NEMO GUARDRAILS: A TOOLKIT FOR CONTROLLABLE AND SAFE LLM APPLICATIONS WITH PROGRAMMABLE RAILS

AUTHORS:NVIDIA

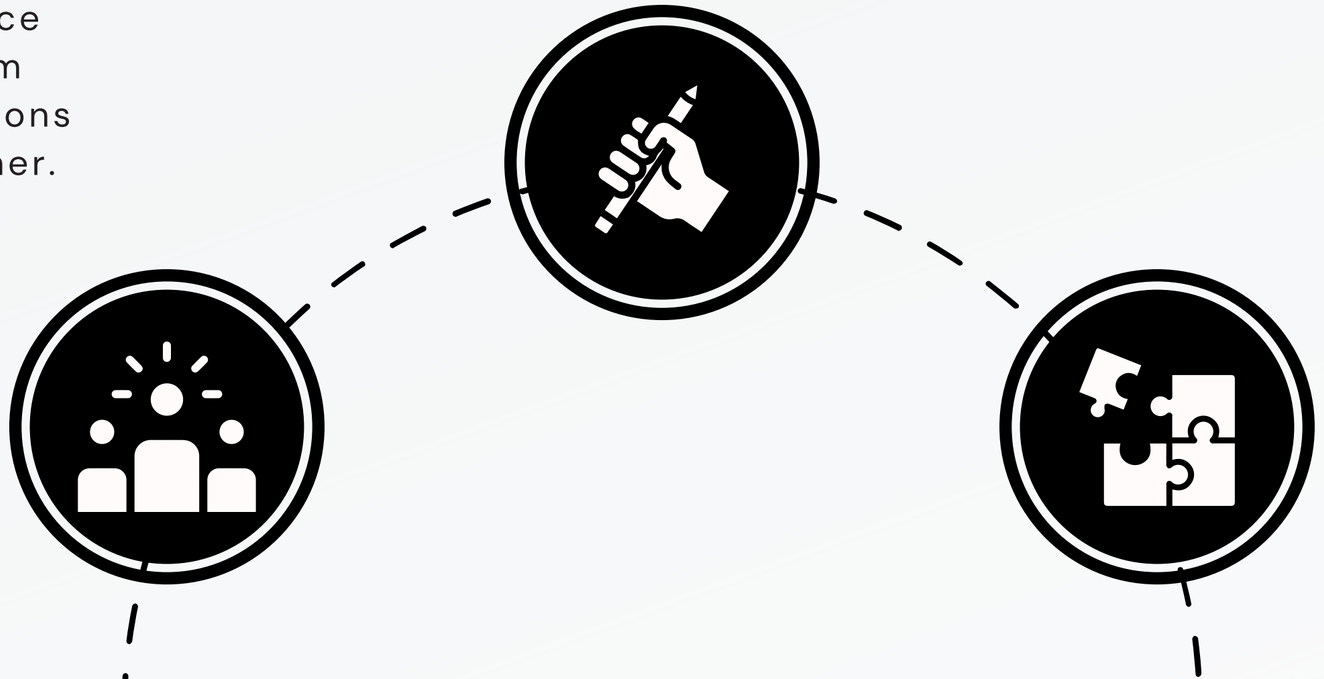
GUARDRAILS

Topical

prevent apps from veering off into undesired areas. For example, they keep customer service assistants from answering questions about the weather.

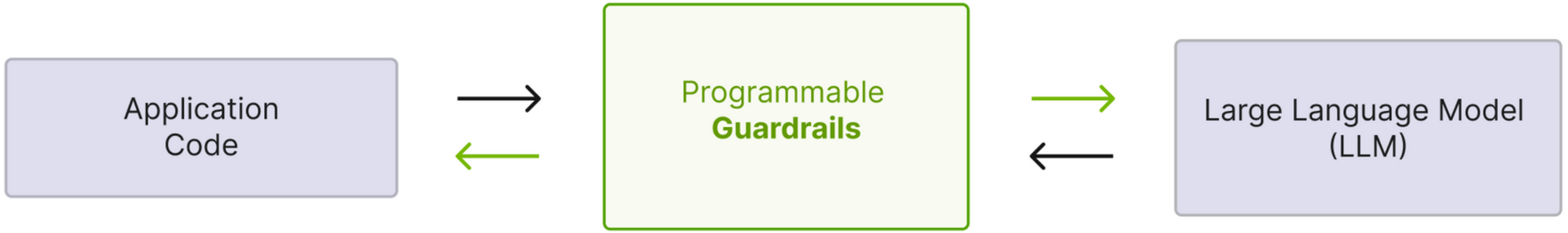
Safety

ensure apps respond with accurate, appropriate information. They can filter out unwanted language and enforce that references are made only to credible sources.

