Innovating Responsibly
For Mental Healthcare

LIVRE BLANC

Le sens sans l'identité

L'IA de pointe sur le texte clinique, sans aucun identifiant lisible hors de la clinique

Abex Beheshti

Cognitai Santé inc., Montréal, Québec, Canada

Version 1.0, novembre 2026

Brevet en instance

Test réalisé : novembre 2026

© 2026 Cognitai Santé inc. Tous droits réservés.
https://cognitai.org/fr/innovation/meaning-without-identity

Résumé

Les cliniques ont dû choisir entre les modèles d'IA les plus performants et la confidentialité des personnes que leurs textes décrivent. Cognitai Hermeta, une couche de confidentialité inventée chez Cognitai et faisant l'objet d'une demande de brevet, supprime ce choix. Nous l'avons comparée à Microsoft Presidio et à OpenAI Privacy Filter selon un protocole rédigé avant le début du test. Six modèles de pointe d'Anthropic et d'OpenAI ont répondu aux mêmes 80 questions sur un dossier clinique de 50 000 mots. Avec Cognitai Hermeta, aucun des 4 112 éléments identifiants du dossier n'est resté lisible, et les modèles ont conservé 99,6 % de la compréhension que permettait le texte original. Chacun des autres outils a sacrifié soit la confidentialité, soit la compréhension. Cognitai Hermeta n'a sacrifié ni l'une ni l'autre.

Principaux résultats

01 Aucun des 4 112 éléments identifiants consignés dans le dossier n'est resté lisible après Cognitai Hermeta. OpenAI Privacy Filter en a laissé 973 lisibles (23,7 %), Microsoft Presidio 185 (4,5 %).

02 Six modèles de pointe ont conservé 99,6 % de la compréhension que permettait le texte original : 78,8 bonnes réponses sur 80 en moyenne, contre 79,2 sur l'original.

03 Toute l'histoire est restée intacte. Les liens entre les personnes, la chronologie et les faits qui relient des parties éloignées d'un dossier ont obtenu les mêmes résultats que sur l'original ; qui a fait quoi, 115 sur 120 contre 119.

04 Chacun des autres outils a sacrifié l'une des deux mesures. OpenAI Privacy Filter a conservé 96,0 % de la compréhension en laissant lisible environ un identifiant sur quatre ; Microsoft Presidio en a laissé moins, mais la compréhension a chuté à 90,7 %, surtout sur les liens entre les personnes et sur qui a fait quoi.

05 Le test a été conçu pour être vérifié : texte et questions figés avant toute exécution des outils, lecteurs de deux fournisseurs, même traitement pour chaque outil, correction par code puis par deux juges à l'aveugle, et chaque résultat publié.

Comment citer (APA 7)

Beheshti, A. (2026). Le sens sans l'identité : l'IA de pointe sur le texte clinique, sans aucun identifiant lisible hors de la clinique [Livre blanc]. Cognitai Santé inc. https://cognitai.org/fr/innovation/meaning-without-identity

À propos de cet article

Il s'agit de la première édition publique. La comparaison a été réalisée les 9 et 10 novembre 2026, sur un dossier en anglais. Un deuxième dossier, en français et selon l'usage québécois, suivra selon le même protocole.

Déclaration

L'auteur a inventé la couche de confidentialité évaluée ici et est un fondateur de Cognitai Santé inc., l'entreprise qui la développe. Il s'agit d'un conflit d'intérêts. Cognitai Hermeta fait l'objet d'une demande de brevet (demande provisoire américaine 64/194,412). Aucun financement externe n'a été reçu. Aucune donnée de client n'a été utilisée. Des modèles de langage ont rédigé le texte du test et ses questions, l'ont lu et ont jugé une partie des réponses ; le protocole et chaque décompte ont été établis et vérifiés par l'auteur. Microsoft Presidio et OpenAI Privacy Filter sont des produits de leurs propriétaires, qui n'ont pas examiné cet article.

