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?

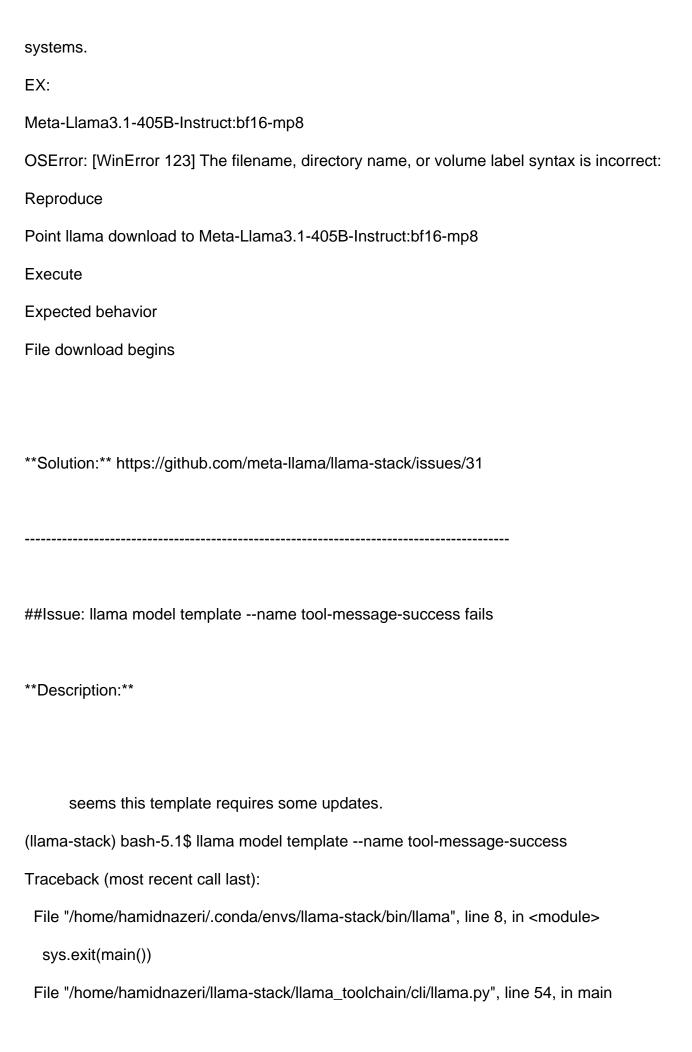
Solution: https://github.com/meta-llama/llama-stack/issues/48

##Issue: Colons in filenames are incompatible with windows

Description:

