servicenow

ServiceNow Text-to-Text SLM Model Card

ServiceNow Text-to-Text SLM Model Card

Intended Use and Functionality

Purpose of the Model

The Model supports ServiceNow users by enabling skills that require rapid inference and high throughput.

It supports Now Assist Guardian and other ServiceNow use cases, seamlessly integrating into application workflows to improve efficiency and user experience.

Autonomy Level

The Model operates at an Assistive autonomy level, providing AI-generated text suggestions that require human review before implementation.

While ServiceNow has aligned the Model to minimize biased and unsafe responses, users must implement human oversight to help ensure outputs are ethical, accurate, and appropriate for their intended use.

This human-in-the-loop approach ensures that users retain full control over AI-generated content, preventing fully autonomous decision-making.

Optimization Scope and Limitations

The Model is based on the Llama-3.18B large language model and has been further fine-tuned with ServiceNow proprietary data to optimize its performance for enterprise applications.

The Model is specifically designed to support high-performance text-to-text processing tasks within the ServiceNow ecosystem, ensuring alignment with enterprise requirements while maintaining efficiency and reliability.

Model Name

Llama 3.1 8B ServiceNow generic SLM ("the Model")

Model Version

v1.0

Model Release Date

November 2024

Model Distribution Method

ServiceNow Platform

Model License

Please refer to your agreement with ServiceNow for all license information.

Products Using this Model

Now Assist skills within application workflows on the ServiceNow Platform and Now Assist Guardian

User Benefits

The Model is fine-tuned for several natural language processing (NLP) tasks, including detecting offensive or harmful content, prompt injection and jailbreaking attacks. It also powers other ServiceNow use cases that aim to enhance user experience and admin productivity.

Risks

The Model may inaccurately predict harmful or offensive content, omit key information, or include irrelevant or redundant data. As a new technology, the Model, like all LLMs, can produce unpredictable, potentially inaccurate, biased, or objectionable outputs.

Factors and limitations

The Model needs sufficient context in a prompt to create an acceptable response.
If the prompt does not contain sufficient instructions for the Model to follow, it may make assumptions and produce a response which is not in line with the expected response.

Ethical considerations

The Model has been fine-tuned to handle contexts that may involve sensitive or potentially harmful topics. However, like other industry-leading large language models, limitations still exist, and the Model may generate unintended or harmful content based on how it is prompted.

When using the model customers should follow ServiceNow's guidelines on intended use available on docs.servicenow.com as well as ServiceNow's AI Acceptable Use Policy.

Please report instances of unintended hallucinations, harmful text (e.g. toxicity, profanity, etc.), or unexpected PII occurrences in the output so that we can evaluate for remediation.

Supported Languages

Primary Language: English
Multilingual Capabilities: The model may respond to non-English inputs, but accuracy is not guaranteed.

Model Architecture

Base Model: Llama-3.18B
Fine-Tuning: The Model has been further tuned by instruction fine tuning for ServiceNow tasks using ServiceNow proprietary data and synthetic data.

Number of Parameters

8B

Maximum Input and Output Size

- **Maximum Input Size:** 128,000 tokens (shared between input and output)
- **Maximum Output Size:** 2,048 tokens
- **Shared Context Window:** 128,000 tokens

Input and Output Modalities

Modality Type

Single-modality: The Model processes and generates **text-to-text** outputs.

Inputs

Input Type: The Model accepts **plain text** input, typically structured as questions, commands, or prompts in natural language (e.g., "Summarize this article" or "Translate this sentence to French").

Input Constraints:

- The Model is designed primarily for natural language inputs and may perform sub-optimally with **non-text inputs** such as raw numerical data or unstructured code.
- Some **long-form inputs** may be truncated based on token limits, affecting response completeness.
- The Model may struggle with **highly ambiguous or domain-specific jargon** if not adequately trained in those areas.

Outputs

Output Type: The Model generates **plain text responses** based on the input prompt. This can include answers to questions, summarizations, translations, code snippets, or structured text.
Output Constraints: Outputs are constrained by **token length limits**, which may truncate longer responses.

Input and Output Formats

- **Input Format:** Inputs must be formatted as **plain text** with no additional structuring required.
- **Output Format:** Outputs are generated as **plain text** with no additional structuring unless explicitly prompted.

Training Data

ServiceNow did not conduct further pre-training.

Fine-Tuning Data

- The Model has been further tuned by instruction fine tuning for ServiceNow tasks using a mix of proprietary and synthetic datasets which are targeted towards enterprise use cases as well as general purpose capabilities.
- The Model has also undergone further alignment designed to improve helpfulness and harmlessness.

Evaluation Data

- The Model undergoes evaluation based on a variety of academic benchmarks to assess conversational ability, ability to follow instructions, reasoning, multi-lingual understanding, potential for harmful output, and other capabilities.
- The Model is also benchmarked on custom-designed synthetic datasets, which resemble enterprise use cases.
- While the benchmarks provide enough insight on model performance in various scenarios, we urge developers to assess the performance for the application being developed.

Metrics

The model was evaluated with the following metrics: F1, Precision, Recall, Correctness, False Positive Rate (PFR) and metrics from other academic benchmarks.

Technical Means for Integration

All interactions and processing occur within the platform's secure architecture.