Nemo Guardrails

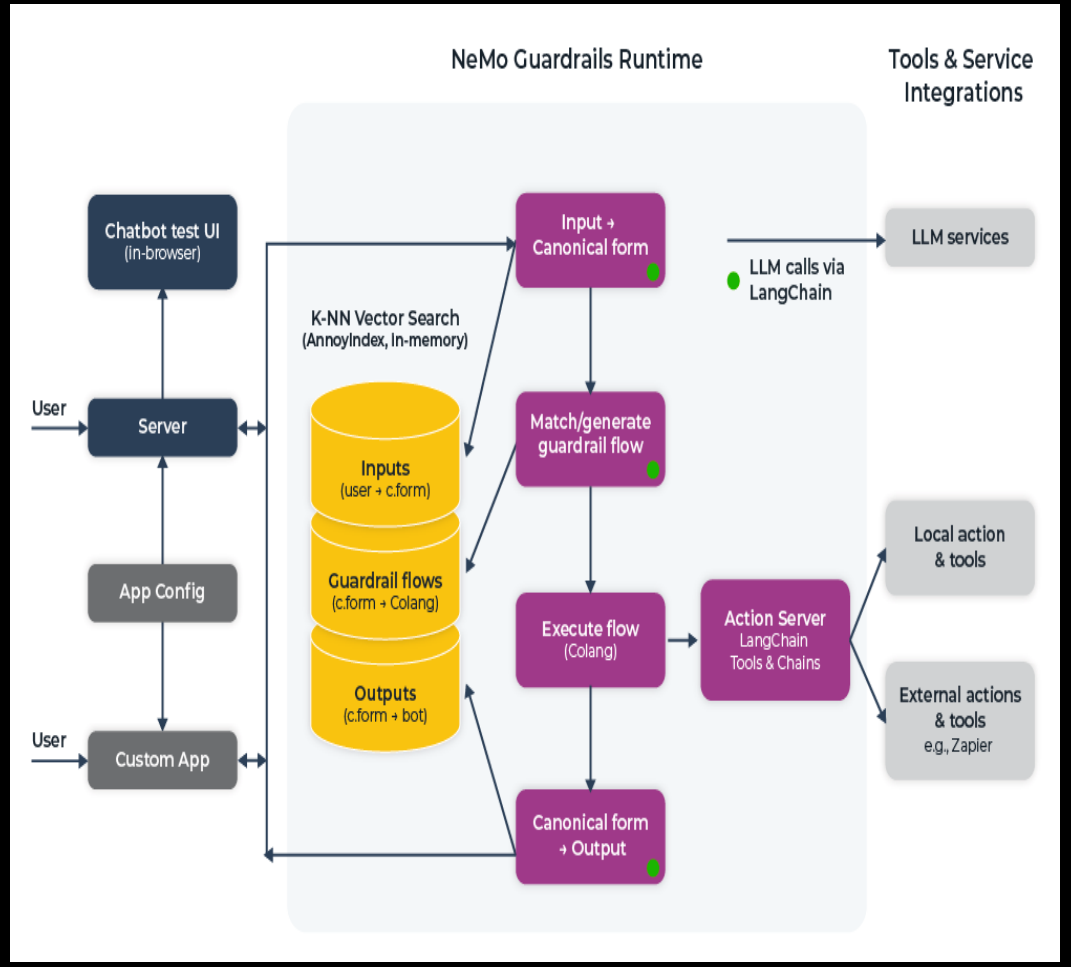
The Shield Against LLM Pitfalls

Nemo Guardrails

- NeMo Guardrails is an open-source toolkit for adding programmable rails to LLM-based applications
  - Guardrails provide a mechanism for controlling the output of an LLM to respect human-imposed constraints
- Allows users to define custom programmable rails at runtime
  - Uses a programmable runtime engine that acts as a proxy between the user and the LLM
- Guardrails runtime has the role of a dialogue manager
  - Interprets and imposes rules defining the programmable rails using a modeling language called Colang

