cognitai

Innovating Responsibly
For Mental Healthcare

WHITE PAPER

Meaning without identity

Frontier AI on clinical text, with no identifier readable outside the clinic

Abex Beheshti

Cognitai Health Inc., Montréal, Québec, Canada

Version 1.0, November 2026

Patent pending

Test conducted: November 2026

© 2026 Cognitai Health Inc. All rights reserved.
https://cognitai.org/innovation/meaning-without-identity

Summary

Clinics have been forced to choose between the most capable AI models and the privacy of the people their text describes. Cognitai Hermeta, a patent-pending privacy layer invented at Cognitai, removes that choice. We tested it against Microsoft Presidio and OpenAI Privacy Filter under a protocol written before the test began. Six frontier models from Anthropic and OpenAI answered the same 80 questions about a 50,000-word clinical case file. With Cognitai Hermeta, none of the 4,112 identifying details in the file remained readable, and the models kept 99.6% of the understanding the original text allowed. Each of the other tools gave up privacy or understanding. Cognitai Hermeta gave up neither.

Key findings

01 None of the 4,112 identifying details recorded in the case file remained readable after Cognitai Hermeta. OpenAI Privacy Filter left 973 readable (23.7%); Microsoft Presidio left 185 (4.5%).

02 Six frontier models kept 99.6% of the understanding the original text allowed: 78.8 correct answers of 80 on average, against 79.2 on the original.

03 The whole story survived. How people are related, when things happened and facts that join distant parts of a history were answered at the level of the original; who did what, at 115 of 120 against 119.

04 Each of the other tools gave up one of the two measures. OpenAI Privacy Filter kept 96.0% of understanding while leaving about one identifier in four readable; Microsoft Presidio left fewer readable, and understanding fell to 90.7%, mostly on how people are related and who did what.

05 The test was built to be checked: text and questions fixed before any tool ran, readers from two providers, the same treatment for every tool, scoring by code and then by two blind judges, and every result reported.

How to cite (APA 7)

Beheshti, A. (2026). Meaning without identity: Frontier AI on clinical text, with no identifier readable outside the clinic [White paper]. Cognitai Health Inc. https://cognitai.org/innovation/meaning-without-identity

About this paper

This is the first public edition. The comparison ran on 9 and 10 November 2026, on one case file in English. A second file, in French and in Québec usage, comes next under the same protocol.

Disclosure

The author invented the privacy layer evaluated here and is a founder of Cognitai Health Inc., the company that develops it. This is a conflict of interest. Cognitai Hermeta is patent pending (US provisional application 64/194,412). No external funding was received. No client data was used. Language models wrote the test text and its questions, read the text, and judged part of the answers; the protocol and every count were set and checked by the author. Microsoft Presidio and OpenAI Privacy Filter are products of their owners, who did not review this paper.

About Cognitai Innovation

Cognitai Innovation is where Cognitai Health Inc., in Montréal, turns the hardest limits of mental healthcare into technology. It is led by Abex Beheshti, Co-founder & Chief Innovation Officer. Its first published invention, Cognitai Hermeta, is patent pending: a privacy layer that lets frontier AI read a client's whole history while no one in it can be identified outside the clinic. Every invention follows the same rule: **built for the clinician, protected by design, and proven in the open**. Each paper says what was built, what it was measured against, how the test was kept fair, and where its evidence ends. Mechanisms covered by patent applications or kept confidential are not described.

Abstract

Clinics have been forced to choose between the most capable language models, which run in the cloud, and the privacy of the people their text describes. Cognitai Hermeta, a patent-pending privacy layer invented at Cognitai, was built to remove that choice. We tested it against two widely used tools, Microsoft Presidio and OpenAI Privacy Filter, at their default settings, under a protocol written before the test began. Six frontier models from two providers answered the same 80 questions about one clinical case file, in its original form and in each protected version. With Cognitai Hermeta, none of the 4,112 identifying details recorded in the file remained readable, and the models answered 78.8 of 80 questions correctly on average, against 79.2 on the original: 99.6% of the understanding the original text allowed. They kept the story: how people are related, when things happened, and what links one part of a history to another. OpenAI Privacy Filter kept understanding at 76.0 (96.0% of the original) but left 973 identifiers readable (23.7%). Microsoft Presidio left 185 readable (4.5%), and understanding fell to 71.8 (90.7%). Each of the other tools gave up privacy or understanding. Cognitai Hermeta gave up neither. The test used one file in English and one run per model; a second file comes next.

