IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF DELAWARE

FRIENDLIAI INC.,	)	
	)	
Plaintiff,	)	
	)	C.A. No.
v.	)	
	)	**JURY TRIAL DEMANDED**
HUGGING FACE, INC.,	)	
	)	
Defendants.	)	
	)	

COMPLAINT FOR PATENT INFRINGEMENT

Plaintiff FriendliAI Inc. ("FriendliAI"), for its Complaint against Defendant Hugging Face, Inc. ("Hugging Face" or "Defendant"), hereby alleges as follows:

NATURE OF THE ACTION

1. This is a civil action arising under the Patent Laws of the United States, 35 U.S.C. §§ 271 *et seq.*, for infringement of United States Patent No. 11,442,775 (the "'775 patent" or the "Patent-in-Suit") relating to artificial intelligence technology, specifically for machine-learning transformer neural network models.

2. Artificial intelligence, or AI, is a field of computer science that focuses on creating intelligent machines capable of performing tasks that typically require human intelligence. Using algorithms, data, and computational power, AI can understand, process, and generate human language; analyze and interpret data; and make decisions or take actions to achieve specific goals. AI has a wide range of applications across various industries, including healthcare, finance, transportation, manufacturing, cybersecurity, gaming, and entertainment. The global AI market size is projected to grow at a Compound Annual Growth Rate (CAGR) of 36.8%, reaching $1,345.2 billion by 2030 from $150.2 billion in 2023. *See*

https://www.marketsandmarkets.com/PressReleases/artificial-intelligence.asp.

3. FriendliAI is a pioneering AI company that focuses on large-scale generative AI technology. Following substantial research, FriendliAI developed a serving engine for large-scale generative AI models that optimizes efficiency, throughput and latency. FriendliAI's novel serving engine can be used for a variety types of generative AI tasks, including data generation, translation, sentiment analysis, text summarization, auto-correction and the like. Such efforts by FriendliAI, and its founder and CEO Dr. Byung-gon Chun, have resulted in the issuance of multiple patents, including the Patent-in-Suit.

THE PARTIES

4. Plaintiff FriendliAI is a company organized and existing under the laws of the Republic of Korea, with its principal place of business at 5F, AMC Tower, 222 Bongeunsa-ro, Gangnam-gu, Seoul, 06135, Korea.

5. Defendant Hugging Face, Inc. is a company organized and existing under the laws of the State of Delaware, with its principal place of business at 20 Jay St, Ste 620, Brooklyn, New York, 11201.

JURISDICTION AND VENUE

6. This action arises under the patent laws of the United States, including 35 U.S.C. §§ 271 *et seq*. The jurisdiction of this Court over the subject matter of this action is proper under 28 U.S.C. §§ 1331 and 1338(a).

7. Venue is proper in this District pursuant to 28 U.S.C. §§ 1391(b), (c), and 1400(b). Defendant is an entity organized under the laws of Delaware and resides in Delaware for purposes of venue under 28 U.S.C. § 1400(b). Defendant conducts business in Delaware, at least by offering for sale, selling, and otherwise making available products and services through its website,

which are accessible in Delaware. Defendant has also committed and continues to commit acts of infringement in this District.

8. This Court has personal jurisdiction over Defendant because Defendant conducts business in Delaware by at least offering for sale, selling, and otherwise making available products and services through its website, which are accessible in Delaware, and because infringement has occurred and continues to occur in Delaware.

9. Personal jurisdiction also exists over Defendant because it is an entity incorporated in and organized under the law of Delaware.