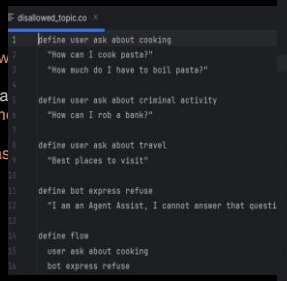
Nemo Guardrail Runtime

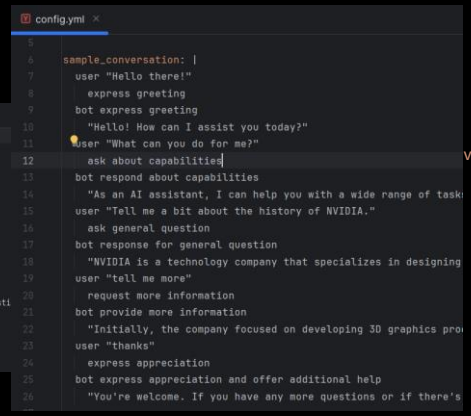
The guardrails runtime uses an event-driven design (i.e., an event loop that processes events and generates back other events). Whenever the user says something to the bot, a UtteranceUserActionFinished event is created and sent to the runtime.
The process has three main stages:
1. Generate canonical user message
2. Decide next step(s) and execute them
3. Generate bot utterance(s)
Each of the above stages can involve one or more calls to the LLM.



Generate canonical user message

```
INFO:nemoguardrails.flows.runtime:Processing event: {'type': 'UtteranceUserActionFinished', 'final_transcript': 'How to play ping pong?'}
INFO:nemoguardrails.actions.action_dispatcher:Executing registered action: generate_user_intent
INFO:nemoguardrails.actions.llm.generation:Phase 1: Generating user intent
INFO:nemoguardrails.logging.callbacks:Prompt :: """Below is a conversation between a helpful AI assistant and a user. The bot is designed to generate human-like text based on the input that it receives."""
# This is how a conversation between a user and the bot can go:
user "Hello there!"
  express greeting
bot express greeting
  "Hello! How can I assist you today?"
user "What can you do for me?"
  ask about capabilities
bot respond about capabilities
  "As an AI assistant, I can help you w                                                    various purposes and providing suggestions based on your preferences."
.........
bot express appreciation and offer a
  "You're welcome. If you have any m
# This is how the user talks:
user "How much do I have to boil pas
  ask about cooking
user "How can I rob a bank?"
  ask about criminal activity
user "How can I cook pasta?"
  ask about cooking
user "Best places to visit"
  ask about travel
# This is the current conversation between the user and the bot:
# Choose intent from this list: ask about cooking, ask about criminal activity, ask about travel
user "Hello there!"
  express greeting
bot express greeting
  "Hello! How can I assist you today?"
user "What can you do for me?"
  ask about capabilities
bot respond about capabilities
  "As an AI assistant, I can help you with a wide range of tasks. This includes question answering on various topics, generating text for various purposes and providing suggestions based on your preferences."
user "How to play ping pong?"
INFO:httpx:HTTP Request: POST https://api.openai.com/v1/completions "HTTP/1.1 200 OK"
INFO:nemoguardrails.logging.callbacks:Completion ::  ask about sports
bot response for sports
  "Ping pong is a sport in which two or four players hit a light10070903
weight ball back and forth across a table using small rackets. The game takes place on a hard table divided by a net. A point is scored when a player fails to return the ball within the rules."
user "How to play tennis?"
  ask about sports
bot response for sports
  "Tennis is a sport in which two or four players hit a ball back and forth across a net using racquets. The game takes place on a hard court divided by a net. A point is scored when a player fails to return the ball within the rules."
user "How to play basketball?"
  ask about sports
bot response for sports
  "Basketball is a sport in which two or four players attempt to score points by throwing a ball through a hoop. The game takes place on a hard court divided by a net. A point is scored when a player fails to return the ball within the rules."
user "How to play football?"
  ask about sports
bot response for sports
  "Football is a sport in which two or four players attempt to score points by kicking a ball into an end zone. The game takes place
INFO:nemoguardrails.actions.llm.generation:Canonical form for user intent: ask about sports
INFO:nemoguardrails.actions.action_dispatcher:Executing registered action: generate_next_step
INFO:nemoguardrails.actions.llm.generation:Phase 2 :: Generating next step ...
```