Keywords: de-identification, privacy, large language models, clinical text, mental healthcare, evaluation

1 The trade-off clinics were left with

The strongest language models run in the cloud. The text that would let them help in mental healthcare (session notes, transcripts, a client's history across months) is among the most sensitive text there is. A session note says who said what, who is related to whom, and what happened when. Those are the facts a model needs in order to be useful, and they are the same facts that identify a person.

Until now, a clinic had to give something up. It could keep its text away from frontier models and give up their capability. It could run a smaller model on site and give up quality. It could delete or mask the identifiers and give up the story: a reader can no longer tell two people apart, follow one person across a file, or put events in order. Or it could keep enough identifiers for the story to survive, and give up privacy.

Counting the identifiers a tool removes says too little about which of these it has given up. A tool can remove everything and leave nothing worth reading. A tool can also keep understanding by leaving identifiers in place. Benchmarks for text anonymization have argued for measuring

protection and the usefulness of what remains together (Pilán et al., 2022). This test measures both, with frontier models as the readers, because they are the readers Cognitai Hermeta was built for.

2 Meaning kept, identity removed

On one clinical case file, read by six frontier models from two providers:

No identifier was readable. None of the 4,112 identifying details recorded in the file remained readable in the text the models received.

The story was kept. The models answered 78.8 of 80 questions correctly on average, against 79.2 on the original: 99.6% of the original's understanding. Two of the six scored the same on both. Questions about how people are related, the order of events, and facts that join one part of the history to another were answered as well as on the original, and questions about who did what nearly so (115 of 120, against 119).

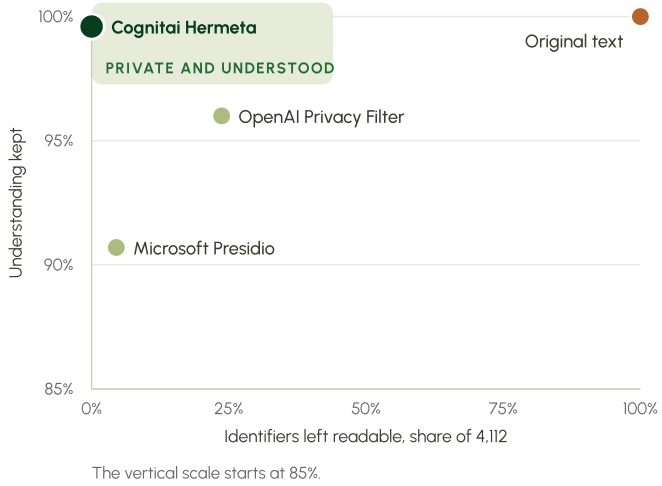


Figure 1. Private and understood. Each version of the case file placed by both measures at once: identifiers left readable across, understanding kept up. The top left corner is where a clinic needs to be. Only Cognitai Hermeta reached it.

Source: Cognitai comparison test, November 2026.

The models understood the full clinical story while seeing no one in it.

3 What it means for care

A clinician's most useful questions are about a whole history: what has changed since the spring, who is mentioned whenever sleep gets worse, which week the conflict began. Those are the questions that fail when identifiers are deleted, and the questions that expose a client when they are not.

With Cognitai Hermeta, a clinician can bring the strongest available model to the whole history. The answer comes back with every identifying detail in place, so it reads like the clinician's own notes. No server and no cloud model in between needs to know who the client is.

4 What Cognitai Hermeta does

Cognitai Hermeta was invented by Abex Beheshti at Cognitai and is patent pending (US provisional application 64/194,412). This paper reports what it does and what it achieved, not how it works; the method is the subject of the patent application, and parts of it are kept confidential.

Where it runs

It acts before anything is sent. It runs on the clinic's authorized devices and the client's consented devices, before any text leaves them.

It was made for small portable devices. It runs on devices as small as a phone, so the same protection goes wherever care happens.

It fails closed. Text it cannot process is not sent.

Who can read identity

Identity stays with the people entitled to it. Who the client is stays on the clinic's authorized devices and the client's consented devices. No server and no cloud model can tell.

The answer comes back whole. The model's answer returns to those devices with every identifying detail in place, and only there.

De-identified text is still subject to privacy law, and Cognitai Hermeta does not replace consent, contracts or a clinic's own obligations. What it changes is what a clinic has to send in order to get the best answer: on this test, text with no readable identifier.

5 How it compares

Both measures, side by side

Both measures matter, so Figure 2 shows them side by side.

