Usage Stats Collection

vLLM collects anonymous usage data by default to help the engineering team better understand which hardware and model configurations are widely used. This data allows them to prioritize their efforts on the most common workloads. The collected data is transparent, does not contain any sensitive information, and will be publicly released for the community's benefit.

What data is collected?

The list of data collected by the latest version of vLLM can be found here: <gh-file:vllm/usage/usage_lib.py>

Here is an example as of v0.4.0:

```json
{
  "uuid": "fbe880e9-084d-4cab-a395-8984c50f1109",
  "provider": "GCP",
  "num_cpu": 24,
  "cpu_type": "Intel(R) Xeon(R) CPU @ 2.20GHz",
  "cpu_family_model_stepping": "6,85,7",
  "total_memory": 101261135872,
  "architecture": "x86_64",
  "platform": "Linux-5.10.0-28-cloud-amd64-x86_64-with-glibc2.31",
  "gpu_count": 2,
  "gpu_type": "NVIDIA L4",
  "gpu_memory_per_device": 23580639232,
```

```
  "model_architecture": "OPTForCausalLM",
  "vllm_version": "0.3.2+cu123",
  "context": "LLM_CLASS",
  "log_time": 1711663373492490000,
  "source": "production",
  "dtype": "torch.float16",
  "tensor_parallel_size": 1,
  "block_size": 16,
  "gpu_memory_utilization": 0.9,
  "quantization": null,
  "kv_cache_dtype": "auto",
  "enable_lora": false,
  "enable_prefix_caching": false,
  "enforce_eager": false,
  "disable_custom_all_reduce": true
}
```


You can preview the collected data by running the following command:

```bash
tail ~/.config/vllm/usage_stats.json
```

Opting out

You can opt-out of usage stats collection by setting the `VLLM_NO_USAGE_STATS` or

`DO_NOT_TRACK` environment variable, or by creating a `~/.config/vllm/do_not_track` file:

```bash
# Any of the following methods can disable usage stats collection
export VLLM_NO_USAGE_STATS=1
export DO_NOT_TRACK=1
mkdir -p ~/.config/vllm && touch ~/.config/vllm/do_not_track
```