PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

First Named Inventor:	Gregory Statton	Confirmation No.:	4998
Serial No.:	18/618,695	Group Art Unit:	2152
Filed:	March 27, 2024	Customer No.:	190019
Examiner:	Evans S Aspinwall		
Docket No.:	1294-004US01/COHE000113-US-ORG1		
Title:	DATA RETRIEVAL USING EMBEDDINGS FOR DATA IN BACKUP SYSTEMS		

CERTIFICATE UNDER 37 CFR 1.8 I hereby certify that this correspondence is being transmitted via the United States Patent and Trademark Office electronic filing system on September 15, 2025.

By: /Aliya R. Khazon/
Name: Aliya R. Khazon

AMENDMENT

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Commissioner:

In response to the Final Office Action mailed June 16, 2025, the period of response for which runs through September 16, 2025, please amend the application. A Request for Continued Examination accompanies this response.

Amendments to the Claims are reflected in the listing of claims which begins on page 2 of this paper.

Remarks begin on page 7 of this paper.

AMENDMENTS TO THE CLAIMS

This listing of claims will replace all prior versions and listings of claims in the application.

Listing of Claims:

Claim 1. (Currently Amended): A computing system comprising:

one or more storage devices; and

processing circuitry having access to the one or more storage devices and configured to:

process an input received from a user or application, the input comprising a natural language query and indicative of a context for one or more queries subsequently expected from a user or application, to dynamically generate a filter~~, wherein the input indicates a context for one or more queries~~;

apply the dynamically generated filter to backup data to obtain filtered data from the backup data;

encode the filtered data to generate an embedding for each item of the filtered data;

generate an on-demand index of embeddings from the generated embeddings;

process, based on the on-demand index of embeddings, a subsequent retrieval augmented generation (RAG) query to generate, by a language model, a context-aware response for the subsequent RAG query; ~~and~~

output the context-aware response; and

based on at least one of a determination that a period of time has elapsed since the generation of the on-demand index of embeddings or a determination that a number of times the on-demand index of embeddings is used over a period of time is below a threshold, delete the on-demand index of the embeddings.

Claim 2. (Canceled).

Claim 3. (Currently Amended): The computing system of claim 1, wherein the dynamically generated filter specifies one or more of a file type, an association with an entity, a date, a time, or a topic.

Claim 4. (Currently Amended): The computing system of claim 1, wherein to dynamically generate the filter the processing circuitry is configured to:

tokenize the backup data to generate preprocessed text data;

apply a CountVectorizer to the preprocessed text data to compute a matrix of token counts;

process the matrix of token counts with a machine learning model to classify items of the preprocessed text data to any of a plurality of classes.

Claim 5. (Currently Amended): The computing system of claim 4, wherein to apply the dynamically generated filter the processing circuitry is configured to:

include, in the filtered data, items of preprocessed text data that are assigned a class that matches the filter.

Claim 6. (Currently Amended): The computing system of claim 1, wherein to dynamically generate the filter the processing circuitry is configured to:

apply a role-based access control to at least one of the subsequent RAG query or to the filter generation.

Claim 7. (Cancelled)

Claim 8. (Currently Amended): The computing system of claim 1, wherein to process the subsequent RAG query the processing circuitry is configured to:

apply retrieval augmented generation using the subsequent RAG query and the on-demand index of embeddings to generate the context-aware response for the subsequent RAG query.

Claim 9. (Currently Amended): The computing system of claim 1, wherein the input further comprises the subsequent RAG query.

Claim 10. (Currently Amended): The computing system of claim 1, wherein the processing circuitry is configured to receive the subsequent RAG query after the on-demand index of embeddings is generated.

Claim 11. (Cancelled)

Claim 12. (Original): The computing system of claim 1, wherein the processing circuitry is configured to store at least a portion of the backup data to a cache.

Claim 13. (Currently Amended): The computing system of claim 12, wherein the processing circuitry is configured to generate or modify an embedding of the on-demand index of embeddings with a reference to corresponding backup data for the embedding stored to the cache.

Claim 14. (Currently Amended): The computing system of claim 12, wherein the processing circuitry is configured to process, based on the on-demand index of embeddings and the cache, a second subsequent RAG query to generate a context-aware response for the second subsequent RAG query.

