résumé en texte brut (facultatif)** – Une **citation en bloc Markdown** contenant une brève description ou une déclaration de vision. Ce "pitch d'ascenseur" donne un contexte dès le départ.
- **Sections d'introduction (facultatif)** – N'importe quel nombre de paragraphes ou de listes (mais *pas* de titres supplémentaires) donnant des détails sur le site ou des instructions pour interpréter les liens suivants. Ceux-ci peuvent être du texte brut, des listes à puces, etc.
- **Sections de liens H2 (facultatif)** – Zéro ou plusieurs sous-sections, chacune précédée d'un H2. Chaque H2 est suivi d'une liste à puces de liens (ancres Markdown `[texte](URL)`), éventuellement avec des notes séparées par des deux-points. Celles-ci compartimentent le contenu du site par catégorie. Par exemple :

  ```
  ## Documentation
  - [Référence API](https://example.com/api) : Documentation API détaillée pour les développeurs.
  - [Guides](https://example.com/guides) : Tutoriels étape par étape.
  ```

  Ces sections sont traitées comme des "listes de fichiers" d'URL dans la spécification ; les LLM ou les outils peuvent les parcourir.
- **Section "Priorité inférieure" facultative** – Il est recommandé (mais non obligatoire) qu'une dernière section intitulée "Optionnel" liste les pages de moindre priorité, afin qu'un LLM puisse les ignorer si sa fenêtre de contexte est limitée.

Cette structure vise à imiter la façon dont les humains pourraient résumer l'architecture de l'information d'un site. Le fichier *lui-même* est écrit en Markdown spécifiquement *parce que* Markdown est facilement analysable par les LLM et les humains (Source: llmstxt.org) (Source: golevels.com). Le format est suffisamment non ambigu pour que les outils automatisés puissent le traiter à l'aide d'une simple analyse de texte (même des méthodes basées sur des expressions régulières ou XML, comme le montre l'exemple de FastHTML) (Source: llmstxt.org) (Source: github.com). Il est essentiel de noter que la spécification souligne que le

contenu de `llms.txt` doit être concis et pertinent — il ne doit pas simplement déverser le contenu entier des pages sans discernement. Au lieu de cela, il **met en évidence** les URL et les faits que le propriétaire du site juge les plus importants pour l'ingestion par l'IA.

Par exemple, la [spécification officielle llmstxt.org] (et la [description GitHub d'AnswerDotAI]) fournit une maquette illustrative :

```
# Titre du site exemple

> Ceci est un résumé concis de l'objectif et des offres clés du site web. Il peut mentionner l'industrie, les produits ou

Les sections suivantes listent les zones de contenu les plus importantes de ce site à prendre en compte par l'IA.

## Guides

- [Démarrage rapide](https://example.com/start) : Une introduction pour les nouveaux utilisateurs.
- [Docs API](https://example.com/api) : La référence API complète.
- [FAQ](https://example.com/faq) : Questions fréquemment posées.

## Projets

- [Projet Alpha](https://example.com/alpha) : Informations détaillées sur le Projet Alpha.
- [Projet Beta](https://example.com/beta) : Aperçu du Projet Beta.

## Optionnel

- [Blog](https://example.com/blog) : Actualités et mises à jour (à ignorer si limité).
```

Cet exemple démontre l'utilisation prévue : une IA lisant `llms.txt` voit un résumé puis des listes clairement structurées d'URL pertinentes avec de courtes étiquettes ou notes. Grâce à cela, les modèles peuvent précharger des résumés de pages clés au lieu d'explorer l'ensemble du site à l'aveuglette.

Un aspect clé de `llms.txt` est qu'il **ne tente pas de remplacer les standards du web**, mais de les compléter pour l'utilisation par l'IA. Par exemple, il pourrait fonctionner *implicitement* comme un sitemap additionnel (listant des pages) mais avec un contexte descriptif. La spécification ne définit *pas* explicitement de règles restrictives ; elle est plutôt informationnelle. Comme le note un explicateur, `llms.txt` est « similaire à robots.txt… mais il offre également un avantage supplémentaire – l'aplatissement complet du contenu » (Source: searchengineland.com). En d'autres termes, alors que robots.txt indique aux machines *ce qu'il ne faut pas explorer*, `llms.txt` indique aux machines *ce qu'il faut explorer* (et pourquoi). Il s'apparente davantage à un **sitemap étendu et organisé par des humains** combiné à de la documentation. En effet, un guide le qualifie formellement de *« le nouveau robots.txt pour l'ère des LLM »* (Source: www.released.so), soulignant qu'il **guide** les LLM pour éviter les approximations.

Sur le plan pratique, la proposition llms.txt et les outils associés envisagent que les pages web contenant du contenu utile proposent également des « versions Markdown propres » de ces pages (par exemple à la même URL mais avec une extension `.md`) (Source: llmstxt.org). Cette suggestion est similaire à la fourniture de HTML pré-traité pour les machines, mais elle n'est pas strictement requise par le standard llms.txt lui-même. Le principal livrable de cette initiative est le fichier llms.txt, qui peut également lister des liens optionnels (dans ses sections) vers de telles ressources Markdown si elles sont disponibles. Certains projets, comme FastHTML, sont allés plus loin en convertissant par programme leurs pages spécifiques aux mm en Markdown, puis en les référençant dans les listes llms.txt (Source: github.com). L'exemple de FastHTML est instructif : son `llms.txt` a été automatiquement étendu en fichiers « llms-ctx.txt » et « llms-ctx-full.txt » qui incorporent le texte des pages liées, adaptés aux besoins de contexte XML du modèle Claude (Source: github.com).

En résumé, `llms.txt` est une **convention** — pas encore un standard IETF formel — pour la publication de métadonnées de site consommables par l'IA. Il prescrit un nom de fichier et un format spécifiques, mais laisse une grande flexibilité aux propriétaires de sites. L'espoir est qu'en annonçant et en documentant cette convention (via llmstxt.org et GitHub), les développeurs et les

entreprises commenceront à l'adopter volontairement. Si suffisamment de fournisseurs de contenu le font, les développeurs d'IA (ou les outils pour utilisateurs finaux) pourraient vérifier par programme `yourwebsite.com/llms.txt` comme une source fiable de contenu intégré à la page.

Relation avec les standards existants (Robots.txt, Sitemaps, etc.)

Pour évaluer la signification de `llms.txt`, il est crucial de le comparer aux standards web plus établis qui servent les moteurs de recherche. La comparaison la plus naturelle est `robots.txt`, qui régit le comportement des crawlers web depuis les années 1990. Bien que `robots.txt` et `llms.txt` partagent l'idée d'un fichier bien connu à la racine du site, leurs fonctions divergent fortement. `robots.txt` est un **ensemble de commandes** pour les robots web : il indique aux moteurs de recherche (via des directives comme `User-agent` et `Disallow`) quelles parties du site ne doivent *pas* être explorées ou indexées. En revanche, `llms.txt` **ne concerne pas le blocage**. Il fournit des indications positives — essentiellement une table des matières rapide — sur *ce qu'il faut inclure* dans le contexte d'un LLM. Comme l'explique Search Engine Land, « les fichiers robots.txt fonctionnent très bien pour les crawlers et n'ont pas besoin d'être modifiés pour les LLM » (Source: searchengineland.com), car le cas d'utilisation de robots.txt (régir les autorisations d'exploration) est orthogonal à celui de llms.txt (améliorer l'ingestion de contenu).

Un autre analogue utile est le sitemap XML (`sitemap.xml`). Un sitemap est simplement une liste d'URL formatées en XML, éventuellement avec des métadonnées comme les dates de dernière modification ou les priorités, entièrement destiné aux moteurs de recherche. Il ne contient pas de contexte descriptif ni de résumés ; il énumère simplement les pages pour la découverte. En revanche, `llms.txt` est comme un **sitemap contextuel**. Il liste toujours des liens, mais sous une forme annotée et lisible par l'homme. Un guide marketing note que « contrairement à un sitemap.xml (qui n'est qu'une liste d'URL), llms.txt fournit un contexte et une structure pour chaque lien » (Source: golevels.com). D'une certaine manière, on peut considérer `llms.txt` comme fusionnant les concepts d'un sitemap et d'une forme de page « À propos » : il énumère les pages clés et explique ce qu'elles sont.

Nous pouvons résumer quelques distinctions clés dans le tableau ci-dessous :

ASPECT / FICHIER	ROBOTS.TXT	SITEMAP.XML	LLMS.TXT
Objectif	Contrôler l'indexation des crawlers (interdire des pages) (Source: searchengineland.com)	Informer les robots de recherche de toutes les URL et métadonnées du site	Guide organisé du contenu important pour les LLM (Source: llms-txt.io) (Source: golevels.com)
Type de contenu	Directives en texte brut (ex. `Disallow:`)	XML avec des entrées `<url>`	Markdown : titres, listes, liens, texte
Public/Agent	Crawlers de moteurs de recherche (Googlebot, etc.)	Crawlers de moteurs de recherche	Systèmes d'IA et agents basés sur les LLM
Différence clé	Indique aux bots *ce qu'il faut ignorer*	Liste *toutes les pages à inclure*	Met en évidence *ce sur quoi se concentrer*
Lisible par l'homme ?	Oui (commandes simples)	Non (format XML machine)	Oui (Markdown simple avec descriptions) (Source: golevels.com)
Exemple d'utilisation	`Disallow: /private/` bloque le chemin	`<loc>https://example.com/page.html</loc>`	`- [FAQ](https://exa.com/faq) : sujets courants`

(Sources : Consultation des propositions llms.txt et des guides SEO (Source: searchengineland.com) (Source: golevels.com) (Source: llms-txt.io).)

