Le LLM de Perplexity : Une analyse technique approfondie de Sonar et PPLX

By rankstudio.net Publié le 17 octobre 2025

Résumé Exécutif

Perplexity AI est une startup basée à San Francisco (fondée en août 2022) qui propose un **moteur de recherche et de réponse alimenté par l'IA**, combinant la recherche web traditionnelle avec des modèles de langage étendus (LLM) pour générer des réponses concises, étayées par des citations et formulées en langage naturel. L'entreprise a rapidement obtenu des financements importants (avec des investisseurs tels que Jeff Bezos, Nvidia, SoftBank, Accel) et a développé sa base d'utilisateurs (plus d'un million d'utilisateurs quotidiens début 2024 (Source: www.theverge.com). La question centrale abordée par ce rapport est de savoir si Perplexity « possède son propre LLM » et quelle pile technologique et architecture elle utilise. La réponse est que Perplexity développe et déploie effectivement des **LLM propriétaires** (collectivement nommés « Sonar » et « PPLX »), tout en exploitant également des modèles externes et open-source (par exemple, la famille GPT d'OpenAI, Claude d'Anthropic, LLaMA de Meta, Mistral, etc.) selon le cas d'utilisation. La technologie de Perplexity intègre ces LLM à son **index de recherche interne** et à des données en temps réel pour fournir des réponses à jour, factuelles et basées sur des sources (Source: www.perplexity.ai) (Source: www.perplexity.ai). L'infrastructure de l'entreprise est hautement optimisée pour la vitesse et l'échelle, utilisant l'inférence GPU (accélérateurs AWS A100, Cerebras, NVIDIA TensorRT-LLM) pour obtenir des réponses à faible latence (Source: www.perplexity.ai) (Source: www.perplexity.ai).

Ce rapport présente une analyse approfondie et étayée de la technologie de Perplexity, de l'historique de l'entreprise et de son impact plus large. Les principales conclusions sont les suivantes : (1) Les *LLM internes* de Perplexity (série « Sonar » et modèles « PPLX online ») sont construits sur des modèles ouverts (Llama 3.3, Mistral, etc.) et affinés pour être ancrés dans les résultats de recherche (Source: www.perplexity.ai) (Source: www.perplexity.ai). (2) La plateforme utilise *également* en option des LLM de pointe d'OpenAI et d'Anthropic : par exemple, le niveau Pro prend explicitement en charge GPT-4/5 et Claude 4.0 en plus de Sonar (Source: www.perplexity.ai). (3) L'architecture de Perplexity suit un pipeline de récupération et de génération en plusieurs étapes : elle émet des requêtes de recherche (souvent via les API Google/Bing ou son propre crawler, extrait le texte pertinent, puis alimente ce contenu dans un LLM pour synthétiser une réponse (Source: primaryposition.com) (Source: www.perplexity.ai). (4) L'entreprise a lancé des services connexes : **PPLX API**, une API publique pour les LLM open-source (Mistral, Llama2, etc.) avec une inférence optimisée (Source: www.perplexity.ai) (Source: www.perplexity.ai) ; **Perplexity Enterprise**, qui peut rechercher à la fois sur le web ouvert et sur des corpus privés (Source: www.axios.com) ; et un **programme pour les éditeurs** visant à partager les revenus publicitaires avec les fournisseurs de contenu (pour répondre aux préoccupations en matière de droits d'auteur) (Source: www.reuters.com) (Source: www.reuters.com). (5) Perplexity est au centre des tendances juridiques et industrielles : elle fait face à des poursuites pour violation de droits d'auteur (Dow Jones/NY Post, NY Times) concernant son utilisation de contenu d'actualités (Source: www.reuters.com) (Source: www.reuters.com), même si elle s'efforce d'intégrer ses outils dans des produits comme Safari d'Apple (apparemment en négociation (Source: www.reuters.com) et d'étendre sa monétisation via des publicités et des fonctionnalités d'achat (Source: www.reuters.com) (Source: www.reuters.com).

En résumé, Perplexity n'est pas un unique « LLM de marque » comme GPT-4 ; c'est plutôt un **moteur de réponse composite** qui orchestre plusieurs LLM (à la fois auto-hébergés et tiers) au-dessus d'un index de recherche propriétaire. Ce rapport couvrira l'histoire de Perplexity, son financement et son équipe, sa pile technologique, les fonctionnalités de ses produits, ses performances et son contexte industriel, avec des détails techniques et des citations approfondis.

Introduction et Contexte

Le paysage de la récupération d'informations en ligne est en train d'être transformé par l'IA générative. Les moteurs de recherche traditionnels (Google, Bing) renvoient des listes de liens ; en revanche, les « moteurs de réponse » basés sur l'IA (comme Perplexity, Microsoft Copilot ou les résumeurs d'IA de Google) visent à fournir une réponse synthétisée directe avec des preuves à l'appui. Perplexity AI (parfois stylisé « perplexity.ai ») est un acteur notable dans ce domaine. Apparue en 2022, Perplexity se positionne comme un « moteur de réponse alimenté par l'IA » qui promet des **réponses rapides, précises et à jour** aux requêtes des utilisateurs, en mettant l'accent sur l'ancrage factuel et les citations de sources (Source: www.axios.com) (Source: www.theverge.com).

L'entreprise a été cofondée en août 2022 à San Francisco par Aravind Srinivas, Denis Yarats, Johnny Ho et Andy Konwinski (Source: cincodias.elpais.com). Srinivas (PDG) est titulaire d'un doctorat de l'UC Berkeley et aurait travaillé chez OpenAI, Google Brain et DeepMind (Source: cincodias.elpais.com) ; Yarats a obtenu un doctorat de NYU et a travaillé chez Meta AI ; Johnny Ho (CSO) a précédemment travaillé chez Quora et a une expérience de programmeur compétitif champion (Source: scaleup.events) ; Andy Konwinski (CTO) a cofondé Databricks et est un créateur d'Apache Spark. Ces fondateurs ont apporté leur expertise en recherche ML et en systèmes à grande échelle (Spark, calcul distribué) à la conception du moteur de Perplexity. La mission de l'entreprise est de **« révolutionner la recherche »** en fournissant des réponses directes et une compréhension contextuelle plutôt qu'une longue liste de liens (Source: cincodias.elpais.com) (Source: www.axios.com). Dès le début, Perplexity a attiré des investisseurs de renom : Jeff Bezos, Nvidia, SoftBank, Y Combinator (Garry Tan), Cyberstarts, et d'autres. Début 2024, elle avait levé plus de 164 millions de dollars en capitaux propres et subventions, atteignant le statut de licorne (valorisation > 1 milliard de dollars) début 2024 (Source: www.theverge.com) (Source: www.reuters.com), et à la mi-2025, certaines sources estimaient sa valorisation entre 9 et 18 milliards de dollars (Source: www.reuters.com) (Source: www.reuters.com). (Un récent rapport du Wall Street Journal a indiqué que Perplexity négocie un tour de financement de 500 millions de dollars à une valorisation de 14 milliards de dollars (Source: www.reuters.com).)

La croissance de Perplexity a été alimentée par l'atteinte rapide de millions d'utilisateurs. En mars 2024, des rapports de presse faisaient état de *plus d'un million d'utilisateurs quotidiens* interagissant avec le moteur d'IA (Source: www.theverge.com). L'utilisation de la plateforme dans les cercles technologiques a également attiré l'attention : le PDG de NVIDIA, Jensen Huang, l'utiliserait « presque tous les jours », et le PDG de Shopify, Tobi Lütke, a déclaré qu'elle avait remplacé Google pour lui (Source: www.theverge.com). Le journaliste d'The Verge, Alex Heath, a constaté que Perplexity excellait sur certaines requêtes nécessitant des réponses spécifiques, bien qu'elle reste limitée par rapport à Google sur d'autres (Source: www.theverge.com). Il est important de noter que Perplexity met l'accent sur la **transparence des sources** : chaque réponse qu'elle génère est accompagnée de *citations cliquables* tirées de documents web (actualités, forums, wikis, etc.), ce qui contraste avec les chatbots LLM typiques qui peuvent halluciner ou omettre la paternité (Source: www.theverge.com) (Source: www.tomsguide.com).

Parallèlement au développement de produits, Perplexity a rapidement élargi son offre. En 2023-2024, elle a introduit :

- **Perplexity (Consommateur)** : Le service gratuit et Pro de chatbot/recherche sur [perplexity.ai] où les utilisateurs peuvent poser des questions et obtenir des réponses avec des sources. (Le niveau « Pro » offre des modèles plus avancés et des limites d'utilisation plus élevées (Source: www.perplexity.ai).)
- **Perplexity Enterprise** : Lancé en avril 2024 (Source: www.axios.com), un produit payant permettant aux entreprises d'indexer à la fois le web ouvert et les données internes privées, fournissant des réponses IA en temps réel à partir de leur propre base de connaissances.
- **PPLX API** : Une API publique (en version bêta) permettant aux développeurs d'utiliser l'infrastructure d'inférence optimisée de Perplexity sur des LLM open-source (par exemple Llama, Mistral). Elle a été lancée fin 2023 (Source: www.perplexity.ai).
- **Perplexity Labs** : Une offre de « terrain de jeu » où les utilisateurs avancés peuvent tester divers modèles open-source et propriétaires au sein de l'interface Perplexity.
- **Programme pour les éditeurs** : À partir de mi-2024, des partenariats avec des éditeurs de médias (Time, LA Times, etc.) pour partager les revenus publicitaires lorsque Perplexity cite leur contenu (Source: www.reuters.com) (Source: www.reuters.com). Cela a été une réponse aux pressions en matière de droits d'auteur de News Corp et d'autres qui ont intenté des actions en justice contre les scrapeurs d'IA (Source: www.reuters.com) (Source: www.reuters.com).