Claim 15. (Currently Amended): A method comprising:

processing, by a computing system, an input received from a user or application, the input comprising a natural language query and indicative of a context for one or more queries subsequently expected from a user or application, to dynamically generate a filter~~, wherein the input indicates a context for one or more queries~~;

applying the dynamically generated filter to backup data to obtain filtered data from the backup data;

encoding the filtered data to generate an embedding for each item of the filtered data;

generating an on-demand index of embeddings from the generated embeddings;

processing, based on the on-demand index of embeddings, a subsequent retrieval augmented generation (RAG) query to generate, by a language model, a context-aware response for the subsequent RAG query; ~~and~~

outputting the context-aware response; and

based on at least one of a determination that a period of time has elapsed since the generation of the on-demand index of embeddings or a determination that a number of times the on-demand index of embeddings is used over a period of time is below a threshold, deleting the on-demand index of the embeddings.

Claim 16. (Cancelled).

Claim 17. (Cancelled)

Claim 18. (Currently Amended): The method of claim 15, wherein processing the subsequent RAG query comprises applying retrieval augmented generation using the subsequent RAG query and the on-demand index of embeddings to generate the context-aware response for the subsequent RAG query.

Claim 19. (Currently Amended): The method of claim 15, further comprising:

storing at least a portion of the backup data to a cache;

generating or modifying an embedding of the on-demand index of embeddings with a reference to corresponding backup data for the embedding stored to the cache; and

processing, based on the on-demand index of embeddings and the cache, a second subsequent RAG query to generate a context-aware response for the second subsequent RAG query.

Claim 20. (Currently Amended): Non-transitory computer-readable media comprising instructions that, when executed by processing circuitry, cause the processing circuitry to:

process an input received from a user or application, the input comprising a natural language query and indicative of a context for one or more queries subsequently expected from a user or application, to dynamically generate a filter~~, wherein the input indicates a context for one or more queries~~;

apply the dynamically generated filter to backup data to obtain filtered data from the backup data;

encode the filtered data to generate an embedding for each item of the filtered data;

generate an on-demand index of embeddings from the generated embeddings;

process, based on the on-demand index of embeddings, a subsequent retrieval augmented generation (RAG) query to generate, by a language model, a context-aware response for the subsequent RAG query; ~~and~~

output the context-aware response; and

based on at least one of a determination that a period of time has elapsed since the generation of the on-demand index of embeddings or a determination that a number of times the on-demand index of embeddings is used over a period of time is below a threshold, delete the on-demand index of the embeddings.

REMARKS

This Amendment is in response to the Office Action dated June 16, 2025. Applicant has amended claims 1, 3-6, 8-10, 13-15, and 18-20. Applicant has cancelled claims 7, 11, and 17. Claims 1, 3-6, 8-10, 12-15, and 18-20 are pending upon entry of this communication.

Interview Summary

Applicant thanks the Examiner for the telephonic interview conducted on Thursday, August 7, 2025. Participating in the interview were Examiner Aspinwall and Applicant's representatives, Michael A. Buschbach (Reg. No. 66,307) and Hunter T. Berry (Reg. No. 82,969). During the interview, Applicant's representatives proposed amendments to, e.g., claim 1. Applicant's representatives discussed the rejection of claim 1 under § 101, in view of the proposed amendments. Applicant's representatives also discussed the rejection of claim 1 under § 103, arguing that the provisional application for U.S. Publication No. 2024/0289863 (hereinafter, "Smith Lewis") relied upon for a priority date does not support the subject matter relied upon in the Final Office Action as prior art. No agreements were reached during the interview. No exhibits were submitted, and no demonstrations were performed.

Amendments and Basis

Applicant has amended claim 1, e.g., to recite "process an input received from a user or application, the input comprising a natural language query and indicative of a context for one or more queries subsequently expected from a user or application, to dynamically generate a filter." Support for this amendment is found at least in Application ¶ [0008], which states, "The response generation platform described herein may deliver an end-to-end cloud operational experience that simplifies and transforms Information Technology (IT) operations using a conversation-centric approach that responds to natural language questions with actionable, targeted responses based on data managed by the data platform" and in Application ¶ [0041], which states, "This generated index of embeddings 164 may be effectively 'scoped' to a context for a set of one or more queries expected from a user or application and, in some cases, may be generated in an on-demand manner based on received inputs."

