UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
**United States Patent and Trademark Office**
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

| APPLICATION NO. | FILING DATE | FIRST NAMED INVENTOR | ATTORNEY DOCKET NO. | CONFIRMATION NO. |
|---|---|---|---|---|
| 18/611,498 | 03/20/2024 | Daniel Hunter | 56710-0002001 | 9701 |

26191 7590 02/19/2025
FISH & RICHARDSON P.C. (TC)
PO BOX 1022
MINNEAPOLIS, MN 55440-1022

| EXAMINER | |
|---|---|
| LE, MICHAEL | |

| ART UNIT | PAPER NUMBER |
|---|---|
| 2163 | |

| NOTIFICATION DATE | DELIVERY MODE |
|---|---|
| 02/19/2025 | ELECTRONIC |

**Please find below and/or attached an Office communication concerning this application or proceeding.**

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

PATDOCTC@fr.com

| Office Action Summary | Application No. 18/611,498 | Applicant(s) Hunter et al. | |
|---|---|---|---|
| | Examiner MICHAEL LE | Art Unit 2163 | AIA (FITF) Status Yes |

*-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --*

**Period for Reply**

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTHS FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

**Status**

1)☑ Responsive to communication(s) filed on 10/17/2024.

☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2a)☑ This action is **FINAL.** 2b) ☐ This action is non-final.

3)☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

4)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

**Disposition of Claims***

5) ☑ Claim(s) 1-6 and 9-23 is/are pending in the application.

5a) Of the above claim(s) _____ is/are withdrawn from consideration.

6) ☐ Claim(s) _____ is/are allowed.

7) ☑ Claim(s) 1-6 and 9-23 is/are rejected.

8) ☐ Claim(s) _____ is/are objected to.

9) ☐ Claim(s) _____ are subject to restriction and/or election requirement

* If any claims have been determined allowable, you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information, please see http://www.uspto.gov/patents/init_events/pph/index.jsp or send an inquiry to **PPHfeedback@uspto.gov.**

**Application Papers**

10)☐ The specification is objected to by the Examiner.

11)☐ The drawing(s) filed on _____ is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

**Priority under 35 U.S.C. § 119**

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

**Certified copies:**

a)☐ All b)☐ Some** c)☐ None of the:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

** See the attached detailed Office action for a list of the certified copies not received.

**Attachment(s)**

1) ☑ Notice of References Cited (PTO-892)

2) ☐ Information Disclosure Statement(s) (PTO/SB/08a and/or PTO/SB/08b) Paper No(s)/Mail Date _____.

3) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date _____.

4) ☐ Other: _____.

## DETAILED ACTION

### *Summary and Status of Claims*

1. The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA.

2. This Office Action is in response to Applicant's reply filed 10/17/2024.

3. Claims 7 and 8 are cancelled.

4. Claims 21-23 are new.

5. Claims 1-6 and 9-23 are pending.

6. Claims 1-6 and 9-23 are rejected under 35 U.S.C. 101.

7. Claims 1-6 and 9-20 are rejected under 35 U.S.C. 103 as being unpatentable over Gupta (US Patent Pub 2022/0092097) in view of Madisetti et al. (US Patent 12,001,462).

8. Claims 21-23 are rejected under 35 U.S.C. 103 as being unpatentable over Gupta (US Patent Pub 2022/0092097) in view of Madisetti et al. (US Patent 12,001,462), further in view of Staub et al. (US Patent Pub 2021/0216576).

9. The text of those sections of Title 35, U.S. Code not included in this action can be found in a prior Office action.

### *Claim Rejections - 35 USC § 101*

10. 35 U.S.C. 101 reads as follows:

> Whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefor, subject to the conditions and requirements of this title.

11. **Claims 1-6 and 9-23 are rejected under 35 U.S.C. 101** because the claimed invention is directed to a judicial exception (i.e., a law of nature, a natural phenomenon, or an abstract idea) without significantly more. Determining whether claims are statutory under 35 U.S.C. 101 involves a two-step

analysis. Step 1 requires a determination of whether the claims are directed to the statutory categories of invention. Step 2 requires a determination of whether the claims are directed to a judicial exception without significantly more. Step 2 is divided into two prongs, with the first prong having a part 1 and part 2. *See* MPEP 2106; *See* 2019 Revised Patent Subject Matter Eligibility Guidance (2019 PEG).

12. Pursuant to Step 1, claims 1-14 are directed to the statutory category of a process. Claims 15-17 would have been directed to a manufacture, if amendment to overcome the nonstatutory subject matter rejection above. Claims 18-20 are directed to a machine.

**13. Claim 1**

14. Pursuant to Step 2A, part 1, claims are analyzed to determine whether they are directed to an abstract idea. Under the 2019 PEG, claims are deemed to be directed to an abstract idea if they fall within one of the enumerated categories of (a) mathematical concepts, (b) certain methods of organizing human activity, and (c) mental processes. Here, claim 1 are directed to an abstract idea categorized under mental processes. Courts consider a claims to be directed to a mental process if it "can be performed in the human mind, or by a human using a pen and paper." MPEP 2016(a)(2)(III). Claim 1 recites a mental process because the recited steps recite the action of observation. See MPEP 2106(a)(2)(III). For example, claim 1 recites limitations of receiving a query related to a document from a user, submitting the document to a first LLM along with a first prompt to generate an outline, receiving the document outline, submitting the document to a second LLM with a second prompt to generate metadata of the document and receiving the metadata, submitting the query, metadata, and outline to a third LLM with a third prompt to generate a natural language response to the query and receiving the response, determining relevant document sentences based on embeddings of sentences from the document and a selected sentence from the natural language response, submitting the relevant document sentences and the selected sentence from the natural language response to a fourth LLM along with a fourth prompt prompting the fourth LLM to generate a ranking of relevant document

sentences based on relevance of respective document sentences to the selected sentence from the natural language response, correlating the sentence from the response to a sentence from the document based on the ranking generated by the fourth LLM, and providing the response including a citation to one or more specific sentences from the document correlated to the selected sentence from the natural language response to the user.