À propos de Cognitai Innovation

Cognitai Innovation est l'endroit où Cognitai Santé inc., à Montréal, transforme les limites les plus difficiles des soins en santé mentale en technologie. Elle est dirigée par Abex Beheshti, cofondateur et chef de l'innovation. Sa première invention publiée, Cognitai Hermeta, fait l'objet d'une demande de brevet : une couche de confidentialité qui permet à l'IA de pointe de lire tout le parcours d'un client, sans que personne puisse y être identifié hors de la clinique. Chaque invention suit la même règle : **conçue pour le clinicien, protégée dès la conception, et prouvée ouvertement**. Chaque article précise ce qui a été construit, à quoi il a été comparé, comment le test est resté équitable et où s'arrêtent ses preuves. Les mécanismes visés par une demande de brevet ou gardés confidentiels ne sont pas décrits.

Résumé

Les cliniques ont dû choisir entre les modèles de langage les plus performants, qui fonctionnent dans le nuage, et la confidentialité des personnes que leurs textes décrivent. Cognitai Hermeta, une couche de confidentialité inventée chez Cognitai et faisant l'objet d'une demande de brevet, a été conçue pour supprimer ce choix. Nous l'avons comparée à deux outils très utilisés, Microsoft Presidio et OpenAI Privacy Filter, à leurs réglages par défaut, selon un protocole rédigé avant le début du test. Six modèles de pointe de deux fournisseurs ont répondu aux mêmes 80 questions sur un dossier clinique, dans sa forme originale et dans chaque version protégée. Avec Cognitai Hermeta, aucun des 4 112 éléments identifiants consignés dans le dossier n'est resté lisible, et les modèles ont bien répondu à 78,8 questions sur 80 en moyenne, contre 79,2 sur l'original : 99,6 % de la compréhension que permettait le texte original. Ils ont gardé l'histoire : les liens entre les personnes, le moment où les choses se sont produites et ce qui relie une partie d'un parcours à une autre. OpenAI Privacy Filter a maintenu la compréhension à 76,0 (96,0 % de l'original), mais a laissé 973 identifiants lisibles (23,7 %). Microsoft Presidio en a laissé 185 lisibles (4,5 %), et la compréhension est tombée à 71,8 (90,7 %). Chacun des autres outils a sacrifié soit la confidentialité, soit la compréhension. Cognitai Hermeta n'a sacrifié ni l'une ni l'autre. Le test portait sur un seul dossier, en anglais, avec une seule lecture par modèle ; un deuxième dossier suivra.

Mots-clés : dépersonnalisation, confidentialité, grands modèles de langage, texte clinique, santé mentale, évaluation

1 Le compromis imposé aux cliniques

Les modèles de langage les plus puissants fonctionnent dans le nuage. Les textes qui leur permettraient d'aider en santé mentale (notes de séance, transcriptions, parcours d'un client sur plusieurs mois) comptent parmi les plus sensibles qui soient. Une note de séance dit qui a dit quoi, qui est lié à qui et ce qui s'est passé quand. Ce sont les faits dont un modèle a besoin pour être utile, et ce sont les mêmes faits qui identifient une personne.

Jusqu'ici, une clinique devait renoncer à quelque chose. Elle pouvait tenir ses textes à l'écart des modèles de pointe et renoncer à leurs capacités. Elle pouvait faire tourner un modèle plus petit sur place et renoncer à la qualité. Elle pouvait supprimer ou masquer les identifiants et renoncer à l'histoire : un lecteur ne peut plus distinguer deux personnes, suivre une même personne dans un dossier ni remettre les événements en ordre. Ou elle pouvait garder assez d'identifiants pour que l'histoire tienne, et renoncer à la confidentialité.

Compter les identifiants qu'un outil retire en dit trop peu sur ce à quoi il a renoncé. Un outil peut tout retirer et ne rien laisser qui vaille la peine d'être lu. Un outil peut aussi préserver la compréhension en laissant des identifiants en place. Des bancs d'essai en anonymisation de textes ont plaidé pour mesurer ensemble la protection et l'utilité de ce qui reste (Pilán et al., 2022). Ce test mesure les deux, avec des modèles de pointe comme lecteurs, parce que ce sont les lecteurs pour lesquels Cognitai Hermeta a été conçue.

2 Le sens conservé, l'identité retirée

Sur un dossier clinique, lu par six modèles de pointe de deux fournisseurs :

Aucun identifiant n'était lisible. Aucun des 4 112 éléments identifiants consignés dans le dossier n'est resté lisible dans le texte reçu par les modèles.