En résumé, Perplexity combine la recherche, l'indexation et l'IA pour répondre aux questions. La question de savoir si elle « possède son propre LLM » est résolue par le fait qu'elle a développé des modèles sur mesure (la série « Sonar ») adaptés à cette tâche, en plus d'utiliser des LLM d'autres fournisseurs. La stratégie technique de l'entreprise est d'intégrer étroitement un composant de recherche/index (leur « moteur de réponse ») avec la génération basée sur les LLM, ce qui donne une architecture de « génération augmentée par la recherche » où les modèles sont ancrés dans du contenu web frais (Source: www.perplexity.ai) (Source: primaryposition.com) (plutôt que de s'appuyer uniquement sur le pré-entraînement).

Présentation de l'entreprise : Historique, Financement et Leadership

Perplexity AI a été constituée en **août 2022** à San Francisco. Son équipe de direction fondatrice réunit des atouts en apprentissage automatique et en systèmes de données à grande échelle. Le PDG Aravind Srinivas a une formation universitaire en ML et une expérience préalable chez OpenAI, Google Brain et DeepMind (Source: cincodias.elpais.com) ; Andy Konwinski (CTO) a cofondé

Databricks (McGlashan, Sagiv, Zhou) et possède une expertise de niveau doctorat en calcul distribué ; Denis Yarats (CTO Produit) est un chercheur en IA de NYU/Meta ; Johnny Ho (CSO) a également cofondé la startup et dirige la stratégie produit (Source: scaleup.events). Ensemble, ils ont imaginé un « moteur de recherche IA » qui synthétise les réponses à la volée, contrastant avec la recherche classique.

Au cours de sa première année, Perplexity a obtenu des investissements de démarrage et de capital-risque précoce. Début 2023, elle avait levé des dizaines de millions (environ 73,6 millions de dollars en janvier 2024 pour une valorisation de 520 millions de dollars selon les rapports (Source: www.reuters.com)). Mi-2023, la frénésie autour de ChatGPT a stimulé l'intérêt des investisseurs, et Perplexity a clôturé une série A (les rapports varient, mais une source indique : 62,7 millions de dollars en avril 2024 (Source: www.reuters.com), portant le financement total à environ 164 millions de dollars (Source: www.axios.com)). En juin 2024, le Vision Fund 2 de SoftBank a accepté d'investir 10 à 20 millions de dollars dans le cadre d'un tour de financement plus important de 250 millions de dollars valorisant Perplexity à environ 3 milliards de dollars (Source: www.reuters.com). Ses soutiens de premier plan continuent d'inclure Nvidia (qui a fourni des crédits GPU), Amazon/Bezos, Y-combinator, Tiger Capital, et d'autres.

Les métriques de croissance de Perplexity ont été impressionnantes : en 2023, elle aurait traité **plus de 500 millions de requêtes utilisateur** même avec un marketing minimal (Source: www.reuters.com). The Verge (mars 2024) a noté le dépassement d'un million d'utilisateurs quotidiens (Source: www.theverge.com), et des milliards de réponses générées avec citation. L'entreprise emploie des centaines de personnes (estimé à 100-250 employés en 2024), dont des ingénieurs, des chercheurs et des curateurs de données pour le réglage fin et l'évaluation. Terrence Townsend (ex-Google) aurait rejoint l'entreprise en tant que responsable de la stratégie produit de recherche. La culture d'entreprise est décrite comme axée sur la mission mais centrée sur les fondateurs ; Srinivas est connu pour ses commentaires publics provocateurs (par exemple, accusant Google de « rattraper son retard » en matière d'IA (Source: www.axios.com) et des coups d'éclat audacieux (tels qu'une **offre d'août 2025 pour acheter le navigateur Google Chrome** pour 42,5 millions de dollars, qui visait en partie à des fins antitrust et de relations publiques (Source: cincodias.elpais.com).

Le modèle économique de Perplexity a évolué. Son principal produit grand public était initialement gratuit, avec un niveau Pro payant introduit pour monétiser les utilisateurs avancés (Source: www.perplexity.ai). L'entreprise a annoncé son intention d'introduire de la publicité de recherche (sans compromettre la qualité des réponses) – en effet, au quatrième trimestre 2024, elle a commencé à tester des publicités et des cartes de contenu sponsorisé via un programme avec des éditeurs comme TIME, Fortune et Der Spiegel (Source: www.reuters.com) (Source: www.reuters.com). De plus, Perplexity vend son produit Enterprise Pro, destiné aux entreprises qui ont besoin d'une recherche de connaissances sécurisée et privée sur leurs documents internes pour environ 40 à 50 $ par utilisateur et par mois (Source: www.axios.com).

Les observateurs de l'industrie continuent de suivre la trajectoire rapide de Perplexity : à la mi-2025, des rapports suggèrent qu'elle lève à nouveau d'importants fonds (par exemple, 500 millions de dollars pour une valorisation estimée entre 14 et 18 milliards de dollars (Source: www.reuters.com) (Source: www.reuters.com). Sa valorisation élevée (jusqu'à 18 milliards de dollars, rumeur fin 2024 (Source: www.reuters.com) (Source: www.reuters.com) et ses mouvements stratégiques (examinant des partenariats avec Apple et proposant l'acquisition de Chrome) indiquent des ambitions au-delà d'un « simple chatbot » pour défier les acteurs établis de la recherche. Des controverses ont également suivi : des géants de l'édition (Dow Jones/NY Post, NY Times) ont poursuivi Perplexity pour violation du droit d'auteur (Source: www.reuters.com) (Source: www.reuters.com), poussant Perplexity à négocier des accords de licence de contenu et de partage des revenus (d'où le programme pour les éditeurs (Source: www.reuters.com). Ces batailles juridiques sont emblématiques des tensions plus larges entre les outils d'IA et les propriétaires de contenu.

Le **Tableau 1** ci-dessous résume les étapes clés de l'histoire de Perplexity, telles que rapportées dans la presse :

DATE	ÉVÉNEMENT	CITATIONS/NOTES
Août 2022	Perplexity AI fondée par Aravind Srinivas, Denis Yarats, Johnny Ho, Andy Konwinski.	Co-fondateurs listés (Source: cincodias.elpais.com)
Janv. 2023	[Financement] Perplexity lève un tour de financement d'amorçage/Série A (financement total d'environ 73,6 M$, valorisation d'environ 520 M$) avec des investisseurs initiaux incluant Bezos, Nvidia, Amazon.	Reuters via SoftBank : tour de janv. 2024 (Source: www.reuters.com)
Mars 2024	Masse critique : plus d'1 million d'utilisateurs quotidiens signalés ; le PDG de Perplexity se vante de réponses IA plus rapides/précises.	Rapport de The Verge (Source: www.theverge.com)
Avr. 2024	Perplexity lance **Enterprise Pro**, une recherche IA pour les entreprises (web + données privées). Un tour de financement (environ 62,7 M$) porte le total à environ 164 M$ de valorisation.	Axios : Enterprise Pro et financement de 164 M$ (Source: www.axios.com)
Avr. 2024	Perplexity lève environ 62,7 M$ (avec Nvidia, Y-Combinator Garry Tan, etc.), valorisation à >1 Md$.	Reuters : « Soutenu par Nvidia, Bezos » (Source: www.reuters.com)
Juin 2024	SoftBank (Vision Fund 2) investit 10 à 20 M$ (sur un tour de 250 M$), fixant la valorisation à environ 3 Md$.	Reuters : SoftBank investit (Source: www.reuters.com)
Juil. 2024	Lancement du programme publicitaire pour les éditeurs (avec des partenaires comme TIME, Fortune, Der Spiegel) pour partager les revenus publicitaires sur le contenu cité par les réponses.	Reuters : programme lancé en juillet (Source: www.reuters.com)
Août 2024	Perplexity annonce qu'elle commencera à afficher des publicités sur sa plateforme (d'ici le 4e trimestre 2024) et partagera les revenus avec des partenaires médias (Time, etc.).	Reuters : publicités sur la plateforme (Source: www.reuters.com)
Oct. 2024	Poursuite : News Corp (Dow Jones/NY Post) poursuit Perplexity pour violation du droit d'auteur (alléguant qu'elle a copié le contenu d'articles mot pour mot).	Rapport juridique de Reuters (Source: www.reuters.com)
Oct. 2024	Perplexity riposte avec le programme pour les éditeurs (janv. 2024 : expansion des partenaires au LA Times, Independent, etc. ; CNBS ?).	Reuters : ajoute de nouveaux éditeurs en déc. 2024 (Source: www.reuters.com) (mentionne les problèmes juridiques)
Nov. 2024	Financement/pistes : Perplexity discute d'une levée de 500 M$ pour une valorisation d'environ 9 Md$ (rapport).	Reuters : levée de 500 M$, valeur de 9 Md$ (Source: www.reuters.com)
Nov. 2024	Lancement de **fonctionnalités d'achat** : cartes de recherche de produits (intégrant Shopify), téléchargement visuel "Snap to Shop".	Reuters : lancement du hub d'achat (Source: www.reuters.com)
Mars 2025	Actualités : Perplexity en pourparlers pour lever environ 500 M$ pour une valorisation de 18 Md$, selon le WSJ.	Rumeurs Reuters (Source: www.reuters.com)
Mai 2025	Fonds : Rapports d'une levée d'environ 500 M$ pour une valorisation de 14 Md$ (Accel en tête). Apple discute de l'intégration d'une IA de type Perplexity dans Safari.	Rapport de financement Reuters/WSJ (Source: www.reuters.com)

DATE	ÉVÉNEMENT	CITATIONS/NOTES
Août 2025	Coup de pub : Perplexity propose d'acheter Google Chrome alors que Google fait face à une action antitrust (offre rapportée à 42,5 M$).	Actualités El País (Source: cincodias.elpais.com)
Août 2025	Tribunal : Perplexity perd sa requête en rejet de la plainte pour droit d'auteur (Dow Jones c. Perplexity), l'affaire se poursuit à NY.	Décision juridique de Reuters (Source: www.reuters.com)

La chronologie ci-dessus montre l'évolution rapide de Perplexity, passant de startup à acteur majeur de la plateforme d'IA en quelques années, mêlant lancements de nouveaux produits (Enterprise, Publicités, Shopping), importantes levées de fonds et controverses très médiatisées.

