IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF DELAWARE

FLINN ARTIFICIAL INTELLIGENCE (AI) TECHNOLOGIES, LLC,	)	
Plaintiff,	)	C.A. No.
v.	)	**JURY TRIAL DEMANDED**
ANTHROPIC, PBC,	)	
Defendant.	)	

COMPLAINT FOR PATENT INFRINGEMENT

Plaintiff Flinn Artificial Intelligence (AI) Technologies, LLC ("Plaintiff" or "Flinn") brings this action for patent infringement against Defendant Anthropic, PBC ("Defendant" or "Anthropic"), and alleges as follows:

NATURE OF THE ACTION

1. This case concerns Anthropic's use of Flinn's patented technologies to build and expand one of the fastest-growing artificial-intelligence businesses in history without paying for them.

2. Steven D. Flinn and Naomi Felina Moneypenny are decision-science and machine-learning practitioners who spent their careers building adaptive, learning-based computer systems for large enterprises.

3. By 2015, Mr. Flinn and Ms. Moneypenny had already spent years developing patented technology directed principally to adaptive enterprise systems, recommendation, personalization, behavioral inference, information discovery, and search. They then began a distinct line of research directed to a different technical problem: how neural-network architectures could use multiple layers of attention to capture, refine, and apply context across successive stages

of processing.

4. On August 3, 2015, and again on January 18, 2016, Mr. Flinn and Ms. Moneypenny filed patent applications describing iterative attention-based neural-network architectures. For example, Mr. Flinn and Ms. Moneypenny disclosed architectures in which multiple attentions are directed to representations of different portions of information; a trained neural network generates probabilities based on the prioritization of those attentions; generated probabilities inform further attention and neural-network processing; and resulting outputs may be evaluated, stored, retrieved, verified, used to generate content, or used to perform actions. The patents asserted in this action (the "Asserted Patents," defined below) arose out of these patent applications.

5. Anthropic was founded in 2021. In the years since, it has become one of the fastest-growing companies in history. Anthropic's annualized revenue run rate grew from approximately $10 billion in 2025 to approximately $65 billion by July 2026. In May 2026, Anthropic closed a $65 billion financing round at a post-money valuation of approximately $965 billion. Anthropic's Claude Code product reached approximately $2.5 billion in annualized revenue run rate within approximately nine months of launch. On October 26, 2026, Anthropic filed an S-1 with the SEC for a proposed initial public offering. https://www.anthropic.com/news/s1-sec.[1]

6. Anthropic built that business based on the architectures recited in the Asserted Patents, offering reasoning, retrieval, memory, coding, agentic, and multimodal capabilities that are likewise recited by claims of the Asserted Patents. This action seeks compensation for Anthropic's unauthorized use of Flinn's patented inventions, together with the other relief

[1] https://www.sec.gov/Archives/edgar/data/2130357/000162828026072484/anthropicpbc-sx1ax2.htm.

available under the patent laws of the United States.

THE PARTIES

7. Plaintiff Flinn is a limited liability company organized and existing under the laws of the State of Texas with its principal place of business at 130 South Charlton Street, Woodville, Texas 75979.

8. Defendant Anthropic is a public benefit corporation organized and existing under the laws of the State of Delaware, with a principal place of business at 500 Howard Street, San Francisco, California 94105. Anthropic can be served via its registered agent for service of process, Corporation Service Company, 251 Little Falls Drive, Wilmington, Delaware 19808.

9. Anthropic operates and owns the claude.ai website and the Claude platform, through which it markets, offers, distributes, and provides technical support for its artificial-intelligence products and services, including the Claude family of large language models and associated products such as Claude Chat, Claude Code, Claude Cowork, Claude Design, Claude Enterprise, Claude API access, and hosted or dedicated Claude deployments throughout the United States, including in this District.

10. Anthropic's infringing products include Claude and its chat, agentic, and task-execution functionality, whether offered separately or through a unified experience, including Claude Chat, Claude Code, Claude Cowork, Claude Design, Claude Enterprise, Claude API access, and hosted or dedicated Claude deployments, that are powered by artificial-intelligence models from Anthropic's "Claude" series, including Claude 1, Claude 1.2, Claude 2, Claude 2.1, Claude 3, Claude 3.5, Claude 3.7, Claude 4, Claude 4.1, Claude 5, and all Opus, Sonnet, Haiku, Fable, Mythos, and other versions, variants, or configurations of those models, as well as any later-released, successor, replacement, updated, modified, derivative, or renamed products, services, platforms, systems, versions, or variants of the foregoing that perform the same or substantially

similar accused functionality (collectively, the "Accused Products").

JURISDICTION AND VENUE

11. This is a civil action against Anthropic for patent infringement under the United States patent laws, as amended, 35 U.S.C. §§ 101 *et seq*. (the "Patent Laws"). This action arises out of Anthropic's making, using, selling, offering for sale, and/or importing into the United States certain products and services covered by Flinn's patents identified in this Complaint and/or inducing or contributing to direct infringement of Flinn's patents by its customers, developers, distributors, enterprise users, and platform partners. Accordingly, Flinn seeks relief under the Patent Laws.

12. This Court has subject matter jurisdiction pursuant to 28 U.S.C. §§ 1331 and 1338.

13. This Court has personal jurisdiction over Anthropic because Anthropic is organized and exists under the laws of Delaware, and Anthropic is subject to general jurisdiction in this District.

14. Venue is proper in this District pursuant to 28 U.S.C. § 1400(b) because Anthropic resides in this District. As a domestic corporation, Anthropic resides in this District because it is organized and exists under the laws of the State of Delaware.

15. Anthropic is not regional in character. Anthropic maintains offices, employees, and technical personnel in multiple states; trains, deploys, and operates the Accused Products using data-center and cloud infrastructure located in multiple states; distributes the Accused Products through nationwide and international channels, including third-party cloud platforms; and derives revenue from customers throughout the United States and abroad, including in this District. As set forth below, the personnel, documents, and source code relevant to the Accused Products are distributed across, and accessible from, multiple judicial districts.

16. Anthropic maintains offices across the world and within the United States,

including significant offices in Seattle, Washington, and New York, New York.

17. On information and belief, Anthropic's employees access, use, test, develop, and deploy the Accused Products from locations throughout the United States, including Anthropic's offices and through remote working arrangements offered by Anthropic. Anthropic, *Careers*, https://www.anthropic.com/careers/jobs (last visited Nov. 10, 2026).

18. On information and belief, as of 2026, Anthropic employs thousands of people globally, with estimates ranging from approximately 3,000 to 5,000 employees.

19. On information and belief, Anthropic's research and development of the Accused Products occurs on both the West Coast and the East Coast, spanning locations throughout the country. Anthropic's AI Research and Engineering team responsible for developing, training, and refining the Accused Products advertises at least sixty-seven open positions across the United States, many of which include the possibility of remote work. Anthropic, *Careers*, https://www.anthropic.com/careers/jobs (last visited Nov. 10, 2026).

20. On information and belief, Anthropic actively staffs technical roles directly responsible for the design, training, and operation of the Accused Products in New York City, including positions on the teams responsible for training, fine-tuning, and developing the Accused Products. Anthropic announced plans to lease a sixteen-story office building at 330 Hudson Street in Lower Manhattan's Hudson Square neighborhood as part of Anthropic's plan to double its New York City workforce to 1,000 employees.

21. On information and belief, Anthropic trains, develops, deploys, and operates the Accused Products using a vast and geographically distributed network of data center facilities located throughout the United States. For example, Anthropic's training infrastructure includes Project Rainier, a data center built in New Carlisle, Indiana. Project Rainier extends beyond

Indiana to additional facilities in Mississippi and Pennsylvania.

22. In November 2025, Anthropic announced plans to build its first company-owned data centers in Texas and New York as part of a $50 billion infrastructure investment, with sites coming online throughout 2026. On information and belief, the Accused Products are made, used, tested, trained, and deployed at Project Rainier facilities and through cloud infrastructure accessible throughout the United States, including in this District.

23. On information and belief, Anthropic stores its source code and records electronically so that they are accessible to Anthropic employees across the country. Anthropic's Deputy Chief Information Security Officer has publicly disclosed that Anthropic's engineers develop software on remote virtual machines. Anthropic engineers use cloud-hosted remote development environments to run multiple AI coding agents in parallel across the company's code repositories from any location, rather than relying on local machines. These cloud-based development and deployment systems enable Anthropic personnel in Seattle, New York, and remote locations throughout the United States to access, modify, test, and deploy the source code underlying the Accused Products. On information and belief, Anthropic's source code and technical records are maintained electronically, and may be accessed, reviewed, and produced from locations throughout the United States.

THE INVENTORS AND THE ASSERTED PATENTS

24. Steven D. Flinn has spent his career at the intersection of decision science, machine learning, and enterprise technology. He holds a bachelor's degree with a double major in mathematics and economics and a concentration in computer science, a Master of Science in Decision Analysis and Mathematical Modeling from Stanford University, and a Master of Business Administration from the Kellogg School of Management at Northwestern University.

25. Before turning to machine learning full time, Mr. Flinn served as Vice President,

Strategy and Strategic Alliances, and as Chief Information Officer at Royal Dutch Shell, where he was responsible for the information technology strategy of one of the largest enterprises in the world.

26. Mr. Flinn is the founder and Chief Executive Officer of ManyWorlds, Inc., a research, development, and innovation company specializing in machine-learning-based adaptive systems, decision intelligence, recommender systems, and organizational learning technologies. Under Mr. Flinn's leadership, ManyWorlds, Inc. developed and commercialized a foundational patent portfolio directed to machine learning, recommendation systems, personalization, probabilistic inference, automated machine learning, and explainable artificial intelligence, and licensed or otherwise transacted portions of that earlier portfolio, which does not include the Asserted Patents, with major enterprises. In 2015, Gartner recognized ManyWorlds, Inc., highlighting the enterprise impact of its technology. Gartner, *Cool Vendors in the Digital Workplace, 2015* (2015).

27. Mr. Flinn is also the author of books on organizational learning and artificial intelligence, including *The Learning Layer* and *Optimizing Data-to-Learning-to-Action*; a keynote speaker and executive educator at international conferences and corporate programs; and an advisor to Global 1000 companies on decision science, machine-learning strategy, and performance improvement. Steven D. Flinn, *The Learning Layer: Building the Next Level of Intellect in Your Organization* (2010); Steven Flinn, *Optimizing Data-to-Learning-to-Action: The Modern Approach to Continuous Performance Improvement for Businesses* (2018).

28. Naomi Felina Moneypenny is named as an inventor on more than forty-five patents directed to machine learning. Ms. Moneypenny was a senior business and information-technology strategist at Royal Dutch Shell, where she worked alongside Mr. Flinn. She

subsequently joined ManyWorlds, Inc. and served as Chief Technology Officer of Synxi, a ManyWorlds, Inc. brand, where she led the development team responsible for its adaptive recommendation and machine-learning engine.

29. Before the inventions asserted in this case, Mr. Flinn and Ms. Moneypenny had already developed an extensive body of patented technology concerning adaptive enterprise systems, personalized recommendations, behavioral inference, information discovery, search, and related technologies. That earlier work traced to applications beginning in 2010 and 2011 and addressed, among other things, how computer systems could learn from user behavior and relationships among stored information to identify and recommend information relevant to particular users and contexts. The inventors obtained numerous patents arising from that work.

30. The Asserted Patents arise from a later and distinct generation of the inventors' research. Beginning with applications filed in August 2015 and January 2016, Mr. Flinn and Ms. Moneypenny turned from principally determining what information an adaptive system should identify or recommend to the underlying architecture through which a neural-network-based system could process information, direct and redirect attention, generate and use probabilities, and iteratively generate, evaluate, and act upon resulting content.

31. That later work disclosed architectures in which multiple attentions are associated with representations of different portions of information; neural-network processing generates probabilities based on the prioritization of those attentions; those probabilities may inform further attention and further neural-network processing; and the resulting representations and outputs may be evaluated, stored, retrieved, verified, used to generate additional content, or used to perform actions. As described below, different claims build on these architectures to address reasoning, memory, retrieval and grounding, training and alignment, executable code, agentic action, and

multimodal generation.

