servicenow

ServiceNow Small Language Model (SLM) Model Card

ServiceNow Small Language Model (SLM) Model Card

Intended Use and Functionality

Purpose of the Model

The **Model** is designed to support **enterprise AI applications** by enhancing text-based automation and content generation within **ServiceNow workflows.** The model is optimized for **text-to-text processing, code generation, workflow generation (text-to-flow), summarization, question answering, and query handling,** ensuring alignment with **ServiceNow-specific enterprise use cases.**

- **Enterprise Optimization:** Tailored for **business process automation,** helping to ensure generated content meets **ServiceNow platform requirements.**
- **Model Consolidation:** Consolidates multiple **ServiceNow-related tasks** into a single, efficient model, reducing system complexity while enhancing performance.
- **Workflow Integration:** Supports AI-driven text generation **for enterprise applications, knowledge retrieval, and automated reporting.**

Autonomy Level

The model operates at an **Assistive autonomy level,** generating AI-powered text suggestions that users can **review, edit, and approve** before finalizing.

This **human-in-the-loop** approach helps to ensure that users retain **full control over AI-generated content, enabling transparent and accurate decision-making.**

The model **does not support fully autonomous decision-making,** requiring **human validation** before applying AI-generated responses in **business-critical workflows.**

Model Name

ServiceNow Small Language Model ("the Model")

Model Version

v1.0

Model Release Date

May 2025

Model Distribution Method

ServiceNow Platform

Model License

Apache 2.0

Products Using this Model

Now Assist skills within app workflows, as well as the content monitoring and evaluation layer on the ServiceNow platform.

Optimization Scope and Limitations

The Model is fine-tuned for **enterprise-specific text processing and automation,** with a focus on:

- **Text Generation:** Supports **content creation, summarization, and text formatting** within enterprise applications.
- **Enterprise Knowledge Retrieval:** Optimized for **question answering and data extraction relevant to business workflows.**
- **Code and Query Processing: Text-to-Code and Text-to-Cypher capabilities for ServiceNow automation tasks.**
- **Workflow Generation (Text-to-Flow):** Translating natural language inputs into workflows within **Flow Designer.**
- **Content Moderation:** Optimized for detecting harmful content, enforcing security policies, and supporting compliance with enterprise AI governance standards.

User Benefits

The Model is fine-tuned for **enterprise natural language processing (NLP) tasks, enabling efficient automation and text generation within ServiceNow workflows.** The model supports various AI-driven capabilities, designed to enhance **user productivity and enterprise automation.**

- **Instruction Adherence and Response Generation:** Helps ensure that AI-generated outputs align with business logic, enterprise standards, and workflow requirements.
- **Summarization and Knowledge Retrieval:** Optimized for content summarization, enterprise search, and knowledge base (KB) generation, enabling faster access to critical information.
- **Question Answering (Q&A):** Enhances enterprise query resolution, allowing users to retrieve relevant information based on structured prompts.
- **Intent Recognition and Classification:** Supports text understanding and automation, improving response accuracy in AI-driven workflows.
- **Code and Query Generation:** Fine-tuned for Text-to-Code and Text-to-Cypher tasks, assisting users in automating programming and database interactions.
- **Workflow Generation (Text-to-Flow):** Automatically generate multi-step flows in Workflow Studio with configured triggers, actions, data pill values, and text instructions, following standard design patterns.

By leveraging these capabilities, users benefit from **improved efficiency, streamlined workflows, and AI-assisted content creation while maintaining control over AI-generated outputs.**

Risks

Potential Limitations: As with any language model, the potential outputs cannot be predicted in advance due to the probabilistic nature of generative AI. The model can produce inaccurate, biased, or otherwise objectionable responses to user prompts.

Factors and limitations

Optimization Scope:

- The Model is a general-purpose AI system optimized for a broad range of ServiceNow applications, including Text-to-Code, Text-to-Cypher, Content Moderation, Workflow Generation (Text-to-Flow), and agent-related use cases.
- It consolidates multiple functionalities into a singular high-performance architecture, reducing complexity while improving efficiency.

Input Requirements:

- The Model relies on **structured, well-defined prompts** to generate high-quality outputs.
- **Ambiguous, incomplete, or overly generic prompts may lead to misinterpretations, incorrect results, or suboptimal performance** in text generation and code-related use cases.
- For **code generation,** the Model performs best when provided **clear problem statements, sufficient context, and examples of expected output.**

Data Scope:

- The Model is primarily trained & fine-tuned on a combination of **open-source data, ServiceNow platform-specific datasets, and a curated selection of mathematical, multilingual, reasoning, and instruction-following tokens.**

Domain-Specific Challenges:

- The Model is **not explicitly designed for niche domains requiring specialized technical knowledge.**

Ethical considerations

The Model has been fine-tuned with the intention of reducing bias, toxicity, and hallucinations, although such limitations may still exist due to the probabilistic nature of generative AI.

Text LLMs can produce harmful text based on how they are prompted, and the Model is not free from such limitations, consistent with other industry LLMs. 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 data occurrences in the model output so that we can evaluate for remediation.

