# Nemotron 3.5 Bolt E4B 4GB: An Open and Efficient Small Language Model for Fast Conversational Agents

**NVIDIA**

**Abstract.** We introduce Nemotron 3.5 Bolt E4B 4GB, an open and efficient small language model further post-trained from Gemma-4-E4B-it and optimized for speed and intelligence per GB of memory used for fast conversational agents. The model carries forward Gemma's strong conversational capabilities and further improves multi-turn tool use, instruction following, and input grounding for fast, local voice/device assistants, interactive personas, and embedded/Edge interfaces. The model combines natural conversation with multi-turn agentic capabilities to power voice/device assistants, interactive characters and robotic/IoT interfaces running on GeForce RTX, Jetson, and RTX/DGX Spark. Experiments on 21 benchmarks show that Nemotron 3.5 Bolt E4B 4GB achieves the highest average accuracy while generating the fewest output tokens in both reasoning and non-reasoning modes, compared with Gemma-4-E4B-it and Qwen3.5-4B. The model is released with a DFlash drafter. With DFlash, Bolt E4B achieves 2.74× the per-user decode speed of inference without speculative decoding at concurrency 1, and 1.21× the generation throughput of Gemma-4-E4B-it with MTP at batch size 1.

## 1. Introduction

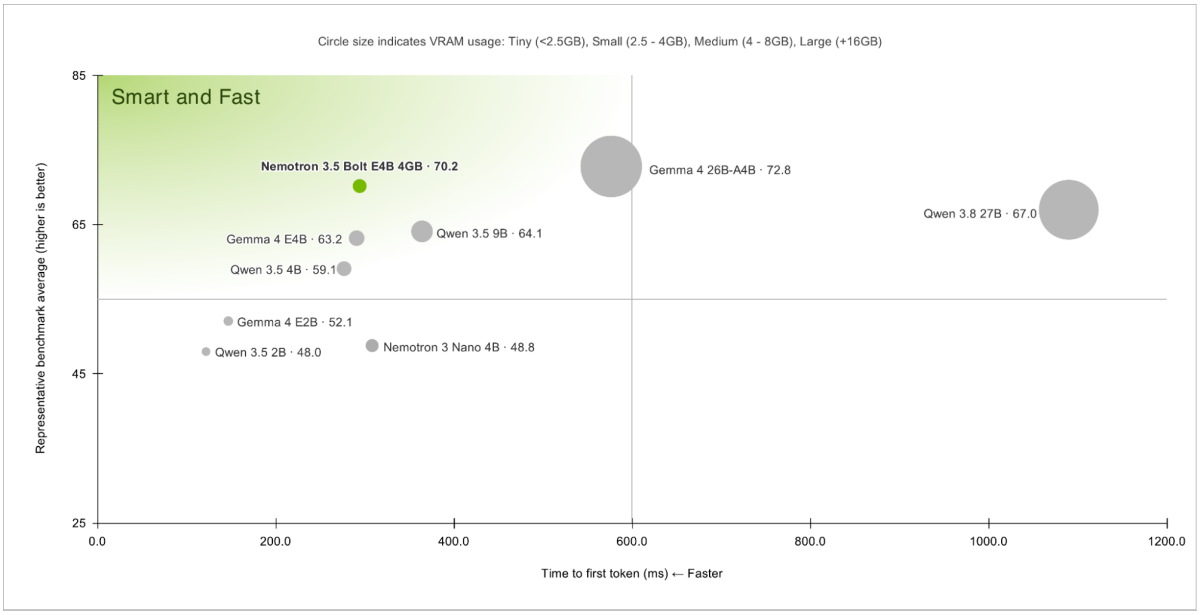



**Figure 1** | Average benchmark score versus time to first token (TTFT). Scores are averaged across MultiChallenge, RoleMRC, BFCL-v3, IHEval, SysBench, and RelyToolBench. Local models use Q4_K_M quantization and are measured on RTX 5090 with input/output sequence lengths of approximately 3,000/50 tokens. Cloud reference results (score; TTFT): Gemini 3.8 Flash (77.1; 2,449.5ms) and Claude Sonnet 5.5 (76.2; 3,002.2ms).

Voice agents, interactive characters, and robotic interfaces need language models that can hold a conversation and take actions reliably. These models must follow instructions, use tools across

© 2026 NVIDIA. All rights reserved.

multiple turns, and make sense of conversation history, application state, and tool results. For local deployment, they must also respond quickly and fit within limited memory.