15. Claims are given their broadest reasonable interpretation in light of the specification. MPEP 2111. Courts consider a mental process if it "can be performed in the human mind, or by a human using a pen and paper." The mental process grouping covers concepts performed in the human mind, including observation, evaluation, judgment, and opinion. MPEP 2016(a)(2)(III). The mathematical concepts grouping is defined as mathematical relationships, mathematical formulas or equations, and mathematical calculations. The Supreme Court has identified a number of concepts falling within this grouping as abstract ideas including: a procedure for converting binary-coded decimal numerals into pure binary form, Gottschalk v. Benson, 409 U.S. 63, 65, 175 USPQ2d 673, 674 (1972); a mathematical formula for calculating an alarm limit, Parker v. Flook, 437 U.S. 584, 588-89, 198 USPQ2d 193, 195 (1978); the Arrhenius equation, Diamond v. Diehr, 450 U.S. 175, 191, 209 USPQ 1, 15 (1981); and a mathematical formula for hedging, Bilski v. Kappos, 561 U.S. 593, 611, 95 USPQ 2d 1001, 1004 (2010). Limitations can also be deemed insignificant extra-solution activity (IESA). The term "extra-solution activity" can be understood as activities incidental to the primary process or product that are merely a nominal or tangential addition to the claim. Extra-solution activity includes both pre-solution and post-solution activity. An example of pre-solution activity is a step of gathering data for use in a claimed process, e.g., a step of obtaining information about credit card transactions, which is recited as part of a claimed process of analyzing and manipulating the gathered information by a series of steps in order to detect whether the transactions were fraudulent. An example of post-solution activity is an element that is not integrated into the claim as a whole, e.g., a printer that is used to output a report of

fraudulent transactions, which is recited in a claim to a computer programmed to analyze and manipulate information about credit card transactions in order to detect whether the transactions were fraudulent. MPEP 2106.05(g). In this case, the limitations of "receiving" and "submitting" amount to IESA in the form of receiving and transmitting data, such as over a network. The additional elements of the particular "large language models" are recited at a high level of generality without reciting specifics of the particular models. These models perform steps that can be performed by a person with the aid of a computer as a tool to perform the abstract idea, which is to summarize a document into an outline and metadata, determine sentences of relevant documents that correlate with the determined response to the query, and providing citations to those sentences after ranking them. The step of "correlating" is a mental step that is not recited with any particularity and as recited, can be performed by a person in the mind through observation and judgment of the documents. Lastly, the providing of the citations is IESA and a mental step because it merely amounts to outputting of information, where the information are citations to sentences in the documents, which is a step practically performed by a person in the mind with the aid of a computer as a tool. Therefore, claim 1 is directed to an abstract idea categorized as a mental process.

16. Pursuant to Step 2A, part 2, claims are analyzed to determine whether the recited abstract idea is integrated into a practical application. This evaluation is performed by (1) identifying whether there are any additional elements recited in the claim beyond the judicial exception, and (2) evaluating those additional elements individually and in combination to determine whether the claim as a whole integrates the exception into a practical application. See MPEP 2106.04(d). One way to determine integration into a practical application is when the claimed invention improves the functioning of a computer or improves another technology or technical field. To evaluate an improvement to a computer or technical field, the specification must set forth an improvement in technology and the claim itself must reflect the disclosed improvement. See MPEP 2106.04(d)(1). In this case, as explained above,

claim 1 merely recite limitations directed to a mental process. These limitations describe generating an outline and metadata of a document in order to generate a natural language response to the query about the document with an indication to a correlation of the response to a part of the document and determining a citation to sentences in the document relevant to the response. These steps do not provide meaningful limits on the abstract idea because summarizing documents in order to answer questions about them is a mental process often performed by people. The additional components of using LLMs to perform the various steps are recited at a high level of generality, which do not add meaningful limits on the recited abstract idea to integrate it into a practical application. Here, the claim recites no details about a particular LLMs. The LLMs are used to generally apply the abstract idea (i.e., perform the outlining, generating metadata, answering a query about a document, and ranking relevant document sentences recited) without placing any limitation on how the LLMs operate. In addition, the limitation would cover every mode of implementing the recited abstract idea using the LLMs. The claim omits any details as to how the LLMs solve a technical problem and instead recites only the idea of a solution or outcome. See MPEP 2106.05(f). Accordingly, for these reasons, none of the limitations integrate the abstract idea by providing an improvement to the functioning of a computer or technology, implementing the abstract idea with a particular machine or manufacture that is integral to the claim, effecting a transformation or reduction of a particular article to a different state or thing, nor applying the abstract idea in some meaningful way beyond linking its use to computer technology. *See* 2019 PEG. At best, they merely tie the abstract idea to the field of technology of machine learning and language models and act as a tool for the person to perform an existing process.

17. Since claim 1 is directed to an abstract idea categorized as a mental process and does not integrate the judicial exception into a practical application, claim 1 is directed to a judicial exception.

18. Pursuant to Step 2B, claims are analyzed to determine whether they recite significantly more than the abstract idea. In other words, it is determined whether the claims provide an inventive

concept. In this case, claim 1 does not recite limitations that amount to significantly more than the abstract idea. The limitations are steps involving processes that can be practically performed by a human with the aid of pen and paper or with the aid of a computer as a tool to execute the abstract idea, as explained above. The steps identified as being directed to IESA are well understood, routine, and conventional. MPEP 2106.05(d). Even when considered in combination, these additional elements represent mere instructions to implement an abstract idea or other exception on a computer and insignificant extra-solution activity, which do not provide an inventive concept. For at least these reasons, claim 1 is nonstatutory because they are directed to a judicial exception without significantly more. For at least these reasons, claim 1 is nonstatutory because they are directed to a judicial exception without significantly more.