L'histoire a été conservée. Les modèles ont bien répondu à 78,8 questions sur 80 en moyenne, contre 79,2 sur l'original : 99,6 % de la compréhension de l'original. Deux des six ont obtenu le même score sur les deux versions. Les questions sur les liens entre les personnes, l'ordre des événements et les faits qui relient une partie du parcours à une autre ont obtenu les mêmes résultats que sur l'original, et celles sur qui a fait quoi presque autant (115 sur 120, contre 119).

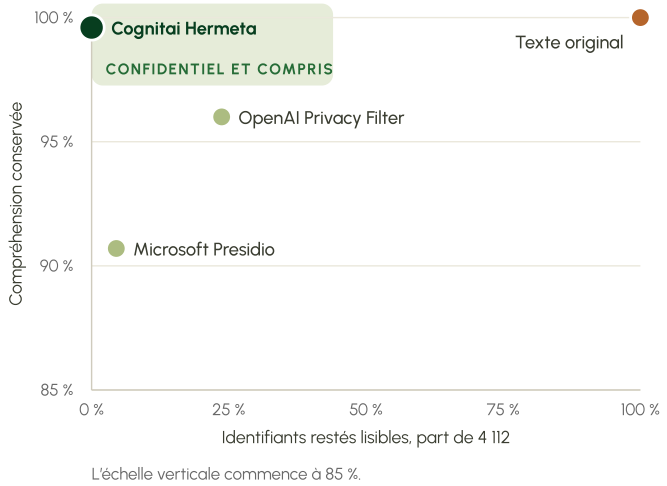


Figure 1. Confidentiel et compris. Chaque version du dossier placée selon les deux mesures à la fois : les identifiants restés lisibles à l'horizontale, la compréhension conservée à la verticale. Le coin supérieur gauche est là où une clinique doit être. Seule Cognitai Hermeta l'a atteint.

Source: Test comparatif de Cognitai, novembre 2026.

Les modèles ont compris toute l'histoire clinique sans y voir personne.

---

3 Ce que cela change pour les soins

Les questions les plus utiles d'un clinicien portent sur tout un parcours : ce qui a changé depuis le printemps, qui est mentionné chaque fois que le sommeil se dégrade, la semaine où le conflit a commencé. Ce sont les questions qui échouent quand les identifiants sont supprimés, et celles qui exposent un client quand ils ne le sont pas.

Avec Cognitai Hermeta, un clinicien peut confier tout le parcours au meilleur modèle disponible. La réponse revient avec chaque élément identifiant à sa place, et se lit comme ses propres notes. Aucun serveur ni aucun modèle infonuagique entre les deux n'a besoin de savoir qui est le client.

---

4 Ce que fait Cognitai Hermeta

Cognitai Hermeta a été inventée par Abex Beheshti chez Cognitai et fait l'objet d'une demande de brevet (demande provisoire américaine 64/194,412). Cet article présente ce qu'elle fait et ce qu'elle a accompli, pas son fonctionnement ; la méthode fait l'objet de la demande de brevet, et certaines parties restent confidentielles.

Où elle fonctionne

Elle agit avant tout envoi. Elle fonctionne sur les appareils autorisés de la clinique et sur les appareils du client qui y a consenti, avant qu'un texte ne les quitte.

Elle a été conçue pour de petits appareils portables. Elle fonctionne sur des appareils aussi petits qu'un téléphone, pour que la même protection suive les soins partout.

En cas de doute, rien ne part. Un texte qu'elle ne peut pas traiter n'est pas envoyé.

Qui peut lire l'identité

L'identité reste aux personnes qui y ont droit. Qui est le client reste sur les appareils autorisés de la clinique et sur les appareils du client qui y a consenti. Aucun serveur ni aucun modèle infonuagique ne peut le savoir.

La réponse revient entière. La réponse du modèle revient sur ces appareils avec chaque élément identifiant à sa place, et seulement là.

Un texte dépersonnalisé reste soumis aux lois sur la protection des renseignements personnels, et Cognitai Hermeta ne remplace ni le consentement, ni les contrats, ni les obligations propres à la clinique. Ce qu'elle change, c'est ce qu'une clinique doit envoyer pour obtenir la meilleure réponse : dans ce test, un texte sans aucun identifiant lisible.