32. This later work is the subject of this action. The Asserted Patents do not arise from the inventors' earlier 2010–2011 patent families concerning adaptive recommendation and discovery systems. They trace instead to the inventors' August 2015 and January 2016 applications directed to this new generation of attention-based neural-network technology that now underpins important functionality of modern generative artificial-intelligence systems.

33. The work of Mr. Flinn and Ms. Moneypenny on attention-based neural-network architectures resulted in an issued patent portfolio comprising at least eighteen United States patents, with additional applications pending, including the patents asserted in this action: United States Patent Nos. 12,327,166 (the "'166 Patent"); 12,223,404 (the "'404 Patent"); 12,511,580 (the "'580 Patent"); 12,462,197 (the "'197 Patent"); and 12,462,198 (the "'198 Patent") (collectively, the "Asserted Patents").

34. Each of the Asserted Patents claims priority to U.S. Application No. 15/000,011, filed January 18, 2016, which issued as U.S. Patent No. 10,510,018, and to U.S. Application No. 14/816,439, filed August 3, 2015.

35. The '166 Patent, titled "Iterative attention-based neural network training and processing," was filed on November 27, 2024, and issued on June 10, 2025. *See* Ex. 1.

36. The '404 Patent, titled "Iterative attention-based neural network training and processing," was filed on January 26, 2023, and issued on February 11, 2025. *See* Ex. 2.

37. The '580 Patent, titled "Iterative attention-based neural network training and processing," was filed on December 6, 2024, and issued on December 30, 2025. *See* Ex. 3.

38. The '197 Patent, titled "Iterative attention-based neural network training and processing," was filed on December 6, 2024, and issued on November 4, 2025. *See* Ex. 4.

39. The '198 Patent, titled "Iterative attention-based neural network training and processing," was filed on January 16, 2025, and issued on November 4, 2025. *See* Ex. 5.

40. Mr. Flinn and Ms. Moneypenny are the named inventors of each of the Asserted Patents. By written assignment, Mr. Flinn and Ms. Moneypenny each assigned to Flinn Artificial Intelligence (AI) Technologies, LLC all right, title, and interest in and to the Asserted Patents, the applications from which they issued, and the earlier applications to which they claim priority, including the right to sue for and recover damages for past, present, and future infringement. Those assignments were recorded with the United States Patent and Trademark Office at Reel 76630, Frame 0145.

41. Flinn is the sole and exclusive owner of all substantial rights in each Asserted Patent, including the exclusive rights to make, use, sell, offer for sale, and import the claimed inventions, to exclude others from doing so, and to license the Asserted Patents. The Asserted Patents are not subject to any license, covenant not to sue, encumbrance, or other grant of rights to any third party. Flinn has constitutional and statutory standing to bring this action in its own name and without joinder of any other party.

42. Each of the Asserted Patents is valid and enforceable and was duly issued in full compliance with Title 35 of the United States Code after a full and fair examination by the United States Patent and Trademark Office. Each enjoys a statutory presumption of validity under 35 U.S.C. § 282.

43. Flinn has complied with any applicable requirements of 35 U.S.C. § 287. Flinn does not make, use, offer for sale, sell, or import within the United States any article covered by any claim of the Asserted Patents, and Flinn has not licensed the Asserted Patents to any person or entity that does so. Accordingly, no marking was required, and Flinn is entitled to recover

damages for infringement occurring before the filing of this Complaint, subject to 35 U.S.C. § 286.

FLINN'S INNOVATIONS IN ARTIFICIAL INTELLIGENCE

A. The State of the Art as of the Priority Dates

44. As of the August 2015 and January 2016 priority dates, computer systems used neural networks and semantic knowledge structures to process language and represent relationships among information. Existing semantic approaches, however, generally depended on structures and relationships defined in advance and lacked mechanisms by which the system could automatically adapt those representations through continued use.

45. The specification of the Asserted Patents identifies the need for semantic-based systems that "automatically adapt to use" and support capabilities including "continuing streams of attention and reflection." '166 Patent at 1:36–44. The identified problem was rooted in computer-based information processing: how a neural-network-based system could identify and prioritize contextually significant portions of information, capture and refine semantic relationships among those portions, and apply the resulting machine information during successive stages of processing.

46. Multiple attentions permit a neural-network-based system to process relationships among different represented portions of an input or prompt. Successive attention and neural-network operations permit those relationships to be refined and applied as context during later processing. By carrying that context forward, the system can generate communications that are coherent and responsive to the meaning of the information supplied by the user.

47. More specifically, the specification explains that there was "a need for a system and method that enables such semantic-based approaches to automatically adapt to use, and further, to extend semantic-based approaches so as to enable the automatic generation of more engaging communications that embody characteristics lacking in existing semantic-based systems such as self-awareness, imagination, introspection, continuing streams of attention and reflection, and

communicating creatively with metaphorical constructs and/or wit." *Id.* at 1:36–44.

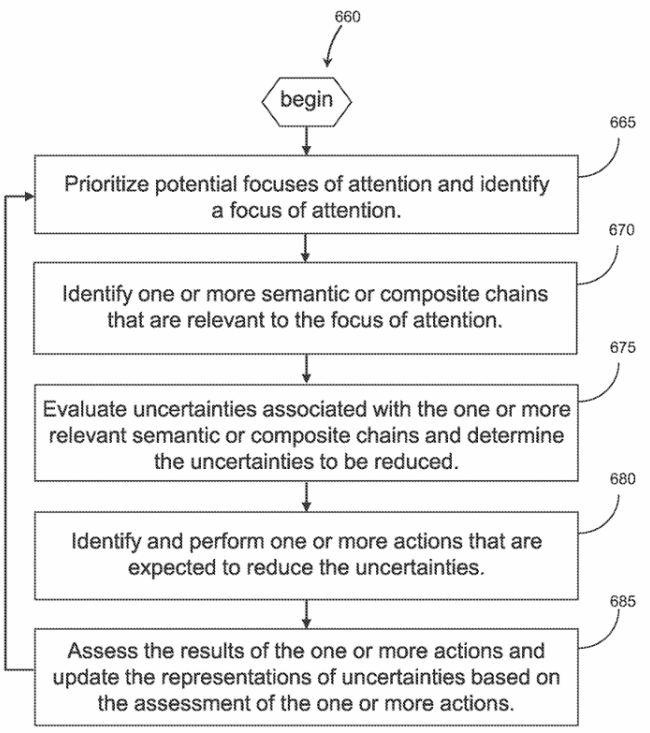
48. Symbolic semantic architectures, including Resource Description Framework ("RDF") systems, that relied on expressly defined structures and relationships for representing and processing knowledge were also known. However, extending or correcting those systems for new subject matter or usage patterns could require additional development rather than allowing the system itself to adapt through continuing use.

B. The Disclosed Solutions

49. The Asserted Patents describe iterative neural-network architectures in which multiple attentions, representations, generated probabilities, and successive processing operations interact. Rather than merely directing a computer to produce a desired response, the disclosed architectures provide mechanisms through which the results of one processing stage may inform and alter later neural-network processing.

50. Depending on the particular embodiment or claim, those mechanisms include prioritizing multiple attentions associated with representations of different portions of information; generating probabilities based on those attentions; using generated probabilities to inform further attention and neural-network processing; generating and evaluating candidate sequences; storing and retrieving machine representations; comparing generated vectors; evaluating and correcting generated content; selecting executable code; performing requested actions; and generating visual content.

51. For example, the specification discusses this iterative, closed-loop attention mechanism in detail through the embodiment summarized in Figure 14C, which has been reproduced below. The specification describes this embodiment as "summariz[ing] the process flow of recursive streams of attention (or 'consciousness') and/or autonomous behaviors of computer-based system 925." '166 Patent at 54:52–55.



Id. at Fig. 14C.

52. The first step comprises "prioritizing potential focuses of attention and selecting a focus of attention based on the prioritization." *Id*. at 54:55–57. The specification explains that "potential focuses of attention can be derived from external sources via, for example, information attained via sensors, from accessible content, or from internally stored information such as saved communications 250c, chains, or images" and that this prioritization may be performed "through application of precedence rules or scoring algorithms" that assign priority based on factors such as "recency considerations" and "probabilistic evaluations and/or value of information considerations." *Id*. at 54:57–55:2. The specification further explains that "[l]anguage-based sources of potential focuses of attention (such as processing speech from a user 200 or processing digitized content) can be directly converted to one or more syntactical elements such as behavioral chains, semantic chains, or composite chains," and that for non-language-based sources such as images, an associated language-based description is "first determined for each image (such as, but

not limited to, by means of the application of a trained neural network to the images . . .)." *Id*. at 55:2–12. A particular focus of attention is then "identified based on the application of the prioritization rules or algorithms." *Id*. at 55:14–16.

53. The second step "comprises identifying one or more chains that are relevant to the identified focus of attention," including by "searching for other chains with the same or similar subjects, predicates, or objects as those of the focus of attention chain(s)," and linking chains that result from that search into composite chains. *Id*. at 55:24–29. The third step "comprises evaluating uncertainties, as represented by corresponding W1 and/or W2-type and/or W3-type weights, of the relevant chains determined by the previous step . . . and determining which uncertainties should be potentially targeted for reduction," a determination made "in accordance with a value of information process." *Id*. at 55:32–38. A "W1-type" weighting "may correspond to a probabilistic confidence level associated with the semantic chain," such that the weighting reflects "the probabilit[y] that the inference is true." *Id*. at 30:56–59, 31:3–5. A "W2-type" weighting may reflect "the probability the associated semantic chain validly applies semantically in the context of interpreting specified content." *Id*. at 31:28–29, 31:50–56. The specification ties these probability constructs directly to neural-network processing, disclosing that "neural network-based systems are trained to make inferences of relevant semantic chains from text and/or images and to inform W1-type weights associated with the inferred semantic chains." *Id*. at 31:14–17.

54. The fourth step "comprises identifying and performing one or more actions that are expected to reduce the potentially targeted uncertainties," a step that may weigh the expected "cost" of a candidate action against "the degree to which the one or more candidate actions can be expected to reduce the uncertainty." *Id*. at 55:39–48. Candidate actions may include

"generating interrogative communications 250c directed to external agents such as users 200 or directed internally to (i.e., self-directed) the computer-based system," formed "by transforming the chain(s) that is associated with the uncertainty that is targeted for reduction into an appropriate question." *Id*. at 55:56–65. The fifth step "comprises assessing the results of the one or more candidate actions that are actually performed and updating the representations of the uncertainties (i.e., W1 and/or W2 and/or W3-type weightings or probabilities) according to the assessment of the results of the actions." *Id*. at 56:23–28. The specification explains the recursive nature of this architecture: "After probabilities that embody uncertainties are updated based on the result of performed actions . . . the first step of the process 665 is again invoked. This closed loop process enables continuous, autonomous learning by [a] computer-based system." *Id*. at 56:53–59.

55. These disclosures describe an integrated computer architecture in which neural-network processing, representations, weights, probabilities, attention prioritization, evaluation, and responsive action may inform one another across successive stages. The disclosed mechanisms do not merely describe desired informational results. They address ways in which outputs from neural-network processing may inform later processing and actions.

C. The Scale and Diversity of the Claimed Subject Matter

56. The Asserted Patents contain over one thousand issued claims. Those claims differ in statutory class, scope, structure, ordered combination, and claimed subject matter. Among these claims are claims directed to one or more of: (1) Foundation Model Architecture, Training, and Alignment; (2) Reasoning and Inference Control; (3) Memory and Personalization; (4) Retrieval, Search, and Grounding; (5) Agentic and Developer Workflows; and (6) Multimodal Processing and Generation.