Ce qui précède souligne que les standards existants répondent à des besoins différents. L'optimisation SEO traditionnelle (via un HTML approprié, des balises meta, des données structurées, des sitemaps, etc.) reste fondamentalement axée sur les utilisateurs humains et les algorithmes de Google (Source: llms-txt.io) (Source: llms-txt.io). `llms.txt` reconnaît explicitement que *ces méthodes sont insuffisantes pour l'IA*. En effet, comme le note une analyse, les LLM « ont une capacité finie à traiter l'information en une seule fois » et « le contenu optimisé par mots-clés ne fournit pas toujours la compréhension complète dont les LLM ont besoin » (Source: llms-txt.io). En d'autres termes, un site fortement optimisé pour le SEO pourrait bien se classer sur Google mais dérouter une IA en lui faisant manquer de contexte ou en ingérant des informations inutiles. `llms.txt` est proposé comme un supplément — *pas un remplacement* — aux pratiques SEO (Source: llms-txt.io) (Source: llms-txt.io). Un bon SEO (pages rapides, titres clairs, etc.) reste nécessaire pour une visibilité générale, tandis que `llms.txt` garantirait en outre que l'*IA* saisisse l'essence de votre contenu.

D'autres idées connexes dans l'industrie soutiennent cette division. Par exemple, certains ont suggéré d'ajouter des balises `<meta name="LLM">` spéciales ou des indications d'en-tête HTTP pour signaler un contenu adapté à l'IA. Un stratège SEO a même proposé un lien `rel="llm"` ou un profil MIME pour le Markdown compatible LLM (Source: news.ycombinator.com). Ces propositions partagent l'objectif de signaler le contenu pertinent à l'IA, mais elles diffèrent dans leur implémentation. `llms.txt` a été choisi (du moins initialement) comme un simple fichier à la racine pour éviter de nécessiter des modifications de la mise en page HTML ou de la configuration du serveur HTTP. Les partisans de `llms.txt` soutiennent qu'un fichier texte autonome est une solution à faible friction : tout site hébergeant du contenu statique peut y déposer un fichier Markdown sans risque de casser la présentation du site.

Il est important de noter que le géant de la recherche web Google a donné son avis sur cet écosystème proliférant. Dans un rapport de Search Engine Land de juillet 2025, Gary Illyes de Google (de l'équipe Search Central) a explicitement déclaré que **Google ne traitera pas les fichiers llms.txt** : « Les aperçus d'IA de Google reposent sur le SEO standard ; vous n'avez pas besoin de llms.txt ou de tout fichier spécial » (Source: searchengineland.com). Illyes a réaffirmé lors d'une discussion publique que Google « ne prend pas en charge LLMs.txt et n'a pas l'intention de le faire » (Source: searchengineland.com). Au lieu de cela, Google demande aux webmasters *d'utiliser simplement le SEO normal* pour être visibles dans les fonctionnalités « Aperçu IA » basées sur l'IA. En revanche, certains produits d'IA de startups plus petites (comme les moteurs d'OpenAI ou Claude) semblent explorer ou même lire activement ces fichiers. Par exemple, un développeur web a signalé que le crawler d'OpenAI frappait les points de terminaison /llms.txt de ses sites toutes les quelques minutes (Source: searchengineland.com). Ainsi, à l'heure actuelle, il semble que llms.txt puisse être pertinent pour les outils d'IA spécialisés, mais pas pour l'indexation de recherche grand public.

En résumé, **llms.txt occupe un nouvel espace** : il est explicitement destiné non pas aux moteurs de recherche, mais aux agents IA. Il complète plutôt qu'il ne remplace `robots.txt` ou `sitemap.xml`. Il est *inspiré par* ces conventions plus anciennes (d'où son surnom de « robots.txt pour l'IA » (Source: www.released.so), mais ses indications sont d'une nature différente. La question centrale est de savoir si les LLM et les entreprises adopteront cette convention (abordée plus tard), mais techniquement, elle comble une niche unique : rendre le contenu complexe d'un site facilement consommable par l'IA générative.

La raison d'être : pourquoi `/llms.txt` pourrait être important

Comprendre l'importance de `llms.txt` nécessite d'examiner les motivations et les avantages anticipés sous plusieurs angles : pour les propriétaires de contenu, pour les développeurs d'IA et pour les utilisateurs finaux.

1. Contrôle de l'interprétation par l'IA : L'avantage le plus souvent cité est de donner aux propriétaires de sites web un certain contrôle sur la manière dont l'IA utilise leur contenu. Dans le paysage actuel, les grands modèles d'IA s'entraînent généralement sur des extractions web massives et non catégorisées (par exemple, Common Crawl) ou récupèrent des pages ad-hoc sans guide (Source: privacyinternational.org). Les auteurs et les entreprises ont exprimé leur inquiétude que ce processus puisse déformer ou mal interpréter leur contenu — ou que l'IA puisse répondre aux questions des utilisateurs sans donner de « citation » ou de contexte approprié. En fournissant `llms.txt`, un site peut *mettre en évidence* les pages et les données exactes qu'il *veut* que les IA lisent. Cela peut garantir, par exemple, que les descriptions de produits ou les conditions légales sont incluses, tandis que les pages sans importance (comme les menus de navigation, les pages de connexion ou les pages d'erreur) sont exclues. Selon les auteurs de la proposition, cette transparence peut être une forme de gestion des droits de contenu : les sites web peuvent signaler efficacement quel contenu ils autorisent un LLM à « ingérer » pour répondre aux requêtes (Source: llmscentral.com) (Source:

www.released.so). Dans cette optique, llms.txt devient un pendant au débat en cours sur les données d'entraînement de l'IA et le droit d'auteur. Comme le note Search Engine Land, les créateurs de contenu y voient « une certaine assurance d'un contrôle accru par le propriétaire, en termes de ce qui doit être accédé et dans quelle mesure » (Source: searchengineland.com).

2. Amélioration de la qualité des réponses de l'IA : Lorsqu'un LLM a un accès direct à une base de connaissances concise, la qualité de sa génération s'améliore. Si un assistant IA répond à des questions sur votre site ou votre domaine, vous voulez qu'il dispose de sources faisant autorité. L'analyse de HTML brut peut entraîner des « hallucinations » ou des omissions sans contexte. En revanche, un fichier `llms.txt` bien conçu résume les faits clés et lie des informations à jour. Des praticiens ont rapporté qu'après avoir alimenté un LLM avec le contenu listé dans `llms.txt`, l'IA fournit des réponses plus précises et pertinentes sur le site. Par exemple, un praticien a testé un fichier `llms.txt` pour une entreprise appelée Enhance Media en utilisant trois modèles (ChatGPT, Gemini, Claude) et a constaté que *les trois étaient capables de résumer correctement l'entreprise à partir de ce seul fichier* (Source: www.linkedin.com). Le format structuré du fichier a aidé les modèles à se concentrer rapidement sur les points saillants. De même, les créateurs de FastHTML ont découvert qu'un contexte soigneusement organisé (via un fichier `llms.txt` étendu) produisait « des résultats considérablement meilleurs » de Claude et d'autres outils que le scraping non ciblé (Source: news.ycombinator.com).

3. Efficacité technique : Les crawlers à grande échelle (surtout pour les modèles d'IA plus petits) sont gourmands en ressources. Les entreprises de LLM doivent équilibrer la fréquence à laquelle elles doivent ré-explorer les sites pour obtenir des données fraîches. Une offre `llms.txt` peut servir de **balise de fraîcheur** : elle peut permettre à un crawler d'IA de vérifier un seul fichier pour les mises à jour plutôt que d'explorer l'ensemble du site. En effet, comme indiqué dans [33], au moins un système OpenAI interrogeait les `llms.txt` des développeurs toutes les 15 minutes pour les mises à jour. Ce type de flux de travail rationalisé peut réduire la charge inutile à la fois sur l'IA et sur les serveurs web. Il peut également garantir que la version du contenu à laquelle l'IA est exposée est la version officielle et aplatie fournie par le site — et non un extrait partiel ou obsolète. En effet, `llms.txt` pourrait servir de sorte d'« API » pour le contenu de site statique, bien que sans la structure formelle d'un appel API.

4. Égalisation des chances : Les petits sites et les nouvelles startups peuvent voir llms.txt comme un moyen de concourir pour l'attention dans la recherche basée sur l'IA. Certains analystes ont établi un parallèle avec les premières stratégies SEO : aux débuts du web, les petites entreprises utilisaient `robots.txt`, les balises Meta et les sitemaps pour se démarquer auprès des moteurs de recherche. Aujourd'hui, si les agents IA deviennent de nouveaux « curateurs » de contenu, n'importe quel site peut utiliser `llms.txt` pour se démarquer auprès d'eux aussi. Cet aspect démocratisant est explicitement mentionné dans les supports promotionnels : en ajoutant `llms.txt` et même en le partageant sur des plateformes comme GitHub, *« vous façonnerez la manière dont l'IA traite votre contenu »* (Source: llmsly.com). L'idée est que les sites web avant-gardistes pourraient acquérir un avantage réputationnel en étant les premiers à s'associer à l'IA.

5. Précédent des « robots » IA : Déjà, certains outils d'IA se présentent comme des agents qui explorent le web. Par exemple, Claude Projects (une intégration IDE) peut prendre des fichiers de documentation en contexte. De tels outils exigent souvent que les utilisateurs les dirigent vers des documents ou des données clés. `llms.txt` peut automatiser ce processus. En offrant un fichier d'ancrage bien connu, les propriétaires de sites peuvent s'inscrire automatiquement dans ces écosystèmes d'IA émergents. C'est similaire au rôle précoce de `robots.txt` : au début, peu de sites l'utilisaient, mais à mesure que Googlebot et d'autres ont appris à le vérifier, il est devenu standard. Les premiers à adopter `robots.txt` (vers 1994-95) l'ont fait pour guider les crawlers d'AltaVista ou de Google. Aujourd'hui, les concepteurs de `llms.txt` espèrent que les « architectes de l'IA » (certaines équipes d'IA de premier plan) feront de même. En effet, les créateurs soulignent souvent que les développeurs d'Anthropic promeuvent `llms.txt` sur leurs documents, et que des entreprises comme Mintlify ont intégré son support (Source: www.released.so). En somme, `llms.txt` est important pour ses défenseurs car **il répond directement à un goulot d'étranglement technique des systèmes d'IA actuels**. Il promet un moyen simple de rendre le Web plus « compatible LLM », ce qui pourrait faciliter le travail de l'IA et améliorer la qualité de ses réponses.