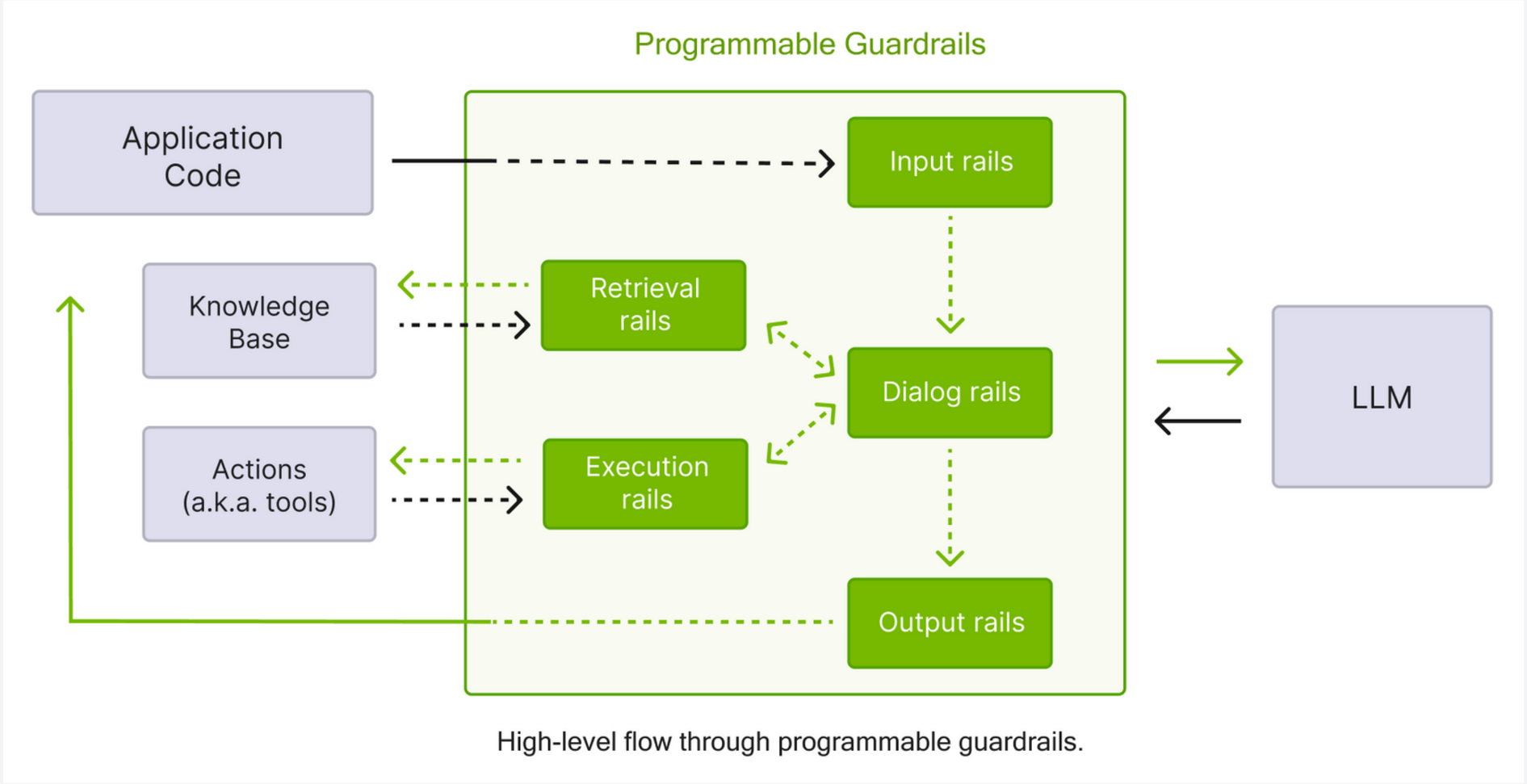


Security

restrict apps to making connections only to external third-party applications known to be safe.



Application code interacting with LLMs through programmable guardrails.



High-level flow through programmable guardrails.