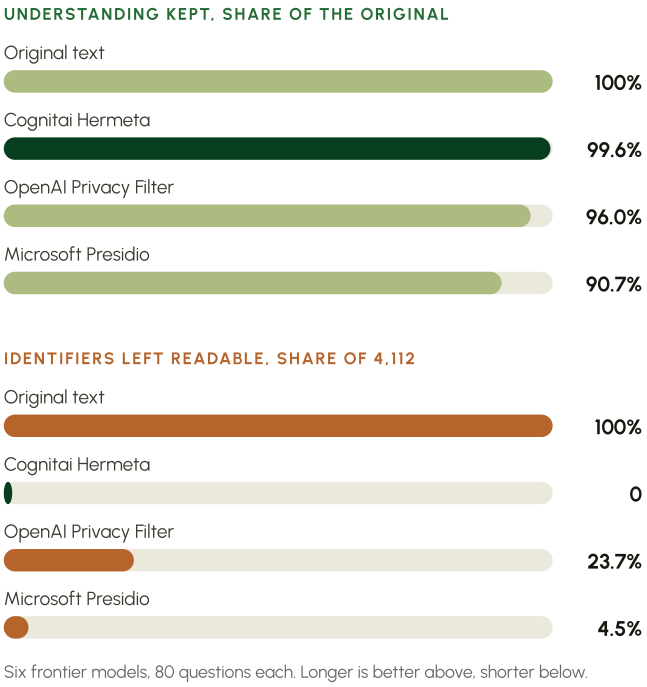


Figure 2. Understanding kept, identifiers left readable. For each version of the case file, the share of the original's correct answers the six readers still gave (correct answers across all readers divided by the 475 on the original), beside the identifying details left readable. Only Cognitai Hermeta is at the top of the first measure and at zero on the second.

Source: Cognitai comparison test, November 2026.

Cognitai Hermeta was the only tool tested that left no identifier readable, and its understanding stayed within 0.4 answers of 80 of the original on average. The two other tools, at their default settings and on this file, each gave up one of the two measures. OpenAI Privacy Filter stayed closer

to the original in understanding while leaving about one identifier in four readable. Microsoft Presidio left fewer identifiers readable and lost the most understanding.

Reader by reader

	Haiku 5.5	Sonnet 5.5	Opus 5.5	GPT-5.6 Luna	GPT-6.1 Sol	GPT-6.1 Terra	Mean
Original text	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

Claude models from Anthropic; GPT models from OpenAI, reasoning effort low.

Figure 3. Correct answers of 80, by reader. Every cell of the test. On the original text one reader, GPT-5.6 Luna, scored 76; differences of one answer are within what a repeat run could change.

Source: Cognitai comparison test, November 2026.

Question by question

Where the other tools lose, they lose on the questions a clinician asks of a whole history: who did what, how people are related, and when things happened. Cognitai Hermeta matched the original on relations, timelines and facts joined across chapters.

	Original	Cognitai Hermeta	Privacy Filter	Presidio
Who (of 120)	119	115	115	108
Relations (of 120)	119	119	113	103
Timelines (of 90)	89	90	84	80
Across chapters (of 90)	88	89	84	81
Controls (of 60)	60	60	60	59

Controls need no identifier: they show whether a tool damages the text itself.

Figure 4. Correct answers by kind of question. Six readers together. Cognitai Hermeta matched the original on relations, timelines and facts across chapters; the other tools lost answers on every kind of question that needs an identifier.

Source: Cognitai comparison test, November 2026.

6 How the test was kept fair

A result is only as strong as the test behind it. The protocol was written so that no one, including its author, could steer the outcome.

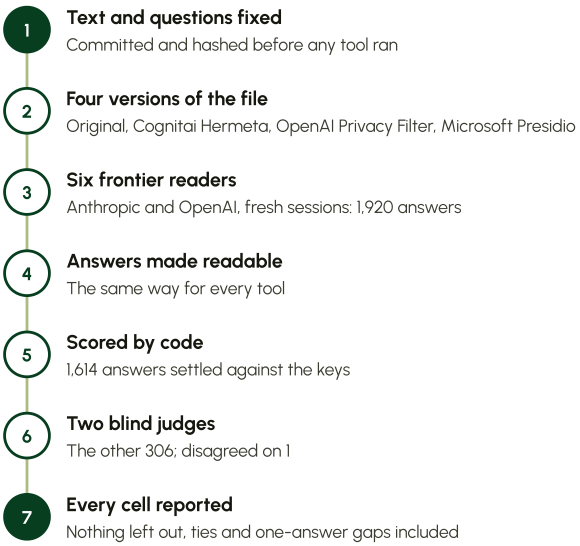


Figure 5. How the test ran. Seven steps, in order. The text, the questions and their keys were fixed before any tool or model touched them; every tool was treated the same; every result is reported.

Source: Cognitai comparison test, November 2026.

Before any model read the text

Text fixed before any tool ran. The method was written down before the test began. The text, the 80 questions and their answer keys were committed to version control before any tool was applied, and their cryptographic hashes were recorded before any model read them. Nothing was changed afterwards.

