PyPI recent updates for vllm https://pypi.org/project/vllm/ Recent updates to the Python Package Index for vllm en 0.7.2 https://pypi.org/project/vllm/0.7.2/ A high-throughput and memory-efficient inference and serving engine for LLMs Thu, 06 Feb 2025 18:26:16 GMT 0.7.1 https://pypi.org/project/vllm/0.7.1/ A high-throughput and memory-efficient inference and serving engine for LLMs Sat, 01 Feb 2025 23:29:27 GMT 0.7.0 https://pypi.org/project/vllm/0.7.0/ A high-throughput and memory-efficient inference and serving engine for LLMs Mon, 27 Jan 2025 17:00:35 GMT 0.6.6.post1 https://pypi.org/project/vllm/0.6.6.post1/ A high-throughput and memory-efficient inference and serving engine for LLMs Fri, 27 Dec 2024 07:09:03 GMT 0.6.6 https://pypi.org/project/vllm/0.6.6/ A high-throughput and memory-efficient inference and serving engine for LLMs Fri, 27 Dec 2024 00:47:10 GMT 0.6.5 https://pypi.org/project/vllm/0.6.5/ A high-throughput and memory-efficient inference and serving engine for LLMs Wed, 18 Dec 2024 00:02:45 GMT 0.6.4.post1 https://pypi.org/project/vllm/0.6.4.post1/ A high-throughput and memory-efficient inference and serving engine for LLMs Fri, 15 Nov 2024 18:43:28 GMT 0.6.4 https://pypi.org/project/vllm/0.6.4/ A high-throughput and memory-efficient inference and serving engine for LLMs Fri, 15 Nov 2024 07:49:05 GMT 0.6.3.post1 https://pypi.org/project/vllm/0.6.3.post1/ A high-throughput and memory-efficient inference and serving engine for LLMs Thu, 17 Oct 2024 18:04:37 GMT 0.6.3 https://pypi.org/project/vllm/0.6.3/ A high-throughput and memory-efficient inference and serving engine for LLMs Mon, 14 Oct 2024 21:27:11 GMT 0.6.2 https://pypi.org/project/vllm/0.6.2/ A high-throughput and memory-efficient inference and serving engine for LLMs Wed, 25 Sep 2024 22:36:49 GMT 0.6.1.post2 https://pypi.org/project/vllm/0.6.1.post2/ A high-throughput and memory-efficient inference and serving engine for LLMs Fri, 13 Sep 2024 18:56:23 GMT

0.6.1.post1 https://pypi.org/project/vllm/0.6.1.post1/ A high-throughput and memory-efficient inference and serving engine for LLMs Fri, 13 Sep 2024 05:01:49 GMT 0.6.1 https://pypi.org/project/vllm/0.6.1/ A high-throughput and memory-efficient inference and serving engine for LLMs Wed, 11 Sep 2024 22:18:35 GMT 0.6.0 https://pypi.org/project/vllm/0.6.0/ A high-throughput and memory-efficient inference and serving engine for LLMs Thu, 05 Sep 2024 04:26:12 GMT 0.5.5 https://pypi.org/project/vllm/0.5.5/ A high-throughput and memory-efficient inference and serving engine for LLMs Fri, 23 Aug 2024 19:28:39 GMT 0.5.4 https://pypi.org/project/vllm/0.5.4/ A high-throughput and memory-efficient inference and serving engine for LLMs Mon, 05 Aug 2024 23:46:53 GMT 0.5.3.post1 https://pypi.org/project/vllm/0.5.3.post1/ A high-throughput and memory-efficient inference and serving engine for LLMs Tue, 23 Jul 2024 17:26:49 GMT 0.5.3 https://pypi.org/project/vllm/0.5.3/ A high-throughput and memory-efficient inference and serving engine for LLMs Tue, 23 Jul 2024 08:55:15 GMT 0.5.2 https://pypi.org/project/vllm/0.5.2/ A high-throughput and memory-efficient inference and serving engine for LLMs Mon, 15 Jul 2024 19:22:43 GMT 0.5.1 https://pypi.org/project/vllm/0.5.1/ A high-throughput and memory-efficient inference and serving engine for LLMs Sat, 06 Jul 2024 02:32:58 GMT 0.5.0.post1 https://pypi.org/project/vllm/0.5.0.post1/ A high-throughput and memory-efficient inference and serving engine for LLMs Fri, 14 Jun 2024 06:10:04 GMT 0.5.0 https://pypi.org/project/vllm/0.5.0/ A high-throughput and memory-efficient inference and serving engine for LLMs Tue, 11 Jun 2024 23:05:48 GMT 0.4.3 https://pypi.org/project/vllm/0.4.3/ A high-throughput and memory-efficient inference and serving engine for LLMs Sat, 01 Jun 2024 04:29:44 GMT 0.4.2 https://pypi.org/project/vllm/0.4.2/ A high-throughput and memory-efficient inference and serving engine for LLMs Sun, 05 May 2024 07:14:31 GMT 0.4.1 https://pypi.org/project/vllm/0.4.1/ A

