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 01/14/2026
FISH & RICHARDSON P.C. (TC)
PO BOX 1022
MINNEAPOLIS, MN 55440-1022

EXAMINER	
LE, MICHAEL	

ART UNIT	PAPER NUMBER
2163	

NOTIFICATION DATE	DELIVERY MODE
01/14/2026	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 (First Inventor to File) 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 12/11/2025.

☐ 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) 1-6,9 and 11-25 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) 1-6,9 and 11-25 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 12/11/2025.

3. Claim 10 is cancelled.

4. Claims 24 and 25 are new.

5. Claims 1-6, 9, and 11-25 are pending.

6. Claims 1-6, 9, 11-23, and 25 are rejected under 35 U.S.C. 103 as being unpatentable over Gupta (US Patent Pub 2022/0092097) in view of Zhou et al. (US Patent Pub 2025/0061286), further in view of Madisetti et al. (US Patent 12,001,462).

7. Claim 24 is rejected under 35 U.S.C. 103 as being unpatentable Gupta (US Patent Pub 2022/0092097) in view of Zhou et al. (US Patent Pub 2025/0061286), further in view of Madisetti et al. (US Patent 12,001,462), further in view of Gardner (US Patent Pub 2024/0296279).

8. 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 Objections

9. Claim 25 is objected to because of the following informalities:

10. In the "augmenting" limitation "large" should not be underlined.

11. In the "submitting the collection" limitation, the punctuations at the end should be corrected.

12. In the last limitation, line 1, a comma and a space should be deleted after "prompt".

13. Appropriate correction is required.

Note on Prior Art Rejections

14. 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

15. 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.

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

17. 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.

18. 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.

19. Claims 1-6, 9, 11-23, and 25 are rejected under 35 U.S.C. 103 as being unpatentable over Gupta (US Patent Pub 2022/0092097) in view of Zhou et al. (US Patent Pub 2025/0061286) (Zhou), further in view of Madisetti et al. (US Patent 12,001,462) (Madisetti).

20. In regards to **claim 1**, Gupta discloses a method performed by one or more computers, the 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);

[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).

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.

21. Gupta does not expressly disclose (1) determining a collection of candidate document sentences based on comparisons of: (i) respective embeddings of each of a plurality of sentences from the document, and (ii) an embedding of the target sentence from the natural language response, (2) submitting the collection of candidate document sentences and the target sentences from the natural language response to a fourth LLM along with a fourth prompt prompting the fourth LLM to generate, as an output of the fourth LLM, a ranking of the collection of candidate document sentences based on a respective relevance of each of the document sentences to the target sentence from the natural language response, (3) selecting one or more of the candidate document sentences as specific supporting sentences from the document that provide support for the target sentence from the natural language response based on the ranking generated using the fourth LLM, (4) augmenting the natural language response generated by the third LLM with one or more citations to the specific supporting sentences from the document that provide support for the target sentence from the natural language response, and (5) outputting the natural language response with one or more citations to the one or more specific supporting sentences from the document.

22. Zhou discloses a system and method for a question and answer system that utilizes LLMs to provide answers to a user provided question and to determine whether a generated answer is supported by the document context that should be used to generate the answer. Zhou at para. 0006.

[4] The extracted entity information and outline are used by the system to generate an answer to a question in the given context.

The method of determining whether an answer is supported includes (1) determining sentences that are similar (i.e., relevant) to the answer (i.e., target sentence) by generating embeddings for the answer and for the sentences of the context information (i.e., documents). The similarity calculation is performed between each respective pair of sentences. Zhou at Fig. 4; paras. 0046-50. (2) The similarity calculations between sentence embedding pairs are analyzed with respect to a predetermined threshold to classify (i.e., generating a ranking) sentence pairs that exceed the predetermined threshold as being sufficiently similar and therefore, provides support for the answer. Zhou at para. 0050. (3) Based on the similarity analysis (i.e., based on the ranking …) the sentences from the context information (i.e., candidate document sentences) exceeding the threshold are determined to support the answer (i.e., selecting one or more of the candidate document sentences as specific supporting sentences from the document that provide support for the target sentence). Zhou at Fig. 4; paras. 0046-50. Zhou further discloses generation of the embeddings and similarity is performed by the processor using a language model (i.e., fourth LLM). Zhou at paras. 0049-50. (4, 5) Zhou further discloses outputting a natural language answer that includes information that identifies which sentences in the natural language context information (i.e., the document) provide support for each sentence in the natural language answer. Indications of the supporting sentences are output alongside the natural language answer. Zhou at para. 0074.