Adoption, réponse de l'industrie et études de cas

Dans quelle mesure `llms.txt` est-il utilisé en pratique, et qui y prête attention ? Depuis que l'idée a fait surface fin 2024, l'adoption a été limitée et inégale, mais certaines concentrations d'activité sont notables.

Premièrement, les **entreprises technologiques et les plateformes de documentation** ont manifesté leur intérêt. En novembre 2024, la plateforme de documentation Mintlify a annoncé la prise en charge intégrée de `llms.txt` pour les projets publiés sur son site (Source: www.released.so). Cela signifiait que, pratiquement du jour au lendemain, des *milliers* de

documentations de projets logiciels sont devenues accessibles via `llms.txt`. Le billet de blog de Jens Schumacher note : « D'un seul coup, ils ont rendu les documentations de milliers d'outils de développement compatibles avec les LLM, comme Anthropic et Cursor » (Source: www.released.so). Les projets d'outils de développement dont les documentations sont hébergées sur Mintlify (par exemple, de nombreuses bibliothèques open source) ont ainsi acquis des fichiers llms.txt sans action individuelle de la part des mainteneurs. De même, certaines entreprises technologiques créent explicitement des fichiers `llms.txt`. Dans [15], Radu Stoian affirme que **Anthropic** (l'entreprise derrière l'IA Claude) et d'autres entités non spécifiées ont publiquement demandé des fichiers llms.txt pour leurs sites : « Des leaders de l'IA comme Anthropic... l'ont initié... ils ont construit leurs modèles en s'attendant à trouver ce fichier » (Source: www.linkedin.com). Nous avons vérifié indépendamment que https://www.anthropic.com/llms.txt (ou le lien équivalent généré statiquement) existe bel et bien et répertorie des dizaines de pages sur le site d'Anthropic (Source: llmstxtgen.com).

Au-delà des développeurs, les **sociétés de conseil et les agences** ont commencé à recommander llms.txt. Par exemple, un auteur de blog orienté business le qualifie de « votre nouvelle arme secrète » pour l'optimisation de l'IA (Source: llmsly.com). D'autres sites web axés sur le SEO et articles LinkedIn saluent `llms.txt` comme « essentiel pour les marques » à l'ère de l'IA (Source: www.linkedin.com), lui conférant une grande visibilité dans les milieux du marketing. Un nombre important de petites entreprises et de fournisseurs de services (des agences SEO aux fournisseurs d'IA) ont publié des articles de blog sur la manière d'implémenter `llms.txt` sur les sites de leurs clients. Cet enthousiasme est en partie exploratoire — beaucoup considèrent le contenu IA comme la prochaine frontière de la visibilité, et traitent llms.txt comme une bonne pratique à tester.

Cependant, lorsque nous examinons l'*utilisation réelle*, le tableau est mitigé. Un répertoire collaboratif de fichiers llms.txt, [llmstxt.site], atteste de centaines de sites web où `llms.txt` a été détecté (Source: llmstxt.site). Ce répertoire liste des dizaines de sites exemples et leur nombre de tokens. Par exemple, l'outil de design populaire **Framer** possède un `llms.txt` d'environ 1 821 tokens (taille du texte) (Source: llmstxt.site). La société de fintech **Klarna** (dans son sous-domaine de documentation) a 17 387 tokens dans son `llms.txt` (Source: llmstxt.site). Même un site de contenu apparemment important, Weather.com (The Weather Company), est répertorié comme ayant un llms.txt (vide ?) (0 tokens) (Source: llmstxt.site), suggérant qu'il a peut-être créé le fichier mais l'a laissé vide. À plus petite échelle, de nombreux blogs personnels, éducatifs et technologiques ont implémenté `llms.txt`, parfois avec des milliers, voire des centaines de milliers de tokens. Par exemple, un blog d'astrologie « LookUpTheStars » signale un `llms.txt` avec environ 385 221 tokens (Source: llmstxt.site). À l'autre extrémité, certains fichiers `llms.txt` ne contiennent que quelques centaines de mots (par exemple, Ideanote.io avait 1 106 tokens) (Source: llmstxt.site). Notre étude du répertoire llmstxt.site révèle une adoption *expérimentale* généralisée : des entreprises de toutes tailles, des produits logiciels au commerce électronique de niche, ont créé ces fichiers (souvent en convertissant des sitemaps existants ou des listes de liens manuelles). Beaucoup semblent suivre précisément le format de la spécification, tandis que quelques-uns ont des implémentations incomplètes ou ascendantes (des exemples de conseils pour les parseurs sont disponibles sur les forums communautaires).

Pour avoir une idée plus large de l'adoption, deux analyses ont été rapportées par des tiers. L'une est un « Rapport sectoriel » d'un site appelé *LLMS Central*, qui affirme avoir analysé 2 147 sites web dans 15 secteurs d'activité début 2025 (Source: llmscentral.com). Leurs statistiques principales indiquent que **68% des sites « autorisent » l'entraînement de l'IA** (avec des politiques entièrement ouvertes ou sélectives), 23% « autorisent tout », 45% ont des « politiques sélectives », 18% bloquent tout, et seulement 14% n'ont pas de `llms.txt` du tout (Source: llmscentral.com). Ils interprètent cela comme signifiant qu'une majorité de sites publient des directives pour les LLM. Il est à noter que, dans leur échantillon d'entreprises technologiques et logicielles (n=387), ils rapportent que 95% ont une politique `llms.txt` explicite d'une sorte ou d'une autre (Source: llmscentral.com). Ces chiffres, cependant, doivent être pris avec prudence. Le rapport ne divulgue pas comment les sites ont été choisis ni s'ils ont simplement recherché toute mention de `llms.txt`. Il est possible que leur ensemble de données soit enrichi d'entreprises déjà impliquées dans l'IA/la technologie, ce qui fausse les pourcentages à la hausse.

En net contraste, une firme d'analyse SEO, *Rankability*, a publié un « Rapport mensuel sur l'adoption de LLMS.txt » axé sur les **1 000 premiers sites web commerciaux** en termes de trafic (Source: www.rankability.com). Ils ont constaté une adoption pratiquement **nulle** : un taux d'adoption de 0,3% (soit effectivement 3 sur 1000) (Source: www.rankability.com). Ils déclarent sans ambages « Zéro adoption actuelle » (Source: www.rankability.com), avec un scan automatisé étendu ne produisant presque aucun résultat positif. Par secteur, leurs données montrent un taux d'adoption de 0,00% dans l'e-commerce, les médias sociaux, la finance, la santé, les secteurs gouvernementaux, avec seulement 0,73% d'adoption dans le secteur de l'éducation (suggérant que peut-être 7 sur 1000 sont des universités ou des valeurs aberrantes similaires) (Source: www.rankability.com). En bref : parmi les plus grands sites du monde, pratiquement aucun n'implémente `llms.txt` à la mi-2025. Cela implique que la norme reste de niche.

Pourquoi une telle divergence ? Il semble que l'adoption se soit concentrée sur les **sites plus petits ou orientés technologie**, et pratiquement aucune parmi les grandes marques grand public. La liste des 500-1000 premiers comprend des géants mondiaux (Amazon, YouTube, etc.) avec des équipes SEO bien établies ; il est évident que cela n'a pas encore pénétré ces cercles. En comparaison, les sites de petite à moyenne taille, les bases de connaissances et les outils de développement s'y sont ralliés. Les données de Rankability suggèrent qu'un ou deux *cas isolés* sur 1000 ont été trouvés (probablement de petits sites qui se sont classés juste dans le top 1000). Pendant ce temps, le rapport LLMS Central a probablement échantillonné des entreprises au moins partiellement engagées dans les discussions sur l'IA, d'où ses chiffres d'adoption plus élevés. Cet écart entre la « communauté d'enthousiastes » et le « marché de masse » sera important pour évaluer l'impact réel que `llms.txt` peut avoir.

Compte tenu de ces chiffres, il est juste de dire que **`llms.txt` a une étincelle mais pas (encore) de flamme**. Il compte dans certains écosystèmes (notamment les documentations logicielles et les commentaires des agences SEO) mais pas largement sur le web. Cela dit, les tendances d'adoption pourraient s'accélérer si de grandes plateformes comme Google ou Bing de Microsoft décidaient de l'exploiter. Alternativement, il pourrait rester une optimisation optionnelle pour un sous-ensemble de propriétaires de sites. Ensuite, nous explorerons quelques exemples détaillés d'utilisation de `llms.txt`, ainsi que les réactions des développeurs d'outils d'IA.

Étude de cas : Documentation technique

Un cas d'utilisation précoce et logique est la documentation technique logicielle. Les documentations pour développeurs génèrent souvent déjà du contenu HTML à partir de balises (par exemple, Markdown) et s'efforcent généralement d'être lisibles par les machines et les humains. Elles bénéficient également grandement de réponses précises. La bibliothèque FastHTML discutée précédemment en est un exemple : ses développeurs ont créé des entrées `llms.txt` pour aider les IA orientées développeurs. Un autre exemple notable est la **documentation développeur de Klarna** (la société européenne de paiements). Selon le répertoire llmstxt, la documentation de Klarna (hébergée sur docs.klarna.com) inclut un `llms.txt` avec environ 17 387 tokens (Source: llmstxt.site).