57. These areas are not allocated one-to-one among the patents. A single Asserted Patent may contain claims addressing several areas, and claims associated with the same area may

recite materially different structures, processing sequences, and operations. The following table identifies, for each Asserted Patent, the number of issued claims, the statutory classes present, and some exemplary technology areas present among its claims:

Patent	Issue Date	Issued Claims	Statutory Classes Present	Exemplary Technology Areas Present
’166 Patent	Jun. 10, 2025	397	Method	Foundation Model Architecture, Training, and Alignment; Reasoning and Inference Control; Memory and Personalization; Retrieval, Search, and Grounding; Agentic and Developer Workflows; Multimodal Processing and Generation
’404 Patent	Feb. 11, 2025	375	Method; System	Foundation Model Architecture, Training, and Alignment; Reasoning and Inference Control; Memory and Personalization; Retrieval, Search, and Grounding; Agentic and Developer Workflows; Multimodal Processing and Generation
’580 Patent	Dec. 30, 2025	65	Method	Reasoning and Inference Control; Retrieval, Search, and Grounding
’197 Patent	Nov. 4, 2025	128	Method	Foundation Model Architecture, Training, and Alignment; Reasoning and Inference Control; Memory and Personalization; Retrieval, Search, and Grounding
’198 Patent	Nov. 4, 2025	115	Method	Foundation Model Architecture, Training, and Alignment; Reasoning and Inference Control; Retrieval, Search, and Grounding; Agentic and Developer Workflows

58. The claims discussed below are identified as examples of the claimed technologies. They are not the complete set of claims that Flinn may identify as infringed, and Flinn will identify asserted claims in accordance with the Court's scheduling order and applicable procedures.

59. Because the discussed claims are examples, the discussion of a claim does not allege or concede that the discussed claim shares the focus, character as a whole, claimed advance, or inventive concept of any other claim, including other claims of the same Asserted Patent or other claims associated with the same technology area.

60. Likewise, the six technology areas reflect applications in which the claimed technologies may be used. They do not establish that the claims grouped within an area are materially similar to one another. Within each area, the claims differ in statutory class, recited operations, relationships among those operations, intermediate results generated and used, and the particular technological problem addressed.

61. Nothing in this section constitutes a proposed construction of any claim term, and Flinn expressly reserves its positions on claim construction.

D. The Claimed Technologies

62. As of the August 2015 and January 2016 priority dates, conventional computer systems did not perform the particular processing arrangements recited in the claims of the Asserted Patents. The claims do not merely place familiar information-processing functions on a generic computer. They recite improvements to the conventional functioning of a computer by employing specific relationships through which machine-generated probabilities, representations, vectors, evaluations, selections, and correspondences influence or supply later computer processing.

63. For example, conventional neural-network systems did not use probabilities generated from the prioritization of one plurality of attentions as the basis for prioritizing a further

plurality of attentions in the ordered arrangement recited by Claim 1 of the '166 Patent. By using the results of one attention-processing stage to direct a later stage, that arrangement enables the system to refine which represented portions of the input receive attention as processing progresses.

64. Conventional systems likewise did not store and retrieve a representation generated or used during attention-based processing for use in prioritizing other attentions in a later generation operation, as recited by Claim 373 of the '166 Patent, for example. By using the retrieved representation to participate in later attention prioritization, the claimed arrangement enables context developed during earlier processing to influence subsequent neural-network processing.

65. Other claims recite different departures from conventional computer operation. For example, Claim 1 of the '580 Patent recites generating a vector representation of an input through multiple-attention-based neural-network application; searching and comparing that representation against stored vector representations generated through the same type of application; identifying content through that comparison; and using the identified content in a further neural-network generation.

66. As a few more examples, Claim 2 of the '197 Patent recites using responsive human input in both the system's evaluation of generated content and its automatic generation of corrected content. Claim 1 of the '198 Patent recites generating probabilities associated with subsets of executable computer-code elements and using those probabilities to select code elements. Claims 357 and 274 of the '404 Patent recite, respectively, using neural-network-generated elements to perform a requested computer action and using learned text-to-pixel correspondences and attention to generate visual output.

67. In each arrangement, newly generated machine information has operative

significance because the computer uses it in a subsequent processing operation. The claimed arrangements thereby change how the computer represents information, identifies contextually significant information, carries context forward, selects information for further processing, and generates subsequent outputs and performs subsequent actions.

68. These particular arrangements, considered both individually and as ordered combinations, were not well-understood, routine, or conventional to a person of ordinary skill in the art as of the August 2015 and January 2016 priority dates.

69. The claims do not preempt attention-based neural-network processing generally. They recite particular ordered arrangements of attention prioritization, probability generation, representation storage and retrieval, vector comparison, code-element selection, and action performance. Neural-network systems may process language, retrieve content, generate code, and perform actions without practicing those recited arrangements—for example, by processing an input through a single undifferentiated attention operation, by retrieving content through literal term matching rather than comparison of neural-network-generated vector representations, or by generating code without generating probabilities associated with corresponding subsets of executable code elements. The claims therefore leave those and other approaches available to others.

70. Anthropic publicly presents the accused reasoning, extended thinking, memory, retrieval, agentic, and multimodal capabilities as technical advances that differentiate its products and justify its model tiers and pricing, and Anthropic's interpretability research analyzes the attention mechanisms at issue as recent technical advances enabling its products and services. A defendant that markets these capabilities as recent technical advances cannot credibly characterize the same arrangements as well-understood, routine, or conventional in 2016.

71. The claimed technologies address problems arising in neural-network architecture, training, reasoning, memory, retrieval, grounding, code processing, action execution, and multimodal processing through particular arrangements of computer operations and intermediate representations. Viewed in light of the claim language and specification, the focus of the claims and their character as a whole concern particular mechanisms for computer and neural-network processing, rather than merely the use of generic computers to perform longstanding human activities.

1. Foundation Model Architecture, Training, and Alignment

72. Foundation-model architecture, training, and alignment determine how generative artificial-intelligence systems process input, carry information across stages, generate output, and adjust that output based on evaluation and feedback. Technical advances in these areas affect whether a neural-network-based system can understand relationships within previously unencountered input, adapt its processing to that input, and generate coherent and responsive output.

73. A technological problem rooted in this area concerns how a neural-network-based system can identify and develop the semantic meaning of an input rather than merely process its syntactical elements in isolation. The specification identifies the need for semantic-based systems that "automatically adapt to use" and support capabilities including "introspection, continuing streams of attention and reflection." '166 Patent at 1:36–44. Successive attention and neural-network operations allow relationships among represented portions of the input to be captured, refined, and applied during later processing. Claims of the Asserted Patents address different aspects of this problem through particular arrangements of attentions, probabilities, representations, generated content, evaluation, and feedback.

74. For example, Claim 1 of the '166 Patent recites:

> 1. A computer-implemented method, comprising:
>
> at a system with access to a trained computer-implemented neural network:
>
> causing access to information including a first plurality of syntactical elements;
>
> causing prioritization of a first plurality of attentions including a first attention associated with a representation of a first subset of the first plurality of syntactical elements, and a second attention associated with a representation of a second subset of the first plurality of syntactical elements;
>
> causing generation, by application of the trained computer-implemented neural network, of a first plurality of probabilities based on the prioritization of the first plurality of attentions;
>
> causing prioritization of a second plurality of attentions including a third attention and a fourth attention, based on the first plurality of probabilities generated by the application of the trained computer-implemented neural network;
>
> causing generation, by application of the trained computer-implemented neural network, of a second plurality of probabilities based on the prioritization of the second plurality of attentions;
>
> causing generation of a second plurality of syntactical elements based on the second plurality of probabilities generated by the application of the trained computer-implemented neural network; and
>
> causing a communication to be sent to a user.

'166 Patent, Cl. 1.

75. Existing symbolic and RDF-based semantic systems depended on predefined structures and relationships and could not adapt those representations through continued use. Processing an input through a single undifferentiated operation provides no mechanism for successively identifying, refining, and carrying forward the contextual relationships that give the input its semantic meaning. Claim 1 addresses that technological deficiency. By carrying context forward through those successive attention and probability-generation operations, the claimed architecture provides a mechanism for developing and applying relationships among

syntactical elements in previously unencountered information.

76. The claimed architecture uses successive rounds of attention and probability generation to develop richer representations of the accessed information as processing progresses. Probabilities generated through one prioritization of attentions inform the prioritization of further attentions, allowing the system to refine which represented portions of the input are contextually significant and apply that developing context during subsequent neural-network processing. The resulting syntactical elements are generated from probabilities produced after this successive attention processing.

77. This architecture allows a computer to develop and apply contextual relationships within an input that the system has not previously encountered. Rather than being limited to predetermined semantic structures or previously programmed responses, the neural-network-based system can adapt its processing to the particular syntactical elements supplied by the user and generate a communication that is coherent and responsive to their meaning. The claimed technology enables new computer functionality through which semantic context is developed across successive processing stages and used to generate responsive output.

78. Other claims of the Asserted Patents in this technology area describe further technical arrangements including prompt-wide representations, positional and relational information, single-data-structure processing, selective attention, or successive weight updating. These claims may address different technical problems or describe different solutions than Claim 1 of the ’166 Patent. Claim 1 of the ’166 Patent is not alleged or conceded to represent these other claims merely because they share aspects of attention-based neural-network processing.

79. The Asserted Patents also include claims addressing training and alignment mechanisms distinct from the foundational processing arrangement described above. For

example, Claim 2 of the '197 Patent recites:

> 2. The method of claim 1, wherein the user includes a human user, the access is caused to the second plurality of syntactical elements in response to a receipt of a prompt from the user, and further comprising:
>
> before causing the determination whether the first scenario represents objective reality, causing an other communication including the third plurality of syntactical elements to be sent to the user;
>
> receiving, from the user, a response to the other communication; and
>
> in response to receiving the response to the other communication, causing the determination whether the first scenario represents objective reality based on the response, such that, based on the determination that the first scenario does not represent objective reality, the fourth plurality of syntactical elements is caused to be automatically generated based on the response.

'197 Patent, Cl. 2.

80. Claim 2, considered together with Claim 1 from which it depends, addresses a technological problem rooted in neural-network generation: how to align generated output with an expected or correct result when the network initially generates an inaccurate response. Neural networks estimate outputs based on learned relationships in training information, creating the possibility of error between the generated result and the result that accurately represents objective reality. On its own, a neural network merely produces an initial estimate and lacks a mechanism for evaluating that estimate against responsive information and using the identified error to guide corrective neural-network generation.

81. The claimed technology integrates human feedback into a continuing neural-network evaluation and correction process. This feedback arrangement allows the computer to use information identifying or reflecting error in an initial neural-network output to produce a better-aligned result. The user's response is not merely collected, displayed, or stored after generation. It becomes operative in both the evaluation of the estimated output and the neural

network's generation of corrected content. The system can compare generated content against responsive information, identify misalignment, and automatically produce a new result directed toward the expected or objectively accurate outcome.

82. Other claims within the Asserted Patents address additional training and alignment mechanisms, including feedback-informed evaluation, correction of generated content, successive updating of weights, and other uses of responsive input during neural-network processing. Those claims may address different technical problems or employ different processing arrangements and are not alleged or conceded to be represented by Claim 2 of the '197 Patent.

2. Reasoning and Inference Control

83. Reasoning and inference control are important to the ability of an artificial-intelligence system to address complex prompts that cannot reliably be resolved through a single generation attempt or initial estimate. These capabilities allow the system to develop, examine, and evaluate intermediate information before producing a response, improving the quality of its output on tasks involving multiple possible approaches, successive analytical steps, or uncertain intermediate conclusions.

84. A technological problem rooted in this area concerns how a neural-network-based system can perform a continuing process of internal evaluation and reflection rather than rely only on an initial generation path. The specification identifies the need for systems capable of "introspection, continuing streams of attention and reflection." '166 Patent at 1:36–44. A system lacking such a mechanism may produce an initial response without developing alternatives, evaluating intermediate information, or using the results of that evaluation to improve later generation. Claims of the Asserted Patents address this problem through different arrangements for generating, evaluating, selecting, representing, and applying intermediate neural-network information.

85. For example, Claim 3 of the ’166 Patent recites:

> 3. The method of claim 1, wherein:
>
> a response perceptible to the user that includes a human user is caused based on a prompt, the prompt includes the first plurality of syntactical elements, and at least one operation is caused based on the second plurality of syntactical elements to generate a representation of a rationale for the second plurality of syntactical elements such that the representation of the rationale is used to generate a third plurality of syntactical elements on which the response is based, the at least one operation including:
>
> generation, utilizing the trained computer-implemented neural network, of a plurality of sequences of elements, where each of the plurality of sequences of elements includes multiple elements;
>
> evaluation of the plurality of sequences of elements;
>
> selection of at least one of the plurality of sequences of elements, based on the evaluation; and
>
> inclusion of a representation of the at least one of the plurality of sequences of elements with the representation of the rationale.