23. Gupta and Zhou are analogous art because they are directed to the same field of endeavor of question and answer systems.

24. 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) determining a collection of candidate document sentences based on comparisons of: (i) respective embeddings of each of a plurality of sentences from the document, and (ii) an embedding of the target sentence from the natural language response, (2) submitting the collection of candidate document sentences and the target

sentences from the natural language response to a fourth LLM along with a fourth prompt prompting the fourth LLM to generate, as an output of the fourth LLM, a ranking of the collection of candidate document sentences based on a respective relevance of each of the document sentences to the target sentence from the natural language response, (3) selecting one or more of the candidate document sentences as specific supporting sentences from the document that provide support for the target sentence from the natural language response based on the ranking generated using the fourth LLM, (4) augmenting the natural language response generated by the third LLM with one or more citations to the specific supporting sentences from the document that provide support for the target sentence from the natural language response, and (5) outputting the natural language response with one or more citations to the one or more specific supporting sentences from the document., as disclosed by Zhou.

25. The motivation for doing so would have been to improve overall efficiency of hallucination detection. Zhou at para. 0071.

26. Gupta in view of Zhou 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. 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. It is also noted that Gupta in view of Zhou discloses determining sentences in contextual documents that provide support for target sentences in the generated natural language answer. Zhou at paras. 0045-50. What is not specifically disclosed is using multiple LLMs to perform these steps.

27. Madisetti discloses a system and method that (1) 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 to rank and determine the best context aware answers to return to the user along with citations from the knowledge documents. 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.

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

29. 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 Zhou 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, as disclosed by Madisetti.

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

31. In regards to **claim 2**, Gupta in view of Zhou and 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]

32. In regards to **claim 3**, Gupta in view of Zhou and 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]

---

[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).

33. In regards to **claim 4**, Gupta in view of Zhou and 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]

34. In regards to **claim 5**, Gupta in view of Zhou and 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 Zhou and Madisetti does not expressly disclose a generating a "numerical outline" of the document.

35. 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 Zhou and 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.

36. 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 Zhou and Madisetti by adding the features of submitting the document to a fifth large language model along with a fifth

[7] The outline is based on topics detected in the document (i.e., topic aware document).

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.

37. In regards to **claim 6**, Gupta in view of Zhou and 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 (Gupta at para. 0024-25; Madisetti at col. 7, lines 30-54);

b. receiving the first outline request and the second outline request from the sixth LLM (Gupta at para. 0024-25; Madisetti at col. 7, lines 30-54);

c. including the first outline request in the first prompt (Gupta at para. 0024-25; Madisetti at col. 7, lines 30-6; col. 8, lines 1-6); and

d. including the second outline request in the fifth prompt. (Gupta at para. 0024-25; Madisetti at col. 7, lines 30-6; col. 8, lines 1-6)[8]

38. In regards to **claim 9**, Gupta in view of Zhou and Madisetti discloses the method of claim 1, wherein determining a collection of candidate document sentences based on comparisons of: (i) respective embeddings of each of a plurality of sentences from the document, and (ii) an embedding of

[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, such as outline requests disclosed by Gupta.

the target sentence of the natural language response comprises: determining relevance based on vector distance between embeddings of document sentences and of the target sentence. Zhou at para. 0050.[9]

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

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).[10]

40. In regards to **claim 12**, Gupta in view of Zhou and 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.[11]

41. In regards to **claim 13**, Gupta in view of Zhou and 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.[12]

42. In regards to **claim 14**, Gupta in view of Zhou and Madisetti discloses the method of claim 1, wherein the document is a plurality of documents. Gupta at para. 0024.[13]

---

[9] The similarity between embeddings is interpreted as using vector distance since the embeddings are vectors.
[10] 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).
[11] 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.
[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] A plurality of documents can be uploaded to the system.