Applicant has amended claim 1, e.g., to recite "based on at least one of a determination that a period of time has elapsed since the generation of the on-demand index of embeddings or a determination that a number of times the on-demand index of embeddings is used over a period of time is below a threshold, delete the on-demand index of the embeddings." Support for this amendment is found at least in Application ¶ [0089], which states, "The system automatically cleans up the index of embeddings 164 after a specific time has elapsed or the number of times the index is used drops below a threshold, ensuring that the system remains efficient and does not become bogged down with outdated or unused indexes."

Independent claims 15 and 19 have been amended similarly.

Claim Rejection Under 35 U.S.C. § 101

The Office Action rejected claims 1, 3-15, and 17-20 under 35 U.S.C. § 101 based on an assertion that these claims are directed to non-statutory subject matter. Applicant respectfully traverses the rejections to the extent the rejections may be considered applicable to the claims as amended.

Claims 1, 3-6, 8-10, 12-15, and 18-20 include additional elements that integrate the alleged judicial exception into practical applications, and are therefore eligible at least under Step 2A, Prong Two of the subject matter eligibility analysis.[1] Amended claim 1, for example, recites "process an input received from a user or application, the input comprising a natural language query and indicative of a context for one or more queries subsequently expected from a user or application, to dynamically generate a filter" and "based on at least one of a determination that a period of time has elapsed since the generation of the on-demand index of embeddings or a determination that a number of times the on-demand index of embeddings is used over a period of time is below a threshold, delete the on-demand index of the embeddings"[2] These additional elements, considering claim 1 as a whole, have the effects of improving the utilization and efficiency of data storage and improving the performance of natural language query processing (more specifically, RAG query processing).

The specification explains, "By generating a unique filter 304 and creating an index of embeddings 164 on-demand or 'on the fly,' the system can quickly and efficiently process large quantities of backup data to be made available for RAG queries or other AI/ML applications,

[1] Step 2A, Prong Two involves: (a) identifying whether there are any additional elements recited in the claim beyond the judicial exception; and (b) evaluating those additional elements individually and in combination to determine whether they integrate the exception into a practical application of that exception. MPEP § 2106.04(d)II.

[2] Emphasis denoting amended claim language.

ensuring that users can access the information they need without significant delays."[3] The specification further explains that cleaning up (e.g., deleting) the on-demand index of embeddings ensures "that the system remains efficient and does not become bogged down with outdated or unused indexes."[4] These descriptions of technical improvements are provided by the amended claim 1 limitations.

These additional elements are not mere insignificant extra-solution activities. While conventional data retrieval systems maintain embeddings for all data to be queried using retrieval augmented generation techniques, requiring large amounts of storage, the amended claim 1 subject matter of "process an input received from a user or application, the input comprising a natural language query and indicative of a context for one or more queries subsequently expected from a user or application, to dynamically generate a filter" and "based on at least one of a determination that a period of time has elapsed since the generation of the on-demand index of embeddings or a determination that a number of times the on-demand index of embeddings is used over a period of time is below a threshold, delete the on-demand index of the embeddings" improves natural language query processing performance and storage utilization by (1) ensuring that the "generated index of embeddings 164 may be effectively 'scoped' to a context for a set of one or more queries expected from a user or application and, in some cases, may be generated in an on-demand manner based on received inputs"[5] and (2) reducing storage of indexes of embeddings that are no longer needed. For example, the "on-demand nature of the index of embeddings 164 … allows users and application to access the data they need when they need it, without having to wait for lengthy processing times or rely on pre-generated indexes that require large amounts of storage" and "the system automatically cleans up the index of embeddings."[6] The additional elements are thus substantive and central to the inventive concept and not insignificant extra-solution activities.

