#

Total Closed Issues: 9

##Issue: Sales

Description:

Hello, I currently work with a multibillion dollar fund doing Concierge Banking. I would like to connect with the Llama Stack specific to Enterprise Sales > Financial Services Team > Llama API Sales to financial institutions > Spanish language Americas including the USA. (I am not a developer). Where can I find my Llama Stack community?

Solution: https://github.com/meta-llama/llama-stack/issues/48

##Issue: Colons in filenames are incompatible with windows

Description:

Describe the bug

The model ID for several of the models include colons making them incompatible with windows systems.

EX:

Meta-Llama3.1-405B-Instruct:bf16-mp8

OSError: [WinError 123] The filename, directory name, or volume label syntax is incorrect:

Reproduce

Point llama download to Meta-Llama3.1-405B-Instruct:bf16-mp8

Execute

Expected behavior

File download begins

Solution: https://github.com/meta-llama/llama-stack/issues/31

##Issue: llama model template --name tool-message-success fails

Description:

seems this template requires some updates.

```
(llama-stack) bash-5.1$ llama model template --name tool-message-success

Traceback (most recent call last):

  File "/home/hamidnazeri/.conda/envs/llama-stack/bin/llama", line 8, in <module>

    sys.exit(main())

  File "/home/hamidnazeri/llama-stack/llama_toolchain/cli/llama.py", line 54, in main
```

```
    parser.run(args)
  File "/home/hamidnazeri/llama-stack/llama_toolchain/cli/llama.py", line 48, in run
    args.func(args)
  File "/home/hamidnazeri/llama-stack/llama_toolchain/cli/model/template.py", line 58, in _run_model_template_cmd
    template, tokens_info = render_jinja_template(args.name)
  File "/home/hamidnazeri/.conda/envs/llama-stack/lib/python3.10/site-packages/llama_models/llama3_1/api/interface.py", line 210, in render_jinja_template
    raise ValueError(f"No template found for `{name}`")
ValueError: No template found for `tool-message-success`
```

Solution: https://github.com/meta-llama/llama-stack/issues/28

##Issue: REAME Updates

Description:

The installation from src instructions needs an update,

```
git clone https://github.com/meta-llama/llama-stack.git
```

```
cd llama-stack

pip install -e .
```

Solution: https://github.com/meta-llama/llama-stack/issues/26

##Issue: Fail to start basic inference server example

Description:

I tried to execute the procedure specified in docs/cli_reference.md but I got an error, I'm gonna paste the console content from the last part of the output of the command llama distribution install --spec local --name local_llama_8b_instruct

```
Successfully setup distribution environment. Configuring...
Configuring API surface: inference
Enter value for model (default: Meta-Llama3.1-8B-Instruct):
Enter value for quantization (optional):
Enter value for torch_seed (optional):
Enter value for max_seq_len (required): 4096
Enter value for max_batch_size (default: 1):


Configuring API surface: safety

Do you want to configure llama_guard_shield? (y/n): n

Do you want to configure prompt_guard_shield? (y/n): n


Configuring API surface: agentic_system


YAML configuration has been written to
/home/ubuntu/.llama/distributions/local_llama_8b_instruct/config.yaml

Distribution `local_llama_8b_instruct` (with spec local) has been installed successfully!

ubuntu@ip-hidden:~$ llama distribution start --name local_llama_8b_instruct --port 5000

> initializing model parallel with size 1

> initializing ddp with size 1

> initializing pipeline with size 1

/opt/conda/envs/local_llama_8b_instruct/lib/python3.10/site-packages/torch/__init__.py:955:

UserWarning: torch.set_default_tensor_type() is deprecated as of PyTorch 2.1, please use
torch.set_default_dtype() and torch.set_default_device() as alternatives. (Triggered internally at
../torch/csrc/tensor/python_tensor.cpp:432.)

  _C._set_default_tensor_type(t)

E0812 17:24:42.239000 139992924231488 torch/distributed/elastic/multiprocessing/api.py:702]
failed (exitcode: -9) local_rank: 0 (pid: 1822) of fn: worker_process_entrypoint (start_method: fork)

E0812 17:24:42.239000 139992924231488 torch/distributed/elastic/multiprocessing/api.py:702]
Traceback (most recent call last):

E0812 17:24:42.239000 139992924231488 torch/distributed/elastic/multiprocessing/api.py:702]

File
```

```
"/opt/conda/envs/local_llama_8b_instruct/lib/python3.10/site-packages/torch/distributed/elastic/multi
processing/api.py", line 659, in _poll
E0812 17:24:42.239000 139992924231488 torch/distributed/elastic/multiprocessing/api.py:702]
self._pc.join(-1)
E0812 17:24:42.239000 139992924231488 torch/distributed/elastic/multiprocessing/api.py:702]
File
"/opt/conda/envs/local_llama_8b_instruct/lib/python3.10/site-packages/torch/multiprocessing/spawn.
py", line 170, in join
E0812 17:24:42.239000 139992924231488 torch/distributed/elastic/multiprocessing/api.py:702]
raise ProcessExitedException(
E0812 17:24:42.239000 139992924231488 torch/distributed/elastic/multiprocessing/api.py:702]
torch.multiprocessing.spawn.ProcessExitedException: process 0 terminated with signal SIGKILL
Process ForkProcess-1:
Traceback (most recent call last):
  File "/opt/conda/envs/local_llama_8b_instruct/lib/python3.10/multiprocessing/process.py", line 314,
in _bootstrap
    self.run()
  File "/opt/conda/envs/local_llama_8b_instruct/lib/python3.10/multiprocessing/process.py", line 108,
in run
    self._target(*self._args, **self._kwargs)
                                                                                                  File
"/opt/conda/envs/local_llama_8b_instruct/lib/python3.10/site-packages/llama_toolchain/inference/me
ta_reference/parallel_utils.py", line 175, in launch_dist_group
    elastic_launch(launch_config, entrypoint=worker_process_entrypoint)(
                                                                                                  File
"/opt/conda/envs/local_llama_8b_instruct/lib/python3.10/site-packages/torch/distributed/launcher/api
```

```
.py", line 133, in __call__
    return launch_agent(self._config, self._entrypoint, list(args))
                                                                                    File
"/opt/conda/envs/local_llama_8b_instruct/lib/python3.10/site-packages/torch/distributed/launcher/api
.py", line 264, in launch_agent
    raise ChildFailedError(
torch.distributed.elastic.multiprocessing.errors.ChildFailedError:
============================================================
worker_process_entrypoint FAILED
------------------------------------------------------------
Failures:
  <NO_OTHER_FAILURES>
------------------------------------------------------------
Root Cause (first observed failure):
[0]:
  time      : 2024-08-12_17:24:42
  host      : ip-hidden
  rank      : 0 (local_rank: 0)
  exitcode  : -9 (pid: 1822)
  error_file: <N/A>
  traceback : Signal 9 (SIGKILL) received by PID 1822
============================================================

Ctrl-C detected. Aborting...
```