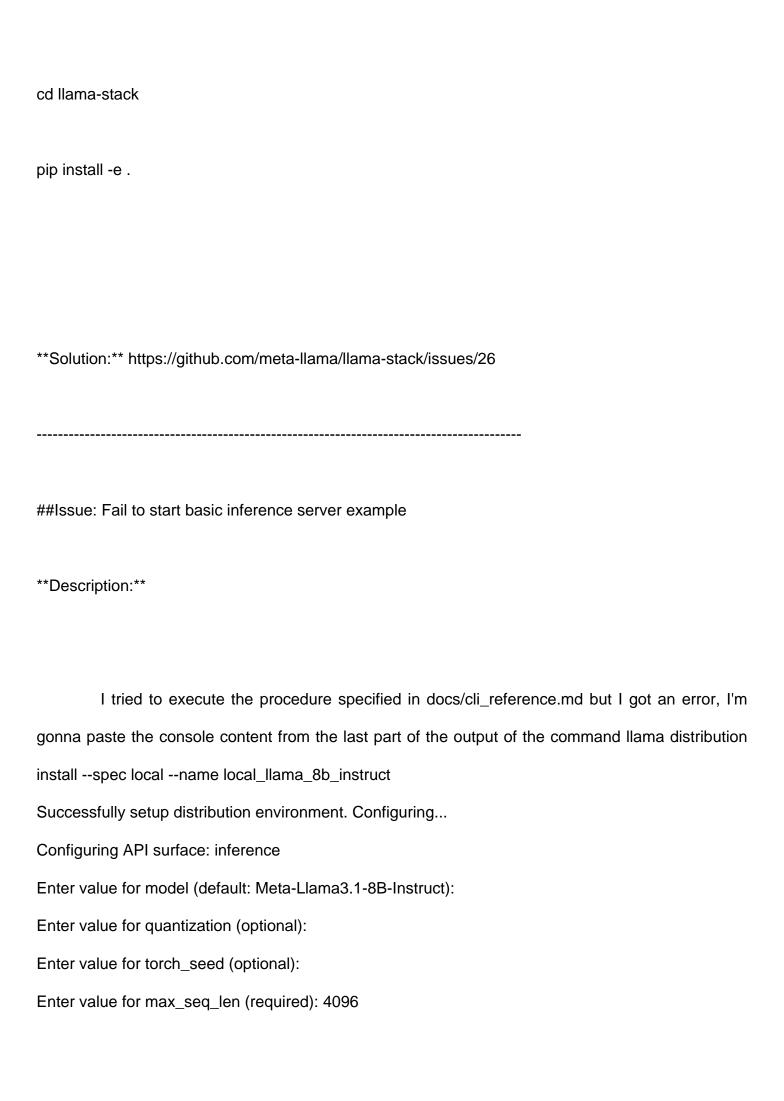
Describe the bug

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



```
parser.run(args)
 File "/home/hamidnazeri/llama-stack/llama_toolchain/cli/llama.py", line 48, in run
  args.func(args)
            "/home/hamidnazeri/llama-stack/llama_toolchain/cli/model/template.py",
     File
                                                                                            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



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

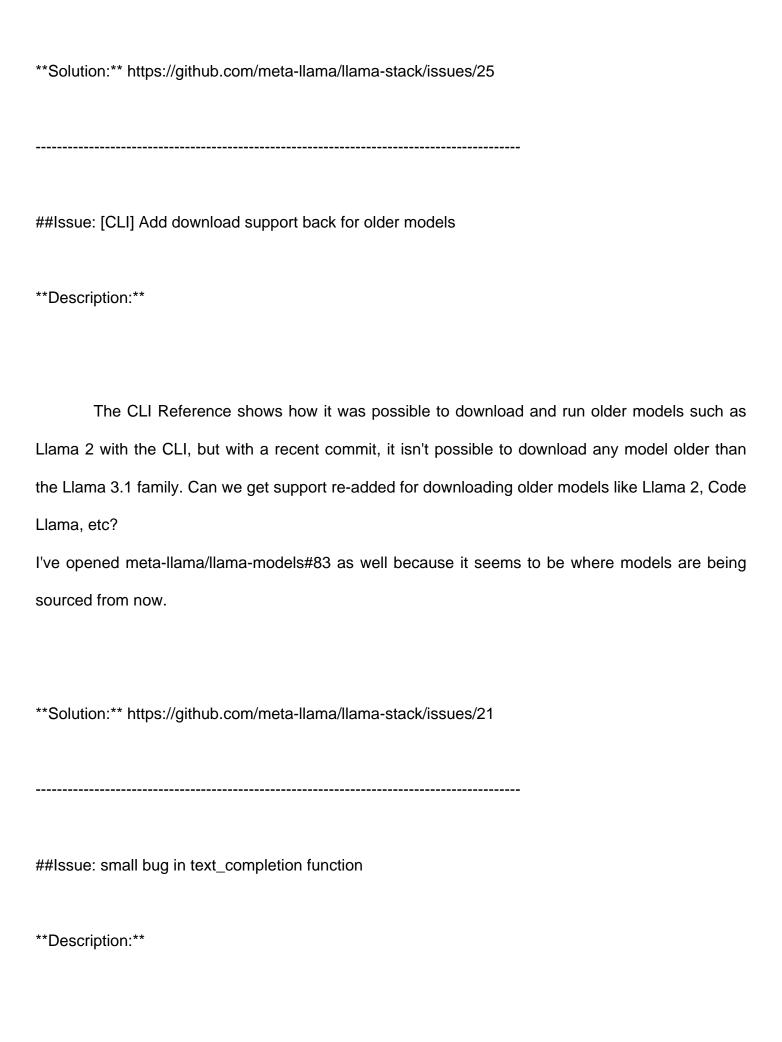
```
.py", line 133, in __call__
  return launch_agent(self._config, self._entrypoint, list(args))
                                                                                                 File
"/opt/conda/envs/local_llama_8b_instruct/lib/python3.10/site-packages/torch/distributed/launcher/api
.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]:
        : 2024-08-12_17:24:42
 time
        : ip-hidden
 host
         : 0 (local_rank: 0)
 rank
 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 535.18		sion: 535.183.01	CUDA Version: 12.2
'	Persistence-M Bus-		
Fan Temp Perf	Pwr:Usage/Cap	Memory-Usage	GPU-Util Compute M.
I	1 1	MIG M.	
		======	:=====+========
0 NVIDIA A10G	On 000000	00:00:1E.0 Off	0
0% 29C P8	9W / 300W 0N	1iB / 23028MiB	0% Default
	1	N/A	
+		+	·+
+			+
Processes:		I	
GPU GI CI F	PID Type Process nar	me	GPU Memory
ID ID		Usage	
==========			
======			
No running processes found			l
+			+

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

Thanks!



Hi, prompt_tokens = self.tokenizer.encode(x, bos=True, eos=False) should take prompt as input , not x, in $Ilama_toolchain/inference/generation.py$. 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 seg 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 Ilama model.

Fix it please. ?

Thank you.