We introduce Nemotron 3.5 Bolt E4B 4GB, an open small language model designed for fast conversational agents with a target memory footprint of approximately 4 GB. Nemotron 3.5 Bolt E4B 4GB supports reasoning and non-reasoning modes. The model can also generate concise conversational responses by enabling *conversational_mode*.

Nemotron 3.5 Bolt E4B 4GB is based on Gemma4-E4B-it (Gemma, 2026) and further trained using RLVR and RLHF through NeMo-RL (NVIDIA, 2025b) and NeMo-Gym (NVIDIA, 2025a). Training covers conversational tool use, tool-call schemas, instruction following, instruction hierarchy, and output formatting. We use every training sample in both reasoning and non-reasoning modes. This choice emphasizes accuracy without explicit reasoning while preserving reasoning capabilities for tasks that benefit from them.

We evaluate Nemotron 3.5 Bolt E4B 4GB on 21 benchmarks spanning conversational proficiency, multi-turn tool calling, instruction following/control, context robustness, and state/action grounding. Across this diverse evaluation suite, Nemotron 3.5 Bolt E4B 4GB achieves the highest average benchmark score compared with Gemma4-E4B-it and Qwen3.5-4B in both reasoning and non-reasoning modes. It also generates fewer output tokens on average than either baseline in both reasoning and non-reasoning modes. These results demonstrate that Nemotron 3.5 Bolt E4B 4GB is a strong fit for fast conversational agents.

For deployment, we quantize the BF16 checkpoint to Q4_K_M with importance-matrix calibration. We also provide a DFlash parallel drafter to accelerate generation through speculative decoding. The model is available via HF and NVIDIA NIM.

### Checkpoints

- `NVIDIA-Nemotron-3.5-Bolt-E4B-4GB` 🤗: Q4_K_M-quantized model in GGUF format.
- `NVIDIA-Nemotron-3.5-Bolt-E4B-BF16` 🤗: BF16 model in Hugging Face format.
- `NVIDIA-Nemotron-3.5-Bolt-E4B-4GB-DFlash` 🤗: DFlash drafter for the GGUF model, in GGUF format.
- `NVIDIA-Nemotron-3.5-Bolt-E4B-BF16-DFlash-NVFP4` 🤗: DFlash drafter for the BF16 model, in HF format.

### NVIDIA NIM

- `build.nvidia.com/nvidia/nemotron-3.5-bolt-e4b` : Start building with a free API endpoint.
- `Nemotron-3.5-bolt-e4b-4gb` : Download NVIDIA NIM for local deployment.

### Data

- `Nemotron-System-User-Instruction-Hierarchy-v1` 🤗: a collection of model-generated evaluation rubrics for assessing whether language-model responses follow the instruction hierarchy between system and user messages.

## 2. Chat Template Design

We extend the chat template of Gemma4-E4B-it to support alignment between tool-use and spoken responses in conversational applications. When an assistant message contains text and tool calls, the template puts the text before the calls and their results. This order lets the assistant acknowledge a request or explain an action before it calls a tool. We keep the existing *enable_thinking* and *preserve_thinking* controls.

### 2.1. Conversational Mode

Nemotron 3.5 Bolt E4B 4GB also supports an optional *conversational_mode* flag. When this flag is *true*, the template adds a system instruction to keep responses concise and without excessive text formatting, making it more suitable for spoken conversations. The instruction is added after an existing system prompt, or if there is no system prompt, this instruction becomes the system prompt. When the *conversational_mode* flag is absent or *false*, no instruction is added to the system prompt.

## 3. Post-Training

Nemotron 3.5 Bolt E4B 4GB is further post-trained from Gemma4-E4B-it with multi-environment RLVR and RLHF. The collection of environments targets instruction following, multi-turn tool-calling and chat capabilities. The exact list of environments and link to the publicly released datasets is shown in Table 1.