Architecture technologique et flux de données

Une caractéristique distinctive de Perplexity est son **architecture hybride combinant la recherche web et l'IA générative**. Plutôt que de s'appuyer uniquement sur une base de connaissances LLM figée, Perplexity effectue une récupération d'informations en direct pour fonder ses réponses. En pratique, lorsqu'un utilisateur soumet une requête, le système de Perplexity effectue généralement les étapes suivantes (telles qu'inférées à partir de sources officielles et d'analyses techniques) :

1. **Compréhension et reformulation de la requête (LLM)** : La requête de l'utilisateur (par exemple, « Quelle est la capitale du pays X ? ») est d'abord comprise par un LLM, qui peut la réécrire ou la décomposer en sous-requêtes ou mots-clés. (Le propre LLM interne de Perplexity peut analyser la question et identifier les phrases clés.)

2. **Recherche Web (API de moteur de recherche ou Index)** : Perplexity émet une ou plusieurs requêtes de recherche pour trouver des documents pertinents. Cela peut utiliser leur *index de recherche et leur robot d'exploration internes* (PerplexityBot) ou des API externes. Selon le blog de Perplexity, ils maintiennent des **robots d'exploration web internes et un index propriétaire** qui est « grand, mis à jour régulièrement » et qui privilégie le contenu faisant autorité (Source: www.perplexity.ai). En pratique, une analyse indépendante suggère que Perplexity peut également « distribuer » des requêtes à des moteurs de recherche externes (Google/Bing) si nécessaire (Source: primaryposition.com). Leur blog met l'accent sur l'intégration de la *recherche web en temps réel* : en récupérant des « extraits » web et des URL à jour pour les fournir aux LLM (Source: www.perplexity.ai).

3. **Récupération de contenu et extraction d'extraits** : À partir des résultats de recherche renvoyés (SERP), Perplexity récupère de manière programmatique le contenu textuel des pages les mieux classées (souvent les 5 à 10 premiers résultats) et en extrait les passages pertinents. Il peut appliquer des filtres pour assurer la diversité et la qualité (en évitant le contenu fortement optimisé pour le SEO, par exemple). Ces passages constituent la base de preuves.

4. **Synthèse de réponse LLM (avec ancrage)** : Les passages collectés (extraits) sont *fournis comme contexte* à un grand modèle linguistique avec une instruction système pour répondre à la question originale de l'utilisateur en utilisant *uniquement* ce texte. Cela garantit que la réponse est directement ancrée dans un contenu actuel et factuel. Le blog de Perplexity décrit cela comme un réglage fin des modèles pour « utiliser efficacement les extraits » afin d'améliorer la **fraîcheur, la factualité et l'utilité** (Source: www.perplexity.ai). Le LLM cite systématiquement les sources (hyperliens vers les extraits) dans sa réponse.

5. **Présentation des résultats** : La réponse finale est formatée et renvoyée à l'utilisateur avec des citations intégrées et (souvent) des puces ou des paragraphes. L'utilisateur voit la réponse ainsi que les sources listées. Les utilisateurs peuvent ensuite cliquer sur n'importe quelle citation pour valider l'information.

Ce pipeline est souvent appelé *génération augmentée par récupération* (RAG). L'innovation de Perplexity réside dans l'optimisation de ce flux de bout en bout : ils disposent d'une infrastructure à haute vitesse pour minimiser la latence (atteignant des « réponses quasi instantanées » (Source: www.perplexity.ai), et d'un étiquetage de données et d'un réglage fin propriétaires pour maximiser la précision. Ils affirment prioriser les résultats **« utiles, factuels et à jour »** (Source: www.perplexity.ai). Les évaluations humaines sur ces axes sont une partie essentielle de leur formation et de leur déploiement de modèles, selon leur blog.

Il est important de noter que cette conception signifie que *le moteur principal de Perplexity n'est pas seulement un LLM*. Il s'agit plutôt d'un « **moteur de réponses** » qui utilise les LLM comme l'un de ses composants. Ses LLM ont généralement d'énormes fenêtres contextuelles (des centaines de milliers de jetons) pour ingérer plusieurs documents à la fois (Source: docs.perplexity.ai). Par exemple, les modèles Sonar de Perplexity prennent en charge un contexte allant jusqu'à 128K jetons (Source: docs.perplexity.ai), bien au-delà des LLM typiques. Ils implémentent également des variantes de raisonnement en chaîne de pensée (CoT) (par exemple, Sonar Reasoning Pro utilise une base spécialisée « DeepSeek-R1 ») pour améliorer l'analyse étape par étape (Source: docs.perplexity.ai). Le diagramme ci-dessous (**Figure 1**) illustre l'architecture de Perplexity :

+----------------------------------+ +------------------+ +--------------+ +--------------+ | | | | | | | | | [Requête utilisateur] |----->| Reformulateur de |----->| Système de |----->| Index | | | | requête | | recherche | |Récupérateur | | Exemple : "Qui a gagné l'affaire | | (Analyse LLM) | | (Robot d'exploration | | | | de la Cour suprême hier soir ?" | | | | interne ou API Google/Bing) | | | +----------------------------------+ +------------------+ +--------------+ +-----+--------+ | | v | +------------------+ +----------------------------------+ | | | | | Synthétiseur de réponses LLM | [Extraits de documents récupérés] | | (ex: modèle Sonar) : Étant donné le texte récupéré, | | | produire une réponse avec des citations. | ex: "Reuters: Le juge X a gagné..." | | | +----------------------------------+ | "NYTimes: L'affaire Y classée..." | +------------------+ | Ajustement de la factualité/ | +----------------------------------+ ^ | lisibilité basé sur les éval. humaines | | +----------------------------------+ | +----------------------------------+ | | | Affichage avec sources | +----------------------------------+

Figure 1. Vue d'ensemble de l'architecture hybride de Perplexity. Les requêtes des utilisateurs sont analysées par des LLM, envoyées à un moteur de recherche (le robot d'exploration/index de Perplexity ou une API) pour récupérer du contenu pertinent, puis synthétisées par un LLM en une réponse finale avec des citations. Il est crucial de noter que la grande fenêtre de contexte permet au modèle de "lire" plusieurs extraits simultanément.{Source: Blogs techniques de Perplexity (Source: www.perplexity.ai) (Source: www.perplexity.ai)}

Cette approche contraste avec un chatbot pur comme ChatGPT, qui s'appuie soit uniquement sur ses connaissances pré-entraînées (statiques jusqu'à une date limite), soit sur un plugin de navigation ajouté. La conception de Perplexity entrelace étroitement la recherche web actuelle avec la génération, ce qui en fait davantage un *moteur de recherche IA* qu'un chatbot LLM autonome. L'équipe de Perplexity qualifie souvent le produit de "moteur de réponses" (Source: www.axios.com), soulignant la sérendipité de la recherche avec la fluidité des LLM.

Infrastructure technique

Perplexity a bâti une infrastructure substantielle pour gérer ces charges de travail à grande échelle. Ils exploitent des clusters d'inférence optimisés principalement sur des GPU NVIDIA (AWS A100 via des instances P4d) et utilisent également du matériel spécialisé (machines à l'échelle de la tranche de Cerebras) pour leurs modèles Sonar. Leur **blog PPLX API** détaille une pile d'inférence utilisant la bibliothèque open-source TensorRT-LLM de NVIDIA pour accélérer l'inférence des LLM, atteignant un débit bien plus élevé que les frameworks de base (Source: www.perplexity.ai). Par exemple, les benchmarks de Perplexity montrent que son système optimisé est jusqu'à 2,9 fois plus rapide que l'inférence de génération de texte (TGI) de Meta et 4,35 fois plus rapide en latence du premier jeton (Source: www.perplexity.ai). Ils atteignent plus de **1 200 jetons par seconde** avec Sonar sur le matériel Cerebras (Source: www.perplexity.ai), permettant aux réponses de s'afficher presque instantanément. Cela représente un débit de décodage environ 10 fois plus rapide que certains modèles concurrents (Source: www.perplexity.ai). L'effet net est que la latence des LLM devient imperceptible par rapport à la vitesse de lecture de l'utilisateur ("la vitesse de lecture humaine moyenne est de 5 jetons/sec", tandis que Perplexity sert 1200 jetons/sec (Source: www.perplexity.ai).