’166 Patent, Cl. 3.

86. Claim 3, considered together with Claim 1 from which it depends, addresses the technological problem of how a neural-network-based system can develop and use evaluated intermediate information when generating a response.

87. The claimed arrangement gives a computer a form of continuing internal evaluation rather than limiting generation to an initial attempt. The system can therefore produce a response informed by evaluated intermediate information, allowing it to handle complex tasks with greater analytical depth and generate higher-quality output than a process dependent solely on a single, unevaluated generation path.

88. Other claims address displayed explanations, iterative reasoning, explanation-based revision, value-of-information controls, and reasoning applied to executable code.

3. Memory and Personalization

89. A model's ability to retain and reuse previously generated information may affect the continuity, efficiency, and usefulness of later processing. Memory and personalization allow the system to adapt through use, preserve context developed during earlier processing, and apply that context when responding to later information rather than treating every generation as an isolated event.

90. A technological problem rooted in this area concerns how information developed during neural-network processing can remain available and become operative during later attention-based generation. Merely preserving text or maintaining a record of an earlier interaction does not provide a mechanism through which a machine representation developed during earlier processing can be retrieved and used to affect how the network processes later information. Claims of the Asserted Patents address different aspects of this problem through storage, retrieval, reuse, persistent model information, prior-interaction processing, and preference-based generation and retrieval.

91. For example, Claim 373 of the '166 Patent recites:

> 373. The method of claim 1, wherein the representation of the first subset of the first plurality of syntactical elements is stored, and the stored representation of the first subset of the first plurality of syntactical elements is retrieved for use in prioritizing one or more other attentions associated with one or more representations of one or more other syntactical elements that are utilized for the generation of the second plurality of syntactical elements.

'166 Patent, Cl. 373.

92. Claim 373, considered together with Claim 1 from which it depends, addresses how a neural-network-based system can adapt through use by preserving a representation developed during attention-based processing and applying that representation in later generation. Without an integrated storage-and-reuse mechanism, context developed during one processing operation

ceases to affect the system after that operation ends.

93. This reuse allows context developed during earlier processing to influence how the system processes and responds to later information. The computer can thereby adapt its subsequent generation based on information developed through prior use rather than beginning each processing operation without the benefit of that machine-developed context. The claimed technology provides a particular mechanism for continuity and adaptation across attention-based neural-network processing.

94. Other claims of the Asserted Patents within this technology area address additional memory and personalization mechanisms, including reuse across multiple prompts, persistence of learned weights, inference of user preferences from prior interactions, generation of preference representations, and preference-based content retrieval. Those claims address different uses of prior information and are not alleged or conceded to be represented by Claim 373.

4. Retrieval, Search, and Grounding

95. Retrieval, search, and grounding expand the information available to an artificial-intelligence system beyond what is encoded in model parameters or included directly in an immediate prompt. These capabilities allow a system to identify contextually relevant external information and use that information in generating new content. They may also improve the currency, specificity, factual support, and source grounding of model output.

96. A technological problem in this area concerns how a neural-network-based system can identify relevant external content and integrate that content into subsequent generation. Conventional search systems may identify documents containing specified words, while a foundation model may generate responsive content from an input. Neither operation alone necessarily provides an integrated mechanism for creating neural-network representations of the input and stored content, comparing those representations, identifying content through that

comparison, and using the identified content in further neural-network generation. Claims of the Asserted Patents address different aspects of this problem through vector representations, searches, comparisons, content identification, graph and cluster structures, post-generation evaluation, and corrective processing.

97. For example, Claim 1 of the '580 Patent recites:

> 1. A computer-implemented method, comprising:
>
> at a system including one or more processors and one or more memories in communication with the one or more processors and with instructions stored therein:
>
> causing access to a first plurality of syntactical elements;
>
> based on the first plurality of syntactical elements, causing generation, by multiple-attention-based neural network application, of a vector representation of the first plurality of syntactical elements, the multiple-attention-based neural network application involving application of a computer-implemented neural network that is trained utilizing hardware that processes representations of training syntactical elements based on multiple training attentions that correspond to representations of different subsets of the training syntactical elements;
>
> causing a search with respect to a plurality of stored vector representations each representing at least one of a plurality of distinct items of content and each generated by multiple-attention-based neural network application;
>
> based on the search, causing a comparison of the vector representation of the first plurality of syntactical elements with at least a portion of the plurality of stored vector representations;
>
> based on the comparison, causing identification of one or more of the plurality of distinct items of content;
>
> based on the one or more of the plurality of distinct items of content, causing generation, by multiple-attention-based neural network application, of a second plurality of syntactical elements; and
>
> causing a communication, that is based on the second plurality of syntactical elements, to be sent to a user.

'580 Patent, Cl. 1.

98. Claim 1 addresses the technological problem of how a neural-network-based system can identify external content through machine-generated representations and use the identified content in further generation. A system that relies only on literal matching may fail to identify content based on relationships reflected in neural-network representations, while a system that only returns retrieved content may not generate a new response incorporating that content.

99. The claimed arrangement enables a computer to produce a new result informed by content identified through attention-based vector comparison, rather than merely returning a preexisting search result or generating exclusively from the immediate input.

100. Other claims of the Asserted Patents in this technology area recite materially different mechanisms, including hybrid semantic and literal retrieval, hierarchical navigation among vector clusters, retrieval within a single document, representations derived from a graph structure, search and correction performed after an initial generation, logical evaluation of generated content, confidence-based assessment, and evaluation through a specialized neural-network subsystem. Those claims recite different representations, different search and comparison operations, and different relationships between retrieval and generation. Claim 1 of the ’580 Patent is not alleged or conceded to represent them merely because each may involve external information.

5. Agentic and Developer Workflows

101. Agentic and developer capabilities allow artificial-intelligence systems to move beyond generating natural-language content to pursue broader goals involving multiple discrete tasks, including software development, tool use, interactions with computer environments, and execution of requested operations. These capabilities are important because useful computer action must reflect the context and objective of the user’s request, not merely resemble code or describe an action at a general level.

102. One technological problem rooted in neural-network software code generation concerns how a system can understand the context reflected in input syntactical elements and use that context to generate or select executable code responsive to a task the system has not previously encountered. A system that treats code generation as an undifferentiated prediction may produce syntactically plausible code that does not correspond to the supplied code, development context, or requested operation. Claims of the Asserted Patents address different aspects of this problem through attentions, probabilities, executable code elements, evaluation, execution results, revision, and communication with other systems.

103. For example, Claim 1 of the '198 Patent recites:

> 1. A computer-implemented method, comprising:
>
> at a system including one or more processors, and one or more memories in communication with the one or more processors and with one or more programs stored therein:
>
> causing access to a computer-implemented neural network trained on computer code so that the trained computer-implemented neural network is configured for probability generation in connection with a plurality of executable computer code elements for selection therefrom;
>
> causing access to a first plurality of syntactical elements;
>
> causing a first attention to be directed to a representation of a first subset of the first plurality of syntactical elements;
>
> causing a second attention to be directed to a representation of a second subset of the first plurality of syntactical elements, for use by the trained computer-implemented neural network;
>
> causing generation, by application of the trained computer-implemented neural network, of a plurality of probabilities that are each associated with a corresponding subset of the plurality of executable computer code elements, based on the second attention;
>
> causing a selection of one or more of the plurality of executable computer code elements based on the plurality of probabilities; and
>
> causing a communication to be sent to a user.

’198 Patent, Cl. 1.

104. Claim 1 addresses the technical problem of how a neural-network-based system can use context developed from input syntactical elements to select useful executable computer code elements. Attention is applied to represented portions of the input so that code generation is informed by the particular relationships and context reflected in the information provided to the network.

105. The claimed arrangement connects that attention-based processing to probabilities associated with corresponding subsets of executable computer code elements. Those probabilities provide the basis for selecting code elements suited to the represented input and its context. The selection therefore emerges from a defined neural-network processing pathway rather than from a general instruction to write code.

106. This architecture enables the computer to create new and useful executable computer code responsive to previously unencountered input. The system can account for the supplied syntactical and code context, generate probability information associated with available code elements, and select code suited to the task presented by the user. It thereby enables the computer to produce new executable content that advances the requested development task, rather than merely retrieving existing code or generating text that superficially resembles code.

107. Other claims of the Asserted Patents within this technology area address different developer mechanisms, including logical evaluation of candidate code, system self-interrogation, search-based evaluation, evaluation of code execution results, vector retrieval of code, revision of generated code, execution of selected code, and transmission of code to another system.

108. Claims within this technology area also address how neural-network processing can enable a computer to act autonomously in furtherance of a user’s objective, rather than merely

generating content describing what should be done.

109. For example, Claim 357 of the '404 Patent, considered together with Claim 10 from which it depends, recites:

> 357. The system of claim 10, wherein the system is configured such that:
>
> the first plurality of syntactical elements represents a request for the system to perform an action other than generating content utilizing the trained computer-implemented neural network and other than causing the communication to be sent to the user that includes a human user;
>
> the action is performed by the system based on the second plurality of syntactical elements; and
>
> the communication reflects the performing of the action.

'404 Patent, Cl. 357.

110. Claim 357, considered together with Claim 10, addresses a technological problem rooted in agentic computing: how a neural-network-based system can translate the meaning and objective of a user's prompt into performance of a responsive computer action without that particular action having been coded in advance. A conventional computer generally performs actions according to predefined instructions specifying the operation to be executed. A system may also generate text describing a proposed action without possessing a mechanism for using its understanding of the prompt to perform the action itself.

111. Claim 10 provides the underlying neural-network architecture through which successive probabilities and attentions process different represented portions of the input and generate syntactical elements responsive to the user's request. Claim 357 connects the syntactical elements generated through that processing to performance of an action distinct from generating content or sending a communication.

112. Other claims of the Asserted Patents within this technology area address potentially different agentic mechanisms, including physical actions, operations not perceptible to a human

user, explanation-informed actions, execution of selected computer code, communication of code to another system for execution, interaction with other systems, and control of a device. Claim 357 of the '404 Patent is not alleged or conceded to represent those claims merely because each may relate to actions extending beyond content generation.

6. Multimodal Processing and Generation

113. Multimodal processing and generation allow artificial-intelligence systems to operate across different forms of information, including language and visual content. These capabilities expand the system beyond text-only inputs and outputs and permit it to understand, relate, and generate information across different representational domains.

114. A technological problem rooted in this area concerns how a neural-network-based system can develop and apply relationships between syntactical information and visual information. Language and images are represented differently within a computer. A textual description of a desired image does not itself provide a mechanism for translating semantic relationships among syntactical elements into patterns of pixels, just as access to an image does not itself allow the system to understand the image in the context of accompanying language. Claims of the Asserted Patents address different aspects of this cross-modal problem through attention-based processing of text and images, learned correspondences, image generation, reference images, visual explanations, video generation, and training-image generation.

115. For example, Claim 274 of the '404 Patent, considered together with Claim 10 from which it depends, recites:

> 274. The system of claim 10, wherein the one or more programs include instructions for causing:
>
> the trained computer-implemented neural network to be further trained utilizing training syntactical elements and a first plurality of images, such that a first plurality of correspondences is identified between multiple syntactical elements and multiple distinct patterns of pixels;

> access to a third plurality of syntactical elements that describes a scenario;
>
> identification, utilizing the trained computer-implemented neural network, of a second plurality of correspondences between the third plurality of syntactical elements and sets of pixels that each include a plurality of pixels; and
>
> generation, utilizing the trained computer-implemented neural network, of one or more images that represents the scenario based on the second plurality of correspondences;
>
> wherein the system is configured such that the one or more images is caused to be generated by applying a first plurality of attentions based on the second plurality of correspondences.

'404 Patent, Cl. 274.

116. Claim 274, considered together with Claim 10, addresses the technical problem of how a neural-network-based system can develop and use relationships between language and visual patterns to generate an image corresponding to a textual scenario. Claim 10 provides the underlying successive attention and probability-based neural-network processing architecture. Claim 274 extends that architecture across modalities by using correspondences between syntactical elements and pixel patterns in attention-based image generation.