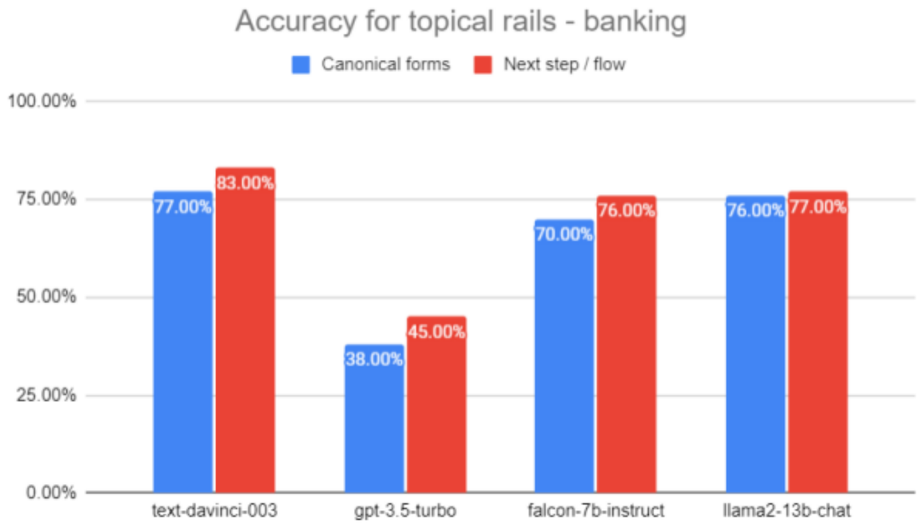


Figure 5: Performance of topical rails on Banking.

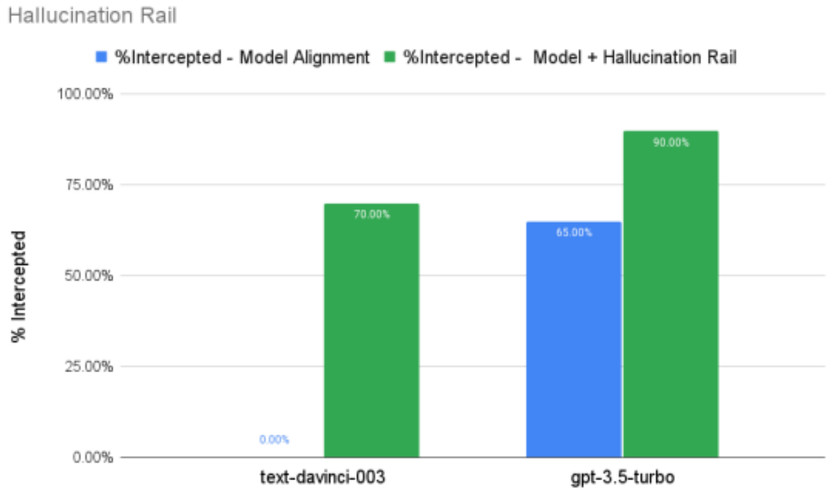


Figure 6: Performance of the hallucination rail.

Model	Us int, no sim	Us int, sim=0.6	Bt int, no sim	Bt int, sim=0.6	Bt msg, no sim	Bt msg, sim=0.6
text-davinci-003, k=all	0.89	0.89	0.90	0.90	0.91	0.91
text-davinci-003, k=3	0.82	N/A	0.85	N/A	N/A	N/A
text-davinci-003, k=1	0.65	N/A	0.73	N/A	N/A	N/A
gpt-3.5-turbo, k=all	0.44	0.56	0.50	0.61	0.54	0.65
dolly-v2-3b, k=all	0.65	0.78	0.68	0.78	0.69	0.78
falcon-7b-instruct, k=all	0.81	0.81	0.81	0.82	0.81	0.82
llama2-13b-chat, k=all	0.87	N/A	0.88	N/A	0.89	N/A

Table 2: Topical evaluation results on chit-chat dataset. **Us int** means accuracy for user intents, **Bt int** is accuracy for next step generation (i.e., the bot intent), **Bt msg** is accuracy for generated bot message. **Sim** denotes if semantic similarity was used for matching (with a specified threshold, in this case 0.6) or exact match.