Supported Languages

The Model was trained on a **diverse multilingual dataset** to help ensure strong **language proficiency, text generation, and enterprise automation support.** The dataset includes **extensive linguistic coverage, with a focus on enterprise-relevant languages.**

Multilingual Coverage:

- Supports **a broad range of P1 and P2 languages (English, French, Dutch, German, Spanish, Italian, Brazilian Portuguese, Portuguese, French Canadian, Japanese),** helping to ensure effective AI-driven text generation across multiple linguistic contexts.
- Includes **specialized datasets to enhance fluency and contextual accuracy, particularly in business and enterprise interactions.**

English Proficiency:

- Optimized for **enterprise text processing tasks, including summarization, question answering, content generation, and instruction-following.**
- Trained on **text datasets to help ensure coherence, factual accuracy, and response generation.**

Supported Coding Languages:

- **JavaScript (GlideScript):** Optimized for **ServiceNow Platform development and code generation tasks.**
- **Python:** Supported for **general-purpose scripting and automation, leveraging task-specific coding datasets.**
- **SQL (Cypher Query Language):** Integrated for **Text-to-Cypher tasks, enabling query generation and database interaction.**

Model Architecture

The Model is based on **Mistral-Nemo-12B, a transformer-based dense language model** with **12 billion parameters.**

It is designed as a **general-purpose AI system** optimized for a broad range of **ServiceNow applications, including Text-to-Text, Text-to-Code, Workflow Generation (Text-to-Flow), Text-to-Cypher, and Content Moderation tasks.**

The model consolidates multiple functionalities into a **single, high-performance architecture,** reducing system complexity while enhancing efficiency.

Model Architecture Key Components

- **Instruction Adherence & Reasoning Module:** Enables response generation with improved **logical reasoning, context retention, and instruction-following capabilities** across diverse query types, including complex prompts, structured data formats, and knowledge-based tasks.
- **Multilingual Processing Engine:** Supports **all P1 & P2 languages,** featuring a dedicated **Japanese language model enhancement** for improved syntactic and semantic comprehension. Optimized for multilingual chat performance and translation accuracy.
- **ServiceNow Agent Optimization Layer:** Tailored for enterprise automation, enabling **case summarization, chat summarization, resolution note generation, and knowledge base content synthesis** with improved precision and contextual awareness.
- **Text-to-Code & Text-to-Cypher Generator:** Provides **specialized model pathways for Glide JavaScript and generic JavaScript editing,** along with query formulation and execution. Enhances **code generation, auto-completion, and refinement** for enterprise-grade automation.
- **Workflow Generation (Text-to-Flow):** Powers automated multi-step flow creation in ServiceNow's Flow Designer using natural language input. Leverages structured prompt interpretation to configure triggers, actions, data pill values, and logic paths aligned with standard workflow design patterns, enabling intuitive process automation for both technical and non-technical users.
- **Unified Model Architecture for Deployment Efficiency:** Designed to **consolidate multiple AI models** into a single deployment framework, reducing **fragmentation, resource utilization, and operational overhead** while maintaining performance across diverse AI-driven workflows.
- **Adaptive Content Moderation & Security Layer:** Integrates **robust adversarial defense mechanisms** to mitigate jail-breaking attempts, harmful content generation, and policy violations. Features a **low false positive filtering system** to help ensure responsible AI deployment.
- **Alignment** was conducted with SimPO (Simple Preference Optimization), which improves the Model's ability to follow instructions, making it more helpful and safer.

Number of Parameters

12B

Maximum Input and Output Size

- **Maximum Input Size:** The model supports a context window of up to **32K tokens**. This includes the prompt, system instructions, and any additional user-provided input.
- **Maximum Output Size:** The total number of output tokens is limited by the remaining space in the 32K token window after accounting for the input length. For example, with a 10,000-token input, the model can generate up to 22K tokens. Actual output limits may be constrained further by runtime settings or system resources.
- **Shared Context Window:** The **32K limit** is shared between **input** and **output**. The model dynamically allocates space for generation based on the size of the prompt.

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

The Model went through continued pre-trained (CPT) on a **diverse dataset** to enhance its capabilities in **multilingual processing, code generation, workflow generation, mathematical reasoning, instruction adherence, and content moderation.** The dataset was carefully curated to help ensure **broad coverage across enterprise applications while maintaining data quality, relevance, and alignment with business automation needs.**

Types of Data Used

- **Multilingual Data:** Included diverse language datasets to improve comprehension, fluency, and contextual accuracy across multiple languages.
- **Code and Query Data:** Integrated datasets for Text-to-Code and Text-to-Cypher, enhancing the model's ability to generate and interpret programming and database queries.
- **Mathematical and Logical Reasoning Data:** Focused on problem-solving, calculations, and logical reasoning relevant to enterprise AI applications.
- **Enterprise-Specific Instruction Data:** Curated datasets tailored for instruction adherence, response generation, and workflow automation within ServiceNow applications.