In addition, the amended claim 1 computing system "process[es], based on the on-demand index of embeddings, a subsequent retrieval augmented generation (RAG) query to generate, by a language model, a context-aware response for the subsequent RAG query."[7] These

[3] Application, ¶ [0089].
[4] *Id.*
[5] Application, ¶ [0005].
[6] *Id.*, ¶ [0089].
[7] Emphasis added.

additional elements integrate any alleged abstract ideas into the practical application of improving data retrieval from a backup system. For example, the on-demand indexes "allow users and application[s] to access the data they need when they need it, without having to wait for lengthy processing times or rely on pre-generated indexes that require large amounts of storage," having the effect of improving the performance of a particular machine (e.g., data platform 150 described in the specification and embodied in the claim 1 computing system) by enabling the system to "quickly and efficiently process large quantities of backup data to be made available for RAG queries or other AI/ML applications"[8] and use a language model to generate context-aware responses that are "knowledgeable but also diverse and relevant to an enterprise's domain-specific content."[9] That is, "by leveraging RAG on top of an enterprise's own dataset, a customer will not need to perform costly fine-tuning or initial training to teach [a language model] 'what' to say," having the additional technical benefits of reducing environmental impact.[10]

With respect to the "process … a subsequent retrieval augmented generation (RAG) query to generate, by a language model, a context-aware response" subject matter, the Examiner contends in the Office Action that the previously presented claim 1 "limitation(s) of processing; generating; processing; outputting, as drafted, is a process that, under its broadest reasonable interpretation, covers performance of the limitation in the mind but for the recitation of generic computer components."[11] Applicant respectfully disagrees.

The August 4, 2025 Memorandum[12] issued reminders on evaluating subject matter eligibility of claims, including a reminder that:

> USPTO subject matter eligibility analysis instructs examiners to determine that a claim recites a mental process when it contains limitation(s) that can practically be performed in the human mind, including, for example, observations, evaluations, judgments, and opinions. On the other hand, a claim does not recite a mental process when it contains limitation(s) that cannot practically be performed in the human mind, for instance when the human mind is not equipped to perform the claim limitation(s).

[8] Application, ¶ [0089].
[9] *Id.*, ¶ [0078].
[10] *Id.*, ¶ [0080].
[11] Office Action, pg. 10.
[12] Memorandum from Deputy Commissioner for Patents: Reminders on evaluating subject matter eligibility of claims under 35 U.S.C. 101, dated August 4, 2025.

> The mental process grouping is not without limits. ***Examiners are reminded not to expand this grouping in a manner that encompasses claim limitations that cannot practically be performed in the human mind.***[13]

Claim 1, as amended, recites "process, ***based on the on-demand the index of embeddings***, a subsequent retrieval augmented generation (RAG) query to generate, by a language model, a context-aware response for the subsequent RAG query." It is improper to characterize this subject matter as a step or process that, "under its broadest reasonable interpretation, covers performance of the limitation in the mind but for the recitation of generic computer components." Under the broadest reasonable interpretation, the language of claim 1 nevertheless recites:

> encode the filtered data to generate **an embedding** for each item of the filtered data;
>
> generate an on-demand **index of embeddings** from the generated embeddings;
>
> process, **based on the on-demand index of embeddings**, a subsequent retrieval augmented generation (RAG) query to generate, by a language model, a context-aware response for the subsequent RAG query.

The human mind is simply not equipped to generate or process embeddings as this term would be understood a person of skill in the art, i.e., "a way of representing data as points in n-dimensional space so that similar data points cluster together,"[14] and is therefore incapable of performing the claim limitations, even when interpreted according to BRI. It is entirely impractical for the human mind to process a RAG query based on an on-demand index of embeddings, and to generate a context-aware response, even with pen and paper, and in this way the subject matter is more analogous to that of claims that have been found to not recite mental processes, such as those directed to encryption, packet analysis, position calculation, and image processing.[15] The process step of claim 1 also does not contain limitations involving observations, evaluations, judgments, and opinions that have been found by courts to contain limitations that can be performed in the human mind, when recited at a "high level of generality." Rather, the "process" step is "based on the on-demand index of embeddings."[16]

[13] Emphasis added.
[14] Application ¶ [0072].
[15] *See id.*
[16] *See id.*

Finally, the claimed improvement is not a claim to improving the alleged abstract idea itself. "[T]he claim must include the components or steps of the invention that provide the improvement described in the specification. However, *the claim itself does not need to explicitly recite the improvement described in the specification*..." MPEP § 2106.05(a). As explained above, the claim provides a disclosed improvement in the technical fields of natural language query processing performance and storage utilization.