En pratique, la flotte d'inférence de Perplexity peut gérer une charge très élevée. Selon des métriques internes, le passage d'une seule fonctionnalité (précédemment servie par une API externe) à leur propre système PPLX a réduit les coûts d'environ 75 % et supporte un trafic quotidien de millions de requêtes (environ 1 milliard de jetons par jour (Source: www.perplexity.ai). Ceux-ci sont cités comme "éprouvés au combat", exécutant des millions de requêtes avec une disponibilité de 99,9 %. La pile est conteneurisée sur Kubernetes pour une mise à l'échelle élastique (Source: www.perplexity.ai).

Côté données, Perplexity investit massivement dans l'infrastructure de recherche. Leurs blogs mettent l'accent sur un **corpus web interne et une pipeline de classement**. Alors que certains analystes ont spéculé que Perplexity s'appuie toujours sur Google/Bing pour les résultats de recherche en direct (Source: primaryposition.com), Perplexity affirme construire son propre index avec des robots qu'ils appellent *PerplexityBot*, en priorisant les sites de haute qualité et en les mettant à jour fréquemment (Source: www.perplexity.ai). Que ce soit via leur propre index ou un système hybride, une chose est claire : la plateforme est

conçue pour ingérer le **web frais**. Les "LLM en ligne" de Perplexity (voir section suivante) explorent explicitement et intègrent le contenu web actuel dans les réponses, permettant d'obtenir des nouvelles ou des faits récents (par exemple, "score du match des Warriors hier soir") que les modèles purement hors ligne ne peuvent pas connaître (Source: www.perplexity.ai).

Pour les développeurs, Perplexity expose également le même environnement d'inférence haute vitesse via l'offre **pplx-api** (LLM-as-a-service) (Source: www.perplexity.ai). Cette API permet à tout utilisateur d'appeler des modèles ouverts (Mistral, Llama2, Code Llama, etc.) sur le backend de Perplexity. Tout le calcul est effectué côté Perplexity – l'utilisateur n'a besoin que d'un simple appel REST, aucun GPU n'est nécessaire. L'API est actuellement gratuite pour les abonnés Perplexity Pro car elle est en bêta publique (Source: www.perplexity.ai). L'infrastructure sous-jacente – des serveurs de modèles conteneurisés avec accélération NVIDIA TensorRT-LLM – est essentiellement le même moteur qui alimente le propre produit de Perplexity.

Dans l'ensemble, la pile technologique de Perplexity peut être résumée comme suit (liste non exhaustive) :

- **Données et Indexation** : Robots d'exploration web propriétaires (PerplexityBot), et potentiellement intégration avec les principales API de recherche. Classement et filtrage sophistiqués pour collecter des extraits de texte pertinents.
- **Modèles LLM** : Un mélange de LLM propriétaires et tiers (détaillés ci-dessous), chacun chargé dans une pipeline d'inférence à contexte élevé (jusqu'à 128K jetons).
- **Matériel d'Inférence** : Principalement des clusters GPU AWS (NVIDIA A100), ainsi que des systèmes Cerebras spécialisés pour une inférence Sonar ultra-rapide (Source: www.perplexity.ai).
- **Logiciel** : NVIDIA TensorRT-LLM pour l'inférence optimisée, orchestration Kubernetes, pipelines de prompts personnalisées. L'API PPLX intègre également des fonctionnalités supplémentaires pour une diffusion efficace.
- **Métriques et Surveillance** : Tests A/B continus avec des utilisateurs réels, surveillance de la satisfaction utilisateur comme métrique clé (Source: www.perplexity.ai), et analyse statistique de la vitesse/précision.

Ensuite, nous examinons les modèles LLM eux-mêmes.

Les modèles LLM de Perplexity

Contrairement à certaines attentes, Perplexity ne s'appuie **pas exclusivement sur un seul LLM géant**. Au lieu de cela, elle utilise une approche de *méta-modèle* : plusieurs modèles sont employés dans différents "modes" (recherche, raisonnement, recherche approfondie), et le système sélectionne souvent le meilleur modèle à la volée. Il est important de noter que Perplexity **développe ses propres LLM** – sous les noms de *Sonar* et *PPLX.* Ce sont des **versions affinées de modèles open-source**, personnalisées pour les cas d'utilisation de Perplexity.

Le modèle interne phare est **Sonar**. Introduit début 2024 et mis à jour à plusieurs reprises, Sonar est *« le modèle interne de Perplexity optimisé pour la qualité des réponses et l'expérience utilisateur. »* En février 2025, Sonar est construit sur le modèle de fondation LLaMA 3.3 70B de Meta, puis entraîné davantage par Perplexity (Source: www.perplexity.ai). L'objectif de l'entraînement était axé sur la *factualité* et la *lisibilité* dans le contexte de la réponse aux recherches. Après l'affinage, Perplexity rapporte que Sonar surpasse significativement d'autres modèles de taille similaire (par exemple, GPT-4o mini, Claude 3.5 Haiku) lors des tests A/B de satisfaction utilisateur (Source: www.perplexity.ai), et approche même les performances de modèles de pointe comme GPT-4o pour une fraction du coût (Source: www.perplexity.ai). Une version mise à jour de Sonar (février 2025) délivre environ 1200 jetons/sec, grâce à l'accélération Cerebras (Source: www.perplexity.ai).

En pratique, "Sonar" n'est pas monolithique : la documentation révèle une famille de variantes Sonar pour différentes tâches :

- **Sonar (base)** – un modèle de recherche léger (non-raisonnement) avec un contexte de 128K, optimisé pour la vitesse et les questions-réponses de base (Source: docs.perplexity.ai) (Source: docs.perplexity.ai).
- **Sonar Pro** – (Mode de recherche avancé) variante à plus grande capacité pour les questions multi-tours ou complexes (détails non publics).
- **Sonar Reasoning** – un modèle de chaîne de pensée (contexte de 128K) pour les problèmes en plusieurs étapes, "alimenté par DeepSeek-R1" (une architecture optimisée) (Source: docs.perplexity.ai) (Source: docs.perplexity.ai).
- **Sonar Reasoning Pro** – un modèle CoT encore plus précis pour les tâches analytiques les plus difficiles (DeepSeek-R1 avec CoT).

- **Sonar Deep Research** – un modèle de niveau expert (probablement un contexte plus large, plus lent) pour les revues de littérature exhaustives et l'analyse approfondie de sujets (Source: docs.perplexity.ai).

Les modèles Sonar de base et Pro sont décrits dans la documentation de Perplexity comme étant adaptés aux requêtes factuelles rapides avec ancrage (Source: docs.perplexity.ai). Ils ont un contexte de 128K jetons et ne sont pas entraînés sur les données des clients (garantissant la confidentialité). La variante "Deep Research" vise à synthétiser plusieurs sources en rapports cohérents. Tous les modèles Sonar sont censés être affinés sur les propres ensembles de données de Perplexity pour les questions-réponses avec un contexte web en temps réel (Source: www.perplexity.ai) (Source: docs.perplexity.ai).

Modèles PPLX-Online : Fin 2023, Perplexity a introduit des modèles "LLM en ligne" sous la marque PPLX : **pplx-7b-online** et **pplx-70b-online** (Source: www.perplexity.ai). Ce sont des modèles de petite et moyenne taille (7B et 70B paramètres) spécifiquement affinés pour exploiter les connaissances web en temps réel. Selon leur blog, `pplx-7b-online` est construit sur **Mistral 7B** comme base, tandis que `pplx-70b-online` utilise **Llama2-70B** comme base (Source: www.perplexity.ai). Les deux sont continuellement ré-entraînés afin de pouvoir récupérer et intégrer des informations à jour ("en ligne" signifie qu'ils intègrent directement des extraits de recherche web) (Source: www.perplexity.ai). Ceux-ci répondent au cas d'utilisation de la gestion des requêtes sensibles au temps (scores, événements d'actualité) en récupérant des faits récents. Le fait qu'ils soient des bases open-source signifie que leurs poids sont plus portables (ces modèles sont également accessibles via le terrain de jeu Perplexity Labs).

Modèles tiers : Perplexity tire également parti des meilleurs modèles du marché. L'abonnement Pro permet explicitement aux utilisateurs de choisir parmi des modèles avancés d'OpenAI et d'Anthropic. Selon l'article d'aide de Perplexity, les abonnés Perplexity Pro peuvent utiliser des modèles tels que les plus avancés d'OpenAI (GPT-4 ou même GPT-5 lors de sa sortie) et Claude 4.0 ("Sonnet") d'Anthropic (Source: www.perplexity.ai). Par exemple, la documentation Pro liste **"GPT-5"** (le prochain modèle d'OpenAI) et **"Claude 4.0 Sonnet"** comme choix disponibles (Source: www.perplexity.ai). (Au minimum, GPT-4a/b est supporté ; la liste suggère qu'ils suivent les dernières versions.) Ces modèles propriétaires ne sont *pas* exécutés sur les propres serveurs de Perplexity ; au lieu de cela, Perplexity utilise des API pour les appeler à la demande dans les modes haut de gamme. Le document d'aide note également que leur propre **Sonar Large** est construit sur LLaMA 3.1 (70B) et "entraîné en interne pour fonctionner de manière transparente avec le moteur de recherche de Perplexity" (Source: www.perplexity.ai), confirmant l'architecture de Sonar.