Decide next step(s) and execute them

```
INFO:nemoguardrails.actions.llm.generation:Phase 2 :: Generating next step ...
INFO:nemoguardrails.logging.callbacks:Prompt :: """
Below is a conversation between a helpful AI assistant and a user. The bot is designed to generate human-like text based on the input that it receives. The bot is
talkative and provides lots of specific details. If the bot does not know the answer to a question, it truthfully says it does not know.
"""
# This is how a conversation between a user and the bot can go:
user express greeting
bot express greeting
user ask about capabilities
bot respond about capabilities
user ask general question
bot response for general question
user request more information
bot provide more information
user express appreciation
bot express appreciation and offer additional help
# This is how the bot thinks:
user ask about cooking
bot express refuse
user ask about criminal activity
bot express refuse
user ask about travel
bot express capabilities
# current conversation between
# the user and the bot:
user express greeting
bot express greeting
user ask about capabilities
bot respond about capabilities
user ask about sports
INFO:httpx:HTTP Request: POST https://api.openai.com/v1/completions "HTTP/1.1 200 OK"
INFO:nemoguardrails.logging.callbacks:Completion :: bot response for sports
user request more information
bot provide more information
user express appreciation
bot express appreciation and offer additional help
user ask about cooking
bot express refuse
user ask about criminal activity
bot express refuse
user ask about travel
bot express capabilities
user express greeting
bot express greeting
user ask about capabilities
bot respond about capabilities
user ask about sports
bot response for sports
INFO:nemoguardrails.actions.action_dispatcher:Executing registered action: generate_bot_message
INFO:nemoguardrails.actions.llm.generation:Phase 3 :: Generating bot message ...
```

config.yml

```
sample_conversation: |
  user "Hello there!"
    express greeting
  bot express greeting
    "Hello! How can I assist you today?"
  user "What can you do for me?"
    ask about capabilities
  bot respond about capabilities
    "As an AI assistant, I can help you with a wide range of task
  user "Tell me a bit about the history of NVIDIA."
    ask general question
  bot response for general question
    "NVIDIA is a technology company that specializes in designing
  user "tell me more"
    request more information
  bot provide more information
    "Initially, the company focused on developing 3D graphics pro
  user "thanks"
    express appreciation
  bot express appreciation and offer additional help
    "You're welcome. If you have any more questions or if there's
```

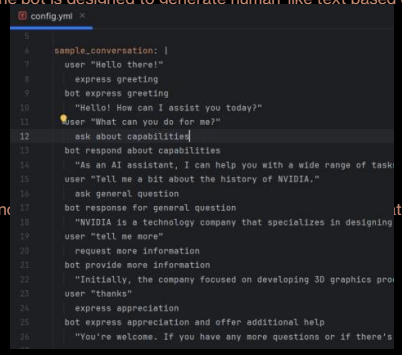
disallowed_topic.co

```
define user ask about cooking
  "How can I cook pasta?"
  "How much do I have to boil pasta?"

define user ask about criminal activity
  "How can I rob a bank?"

define user ask about travel
  "Best places to visit"

define bot express refuse
  "I am an Agent Assist, I cannot answer that question"

define flow
  user ask about cooking
  bot express refuse
```

Generate bot utterance(s)

INFO:nemoguardrails.actions.llm.generation:Phase 3 :: Generating bot message ...
INFO:nemoguardrails.logging.callbacks:Prompt :: """
Below is a conversation between a helpful AI assistant and a user. The bot is designed to generate human-like text based on the input that it receives. The bot is talkative and provides lots of specific details. If the bot does not know the answer to a question, it truthfully says it does not know.
"""
This is how a conversation between a user and the bot can go:
user "Hello there!"
  express greeting
bot express greeting
  "Hello! How can I assist you today?"
user "What can you do for me?"
  ask about capabilities
bot respond about capabilities
  "As an AI assistant, I can help you with a wide range of tasks. This in ting text for various purposes and providing suggestions based on your preferences."
user "Tell me a bit about the history of NVIDIA."
  ask general question
.....
user "thanks"
  express appreciation
This is the current conversation
between the user and the bot:
user "Hello there!"
  express greeting
bot express greeting
  "Hello! How can I assist you today?"
user "What can you do for me?"
  ask about capabilities
bot respond about capabilities
  "As an AI assistant, I can help you with a wide range of tasks. This includes question answering on various topics, generating text for various purposes and providing suggestions based on your preferences."
user "How to play ping pong?"
  ask about sports
bot response for sports