**19. Claims 2-4**

20. Pursuant to step 2A, part 1, claims 2-4 depend on claim 1 and therefore recites the same abstract idea. Pursuant to step 2A, part 2, claims 2-4 recites the additional limitations of the metadata comprises a title, date, or information indicating one or more parties to the document, wherein the second LLM generate the metadata based on a predetermined portion of the document, and the first prompt is to generate as the outline of the document a topic aware outline of the document. These limitations do not provide meaningful limits on the abstract idea because they merely describe extra solution activity or features, such as type of metadata, where a person should look to generate metadata, and that the outline is a topic aware outline. Therefore, these additional limitations do not integrate the abstract idea into a practical application. Pursuant to step 2B, the additional limitations do not amount to significantly more than the abstract idea because the limitations are not recited in a manner that provides improvements to the functioning of a computer or any other technology or technical field.

**21. Claim 5**

22. Pursuant to step 2A, part 1, claim 5 depends on claim 4 and therefore recites the same abstract idea. Pursuant to step 2A, part 2, claim 5 recites the additional limitations of submitting the document to a fifth LLM with a fifth prompt to generate a numerical outline of the document, receiving it, and submitting it with the query, metadata, and topic outline to the third LLM. These limitations describe limitations that can be performed by a person with the aid of a computer as a tool, which in this case, is an LLM, recited at a high level of generality as explained in the rejection of claim 1 above. Therefore, these additional limitations do not integrate the abstract idea into a practical application. Pursuant to step 2B, the additional limitations do not amount to significantly more than the abstract idea because the limitations are not recited in a manner that provides improvements to the functioning of a computer or any other technology or technical field for the same reasons.

**23. Claim 6**

24. Pursuant to step 2A, part 1, claim 6 depends on claim 5 and therefore recites the same abstract idea. Pursuant to step 2A, part 2, claim 6 recites the additional limitations of submitting the query to a sixth LLM with a sixth prompt to transform the query into outline requests for each of the topic aware outline and the numerical outline, receiving the outline requests, and including them in the first and fifth prompts respectively. These limitations merely describe using a computer as a tool to convert a query into a format that may be used for each of the first and fourth LLMs respectively. However, the additional elements are recited at a high level of generality that does not provide meaningful limits on the abstract idea. Therefore, these additional limitations do not integrate the abstract idea into a practical application. Pursuant to step 2B, the additional limitations do not amount to significantly more than the abstract idea because the limitations are not recited in a manner that provides improvements to the functioning of a computer or any other technology or technical field for the same resaons.

**25. Claim 9**

26. Pursuant to step 2A, part 1, claim 9 depends on claim 1 and therefore recites the same abstract idea. Pursuant to step 2A, part 2, claim 9 recites the additional limitations of determining relevant document sentences based on vector distance of the embeddings. As stated above, these limitations merely recite steps that can be performed by a person with the aid of a computer as a tool, which can be used to create an embedding of a sentence and perform vector distances between them to determine relevancy scores. Vector distance is a mathematical calculation that can be performed by a person with the aid of pen and paper or with a computer as a tool. The limitations are not recited with meaningful limits on the abstract idea. Therefore, these additional limitations do not integrate the abstract idea into a practical application. Pursuant to step 2B, the additional limitations do not amount to significantly more than the abstract idea because the limitations are not recited in a manner that provides improvements to the functioning of a computer or any other technology or technical field for the same reasons.

**27. Claim 10**

28. Pursuant to step 2A, part 1, claim 10 depends on claim 6 and therefore recites the same abstract idea. Pursuant to step 2A, part 2, claim 10 recites the additional limitations of the LLMs are selected based on one or more of latency, maximum context window size, accuracy of results, quality of results, or resource usage. These limitations merely describe selecting computer components based on performance or specific attributes. Such decisions can be made by a person when reviewing possible choices. Therefore, these additional limitations do not integrate the abstract idea into a practical application. Pursuant to step 2B, the additional limitations do not amount to significantly more than the abstract idea because the limitations are not recited in a manner that provides improvements to the functioning of a computer or any other technology or technical field.

**29. Claim 11**

30. Pursuant to step 2A, part 1, claim 11 depends on claim 1 and therefore recites the same abstract idea. Pursuant to step 2A, part 2, claim 11 recites the additional limitations of using a 7th LLM to classify the query into a first query classification of two possible classifications, determining a 7th prompt based on the first query classification to prompt the 3rd LLM based on the metadata and the document outline and provide the 7th prompt to the third LLM. These limitations do not provide meaningful limits on the abstract idea because a person can classify a query based on what type of query it is and accordingly, mentally determine what type of query is being asked to best get a response. Therefore, these additional limitations do not integrate the abstract idea into a practical application. Pursuant to step 2B, the additional limitations do not amount to significantly more than the abstract idea because the limitations are not recited in a manner that provides improvements to the functioning of a computer or any other technology or technical field.

**31. Claims 12 and 13**

32. Pursuant to step 2A, part 1, claims 12 and 13 depend on claim 11 and therefore recites the same abstract idea. Pursuant to step 2A, part 2, claims 12 and 13 recites the additional limitations of wherein two or more of the LLMs are the same or different, which does not provide meaningful limits on the abstract idea. Therefore, these additional limitations do not integrate the abstract idea into a practical application. Pursuant to step 2B, the additional limitations do not amount to significantly more than the abstract idea because the limitations are not recited in a manner that provides improvements to the functioning of a computer or any other technology or technical field.

