Résumés de la vidéo : l’IA ultime est là, avec NotDiamond

Résumé Complet du Document avec Not Diamond : Version Perplexity

Not Diamond est une plateforme innovante qui utilise un routeur d'IA pour diriger automatiquement les requêtes vers le meilleur modèle de langage disponible, optimisant ainsi la qualité des réponses et réduisant les coûts et les latences.

1. **Principe de Fonctionnement**
Not Diamond repose sur un modèle meta qui combine plusieurs modèles de langage pour déterminer lequel est le plus approprié pour répondre à une question spécifique. Cela permet de maximiser la qualité des réponses tout en réduisant les coûts et les latences associés à l'utilisation de modèles moins performants pour des tâches simples.

2. **Modèles Intégrés**
La plateforme supporte une variété de modèles de pointe, tels que GPT-4, GPT-3.5, Claude 3.5, Gemini, et LLaMA 3.1. Cette intégration permet de choisir le modèle le plus adapté à chaque type de requête, améliorant ainsi la pertinence des réponses.

3. **Arena Mode**
Not Diamond propose un mode appelé "Arena Mode" qui permet de comparer les réponses de plusieurs modèles pour une même question. Cela permet aux utilisateurs de choisir la réponse la plus appropriée et de fournir des retours pour affiner les choix du modèle.

4. **Personnalisation des Réponses**
Les utilisateurs peuvent créer des instructions personnalisées pour les modèles, leur permettant de spécialiser les réponses en fonction de leur domaine d'expertise (par exemple, marketing digital ou finances). Ces instructions personnalisées améliorent la qualité et la pertinence des réponses.

5. **Avantages et Utilisation**
Not Diamond offre des avantages significatifs, notamment une grande simplicité d'utilisation, une interface familière pour les utilisateurs de chatbots comme ChatGPT, et une possibilité de comparer les performances de différents modèles. La plateforme est actuellement gratuite avec une limite de 100 000 requêtes par mois, ce qui en fait une ressource précieuse pour les développeurs et les utilisateurs.

6. **Bénéfices pour les Entreprises**
Not Diamond aide les entreprises à surmonter le dilemme coût-performance des modèles de langage. En utilisant des modèles spécialisés pour des tâches spécifiques, les entreprises peuvent réduire les coûts et les latences tout en améliorant la qualité des réponses. Cela a déjà permis à certaines entreprises de voir une amélioration de 10% dans la qualité des sorties et une réduction de 10% des coûts et latences.

7. **Technologie et Développement**
La technologie de Not Diamond repose sur un algorithme de classement des modèles qui utilise un ensemble de données d'évaluation pour déterminer le modèle le plus approprié pour chaque requête. La plateforme est en constante évolution, avec de nouvelles fonctionnalités prévues pour accélérer son adoption.

En résumé, Not Diamond est une innovation majeure dans le domaine des modèles de langage, offrant une solution flexible, efficace et gratuite pour répondre aux besoins diversifiés des utilisateurs, tout en maximisant la qualité des réponses et en réduisant les coûts associés.

```
model: Perplexity
latency: 2832ms
cost: $0.0041
```

Résumé Complet du Document sur Not Diamond : Version Gpt 4 Turbo

La transcription décrit un nouveau service d'IA appelé NotDiamond, qui promet d'utiliser les modèles d'IA les plus avancés disponibles actuellement, tels que Chat GPT-4, Clo 3.5, Gemini, Llama 3.1, et Perplexity pour fournir des réponses spécifiquement adaptées aux questions posées. Voici un résumé détaillé des points clés, des insights et des arguments présentés :

1. **Introduction de NotDiamond** : NotDiamond est décrit comme une IA révolutionnaire qui sélectionne le meilleur modèle pour répondre aux questions posées, en fonction de leur nature.

Elle est gratuite et vise à offrir des réponses optimisées et pertinentes.

- **Explication** : L'essence de NotDiamond repose sur sa capacité à évaluer la question de l'utilisateur et à choisir le modèle d'IA le plus approprié parmi une sélection des plus récentes et performantes. Cela permet d'obtenir des réponses plus précises et de haute qualité.

2. **Fonctionnement et interface utilisateur** : L'interface de NotDiamond est similaire à celle de Chat GPT, ce qui la rend facile à utiliser pour ceux qui sont déjà familiers avec les chatbots. L'interface simple et l'accès rapide aux différents modèles d'IA sont des points forts.