43. **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.

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

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)[14];

ii. submitting the document to the system to generate an outline of the document (Gupta at paras. 0024-25)[15];

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)[16]

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)[17];

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

---

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

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

45. Gupta does not expressly disclose (1) determining a collection of candidate document sentences based on comparisons of: (i) respective embeddings of each of a plurality of sentences from the document, and (ii) an embedding of the target sentence from the natural language response, (2) submitting the collection of candidate document sentences and the target sentences from the natural language response to a fourth LLM along with a fourth prompt prompting the fourth LLM to generate, as an output of the fourth LLM, a ranking of the collection of candidate document sentences based on a respective relevance of each of the document sentences to the target sentence from the natural language response, (3) selecting one or more of the candidate document sentences as specific supporting sentences from the document that provide support for the target sentence from the natural language response based on the ranking generated using the fourth LLM, (4) augmenting the natural language response generated by the third LLM with one or more citations to the specific supporting sentences from the document that provide support for the target sentence from the natural language response, and (5) outputting the natural language response with one or more citations to the one or more specific supporting sentences from the document.

46. Zhou discloses a system and method for a question and answer system that utilizes LLMs to provide answers to a user provided question and to determine whether a generated answer is supported by the document context that should be used to generate the answer. Zhou at para. 0006. The method of determining whether an answer is supported includes (1) determining sentences that are similar (i.e., relevant) to the answer (i.e., target sentence) by generating embeddings for the answer and for the sentences of the context information (i.e., documents). The similarity calculation is performed between each respective pair of sentences. Zhou at Fig. 4; paras. 0046-50. (2) The similarity calculations between sentence embedding pairs are analyzed with respect to a predetermined threshold

to classify (i.e., generating a ranking) sentence pairs that exceed the predetermined threshold as being sufficiently similar and therefore, provides support for the answer. Zhou at para. 0050. (3) Based on the similarity analysis (i.e., based on the ranking …) the sentences from the context information (i.e., candidate document sentences) exceeding the threshold are determined to support the answer (i.e., selecting one or more of the candidate document sentences as specific supporting sentences from the document that provide support for the target sentence). Zhou at Fig. 4; paras. 0046-50. Zhou further discloses generation of the embeddings and similarity is performed by the processor using a language model (i.e., fourth LLM). Zhou at paras. 0049-50. (4, 5) Zhou further discloses outputting a natural language answer that includes information that identifies which sentences in the natural language context information (i.e., the document) provide support for each sentence in the natural language answer. Indications of the supporting sentences are output alongside the natural language answer. Zhou at para. 0074.

47. Gupta and Zhou are analogous art because they are directed to the same field of endeavor of question and answer systems.

48. 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) determining a collection of candidate document sentences based on comparisons of: (i) respective embeddings of each of a plurality of sentences from the document, and (ii) an embedding of the target sentence from the natural language response, (2) submitting the collection of candidate document sentences and the target sentences from the natural language response to a fourth LLM along with a fourth prompt prompting the fourth LLM to generate, as an output of the fourth LLM, a ranking of the collection of candidate document sentences based on a respective relevance of each of the document sentences to the target sentence from the natural language response, (3) selecting one or more of the candidate document sentences as specific supporting sentences from the document that provide support for the target

sentence from the natural language response based on the ranking generated using the fourth LLM, (4) augmenting the natural language response generated by the third LLM with one or more citations to the specific supporting sentences from the document that provide support for the target sentence from the natural language response, and (5) outputting the natural language response with one or more citations to the one or more specific supporting sentences from the document., as disclosed by Zhou.

49. The motivation for doing so would have been to improve overall efficiency of hallucination detection. Zhou at para. 0071.

50. Gupta in view of Zhou 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. 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. It is also noted that Gupta in view of Zhou discloses determining sentences in contextual documents that provide support for target sentences in the generated natural language answer. Zhou at paras. 0045-50. What is not specifically disclosed is using multiple LLMs to perform these steps.