5 La comparaison

Les deux mesures, côte à côte

Les deux mesures comptent ; la figure 2 les montre donc côte à côte.

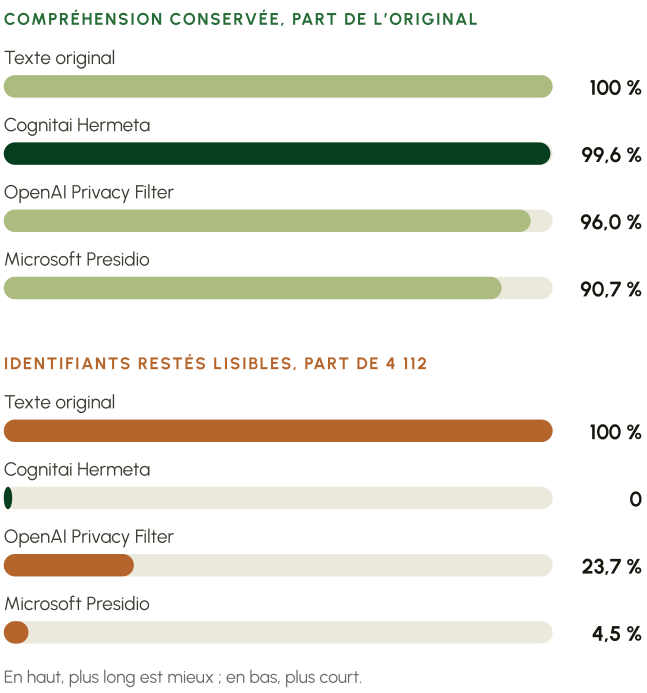


Figure 2. Compréhension conservée, identifiants restés lisibles. Pour chaque version du dossier, la part des bonnes réponses de l'original que les six lecteurs ont encore données (bonnes réponses de tous les lecteurs divisées par les 475 obtenues sur l'original), à côté des éléments identifiants restés lisibles. Seule Cognitai Hermeta est en tête de la première mesure et à zéro sur la seconde.

Source: Test comparatif de Cognitai, novembre 2026.

Cognitai Hermeta est le seul outil testé à n'avoir laissé aucun identifiant lisible, et sa compréhension est restée à 0,4 réponse sur 80 de l'original, en moyenne. Les deux autres outils, à leurs réglages par défaut et sur ce dossier, ont chacun sacrifié l'une des deux mesures. OpenAI Privacy Filter est resté plus près de l'original pour la compréhension, tout en laissant lisible environ un identifiant sur quatre. Microsoft Presidio a laissé moins d'identifiants lisibles et a perdu le plus de compréhension.

Lecteur par lecteur

	Haiku 5.5	Sonnet 5.5	Opus 5.5	GPT-5.6 Luna	GPT-6.1 Sol	GPT-6.1 Terra	Moyenne
Texte original	80	80	80	76	80	79	79,2
Cognitai Hermeta	**78**	**80**	**80**	**77**	**79**	**79**	**78,8**
OpenAI Privacy Filter	74	77	78	72	79	76	76,0
Microsoft Presidio	72	76	76	56	75	76	71,8

Modèles Claude d'Anthropic ; modèles GPT d'OpenAI, effort de raisonnement faible.

Figure 3. Bonnes réponses sur 80, par lecteur. Toutes les cases du test. Sur le texte original, un lecteur, GPT-5.6 Luna, a obtenu 76 ; un écart d'une réponse reste dans ce qu'une reprise pourrait changer.

Source: Test comparatif de Cognitai, novembre 2026.

Question par question

Là où les autres outils perdent, ils perdent sur les questions qu'un clinicien pose à tout un parcours : qui a fait quoi, les liens entre les personnes et le moment où les choses se sont produites. Cognitai Hermeta a égalé l'original sur les liens, la chronologie et les faits répartis entre chapitres.

	Original	Cognitai Hermeta	Privacy Filter	Presidio
Qui (sur 120)	119	115	115	108
Liens (sur 120)	119	119	113	103
Chronologie (sur 90)	89	90	84	80
Entre chapitres (sur 90)	88	89	84	81
Contrôles (sur 60)	60	60	60	59

Contrôles : sans identifiant, ils montrent si un outil abîme le texte.