- **Explication** : L'interface utilisateur de NotDiamond est conçue pour être intuitive. Elle comprend une barre de chat et un menu facilement accessible pour choisir parmi les différents modèles d'IA, ce qui simplifie l'interaction et rend l'outil plus accessible.

3. **Flexibilité et personnalisation** : NotDiamond peut être personnalisé pour répondre à des besoins spécifiques en sélectionnant manuellement un modèle d'IA particulier. Cela inclut aussi un mode "Arena", où deux modèles peuvent être comparés en donnant deux réponses à une question pour que l'utilisateur choisisse la meilleure.

- **Explication** : Cette fonctionnalité ajoute une couche de personnalisation permettant aux utilisateurs d'expérimenter et de choisir le modèle qui répond le mieux à leurs besoins spécifiques, renforçant ainsi l'utilité et la flexibilité de l'IA.

4. **Analyse et résumé de documents** : NotDiamond peut analyser et résumer des documents, en choisissant le modèle le plus approprié pour cette tâche spécifique. Cela a été illustré par l'analyse d'un livre PDF de 131 pages, où l'IA a choisi le modèle Perplexity pour un résumé précis.

- **Explication** : La capacité de NotDiamond à analyser de grands volumes de texte et à en extraire des résumés pertinents montre son potentiel dans des applications académiques et professionnelles, offrant un gain de temps significatif pour ses utilisateurs.

5. **Gratuité et accessibilité** : Malgré ses fonctionnalités avancées, NotDiamond est offert gratuitement. Cette gratuité est soulignée comme un avantage majeur, surtout pour les utilisateurs qui ne souhaitent pas payer pour des abonnements à d'autres services d'IA.
   - **Explication** : La gratuité de NotDiamond la rend particulièrement attractive pour un large éventail d'utilisateurs, des étudiants aux professionnels, qui peuvent accéder aux technologies d'IA avancées sans barrière financière.

En résumé, NotDiamond se présente comme une innovation majeure dans le domaine de l'IA conversationnelle, offrant personnalisation, facilité d'utilisation, et accessibilité sans coût, tout en utilisant les technologies d'IA parmi les plus avancées et efficaces disponibles aujourd'hui.

model: GPT-4 Turbo
latency: 1916ms
cost: $0.1353

N.D.L.R
Je préfère, et NotDiamond, aussi la version de Perplexity beaucoup plus rapide (que la version turbo de ChatGpt 4, un comble) et plus complète.

La transcription originale, pour vous montrer à partir de quoi travaille les I.A

imagine l'a ultime unea qui se
reposerait sur les tout derniers modèles
les plus performants chat gpt4o Clo 3.5
Gemini liama 3.1 perplexity bref le
meilleur de ce qui se fait aujourd'hui
mais encore mieux imagine que ca puisse
faire appel au meilleur de ses modèles
en fonction de la question que tu lui
poses pour pouvoir te proposer la
réponse la plus pertinente possible et
bien écoute j'ai une bonne nouvelle
cette elle existe depuis un mois à peine
elle s'appelle not Diamond et elle est
complètement gratuite dans cette vidéo
je vais te montrer comment s'en servir
comment maximiser son potentiel et
pourquoi c'est sans doute l' a du moment
si tu dois en choisir une j'ai hâte de
te montrer tout ça alors ne perdons pas
plus de temps et on commence tout de
suite alors je suis très content de vous
montrer cette hier aujourd'hui parce que
vous allez voir elle est super simple
d'utilisation et elle permet de faire
toutes les choses qu'on peut attendre
dans chatbot au niveau de l'interface
c'est hyper simple si tu utilises chat
GPT régulièrement ou n'importe quel
autre il a tu dois pas être dépaysé je
trouve d'ailleurs que ça ressemble
beaucoup à chat GPT et honnêtement c'est
pas plus mal au moins on est dans un
écosystème qui nous est familier donc
comme d'habitude tu as la barre de chat
ici juste en bas tu peux poser tes
questions sur le côté comme d'habitude
le volet de gauche avec l'ensemble des
fenêtres de chat donc là en l'occurrence
on va démarrer on en a pas encore en
haut à droite quelques petits paramètres