**Table 1** | Environments and source datasets used in post-training for Nemotron-3.5-Bolt-E4B.

| Environments | Source data |
|---|---|
| Tool-Calling | Conversational Tool Use Pivot |
| Tool-call schema | Function Calling Pivot |
| Instruction Following | Instruction Following, SysBench, InverseIFEval, CFBench, MultiTurnChat, Calendar |
| Instruction Hierarchy | System-User-Instruction-Hierarchy |
| Structured Outputs | Structured Outputs |
| Free-form formatting | Free-Form Formatting |
| GenRM | Updated version of Nemotron-RLHF-GenRM-v1 |

The RLVR and RLHF post-training is done using NeMo-RL and NeMo-Gym. We use a global batch size of 256 with 8 rollouts per prompt, a maximum generation length of 4096 tokens for reasoning samples, and 1024 for non-reasoning samples.

### 3.1. Non-Reasoning Focus

To emphasize non-reasoning accuracy in post-training, we use each sample in our RLVR/RLHF training blend with both reasoning (*enable_thinking=true*) and non-reasoning (*enable_thinking=false*) inference modes. We independently perform reward profiling for the datasets in both reasoning modes, and maintain a deliberate $2:1$ non-reasoning to reasoning samples ratio in each training batch.

### 3.2. Verbosity Control

We observe that as RL post-training progressed, there was a considerable increase in verbosity for both reasoning and final answer tokens. We were able to control verbosity by increasing group-relative length penalties for reasoning and answer length in the GenRM environment, as well as disabling overlong filtering for all other environments. While these interventions were effective in controlling

verbosity, we found that tightening and loosening the length penalties, along with enabling and disabling overlong filtering in a cycle as the training progressed helped improve accuracy while still maintaining lower verbosity.

### 3.3. SFT + RL vs. RL without SFT

In an attempt to improve tool-calling behaviors of Gemma4-E4B-it, we experimented with SFT using datasets from Nemotron-Post-Training-Datasets. We observed that we were able to improve multi-turn tool-calling with SFT and then further RLVR to a higher ceiling and with less RL training steps than with just RL-based post-training. However, SFT on data from diverse teachers resulted in a regression on conversational qualities and verbosity of Gemma4-E4B-it, for which it was used as a starting point. Hence, Nemotron-3.5-Bolt-E4B did not include an SFT stage and was post-trained purely with RL.

## 4. Post-Trained Model Evaluations

### 4.1. Benchmarks and Evaluation Setup

We evaluate Nemotron 3.5 Bolt E4B 4GB on a comprehensive suite of benchmarks covering conversational proficiency, multi-turn tool use, instruction following/control, context robustness, state/action grounding, and response conciseness. All evaluations for Nemotron 3.5 Bolt E4B 4GB and baselines are conducted with NeMo-Gym. For each benchmark, we estimate the number of repeats needed to limit score variation to 1–2 percentage points and evaluate each model using that number of repeats. The following sections describe our benchmark selection and evaluation methodology across these capabilities, and Table 2 presents the evaluation settings.

**Conversational Proficiency.** We evaluate conversational proficiency using **EQ-Bench v3** (Paech, 2025, 2024) for emotional understanding, empathy, and social reasoning, with an ensemble of Claude-Sonnet-5.5, Gemini-4-Argon, and GPT-6.1-Sol judges; **MultiChallenge** (Sirdeshmukh et al., 2025) for instruction retention, context tracking, and response consistency across multiple turns, judged by GPT-4o; **RoleBench** (Wang et al., 2024) for role-playing fidelity and character-consistent responses, judged by GPT-5.5; and **RoleMRC** (Lu et al., 2025b) for role-conditioned reading comprehension and instruction following, judged by GPT-4.1. For RoleBench, we use a custom pairwise evaluation against a fixed Gemma4-E4B-it reference, with each response pair judged in both answer orders to reduce position bias. For RoleMRC, we report the average judge score excluding the multi-turn subset.

**Response Conciseness.** As part of conversational proficiency, we also evaluate response conciseness using an internal benchmark for brevity which covers personal assistant, customer service, and passage-based question-answering tasks. The benchmark compares response length, excluding reasoning tokens, with human-written reference responses, awarding high scores to responses that are no longer than the references. System prompts request concise answers through general instructions such as “brief” or “suitable for spoken conversation”, making conciseness an instruction-following task. Answer quality is assessed separately by an LLM judge to check that shorter responses preserve quality. We report the average of answer quality and conciseness scores. The benchmark will be publicly released in the future. We also use the results from this benchmark to determine the system prompt to be used for *conversational_mode* described in the chat template design section.