De même, un projet GitHub « pgai/llms.txt » indique que le projet Postgres AI (Timescale) a ajouté un fichier `llms.txt` à son dépôt, suggérant une implémentation dans un produit de base de données réel (Source: github.com). Les API d'entreprise, les bibliothèques open source et les plateformes cloud (le répertoire liste des entrées pour AWS, les documentations Azure, etc.) ont également commencé à adopter le format. Ces utilisations sont logiques : les utilisateurs techniques sont susceptibles de bénéficier de résumés de documentation clairs et lisibles par l'IA.

Étude de cas : Sites de services et praticiens

Toute l'adoption ne se trouve pas dans la haute technologie. Par exemple, le répertoire SEO liste **HoodChefs** (un service de location de cuisines) avec 44 494 tokens (Source: llmstxt.site), et un site web de **concessionnaire automobile** « AutoChampion24 » en Allemagne avec 6 750 tokens (Source: llmstxt.site). Ces entrées montrent que même les petites entreprises y voient un potentiel. « GalaxxiaMarketing » (une firme de marketing brésilienne) a 676 tokens (Source: llmstxt.site), présentant apparemment ses services via llms.txt. Des sites religieux et spirituels, des blogs personnels et des fournisseurs d'e-learning ont également été repérés. L'existence d'un site comme « lookupthestars.com » avec 385k tokens (Source: llmstxt.site) est notable : il semble s'agir d'un site sur le thème de l'astrologie qui a pleinement adopté la norme.

Il est difficile de vérifier la motivation commerciale de chaque `llms.txt` ad hoc, mais beaucoup l'ont probablement fait par curiosité ou pour expérimenter avec le SEO. Les contributions communautaires aux répertoires llmstxt suggèrent que des plugins WordPress ont été créés pour générer automatiquement des fichiers llms.txt, et des développeurs sur les forums mentionnent des moments où leurs bots de tutorat IA ont pour la première fois pris en charge llms-txt.

Soutiens de l'industrie

Certains acteurs majeurs ont au moins **reconnu** le concept. Le blog de Cloudflare (mai 2025) discute de la manière dont leurs nouveaux services AI Gateway s'intègrent avec d'autres outils d'IA, bien qu'il ne mentionne pas directement llms.txt (Source: www.cloudflare.net). Plus pertinent est Anthropic : leur portail de documentation inclut désormais un lien visible vers le fichier « LLMS.txt », et ils ont « posté sur X » à propos de leur soutien (Source: www.released.so). En bref, les **entreprises orientées IA**

sont au moins curieuses. En revanche, les grandes entreprises technologiques ou médiatiques sont restées silencieuses. Nous n'avons connaissance d'aucun rapport d'adoption de llms.txt par Google, Amazon (au-delà de ceux figurant dans le répertoire public) ou Facebook.

Métriques et analyses

Peu de données existent sur l'*efficacité* de llms.txt. Une métrique approximative provient d'un auteur LinkedIn qui a examiné les analyses de Google Search Console. Il a affirmé que Google avait déjà indexé un fichier llms.txt d'un site de test (Source: www.linkedin.com), bien que Google déclare ne pas les *utiliser*. Une autre trace citée est celle des journaux de serveur : un webmaster a remarqué que les crawlers d'*OpenAI* interrogeaient les fichiers llms.txt de ses sites toutes les 15 minutes pour vérifier leur fraîcheur (Source: searchengineland.com). Cette anecdote suggère qu'au moins certains services avancés de recherche/IA y prêtent attention.

D'autres métriques pourraient inclure les changements dans les réponses aux requêtes ou le trafic de référence. Au moment de la rédaction de cet article, ces données ne sont pour la plupart pas publiques. En théorie, on pourrait suivre le trafic provenant des interfaces de chat IA (via des balises UTM spéciales ou des « références » d'API), mais peu de propriétaires de sites disposent d'un tel suivi. Certains articles SEO suggèrent d'utiliser des API personnalisées pour surveiller le trafic généré par les LLM, mais les exemples concrets sont rares (le guide golevels.com en discute de manière conceptuelle). Des signes précoces dans les résultats de recherche pourraient également indiquer une utilisation. Un post LinkedIn d'un consultant SEO a montré des résultats de recherche Google mettant en évidence un fichier `llms-full.txt` dans les résultats, suggérant une indexation (Source: distinctly.co), mais il n'est pas clair si cela est officiel ou un bug.

Adoption par région ou secteur

Les données de Rankability montrent que l'**éducation** est le seul secteur avec une présence mesurable (0,73%) parmi les meilleurs sites (Source: www.rankability.com). Cela pourrait être dû à des universités ou des projets universitaires expérimentant le format. En revanche, des secteurs comme l'e-commerce, les médias sociaux, la finance, la santé et le gouvernement affichaient 0% dans le top 1000 (Source: www.rankability.com). Le rapport LLMS Central (bien que moins faisant autorité) indique que les entreprises de **technologie/logiciels** sont leaders en matière d'adoption, avec « 95% ayant des politiques explicites » au sein de ce segment (Source: llmscentral.com). Cela correspond à l'intuition : les éditeurs de technologie sont les premiers bancs d'essai de la technologie de l'IA.

Critiques, préoccupations et perspectives alternatives

Pour l'équilibre, nous devons aborder les raisons pour lesquelles `/llms.txt` pourrait *ne pas* prendre ou être problématique. Plusieurs critiques ont émergé de la part de développeurs, de spécialistes SEO et de sceptiques. Nous les organisons ici :

A. Duplication des efforts et préoccupations concernant l'expérience utilisateur (UX) : Les critiques observent que si un site est déjà bien structuré et possède des pages « aide » ou « à propos », l'ajout de llms.txt peut être redondant. Comme l'a souligné une discussion sur Hacker News : *« Ce n'est pas une bonne UX pour les machines. C'est un correctif pour une mauvaise UX afin d'aider les LLM... Certains sites web ont le même correctif pour les humains sous la forme d'une section 'Aide' ou 'À propos' »* (Source: news.ycombinator.com). En d'autres termes, idéalement, un site bien conçu devrait déjà rendre les informations essentielles accessibles, et un lecteur (humain ou bot) devrait les trouver naturellement. Si le contenu réel du site était plus simple ou plus textuel (par exemple, via un « mode lecture »), une IA pourrait ne pas avoir besoin de llms.txt. Cette critique revient essentiellement à dire : « Réparez le site web, ne masquez pas ses défauts. » Elle avertit également que llms.txt est une sorte de raccourci qui pourrait décourager l'amélioration de la conception sous-jacente du site (comme entasser du contenu dans un bloc de citation SEO plutôt que de concevoir une interface utilisable).

B. Portée limitée (entraînement vs. inférence) : Il est important de clarifier que llms.txt affecte principalement l'utilisation des sites web par l'IA au moment de l'*inférence*, et non l'entraînement initial du modèle. De nombreux propriétaires de contenu souhaitent contrôler la manière dont leur contenu est utilisé pour entraîner de nouveaux modèles (un débat juridique et éthique), mais llms.txt tel que spécifié n'applique ni n'enregistre directement les autorisations d'entraînement. Il aide simplement un LLM à *récupérer* du contenu pour répondre aux requêtes. Comme le soutient Search Engine Land, les différences clés tournent autour de l'indexation vs. l'utilisation : *« Robots.txt concerne la gestion de l'exploration tandis que la discussion sur le droit d'auteur concerne la manière dont les données sont utilisées »* (Source: searchengineland.com). Les critiques pourraient dire : si une entreprise ne

veut pas du tout que son site apparaisse dans les sorties d'IA, llms.txt n'arrête personne (il ne fait que guider). Inversement, si l'entreprise licencie déjà explicitement son contenu (par exemple, avec Creative Commons), llms.txt n'ajoute que peu de choses. L'article GEO de Konstantinos Zoulas de 2023 suggère que les licences Creative Commons (CC0, CC-BY, etc.) pourraient régir l'utilisation de l'IA plus directement que les directives robots ou llms (Source: searchengineland.com). Cette vision implique que llms.txt ne résout que le symptôme (la découverte de données) et non le cœur du problème des droits de contenu.

C. Manque de standardisation et d'application : Actuellement, `/llms.txt` est une proposition volontaire sans RFC ni registre formel. Comme Jeremy Howard l'a lui-même admis sur Hacker News, il n'a pas été enregistré sous le registre URI .well-known de l'IANA (une étape requise pour le statut de norme officielle) (Source: news.ycombinator.com). Sans décision formelle ou approbation à l'échelle de l'industrie, il n'y a aucune garantie que les logiciels le rechercheront de manière fiable. Les critiques soulignent que même robots.txt n'est pas strictement appliqué — c'est une convention — et Google a montré qu'il peut ignorer « robots.txt » si nécessaire pour des raisons légales. Avec llms.txt encore plus en évolution, certains soutiennent qu'il pourrait s'essouffler si les acteurs clés restent en marge. (La position de Google de l'ignorer a peut-être déjà freiné l'enthousiasme.)

D. Potentiel de mauvaise utilisation ou de manipulation : Comme pour tout signal de type SEO, on pourrait s'inquiéter du spam ou de la « manipulation » de llms.txt. En principe, un site malveillant pourrait créer un llms.txt contenant des liens trompeurs ou malveillants, ou dissimuler des URL de traqueurs ou de publicités. Cependant, comme llms.txt n'injecte pas automatiquement de contenu dans les données d'entraînement de l'IA, ce risque est limité. Il s'agit plutôt du risque qu'un site peu scrupuleux puisse remplir son llms.txt de liens non pertinents juste pour pousser les utilisateurs (via les réponses de l'IA) vers eux. La spécification actuelle ne précise aucune validation ou limitation de débit. Comment un outil d'IA saurait-il si un llms.txt est légitime ? C'est une question non résolue. En pratique, étant donné que le format est lisible par l'homme et vraisemblablement curaté, les abus flagrants seraient probablement repérés et discrédités par la communauté avant de proliférer.