Pour résumer l'utilisation des modèles par Perplexity :

- **Sonar Large (70B, LLaMA 3.x)** – LLM interne orienté recherche (mode par défaut pour de nombreuses requêtes). Inférence rapide (1200 jetons/s) sur Cerebras.
- **Sonar Pro/Reasoning/Deep Research** – LLM internes spécialisés pour les tâches de raisonnement complexe ou de recherche approfondie. Entraînés CoT.
- **PPLX-7b-online (7B, Mistral)** – Base open-source, pour la fraîcheur.
- **PPLX-70b-online (70B, Llama2)** – Base open-source, pour la fraîcheur.
- **OpenAI GPT-4/4.5/5 (estimé)** – Via API pour la plus haute capacité (fonctionnalité Pro).
- **Anthropic Claude v3/v4** – (coûteux, pour les tâches de nuance, également via API).
- **Autres modèles open-source via l'API PPLX** – (Mistral 7B, Code Llama 34B, etc.) selon les annonces de PPLX (Source: www.perplexity.ai).

Tableau 2 ci-dessous résume ces modèles et leurs rôles :

MODÈLE	TYPE	MODÈLE DE BASE ET TAILLE	RÔLE/UTILISATION
Sonar (interne)	Modèle de réponse de recherche	LLaMA 3.x × 70B (affinage)	LLM par défaut pour les Q&A de recherche ; optimisé pour des réponses factuelles et concises (Source: www.perplexity.ai) (Source: www.perplexity.ai).
Sonar Reasoning	Modèle de chaîne de pensée	Dérivé de Sonar / DeepSeek-R1	Requêtes de raisonnement complexes en plusieurs étapes (avec grand contexte) (Source: docs.perplexity.ai).
Sonar Deep Research	Modèle de recherche exhaustive	Dérivé de Sonar	Rapports thématiques approfondis et synthèse de littérature.
pplx-7b-online	LLM en ligne (ouvert)	Mistral 7B (open-source)	Axé sur la fraîcheur, réponses à jour pour les requêtes opportunes (Source: www.perplexity.ai).
pplx-70b-online	LLM en ligne (ouvert)	LLaMA 2 70B (open-source)	Similaire au précédent, mais contexte plus large pour les requêtes opportunes complexes (Source: www.perplexity.ai).
GPT-4 / GPT-4o / GPT-5	LLM propriétaire	OpenAI (taille inconnue)	Raisonnement/créativité haut de gamme (via API) pour les utilisateurs Pro (Source: www.perplexity.ai).
Claude 3.5/4.0	LLM propriétaire	Anthropic (Sonnet, etc.)	Tâches linguistiques avancées via API (fonctionnalité Pro).
Autres open-source	ex: série Llama 2, Code Llama	Divers (13B, 34B, 70B)	Utilisé via l'API PPLX ou Labs pour le codage, la génération générale (ouvert).
Index PerplexityBot	Pas un LLM, un index de recherche	Index global interne	Alimente la récupération de contenu à jour (encore en développement).

Tableau 2 : Modèles et composants clés utilisés par Perplexity. Sonar et PPLX-Online sont les propres variantes affinées de Perplexity ; GPT et Claude sont des modèles externes intégrés ; d'autres (Llama, Mistral, etc.) sont des modèles open-source déployés via l'API de Perplexity (Source: www.perplexity.ai) (Source: www.perplexity.ai).

La preuve que Perplexity utilise ces modèles provient à la fois de sources officielles et d'analyses externes. Le **blog PPLX-API** liste explicitement les LLM open-source qu'ils proposent (Mistral 7B, Llama 2 13B/70B, Code Llama 34B, etc.) (Source: www.perplexity.ai). Le **blog Online LLMs** indique clairement que pplx-7b-online = base Mistral-7B et pplx-70b-online = base Llama2-70B (Source: www.perplexity.ai). Le **blog « Meet Sonar »** confirme que Sonar est basé sur Llama 3.3-70B (Source: www.perplexity.ai) et fait état de gains de performance. Des actualités technologiques indépendantes reconnaissent que Perplexity utilise des modèles OpenAI en coulisses pour certaines tâches (Source: www.reuters.com), et les FAQ de Perplexity listent elles-mêmes GPT-5/Claude-4, etc. On peut donc conclure : **Perplexity possède ses propres LLM (Sonar/PPLX) mais utilise également d'autres modèles de manière flexible.**

Capacités et évaluation des modèles

Perplexity met l'accent sur une évaluation rigoureuse de ses modèles selon plusieurs axes. Selon leur blog, ils évaluent l'**utilité, la factualité et l'actualité** via des ensembles de données sélectionnés et des évaluateurs humains (Source: www.perplexity.ai). L'actualité est évaluée en vérifiant si la réponse contient des informations à jour. L'équipe Sonar rapporte qu'après le fine-tuning, Sonar a significativement amélioré sa factualité et sa lisibilité (concision, clarté) par rapport à son modèle de base (Source: www.perplexity.ai). Ils affirment que Sonar surpasse même ses concurrents propriétaires : lors de tests A/B en aveugle, les

utilisateurs ont préféré les réponses de Sonar à celles de GPT-4o mini et Claude 3.5 Haiku par une marge substantielle, et l'ont trouvé comparable aux réponses de GPT-4o (Source: www.perplexity.ai) (Source: www.perplexity.ai). De plus, sur les benchmarks standards (suivi d'instructions, connaissances générales), Sonar « surpasse les modèles de sa catégorie » comme GPT-4o mini et Claude 3.5 (Source: www.perplexity.ai).

Bien que ces résultats soient internes, ils suggèrent que les modèles de Perplexity sont hautement optimisés pour leur cas d'utilisation. Des comparaisons indépendantes renforcent ce tableau : une critique de Tom's Guide a révélé que le moteur de Perplexity « surpassait constamment » la nouvelle recherche IA de Google dans la plupart des requêtes de test (Source: www.tomsguide.com). Un autre rapport d'utilisateur a loué Perplexity pour avoir agrégé diverses sources (y compris Reddit et des revues) afin de fournir des réponses détaillées et précises sans hallucination (Source: www.tomsguide.com). Ces observations anecdotiques, associées aux témoignages d'utilisateurs (par exemple, du PDG de Shopify et d'autres (Source: www.theverge.com), indiquent que la plateforme est compétitive dans la recherche basée sur l'IA.

Cependant, aucun score de benchmark public (comme GPT4Eval ou les métriques F1) n'est disponible pour les modèles de Perplexity. L'entreprise se concentre davantage sur la satisfaction de l'utilisateur final que sur les scores académiques. Les seuls chiffres publics disponibles concernent les performances/la latence : comme indiqué, Sonar sur Cerebras est environ 10 fois plus rapide en décodage que Gemini 2.0 Flash (Source: www.perplexity.ai). Le blog de l'API PPLX quantifie les améliorations de débit (par exemple, une génération de jetons 1,9 à 6,75 fois plus rapide que les bases de référence TensorFlow/GEMM (Source: www.perplexity.ai). À l'échelle, Perplexity affirme que le système peut supporter plus d'un million de requêtes par jour et près d'un milliard de jetons traités quotidiennement (Source: www.perplexity.ai), illustrant sa robustesse en production.

Récupération et actualité

Une innovation critique de Perplexity est la récupération « **en ligne** » : l'extraction active de nouvelles informations. Cela résout deux problèmes récurrents des LLM : les connaissances obsolètes et l'hallucination. Les blogs de Perplexity soulignent qu'ils disposent d'**ingénieurs de données et de spécialistes de la recherche** dédiés qui parcourent le web, indexent des millions de pages et mettent à jour l'index régulièrement (Source: www.perplexity.ai). Ils affinent même les LLM pour intégrer ces extraits. En pratique, cela signifie que leurs LLM peuvent répondre à des requêtes sur des événements très récents en utilisant le contenu web en temps réel inclus dans l'invite. Par exemple, les modèles PPLX-Online peuvent répondre à « Qui a gagné le match hier soir ? » en recherchant les scores en ligne. Cela contraste avec la plupart des LLM dont les connaissances s'arrêtent à une date limite d'entraînement (par exemple, la date limite de GPT-4 est 2021).

De l'extérieur, cette récupération dynamique fonctionne comme suit (conformément à tout système RAG). Considérons la requête « Que s'est-il passé lors de la décision de la Cour suprême sur X le 15 août 2025 ? ». Le système probablement :

- Utilise un modèle pour générer des requêtes de recherche comme « Résumé de la décision de la Cour suprême X 15 août 2025 ».
- Interroge l'index de recherche Penguin ou Google pour les derniers résultats.
- Récupère les articles de presse ou les textes juridiques liés à partir des résultats.
- Transmet ces extraits de texte (avec URL) à Sonar avec une instruction de répondre de manière factuelle.
- Sonar répond, en citant les sources des extraits.