**33. Claim 14**

34. Pursuant to step 2A, part 1, claim 14 depends on claim 1 and therefore recites the same abstract idea. Pursuant to step 2A, part 2, claim 14 recites the additional limitations of the document is a plurality of documents. This limitation does not provide meaningful limits on the abstract idea. Therefore, these additional limitations do not integrate the abstract idea into a practical application.

Pursuant to step 2B, the additional limitations do not amount to significantly more than the abstract idea because the limitations are not recited in a manner that provides improvements to the functioning of a computer or any other technology or technical field.

35. **Claims 15-17** recite essentially the same subject matter as claims 1-3, respectively, in the form of a computer readable storage media. Therefore, they are rejected for the same reasons.

36. **Claims 18-20** recite essentially the same subject matter as claims 1-3, respectively, in the form of a system. Therefore, they are rejected for the same reasons.

37. **Claims 21-23**

38. Pursuant to step 2A, part 1, claims 21-23 depend on claim 1 and therefore recites the same abstract idea. Pursuant to step 2A, part 2, claims 21-23 recites the additional limitations of the specific sentences from the document comprise less than ten, five, or three sentences. These limitations do not provide meaningful limits on the abstract idea. They merely recite a value by which to guide returning of the citation. The limitation in claim 1 also only requires that one sentence be cited to, which reads on each of these sentences. Accordingly, the limitations do not add meaningful limits on the abstract idea discussed with regards to claim 1. Therefore, these additional limitations do not integrate the abstract idea into a practical application. Pursuant to step 2B, the additional limitations do not amount to significantly more than the abstract idea because the limitations are not recited in a manner that provides improvements to the functioning of a computer or any other technology or technical field for the same reasons.

39. **Claims 1-6 and 9-23** are therefore not drawn to eligible subject matter as they are directed to an abstract idea without significantly more.

40. To expedite a complete examination of the instant application, the claims rejected under 35 U.S.C. 101 (nonstatutory) above are further rejected as set forth below in anticipation of applicant amending these claims to overcome the rejection.

***Note on Prior Art Rejections***

41. In the event the determination of the status of the application as subject to AIA 35 U.S.C. 102 and 103 (or as subject to pre-AIA 35 U.S.C. 102 and 103) is incorrect, any correction of the statutory basis for the rejection will not be considered a new ground of rejection if the prior art relied upon, and the rationale supporting the rejection, would be the same under either status.

***Claim Rejections - 35 USC § 103***

42. The following is a quotation of 35 U.S.C. 103 which forms the basis for all obviousness rejections set forth in this Office action:

> A patent for a claimed invention may not be obtained, notwithstanding that the claimed invention is not identically disclosed as set forth in section 102 of this title, if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the claimed invention pertains. Patentability shall not be negated by the manner in which the invention was made.

43. The factual inquiries set forth in *Graham v. John Deere Co.*, 383 U.S. 1, 148 USPQ 459 (1966), that are applied for establishing a background for determining obviousness under 35 U.S.C. 103 are summarized as follows:

1. Determining the scope and contents of the prior art.

2. Ascertaining the differences between the prior art and the claims at issue.

3. Resolving the level of ordinary skill in the pertinent art.

4. Considering objective evidence present in the application indicating obviousness or nonobviousness.

44. This application currently names joint inventors. In considering patentability of the claims the examiner presumes that the subject matter of the various claims was commonly owned as of the effective filing date of the claimed invention(s) absent any evidence to the contrary. Applicant is advised

of the obligation under 37 CFR 1.56 to point out the inventor and effective filing dates of each claim that was not commonly owned as of the effective filing date of the later invention in order for the examiner to consider the applicability of 35 U.S.C. 102(b)(2)(C) for any potential 35 U.S.C. 102(a)(2) prior art against the later invention.

**45. Claims 1-6 and 9-20 are rejected under 35 U.S.C. 103 as being unpatentable over Gupta (US Patent Pub 2022/0092097) in view of Madisetti et al. (US Patent 12,001,462) (Madisetti).**

46. In regards to **claim 1**, Gupta discloses a method comprising:

a. receiving, from a user, a query related to a document (Gupta at para. 0034)[1];

b. submitting the document to the system to generate an outline of the document (Gupta at paras. 0024-25)[2];

c. receiving a document outline of the document from the system (Gupta at para. 0024-25);

d. submitting the document to generate metadata of the document (Gupta at paras. 0032-34)[3]

e. receiving document metadata of the document from the system (Gupta at paras. 0032-34);

f. submitting at least a portion of the query, the document metadata, and the document outline to the system to generate a natural language response to the query based at least in part on the document metadata and the document outline (Gupta at para. 0034)[4];

g. receiving the natural language response from the system (Gupta at para. 0034);

i. providing the natural language response to the user. Gupta at paras. 0034, 0047.

---

[1] A user submits a question (i.e., query) related to a document.
[2] Documents are submitted to the system to generate outlines for them.
[3] The system extracts entities of the document, such as dates and important clauses (i.e., metadata).
[4] The extracted entity information and outline are used by the system to generate an answer to a question in the given context.