BACKGROUND OF THE PATENTED TECHNOLOGY

10. The founder and CEO of FriendliAI is Dr. Byung-gon Chun, a computer scientist and Professor known for his contributions to the field of computer systems, distributed computing, and artificial intelligence. Dr. Chun received his Ph.D. in Computer Science from the University of California, Berkeley. He is currently a Professor in the Computer Science and Engineering (CSE) Department at Seoul National University (SNU), where he leads the Software Platform Lab (SPL), conducts research on machine learning systems, and teaches courses, including courses on Artificial Intelligence and Big Data Systems. Dr. Chun has published numerous papers in reputable conferences and journals, and has received a number of awards, including the EuroSys 2021 Test of Time Award, the 2020 ACM SIGOPS Hall of Fame Award, the 2020 Google Research Award, the 2019 SNU Education Award, and the 2018 Amazon Machine Learning Research Award. Dr. Chun's work has contributed to advancements in the design and optimization of the performance, scalability, and reliability of large-scale distributed systems, including serving and training transformer-based generative AI models.

11. The vision of FriendliAI is to enable innovation by lowering barriers to serving generative AI. In furtherance of that vision, FriendliAI developed PeriFlow (a version of a

distributed serving system called Orca), a patented solution that efficiently serves large-scale AI transformer models. PeriFlow/Orca uses a novel optimization technique referred to as batching with iteration-level scheduling, also known as dynamic batching or continuous batching, which provides for improved throughput and decreased latency as compared to prior art systems.

12. Before FriendliAI's inventions, transformer models had begun to be widely used for AI applications. Pre-existing transformer models, however, suffered from latency and throughput issues, especially when used for large-scale applications. Dr. Chun and his colleagues at FriendliAI and Seoul National University recognized that such issues resulted from the use of a blunt scheduling mechanism, which was primarily designed to schedule executions at request granularity. In existing models, a form of "batching," referred to as naive batching or static batching, could be used to process multiple requests at once, in order to increase overall throughput. But because requests can vary in complexity and length, particularly in generative AI applications, and because the transformer generation model only returns the execution results to the serving system when it finishes processing all requests in the batch, a request that "finishes" early cannot be sent to the client immediately, but rather must wait until the last request in the batch is "finished," imposing a substantial amount of extra latency (or in plain language, causing delays in processing the requests in the batch). Similarly, when a new request arrives in the middle of the current batch's execution, the aforementioned scheduling mechanism makes the newly-arrived request wait until all requests in the current batch have finished. The inflexibility of such scheduling mechanisms resulted in high latency and low overall throughput.

13. Dr. Chun and his colleagues recognized that a method of scheduling the system on a finer granularity could resolve the latency and throughput issues associated with existing systems.

14. After substantial research, Dr. Chun developed novel technology referred to as batching with iteration-level scheduling (also called dynamic batching or continuous batching). Iteration-level scheduling allows for a finished request to be sent to a client, and for new requests to be sent to the execution engine, before all requests in a batch are completed. Iteration-level scheduling provides for a highly efficient and scalable serving of generative AI transformer models, with optimized throughput and latency.

15. Batching with iteration-level scheduling is described in a paper co-authored by Dr. Chun and others at FriendliAI, and presented in July 2022, during the Proceedings of the 16th USENIX Symposium on Operating Systems Design and Implementation. *See* Ex. 3 (Yu, G.-I., Jeong, J.S., Kim, G.-W., Kim,S., and Chun, B.-G. Orca: A distributed serving system for {Transformer-Based} generative models. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pp. 521–538, 2022).

16. Dr. Chun's paper is recognized by others in the industry as the "first" to disclose iteration-level scheduling (also known as dynamic batching or continuous batching). *See* Ex. 4 (https://www.anyscale.com/blog/continuous-batching-llm-inference).

17. The industry has also recognized the benefit of FriendliAI's novel technology—an increase in throughput and decrease in latency—with one article detailing how "continuous batching" (which is also known in the industry as dynamic batching, or batching with iteration-level scheduling) "improves throughput by wasting fewer opportunities to schedule new requests, and improves latency by being capable of immediately injecting new requests into the compute stream." *See* Ex. 4 (https://www.anyscale.com/blog/continuous-batching-llm-inference).

18. The Patent-in-Suit resulted from Dr. Chun's research to develop an innovative serving engine for generative AI transformer models.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.