This is the output of nvidia-smi:

```
Mon Aug 12 17:26:34 2024

+---------------------------------------------------------------------------------------+

| NVIDIA-SMI 535.183.01             Driver Version: 535.183.01   CUDA Version: 12.2     |

|-----------------------------------------+----------------------+----------------------+

| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |

| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |

|                                         |                      |               MIG M. |

|=========================================+======================+=====================

=|

|   0  NVIDIA A10G                    On  | 00000000:00:1E.0 Off |                    0 |

|  0%   29C    P8               9W / 300W |      0MiB / 23028MiB |      0%      Default |

|                                         |                      |                  N/A |

+-----------------------------------------+----------------------+----------------------+


+---------------------------------------------------------------------------------------+

| Processes:                                                                            |

|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |

|        ID   ID                                                             Usage      |

|=======================================================================================

=|

|  No running processes found                                                           |

+---------------------------------------------------------------------------------------+
```

Is there any other software/hardware requirement not specified in the doc that can cause this?

Thanks!

Solution: https://github.com/meta-llama/llama-stack/issues/25

##Issue: [CLI] Add download support back for older models

Description:

The CLI Reference shows how it was possible to download and run older models such as Llama 2 with the CLI, but with a recent commit, it isn't possible to download any model older than the Llama 3.1 family. Can we get support re-added for downloading older models like Llama 2, Code Llama, etc?

I've opened meta-llama/llama-models#83 as well because it seems to be where models are being sourced from now.

Solution: https://github.com/meta-llama/llama-stack/issues/21

##Issue: small bug in text_completion function

Description:

Hi, prompt_tokens = self.tokenizer.encode(x, bos=True, eos=False) should take prompt as input , not x, in llama_toolchain/inference/generation.py. Encountered it while using directly the Llama class instead of recommended client server route.

```
def text_completion(
    self,
    prompt: str,
    temperature: float = 0.6,
    top_p: float = 0.9,
    max_gen_len: Optional[int] = None,
    logprobs: bool = False,
    echo: bool = False,
) -> Generator:
    if (
        max_gen_len is None
        or max_gen_len == 0
        or max_gen_len >= self.model.params.max_seq_len
    ):
        max_gen_len = self.model.params.max_seq_len - 1

    **prompt_tokens = self.tokenizer.encode(x, bos=True, eos=False)**

    yield from self.generate(
        model_input=ModelInput(tokens=prompt_tokens),
        max_gen_len=max_gen_len,
        temperature=temperature,
        top_p=top_p,
        logprobs=logprobs,
        echo=echo,
    )
```

Solution: https://github.com/meta-llama/llama-stack/issues/16

##Issue: Meta API+OpenRouter API has language encoding issues. #bugreport

Description:

I dont know if the problem exist only for OpenRouter end or Meta end.
Just incase I am reporting this problem to both parties.
https://sinanisler.com/wp-content/uploads/2024/07/2024-07-24-17-33-17.mp4
this problem doesn't happen on site only it happens on OpenRouter API too sadly we first discovered on API and started investigating and foundout problem happening on OpenRouter llama model.
Fix it please. ?
Thank you.

Solution: https://github.com/meta-llama/llama-stack/issues/10

##Issue: RFC-0001 - Llama Stack

Description:

As part of the Llama 3.1 release, Meta is releasing an RFC for ?Llama Stack?, a comprehensive set of interfaces / API for ML developers building on top of Llama foundation models. We are looking for feedback on where the API can be improved, any corner cases we may have missed and your general thoughts on how useful this will be.

Ultimately, our hope is to create a standard for working with Llama models in order to simplify the developer experience and foster innovation across the Llama ecosystem.

Solution: https://github.com/meta-llama/llama-stack/issues/6