Dans certaines comparaisons rapportées, les aperçus IA de Google se sont limités à des résultats web statiques ou ont donné des réponses minimales, tandis que l'IA de Perplexity a répondu avec un texte synthétisé plus riche (Source: www.tomsguide.com). Un blogueur détaillé (« How Perplexity Crawls and Ranks ») émet l'hypothèse que l'implémentation *en coulisses* de Perplexity pourrait impliquer des appels à Google/Bing pour récupérer des pages (Source: primaryposition.com). Que Perplexity s'appuie sur son propre index ou qu'il utilise des moteurs de recherche plus importants comme proxy, l'**effet** est qu'il fournit des informations actuelles dans ses réponses. L'entreprise souligne avec insistance que ses modèles excellent dans les requêtes où l'« actualité » est cruciale, un objectif de conception intentionnel (Source: www.perplexity.ai) (Source: www.tomsguide.com).

Cet accent mis sur l'actualité et la factualité influence l'entraînement des modèles. Les LLM de Perplexity sont explicitement affinés pour préférer les réponses basées sur des preuves plutôt que sur des écrits spéculatifs. Par exemple, Sonar a été entraîné à privilégier l'« ancrage » (faits basés sur des preuves) et la clarté (Source: www.perplexity.ai). L'évaluation des réponses met

l'accent sur l'exactitude factuelle (moins d'hallucinations) plutôt que sur la créativité. Les commentateurs de l'industrie notent que les réponses de Perplexity ont tendance à pécher par excès de complétude (citant parfois trop) plutôt que par concision, ce qui, selon eux, peut être un compromis (Source: www.tomsguide.com).

Produits et fonctionnalités de Perplexity

Au-delà de la technologie de base, Perplexity propose une suite de produits destinés aux utilisateurs :

Perplexity Consumer (Chatbot/Moteur de réponses) : Le service phare est l'interface web (perplexity.ai) et l'application mobile où les utilisateurs tapent leurs questions. L'interface est minimaliste : une boîte de discussion et une liste de citations de réponses. Les utilisateurs voient des réponses qui incluent souvent des puces ou des explications, chacune liée à des sources. En mode gratuit, les utilisateurs ont une limite de requêtes quotidienne (variable ; par exemple, 10 questions/jour lorsqu'ils demandent des réponses de type GPT-4). Un niveau payant « Perplexity Pro » (20 $/mois en 2024) débloque des limites plus élevées et la possibilité d'utiliser des modèles avancés (par exemple, GPT-4) pour certaines requêtes (Source: www.perplexity.ai), ainsi qu'une clé API. Selon les retours, les utilisateurs Pro constatent des résultats plus rapides et plus pertinents.

Perplexity Enterprise : Annoncé en avril 2024 (Source: www.axios.com), il s'agit d'un abonnement pour les entreprises. Il permet de connecter le moteur Perplexity à des ensembles de données internes (documents, intranets, Slack, etc.) ainsi qu'au web public. Les utilisateurs d'entreprise peuvent poser des requêtes qui mélangent connaissances internes et externes. L'interface fournit toujours des réponses citées, mais peut désormais inclure des documents d'entreprise. Le prix a été rapporté à environ 40 $/mois/utilisateur. Ce produit est en concurrence avec des services d'IA d'entreprise comme Microsoft Copilot pour les entreprises ou même des outils spécialisés d'eDiscovery. Perplexity le présente comme un moyen d'« accélérer la recherche » en agrégeant les connaissances web et privées (Source: www.axios.com).

API PPLX : Comme décrit, il s'agit d'une API orientée développeurs. Elle permet un accès programmatique à la pile de modèles de Perplexity. Les développeurs peuvent spécifier un modèle (par exemple, `pplx-7b-online`) et obtenir des complétions. Les arguments de vente sont une faible latence, un débit élevé et une interface REST simple. Perplexity compare l'API et la trouve beaucoup plus rapide que les alternatives (par exemple, Anyscale, Replicate GPUs) (Source: www.perplexity.ai). Les cas d'utilisation incluent la création de chatbots personnalisés, d'applications ou l'intégration de LLM dans des produits sans gérer de GPU. L'API PPLX est actuellement en version bêta et gratuite pour les abonnés Pro, avec des plans pour des niveaux payants ultérieurement (Source: www.perplexity.ai). Elle représente l'entrée de Perplexity sur le marché de l'infrastructure IA, à l'instar de l'API d'OpenAI.

Programmes pour éditeurs/partenaires : Pour atténuer les problèmes de droits d'auteur et générer des revenus, Perplexity a lancé un **programme pour éditeurs**. À partir de mi-2024, il a offert aux sites d'information/médias participants une part des revenus publicitaires chaque fois que le moteur d'IA cite leur contenu (Source: www.reuters.com) (Source: www.reuters.com). Parmi les partenaires initiaux notables figuraient TIME, SPIN Media (Spin, Slate magazine), Fortune, et des médias étrangers comme Der Spiegel (Source: www.reuters.com). Fin 2024, il a élargi la liste aux grands journaux américains (LA Times) et aux titres britanniques/européens (Source: www.reuters.com). Ce programme donne également à ces éditeurs accès à des analyses sur la fréquence et l'endroit où leur contenu est cité, transformant ainsi les statistiques d'utilisation de Perplexity en un nouveau canal de trafic. Les unités publicitaires sont placées avec soin afin de ne pas perturber les résultats de la question de l'utilisateur. Les publicités/parrainages de recherche devraient arriver au quatrième trimestre 2024 (Source: www.reuters.com) et il est explicitement indiqué qu'ils n'influencent pas la réponse (tout comme Google affirme que les publicités n'affectent pas le classement de recherche). Cette initiative ouvre non seulement une source de revenus, mais répond également en partie aux poursuites pour violation de droits d'auteur en offrant des licences et des paiements aux producteurs de contenu. En fait, des rapports de Reuters mentionnent des « partenariats musicaux » initiés parallèlement à des litiges juridiques (Source: www.reuters.com).

Fonctionnalités d'achat : Fin 2024, Perplexity a ajouté des capacités de commerce électronique. Un « hub d'achat » peut répondre aux requêtes de produits en affichant des fiches produits avec des images et des détails (via une intégration avec Shopify) (Source: www.reuters.com). Il a également introduit une fonction « Snap to Shop » basée sur l'image : les utilisateurs peuvent télécharger une photo d'un article et Perplexity recherchera les produits correspondants. Ces fonctions sont probablement soutenues par des modèles de reconnaissance/intégration d'images et des API vers les catalogues des détaillants. L'objectif est de capter les requêtes à intention d'achat et de générer des revenus d'affiliation/de parrainage. Reuters a noté ces fonctionnalités comme faisant partie des efforts de Perplexity pour concurrencer la domination de Google dans la recherche (Source: www.reuters.com). Initialement réservées aux États-Unis, les fonctionnalités d'achat pourraient s'étendre à l'international.

Pièces jointes et navigation (Fonctionnalités utilisateur) : Selon un article de Tom's Guide, Perplexity permet aux utilisateurs de télécharger des pièces jointes (PDF, diapositives, plans d'étage) pour obtenir des réponses pertinentes à ce contenu (Source: www.tomsguide.com). Il s'agit d'une capacité relativement unique (le chat de Google ne permet pas les pièces jointes en 2025). Cela suggère que Perplexity a intégré des pipelines d'ingestion de données pour inclure les documents fournis par l'utilisateur dans le contexte de récupération. Cette capacité serait très utile dans les scénarios de recherche.

Dans tous ces produits, l'**expérience utilisateur** est similaire : une interface de chat, des résultats immédiats, des citations et la possibilité de poser des questions de suivi sans perdre le contexte (mode conversationnel avec état). Contrairement à de nombreux chatbots LLM, Perplexity réinitialise délibérément le contexte à chaque session (il n'a pas de mémoire à long terme), mettant l'accent sur la confidentialité et la véracité (Source: www.theverge.com). Chaque nouvelle conversation est sans état, ce qui, selon eux, aide à éviter la confusion et les hallucinations. Cependant, les utilisateurs ont noté que cela signifie « qu'il faut reformuler le contexte à chaque session », un compromis de leur conception épurée (Source: www.theverge.com).

Données, statistiques et performances

Perplexity a publié et rapporté diverses métriques de performance, et certaines ont été vérifiées indépendamment par des journalistes. Les points de données notables incluent :