high-throughput and memory-efficient inference and serving engine for LLMs

Wed, 24 Apr 2024 04:37:09 GMT 0.4.0.post1

https://pypi.org/project/vllm/0.4.0.post1/ A high-throughput and memory-efficient inference and serving engine for LLMs Wed, 03 Apr 2024 17:18:29 GMT

0.4.0 https://pypi.org/project/vllm/0.4.0/ A high-throughput and memory-efficient inference and serving engine for LLMs Sun, 31 Mar 2024 03:46:25 GMT

0.3.3 https://pypi.org/project/vllm/0.3.3/ A high-throughput and memory-efficient inference and serving engine for LLMs Fri, 01 Mar 2024 22:41:10 GMT

0.3.2 https://pypi.org/project/vllm/0.3.2/ A high-throughput and memory-efficient inference and serving engine for LLMs Wed, 21 Feb 2024 21:13:34 GMT

0.3.1 https://pypi.org/project/vllm/0.3.1/ A high-throughput and memory-efficient inference and serving engine for LLMs Sat, 17 Feb 2024 01:15:05 GMT

0.3.0 https://pypi.org/project/vllm/0.3.0/ A high-throughput and memory-efficient inference and serving engine for LLMs Wed, 31 Jan 2024 10:02:47 GMT

0.2.7 https://pypi.org/project/vllm/0.2.7/ A high-throughput and memory-efficient inference and serving engine for LLMs Thu, 04 Jan 2024 02:00:15 GMT

0.2.6 https://pypi.org/project/vllm/0.2.6/ A high-throughput and memory-efficient inference and serving engine for LLMs Sun, 17 Dec 2023 19:06:30 GMT

0.2.5 https://pypi.org/project/vllm/0.2.5/ A high-throughput and memory-efficient inference and serving engine for LLMs Thu, 14 Dec 2023 08:20:14 GMT

0.2.4 https://pypi.org/project/vllm/0.2.4/ A high-throughput and memory-efficient inference and serving engine for LLMs Mon, 11 Dec 2023 20:12:38 GMT

0.2.3 https://pypi.org/project/vllm/0.2.3/ A high-throughput and memory-efficient inference and serving engine for LLMs Sun, 03 Dec 2023 21:04:46 GMT

0.2.2 https://pypi.org/project/vllm/0.2.2/ A high-throughput and memory-efficient inference and serving engine for LLMs Sun, 19 Nov 2023 06:26:10 GMT

0.2.1.post1 https://pypi.org/project/vllm/0.2.1.post1/ A high-throughput and

memory-efficient inference and serving engine for LLMs Tue, 17 Oct 2023 17:18:45 GMT 0.2.1 https://pypi.org/project/vllm/0.2.1/ A high-throughput and memory-efficient inference and serving engine for LLMs Mon, 16 Oct 2023 20:39:33 GMT