E. Impact sur les performances des sites web : Une autre préoccupation (principalement hypothétique) est de savoir si l'exploration et la diffusion de ces fichiers texte potentiellement volumineux pourraient surcharger les serveurs web. Comme noté, certains fichiers `llms.txt` atteignent des centaines de kilo-octets, voire des méga-octets, comparables à une petite page HTML. Si un système d'IA les interroge fréquemment (toutes les 15 minutes, comme l'indiquait un journal (Source: searchengineland.com), cela pourrait imposer une charge non négligeable. Les opérateurs de sites devraient en être conscients — bien que ce problème soit parallèle au concept préexistant d'interrogation de « sitemap.xml ». Les serveurs pourraient toujours mettre en cache et limiter le débit ; c'est un détail technique, mais qui doit être mis en œuvre par les administrateurs web si `llms.txt` gagne en popularité.

F. Confusion autour des noms et du versionnement : Il existe une certaine ambiguïté terminologique : la proposition originale utilise « `llms.txt` », mais de nombreux articles (et articles LinkedIn) l'écrivent comme « LLMS.txt » (avec des LLM en majuscules ou au pluriel). La communauté s'est généralement accordée sur « llms.txt » (nom de fichier en minuscules). De plus, différents outils parlent de `llms-full.txt` (qui contient le texte complet concaténé des pages) par opposition à `llms.txt` (qui liste les liens). Cela peut dérouter les nouveaux venus. La standardisation ou la dénomination pourrait évoluer, mais pour l'instant, cette confusion pourrait freiner l'adoption occasionnelle.

G. Approches alternatives (pas de nouveau fichier) : Enfin, la critique la plus fondamentale : *Avons-nous même besoin d'un nouveau fichier ?* Certains experts en SEO soutiennent que les mêmes objectifs pourraient être atteints en revitalisant des idées plus anciennes. Par exemple, les premières discussions d'OpenAI mentionnaient l'utilisation de « noindex » ou « nofollow » dans les fichiers robots pour différencier la recherche régulière de l'utilisation par l'IA (Source: searchengineland.com). D'autres proposent des signaux entièrement intégrés : par exemple, Google (mi-2023) a suggéré d'utiliser simplement des liens normaux et des pratiques SEO afin que l'IA (comme les propres Aperçus de Google) trouve naturellement le contenu (Source: searchengineland.com). Il existe également le concept d'un en-tête HTTP ou d'un élément qui identifie un fichier ou un format pour les LLM, plutôt qu'un fichier texte brut. Certains commentateurs affirment que cela serait plus sémantiquement « web-compatible » que d'inventer encore un autre type de fichier. Les partisans de llms.txt répondent généralement que rien n'empêche d'utiliser plusieurs approches (en-tête *et* llms.txt), mais cela reste un sujet de discussion.

En somme, les critiques se concentrent sur la **praticité et la nécessité** : Si Google (et Bing) obtiennent tout le contenu via d'anciennes méthodes, llms.txt pourrait être superflu. Si les développeurs d'IA pouvaient simplement mieux extraire le HTML ou utiliser des embeddings à partir d'index de recherche existants, ils n'en auraient peut-être pas strictement besoin. Dans le même temps, les partisans soulignent que ces problèmes n'ont pas découragé les expériences initiales ou la formation de normes. Que ces préoccupations se révèlent fatales ou surmontables dépendra probablement de l'utilisation concrète et de l'élan de la communauté.

Données et analyses

Une analyse approfondie de `/llms.txt` nécessite non seulement des descriptions, mais aussi des informations basées sur des données. Cependant, à la mi-2025, l'écosystème est encore naissant. Nous résumons ci-dessous les données disponibles et les résultats quantitatifs :

- **Adoption dans les classements de trafic web :** L'étude Rankability est l'une des rares analyses d'adoption rapportées publiquement. Elle a sondé les 1 000 sites web les plus visités (mondialement) à la mi-2025 et a constaté **0 %** d'utilisation de `llms.txt` (seulement ~0,3 % selon un décompte, arrondi à 0 %) (Source: www.rankability.com). En ventilant par secteur, elle a rapporté 0,00 % d'adoption dans chaque catégorie industrielle majeure (e-commerce, médias sociaux, finance, etc.), à l'exception d'un léger pic de 0,73 % dans l'Éducation (Source: www.rankability.com). Cela suggère que, parmi les poids lourds du Web, pratiquement aucun n'a implémenté `llms.txt`. En termes pratiques, si vous recherchez un grand site (par exemple, Wikipédia, CNN, Amazon) sur Google, vous ne trouverez pas de `llms.txt` à moins que quelqu'un n'en ait explicitement mis un en place juste pour tester. (Il est à noter que la définition de l'« adoption » par Rankability exigeait probablement une réponse HTTP 200 pour /llms.txt. Certains sites renvoyant une erreur 404 ou une autre erreur seraient considérés comme n'ayant pas adopté.)

- **Adoption parmi les sites sondés :** En revanche, une autre analyse portant sur un ensemble plus large de 2 147 sites web (le rapport « LLMS Central ») a affirmé que 86 % des sites avaient **du contenu llms.txt** (68 % autorisant l'entraînement de l'IA entièrement ou sélectivement, et seulement 14 % n'en ayant pas) (Source: llmscentral.com). Leur méthodologie n'est pas entièrement transparente, mais ils ont regroupé les politiques des sites en « Tout autoriser », « Sélectif », « Tout bloquer » ou « Pas de fichier ». Voir une catégorie comme « Tout autoriser » (23 %) implique que ces sites ont un fichier llms.txt déclarant explicitement autoriser l'utilisation de l'IA. Si l'on prend ce rapport au pied de la lettre, il suggère que **plus des deux tiers des sites de taille moyenne** de leur échantillon ont publié un fichier llms.txt. Il constate également que les entreprises technologiques sont particulièrement enthousiastes : 95 % des entreprises de technologie/logiciels qu'ils ont sondées avaient un fichier llms (Source: llmscentral.com), contre des pourcentages plus faibles dans d'autres industries. Cependant, sans connaître leur sélection d'échantillon, cela peut refléter un biais d'auto-sélection (peut-être ont-ils exploré des sites qui mentionnaient déjà l'IA sur leurs blogs).

- **Tailles et contenu des fichiers :** En examinant le contenu réel des fichiers llms.txt, nous constatons une énorme variation. L'exemple du Tableau 2 ci-dessous présente des décomptes de tokens représentatifs pour quelques sites (issus du répertoire llmstxt.site). Ces chiffres donnent une idée de l'échelle. Notamment, certains sites de documentation technique génèrent d'énormes fichiers llms : par exemple, *M-Source* (une entreprise de bases de données) a **328 716 tokens** listés (Source: llmstxt.site), et *LookupTheStars* a **385 221 tokens** (Source: llmstxt.site). (Pour le contexte, la limite de contexte de GPT-4 est d'environ 32k tokens, donc un seul fichier llms.txt de 300k tokens devrait être découpé.) D'autres sont plus légers en tokens : le fichier llms.txt d'Ideanote.io fait 1 106 tokens (llmstxt.site, HoodChefs 44 494 tokens (llmstxt.site, Framer 1 821, Klarna 17 387, etc. Un cas extrême est *X-CMD*, dont le fichier llms-full fait 590 515 tokens (Source: llmstxt.site) (impliquant un site colossal ou peut-être une particularité de sa génération). La variabilité indique que les sites interprètent différemment la quantité d'informations à inclure.

- **Insights sur l'exploration et le trafic :** Il y a peu de données publiques sur le trafic. Un tableau du site de reporting SEO [33] souligne que les requêtes de Googlebot pour llms.txt se produisent *zéro* fois (« Google n'explorera pas votre fichier LLMS.txt » (Source: searchengineland.com). En revanche, l'utilisateur Ray Martinez a rapporté dans les journaux de son site qu'« OpenAI explore mon fichier LLMs.txt sur quelques sites... interrogeant nos serveurs toutes les 15 minutes à la recherche de nouveautés » (Source: searchengineland.com). Cette analyse des journaux suggère que, du moins pour ses sites, **les systèmes d'OpenAI vérifient activement et souvent llms.txt** (supposant peut-être qu'ils le devraient). John Mueller de Google a également déclaré lors d'un précédent hangout de la Search Console qu'« aucun système d'IA n'utilise actuellement le fichier LLMS.txt » (Source: searchengineland.com) (citation de seroundtable). En somme, la seule information empirique dont nous disposons est anecdotique : la recherche Google l'ignore, certains laboratoires d'IA l'interrogent.

- **Corrélation avec les performances SEO :** Aucune donnée agrégée crédible n'existe liant llms.txt à une amélioration du classement de recherche ou du trafic. Google déclare explicitement que le SEO normal est suffisant (Source: searchengineland.com), ce qui implique qu'ils n'ont trouvé aucun avantage. Il reste à voir si, par exemple, l'inclusion de llms.txt

affectera positivement les extraits ou les « réponses » dans les interfaces de chat IA. En principe, si un assistant IA cite directement le contenu de llms.txt, un marketeur avisé tentera de le détecter et d'optimiser en conséquence. Mais à la mi-2025, cela reste hypothétique.

- **Support des outils LLM :** Au-delà de Google, des produits LLM notables ont commencé à reconnaître llms.txt. La documentation d'*Anthropic* (Claude) l'inclut ; le MCP (plugin multi-contexte) de *LangChain* prend en charge la lecture de llms.txt depuis les IDE (Source: github.com). Certains frameworks de chatbots basés sur des LLM open-source incluent désormais du code passe-partout pour rechercher llms.txt. L'existence même d'un dépôt GitHub (AnswerDotAI/llms-txt) et de tests CI automatisés indique un intérêt des développeurs. D'autre part, les plateformes majeures comme ChatGPT (interface d'OpenAI) n'ont pas annoncé de support formel, mis à part l'indexation en arrière-plan. Des rapports d'analystes de Distinctly (actualités SEO) ont noté une capture d'écran de ChatGPT extrayant du contenu d'un « llms-full.txt » (Source: distinctly.co), mais les détails manquent et cela pourrait être un cas isolé.