**Table 2** | Evaluation configuration for the benchmark suite. Reasoning mode indicates whether reasoning is enabled (On) or disabled (Off). Judge and User identify the evaluator and simulated user models, respectively.

| Benchmark | Metric | Reasoning Mode | Judge/User Model | Repeats |
|---|---|---|---|---|
| **Conversational Proficiency** | | | | |
| MultiChallenge | Overall Score | On | Judge: GPT-4o | 16 |
| RoleBench Custom | Win % vs. Gemma4-E4B-it | Off | Judge: GPT-5.5 | 4 |
| RoleMRC | Judge Average w/o Multi-Turn | On | Judge: GPT-4.1 | 8 |
| EQ-Bench v3 | Rubric Score | Off | Judge: Ensemble[a] | 4 |
| Brevity Benchmark | Implicit - Average of Answer Quality & Conciseness | On | User: Qwen3-235B-A22B-Instruct-2507 | 4 |
| **Multi-Turn Tool Use** | | | | |
| BFCL-v3 | Average Accuracy | On | N/A | 4 |
| When2Call Multi-Turn | Overall Accuracy | On | Judge: GPT-5.2 | 4 |
| ToolSandbox | Average Accuracy | On | User: GPT-4o | 4 |
| Tau2-Bench | Average Accuracy | On | User: GPT-5.2 | 4 |
| **Instruction Following & Control** | | | | |
| IHEval | Conflict Accuracy | On | N/A | 2 |
| AgentIF | Instruction Success Rate (ISR) | On | Judge: GPT-5.1 | 4 |
| StructEval-T | Average Accuracy | On | N/A | 4 |
| IFBench | Prompt Loose Accuracy | On | N/A | 4 |
| InverseIFEval | Overall Accuracy | On | Judge: GPT-5.1 | 8 |
| **Context Robustness** | | | | |
| RULER (64K) | Overall Accuracy | Off | N/A | 4 |
| LongMemEval | Overall Accuracy | On | Judge: GPT-4o | 8 |
| LongBench v2 (48K) | Overall Accuracy | On | N/A | 4 |
| SysBench | Instruction Success Rate (ISR) | On | Judge: GPT-5.1 | 4 |
| **State & Action Grounding** | | | | |
| GaRAGe | Relevance-Aware Factuality (RAF) | On | Judge: GPT-5.1 | 4 |
| RAGTruth | F1 | On | N/A | 4 |
| RelyToolBench w/ Native Function-Calling | Real Pass Rate (RePR) | On | Judge: GPT-5.1 | 8 |

[a]Ensemble: Claude-Sonnet-5.5, Gemini-4-Argon, and GPT-6.1-Sol.

**Multi-Turn Tool Use.** We evaluate multi-turn tool use using **BFCL-v3** (Patil et al., 2025; Mao et al., 2024) for fundamental tool calling; $\tau^2$**-Bench** (Barres et al., 2025) for task completion through tool use and dialogue, with GPT-5.2 as the user model; and **ToolSandbox** (Lu et al., 2025a) for stateful tool use in conversational settings, with GPT-4o as the user model. We also evaluate **When2Call-MT**, our multi-turn extension of When2Call (Ross et al., 2025), which tests whether an assistant selects the correct response policy as a conversation develops, including when to call a tool, ask for clarification, refuse due to lack of tools, or respond directly. When2Call-MT uses GPT-5.2 as the judge and will be publicly released in the future.

**Instruction Following & Control.** We evaluate instruction following and control using **AgentIF** (Qi et al., 2025) for following complex instructions in agentic settings, judged by GPT-5.1; **IFBench** (Pyatkin et al., 2025) for generalization to verifiable instruction-following constraints; **IHEval** (Zhang et al., 2025b) for resolving conflicting instructions according to the instruction hierarchy; InverseIFEval (Zhang et al., 2025a) for following counter-intuitive instructions, judged by GPT-5.1; and StructEval-T (Yang et al., 2026) for generating outputs that satisfy specified structural and formatting requirements. For **InverseIFEval**, we evaluate using the methodology as originally

**Table 3** | Evaluation results for BF16 checkpoints. Evaluation metrics and reasoning settings are listed in Table 2. Bold marks the best score in each row.