- **Latence et Débit :** Comme mentionné, Sonar sur Cerebras : ~1200 jetons/sec (Source: www.perplexity.ai). Le benchmark de l'API PPLX : jusqu'à 2,9 fois plus rapide en latence globale par rapport au TGI de Meta sur le même matériel (Source: www.perplexity.ai), et 4,35 fois plus rapide en latence de première réponse lors des tests (pour un modèle Llama-2-13B). Le débit de jetons était 1,9 à 6,75 fois plus rapide que le TGI sous charge (Source: www.perplexity.ai).
- **Échelle :** Perplexity déclare que l'API PPLX « pourrait supporter une charge quotidienne de plus d'un million de requêtes, totalisant près d'un milliard de jetons traités quotidiennement » sans dégradation de la qualité (Source: www.perplexity.ai). En interne, leurs clients (via l'API PPLX) incluent au moins une fonctionnalité dans leur produit principal, qui coûtait auparavant 0,62 M$/an via OpenAI, désormais remplacée par leur API (Source: www.perplexity.ai).
- **Volume de requêtes :** Le rapport de SoftBank mentionne que Perplexity a « traité plus de 500 millions de requêtes en 2023 » (Source: www.reuters.com). The Verge mentionne un nombre estimé d'un million d'utilisateurs quotidiens début 2024 (Source: www.theverge.com). Si cela se maintient, cela impliquerait de l'ordre de ~300+ millions de questions par an (en supposant qu'un utilisateur moyen pose quelques dizaines de questions).
- **Évaluations des modèles :** Bien que Perplexity ne publie pas de métriques de classement publiques, l'entreprise cite des résultats de tests A/B internes. Par exemple, dans l'annonce de Sonar [8], des barres d'échelle montrent que Sonar est bien mieux classé que GPT-4o mini/Claude Haiku en termes de satisfaction utilisateur. (Les chiffres exacts ne sont pas fournis dans le texte, mais les graphiques indiquent que Sonar bénéficie souvent d'une préférence majoritaire de plus de 50 %). Ils mentionnent également avoir surpassé Llama-3.3 de base en termes de factualité/lisibilité.
- **Études utilisateur :** Le blog de Perplexity [8] décrit des tests A/B en ligne approfondis avec de vrais utilisateurs. Ils ont constaté des améliorations statistiquement significatives de la satisfaction avec Sonar par rapport aux modèles de référence, sans compromettre la vitesse. Ils notent également qu'il n'y a pas eu de « différence statistiquement significative » en termes de qualité lorsqu'ils ont basculé une fonctionnalité d'une API externe vers leur propre API PPLX (Source: www.perplexity.ai), ce qui signifie que les réponses de leurs modèles étaient équivalentes à celles des grands modèles externes lors de tests en aveugle.
- **Benchmarks techniques :** Sonar a affirmé dépasser des modèles comme Google Gemini et Claude en termes de vitesse de décodage (Source: www.perplexity.ai). Bien que ces entreprises publient rarement des chiffres bruts, l'affirmation d'être « 10 fois plus rapide que Gemini 2.0 Flash » suggère une focalisation sur la performance comme facteur de différenciation du produit. Pour contextualiser, Gemini 2.0 Flash de Google est lui-même optimisé, donc cette affirmation de vitesse (si vérifiée) dénote un travail d'ingénierie significatif.

Dans les rapports publics, les utilisateurs ont noté la rapidité de Perplexity. Tom's Guide a observé que les réponses de Perplexity apparaissent presque instantanément, même pour des requêtes complexes, là où l'IA de Google avait souvent une réponse plus lente ou nécessitait de faire défiler une liste de liens d'articles (Source: www.tomsguide.com) (Source: www.tomsguide.com). De

manière anecdotique, les réponses longues peuvent prendre 1 à 2 secondes, ce qui est très performant pour un système LLM. En résumé, l'*enveloppe de performance* de Perplexity est élevée : réponses en moins d'une seconde, disponibilité de 99,9 % et capacité à servir des millions d'utilisateurs avec des citations.

Une autre métrique pertinente est la **précision factuelle**. Bien que difficile à quantifier, l'accent mis par Perplexity sur les réponses étayées par des sources suggère des taux d'hallucination plus faibles que les chatbots non contraints. L'article de Tom's Guide faisant l'éloge de Perplexity a souligné qu'il « fournit des réponses plus précises en évitant les hallucinations de l'IA et en s'appuyant sur un contenu web fiable » par rapport à l'IA de Google (Source: www.tomsguide.com). Ils ont également noté l'avantage de Perplexity qui permet aux utilisateurs de vérifier les informations via les URL citées. Les preuves anecdotiques de la communauté s'accordent généralement avec cela : lorsque Perplexity échoue ou hallucine, c'est souvent lorsque sa récupération ne trouve pas de bonne source ou lorsque la question nécessite plus de raisonnement que le contenu de l'extrait ne fournit. En revanche, les LLM typiques pourraient inventer des détails avec assurance.

En bref, la performance de Perplexity se caractérise par une **réponse rapide**, une **connaissance étendue (via la récupération)** et une **grande précision en situation réelle** sur les requêtes spécifiques à un domaine. Son débit et son architecture suggèrent qu'il peut évoluer, et ses processus d'évaluation indiquent qu'il vise un niveau de fiabilité au-delà d'un « LLM » générique.

Études de cas et retours utilisateurs

Bien que les études de cas formelles soient limitées, plusieurs exemples illustrent l'utilisation de Perplexity :

- **Aide à la recherche :** Des universitaires et des étudiants ont déclaré utiliser Perplexity pour obtenir des aperçus rapides sur des sujets. Parce que Perplexity cite ses sources, il peut servir d'outil rapide de découverte de littérature. Des blogs industriels mentionnent des bibliothécaires le testant sur des corpus académiques (Source: medium.com). (Par exemple, en combinant Perplexity avec des API académiques comme CORE ou SemanticScholar, on peut interroger des articles et obtenir des réponses résumées). La capacité à télécharger des PDF (comme noté par Tom's Guide (Source: www.tomsguide.com) étend cette fonctionnalité à l'analyse de documents spécifiques.

- **Questions-réponses techniques :** Pour l'aide au codage ou les problèmes de configuration, les développeurs préfèrent parfois Perplexity à la recherche car il synthétise des solutions à partir de plusieurs fils de discussion de forums. (Ceci est anecdotique mais cohérent avec la manière dont les questions-réponses de StackOverflow pourraient être agrégées par les LLM). La mention de l'intégration de Llama2 et Code Llama suggère que Perplexity pourrait également répondre à des questions spécifiques au code, bien que nous n'ayons aucune référence directe à cette fonctionnalité. Les laboratoires PPLX de leur site web incluent des modèles de code (comme le modèle de code de Replit), indiquant un cas d'utilisation dans l'assistance à la programmation (Source: www.perplexity.ai).

- **Intelligence économique :** Perplexity Enterprise permet aux entreprises d'interroger leurs données internes. Bien qu'aucun cas client public ne soit cité dans la presse, on peut imaginer son utilisation par des analystes souhaitant des résumés rapides de rapports internes. L'existence du produit a été rapportée (Source: www.axios.com), mais les témoignages d'utilisateurs ne sont pas publiquement connus. Cependant, l'idée générale est qu'un analyste financier, par exemple, pourrait demander « Quelles ont été nos 3 meilleures campagnes marketing le trimestre dernier, basées sur les métriques internes et les tendances externes ? » et obtenir une réponse semi-structurée tirant des informations des CRM et des actualités.

- **Utilisation pour l'apprentissage/le divertissement :** Les utilisateurs grand public se sont tournés vers Perplexity comme un « second cerveau » pour des questions curieuses (comme « Pourquoi le pain lève-t-il à la cuisson ? » ou « Quelle est l'histoire du café ? »). L'étendue des requêtes uniques est élevée – la plateforme inclut des invites complexes (comme un planificateur d'itinéraire ou des anecdotes juridiques). Le test de Tom's Guide avec 7 requêtes couvrait les voyages, l'histoire de la technologie IA, l'économie, etc., et a constaté que Perplexity donnait des réponses plus riches que la version de Google (Source: www.tomsguide.com). À titre d'« exemple », un résultat a été que Perplexity a résumé succinctement les connaissances d'experts sur la technologie de réduction du bruit, tandis que Google a principalement renvoyé des liens de listes.

- **Contexte concurrentiel :** Comment les utilisateurs comparent-ils Perplexity aux alternatives ? Tom's Guide suggère un changement croissant, Perplexity l'emportant sur les « réponses détaillées » (Source: www.tomsguide.com). Un autre article (Tom's Guide, oct. 25) a dressé une liste de « 4 raisons d'abandonner Google » pour Perplexity, notant sa capacité à puiser de manière exhaustive dans Reddit, les actualités, les revues (Source: www.tomsguide.com), sa précision et ses réponses

instantanées. Pendant ce temps, The Verge a noté que la conception de Perplexity présentait des compromis : il est « sans état » (donc pas de mémoire continue) et nécessite parfois que les requêtes soient formulées correctement (Source: www.theverge.com). Certains critiques affirment que les outils de recherche basés sur l'IA sont encore à leurs débuts pour comprendre la véritable intention de recherche (voir l'article de No BS Marketplace), mais le consensus général est que Perplexity représente un pas en avant majeur pour la recherche quotidienne.

Les retours utilisateurs soulignent également des limitations : occasionnellement, Perplexity peut omettre certains contextes, ou sa réponse peut être volontairement trop brève pour encourager à cliquer sur les sources. Son utilité se manifeste généralement lorsque la requête est factuelle/de niche ; les questions philosophiques ou très ouvertes peuvent le dérouter. La transparence de Perplexity (citations, pas de manipulation cachée du modèle) est largement appréciée.

Implications, défis et orientations futures

L'essor de Perplexity et d'outils similaires a de multiples implications :

- **Pour la recherche :** Perplexity représente un **nouveau paradigme de recherche**. Si des outils comme celui-ci se généralisent (par exemple, intégrés aux navigateurs ou sous forme d'application), les moteurs de recherche traditionnels devront s'adapter. Google ajoute déjà des aperçus d'IA aux résultats de recherche, Microsoft intègre la technologie OpenAI dans Bing, et Apple serait en pourparlers pour intégrer la recherche IA (Perplexity se serait proposée pour être incluse dans Safari (Source: www.reuters.com). Le succès de Perplexity pourrait pousser Google à améliorer la qualité de ses propres réponses ou à s'associer à d'autres.