For at least these reasons, claims 1 is subject matter eligible under 35 U.S.C. § 101. Independent claims 15 and 20, as amended, recite features that are similar to those of amended claim 1, but recites such features in the context of another category of patent eligible subject matter. Accordingly, the arguments with respect to amended independent claim 1 above also apply to amended independent claims 15 and 20. The dependent claims, i.e., claims 3-6, 8-10, 12-14, and 18-19 also integrate any alleged abstract ideas into a practical application, as each of the dependent claims incorporate the subject matter of their respective independent claims by virtue of their dependency. Accordingly, the dependent claims are likewise patentable. For at least these reasons, claims 1, 3-6, 8-10, 12-15, and 18-20 recite patentable subject matter under 35 U.S.C. § 101. Applicant therefore respectfully requests withdrawal of this rejection.

Claim Rejection Under 35 U.S.C. § 103

The Office Action rejected claims 1, 3-15, and 17-20 under 35 U.S.C. § 103 as follows:

- Claims 1, 3, 7-12, 15, 17, 18, and 20 were rejected as allegedly being unpatentable over Chatterjee et al., U.S. Publication No. 2021/0256966 A1 (hereinafter, "Chatterjee"), in view of Bedadala et al., U.S. Publication No. 2018/0329993 A1 (hereinafter, "Bedadala"), in view of Smith Lewis;
- Claims 4 and 5 were rejected as allegedly being unpatentable over Chatterjee, in view of Bedadala, in view of Smith Lewis, in view of Shailabh et al., U.S. Publication No. 2023/0169271 A1 (hereinafter, "Shailabh");
- Claim 6 was rejected as allegedly being unpatentable over Chatterjee, in view of Bedadala, in view of Smith Lewis, in view of Naufel, U.S. Publication No. 2024/0362208 A1 (hereinafter, "Naufel"); and

- Claims 13, 14, and 19 were rejected as allegedly being unpatentable over Chatterjee, in view of Bedadala, in view of Smith Lewis, in view of Oliner et al., U.S. Publication No. 2023/0069958 A1 (hereinafter, "Oliner").

Applicant respectfully traverses the rejections to the extent the rejections may be considered applicable to the claims as amended. The applied references, alone or in any combination, fail to disclose or suggest the features defined by Applicant's claims, and there would have been no apparent reason that would have caused one of ordinary skill in the art to modify the applied references to arrive at the claimed features.

Applicant has amended claim 1, for example, to recite "process an input received from a user or application, the input comprising a natural language query and indicative of a context for one or more queries subsequently expected from a user or application, to dynamically generate a filter; apply the dynamically generated filter to backup data to obtain filtered data from the backup data; encode the filtered data to generate an embedding for each item of the filtered data; generate an on-demand index of embeddings from the generated embeddings." The cited references, whether considered alone or in combination, do not disclose generation of an **on-demand** index of embeddings as claimed.

Applicant has also amended claim 1 to recite "based on at least one of a determination that a period of time has elapsed since the generation of the on-demand index of embeddings or a determination that a number of times the on-demand index of embeddings is used over a period of time is below a threshold, delete the on-demand index of the embeddings." The cited references, whether considered alone or in combination, do not disclose this subject matter.

In addition, at least some cited portions of newly-cited reference Smith-Lewis have an effective filing date that is after the effective filing date of Applicant's claims. MPEP § 2154.01(b) states, "[t]he provisional or other earlier application(s) to which the reference patent document claims a right of priority or benefit must '**describe the subject matter**' relied upon in the reference patent document as prior art.[17] That is, a provisional application to which a utility application claims priority to must adequately describe all subject matter being cited as prior art in the later filed utility application in order for the cited portions to benefit from the priority

[17] MPEP § 2154.01(b) (citing *Penumbra, Inc. v. RapidPulse, Inc.,* 2023 USPQ2d 292, IPR2021-01466, Paper 34 (March 10, 2023) (precedential as to section II.E.3)) (*emphasis added*).

claim. This concept was recently discussed in the Federal Circuit case *In Re Riggs* (hereinafter "Riggs").