47. Gupta does not expressly disclose (1) submitting the document to a first LLM with a first prompt to generate an outline of the document and using a second LLM with a second prompt to generate metadata of the document and a third LLM with a third prompt to generate a natural language response to the query, (2) determining relevant document sentences based on embeddings of sentences from the document and a selected sentence from the natural language response, (3) submitting the relevant document sentences and the selected sentences form the natural language response to a fourth LLM along with a fourth prompt prompting the fourth LLM to generate a ranking of relevant document sentences based on the relevance of respective relevant document sentences to the selected sentence from the natural language response, (4) determining relevant document sentences based on embeddings of sentences from the document and a selected sentence from the natural language response, (5) submitting the relevant document sentences and the selected sentence from the natural language response to a fourth LLM along with a fourth prompt prompting the fourth LLM to generate a ranking of relevant document sentences based on relevance of respective relevant document sentences to the selected sentence from the natural language response, (6) correlating the selected sentence from the natural language response to a first sentence from the document based on the ranking of relevant document sentences generated by the fourth LLM, and (7) providing the natural language response with a citation to a specific sentence or to specific sentences from the document correlated to the selected sentence from the natural language response. It is noted that Gupta does disclose using an artificial intelligence with natural language processing capabilities to generate the outline of the document and answer questions about the document in natural language. Gupta at paras. 0004, 0034. What is not specifically disclosed is using multiple LLMs to perform these steps, determining relevant document sentences to a selected sentence using embeddings, ranking the relevant sentences, and correlating the selected sentence from the natural language response to a sentence from the document and providing a citation to a sentence from the document that is correlated to the selected sentence.

48. Madisetti discloses a system and method that utilizes a plurality of LLMs for processing documents, such as knowledge documents, and providing answers to queries pertaining to the knowledge documents. In response to a user prompt (i.e., query), the system performs similarity analysis using prompt embeddings (i.e., correlating) to rank and determine the best context aware answers to return to the user along with citations from the knowledge documents (i.e., providing a natural language response to the user with a citation ...). Madsetti at col. 7, lines 30-67; col. 8, lines 1-7. Madisetti expressly states "the best results are sent to the user along with citations from the knowledge documents." Here, "citations from the knowledge documents" are interpreted as one or more specific sentences from the document.

49. Gupta and Madisetti are analogous art because they are both directed to the same field of endeavor of artificial intelligence processing of documents.

50. At the time before the effective filing date of the instant application, it would have been obvious to one of ordinary skill in the art to modify Gupta by adding the features of (1) submitting the document to a first LLM with a first prompt to generate an outline of the document and using a second LLM with a second prompt to generate metadata of the document and a third LLM with a third prompt to generate a natural language response to the query, (2) determining relevant document sentences based on embeddings of sentences from the document and a selected sentence from the natural language response, (3) submitting the relevant document sentences and the selected sentences form the natural language response to a fourth LLM along with a fourth prompt prompting the fourth LLM to generate a ranking of relevant document sentences based on the relevance of respective relevant document sentences to the selected sentence from the natural language response, (4) determining relevant document sentences based on embeddings of sentences from the document and a selected sentence from the natural language response, (5) submitting the relevant document sentences and the selected sentence from the natural language response to a fourth LLM along with a fourth prompt prompting the

fourth LLM to generate a ranking of relevant document sentences based on relevance of respective relevant document sentences to the selected sentence from the natural language response, (6) correlating the selected sentence from the natural language response to a first sentence from the document based on the ranking of relevant document sentences generated by the fourth LLM, and (7) providing the natural language response with a citation to a specific sentence or to specific sentences from the document correlated to the selected sentence from the natural language response, as disclosed by Madisetti.

51. The motivation for doing so would have been to provide more accuracy with responses. Madisetti at col. 7, lines 30-54.

52. In regards to **claim 2**, Gupta in view of Madisetti discloses the method of claim 1, wherein the document metadata comprises a document title, a document date, or information indicating one or more parties party to the document. Gupta at para. 0032.[5]

53. In regards to **claim 3**, Gupta in view of Madisetti discloses the method of claim 1, wherein the second large language model generates the document metadata based on a predetermined portion of the document. Gupta at para. 0034.[6]

54. In regards to **claim 4**, Gupta in view of Madisetti discloses the method of claim 1, wherein the first prompt prompts the first large language model to generate, as the outline of the document, a topic aware outline of the document. Gupta at para. 0025.[7]

[5] Extracted entities include important dates (i.e., a document date).
[6] The system searches a document outline with the hierarchical structure for entities based on a predefined configuration to search a section (i.e., predetermined portion of the document).
[7] The outline is based on topics detected in the document (i.e., topic aware document).

55. In regards to **claim 5**, Gupta in view of Madisetti discloses the method of claim 4, but does not expressly disclose further comprising: submitting the document to a fifth large language model along with a fifth prompt prompting the fifth LLM to generate a numerical outline of the document, receiving the numerical outline of the document, and submitting the numerical outline along with the query, the document metadata, and the topic aware outline to the third LLM. In particular, Gupta in view of Madisetti does not expressly disclose a generating a "numerical outline" of the document.

56. As set forth in the rejection of claim 1 above, Gupta discloses using artificial intelligence with natural language processing to generate outlines of submitted documents in order to provide answers to an input question. In doing so, Gupta in view of Madisetti discloses feeding LLMs a document outline, the query, and any additional metadata context in order to generate a response for the user's query.

57. A numerical outline seems to be an outline of a document's numerical values. Although it is unclear based on Applicant's specification. Given this interpretation, Gupta discloses extracting entities, which can include important dates. Gupta at para. 0032. Dates are often represented with numbers. Since Gupta discloses extracting dates of a document and producing an outline of a document based on extracted information, at the time before the effective filing date of the instant application, it would have been obvious to one of ordinary skill in the art to modify Gupta in view of Madisetti by adding the features of submitting the document to a fifth large language model along with a fifth prompt prompting the fifth LLM to generate a numerical outline of the document, receiving the numerical outline of the document, and submitting the numerical outline along with the query, the document metadata, and the topic aware outline to the third LLM. The motivation for doing so would have been to increase the accuracy of a response. Madisetti at col. 7, lines 30-45.