Figure 4. Bonnes réponses par type de question. Les six lecteurs ensemble. Cognitai Hermeta a égalé l'original sur les liens, la chronologie et les faits répartis entre chapitres ; les autres outils ont perdu des réponses sur chaque type de question qui demande un identifiant.

Source: Test comparatif de Cognitai, novembre 2026.

6 Un test conçu pour être équitable

Un résultat ne vaut que ce que vaut le test qui le produit. Le protocole a été rédigé pour que personne, son auteur compris, ne puisse orienter le résultat.



Figure 5. Le déroulement du test. Sept étapes, dans l'ordre. Le texte, les questions et leurs corrigés ont été figés avant qu'un outil ou un modèle y touche ; chaque outil a été traité de la même façon ; chaque résultat est publié.

Source: Test comparatif de Cognitai, novembre 2026.

Avant qu'un modèle lise le texte

Texte figé avant tout outil. La méthode a été mise par écrit avant le début du test. Le texte, les 80 questions et leurs corrigés ont été versionnés avant l'application de tout outil, et leurs empreintes cryptographiques ont été consignées avant qu'un modèle ne les lise. Rien n'a été modifié ensuite.

Des questions rédigées à l'écart de la couche. Les questions et les corrigés ont été rédigés par un agent distinct, à partir du texte seul, puis vérifiés par un autre lecteur qui y a répondu à partir de l'original sans voir les corrigés. Ils couvrent cinq types de compréhension : qui a fait ou dit quelque chose (20), les liens entre les personnes (20), l'ordre et le moment des événements (15), les faits qui demandent plus d'une partie du parcours (15), et des contrôles qui ne demandent aucun identifiant (10). Les contrôles montrent si un outil abîme un texte qui ne contient aucun identifiant.

Chaque identifiant connu d'avance. Les éléments identifiants du dossier ont été répertoriés avant qu'un modèle ne lise le texte ; la confidentialité a donc été comptée exactement, et non estimée, et de la même façon pour chaque outil.

Pendant la lecture

Des lecteurs de deux fournisseurs. Six modèles de pointe, trois d'Anthropic et trois d'OpenAI, ont lu chaque version dans des sessions neuves, avec les mêmes consignes et sans contexte partagé. Aucun résultat ne repose sur un seul modèle ni sur un seul fournisseur.

Le même traitement pour chaque outil. Les réponses de chaque outil ont été rendues lisibles de la même façon avant la correction. Aucun outil n'a été ajusté pour ce texte, Cognitai Hermeta comprise.

Après les réponses

Le code d'abord, puis des juges à l'aveugle. Les réponses ont été corrigées selon les corrigés par du code, avec une seule règle pour chaque version. La règle et les consignes des juges ont été rédigées après la collecte des réponses et avant qu'un score ne soit vu. Les 306 réponses que le code ne pouvait pas trancher sont allées à deux juges, qui ne voyaient que la question, le corrigé et la réponse rendue lisible, jamais la version ni le lecteur. Une réponse n'était jugée bonne que si les deux juges s'entendaient. Ils ont divergé sur 1 réponse sur 306.

Chaque case publiée. Dans quelques cases, Cognitai Hermeta a obtenu une réponse de plus que l'original (chronologie, faits entre chapitres et GPT-5.6 Luna). Avec une seule lecture par modèle, un écart d'une réponse reste dans ce qu'une reprise pourrait changer ; nous les lisons comme égaux à l'original, pas meilleurs. L'original est le plafond.

Ouvert à la contestation. Les autres outils ont fonctionné à leurs réglages par défaut, avec seulement la petite adaptation nécessaire pour corriger leurs réponses de la même façon. D'autres réglages pourraient donner d'autres scores. Nous invitons leurs développeurs à reprendre le protocole.

---

7 Protocole, modèles et outils

Texte. Un dossier clinique d'environ 50 000 mots en anglais, en dix chapitres, rédigé pour cette évaluation par Claude Opus 5.5 à partir de personnages, de lieux et d'une chronologie fixés d'abord ; les questions ont été rédigées et vérifiées par des agents Claude Opus 5.5 distincts. Il ne contient aucune information sur une personne réelle. Ses éléments identifiants ont été répertoriés avant qu'un modèle ne le lise (140 éléments distincts, 4 112 occurrences).