Fine-Tuning Data

The Model was fine-tuned using **task-specific datasets** to improve its ability to **generate accurate responses, support multilingual interactions, and enhance security in enterprise AI workflows.** The datasets cover a broad range of **business automation, coding tasks, content moderation, and language processing** to help ensure adaptability across **text generation, code processing, and workflow integration.**

Types of Data Used

Enterprise AI Use Case Data:

- Included datasets for **case summarization, chat summarization, knowledge base (KB) prompting, conversation cataloging, skill discovery, and classification** to support AI-powered enterprise automation.
- Focused on response generation and **workflow integration.**

Code Generation and Query Data:

- Integrated datasets for **Text-to-Code and Text-to-Cypher, covering JavaScript, Python, SQL, and Cypher to improve code generation, function calling, and complex reasoning tasks.**

Content Moderation and Security Data:

- Included datasets for **harmful content filtering, bias mitigation, and adversarial robustness,** helping to ensure AI compliance with **enterprise safety standards.**
- Incorporated **red teaming datasets** to enhance security against adversarial misuse.

Mathematical and Logical Reasoning Data:

- Contained **numerical reasoning, structured logic, and domain-specific STEM-based datasets** to improve AI-driven problem-solving for **enterprise decision-making tasks.**

Long-Context and Multi-Turn Interaction Data:

- Included datasets optimized for **handling extended conversations, workflows, and multi-step task processing,** improving AI usability for **business-critical applications.**

Instruction Adherence and Data Processing Data:

- Focused on **enhancing response generation,** helping to ensure AI outputs align with **business logic, workflow automation, and instruction-following requirements.**

Multilingual and Adaptive Use Case Data:

- Covered **P1 and P2 languages, supporting language adaptability and improved context awareness across multilingual interactions** in enterprise applications.

ServiceNow Platform Workflow Data:

- Platform-specific data from **Flow Designer**, ServiceNow's process automation builder with reusable actions and natural language descriptions that enable both technical and non-technical users to build workflows. This data was used to train the model to generate multi-step flows from user instructions, optimizing it for enterprise automation tasks.

Evaluation Data

The Model was evaluated using **datasets covering instruction adherence, multilingual capabilities, enterprise automation, code generation, and content moderation.** The datasets were selected to reflect **real-world enterprise applications,** helping to ensure that the model performs reliably across **business-critical use cases.**

Types of Data Used for Evaluation

Instruction Adherence Data:

- Included datasets assessing the model's ability to **process instructions, generate accurate responses, and follow enterprise automation protocols.**
- Evaluated against **benchmarks for instruction execution and response generation.**

Conversational and Multilingual Data:

- Assessed **coherence, contextual understanding, and accuracy** across **multiple languages and enterprise-specific dialogues.**
- Included **multi-turn conversation datasets and multilingual test cases.**

Enterprise-Specific Use Cases:

- Covered **case summarization, chat summarization, knowledge base (KB) generation, and business process automation.**
- Designed to measure AI effectiveness in **customer service, IT, and HR applications.**

Code Generation and Query Processing Data:

- Evaluated the model's ability to **generate, refine, and edit code for JavaScript, Python, SQL, and Cypher.**
- Included **Text-to-Code and Text-to-Cypher datasets,** ensuring AI-driven automation meets **enterprise coding standards.**

Security and Content Moderation Data:

- **Adversarial Robustness Data:** Integrated **red teaming datasets** to test the model's resilience against **adversarial attacks, prompt manipulation, and security vulnerabilities.** Included evaluations on **jailbreaking attempts, role-playing exploits, and unauthorized system access scenarios.**
- **Content Moderation and Safety Data:** Trained on datasets focused on **identifying and filtering unsafe or policy-violating content.** Evaluated responses for **harmful language, inappropriate content, and security risks** to help ensure compliance with enterprise standards.

ServiceNow Platform Workflow Data:

- Evaluation was conducted using two primary datasets: a synthetic set created by subject matter experts to test structured, controlled scenarios, and a customer dataset consisting of real-world workflows. These datasets were used to assess the model's ability to generate accurate, complete, and logically consistent flows by comparing AI-generated outputs to ground truth using workflow tree similarity metrics.

Metrics

The Model was evaluated using a combination of **automated and human evaluation methods** to measure its **instruction adherence, text generation quality, coding capabilities, query processing, and security performance.** These metrics were selected to ensure the model aligns with **enterprise AI needs and real-world business applications.**

Key Evaluation Metrics

- **Instruction Adherence:** Measures the model's ability to follow structured prompts and generate expected responses.
- **Text Generation Quality:** Assesses coherence, informativeness, groundedness, hallucination, and engagement in generated outputs across different languages.
- **Task-Specific Accuracy:** Evaluates performance in key enterprise AI use cases such as **summarization, question answering, and knowledge retrieval.**
- **Code Generation Performance:** Measures accuracy and efficiency in **code completion, code editing, and query processing** for enterprise automation tasks.
- **Security and Content Moderation:** Assesses robustness against malicious output and adversarial prompts, e.g., jailbreaking, role playing, persuasion.
- **Workflow Generation Quality:** Evaluates the accuracy, completeness, and logical consistency of AI-generated workflows, helping to ensure they align with user intent and enterprise automation best practices.

Technical Means for Integration

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