58. In regards to **claim 6**, Gupta in view of Madisetti discloses the method of claim 5, further comprising:

a. submitting the query to a sixth LLM along with a sixth prompt prompting the sixth LLM to transform the query to a first outline request for generate the topic aware outline of the document and into a second outline request for generating the numerical outline of the document (Madisetti at col. 7, lines 30-54);

b. receiving the first outline request and the second outline request from the sixth LLM (Madisetti at col. 7, lines 30-54);

c. including the first outline request in the first prompt (Madisetti at col. 7, lines 30-6; col. 8, lines 1-6); and

d. including the second outline request in the fifth prompt. (Madisetti at col. 7, lines 30-6; col. 8, lines 1-6)[8]

59. In regards to **claim 9**, Gupta in view of Madisetti discloses the method of claim 1, wherein determining relevant document sentences based on embeddings of sentences from the document and of the selected sentence of the natural language response comprises determining relevance based on vector distance between embeddings of the document sentences and of the selected sentence. Madisetti at col. 7, lines 55-67.[9]

60. In regards to **claim 10**, Gupta in view of Madisetti discloses the method of claim 6, wherein the first LLM, the second LLM, the third LLM, the fourth LLM, the fifth LLM, and the sixth LLM are selected based on one or more of latency, maximum context window size, accuracy of results, quality of results or resource usage. Madisetti at col. 3, lines 30-37.[10]

61. In regards to **claim 11**, Gupta in view of Madisetti discloses the method of claim 1, further comprising:

---

[8] Madisetti discloses deriving prompts based on a user's input prompt. This is interpreted as transforming the query into various prompts for appropriate LLMs, such as for creating a first outline and for creating a second outline.

[9] The similarity between embeddings is interpreted as using vector distance since the embeddings are vectors.

[10] LLMs are chosen based on levels of accuracy.

a. using a seventh LLM to classify the query into a first query classification of at least two possible query classifications (Madisetti at col. 7, lines 30-54);

b. determining a seventh prompt, based on the first query classification, for prompting the third LLM to generate a natural language response to the query based at least in part on the document metadata and the document outline (Madisetti at col. 7, lines 30-54); and

c. providing the seventh prompt to the third LLM. (Madisetti at col. 7, lines 30-54).[11]

62. In regards to **claim 12**, Gupta in view of Madisetti discloses the method of claim 11, wherein two or more of the first LLM, the second LLM, the third LLM, the fourth LLM, the fifth LLM, the sixth LLM, and seventh LLM are a same LLM. Madisetti at col. 6, lines 53-67; col. 1, lines 1-12.[12]

63. In regards to **claim 13**, Gupta in view of Madisetti discloses the method of claim 11, wherein two or more of the first LLM, the second LLM, the third LLM, the fourth LLM, the fifth LLM, , the sixth LLM, and seventh LLM are different LLMs. Madisetti at col. 6, lines 53-67; col. 1, lines 1-12.[13]

64. In regards to **claim 14**, Gupta in view of Madisetti discloses the method of claim 1, wherein the document is a plurality of documents. Gupta at para. 0024.[14]

65. **Claims 15-18** are essentially the same as claims 1-3, respectively, in the form of a computer readable storage media. Gupta at para. 0030. Therefore, they are rejected for the same reasons.

66. In regards to **claim 18**, Gupta discloses a system comprising:

---

[11] Madisetti discloses deriving prompts from a user's query. The prompts are derived based on different categories (i.e., at least two possible query classifications). Each LLM receives a different type of prompt and thus requires a different derived prompt. The system determines which prompt to derive based on the user's query (i.e., classify the query into a first query classification; determining a seventh prompt based on the first query classification).
[12] Various LLMs in the hLLM family can be used. Each can be trained on different training sets or not. This makes two or more the same or different.
[13] Various LLMs in the hLLM family can be used. Each can be trained on different training sets or not. This makes two or more the same or different.
[14] A plurality of documents can be uploaded to the system.

a. one or more computers and one or more storage devices on which are stored instructions that are operable, when executed by the one or more computers, to cause the one or more computers to perform operations (Gupta at para. 0030) comprising:

i. receiving, from a user, a query related to a document (Gupta at para. 0034)[15];

ii. submitting the document to the system to generate an outline of the document (Gupta at paras. 0024-25)[16];

iii. receiving a document outline of the document from the system (Gupta at para. 0024-25);

iv. submitting the document to generate metadata of the document (Gupta at paras. 0032-34)[17]

v. receiving document metadata of the document from the system (Gupta at paras. 0032-34);

vi. submitting at least a portion of the query, the document metadata, and the document outline to the system to generate a natural language response to the query based at least in part on the document metadata and the document outline (Gupta at para. 0034)[18];

vii. receiving the natural language response from the system (Gupta at para. 0034);

viii. providing the natural language response to the user. Gupta at paras. 0034, 0047.

67. Gupta does not expressly disclose (1) submitting the document to a first LLM with a first prompt to generate an outline of the document and using a second LLM with a second prompt to generate

[15] A user submits a question (i.e., query) related to a document
[16] Documents are submitted to the system to generate outlines for them.
[17] The system extracts entities of the document, such as dates and important clauses (i.e., metadata).
[18] The extracted entity information and outline are used by the system to generate an answer to a question in the given context.

metadata of the document and a third LLM with a third prompt to generate a natural language response to the query, (2) determining relevant document sentences based on embeddings of sentences from the document and a selected sentence from the natural language response, (3) submitting the relevant document sentences and the selected sentences form the natural language response to a fourth LLM along with a fourth prompt prompting the fourth LLM to generate a ranking of relevant document sentences based on the relevance of respective relevant document sentences to the selected sentence from the natural language response, (4) determining relevant document sentences based on embeddings of sentences from the document and a selected sentence from the natural language response, (5) submitting the relevant document sentences and the selected sentence from the natural language response to a fourth LLM along with a fourth prompt prompting the fourth LLM to generate a ranking of relevant document sentences based on relevance of respective relevant document sentences to the selected sentence from the natural language response, (6) correlating the selected sentence from the natural language response to a first sentence from the document based on the ranking of relevant document sentences generated by the fourth LLM, and (7) providing the natural language response with a citation to a specific sentence or to specific sentences from the document correlated to the selected sentence from the natural language response. It is noted that Gupta does disclose using an artificial intelligence with natural language processing capabilities to generate the outline of the document and answer questions about the document in natural language. Gupta at paras. 0004, 0034. What is not specifically disclosed is using multiple LLMs to perform these steps, determining relevant document sentences to a selected sentence using embeddings, ranking the relevant sentences, and correlating the selected sentence from the natural language response to a sentence from the document and providing a citation to a sentence from the document that is correlated to the selected sentence.