117. This arrangement enables the computer to understand and apply relationships between language and visual information in a way that text-only processing cannot. The system can translate the semantic content of a previously unencountered scenario description into visual patterns and create new image content representing that scenario. The resulting image is not merely retrieved from stored content or selected because it contains an associated label. It is generated through cross-modal neural-network processing that applies learned text-to-pixel relationships and attention to produce new visual output.

118. Other claims of the Asserted Patents within this technology area address different multimodal mechanisms, including processing a user-provided image together with text, generating

an image based on text and a reference image, generating a sequence of images forming a video, generating an explanation for an image and then generating another image based on that explanation, and generating images used to train another neural network. Those claims recite different inputs, different training operations, and different outputs. Claim 274 of the '404 Patent is not alleged or conceded to represent them merely because each involves visual content.

ANTHROPIC'S INFRINGEMENT

119. Anthropic infringes the Asserted Patents by making, using, offering for sale, selling, and/or importing its Accused Products. Exemplary accused infringing products and services include, but are not limited to, Claude Chat, Claude Cowork, Claude Code, Claude Design, Claude Enterprise, Claude API access, and hosted or dedicated Claude deployments.

120. The internal operation of the accused models—including their architecture, source code, training procedures, configurations, and intermediate processing—is not fully public. Flinn's allegations are based on presently available public information and the accompanying claim charts, and are expected to be confirmed and supplemented through discovery, including Anthropic's source code, technical documentation, design materials, testing records, and testimony.

A. Foundation Model Architecture, Training, and Alignment

121. At the core of Anthropic's business is the Claude family of models, including Claude Fable, Mythos, Opus, Sonnet, and Haiku. Anthropic, *Models Overview*, https://platform.claude.com/docs/en/about-claude/models/overview (last visited Nov. 10, 2026). Anthropic describes the Claude family in terms of advanced reasoning, vision analysis, code generation, multilingual processing, and very large context windows for enterprise workloads. Anthropic, *Introducing the Next Generation of Claude* (Mar. 4, 2024), https://www.anthropic.com/news/claude-3-family.

122. Anthropic's Model Card for Claude 2 confirms that "[b]oth Claude 2 and previous

Claude models are general purpose large language models" that "use a transformer architecture and are trained via unsupervised learning . . . ." Anthropic, *Model Card and Evaluations for Claude Models 2* (2023), https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf. Claude 2 was released in July 2023 as a "continuous evolution" of the earlier Claude 1 models, all sharing that same transformer-based architecture. *Id*.

123. In March 2024, Anthropic released the Claude 3 model family (Claude 3 Haiku, Claude 3 Sonnet, and Claude 3 Opus), which Anthropic described as a "new family of large multimodal models" that, "[l]ike its predecessors," employ "various training methods, such as unsupervised learning and Constitutional AI." Anthropic, *The Claude 3 Model Family: Opus, Sonnet, Haiku* 1 (2024), https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. When asked by the Stanford Center for Research on Foundation Models to disclose the architecture of Claude 3, Anthropic directed the researchers to the Claude 2 Model Card's disclosure where it mentioned the "use of [a] transformer architecture." Stanford Ctr. for Rsch. on Found. Models, *Foundation Model Transparency Index: Anthropic Claude 3 Report* (May 2024), https://crfm.stanford.edu/fmti/May-2024/company-reports/Anthropic_Claude%203.html.

124. On information and belief, Anthropic's later Claude models, including Claude 3 and its successors, also employ an attention-based transformer architecture. In May 2024, Anthropic's interpretability team published research discussing their analysis of Claude 3 Sonnet's use of "attention heads," the same attention mechanisms that are the foundational concept underlying the transformer architecture. Adly Templeton et al., *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet*, Anthropic (May 21, 2024), https://transformer-

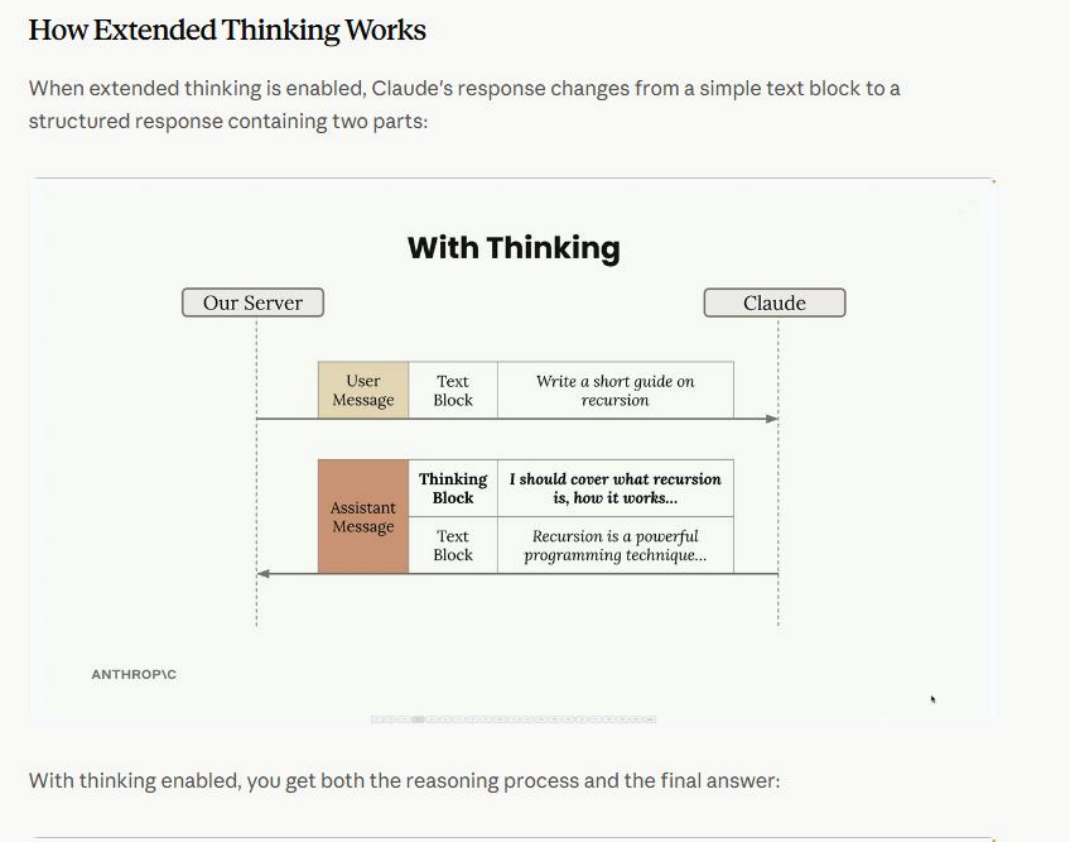
circuits.pub/2024/scaling-monosemanticity/.

125. Anthropic also publicly describes its use of human and model-generated feedback, including through Constitutional AI and reinforcement learning from human feedback, to evaluate and adjust Claude's generated content. Anthropic's Model Card for Claude 2 states that Claude 2 and prior Claude models are general-purpose large language models that use a transformer architecture and are trained using unsupervised learning, reinforcement learning from human feedback, and Constitutional AI. *See* Ex. A at 8.

126. This functionality is commercially significant to Anthropic. How the underlying model processes relationships among input elements and uses intermediate outputs may affect model capability, training and inference efficiency, scalability, latency, throughput, context handling, and the quality of every downstream Claude service. Training and alignment mechanisms may likewise affect factual reliability, safety, instruction following, customer trust, and enterprise deployment. Anthropic markets these characteristics as features of the Accused Products and differentiates its model tiers and pricing on that basis. Anthropic maintains a tiered API pricing structure that charges developers progressively higher per-token rates for Claude models that Anthropic markets as offering greater reasoning, coding, and agentic capabilities. Anthropic's most capable models are priced at ten times the rate charged for the least expensive model in the family. Anthropic, *Pricing*, https://platform.claude.com/docs/en/about-claude/pricing (last visited Nov. 10, 2026); Anthropic, *Models Overview*, https://platform.claude.com/docs/en/about-claude/models/overview (last visited Nov. 10, 2026). That graduated pricing structure reflects Anthropic's assessment that more advanced reasoning, attention-based processing, and model capabilities command a premium, confirming that the accused functionality is a commercially significant differentiator for the Accused Products.

B. Reasoning and Inference Control

127. Anthropic markets identified Claude models and services for multistep reasoning, analysis, coding, mathematics, scientific research, and planning, and provides interfaces, documentation, and APIs through which users and developers invoke those capabilities. Anthropic describes Claude's extended thinking as a user- and developer-controllable reasoning mode in which Claude may "think more deeply about trickier questions," explore different reasoning branches, and double- or triple-check its answers before producing a response. Anthropic, *Claude's Extended Thinking*, https://www.anthropic.com/news/claude-extended-thinking (last visited Nov. 10, 2026).



Anthropic, *Extended Thinking*, https://academy.claude.com/courses/building-with-the-claude-api/extended-thinking (last visited Nov. 10, 2026).

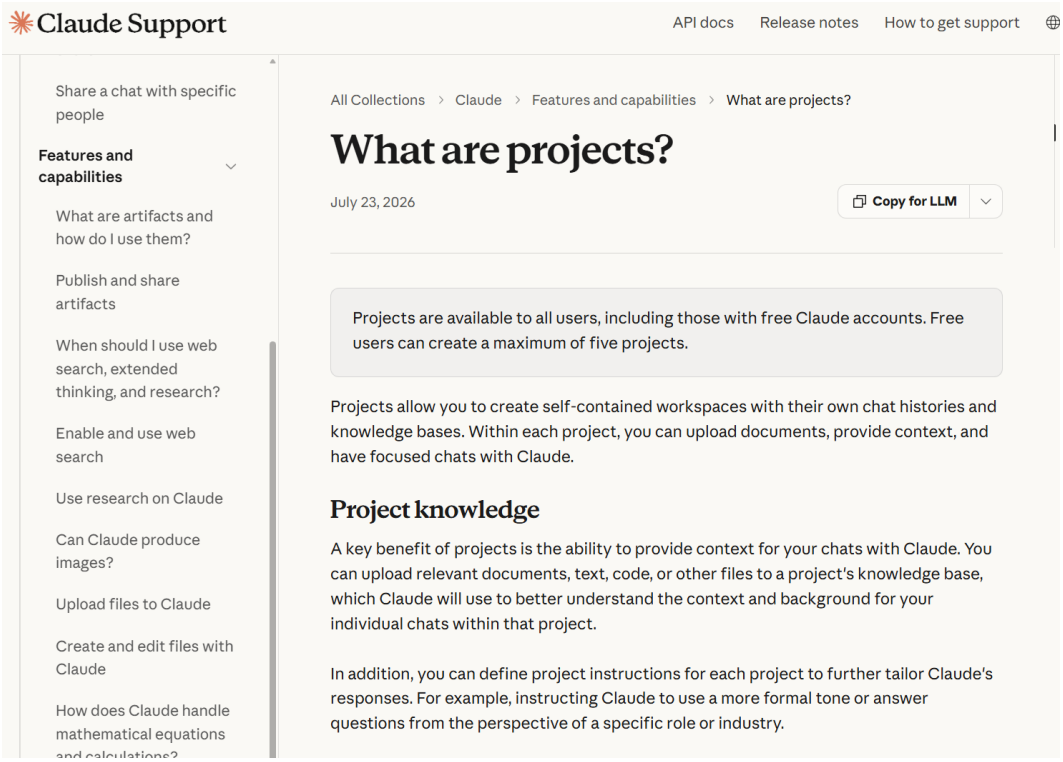
128. On information and belief, the Accused Products configured to provide extended thinking generate, evaluate, and select among intermediate sequences during inference, and use

representations associated with that processing in generating the content on which responses are based.

129. This functionality is commercially significant to Anthropic. Reasoning capability may affect performance on complex tasks, the amount of computation applied at inference time, reliability in evaluating candidate solutions, and the transparency of generated results. Anthropic markets reasoning capability as a basis for selecting among Claude models and service tiers. Anthropic suggests using its higher-priced models, such as Claude Opus, to perform "the most complex reasoning" tasks. Anthropic, *Pricing*, https://platform.claude.com/docs/en/about-claude/pricing (last visited Nov. 10, 2026).