51. Madisetti discloses a system and method that (1) 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 to rank and determine the best context aware answers to return to the user along with citations from the knowledge documents. 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.

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

53. 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 Zhou 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, as disclosed by Madisetti.

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

55. **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.

56. In regards to **claims 21-23**, Gupta in view of Zhou and Madisetti discloses the method of claim 1, 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. Zhou also discloses identifying at least one sentence corresponding to the answer (i.e., specific sentences … that provide support for the target sentence). Accordingly, the combination of Gupta in view of Zhou and Madisetti discloses providing citations to at least one sentence from the documents that provide support for the target sentence.

57. In regards to **claim 25**, Gupta discloses a method performed by one or more computers, the method comprising:

a. receiving, from a user, a query related to a document (Gupta at para. 0034)[18];

b. submitting the document to the system to generate an outline of the document (Gupta at paras. 0024-25)[19], wherein the outline of the document is a topic aware outline of the document (Gupta at para. 0025)[20]

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)[21]

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)[22];

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.

58. Gupta does not expressly disclose (1) determining a collection of candidate document sentences based on comparisons of: (i) respective embeddings of each of a plurality of sentences from the document, and (ii) an embedding of the target sentence from the natural language response, (2) submitting the collection of candidate document sentences and the target sentences from the natural language response to a fourth LLM along with a fourth prompt prompting the fourth LLM to generate, as

[18] A user submits a question (i.e., query) related to a document.
[19] Documents are submitted to the system to generate outlines for them.
[20] The outline is based on topics detected in the document (i.e., topic aware document).
[21] The system extracts entities of the document, such as dates and important clauses (i.e., metadata).
[22] The extracted entity information and outline are used by the system to generate an answer to a question in the given context.

an output of the fourth LLM, a ranking of the collection of candidate document sentences based on a respective relevance of each of the document sentences to the target sentence from the natural language response, (3) selecting one or more of the candidate document sentences as specific supporting sentences from the document that provide support for the target sentence from the natural language response based on the ranking generated using the fourth LLM, (4) augmenting the natural language response generated by the third LLM with one or more citations to the specific supporting sentences from the document that provide support for the target sentence from the natural language response, and (5) outputting the natural language response with one or more citations to the one or more specific supporting sentences from the document.

59. Zhou discloses a system and method for a question and answer system that utilizes LLMs to provide answers to a user provided question and to determine whether a generated answer is supported by the document context that should be used to generate the answer. Zhou at para. 0006. The method of determining whether an answer is supported includes (1) determining sentences that are similar (i.e., relevant) to the answer (i.e., target sentence) by generating embeddings for the answer and for the sentences of the context information (i.e., documents). The similarity calculation is performed between each respective pair of sentences. Zhou at Fig. 4; paras. 0046-50. (2) The similarity calculations between sentence embedding pairs are analyzed with respect to a predetermined threshold to classify (i.e., generating a ranking) sentence pairs that exceed the predetermined threshold as being sufficiently similar and therefore, provides support for the answer. Zhou at para. 0050. (3) Based on the similarity analysis (i.e., based on the ranking …) the sentences from the context information (i.e., candidate document sentences) exceeding the threshold are determined to support the answer (i.e., selecting one or more of the candidate document sentences as specific supporting sentences from the document that provide support for the target sentence). Zhou at Fig. 4; paras. 0046-50. Zhou further discloses generation of the embeddings and similarity is performed by the processor using a language

model (i.e., fourth LLM). Zhou at paras. 0049-50. (4, 5) Zhou further discloses outputting a natural language answer that includes information that identifies which sentences in the natural language context information (i.e., the document) provide support for each sentence in the natural language answer. Indications of the supporting sentences are output alongside the natural language answer. Zhou at para. 0074.

60. Gupta and Zhou are analogous art because they are directed to the same field of endeavor of question and answer systems.