Ces données brossent le tableau : **émergent mais mineur**. Des dizaines ou des centaines de sites plus petits ont un fichier llms.txt, mais pas de masse critique. Si l'adoption était cartographiée au fil du temps, nous pourrions observer une lente augmentation parmi les sites de niveau intermédiaire fin 2024 et en 2025, avant de plafonner. Un point de bascule nécessiterait probablement qu'une ou plusieurs plateformes d'IA dominantes déclarent « oui, nous utilisons llms.txt ». Autrement, cela pourrait rester une meilleure pratique de niche.

Voici un tableau récapitulatif de quelques statistiques et exemples d'adoption :

MÉTRIQUE / CATÉGORIE DE SITE	VALEUR / EXEMPLES	SOURCE
Top 1000 des sites utilisant llms.txt	~0% (0,3%)	(Source: www.rankability.com)
Entreprises technologiques/logiciels (sondées)	95% (les sites de ces catégories ont des politiques llms dans un rapport)	(Source: llmscentral.com) (Source: llmscentral.com)
Tout autoriser (tout le contenu est ouvert)	23% des sites (selon un rapport)	(Source: llmscentral.com)
Politiques sélectives (certaines pages)	45% des sites	(Source: llmscentral.com)
Tout bloquer (aucune utilisation par l'IA autorisée)	18% des sites	(Source: llmscentral.com)
Pas de fichier `llms.txt`	14% des sites	(Source: llmscentral.com)
Exemples de sites avec llms.txt	Framer.com (1 821 tokens), Klarna docs (17 387), M-Source (328 716) (Source: llmstxt.site) (Source: llmstxt.site)	(Source: llmstxt.site) (Source: llmstxt.site)
Plus grande taille de llms rapportée	~385 221 tokens (lookupthestars.com)	(Source: llmstxt.site)
Fréquence d'exploration d'OpenAI	~toutes les 15 minutes (journal de site)	(Source: searchengineland.com)
Requêtes Googlebot pour llms.txt	Aucune rapportée ; Google dit qu'il n'explorera pas llms.txt	(Source: searchengineland.com)

Tableau : Chiffres sélectionnés relatifs à l'adoption et à l'utilisation de llms.txt.

Perspectives et avis d'experts

Pour saisir pleinement les enjeux de `/llms.txt`, nous examinons ce que divers experts et parties prenantes ont dit — parfois haut et fort — à propos de la proposition.

- **Jeremy Howard (Answer.AI, fast.ai) :** *Partisan et proposant de llms.txt.* Il argumente principalement du point de vue de la facilité d'utilisation pour les développeurs. Dans les fils de discussion, Howard a souligné que l'objectif est d'aider « les utilisateurs finaux à utiliser les informations des sites web avec l'aide de l'IA » (Source: news.ycombinator.com). Il a donné des exemples concrets : lorsqu'il a publié la bibliothèque FastHTML, de nombreux utilisateurs potentiels se sont plaints que les outils d'IA (cursor, etc.) ne pouvaient pas répondre aux questions à son sujet car les modèles étaient postérieurs à leurs connaissances. Sa solution : organiser manuellement la documentation une fois dans un fichier llms.txt afin que les outils d'IA l'aient facilement à disposition au moment de l'inférence. Howard présente llms.txt comme une aide pour l'utilisateur final/la communauté plutôt que comme une préoccupation de scraping : « llms.txt n'est pas vraiment conçu pour aider au scraping ; il est conçu pour aider les utilisateurs finaux à utiliser les informations des sites web avec l'aide de l'IA » (Source: news.ycombinator.com). Il souligne également que fournir un fichier llms.txt économise des efforts à *tout le monde* : au lieu que chaque ingénieur choisisse individuellement le contexte pour les prompts, le propriétaire du site le fait une seule fois. Dans les interviews et les articles de blog, il mentionne fréquemment les cas d'utilisation pour la documentation des développeurs, et le fait que de nombreuses documentations fast.ai/nbdev génèrent désormais automatiquement du markdown pour satisfaire ce besoin (Source: github.com).

- **Analystes SEO/Marketing (par ex. SearchEngineLand, Agences SEO attendues) :** De manière générale, les publications SEO ont adopté une vision prudemment optimiste. L'article de Roger Montti de mars 2025 dans SEL a examiné llms.txt et a noté à la fois des « créateurs de contenu intéressés » et des « détracteurs » (Source: searchengineland.com). La position de Montti est neutre à curieuse ; il explique la spécification et suggère qu'elle « augmente le contrôle du propriétaire ». Roger souligne l'angle de l'économie de ressources (les LLM se concentrent sur l'intelligence, pas sur l'exploration) . Pendant ce temps, d'autres acteurs de la communauté SEO présentent llms.txt comme un incontournable pour les marques. Par exemple, l'article LinkedIn de Radu Stoian le titre sans détour « non-négociable pour votre marque » (Source: www.linkedin.com). De tels articles promettent une amélioration de la narration de marque et affirment même que Google indexe désormais llms.txt. Cependant, s'agissant d'un blog non vérifié, ces informations doivent être lues avec scepticisme. Des voix plus mesurées (en dehors de SEL) suggèrent que llms.txt est une technique incrémentale de « SEO pour l'IA » : une optimisation possible mais peu susceptible de supplanter le SEO traditionnel (Source: llms-txt.io).

- **Ingénieurs de recherche Google :** Les déclarations les plus claires sont venues de Google lui-même, bien qu'indirectement. Lors d'un événement Google Search Central en juillet 2025, Gary Illyes (analyste de recherche) l'a explicitement déclaré : « *Pour que votre contenu apparaisse dans l'aperçu de l'IA, utilisez simplement les pratiques SEO normales... Il a également dit que Google n'explorera pas le fichier LLMS.txt. »* (Source: searchengineland.com). En effet, le message de Google est : *Ignorez llms.txt en termes de classement de recherche – nous ne l'utilisons pas.* Cela a été repris par John Mueller, qui a déclaré lors d'un Webmaster Hangout qu'« aucun système d'IA n'utilise actuellement le fichier LLMS.txt » (Source: searchengineland.com). Ces affirmations signifient que, du point de vue de Google, llms.txt n'a aucune incidence sur le SEO. Cela peut décourager les éditeurs qui se soucient principalement de la « Googleabilité ». Cela soulève également une question plus vaste : même si llms.txt est bénéfique pour la rencontre de votre contenu avec *certaines* IA, si cette IA n'est pas celle qui domine les recherches (Google Search), l'impact sur le trafic réel pourrait être minime.

- **OpenAI (développeurs de ChatGPT) :** OpenAI n'a pas commenté publiquement llms.txt, mais des preuves limitées suggèrent qu'ils l'ont au moins *testé* ou autorisé son utilisation. L'analyse des journaux par Ray Martinez est une preuve irréfutable qu'une partie de l'infrastructure d'OpenAI interroge llms.txt pour les changements (Source: searchengineland.com). Cela suggère que les agents d'OpenAI ont reconnu llms.txt dans la nature et le traitent comme un « point d'extrémité de fraîcheur ». Cependant, les porte-parole d'OpenAI n'ont annoncé aucune position officielle. De manière anecdotique, les utilisateurs d'outils comme le plugin « Naviguer avec Bing » de ChatGPT ou d'agents tiers essaient d'exploiter llms.txt, mais aucune documentation officielle n'est disponible.

- **Anthropic (développeurs de Claude) :** Anthropic est largement considéré comme un partisan de llms.txt. Leur équipe de documentation l'a ajouté très tôt, et les ingénieurs d'Anthropic ont manifesté leur intérêt pour la standardisation. Claude Projects (le plugin IDE de code de Claude) traite llms.txt comme un citoyen de première classe : les utilisateurs chargeant une base de code peuvent spécifier un llms.txt. Un extrait de la communauté sur GitHub montre des instructions pour configurer Claude Desktop/Cursor afin de lire llms.txt (Source: gist.github.com), ce qui implique un support intégré. Kohl Marcus (dans Distinctly news) a mentionné qu'« Aimee Jurenka montre ChatGPT accédant au contenu d'un fichier llms-full.txt » (Source: distinctly.co), donc on peut supposer que cela s'est fait via les frameworks d'Anthropic. Tout cela indique qu'au moins les produits d'IA avancés (comme Claude) prennent llms.txt au sérieux.

- **Experts universitaires et en confidentialité :** Les organisations préoccupées par la confidentialité des données notent que llms.txt touche au récit du scraping. Privacy International, dans un explicatif sur les LLM, souligne que « plus les LLM peuvent obtenir de langage écrit, mieux c'est » et que le scraping web est souvent « indiscriminé » (Source: privacyinternational.org). Bien qu'ils ne mentionnent pas spécifiquement llms.txt, l'implication est que tout ce qui rend le scraping plus ciblé (c'est-à-dire guidé par les propriétaires) pourrait s'aligner sur la gouvernance des données. Aucune loi formelle sur la confidentialité ne reconnaît llms.txt, mais des défenseurs comme Jay Graber (PDG de Bluesky) qui mènent les débats sur les droits des créateurs d'IA ont souligné que llms.txt et d'autres initiatives (comme la « Déclaration de Bletchley ») font partie des normes émergentes pour le contrôle des données dans l'IA. En bref, certains voient llms.txt comme un geste constructif envers le respect de la propriété du contenu, même s'il n'est pas contraignant.

- **Critiques et pragmatiques :** De nombreux codeurs et experts SEO ont abordé llms.txt de manière pragmatique. Sur Hacker News et les blogs, les commentateurs ont exprimé leur scepticisme : l'un a noté que si l'expérience utilisateur d'un site est bonne, une « page d'instructions » pourrait suffire et llms.txt serait inutile (Source: news.ycombinator.com). D'autres ont déclaré que la maintenance d'un fichier supplémentaire est une surcharge ; ils préféreraient s'appuyer sur `rel=search` ou des approches basées sur des API. Du point de vue des standards, un commentateur a souligné qu'une balise `<link rel="llm">` ou une négociation de type de contenu HTTP pourrait être plus élégante qu'un fichier texte (Source: news.ycombinator.com). Ces suggestions reflètent un désir de solutions qui s'intègrent en douceur dans l'architecture Web existante, plutôt que d'ajouter un silo parallèle.