C. Memory and Personalization

130. Anthropic offers and enables functionality through which Claude products retain, retrieve, and reuse information from prior content, interactions, projects, files, profiles, or connected sources. While in operation, Claude Chat receives a prompt or other input from a user and accesses persistent project context, conversation history, uploaded files, project instructions, and other information associated with the user's Project. Claude uses that context to generate and return a responsive output to the user. Anthropic markets Projects as allowing users to store relevant documents and instructions around ongoing work so that Claude can retrieve relevant context across interactions and use project-specific information in later responses. *See* Ex. C at 3, 10.



131. On information and belief, the Accused Products store representations generated during attention-based processing and later retrieve and use those representations in subsequent attention prioritization and generation. Discovery is expected to establish the particular representations, storage systems, retrieval operations, and interactions with later model processing.

132. This functionality is commercially significant to Anthropic. Memory and personalization may affect conversational continuity, user-specific responses, context retention across sessions, efficiency on repeated tasks, and the usefulness of agents operating over extended interactions. Anthropic markets these capabilities as features affecting customer retention, subscription value, and enterprise productivity. Anthropic's memory and personalization functionality is described as enhancing enterprise plans because it allows Claude to retain project and business context, reducing the need for users to repeatedly redefine concepts or provide background information, thereby improving productivity. Anirban Ghoshal, *Anthropic Adds Memory to Claude for Team and Enterprise Plan Users*, Computerworld (Sept. 12, 2025),

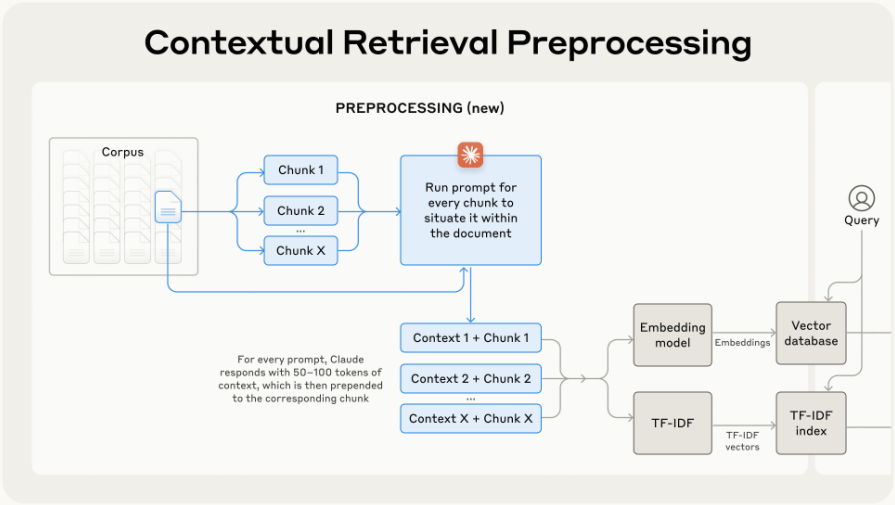
https://www.computerworld.com/article/4056366/anthropic-adds-memory-to-claude-for-team-and-enterprise-plan-users.html. This is further reinforced by Anthropic's pricing page, which ties the number of available projects to higher-priced plans. Anthropic, *Pricing*, https://claude.com/pricing (last visited Nov. 10, 2026).

D. Retrieval, Search, and Grounding

133. Anthropic offers and enables Claude functionality involving web search, connected data sources, document and file access, enterprise information, and retrieval-supported generation, and provides documentation and workflows instructing users and developers how to supply, connect, index, retrieve, and use external content with Claude. Anthropic markets Projects as allowing users to upload documents and other information to a project knowledge base and explains that, when Retrieval-Augmented Generation ("RAG") is enabled, Claude searches and retrieves relevant information from that knowledge to answer users' questions.

134. For example, when Claude Chat operates through RAG-enabled Projects, the Accused Products cause generation of a vector representation based on syntactical elements contained in a user's query for use in searching stored content. Anthropic explains that, when RAG is enabled for a Project, Claude uses a project-knowledge search tool to search uploaded documents and retrieve information relevant to answering the user's question. *See* Ex. C at 14.

Here's what the preprocessing flow looks like in practice:



Contextual Retrieval is a preprocessing technique that improves retrieval accuracy.

If you're interested in using Contextual Retrieval, you can get started with our cookbook.

Using Prompt Caching to reduce the costs of Contextual Retrieval

Contextual Retrieval is uniquely possible at low cost with Claude, thanks to the special prompt caching feature we mentioned above. With prompt caching, you don't need to pass in the reference document for every chunk. You simply load the document into the cache once and then reference the previously cached content. Assuming 800 token chunks, 8k token documents, 50 token context instructions, and 100 tokens of context per chunk, **the one-time cost to generate contextualized chunks is $1.02 per million document tokens.**

135. On information and belief, the Accused Products, including those with RAG-enabled configurations, generate vector representations through multiple-attention-based neural-network application, search and compare against stored vector representations of distinct content items, identify content based on that comparison, and generate new content based on the identified content.

136. This functionality is commercially significant to Anthropic. Retrieval and grounding allow Claude to use proprietary, enterprise, user-provided, or current external information rather than relying exclusively on model parameters, and may affect factual reliability, enterprise adoption,

document analysis, research workflows, source attribution, and data connectivity. Anthropic markets and separately monetizes search, file-access, and connected-data capabilities through subscription product tiers. Anthropic, *Pricing*, https://claude.com/pricing#team-&-enterprise (last visited Nov. 10, 2026). Anthropic includes Projects, Connectors, Enterprise search, file access, and organizational knowledge features in its paid plans and enterprise offerings. *Id*. Anthropic's pricing structure reflects that the retrieval, search, and grounding functionality is a commercially significant component of the Accused Products.

E. Agentic and Developer Workflows

137. Anthropic markets Claude products for generating, evaluating, revising, and executing computer code, and offers functionality through which Claude invokes tools, interacts with files and computer environments, communicates with external systems, and performs actions requested by users. For example, Claude Code causes access to one or more trained Claude models by transmitting a user's coding instruction, together with relevant contextual information that may include selected codebase content, system instructions, tool definitions, tool results, and conversation history, and receiving model-generated outputs in response. Claude Code may repeat this access during its agentic loop as the model evaluates the available context, requests tool use, receives tool results, and generates subsequent outputs. *See* Ex. E at 13.

CORE CONCEPTS

How Claude Code works

Copy page

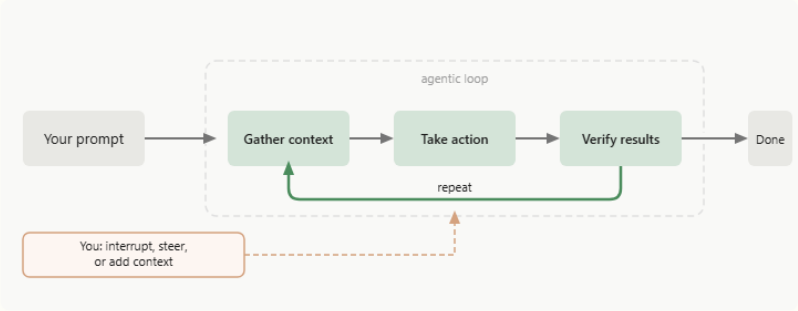
Understand the agentic loop, built-in tools, and how Claude Code interacts with your project.

Claude Code is an agentic assistant that runs in your terminal. While it excels at coding, it can help with anything you can do from the command line: writing docs, running builds, searching files, researching topics, and more.

This guide covers the core architecture, built-in capabilities, and **tips for working effectively**. For step-by-step walkthroughs, see **Common workflows**. For extensibility features like skills, MCP, and hooks, see **Extend Claude Code**.

The agentic loop

When you give Claude a task, it works through three phases: **gather context**, **take action**, and **verify results**. These phases blend together. Claude uses tools throughout, whether searching files to understand your code, editing to make changes, or running tests to check its work.



The loop adapts to what you ask. A question about your codebase might only need context gathering. A bug fix cycles through all three phases repeatedly. A refactor might involve extensive verification. Claude decides what each step requires based on what it learned from the previous step, chaining dozens of actions together and course-correcting along the way.

138. Additionally, Anthropic markets Claude products that use agents to perform a wide variety of workflows. Anthropic identifies several key capabilities of its agents, including understanding complex inputs, engaging in reasoning and planning, reliably using tools, and recovering from errors. For example, Anthropic markets general computer use agents that accomplish open-ended tasks across tools, applications, and environments rather than following a fixed, predetermined path. Erik Schluntz & Barry Zhang, *Building Effective Agents*, Anthropic (Dec. 19, 2024), https://www.anthropic.com/research/building-effective-agents.

139. On information and belief, the Accused Products, including Claude Code and other agent configurations, generate probabilities associated with corresponding subsets of executable

computer code elements based on an attention directed to a representation of a subset of input and select code elements based on those probabilities, and perform requested actions other than content generation based on generated syntactical elements together with a communication reflecting performance of the action.

140. This functionality is commercially significant to Anthropic. Coding capability reduces software development time and assists users in generating, explaining, testing, debugging, revising, and executing code, and agentic capability allows Claude to interact with tools, applications, files, browsers, computer environments, and external services. Anthropic monetizes these capabilities through subscriptions, APIs, enterprise offerings, and developer products, including Claude Code. Anthropic's pricing materials further show that Claude Code is available through paid subscriptions or API access, and Anthropic markets its higher-priced Claude models, Opus and Sonnet, for use with Claude Code and other agentic workflows. Anthropic, *Pricing*, https://claude.com/pricing#api (last visited Nov. 10, 2026).

F. Multimodal Processing and Generation

141. Anthropic offers and enables Claude workflows involving text and visual information, including identified functionality for image understanding, visual analysis, generation or revision of visual content through integrated tools or services, and design-oriented workflows. *See* Ex. B.

142. This functionality is commercially significant to Anthropic because it expands the types of content Claude may process and create and the workflows Claude may support, including design, presentation, documentation, and mixed text-and-image enterprise workflows. Anthropic monetizes multimodal functionality through its paid subscription and enterprise offerings. Anthropic, *Pricing*, https://claude.com/pricing (last visited Nov. 10, 2026). Anthropic's pricing page identifies Claude Design, Slides, and Docs as Pro-plan features, and Anthropic separately announced that Claude Design is available for Claude Pro, Max, Team, and Enterprise subscribers. *Id.*;

Anthropic, *Introducing Claude Design by Anthropic Labs* (Apr. 17, 2026), https://www.anthropic.com/news/claude-design-anthropic-labs.

G. Commercialization and Distribution

143. Anthropic monetizes both end-user access and enterprise or developer adoption of its model platform through consumer subscriptions, enterprise offerings, APIs, hosted and dedicated deployments, developer services, usage-based pricing, and third-party cloud channels. Anthropic has initiated the process for a U.S. initial public offering. On October 26, 2026, Anthropic filed an S-1 with the SEC for a proposed initial public offering. https://www.anthropic.com/news/s1-sec; https://www.sec.gov/Archives/edgar/data/2130357/000162828026072484/anthropicpbc-sx1ax2.htm. With each iteration and expansion of the Claude model family and its associated products, Anthropic's systems offer more advanced attention-based reasoning, retrieval, memory, agentic, and multimodal functionality that infringes the Asserted Patents.

144. On information and belief, Anthropic also provides the Accused Products to third parties for distribution to enterprise customers through fully managed Model-as-a-Service offerings on third-party cloud infrastructure. Through those channels, Anthropic provides software-development kits, API integration documentation, quickstart walkthroughs, integration guides, and sample code repositories instructing enterprise customers and developers how to access and invoke the Accused Products within third-party cloud environments. Anthropic also markets and supports enterprise adoption of the Accused Products through cloud-based marketplace listings, customer materials, webinars, tutorials, training courses, and consumption-based and provisioned-throughput pricing models.

H. Anthropic's Knowledge

145. Anthropic has had knowledge of each Asserted Patent, and of the infringement allegations set forth in this Complaint and the accompanying claim charts, at least as of the filing and

service of this Complaint. The claim charts attached hereto identify, on an element-by-element basis, the manner in which the Accused Products satisfy at least one claim of each Asserted Patent.

146. Notwithstanding that knowledge, Anthropic continues to make, use, test, operate, sell, offer to sell, import, and provide the Accused Products, and to instruct, encourage, and enable its customers, developers, distributors, enterprise users, and platform partners to use the Accused Products in the manner charted.