Lisible. Un identifiant était compté comme lisible quand sa chaîne apparaissait comme mot entier dans le texte protégé, au plus autant de fois qu'elle apparaît dans l'original.

Lecteurs. Claude Haiku 5.5, Claude Sonnet 5.5 et Claude Opus 5.5 (Anthropic), comme agents Claude Code aux réglages par défaut, le 9 novembre 2026 ; GPT-5.6 Luna (9 novembre), GPT-6.1 Sol et GPT-6.1 Terra (10 novembre) (OpenAI), par Codex avec un effort de raisonnement faible. Juges : Claude Sonnet 5.5 et Claude Opus 5.5, qui ne voyaient que la question, le corrigé et la réponse rendue lisible.

Réponses. 1 920 au total (quatre versions, six lecteurs, 80 questions). Chaque réponse a été rendue lisible selon une même règle, puis comparée au corrigé par du code ; les 306 que le code ne pouvait pas trancher ont été jugées bonnes ou mauvaises par les deux juges.

Outils comparés, exécutés localement. Microsoft Presidio (presidio-analyzer et presidio-anonymizer 2.2.366, licence MIT) avec spaCy 3.8.18 et en_core_web_lg 3.8.0, reconnaisseurs prédéfinis, tous les types d'entités, seuil de score 0,0. OpenAI Privacy Filter (openai/privacy-filter, révision Hugging Face 7ffa9a04, licence Apache-2.0), toutes les étiquettes, au point de fonctionnement par défaut de sa fiche de modèle. Dans ce test, Cognitai Hermeta fonctionnait sur un seul ordinateur, et rien d'identifiant ne l'a quitté.

---

8 La suite

Ce test est une étape d'un programme de preuves plus long.

Un deuxième dossier. En français et selon l'usage québécois, selon le même protocole figé.

Un banc d'essai public. Un jeu de recherche public avec un partage publié, évalué une seule fois sur sa partie réservée, comme le Text Anonymization Benchmark (Pilán et al., 2022).

Une revue externe. Du protocole, par des personnes qui n'ont aucun intérêt dans le résultat.

De vrais textes cliniques, avec consentement. Dans une clinique, là où la loi et le consentement le permettent.

9 Limites

Un dossier, une langue, une lecture. Un seul dossier, en anglais, rédigé pour cette évaluation par un modèle de langage ; l'écriture clinique réelle est plus désordonnée. Chaque modèle a lu chaque version une seule fois ; l'écart entre deux lectures n'a donc pas été mesuré.

Ce que la confidentialité a compté. La confidentialité est comptée sur les 4 112 éléments identifiants répertoriés d'avance. D'autres détails pouvant désigner une personne n'ont pas été comptés, et le test ne mesure pas si une personne pourrait être reconnue à partir d'une combinaison de détails qui ne sont pas des identifiants à eux seuls, comme un métier, un lieu de travail et une situation familiale.

Des modèles dans plus d'un rôle. Le dossier et ses questions ont été rédigés par Claude Opus 5.5, qui était aussi l'un des lecteurs et l'un des juges, et les juges viennent du même fournisseur que trois des lecteurs. L'aveugle, les deux juges et les questions de contrôle limitent cet effet ; ils ne l'éliminent pas.

Correction et réglages. La règle de correction a été rédigée après les réponses et n'a pas été enregistrée d'avance. Les autres outils ont fonctionné à leurs réglages par défaut, et les modèles GPT avec un effort de raisonnement faible seulement.

10 Conclusion

Sur un dossier clinique lu par six modèles de pointe, Cognitai Hermeta n'a laissé lisible aucun des 4 112 éléments identifiants et a maintenu la compréhension au niveau du texte original, alors que chacun des autres outils testés a sacrifié l'une ou l'autre. Le sens sans l'identité est possible. Les prochains dossiers montreront jusqu'où cela tient.

Références

Pilán, I., Lison, P., Øvrelid, L., Papadopoulou, A., Sánchez, D., & Batet, M. (2022). The Text Anonymization Benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics, 48*(4), 1053–1101. https://doi.org/10.1162/coli_a_00458