Natural Language Processing and Large Language Models

Corso di Laurea Magistrale in Ingegneria Informatica

Lesson 22

Guardrails for LLMs

Nicola Capuano and Antonio Greco

DIEM – University of Salerno

Outline

- Adding guardrails to LLMs
- Techniques for adding guardrails
- Frameworks for implementing guardrails

Adding guardrails to LLMs

Guardrails

- Guardrails are mechanisms or policies that regulate the behavior of LLMs. They help to ensure that responses are safe, accurate, and context-appropriate.

- They can:
  - Prevent harmful, biased, or inaccurate outputs.
  - Align responses with ethical and operational guidelines.
  - Build trust and reliability for real-world applications.

- Examples are:
  - Blocking harmful content
  - Restricting outputs to specific domains.

Types of guardrails

- **Safety Guardrails**: Prevent generation of harmful or offensive content.
- **Domain-Specific Guardrails**: Restrict responses to specific knowledge areas.
- **Ethical Guardrails**: Avoid bias, misinformation, and ensure fairness.
- **Operational Guardrails**: Limit outputs to align with business or user objectives.

Techniques for adding guardrails

Techniques for adding guardrails

- Rule based filters
- Fine tuning with custom data
- Prompt Engineering
- External validation layers
- Real-time monitoring and feedback

Rule based filters

- Predefined rules to block or modify certain outputs.
- Examples:
  - Keyword blocking (e.g., offensive terms).
  - Regex-based patterns for filtering sensitive information.
- Simple and efficient for basic content filtering.

Fine tuning with custom data

- Train the model on domain-specific, curated datasets.
- Adjust weights to produce outputs aligned with guidelines.
- Examples:
  - Fine-tune for medical advice to restrict responses to accurate and safe recommendations.
  - Fine-tune for question answering on the topics of the course

Prompt Engineering

- Craft and/or refine prompts to guide the LLM behavior within desired boundaries.
- Examples:
  - "Respond only with factual, non-controversial information.“
  - ”Avoid speculative or unverifiable statements."

External validation layers

- Additional systems or APIs that post-process the model's outputs.
- Examples:
  - Toxicity detection APIs.
  - Fact-checking models.
- Allows modular and scalable implementation of guardrails.

Real time monitoring and feedback

- Monitor outputs continuously for unsafe or incorrect content.
- Flag or block problematic outputs in real-time.
- Tools:
  - Human-in-the-loop systems.
  - Automated anomaly detection.

Best practices

- Combine multiple techniques for robust safeguards.
- Example: Rule-based filtering + External validation + Fine-tuning.

Frameworks for implementing guardrails

Frameworks for implementing guardrails

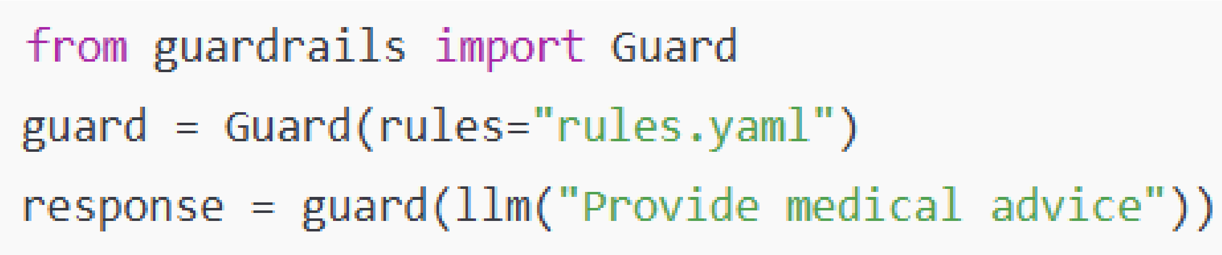
- The existing frameworks for implementing guardrails offer:
  - Easy integration with LLM APIs.
  - Predefined and customizable rulesets.

- Popular tools are:
  - **Guardrails AI**: A library for implementing safeguards.
  - **LangChain**: For chaining prompts and filtering outputs.
  - **OpenAI Moderation**: A prebuilt API to detect unsafe content.

Guardrails AI

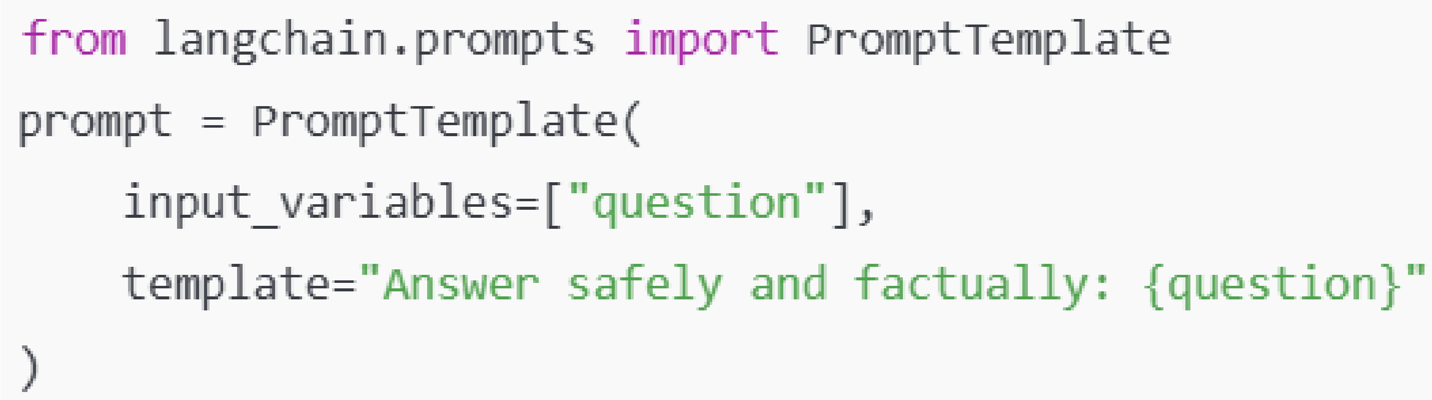
https://www.guardrailsai.com/

- **Validation**: Ensures outputs are within specified guidelines.
- **Formatting**: Controls the output structure.
- **Filters**: Removes or blocks unsafe content.



```
from guardrails import Guard
guard = Guard(rules="rules.yaml")
response = guard(llm("Provide medical advice"))
```

Langchain



```python
from langchain.prompts import PromptTemplate
prompt = PromptTemplate(
    input_variables=["question"],
    template="Answer safely and factually: {question}"
)
```

- Chains prompts with checks and filters.
- Verifies outputs against predefined criteria.

- Integrable with Guardrails: https://www.guardrailsai.com/docs/integrations/langchain

Try it yourself

Try it yourself

- Evaluate which are the techniques to add guardrails that are more suited for your purposes

- A possible suggestion may be to proceed by incrementally add complexity to the guardrails if you are not able to achieve a satisfying result with a simpler approach

- Give a careful look to the documentation of the existing frameworks

- Study similar examples that are available in the documentation of existing frameworks

- Try to apply guardrails to your project

Natural Language Processing and Large Language Models

Corso di Laurea Magistrale in Ingegneria Informatica

Lesson 22

Guardrails for LLMs

Nicola Capuano and Antonio Greco

DIEM – University of Salerno