68. Madisetti discloses a system and method that utilizes a plurality of LLMs for processing documents, such as knowledge documents, and providing answers to queries pertaining to the

knowledge documents. In response to a user prompt (i.e., query), the system performs similarity analysis using prompt embeddings (i.e., correlating) to rank and determine the best context aware answers to return to the user along with citations from the knowledge documents (i.e., providing a natural language response to the user with a citation ...). Madsetti at col. 7, lines 30-67; col. 8, lines 1-7. Madisetti expressly states "the best results are sent to the user along with citations from the knowledge documents." Here, "citations from the knowledge documents" are interpreted as one or more specific sentences from the document.

69. Gupta and Madisetti are analogous art because they are both directed to the same field of endeavor of artificial intelligence processing of documents.

70. At the time before the effective filing date of the instant application, it would have been obvious to one of ordinary skill in the art to modify Gupta by adding the features of (1) submitting the document to a first LLM with a first prompt to generate an outline of the document and using a second LLM with a second prompt to generate metadata of the document and a third LLM with a third prompt to generate a natural language response to the query, (2) determining relevant document sentences based on embeddings of sentences from the document and a selected sentence from the natural language response, (3) submitting the relevant document sentences and the selected sentences form the natural language response to a fourth LLM along with a fourth prompt prompting the fourth LLM to generate a ranking of relevant document sentences based on the relevance of respective relevant document sentences to the selected sentence from the natural language response, (4) determining relevant document sentences based on embeddings of sentences from the document and a selected sentence from the natural language response, (5) submitting the relevant document sentences and the selected sentence from the natural language response to a fourth LLM along with a fourth prompt prompting the fourth LLM to generate a ranking of relevant document sentences based on relevance of respective relevant document sentences to the selected sentence from the natural language response, (6)

correlating the selected sentence from the natural language response to a first sentence from the document based on the ranking of relevant document sentences generated by the fourth LLM, and (7) providing the natural language response with a citation to a specific sentence or to specific sentences from the document correlated to the selected sentence from the natural language response, as disclosed by Madisetti.

71. The motivation for doing so would have been to provide more accuracy with responses. Madisetti at col. 7, lines 30-54.

72. **Claims 19 and 20** are essentially the same as claims 2 and 3, respectively, in the form of a system. Therefore, they are rejected for the same reasons.

73. **Claims 21-23 are rejected under 35 U.S.C. 103 as being unpatentable over Gupta (US Patent Pub 2022/0092097) in view of Madisetti et al. (US Patent 12,001,462) (Madisetti), further in view of Staub et al. (US Patent Pub 2021/0216576) (Staub).**

74. In regards to **claims 21-23**, Gupta in view of Madisetti discloses the method of claim 1, but does not expressly disclose wherein the specific sentences from the document comprise less than ten sentences, less than five sentences, or less than three sentences. It is noted that Madisetti does discloses returning a citation from the knowledge documents to the user and the citation is interpreted as at least one sentence in the knowledge document. However, since Madisetti does not expressly disclose a particular number, Staub is relied upon for disclosing these limitations.

75. Staub discloses a system and method for providing answers to a query about documents. Staub at abstract. The process includes embeddings and language models. Staub at paras. 0027-28. The result of the process sentences with a citation to the document from which the sentences come from.

Staub at Fig. 6; para. 0029. As shown in Fig. 6, each citation corresponds to two sentences. Two sentences are less than three, five, or ten sentences as required by these limitations.

76. Gupta, Madisetti and Staub are analogous art because they are directed to processing of documents to answer a question.

77. At the time before the effective filing date of the instant application, it would have been obvious to one of ordinary skill in the art to modify Gupta in view of Madisetti by adding the features of wherein the specific sentences from the document comprise less than ten sentences, less than five sentences, or less than three sentences, as disclosed by Staub.

78. The motivation for doing so would have been to keep responses uniform for the user interface.

***Response to Amendment***

**Rejection of Claims 4-10 under 35 U.S.C 112(b)**

79. Claims 7 and 8 are cancelled rendering their rejections moot.

80. Applicant's amendment to claims 4-6, 9, and 10 is acknowledged. The rejection to claims 4-6, 9, and 10 under 35 U.S.C. 112(b) is withdrawn.

***Response to Arguments***

**Rejection of Claims 1-20 under 35 U.S.C 101**

81. Claims 7 and 8 are cancelled rendering their rejections moot.

82. Applicant's amendment to claims 15-20 is acknowledged. Consequently, the rejection to claims 15-20 under 35 U.S.C. 101 is withdrawn with regards to the non-statutory medium rejection.

83. Applicant's amendments to claims 1-6 and 9-20 are also acknowledged. However, the amendments are insufficient to overcome the rejections. The amendments essentially incorporate limitations previously rejected in claims 7 and 8. Additionally, Applicant has not presented arguments to

explain why these amendments are sufficient to overcome the rejection. For at least these reasons, claims 1-6 and 9-20 remain rejected under 101 for being directed to an abstract idea without significantly more as set forth in the rejection above.