COUNT 1 - INFRINGEMENT OF U.S. PATENT NO. 12,327,166

147. Flinn incorporates the foregoing paragraphs by reference as if fully set forth herein.

148. This cause of action arises under the Patent Laws of the United States, and, in particular, 35 U.S.C. §§ 271, *et seq*.

149. A true and accurate copy of the '166 Patent is attached hereto as Exhibit 1.

150. Flinn is the assignee of the '166 Patent, with ownership of all substantial rights in the '166 Patent, including the right to exclude others and to enforce, sue, and recover damages for past, present, and future infringement.

151. The '166 Patent is valid, enforceable, and was duly issued in full compliance with Title 35 of the United States Code after a full and fair examination, and enjoys a statutory presumption of validity under 35 U.S.C. § 282.

152. Anthropic's Accused Products meet all limitations of at least Claim 1 of the '166 Patent.

153. Exhibit A sets forth an exemplary claim chart mapping Claim 1 of the '166 Patent to the Accused Products on an element-by-element basis and demonstrating how each limitation of that claim is met. The claim chart contained in Exhibit A is illustrative and not exhaustive; Flinn does not limit its infringement contentions to the particular claims or theories of infringement identified therein.

154. Anthropic directly infringes at least Claim 1 of the '166 Patent by making, using, offering for sale, selling, and/or importing products, including at least the Accused Products, and its products and services containing the same that incorporate the fundamental technologies covered by the '166 Patent and meet every limitation of at least Claim 1, and by performing the claimed methods using the Accused Products, either literally or under the doctrine of equivalents, in violation of 35 U.S.C. § 271(a). *See* Ex. A.

155. Anthropic induces infringement of at least Claim 1 of the '166 Patent under 35 U.S.C. § 271(b). Since at least the filing and service of this Complaint, Anthropic has induced infringement with knowledge of the '166 Patent and of its infringement, and with specific intent to induce infringement or, at minimum, with willful blindness to the infringing nature of its Accused Products. Anthropic has provided third parties with access to its models and granted those third parties permission to operate those models both for the third parties' own purposes and to provide the Accused Products to others. Those third parties directly infringe at least Claim 1 of the '166 Patent by operating the Accused Products. Anthropic encourages that direct infringement by providing access to the models' software, providing permission to operate the Accused Products, and, upon information and belief, entering into contractual arrangements regarding the same.

156. In addition, since at least the filing and service of this Complaint, when Anthropic was placed on notice of its infringement, Anthropic has contributorily infringed, under 35 U.S.C. § 271(c), at least Claim 1 of the '166 Patent. Anthropic has sold and offered to sell a provisioned and configured Claude deployment, as delivered through, for example, hosted deployments and Model-as-a-Service arrangements. That provisioned and configured deployment consists of the trained Claude model together with the configuration, tuning, serving parameters, and enabled functionality that Anthropic supplies with it.

157. The provisioned and configured deployment constitutes a material part of the invention of the '166 Patent as set forth in Exhibit A. Anthropic knows the same to be especially made or especially adapted for use in infringement of the '166 Patent because Anthropic trains, tunes, configures, and enables the deployment to perform the accused functionality.

158. The provisioned and configured Claude deployment has no substantial non-infringing use. The Claude deployment is not a staple article or commodity of commerce. It is a purpose-built deployment that Anthropic configures to perform the charted operations whenever the deployment is used for its intended purpose. Any asserted alternative use either does not avoid the charted operations or is not a substantial use of the component as supplied and configured by Anthropic. Anthropic knowingly contributes to the direct infringement of at least Claim 1 of the '166 Patent by distributors, developers, enterprise customers, and end users that incorporate and use the Accused Products in their own applications and services, including through hosted deployments and Model-as-a-Service arrangements with third-party platforms.

159. Anthropic has infringed and continues to infringe, directly and/or indirectly (by inducing infringement and/or contributing to infringement), at least Claim 1 of the '166 Patent in this District and elsewhere in the United States.

160. Anthropic's infringement of the '166 Patent is willful. Anthropic has had knowledge of the '166 Patent and of the infringement allegations and element-by-element claim chart directed to it since at least the filing and service of this Complaint as set forth above. With that knowledge, Anthropic continues and will continue the accused conduct as alleged above. Flinn is entitled to enhanced damages under 35 U.S.C. § 284 and to its attorneys' fees, expenses, and costs under 35 U.S.C. § 285.

COUNT 2 - INFRINGEMENT OF U.S. PATENT NO. 12,223,404

161. Flinn incorporates the foregoing paragraphs by reference as if fully set forth herein.

162. This cause of action arises under the Patent Laws of the United States, and, in particular, 35 U.S.C. §§ 271, *et seq*.

163. A true and accurate copy of the '404 Patent is attached hereto as Exhibit 2.

164. Flinn is the assignee of the '404 Patent, with ownership of all substantial rights in the '404 Patent, including the right to exclude others and to enforce, sue, and recover damages for past, present, and future infringement.

165. The '404 Patent is valid, enforceable, and was duly issued in full compliance with Title 35 of the United States Code after a full and fair examination, and enjoys a statutory presumption of validity under 35 U.S.C. § 282.

166. Anthropic's Accused Products meet all limitations of at least Claims 10 and 21 of the '404 Patent.

167. Exhibit B sets forth an exemplary claim chart mapping Claims 10 and 21 of the '404 Patent to the Accused Products on an element-by-element basis and demonstrating how each limitation of those claims is met. The claim chart contained in Exhibit B is illustrative and not exhaustive; Flinn does not limit its infringement contentions to the particular claims or theories of infringement identified therein.

168. Anthropic directly infringes at least Claims 10 and 21 of the '404 Patent by making, using, offering for sale, selling, and/or importing products, including at least the Accused Products, and its products and services containing the same, that incorporate the fundamental technologies covered by the '404 Patent, that meet every limitation of at least Claims 10 and 21, either literally or under the doctrine of equivalents, in violation of 35 U.S.C. § 271(a). *See* Ex. B.

169. Anthropic induces infringement of at least Claims 10 and 21 of the '404 Patent under 35 U.S.C. § 271(b). Since at least the filing and service of this Complaint, Anthropic has induced

infringement with knowledge of the '404 Patent and of its infringement, and with specific intent to induce infringement or, at minimum, with willful blindness to the infringing nature of its Accused Products. Anthropic has provided third parties with access to its models and granted those third parties permission to operate those models both for the third parties' own purposes and to provide the Accused Products to others. Those third parties directly infringe at least Claims 10 and 21 of the '404 Patent by operating the Accused Products. Anthropic encourages that direct infringement by providing access to the models' software, providing permission to operate the Accused Products, and, upon information and belief, entering into contractual arrangements regarding the same.

170. In addition, since at least the filing and service of this Complaint, when Anthropic was placed on notice of its infringement, Anthropic has contributorily infringed, under 35 U.S.C. § 271(c), at least Claims 10 and 21 of the '404 Patent. Anthropic has sold and offered to sell a provisioned and configured Claude deployment, as delivered through, for example, hosted deployments and Model-as-a-Service arrangements. That provisioned and configured deployment consists of the trained Claude model together with the configuration, tuning, serving parameters, and enabled functionality that Anthropic supplies with it.

171. The provisioned and configured deployment constitutes a material part of the invention of the '404 Patent as set forth in Exhibit B. Anthropic knows the same to be especially made or especially adapted for use in infringement of the '404 Patent because Anthropic trains, tunes, configures, and enables the deployment to perform the accused functionality.

172. The provisioned and configured Claude deployment has no substantial non-infringing use. The Claude deployment is not a staple article or commodity of commerce. It is a purpose-built deployment that Anthropic configures to perform the charted operations whenever the deployment is used for its intended purpose. Any asserted alternative use either does not avoid the

charted operations or is not a substantial use of the component as supplied and configured by Anthropic. Anthropic knowingly contributes to the direct infringement of at least Claims 10 and 21 of the '404 Patent by distributors, developers, enterprise customers, and end users that incorporate and use the Accused Products in their own applications and services, including through hosted deployments and Model-as-a-Service arrangements with third-party platforms.

173. Anthropic has infringed and continues to infringe, directly and/or indirectly (by inducing infringement and/or contributing to infringement), at least Claims 10 and 21 of the '404 Patent in this District and elsewhere in the United States.

174. Anthropic's infringement of the '404 Patent is willful. Anthropic has had knowledge of the '404 Patent and of the infringement allegations and element-by-element claim chart directed to it since at least the filing and service of this Complaint as set forth above. With that knowledge, Anthropic continues and will continue the accused conduct as alleged above. Flinn is entitled to enhanced damages under 35 U.S.C. § 284 and to its attorneys' fees, expenses, and costs under 35 U.S.C. § 285.

COUNT 3 - INFRINGEMENT OF U.S. PATENT NO. 12,511,580

175. Flinn incorporates the foregoing paragraphs by reference as if fully set forth herein.

176. This cause of action arises under the Patent Laws of the United States, and, in particular, 35 U.S.C. §§ 271, *et seq*.

177. A true and accurate copy of the '580 Patent is attached hereto as Exhibit 3.

178. Flinn is the assignee of the '580 Patent, with ownership of all substantial rights in the '580 Patent, including the right to exclude others and to enforce, sue, and recover damages for past, present, and future infringement.

179. The '580 Patent is valid, enforceable, and was duly issued in full compliance with Title 35 of the United States Code after a full and fair examination, and enjoys a statutory presumption

of validity under 35 U.S.C. § 282.

180. Anthropic's Accused Products meet all limitations of at least Claim 1 of the '580 Patent.

181. Exhibit C sets forth an exemplary claim chart mapping Claim 1 of the '580 Patent to the Accused Products on an element-by-element basis and demonstrating how each limitation of that claim is met. The claim chart contained in Exhibit C is illustrative and not exhaustive; Flinn does not limit its infringement contentions to the particular claims or theories of infringement identified therein.

182. Anthropic directly infringes at least Claim 1 of the '580 Patent by making, using, offering for sale, selling, and/or importing products, including at least the Accused Products, and its products and services containing the same that incorporate the fundamental technologies covered by the '580 Patent and meet every limitation of at least Claim 1, and by performing the claimed methods using the Accused Products, either literally or under the doctrine of equivalents, in violation of 35 U.S.C. § 271(a). *See* Ex. C.

183. Anthropic induces infringement of at least Claim 1 of the '580 Patent under 35 U.S.C. § 271(b). Since at least the filing and service of this Complaint, Anthropic has induced infringement with knowledge of the '580 Patent and of its infringement, and with specific intent to induce infringement or, at minimum, with willful blindness to the infringing nature of its Accused Products. Anthropic has provided third parties with access to its models and granted those third parties permission to operate those models both for the third parties' own purposes and to provide the Accused Products to others. Those third parties directly infringe at least Claim 1 of the '580 Patent by operating the Accused Products. Anthropic encourages that direct infringement by providing access to the models' software, providing permission to operate the Accused Products, and, upon

information and belief, entering into contractual arrangements regarding the same.

184. In addition, since at least the filing and service of this Complaint, when Anthropic was placed on notice of its infringement, Anthropic has contributorily infringed, under 35 U.S.C. § 271(c), at least Claim 1 of the '580 Patent. Anthropic has sold and offered to sell a provisioned and configured Claude deployment, as delivered through, for example, hosted deployments and Model-as-a-Service arrangements. That provisioned and configured deployment consists of the trained Claude model together with the configuration, tuning, serving parameters, and enabled functionality that Anthropic supplies with it.

185. The provisioned and configured deployment constitutes a material part of the invention of the '580 Patent as set forth in Exhibit C. Anthropic knows the same to be especially made or especially adapted for use in infringement of the '580 Patent because Anthropic trains, tunes, configures, and enables the deployment to perform the accused functionality.

186. The provisioned and configured Claude deployment has no substantial non-infringing use. The Claude deployment is not a staple article or commodity of commerce. It is a purpose-built deployment that Anthropic configures to perform the charted operations whenever the deployment is used for its intended purpose. Any asserted alternative use either does not avoid the charted operations or is not a substantial use of the component as supplied and configured by Anthropic. Anthropic knowingly contributes to the direct infringement of at least Claim 1 of the '580 Patent by distributors, developers, enterprise customers, and end users that incorporate and use the Accused Products in their own applications and services, including through hosted deployments and Model-as-a-Service arrangements with third-party platforms.