on va le voir c'est très simple et c'est vraiment ce que j'adore avec cette a c'est que c'est très basique ils ont pas voulu faire quelque chose de compliqué mais ça répond parfaitement aux demandes il y a pas des choses hyper avancées mais dans 99 % des cas ça va largement répondre à tes besoins et surtout tout ça gratuitement alors évidemment ça c'est à l'heure où on se parle notre Diamond c'est très récent c'est sorti il y a à peine un mois il faut voir comment le modèle évolue dans la durée si à terme le modèle de facturation évolue mais aujourd'hui les développeurs ont communiqué sur le fait que c'était complètement gratuit et au niveau des limites il y en a quasiment pas je crois que tu peux monter jusqu'à 100 100000 requettes par mois donc autant dire que c'est quasiment liimité donc la principale force de cette hier on va la retrouver si on clique sur la roue juste ici en haut à droite et ici tu vas retrouver l'ensemble des modèles sur lesquel li a s'appuie donc on retrouve évidemment les tous derniers modèles de chat GPT GPT 4 turbo 4 mini clot 3 opus clot 3 ECU et Clo 3.5 son que j'adore personnellement on a Gini on a liama 3.1 et également perplexity ce qui va être hyper intéressant c'est que tu vas pouvoir poser ta question dans le chat je vais te montrer juste ici donc on firerme la fenêtre de droite on pose notre question et on va lui demander explique-moi simplement le mécanisme de fusion nucléaire comme sur Chat GPT comme sur Claude on lance la demande et ce qui va être hyper intéressant ça va être la manière dont il nous répond là jusqu'ici pas de gros changement par rapport à unea classique mais tout va jouer juste en bas juste ici au niveau

de la section modèle ici tu te rends compte que pour répondre à cette question sur le mécanisme de fusion nucléaire notre Diamond a voulu s'appuyer sur le modèle CLA 3.5 sonné et c'est ça qui est incroyable avec C tia c'est qu'en fonction de la question que tu vas lui poser en fonction de la tâche il va définir le modèle le plus pertinent pour pouvoir proposer la meilleure réponse donc là si on regarde les éléments qu'il nous a proposé globalement c'est intéressant il nous explique en six points les différents éléments du mécanisme avec les atomes LG la haute température et pression il faut surmonter la répulsion la fusion la libération d'énergie et les produits et pour proposer la réponse la plus qualitative il a estimé que Clo 3.5 sonné serait plus performant par exemple qu'un j JPT ou qu'un j mini ce qui est également intéressant c'est qu'ici tu vas voir la latence le temps que ça a pris pour répondre à cette question également le coût lié à cette requête qui encore une fois ne tit pas facturé et c'est pour te donner un ordre idée et pouvoir justement comparer ces différents modèles ça peut être utile pour imaginer combien ça tour écouter si tu étais passé par la p et surtout et c'est vraiment l'intérêt de notre Diamond c'est de pouvoir comparer les modèles entre eux donc donc maintenant on va poser une question d'un tout autre type et tu vas voir qu'il va sans doute utiliser un modèle différent pour ça on va ouvrir un nouveau chat juste ici encore une fois l'interface est hyper simple et c'est vraiment agréable et cette fois-ci on va lui demander d'analyser une image et on va voir si le modèle qu' l'utilise est différent donc

là ce que je vais charger c'est le PDF d'un livre que j'ai récupéré sur Internet et on va voir comment il analyse ce fichier et avec quel modèle il le fait du coup on clique sur la piè jointe on uplo le PDF et on va lui dire analyse ce livre et résume-moi les cinq points les plus important on lance la demande ça prend quelques secondes car il va falloir analyser l'ensemble du PDF et des 131 pages et voilà on a notre réponse avec les cinq points qui résument le livre donc 1 l'importance de l'instant présent 2 au-delà de l'identification mentale 3 le corps subtil et le sentir de lettre 4 le lâcher prise et l'acceptation et 5 les relations éclairées et ce qui est intéressant c'est que la réponse elle est satisfaisante mais surtout on a changé de modèle si tu regardes juste ici en bas on est passé sur perplexity et là du coup notre Diamond a estimé que perplexity était le modèle le plus performant pour répondre à notre question et du coup au-dessus de Clo 3.5 sonné Gini ou encore chat jpt4ro pour le coup au niveau de la latence on est beaucoup plus long mais parce qu'il fallait analyser l'ensemble du livre et par contre au niveau du coup ça reste toujours super satisfaisant ce qui va être intéressant maintenant ça va être de regarder ici l'ensemble des options qu'on a à notre disposition on a la possibilité de coller la réponse donc ça c'est assez classique c'est ce qu'on peut retrouver sur CLA chat GPT par exemple si tu cliques ici tu peux regénérer la réponse si elle te convient pas peut-être d'ailleurs qu'il peut utiliser un autre modèle de par lui-même donc là je viens de cliquer sur la flèche il me refait une toute nouvelle