In *Riggs*, an Examiner rejected claims of a utility patent application as being anticipated under § 102 by Lettich, a U.S. non-provisional patent application claiming priority to a provisional application.[18] The application at issue was filed on December 7, 2004, and properly claimed priority to a provisional application filed on July 28, 2000, while the Lettich application was filed on April 26, 2001, and properly claimed priority to a provisional application filed on April 27, 2000.[19] Therefore, while the Lettich non-provisional application did not pre-date the effective filing date of the applicant's effective filing date, Lettich's priority claim allowed the non-provisional application to have an effective filing date pre-dating that of the application at issue.[20]

While the Examiner in *Riggs* properly showed support for one claim of Lettich in the Lettich provisional, and then relied on other portions of the Lettich non-provisional specification to support the rejection, per the holding in *Dynamic Drinkware, LLC v. Nat'l Graphics, Inc.*, 800 F.3D 1375, 1378 (Fed. Cir. 2015),[21] the *Riggs* court nevertheless held that a provisional must support all paragraphs being cited as prior art for the nonprovisional application to benefit from the filing date of the provisional.[22] The *Riggs* court held that "[e]ven if one demonstrates that a provisional application provides written description support for one claim on the non-provisional application or patent, the provisional application must also provide written description support for the specific portions of the patent specification identified and relied on in the prior art rejection."[23] That is, "to claim priority to the provisional filing date, the portion of the application relied on by the examiner as prior art must be supported by the provisional application,"[24] with the *Riggs* court remanding the case to determine whether the Lettich provisional properly supported the paragraphs cited by the Examiner in the Office Action.[25]

[18] *In Re Riggs* (Fed. Cir. 2025), *available at* https://www.cafc.uscourts.gov/opinions-orders/22-1945.OPINION.3-24-2025_2486478.pdf, at 2.
[19] *Id.* at 10.
[20] *See id.*
[21] *See id.* at 11-12.
[22] *Id.* at 12-14.
[23] *Id.* at 12.
[24] *Id.*
[25] *Id.* at 14.

In citing Smith Lewis, the Office Action relies upon Smith Lewis's priority claim to provisional application 63/448,117 (hereinafter "Smith Lewis provisional), filed on February 24, 2023, to predate Applicant's application effective filing date of May 22, 2023 that is based on the priority claim to provisional application 63/503,631 (hereinafter "Applicant's provisional") filed on May 22, 2023. Here, the Office Action cites Smith Lewis as disclosing the claim 1 limitations in which the computing system "encode[s] the filtered [DATA] to generate an embedding for each item of the filtered data," "generate[s] an index of embeddings from the generated embeddings," and "process[es], based on the index of embeddings, a subsequent retrieval augmented generation (RAG) query to generate a response for the subsequent RAG query."[26] More specifically, the Office Action cites Smith Lewis, paragraphs [0045], [0052], and [0076].

Smith Lewis, paragraphs [0052], cited as disclosing the previously presented claim 1 subject matter of "process[ing] based on the index of embeddings, a subsequent retrieval augmented generation (RAG) query to generate a response for the subsequent RAG query,"[27] states:

> In some embodiments, additional filtering steps, for example against the metadata for each chunk, may be performed prior to vector comparison, so that manual or automated tags and other metadata may be taken into account alongside the meaning and content of the text. For example, in some embodiments, this embedding-retrieval pipeline may be applied to Retrieval Augmented Generation (RAG) whereby a targeted search across the embedded vector database is performed in response to a user query, e.g., in order to produce context for a conversational agent to then generate a response, for example by inserting the retrieved text chunks into a system prompt or message used to generate the agent's response to the user.

In reviewing the Smith Lewis provisional, however, Applicant finds no reference to "additional filtering steps, for example against the metadata for each chunk," such that "other metadata may be taken into account alongside the meaning and content of the text." Additionally, Applicant finds no reference in the Smith Lewis provisional to the retrieval of embeddings (e.g., an "embedding-retrieval pipeline"), nor does Applicant find any reference to a retrieval augmented generation (RAG) query, or any general query which may be retrieved and/or used in, or to trigger, a "targeted search across the embedded vector database." As such, the Smith Lewis provisional does not support the subject matter discussed in paragraph [0052] of

[26] Office Action at pgs. 34-37.
[27] *Id.* at pg. 34.