187. Anthropic has infringed and continues to infringe, directly and/or indirectly (by inducing infringement and/or contributing to infringement), at least Claim 1 of the '580 Patent in this

District and elsewhere in the United States.

188. Anthropic's infringement of the '580 Patent is willful. Anthropic has had knowledge of the '580 Patent and of the infringement allegations and element-by-element claim chart directed to it since at least the filing and service of this Complaint as set forth above. With that knowledge, Anthropic continues and will continue the accused conduct as alleged above. Flinn is entitled to enhanced damages under 35 U.S.C. § 284 and to its attorneys' fees, expenses, and costs under 35 U.S.C. § 285.

COUNT 4 - INFRINGEMENT OF U.S. PATENT NO. 12,462,197

189. Flinn incorporates the foregoing paragraphs by reference as if fully set forth herein.

190. This cause of action arises under the Patent Laws of the United States, and, in particular, 35 U.S.C. §§ 271, *et seq*.

191. A true and accurate copy of the '197 Patent is attached hereto as Exhibit 4.

192. Flinn is the assignee of the '197 Patent, with ownership of all substantial rights in the '197 Patent, including the right to exclude others and to enforce, sue, and recover damages for past, present, and future infringement.

193. The '197 Patent is valid, enforceable, and was duly issued in full compliance with Title 35 of the United States Code after a full and fair examination, and enjoys a statutory presumption of validity under 35 U.S.C. § 282.

194. Anthropic's Accused Products meet all limitations of at least Claim 1 of the '197 Patent.

195. Exhibit D sets forth an exemplary claim chart mapping Claim 1 of the '197 Patent to the Accused Products on an element-by-element basis and demonstrating how each limitation of that claim is met. The claim chart contained in Exhibit D is illustrative and not exhaustive; Flinn does not limit its infringement contentions to the particular claims or theories of infringement identified

therein.

196. Anthropic directly infringes at least Claim 1 of the '197 Patent by making, using, offering for sale, selling, and/or importing products, including at least the Accused Products, and its products and services containing the same that incorporate the fundamental technologies covered by the '197 Patent and meet every limitation of at least Claim 1, and by performing the claimed methods using the Accused Products, either literally or under the doctrine of equivalents, in violation of 35 U.S.C. § 271(a). *See* Ex. D.

197. Anthropic induces infringement of at least Claim 1 of the '197 Patent under 35 U.S.C. § 271(b). Since at least the filing and service of this Complaint, Anthropic has induced infringement with knowledge of the '197 Patent and of its infringement, and with specific intent to induce infringement or, at minimum, with willful blindness to the infringing nature of its Accused Products. Anthropic has provided third parties with access to its models and granted those third parties permission to operate those models both for the third parties' own purposes and to provide the Accused Products to others. Those third parties directly infringe at least Claim 1 of the '197 Patent by operating the Accused Products. Anthropic encourages that direct infringement by providing access to the models' software, providing permission to operate the Accused Products, and, upon information and belief, entering into contractual arrangements regarding the same.

198. In addition, since at least the filing and service of this Complaint, when Anthropic was placed on notice of its infringement, Anthropic has contributorily infringed, under 35 U.S.C. § 271(c), at least Claim 1 of the '197 Patent. Anthropic has sold and offered to sell a provisioned and configured Claude deployment, as delivered through, for example, hosted deployments and Model-as-a-Service arrangements. That provisioned and configured deployment consists of the trained Claude model together with the configuration, tuning, serving parameters, and enabled functionality

that Anthropic supplies with it.

199. The provisioned and configured deployment constitutes a material part of the invention of the '197 Patent as set forth in Exhibit D. Anthropic knows the same to be especially made or especially adapted for use in infringement of the '197 Patent because Anthropic trains, tunes, configures, and enables the deployment to perform the accused functionality.

200. The provisioned and configured Claude deployment has no substantial non-infringing use. The Claude deployment is not a staple article or commodity of commerce. It is a purpose-built deployment that Anthropic configures to perform the charted operations whenever the deployment is used for its intended purpose. Any asserted alternative use either does not avoid the charted operations or is not a substantial use of the component as supplied and configured by Anthropic. Anthropic knowingly contributes to the direct infringement of at least Claim 1 of the '197 Patent by distributors, developers, enterprise customers, and end users that incorporate and use the Accused Products in their own applications and services, including through hosted deployments and Model-as-a-Service arrangements with third-party platforms.

201. Anthropic has infringed and continues to infringe, directly and/or indirectly (by inducing infringement and/or contributing to infringement), at least Claim 1 of the '197 Patent in this District and elsewhere in the United States.

202. Anthropic's infringement of the '197 Patent is willful. Anthropic has had knowledge of the '197 Patent and of the infringement allegations and element-by-element claim chart directed to it since at least the filing and service of this Complaint as set forth above. With that knowledge, Anthropic continues and will continue the accused conduct as alleged above. Flinn is entitled to enhanced damages under 35 U.S.C. § 284 and to its attorneys' fees, expenses, and costs under 35 U.S.C. § 285.

COUNT 5 - INFRINGEMENT OF U.S. PATENT NO. 12,462,198

203. Flinn incorporates the foregoing paragraphs by reference as if fully set forth herein.

204. This cause of action arises under the Patent Laws of the United States, and, in particular, 35 U.S.C. §§ 271, *et seq*.

205. A true and accurate copy of the '198 Patent is attached hereto as Exhibit 5.

206. Flinn is the assignee of the '198 Patent, with ownership of all substantial rights in the '198 Patent, including the right to exclude others and to enforce, sue, and recover damages for past, present, and future infringement.

207. The '198 Patent is valid, enforceable, and was duly issued in full compliance with Title 35 of the United States Code after a full and fair examination, and enjoys a statutory presumption of validity under 35 U.S.C. § 282.

208. Anthropic's Accused Products meet all limitations of at least Claim 1 of the '198 Patent.

209. Exhibit E sets forth an exemplary claim chart mapping Claim 1 of the '198 Patent to the Accused Products on an element-by-element basis and demonstrating how each limitation of that claim is met. The claim chart contained in Exhibit E is illustrative and not exhaustive; Flinn does not limit its infringement contentions to the particular claims or theories of infringement identified therein.

210. Anthropic directly infringes at least Claim 1 of the '198 Patent by making, using, offering for sale, selling, and/or importing products, including at least the Accused Products, and its products and services containing the same that incorporate the fundamental technologies covered by the '198 Patent and meet every limitation of at least Claim 1, and by performing the claimed methods using the Accused Products, either literally or under the doctrine of equivalents, in violation of 35 U.S.C. § 271(a). *See* Ex. E.

211. Anthropic induces infringement of at least Claim 1 of the '198 Patent under 35 U.S.C. § 271(b). Since at least the filing and service of this Complaint, Anthropic has induced infringement with knowledge of the '198 Patent and of its infringement, and with specific intent to induce infringement or, at minimum, with willful blindness to the infringing nature of its Accused Products. Anthropic has provided third parties with access to its models and granted those third parties permission to operate those models both for the third parties' own purposes and to provide the Accused Products to others. Those third parties directly infringe at least Claim 1 of the '198 Patent by operating the Accused Products. Anthropic encourages that direct infringement by providing access to the models' software, providing permission to operate the Accused Products, and, upon information and belief, entering into contractual arrangements regarding the same.

212. In addition, since at least the filing and service of this Complaint, when Anthropic was placed on notice of its infringement, Anthropic has contributorily infringed, under 35 U.S.C. § 271(c), at least Claim 1 of the '198 Patent. Anthropic has sold and offered to sell a provisioned and configured Claude deployment, as delivered through, for example, hosted deployments and Model-as-a-Service arrangements. That provisioned and configured deployment consists of the trained Claude model together with the configuration, tuning, serving parameters, and enabled functionality that Anthropic supplies with it.

213. The provisioned and configured deployment constitutes a material part of the invention of the '198 Patent as set forth in Exhibit E. Anthropic knows the same to be especially made or especially adapted for use in infringement of the '198 Patent because Anthropic trains, tunes, configures, and enables the deployment to perform the accused functionality.

214. The provisioned and configured Claude deployment has no substantial non-infringing use. The Claude deployment is not a staple article or commodity of commerce. It is a purpose-

built deployment that Anthropic configures to perform the charted operations whenever the deployment is used for its intended purpose. Any asserted alternative use either does not avoid the charted operations or is not a substantial use of the component as supplied and configured by Anthropic. Anthropic knowingly contributes to the direct infringement of at least Claim 1 of the '198 Patent by distributors, developers, enterprise customers, and end users that incorporate and use the Accused Products in their own applications and services, including through hosted deployments and Model-as-a-Service arrangements with third-party platforms.

215. Anthropic has infringed and continues to infringe, directly and/or indirectly (by inducing infringement and/or contributing to infringement), at least Claim 1 of the '198 Patent in this District and elsewhere in the United States.

216. Anthropic's infringement of the '198 Patent is willful. Anthropic has had knowledge of the '198 Patent and of the infringement allegations and element-by-element claim chart directed to it since at least the filing and service of this Complaint as set forth above. With that knowledge, Anthropic continues and will continue the accused conduct as alleged above. Flinn is entitled to enhanced damages under 35 U.S.C. § 284 and to its attorneys' fees, expenses, and costs under 35 U.S.C. § 285.

JURY DEMAND

Flinn demands a trial by jury on all issues so triable.

PRAYER FOR RELIEF

WHEREFORE, Flinn prays for the following relief:

a. A judgment that Anthropic has infringed and continues to infringe, directly, indirectly, and willfully, one or more claims of each of the Asserted Patents;

b. A judgment awarding Flinn damages adequate to compensate for Anthropic's infringement of the Asserted Patents, in no event less than a reasonable royalty for

Anthropic's use of the claimed inventions, together with interest and costs as fixed by the Court under 35 U.S.C. § 284;

c. A judgment that Anthropic's infringement of the Asserted Patents has been willful, and an award of enhanced damages up to three times the amount found or assessed under 35 U.S.C. § 284;

d. An accounting of all infringing acts, including those not presented at trial, and an award of supplemental damages for all infringement occurring after the date of the damages evidence presented at trial through entry of final judgment, together with interest;

e. An award of an ongoing royalty for Anthropic's continued infringement of the Asserted Patents after entry of final judgment, in an amount to be determined by the Court, and such further relief as is necessary to compensate Flinn for that continued infringement;

f. An award of pre-judgment interest, calculated from the date of first infringement at the highest rate permitted by law, and post-judgment interest under 28 U.S.C. § 1961;

g. An award of Flinn's costs of suit under 28 U.S.C. § 1920 and Rule 54(d) of the Federal Rules of Civil Procedure;

h. A judgment that this is an exceptional case under 35 U.S.C. § 285 and an award to Flinn of its reasonable attorneys' fees, expenses, and costs; and

i. Such other and further relief, at law or in equity, as the Court deems just and proper.

ASHBY & GEDDES

/s/ *Andrew C. Mayo*

Andrew C. Mayo (#5207)
Jeffrey J. Lyons (#6437)
500 Delaware Avenue, 8th Floor
P.O. Box 1150
Wilmington, DE 19899
(302) 654-1888
amayo@ashbygeddes.com
jlyons@ashbygeddes.com

Attorneys for Plaintiff
Flinn Artificial Intelligence (AI) Technologies, LLC

Of Counsel:

Brett C. Govett
NORTON ROSE FULBRIGHT US LLP
2200 Ross Avenue, Suite 3600
Dallas, TX 75201
(214) 855-8000
brett.govett@nortonrosefulbright.com

Daniel S. Leventhal
NORTON ROSE FULBRIGHT US LLP
1550 Lamar Street, Suite 2000
Houston, TX 77010
(713) 651-5151
daniel.leventhal@nortonrosefulbright.com

John P. Poulos
NORTON ROSE FULBRIGHT US LLP
1045 W. Fulton Market, Suite 1200
Chicago, IL 60607
(312) 964-7766
john.poulos@nortonrosefulbright.com

Dated: November 12, 2026