61. 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) determining a collection of candidate document sentences based on comparisons of: (i) respective embeddings of each of a plurality of sentences from the document, and (ii) an embedding of the target sentence from the natural language response, (2) submitting the collection of candidate document sentences and the target sentences from the natural language response to a fourth LLM along with a fourth prompt prompting the fourth LLM to generate, as an output of the fourth LLM, a ranking of the collection of candidate document sentences based on a respective relevance of each of the document sentences to the target sentence from the natural language response, (3) selecting one or more of the candidate document sentences as specific supporting sentences from the document that provide support for the target sentence from the natural language response based on the ranking generated using the fourth LLM, (4) augmenting the natural language response generated by the third LLM with one or more citations to the specific supporting sentences from the document that provide support for the target sentence from the natural language response, and (5) outputting the natural language response with one or more citations to the one or more specific supporting sentences from the document., as disclosed by Zhou.

62. The motivation for doing so would have been to improve overall efficiency of hallucination detection. Zhou at para. 0071.

63. Gupta in view of Zhou 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) submitting the query to a sixth LLM along with a sixth prompt prompting the sixth LLM model to transform the query into a first outline request for generate the topic aware outline of the document, receiving the first outline request from the sixth LLM and including the first outline request in the first prompt, and (3) 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 content window size, accuracy of results, quality of results, or resource usage. 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 (i.e., outline request). Gupta at paras. 0004, 0034. It is also noted that Gupta in view of Zhou discloses determining sentences in contextual documents that provide support for target sentences in the generated natural language answer. Zhou at paras. 0045-50. What is not specifically disclosed is using multiple LLMs to perform these steps and selecting the LLMs based on one or more of the listed criteria.

64. Madisetti discloses a system and method that (1) 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 to rank and determine the best context aware answers to return to the user along with citations from the knowledge documents. 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. Madisetti further discloses choosing LLMs based on their levels of accuracy and workload of the models. Madisetti at col. 3, lines 30-37.

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

66. 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 Zhou 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) submitting the query to a sixth LLM along with a sixth prompt prompting the sixth LLM model to transform the query into a first outline request for generate the topic aware outline of the document, receiving the first outline request from the sixth LLM and including the first outline request in the first prompt, and (3) 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 content window size, accuracy of results, quality of results, or resource usage, as disclosed by Madisetti.

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

68. Gupta in view of Zhou and Madisetti 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, submitting the query to the sixth LLM with the sixth prompt to generate a second outline request for generating the numerical outline of the document, receiving the second outline request from the sixth LLM, and including the second outline request in the fifth prompt. In

particular, Gupta in view of Zhou and Madisetti does not expressly disclose a generating a "numerical outline" of the document. As explained above, Gupta discloses requesting generation of outlines of a document including a topic aware outline. As discussed further below, Gupta seems to disclose subject matter that is a numerical outline. Madisetti, as discussed above, utilizes multiple LLMs of varying levels of accuracy and particular tasks to generate an answer about a document to the user. Madisetti at col. 7, lines 30-67; col. 8, lines 1-6. This includes using an LLM to perform a particular task as needed, such as generating outlines, which can be used in a subsequent prompt to another LLM in sequence. Madisetti at Fig. 14.

69. As set forth 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 Zhou and 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.

70. 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 Zhou and 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, submitting the query to the sixth LLM with the sixth prompt to generate a second outline request for generating the numerical outline of the

document, receiving the second outline request from the sixth LLM, and including the second outline request in the fifth prompt, for the reasons discussed above.

71. The motivation for doing so would have been to increase the accuracy of a response. Madisetti at col. 7, lines 30-45.

72. Claim 24 is rejected under 35 U.S.C. 103 as being unpatentable over Gupta (US Patent Pub 2022/0092097) in view of Zhou et al. (US Patent Pub 2025/0061286) (Zhou), further in view of Madisetti et al. (US Patent 12,001,462) (Madisetti), further in view of Gardner (US Patent Pub 2024/0296279).