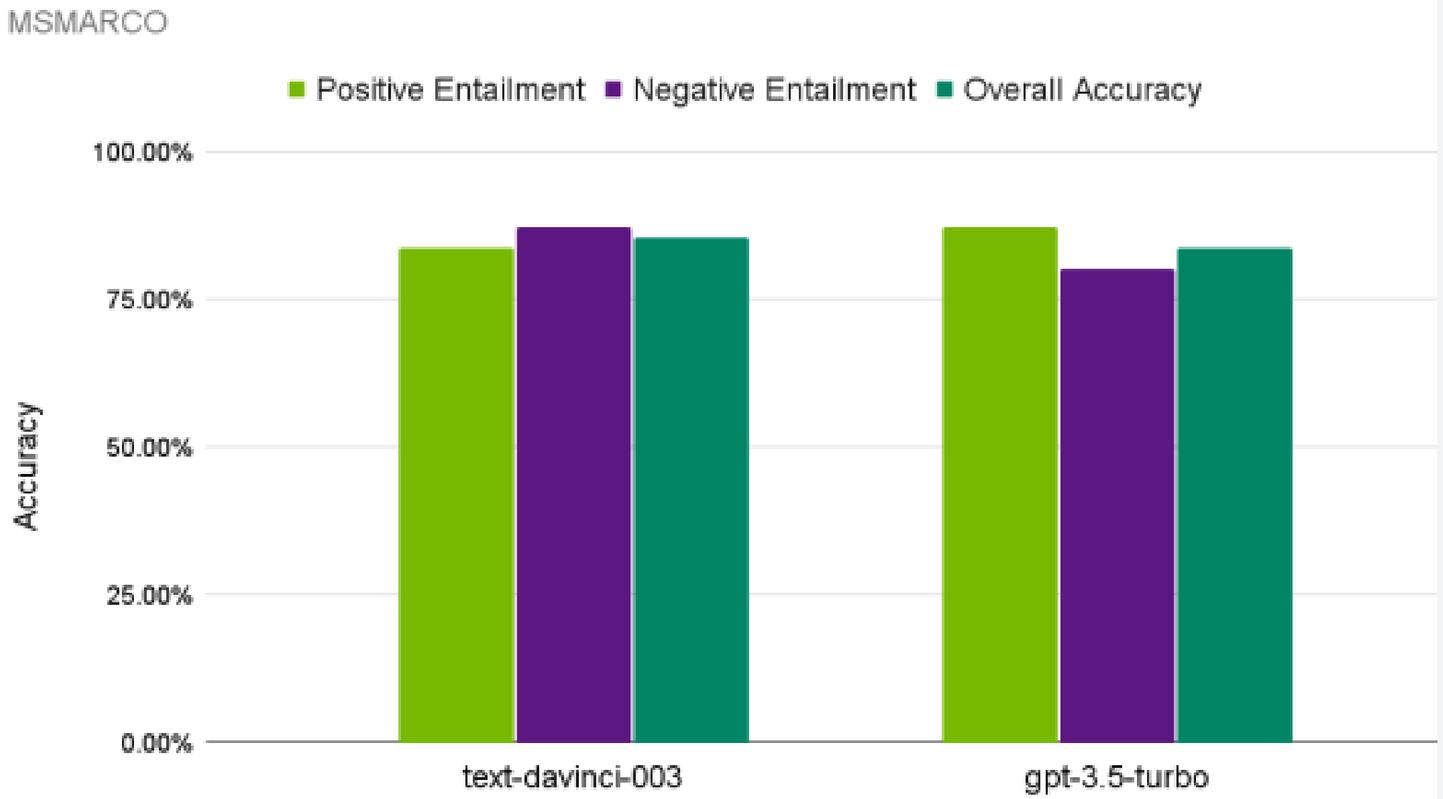
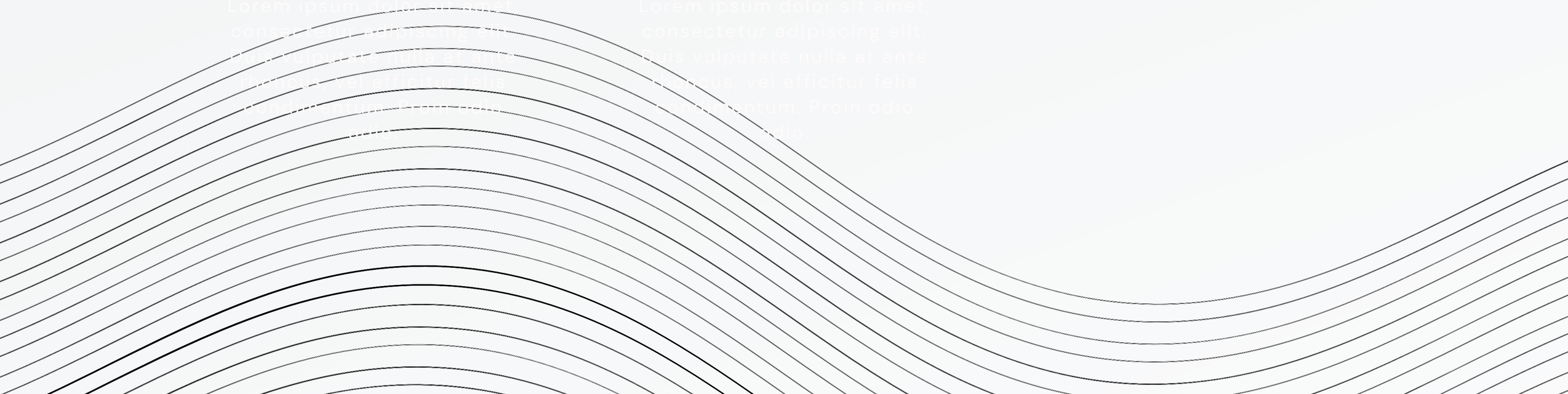


Figure 11: Performance of the fact-checking rail.

DEMO



THANK YOU

PAPER: https://arxiv.org/pdf/2310.10501.pdf
Toolkit: https://github.com/NVIDIA/NeMo-Guardrails/
Slides + Code :https://github.com/kesbeast23/guardrails-bot