| Benchmark | Nemotron-3.5 Bolt-E4B | Gemma4 E4B-it | Qwen3.5 4B | Nemotron-3 Nano-4B |
|---|---|---|---|---|
| **Conversational Proficiency** | | | | |
| MultiChallenge | **53.10** | 42.08 | 51.76 | 35.30 |
| RoleBench Custom | **67.05** | 49.95 | 53.66 | 12.02 |
| RoleMRC | **80.63** | 74.62 | 65.81 | 47.00 |
| EQ-Bench v3 | 39.40 | **41.17** | 32.92 | 19.31 |
| Brevity Benchmark | **82.52** | 79.99 | 75.35 | 73.60 |
| **Multi-Turn Tool Use** | | | | |
| BFCL-v3 | 68.51 | 66.39 | **70.95** | 61.10 |
| When2Call Multi-Turn | **63.36** | 62.90 | 37.85 | 27.70 |
| ToolSandbox | 68.77 | 65.43 | **69.54** | 57.88 |
| Tau2-Bench | 55.10 | 38.50 | **75.06** | 29.20 |
| **Instruction Following & Control** | | | | |
| IHEval | **84.79** | 67.04 | 60.57 | 51.31 |
| AgentIF | **25.81** | 21.89 | 20.44 | 20.08 |
| StructEval-T | **80.40** | 79.27 | 76.01 | 63.50 |
| IFBench | **63.79** | 42.84 | 58.91 | 60.10 |
| InverseIFEval | **46.84** | 42.00 | 30.98 | 35.25 |
| **Context Robustness** | | | | |
| RULER (64K) | 86.99 | 86.55 | **95.15** | 91.51 |
| LongMemEval | **72.10** | 70.95 | 66.85 | 25.20 |
| LongBench v2 (48K) | 50.50 | 49.34 | **59.93** | 31.40 |
| SysBench | **67.88** | 65.85 | 67.34 | 60.88 |
| **State & Action Grounding** | | | | |
| GaRAGe | **42.61** | 34.73 | 24.66 | 25.18 |
| RAGTruth | **77.59** | 76.97 | 75.60 | 71.40 |
| RelyToolBench | **63.16** | 61.60 | 43.35 | 41.20 |

described, but use only English data (506 of 1012 total samples) and translate the judge prompt to English.

**Context Robustness.** We evaluate context robustness using **LongBench v2** (Bai et al., 2025) for understanding and reasoning over long contexts, restricted to the subset with fewer than 48K tokens due to the model's context limit; **LongMemEval** (Wu et al., 2025) for retaining and retrieving information across conversation histories, judged by GPT-4o; **RULER** (Hsieh et al., 2024) for retrieval and reasoning at a 64K-token context length; and **SysBench** (Qin et al., 2024) for following system instructions across varied user inputs, judged by GPT-5.1. For SysBench, we translate the benchmark data into English and use an adapted English judge prompt.

**State & Action Grounding.** We evaluate state and action grounding using **GaRAGe** (Sorodoc et al., 2025) for retrieved evidence-grounded answers, judged by GPT-5.1; **RAGTruth** (Niu et al., 2024) for detecting unsupported claims in retrieval-augmented responses, measured by span-level F1; and **RelyToolBench** (Xu et al., 2025) for reliable tool use under tool failures and misleading tool responses. We evaluate RelyToolBench with native function calling instead of the ReAct format

and use GPT-5.1 as the judge. For GaRAGe, we follow the evaluation methodology as originally described and use our own implementation due to a lack of publicly available code.

### 4.2. Results

We show all benchmark evaluations in Table 3. Our model outperforms the competitive models on 15 out of 21 benchmarks: Conversational Proficiency (4/5), Multi-Turn Tool Use (1/4), Instruction Following & Control (5/5), Context Robustness (2/4), and State & Action Grounding (3/3).

**Non-Reasoning Evaluations.** We also conduct all benchmark evaluations in non-reasoning mode across models to validate the effect of focused post-training. The results are presented in Table 4. The results show a similar trend as reasoning, and Nemotron 3.5 Bolt E4B 4GB shows the highest accuracy on 14 out of 21 benchmarks.