73. In regards to **claim 24**, Gupta in view of Zhou and Madisetti discloses the method of claim 1, wherein augmenting the natural language response generated by the third large language model with one or more citations to the specific supporting sentences from the document that provide support for the target sentence from the natural language response; and outputting the natural language response with the one or more citations to the one or more specific supporting sentences from the document, as set forth in the rejection of claim 1 above, but does not expressly disclose presenting one or more citations as interactive hyperlinks in the outputted natural language response that, when selected, navigate to and highlight the corresponding specific supporting sentences in the document. It is noted that Gupta discloses outline elements having reference links, which when selected, navigate the user to the corresponding section of the document. Gupta at para. 0039.

74. Gardner discloses a system and method of enhancing responses by document based large language models. The system provides for natural language responses based on data items, such as documents, to the user. The natural language responses are produced by one or more large language models and contain one or more citations that are presented as hyperlinks to a cited portion of a data

item. Gardner at paras. 0035-36, 0039-40, 0085. Gardner also discloses showing a quote of the relevant portion (i.e., highlights). Gardner at para. 0092.

75. Gupta, Zhou, Madisetti, and Gardner are analogous art because they are directed to the same field of endeavor of model based document processing.

76. 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 Zhou and Madisetti by adding the features of presenting one or more citations as interactive hyperlinks in the outputted natural language response that, when selected, navigate to and highlight the corresponding specific supporting sentences in the document, as disclosed by Gardner.

77. The motivation for doing so would have been to mitigate hallucinations and enhance user confidence in the AI Model answers. Gardner at para. 0080.

Response to Arguments

Rejection of claims 1-6 and 9-23 under 35 U.S.C. 103

78. Applicant's arguments in regards to the rejections to claims 1-6 and 9-23 under 35 U.S.C. 103, have been fully considered but they are not persuasive.

79. In regards to claim 1, Applicant alleges Gupta in view of Zhou and Madisetti fails to disclose (1) "determining a collection of candidate document sentences based on comparisons of: (i) respective embeddings of each of a plurality of sentences from the document, and (ii) an embedding of the target sentence from the natural language response" and (2) "submitting the collection of candidate document sentences and the target sentence from the natural language response to a fourth large language model along with a fourth prompt prompting the fourth large language model to generate, ... a ranking of the collection of the candidate document sentences based on a respective relevance of each of the candidate document sentences to the target sentence from the natural language response." Remarks at

13. In regards to both limitations, Applicant argues while Zhou discloses sentence similarity calculations using embeddings and thresholds to identify hallucinated content, Zhou does not teach using a fourth LLM to generate a ranking of the collection of candidate document sentences based on a respective relevance of each of the candidate document sentences to the target sentence from the natural language response for the purpose of augmenting a natural language response with citations to supporting sentences. Remarks at 13. The Examiner respectfully disagrees.

80. 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.

81. Applicant acknowledges Zhou discloses sentence similarity calculations using sentence embeddings, but does not address that Zhou discloses performing the embedding based similarity calculations using a processor that utilizes a language model, such as GPT, which is an LLM. Zhou at paras. 0049-50. These types of LLMs require use of prompts. Thus, Zhou does disclose using an LLM (i.e., fourth LLM) to generate a ranking of the collection of candidate document sentences as claimed. Moreover, Madisetti discloses a plurality of LLMs, each with various optimizations for different tasks, such as sentence similarity. Madisetti at col. 7, lines 40-45. Madisetti also discloses generating embeddings using embedding models. Madisetti at col. 7, lines 55-65. Therefore, even if Zhou did not disclose use of an LLM to perform the embedding based similarity, Madisetti would make up for the deficiency. For at least these reasons, Gupta in view of Zhou and Madisetti discloses limitations (1) and (2).

82. 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 9-23 under 35 U.S.C. 103 is maintained.

Additional Prior Art

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

84. Khosla et al. (US Patent Pub 2025/0005057) discloses a system and method for processing natural language queries using LLMs.

85. Zhouludev et al. (US Patent Pub 2019/0213221) discloses a system and method for linking documents using citations.

86. Schafer et al. (US Patent 11,620,441) discloses a system and method for inserting citations into a textual document.

Conclusion

87. 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).

88. 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.

89. 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.

90. 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.

91. 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.

92. 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