réponse en l'occurrence il a utilisé le même modèle perplexity il doit vraiment estimer que c'est le plus performant pour ce genre de tâche et après ici tu as la possibilité de forcer l'utilisation d'un modèle par exemple tu peux le forcer utiliser clot 3.5 soné ou alors Jini 1.5 ou alors liama et cetera ça c'est siéta es vraiment convaincu que tel ou tel modèle est plus performant pour tel type de tche tu peux également vouloir faire ça par curiosité et dire ok quelle réponse aurait été fournie par Claude quelle réponse aurait été fournie par Yama quelle réponse aurait été fournie par chat GPT et ainsi te faire ton propre avis les dernières options qui sont disponibles c'est le pouce vers le haut et le pouce vers le bas donc comme sur Chat GPT à la différence prix que ici ça va permettre d'affiner le choix du modèle donc c'est-à-dire que si tu mets un pouce vers le haut ça va renforcer l'idée que perplexity est le bon modèle pour ce type de tâche en revanche si tu mets un pouce vers le bas il va comprendre que pour la prochaine fois pour des taches similaires il va falloir utiliser un autre modèle donc avec les pouces c'est également une manière de le signifier que tu préfèrees un modèle ou un autre à présent j'aimerais retourner dans les paramètres et te montrer deux trois petites choses intéressantes on va retourner en haut à droite on va cliquer ici comme tu l'as vu tout à l'heure tu peux évidemment décocher des modèles donc par exemple s'il y a des modèles que tu aimes pas comme Jini tu peux le décocher et t'assurer qu'il n'utilisera pas ce modèle pour l'ensemble des prochaines réponses tu peux désactiver les modèles les moins performants comme clot 3 opus

ou Clo 3coup par exemple ça peut être une idée mais des fois il y a des cas d'usages sur lesquels ils sont plus pertinents donc je te conseille de tous les conserver et faire confiance à notre Diamond et surtout ce que j'adore c'est l'Arena mode juste ici donc on va l'activer tu peux également l'activer en cliquant ici donc par exemple si on clique sur la croix tu peux l'activer le désactiver sur ce bouton c'est exactement la même chose quand tu regardes voilà on l'active on le désactive et ici en fait quand tu vas poser une question systématiquement not Diamond va utiliser deux as différentes et te proposer deux réponses distinctes et derrière toi tu vas pouvoir trancher pour la réponse qui te convient le mieux donc on va faire un test tout de suite on va quitter les settings et on va lui dire explique-moi simplement en quoi la carrière du Shen Bolt a été exceptionnelle on lance notre demande et là tu vois tout de suite vu qu'on a activé l'Arena mode avec le bouton juste ici on a deux réponses au lieu d'une quel modèle est utilisé pour chacune des réponses donc on va consulter chacune de nos réponses on a la première à gauche qui a l'air quand même plus détaillée avec les records éthes les performances consistance et longévité impact sur le sport personnalité sur la droite on a également la partie record domination olympique longévité au sommet performance sous pression charisme et impact sur l'athlétisme donc déjà on voit qu'on a CIN points d'un côté 6 points de l'autre on a un formatage qui est un peu plus intéressant sur la gauche moi je trouve que la réponse de gauche est plus intéressante tu tu peux décider que telle ou telle réponse est

meilleure tu peux également dire que les deux sont bonnes ou alors que les deux sont mauvaises en l'occurrence moi je vais choisir la réponse de gauche on clique sur la réponse et le modèle nous est dévoilé donc ici c'était perplexity de l'autre côté on avait clot 3.5 sonné donc tu vois que c'est intéressant pour comparer les IA aujourd'hui j'ai tendance à me dire que clot 3.5 sonné est lié à la plus performante à l'heure où on se parle bien sûr mais en fait tu te rends compte qu'en fonction de la demande perplexity peut être supérieur ou également chat gpt4ro ça dépend vraiment du cas d'usage et ce qui est intéressant avec cette arène à mode c'est que tu peux comparer l'aveugle et t'assurer de pas être biaisé dans ton jugement pour avoir les réponses plus qualitative derrière si on continue et qu'on lui demande comment s'entraîner-t-il on va voir ce qu'il nous propose encore une fois il est en train de charger deux réponses avec comme d'habitude une à gauche une à droite on voit que ce modèle là à droite était plus rapide on peut lui dire pour lui faire plaisir que les deux réponses sont bonnes et on va voir effectivement à droite on avait chat gpt4ro et à gauche clot 3.5 sonné donc on peut voir que Chad pt4o pour ce type de réponse est plus rapide propose un petit peu plus de détail là pour le coup à gauche avec Clare c'était plus concis ça te permet très rapidement de trancher quelle y a est la plus performante dernière chose avec notre Diamond et pas des moindres et je trouve que c'est une chose hyper intéressante parce que les développeurs ont voulu mettre seulement les fonctionnalités essentielles là en l'occurrence en a une je trouve on va