- **Aspects juridiques et économiques :** Les **poursuites pour violation de droits d'auteur** contre Perplexity (par Dow Jones/NY Post de News Corp fin 2024 (Source: www.reuters.com) et par The New York Times) soulignent la tension entre les modèles d'IA et le droit de la propriété intellectuelle. Le modèle de Perplexity s'entraîne sur du contenu extrait et génère des citations, ce que les entreprises de médias considèrent comme une copie non autorisée. Perplexity a réagi en établissant des programmes de partage de revenus (Source: www.reuters.com). L'issue de ces poursuites (en août 2025, un tribunal a autorisé la poursuite de l'affaire de New York (Source: www.reuters.com) pourrait créer des précédents pour les fournisseurs d'IA : auront-ils besoin de licences pour le contenu ? L'approche de Perplexity consistant à s'associer avec des éditeurs pourrait devenir plus courante.

- **Modèle économique :** L'ouverture de Perplexity aux publicités et au shopping indique comment la recherche générative pourrait être monétisée. Ils visent à maintenir la fiabilité de leurs résultats même en insérant des unités publicitaires, affirmant que les publicités « n'influenceront pas les réponses » (Source: www.reuters.com). Les observateurs vérifieront si cette affirmation tient, car l'intégration du commerce avec des réponses impartiales est délicate. La proposition de 42,5 millions de dollars de l'entreprise pour acheter Chrome (août 2025) était plus symbolique, mais elle souligne leur stratégie de perturbation du monopole de Google (à l'instar des problèmes antitrust de Google).

- **Écosystème de l'IA :** L'infrastructure construite par Perplexity (par exemple, l'API PPLX) pourrait alimenter l'écosystème plus large des développeurs d'IA, offrant une alternative concurrentielle à OpenAI/Anthropic. En optimisant les modèles ouverts et en rendant open source les améliorations de latence, ils pourraient aider à pousser l'industrie vers une inférence plus efficace. L'API PPLX montre également une tendance de passage des modèles uniquement fermés aux systèmes ouverts hybrides.

- **Aspects éthiques :** La conception de Perplexity (citations de sources, aucune rétention de données) s'aligne sur les appels actuels à l'éthique de l'IA en matière de traçabilité. Ils affirment également ne pas s'entraîner sur les données des utilisateurs par défaut. Cependant, l'outil peut toujours produire des extraits protégés par le droit d'auteur mot pour mot (ce qui a déclenché des poursuites). La manière dont Perplexity gérera l'utilisation équitable, les licences et la confidentialité des utilisateurs dans les futures mises à jour sera importante.

- **Évolution technique :** À la pointe, Perplexity a fait allusion à l'intégration de GPT-4.5 (certains médias ont rapporté « GPT-4.5 est maintenant en ligne sur Perplexity ») et éventuellement à d'autres mises à niveau de LLM (Source: www.linkedin.com). Leur propre Sonar continue d'évoluer (par exemple, Llama 3.3 de base, peut-être Llama4 bientôt). À mesure que les modèles open source (comme Llama3, Mistral2, etc.) s'améliorent, Perplexity est susceptible de les intégrer rapidement (ils mentionnent l'intégration de nouveaux modèles dans les heures suivant leur publication (Source: www.perplexity.ai). La prolifération de modèles Perplexity spécialisés (comme « Sonar-Coder » pour les programmeurs ou les sonars multimodaux) est concevable.

- **Paysage plus large de la recherche IA :** Le succès de Perplexity suggère que la **« recherche basée sur les LLM »** est un thème majeur pour l'avenir. Les concurrents incluent « Copilot » de Microsoft (intégré à Bing et Office), d'autres startups de recherche IA (Neeva/Community Search), et des bots de recherche internes par Apple, Meta, etc. Chacun adoptera une approche légèrement différente (certains s'appuient davantage sur le résumé PNG, d'autres sur un ensemble d'API). Le modèle hybride de Perplexity semble actuellement l'un des plus matures. Si Apple intègre effectivement un moteur de recherche IA (comme le bruit court (Source: www.reuters.com), Perplexity souhaite faire partie de leur liste de fournisseurs.

- **Comportement des utilisateurs :** Une question ouverte est de savoir comment les gens passeront de la recherche traditionnelle aux réponses générées par l'IA. Les articles de Tom's Guide suggèrent que certains des premiers utilisateurs préfèrent Perplexity pour des réponses détaillées et prévoient d'« abandonner Google » (Source: www.tomsguide.com). En entreprise, si la recherche de données internes devient considérablement plus facile, les flux de travail d'information pourraient changer. Même en dehors de la recherche, des modèles de type Perplexity pourraient augmenter les assistants personnels (imaginez Siri avec Perplexity en coulisses).

Globalement, la trajectoire de Perplexity illustre comment les **grands modèles linguistiques sont intégrés aux données en temps réel et à la recherche** pour former des outils pratiquement utiles. Leur investissement continu dans des modèles personnalisés (Sonar), des API ouvertes et des fonctionnalités centrées sur l'utilisateur les positionne pour influencer l'avenir des applications d'IA. Des défis subsistent : la conformité légale, l'assurance de la précision des réponses, la mise à l'échelle responsable. Mais la tendance est claire : les « moteurs de recherche » basés sur l'IA ne sont plus de la science-fiction.

Conclusions

Perplexity AI est à la fois une entreprise d'IA bénéficiant d'un financement de capital-risque important et un pionnier technologique dans le domaine émergent de la recherche générative. Ce rapport a montré que **Perplexity possède bien ses propres LLM**, principalement la famille « Sonar » (basée sur LLaMA, affinée pour les questions-réponses factuelles) et les modèles « PPLX Online » (basés sur Mistral et Llama2) (Source: www.perplexity.ai) (Source: www.perplexity.ai). Ces modèles internes alimentent la fonctionnalité principale de recherche-réponse. En même temps, la plateforme de Perplexity est un **méta-système** : elle exploite également des LLM leaders de l'industrie d'OpenAI (GPT-4/4.5/5) et d'Anthropic (Claude v3/v4) pour certains cas d'utilisation, et elle propose des modèles open source sur son API (Source: www.perplexity.ai) (Source: www.perplexity.ai). La stratégie de l'entreprise est de combiner la génération de LLM avec un index de recherche à jour, une inférence optimisée par le matériel et une amélioration constante des données pour surpasser la recherche traditionnelle.

En termes techniques détaillés, la pile technologique de Perplexity comprend :

- **Index de recherche et crawler propriétaires** (acquérant et classant continuellement le contenu web).
- **Pipeline hybride de récupération-génération** qui alimente les LLM avec les extraits de documents les plus récents.
- **LLM personnalisés et affinés (Sonar, etc.)** construits sur de grands modèles ouverts pour optimiser les réponses factuelles.
- **Intégration avec des API de LLM commerciaux** pour des capacités premium.
- **Infrastructure d'inférence haute performance** (GPU AWS A100, NVIDIA TensorRT-LLM, puces Cerebras) pour garantir des réponses à faible latence.
- **API développeur (PPLX)** et laboratoires qui étendent la technologie à une utilisation externe.

Nous avons étayé tous ces points par des **citations explicites** provenant des propres communications de Perplexity (blogs, documents) et de sources d'information fiables (Reuters, Axios, The Verge, Tom's Guide, Reuters, etc.). Par exemple, le blog officiel de Perplexity annonce les technologies de base (Source: www.perplexity.ai) (Source: www.perplexity.ai), et de nombreux médias confirment l'utilisation de modèles de classe GPT et les fonctionnalités internes de la plateforme (Source: www.perplexity.ai) (Source: www.reuters.com). Les citations de conflits (poursuites pour copie) et d'expansion (intégration du shopping) illustrent l'impact plus large de la technologie de Perplexity (Source: www.reuters.com) (Source: www.reuters.com).

En résumé, Perplexity est à la pointe de la recherche pilotée par l'IA. Il ne se contente pas d'utiliser d'autres LLM, mais construit et affine activement les siens. Son mélange de modèles internes et externes, ainsi que son index de recherche, en font davantage un *moteur de méta-réponses* qu'un LLM monolithique. L'entreprise continue d'innover (par exemple, sa récente mise à niveau Sonar 3.3 (Source: www.perplexity.ai) et d'attirer l'attention des géants de la technologie (par exemple, les discussions avec Apple (Source: www.reuters.com). Les implications pour la recherche, l'éthique de l'IA et les médias numériques sont importantes, comme

ce rapport l'a détaillé. À l'avenir, il faudra observer comment Perplexity équilibre sa croissance (revenus publicitaires, nouvelles fonctionnalités) avec les contraintes légales et factuelles. Mais pour l'instant, il s'agit de l'un des exemples les plus avancés d'application des LLM au problème de la récupération d'informations en temps réel et fondées.

Références : Les informations ci-dessus sont tirées des propres publications de Perplexity (Source: www.perplexity.ai) (Source: www.perplexity.ai) (Source: www.perplexity.ai) (Source: www.perplexity.ai), d'articles de presse réputés (Reuters (Source: www.reuters.com) (Source: www.reuters.com), Axios (Source: www.axios.com), The Verge (Source: www.theverge.com), Tom’s Guide (Source: www.tomsguide.com) (Source: www.tomsguide.com), et d'analyses techniques (Source: primaryposition.com). Chaque affirmation de ce rapport est étayée par des citations spécifiques, comme indiqué.

Étiquettes: perplexity-ia, grand-modele-langage, pplx, sonar-llm, moteur-reponse-ia, architecture-llm, llm-open-source

---