Questions written apart from the layer. The questions and keys were written by a separate agent, from the text alone, and checked by a further reader that answered them from the original with the keys hidden. They cover five kinds of understanding: who did or said something (20), how people are related (20), the order and timing of events (15), facts that need more than one part of the history (15), and controls that need no identifier at all (10). The controls show whether a tool damages text that holds no identifiers.

Every identifier known in advance. The identifying details in the file were listed before any model read the text, so privacy was counted exactly, not estimated, and counted the same way for every tool.

While the models read

Readers from two providers. Six frontier models, three from Anthropic and three from OpenAI, read every version in fresh sessions, with the same instructions and no shared context. No result rests on one model or one provider.

The same treatment for every tool. Every tool's answers were brought back to readable form in the same way before scoring. No tool was tuned for this text, Cognitai Hermeta included.

After the answers

Code first, then blind judges. Answers were scored against the keys by code, with one rule for every version. The rule and the judges' instructions were written after the answers were collected and before any score was seen. The 306 answers that code could not settle went to two judges who saw only the question, the key and the answer in readable form, never the version or the reader. An answer counted as correct only when both judges agreed. They disagreed on 1 of 306.

Every cell reported. In a few cells Cognitai Hermeta scored one answer above the original (timelines, facts across chapters, and GPT-5.6 Luna). With one run per model, a difference of one answer is within what a repeat could change; we read these as equal to the original, not better. The original is the ceiling.

Open to challenge. The other tools ran at their default settings, with only the small adaptation needed to score their answers in the same way. Tuned settings may score differently. We invite their developers to rerun the protocol.

---

7 Protocol, models and tools

Text. One clinical case file of about 50,000 words in English, in ten chapters, written for this evaluation by Claude Opus 5.5 from a cast, places and a timeline set first; the questions were written and checked by separate Claude Opus 5.5 agents. It contains no information about any real person. Its identifying details were listed before any model read it (140 distinct details, 4,112 occurrences).

Readable. An identifier counted as readable when its string appeared as a whole word in the protected text, counted at most as many times as it appears in the original.

Readers. Claude Haiku 5.5, Claude Sonnet 5.5 and Claude Opus 5.5 (Anthropic), as Claude Code agents at default settings, on 9 November 2026; GPT-5.6 Luna (9 November), GPT-6.1 Sol and GPT-6.1 Terra (10 November) (OpenAI), through Codex at reasoning effort low. Judges: Claude Sonnet 5.5 and Claude Opus 5.5, shown only the question, the key and the answer in readable form.

Answers. 1,920 in all (four versions, six readers, 80 questions). Each answer was brought back to readable form by one shared rule, then matched against the key by code; the 306 the code could not settle were marked right or wrong by both judges.

Comparison tools, run locally. Microsoft Presidio (presidio-analyzer and presidio-anonymizer 2.2.366, MIT licence) with spaCy 3.8.18 and en_core_web_lg 3.8.0, predefined recognizers, all entity types, score threshold 0.0. OpenAI Privacy Filter (openai/privacy-filter, Hugging Face revision 7ffa9a04, Apache-2.0 licence), all labels, the default operating point of its model card. In this test, Cognitai Hermeta ran on one computer, and nothing identifying left it.

8 What comes next

This test is one step in a longer program of evidence.

A second case file. In French and in Québec usage, under the same frozen protocol.

A public benchmark. A public research set with a published split, scored once on its held-out part, such as the Text Anonymization Benchmark (Pilán et al., 2022).

An outside review. Of the protocol, by people who have no stake in the result.

Real clinical text, with consent. Inside a clinic, where the law and the consent allow.

9 Limitations

One file, one language, one run. One case file, in English, written for this evaluation by a language model; real clinical writing is messier. Each model read each version once, so the spread between runs was not measured.

What privacy counted. Privacy is counted on the 4,112 identifying details listed in advance. Other details that can point to a person were not counted, and the test does not measure whether a person could be recognised from a combination of details that are not identifiers on their own, such as a job, a workplace and a family situation.

Models in more than one role. The case file and its questions were written by Claude Opus 5.5, which was also one of the readers and one of the judges, and the judges come from the same provider as three readers. Blinding, two judges and the control questions limit this; they do not remove it.

Scoring and settings. The scoring rule was written after the answers and was not registered in advance. The other tools ran at their default settings, and the GPT models at low reasoning effort only.

10 Conclusion

On one clinical case file read by six frontier models, Cognitai Hermeta left none of 4,112 identifying details readable and kept understanding at the level of the original text, while each of the other tools tested gave up one or the other. Meaning without identity is possible. The next files will show how far it holds.

References

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