pouvoir y accéder encore une fois sur la roulette en haut à droite et comme tu l'as vu peut-être tout à l'heure c'est ici la partie custom system Prom qui est l'équivalent des customs instruction de chat GPT ici comme d'habitude avec cet outil c'est hyper pratique tu cliques sur plus pour ajouter une instruction personnalisée tu ajoutes un nom à l'instruction donc là on va dire qu'on a besoin de spécialiser lia comme si c'était un expert en marketing digital on met ça pour le nom pour le système prompt on va tout simplement copier-coller une instruction donc là en l'occurrence j'ai mis tué Unia spécialist marketing digital ton rô est d'assister les professionnels dans l'optimisation de leur stratégie marketing en ligne avec l'expertise et cetera on va pas tout lire mais globalement c'est des instructions assez simples on sauvegarde et là tout de suite tu vas te rendre compte que tu peux sélectionner n'importe quelle instruction personnalisée donc là on a notre première instruction qu'on va décider de sélectionner comme ça tu peux encréer une nouvelle en cliquant sur plus tu peux encréer une deuxième un expert finance par exemple tu mets un système prompte de ton choix voici l'instruction pour l'expert en finances donc tu es une IA spécialisé en finan t rôle est d'assister les professionnels et les particuliers dans la gestion l'analyse et la prise de décision fin entière on sauvegarde et là tu te rends compte que tu peux sélectionner ton expert en fonction de tes besoins donc là par exemple on va lui poser une question sur le marketing digital on conserve l'instruction expert marketing digital on clique sur la croix pour

aller poser notre question nouveau chat et on va lui dire fais-moi une stratégie simple pour lancer ma marque de vêtement pour sportif on lance notre demande et là forcément la réponse va être beaucoup plus adaptée que si on avait pas utilisé d'instruction personnalisée parce qu'on a précisé des éléments on voit que c'est quand même assez détaillé là je trouve que lia à gauche a l'air plus intéressante à priori on va quand même attendre la fin en l'occurrence à gauche on a une partie sur le budget et sur le calendrier ce qui a pas l'air d'être le cas à droite donc on va sélectionner cette réponse on est resté sur le mode Arena c'est pour ça qu'on a deux réponses et du coup là le meilleur modèle c'était liama 3.1 versus gpt4 turbo donc en l'occurrence j'ai trouvé que c'était le plus efficace et c'était également le moins cher on se rend compte que c'est quand même beaucoup moins cher que gpt4 turbo donc c'est également une force si on fait exactement la même requête on va la copier-coller comme ça en changeant l'instruction donc on va retourner sur les instructions personnalisées on passe en expert finance là j'imagine que ça va être beaucoup moins performant on va faire la même chose lancer la même requête on clique sur le lancement de la demande et là où est tout de suite on voit dès les premiers points qu'on parle de budget on parle de financement on parle de tarification alors que si on comparait plus haut sur un expert marketing plus que finance et ben on a pas du tout ces sujets financiers au début de la réponse là on parlait de public cible on parlait de marque de stratégie de contenu c'est complètement différent et forcément tu te rends

compte à quel point les instruction personnalisée joue un rôle prépondérant dans la réponse voilà j'espère que cette démonstration a pu te montrer à quel point not Diamond est tout simplement incroyable et surtout garde en tête que cet outil est complètement gratuit personnellement je trouve ça super parce que ça te permet d'accéder au tout derniers modèles les plus performants encore une fois gratuitement donc tu as pas besoin de payer même si le coût de la requête est indiqué à la fin de chaque réponse ça peut être super d'utiliser cette IA si tu as pas forcément envie d'aller payer un abonnement pro pour CLA ou chat GPT ou alors pour pouvoir les comparer en fonction des cas d'usage et derrière trancher sur laquelle est la plus utile pour toi et tes besoins si la vidéo t'a plu tu peux la liker tu peux également t'abonner à la chaîne et surtout dis-moi dans les commentaires ce que tu penses de not Diamond j'ai hâte d'avoir ton avis et de mon côté je te dis à la prochaine pour une nouvelle vidéo