**Rejection of claims 1-20 under 35 U.S.C. 103**

84. Claims 7 and 8 are cancelled rendering their rejections moot.

85. Applicant's arguments in regards to the rejections to claims 1-6 and 9-20 under 35 U.S.C. 103, have been fully considered but they are not persuasive. In regards to claim 1, Applicant alleges Gupta in view of Madisetti fails to disclose (1) "determining relevant document sentence embeddings of sentences from the document and a selected sentence from the natural language response" and (2) "providing the natural language response to the user with a citation to a specific sentence or to specific sentences from the document correlated to the selected sentence from the natural language response." Remarks at 11.

86. Examiner is required to give claim limitations their broadest reasonable interpretation in light of the specification. However, limitations from the specification are not read into the claims. MPEP 2111.

87. In regards to limitation (1), Applicant argues that Madisetti does not disclose the "document" of the limitation but instead utilizes a "prompt" and therefore does not disclose determining relevant sentences of an input document based on embeddings of those sentences and embeddings of a natural language response. Remarks at 11. Applicant's arguments against Madisetti individually is unpersuasive. One cannot show nonobviousness by attacking references individually where the rejections are based on combinations of references. See *In re Keller*, 642 F.2d 413, 208 USPQ 871 (CCPA 1981); *In re Merck & Co.,* 800 F.2d 1091, 231 USPQ 375 (Fed. Cir. 1986). In this case, Gupta discloses receiving a question (i.e., query) about a document. The document is provided to the system AI, which creates an outline for the document, extracts metadata, and utilizes the outline and metadata to return

a natural language response to the question. Gupta at paras. 0024-25, 0032-34, 0047. What Gupta does not expressly disclose, and what is disclosed by Madisetti, is the use of multiple LLMs, determining relevance between a sentence of a natural language response and sentences in a document using embeddings, ranking the determined relevant sentences, correlating the sentence of the natural language response to a relevant sentence of the document based on rankings, and providing the natural language response and a citation to sentence of the document that was correlated to the sentence of the natural language response. As discussed in the rejection above, Madisetti discloses utilizing multiple LLMs for processing documents, such as knowledge documents, and providing answers to a query pertaining to the knowledge document. Contrary to Applicant's allegation, Madisetti discloses a prompt and "input documents" (i.e., the document of the claim). Madisetti at col. 7, lines 55-57. The prompt is converted to an embedding and compared to embeddings of the input documents to determine relevant documents, which comprise sentences. These relevant documents and their sentences are ranked to determine the best results, which are sent to the user with citations from the relevant documents. Madisetti at col. 7, lines 55-67; col. 8, lines 1-7. For at least these reasons, Gupta in view of Madisetti discloses limitation (1).

88. In regards to limitation (2), as discussed above in regards to limitation (1), Madisetti discloses returning a citation from the knowledge documents, which is interpreted as to one or more "specific sentences," which is what the limitation requires. For at least these reasons, Gupta in view of Madisetti discloses limitation (2).

89. Applicant does not present additional arguments with regards to the remaining limitations. Therefore, Examiner asserts the cited prior art discloses all the limitations of claim 1 for the reasons explained above. In regards to the remaining claims, Applicant refers to the arguments presented in regards to claim 1, which are addressed above. Consequently, the rejection to claims 1-6 and 9-20 under 35 U.S.C. 103 is maintained.

***Additional Prior Art***

90. Additional relevant prior art are listed on the attached PTO-892 form. Some examples are:

91. McElvain et al. (US Patent Pub 2019/0340172) discloses a system and method for generating contextually correct response to a query.

92. Kim et al. (US Patent Pub 2016/0196342) discloses a system and method for plagiarism detection in documents.

93. Khosla et al. (US Patent Pub 2026/0005051) discloses a system and method for processing natural language queries using inline citations.

94. Khamis et al. (US Patent Pub 2017/0300470) discloses a system and method for identifying evidentiary information with citation output.

***Conclusion***

95. Applicant's amendment necessitated the new ground(s) of rejection presented in this Office action. Accordingly, **THIS ACTION IS MADE FINAL**. See MPEP § 706.07(a). Applicant is reminded of the extension of time policy as set forth in 37 CFR 1.136(a).

96. A shortened statutory period for reply to this final action is set to expire THREE MONTHS from the mailing date of this action. In the event a first reply is filed within TWO MONTHS of the mailing date of this final action and the advisory action is not mailed until after the end of the THREE-MONTH shortened statutory period, then the shortened statutory period will expire on the date the advisory action is mailed, and any extension fee pursuant to 37 CFR 1.136(a) will be calculated from the mailing date of the advisory action. In no event, however, will the statutory period for reply expire later than SIX MONTHS from the date of this final action.

97. Any inquiry concerning this communication or earlier communications from the examiner should be directed to Examiner Michael Le whose telephone number is 571-272-7970 and fax number is 571-273-7970. The examiner can normally be reached Mon-Fri 9:30 AM – 6 PM.

98. Examiner interviews are available via telephone, in-person, and video conferencing using a USPTO supplied web-based collaboration tool. To schedule an interview, applicant is encouraged to use the USPTO Automated Interview Request (AIR) at http://www.uspto.gov/interviewpractice.

99. If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Tony Mahmoudi can be reached on 571-272-4078. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

100. Information regarding the status of published or unpublished applications may be obtained from Patent Center. Unpublished application information in Patent Center is available to registered users. To file and manage patent submissions in Patent Center, visit: https://patentcenter.uspto.gov. Visit https://www.uspto.gov/patents/apply/patent-center for more information about Patent Center and https://www.uspto.gov/patents/docx for information about filing in DOCX format. For additional questions, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/MICHAEL LE/
Examiner, Art Unit 2163

/TONY MAHMOUDI/
Supervisory Patent Examiner, Art Unit 2163