Malgré ces opinions mitigées, **le fil conducteur** est le suivant : `llms.txt` force la question « Le Web doit-il s'adapter à l'IA ? ». De nombreuses voix interrogées se targuent d'être des adopteurs précoces. Les partisans soutiennent que cela permet aux sites web de *participer à la conversation* plutôt que d'être des mines de données passives (Source: www.linkedin.com), tandis que les détracteurs affirment que cela perturbe l'interface uniforme du Web. En fin de compte, la plupart y voient une *expérience* : une idée qui mérite d'être testée maintenant, les retours de la communauté guidant si elle devient une norme de facto ou s'estompe.

Considérations et outils d'implémentation

Pour un propriétaire de site web envisageant d'ajouter `llms.txt`, des questions pratiques se posent : *Comment le créer ? Quel contenu inclure ? Comment le maintenir ?* Heureusement, plusieurs outils et guides ont émergé pour y répondre.

- **Guides et exemples :** Le site communautaire llmstxt (llmstxt.org) propose des exemples de fichiers llms.txt et un guide étape par étape. Il existe également de nombreux articles de blog et dépôts GitHub avec des implémentations llms.txt exemples. Les conseils clés incluent : commencer par la page d'accueil/le titre, rédiger un résumé succinct (environ 1 à 3 phrases) dans un bloc de citation, puis lister les pages cruciales. Certains blogs SEO recommandent d'ajouter des informations sur l'entreprise (contact, adresse), des FAQ, des documents pour développeurs, des pages produits – en gros, tout ce dont une IA utile pourrait avoir besoin pour répondre aux requêtes des utilisateurs (Source: llmsly.com) (Source: golevels.com). Il est souvent suggéré de garder le fichier sous quelques mégaoctets ; un article a mentionné que les fichiers llms.txt peuvent varier de quelques Ko à des centaines de Ko (Source: searchengineland.com). Le format est flexible : vous pouvez utiliser des images (sous forme de liens), des listes à puces ou de courts paragraphes. Certains sites divisent même le contenu llms en plusieurs fichiers : la variante *llms-full.txt* peut contenir des sections entières de texte si nécessaire.

- **Outils existants :** Plusieurs outils open-source aident à générer ou à valider llms.txt. Par exemple :

  - **llms.txt Generator (llmstxtgen.com) :** Une application web où vous collez votre sitemap ou votre liste d'URL ; elle explore et génère un brouillon de llms.txt en quelques secondes. La capture d'écran [10] montre la sortie auto-générée d'un outil (pour anthropic.com).

  - **Utilitaires CLI :** Le dépôt GitHub (AnswerDotAI/llms-txt) inclut des scripts comme `llms_txt2ctx` qui peuvent combiner llms.txt et le markdown lié dans un fichier de contexte consommable par machine (Source: github.com). D'autres (comme l'outil de Firecrawl référencé dans [66]) peuvent explorer et assembler du contenu en listes de balisage.
  - **Plugins CMS :** Il existe des plugins pour WordPress et d'autres CMS qui génèrent llms.txt à partir des menus ou des articles du site (comme suggéré par [59]). Ceux-ci permettent des mises à jour dynamiques lorsque le contenu du site change.
  - **Intégrations IDE/LLM :** Des outils comme `mcpdoc` de LangChain peuvent extraire automatiquement un llms.txt lors de la configuration de l'IA, de sorte que les développeurs n'ont pas à le récupérer manuellement (Source: github.com). Cela montre que les frameworks LLM commencent à reconnaître le fichier.
- **Maintenance :** Étant donné que les sites changent, llms.txt a besoin de mises à jour. Contrairement à sitemap.xml (qui peut être automatisé), llms.txt est plus manuellement géré. Cependant, certains flux de travail le créent à partir de données de site existantes : par exemple, un script peut scanner les menus de navigation pour lister les URL, ou compiler des fichiers README. Le projet de documentation Ethereum, par exemple, utilise un processus CI pour reconstruire llms.md chaque fois que les documents changent (dans le cadre de sa génération de site statique). De manière générale, il est recommandé de revoir llms.txt chaque fois que le contenu majeur du site change, car des liens ou des résumés obsolètes pourraient induire l'IA en erreur. La surveillance implique simplement de vérifier la disponibilité de ce seul fichier (par exemple, les vérifications de l'état du site).

- **Hébergement et performances :** Comme pour tout actif statique, la meilleure pratique consiste à servir llms.txt avec la mise en cache activée (cache-control HTTP) et la compression gzip, car il s'agit généralement de texte. Les fichiers llms.txt volumineux (des centaines de Ko) peuvent peser sur la bande passante s'ils sont explorés trop fréquemment, donc une mise en cache appropriée aide. Certains ont suggéré d'héberger llms.txt sur un CDN ou de l'exposer via .well-known/llms.txt afin que les proxys puissent le mettre en cache globalement.

Études de cas approfondies

FastHTML (Framework hypermédia) : L'expérience du projet FastHTML est illustrative. FastHTML est une petite bibliothèque pour créer des API et des documents. Ses développeurs ont reconnu que les modèles de langage typiques (comme Claude) n'avaient aucune connaissance de FastHTML (il a été publié après leur date limite de formation). Pour compenser, ils ont rédigé un llms.txt pour leur site de documentation. Ensuite, en utilisant `llms_txt2ctx`, ils ont généré deux versions de fichiers de contexte : `llms-ctx.txt` (contenu principal) et `llms-ctx-full.txt` (étendu avec des liens optionnels) (Source: github.com). Cela leur a permis de fournir à Claude une vue concise mais complète des documents chaque fois qu'il répondait à des questions. Le résultat : ils ont signalé des réponses assistées par l'IA considérablement améliorées dans leur IDE et leurs bots de documentation, sans que chaque utilisateur n'ait à copier manuellement les liens. Cela démontre que `llms.txt` sert la « longue traîne » du contenu (les documents de FastHTML n'étaient pas indexés par Google, selon [4]). Leur cas montre comment un projet modeste peut tirer parti de llms.txt pour se rendre « recherchable par l'IA » dès le premier jour.

Anthropic (Société d'IA) : L'adoption de llms.txt par Anthropic est plus symbolique que spécifique à un cas. En tant que grande entreprise d'IA, ils ont sans doute moins besoin d'être trouvables par l'IA, mais ils ont néanmoins créé llms.txt pour la transparence et la signalisation communautaire. Leur llms.txt répertorie des introductions à leurs produits (Claude), des documents de recherche, des canaux de développeurs, et plus encore (la sortie [10] montre des pages comme « Claude in Slack », « API », « Customers »). Leur participation confère de la crédibilité : un leader de l'industrie incluant llms.txt suggère qu'il faut le prendre au sérieux. Cela alimente également probablement les propres modèles d'Anthropic (s'ils l'indexent en interne).

Institution académique (exemple) : Certaines universités ont de grands sites web avec des catalogues de cours, des recherches, etc. Un exemple est « Juris Education » qui a un llms.txt de taille considérable (22 885 jetons) (Source: llmstxt.site). La raison peut être d'aider les futurs étudiants ou les tuteurs/chatbots IA à regrouper rapidement les informations sur les cours. De nombreuses universités ont expérimenté des portails IA pour les questions-réponses des étudiants, et llms.txt pourrait servir de ressource backend.

Gouvernement et réglementations : À ce jour, il ne semble pas y avoir de directives gouvernementales officielles sur llms.txt. Cependant, cela résonne avec les débats politiques. Par exemple, l'article de la directive européenne sur le droit d'auteur concernant l'exploration de textes et de données prévoit des exceptions pour la recherche, ce qui implique que les sites web n'auraient pas besoin d'y adhérer explicitement pour cette utilisation si elle est dans le champ d'application. Llms.txt se situe dans

une zone grise : il s'agit de métadonnées volontaires pour l'utilisation des données par l'IA, et non d'une licence contraignante. Certains décideurs politiques préconisent des mécanismes plus exécutoires (par exemple, des lois sur les robots de scraping web). Aucun gouvernement connu n'a mandaté quoi que ce soit de similaire à llms.txt.

Implications et orientations futures

Pour l'avenir, le succès ou l'échec de `llms.txt` dépendra probablement de quelques facteurs clés :

- **Adoption par les plateformes d'IA :** Si les principaux modèles ou outils d'IA viennent à reconnaître et à faire confiance à llms.txt, son adoption pourrait monter en flèche. Par exemple, si OpenAI le soutenait officiellement (par exemple via ChatGPT instruisant GPT sur un lien llms.txt), ou si Google changeait de cap et indexait llms.txt, cela créerait une forte incitation. Inversement, si les développeurs d'IA préfèrent s'appuyer sur les index de recherche ou les embeddings (comme Bing Chat utilise déjà les résultats de recherche en coulisses), la demande pour llms.txt pourrait rester limitée. Le fait que Google le rejette actuellement suggère que la « recherche IA » grand public sera lente à l'adopter. Mais le paysage peut changer rapidement : la dernière fois que nous avons vérifié (juin 2025), Google a déclaré que le SEO normal était suffisant (Source: searchengineland.com), mais un an plus tard, cela pourrait changer si le comportement des utilisateurs évolue vers les résumés IA.

- **Écosystème d'outils et de frameworks :** La croissance des outils de développement autour de llms.txt pourrait faciliter son adoption. Par exemple, si GitHub Pages génère automatiquement llms.txt, ou si WordPress et d'autres CMS l'incluent par défaut, une multitude de nouveaux sites pourraient être « prêts pour llms.txt » du jour au lendemain. Nous en avons déjà vu les débuts : un plugin WordPress existe, certains générateurs de sites statiques ont des add-ons. Si les principaux systèmes de gestion de contenu intègrent un support, l'adoption pourrait grimper indépendamment des grands acteurs de la recherche.