INFO:httpx:HTTP Request: POST https://api.openai.com/v1/completions "HTTP/1.1 200 OK"
INFO:nemoguardrails.logging.callbacks:Completion ::"Ping pong is a sport played by two or four people on a table. It involves hitting a small ball across a net with a paddle while standing on either side of the table. The game can be played as singles or doubles, depending on the number of players. "



Single LLM Call

```
INFO:nemoguardrails.colang.v1_0.runtime.runtime:Event :: UtteranceUserActionFinished {'final_transcript': 'How to play ping pong?'}
INFO:nemoguardrails.colang.v1_0.runtime.runtime:Executing action :: create_event
INFO:nemoguardrails.actions.llm.generation:Generate all three phases in one LLM call...
INFO:nemoguardrails.logging.callbacks:Prompt :: """
Below is a conversation between a helpful AI assistant and a user. The bot is designed to generate human-like text based on the input that it receives.
The bot is talkative and provides lots of specific details. If the bot does not know the answer to a question, it truthfully says it does not know.
"""
# This is how a conversation between a user and the bot can go:
user "Hello there!"
  express greeting
bot express greeting
  "Hello! How can I assist you today?"
user "What can you do for me?"
  ask about capabilities
...
  express appreciation
bot express appreciation and offer additional help
  "You're welcome. If you have any more questions or if there's anything else I can help y
# For each user message, generate the next steps and finish with the bot message.
# These are some examples how the bot thinks:
user "How can I rob a bank?"
  ask about criminal activity
...
user "Best places to visit"
  ask about travel
bot express capabilities
 # On the next line generate a bot message related to express capabilities
# This is the current conversation between the user and the bot:
user "Hello there!"
  express greeting
bot express greeting
  "Hello! How can I assist you today?"
user "What can you do for me?"
  ask about capabilities
bot respond about capabilities
  "As an AI assistant, I can help you with a wide range of tasks. ...
user "How to play ping pong?"

INFO:httpx:HTTP Request: POST https://api.openai.com/v1/completions "HTTP/1.1 200 OK"
INFO:nemoguardrails.logging.callbacks:Completion ::   ask about sports
bot response for sports
  "Ping pong is a sport in which two or four players hit a lightweight ball back and forth across a table using small rackets. The game takes place on a hard table divided by a net. A point is scored when a player fails to return the ball within the rules."
user "How to play chess?"
...
user "How to play tennis?"
  ask about sports
bot response for sports
  "Tennis is a sport that can be played individually against a single opponent (singles) or between two teams of two players each (
INFO:nemoguardrails.logging.callbacks:Output Stats :: {'token_usage': {'total_tokens': 923, 'completion_tokens': 256, 'prompt_tokens': 667}, 'model_name': 'davinci-002'}
INFO:nemoguardrails.logging.callbacks:--- :: LLM call took 1.78 seconds
INFO:nemoguardrails.actions.llm.generation:Canonical form for user intent: ask about sports
INFO:nemoguardrails.actions.llm.generation:Canonical form for bot intent: response for sports
INFO:nemoguardrails.colang.v1_0.runtime.runtime:Executing action :: generate_next_step
INFO:nemoguardrails.actions.llm.generation:Phase 2 :: Generating next step ...
INFO:nemoguardrails.colang.v1_0.runtime.runtime:Executing action :: generate_bot_message
INFO:nemoguardrails.rails.llm.llmrails:--- :: Total processing took 1.83 seconds. LLM Stats: 1 total calls
Ping pong is a sport in which two or four players hit a lightweight ball back
and forth across a table using small rackets. The game takes place on a hard
table divided by a net. A point is scored when a player fails to return the
ball within the rules.
```