Smith Lewis and relied upon in the rejection of claim 1. The contents of paragraph [0052] of Smith Lewis, including the RAG query, use of metadata, and an embedding-retrieval pipeline, are therefore considered new matter vis-à-vis the Smith Lewis provisional and have a later effective filing date of February 26, 2024.

Additionally, Smith Lewis, paragraphs [0076], cited as disclosing "encod[ing] the filtered data to generate an embedding for each item of the filtered data,"[28] states:

> In some embodiments, conversational history may be filtered or summarized prior to embedding to optimize storage, enhance retrieval, or increase privacy. For example, conversation history may be turned off or restricted for a user based on some settings they control. In some embodiments, conversation history may be stored hierarchically, e.g., by summarizing a full conversation (series of messages within some time frame such as the past hour, or about some related set of topics), and embedding this either in place of or in addition to the messages comprising that conversation. In this way longer histories may be efficiently searched by reference to the conversation summaries, or full relevant conversations may be retrieved and inserted into the system prompt rather than only snippets and individual messages.

In reviewing the Smith Lewis provisional, however, Applicant finds no reference to a conversational history being "filtered or summarized prior to embedding to optimize storage, enhance retrieval, or increase privacy.". Additionally, Applicant finds no reference to the ability of the conversational history to be "turned off or restricted for a user based on some settings they control" or "stored hierarchically," such as by "summarizing a full conversational… and embedding this either in place of or in addition to the messages comprising that conversation." As such, the Smith Lewis provisional does not support the filtering discussed in paragraph [0076] of Smith Lewis. The contents of paragraph [0076], including the filtering and summarizing of conversational history, are therefore considered new matter vis-à-vis the Smith Lewis provisional and have a later effective filing date of February 26, 2024.

Therefore, at least the Smith Lewis paragraphs [0052] and [0076] cited by the Examiner do not appear in and are not supported by the Smith Lewis provisional. As such, the paragraphs are considered new matter vis-à-vis the Smith Lewis provisional and have a later effective filing date of February 26, 2024. Applicant's application claims priority to Applicant's provisional, which supports claim 1. Claim 1 therefore has an effective filing date of May 22, 2023, which pre-dates the February 26, 2024 effective filing date of the cited portions of Smith Lewis. Since

[28] *Id.* at pg. 34.

cited portions of a reference must have an effective filing date before the filing date of the claimed invention to qualify as prior art, the cited portions of Smith Lewis are disqualified as prior art under § 102(a)(1) or § 102(a)(2) with respect to Applicant's claims and cannot be used in the rejection of claim 1 under § 103.

Because of the significant differences between Smith Lewis and the Smith Lewis provisional, Applicant requests that any future reliance by the Office on Smith Lewis identify supporting paragraphs in the Smith Lewis provisional.

For at least the reasons discussed above, independent claim 1 is patentable over Chatterjee in view of Bedadala in further view of Smith Lewis. Claims 15 and 20 are amended similarly to claim 1 and are patentable at least for reasons similar to those discussed above. The dependent claims, i.e., claims 3-6, 8-10, 12-14, and 18-19, incorporate the requirements of the respective independent claims.[29] Accordingly, the dependent claims are likewise patentable. Applicant therefore respectfully requests reconsideration and withdrawal of this rejection.

[29] 35 U.S.C. § 112(d).

CONCLUSION

All claims in this application are in condition for allowance. Applicant does not necessarily acquiesce as to any assertion made in the Office Action, and Applicant's silence with respect to any such assertion in the Office Action should not be interpreted as Applicant's acquiescence thereto. Further, Applicant does not concede that the art cited in the record is relevant art. Applicant reserves the right to comment further with respect to the applied references and any pending claim in a future Amendment, Response, on appeal, in any other proceeding, or otherwise. Applicant respectfully requests reconsideration and prompt allowance of all pending claims.

Please charge any additional fees or credit any overpayment to deposit account number 50-1778. The Examiner is invited to telephone the below-signed representative to discuss this application.

Date:
September 15, 2025
SHUMAKER & SIEFFERT, P.A.
Telephone: 651.286.8355

By:
/Hunter T. Berry/
Name: Hunter Berry, Reg. No. 82,969