- **Standardisation :** Le passage d'une proposition à une norme nécessite normalement un consensus et un enregistrement. Les auteurs ont laissé entendre qu'ils pourraient l'enregistrer comme URI bien connu (par exemple, `/.well-known/llms.txt`) si la norme prend pied (Source: news.ycombinator.com). Une telle démarche faciliterait l'orientation des bots. De plus, la publication d'un RFC ou d'une note du W3C pourrait cimenter le format. Si llms.txt obtient un soutien formel, cela pourrait signaler un « statut officiel », encourageant une adhésion plus large (tout comme RSS est devenu omniprésent une fois standardisé).

- **Approches alternatives :** Il est possible que de meilleures solutions émergent. Par exemple, Google pourrait développer son propre « sitemap IA » ou des balises meta pour contrôler l'indexation par l'IA, rendant llms.txt obsolète. Ou les assistants IA pourraient utiliser des signaux contextuels (balisage schema.org, données du Knowledge Graph, schémas d'assistants vocaux) pour glaner des informations de manière plus sémantique. Il y a une discussion en cours sur des standards comme les fonctionnalités SERP ou les « indices de prompt IA » intégrés dans le HTML. Dans le pire des cas, llms.txt pourrait devenir *une* parmi de nombreuses propositions similaires, et peut-être être supplanté par un protocole plus élégant.

- **Influence réglementaire :** Si les régulateurs exigent des entreprises d'IA qu'elles respectent robots.txt (dans le cadre de la réglementation des scrapers), une extension logique pourrait être d'exiger le respect des directives llms.txt. Cela pourrait se produire par autorégulation de l'industrie ou par la loi, d'autant plus que les débats sur les données d'entraînement de l'IA et le droit d'auteur s'intensifient. Par exemple, si l'UE ou un pays légiférait que les systèmes d'IA doivent honorer les préférences d'utilisation de contenu publiées par les propriétaires de sites web, ils pourraient explicitement mentionner llms.txt comme un canal reconnu. C'est spéculatif mais dans le domaine de la gouvernance émergente de l'IA.

- **Effets de réseau sur la découverte de contenu :** Nous n'en sommes qu'aux premiers stades de la « découverte de contenu basée sur l'IA ». Si un ou deux assistants IA populaires commencent à utiliser par défaut les listes llms.txt, les utilisateurs pourraient commencer à le voir indirectement. Par exemple, si les réponses de Gemini ou Claude citent régulièrement du contenu d'une page llms.txt, les équipes de contenu avisées le remarqueront et optimiseront leurs fichiers. C'est similaire à la façon dont les experts SEO ont réagi lorsque les extraits optimisés ont commencé à tirer parti de structures HTML particulières (ils ont ensuite modifié leur contenu pour alimenter les extraits). Au fil du temps, une bonne pratique llms.txt pourrait générer des avantages partiels en matière de SEO pour l'IA non capturés par les métriques traditionnelles.

- **Bonnes pratiques communautaires :** L'écosystème `llms.txt` lui-même évoluera grâce à l'expérience partagée. Au fur et à mesure que les adopteurs précoces publieront leurs expériences, les bonnes pratiques communautaires se développeront. Les ressources GitHub et les blogs documentent déjà les choses à faire et à ne pas faire (par exemple, des suggestions sur la façon de structurer les blocs de citation afin qu'ils ne confondent pas un LLM). Au fil des mois, nous nous attendons à voir apparaître

des outils de linting pour llms.txt (vérifiant les liens brisés, la clarté, etc.). Des conventions de versioning pourraient également émerger (tout comme robots.txt n'a pas de version officielle, llms.txt pourrait soit fixer la spécification, soit autoriser des variations comme llms-full.txt).

En conclusion, **l'avenir de `llms.txt` est incertain**. De nombreux observateurs ont noté qu'aucune technologie ne peut garantir l'évolution de l'IA — que le secteur du « comportement de contenu » se consolide autour des seuls éditeurs (comme llms.txt) ou reste décentralisé. Pour l'instant, llms.txt occupe un coin de niche mais actif du web de l'IA. S'il prend de l'ampleur, il pourrait conduire à une nouvelle couche de standards de fichiers web ; sinon, il pourrait tranquillement se retirer en tant qu'expérience intéressante.

Conclusion

Notre enquête sur `/llms.txt` révèle qu'il s'agit d'une proposition bien définie avec des objectifs spécifiques : rendre les sites web plus accessibles aux grands modèles de langage au moyen d'une carte de contenu créée par l'homme. Les spécifications techniques (utilisation de Markdown, listes de liens, etc.) sont claires et relativement faciles à mettre en œuvre. Des études de cas précoces dans la documentation logicielle ont montré que llms.txt *peut* améliorer les performances des agents d'IA sur des tâches de niche (Source: searchengineland.com) (Source: www.released.so). Pourtant, en même temps, il y a une égale mesure de scepticisme. Les principaux moteurs de recherche ont jusqu'à présent publiquement proclamé qu'ils ignoreraient ce fichier (Source: searchengineland.com), et une analyse empirique suggère que les sites grand public ne l'ont pas encore adopté de manière significative (Source: www.rankability.com).

Est-ce important ? Pour l'instant, la réponse est : *Cela dépend de vos priorités*. Si vous êtes un éditeur de technologie, un développeur ou un spécialiste du marketing averti en SEO qui souhaite expérimenter toutes les optimisations de pointe à l'ère de l'IA, `llms.txt` semble valoir la peine d'être exploré. Il impose un coût relativement faible, est réversible, et si les outils d'IA commencent à le prendre en charge de manière extensive, vous aurez pris de l'avance. Il est particulièrement important pour les domaines où les questions-réponses basées sur l'IA peuvent stimuler le support technique ou l'intégration des utilisateurs : documents de développeurs, API, manuels de produits, etc.

Cependant, si vous vous concentrez uniquement sur la recherche traditionnelle ou si vous disposez de ressources limitées, alors `llms.txt` peut être considéré comme facultatif. Le consensus de l'équipe SEO de Google est que le « SEO normal » suffit pour être trouvé dans les résultats de l'IA (Source: searchengineland.com). Les organisations désintéressées par l'entraînement de leurs données par l'IA (ou opposées) pourraient préférer des mécanismes légaux plus concrets (licences, blocages robots) plutôt qu'une liste amicale. Comme le rapport du LLMS Central l'a implicitement suggéré, de nombreux propriétaires de contenu considèrent `llms.txt` comme faisant partie de la **transparence de l'entraînement de l'IA** – mais la question de savoir si une IA le respecte réellement (ou le compense) reste largement non testée.

Pour l'avenir, l'effet le plus immédiat de llms.txt est de susciter des conversations vitales parmi les webmasters sur la conception de contenu pour l'IA. En essayant ce nouvel outil, la communauté peut découvrir où les LLM réussissent ou échouent lors de la digestion de sites réels. Cela informe les deux parties (développeurs de sites et d'IA) sur ce qui fonctionne. En ce sens, llms.txt a déjà eu un certain impact : il a sensibilisé les formateurs d'IA aux problèmes de fenêtre de contexte, et les experts SEO au fait que les moteurs de recherche ne sont pas encore des agents d'IA, etc.

En fin de compte, le récit autour de llms.txt fait écho à des discussions plus larges sur l'avenir du Web : les créateurs de contenu exerceront-ils un contrôle explicite sur l'utilisation de leurs données par l'IA, ou le Web restera-t-il un corpus de texte passif ? Verra-t-on un « web de l'IA » avec de nouvelles mini-normes superposées au HTML (tout comme il existe maintenant des conventions AJAX et JSON), ou l'IA se superposera-t-elle simplement à l'infrastructure existante (annotations sémantiques, exploration améliorée) ? Le jury est toujours en délibération.

Ce qui est clair, c'est que **`llms.txt` est important dans la mesure où l'industrie et la communauté décident qu'il l'est**. Si l'on le considère comme analogue à la façon dont `robots.txt` et `sitemap.xml` ont gagné du terrain, alors son importance augmentera dès qu'un nombre suffisant de contenus et de systèmes d'IA convergeront vers lui. Il est encore tôt, et pour chaque avantage technique sous-jacent revendiqué, il y a des préoccupations tout aussi importantes concernant la nécessité et la viabilité.

À notre avis, `llms.txt` est une expérience proactive et constructive : elle vise à prévenir les malentendus liés à l'IA sur le web. Nos recherches suggèrent qu'il s'agit d'une solution bien intentionnée qui répond à de réels défis techniques (Source: searchengineland.com) (Source: www.released.so). Son succès futur dépendra à la fois de son adoption technique (par les

plateformes d'IA) et de son adoption par la communauté (par les propriétaires de sites). Nous soutenons la poursuite de son exploration – après tout, une approche avec des inconvénients négligeables combinée à un même petit avantage en termes de fidélité de l'IA semble être un pari qui en vaut la peine. Qu'il devienne un élément de la boîte à outils standard d'Internet, ou simplement une note de bas de page dans l'histoire de l'évolution du Web, seul le temps (et les données) le dira.

Références : Toutes les affirmations et les chiffres ci-dessus proviennent des sources citées dans le texte (Source: searchengineland.com) (Source: searchengineland.com) (Source: llmscentral.com) (Source: www.rankability.com) (Source: llms-txt.io) (Source: www.released.so) (Source: llmstxt.site) (Source: www.kdjingpai.com), ainsi que de rapports industriels supplémentaires et de commentaires d'experts, comme détaillé. Chaque citation identifie la source de l'information décrite.

Étiquettes: llmstxt, grands-modeles-de-langage, ia-generative, optimisation-ia, seo, robotstxt, normes-web, optimisation-de-contenu

---