**Table 4** | Non-reasoning evaluation results for BF16 checkpoints. Bold marks the best score in each row, including ties.

| Benchmark | Nemotron-3.5 Bolt-E4B | Gemma4 E4B-it | Qwen3.5 4B | Nemotron-3 Nano-4B |
|---|---|---|---|---|
| **Conversational Proficiency** | | | | |
| MultiChallenge | **48.81** | 40.72 | 38.60 | 23.27 |
| RoleBench Custom | **67.05** | 49.95 | 53.66 | 12.02 |
| RoleMRC | **77.18** | 72.46 | 58.91 | 46.11 |
| EQ-Bench v3 | 39.40 | **41.17** | 32.92 | 19.31 |
| Brevity Benchmark | **78.71** | 75.39 | 75.21 | 72.43 |
| **Multi-Turn Tool Use** | | | | |
| BFCL-v3 | 64.68 | 58.08 | **66.04** | 45.23 |
| When2Call Multi-Turn | 51.25 | **54.80** | 46.02 | 38.67 |
| ToolSandbox | **61.54** | 35.72 | 52.19 | 55.32 |
| Tau2-Bench | 34.34 | 25.99 | **73.46** | 26.78 |
| **Instruction Following & Control** | | | | |
| IHEval | **66.90** | 52.95 | 34.25 | 23.78 |
| AgentIF | **23.76** | 22.98 | 21.22 | 16.69 |
| StructEval-T | **78.76** | 76.04 | 75.84 | 10.38 |
| IFBench | **53.33** | 38.83 | 34.33 | 41.25 |
| InverseIFEval | **30.24** | 22.13 | 22.94 | 18.97 |
| **Context Robustness** | | | | |
| RULER (64K) | 86.99 | 86.55 | **95.15** | 91.51 |
| LongMemEval | 60.20 | **65.79** | 65.00 | 31.20 |
| LongBench v2 (48K) | 41.06 | 41.89 | **51.70** | 31.13 |
| SysBench | **65.12** | 64.09 | **65.12** | 55.28 |
| **State & Action Grounding** | | | | |
| GaRAGe | **46.04** | 37.33 | 22.19 | 23.61 |
| RAGTruth | **67.38** | 58.20 | 65.08 | 30.87 |
| RelyToolBench | **61.25** | 58.87 | 19.43 | 36.18 |

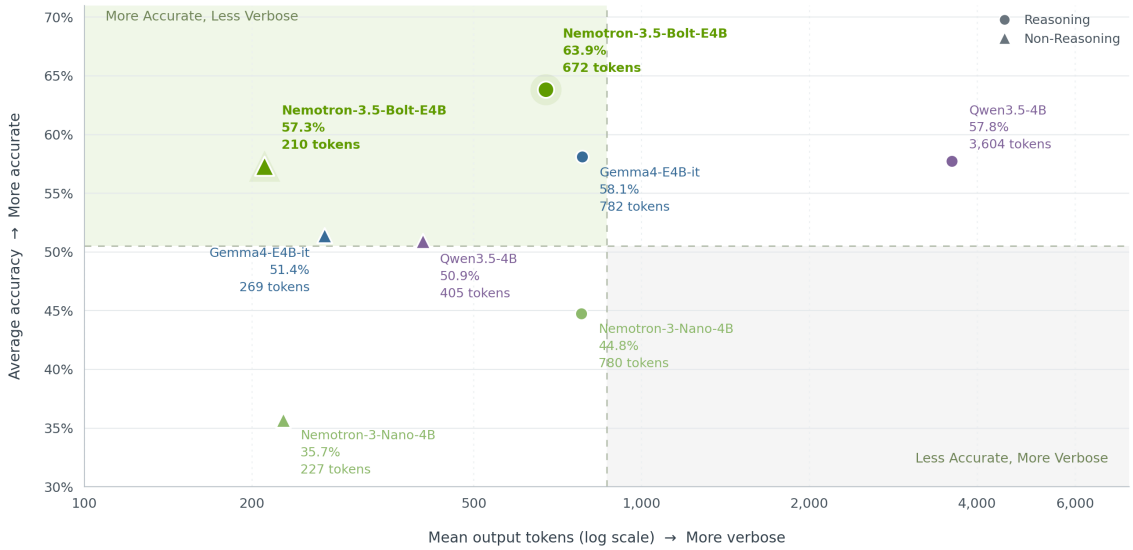



**Figure 2** | Average benchmark score versus mean output token count across 21 benchmarks in reasoning and non-reasoning settings. Higher scores and fewer output tokens indicate higher quality and greater response conciseness, respectively. The horizontal axis uses a logarithmic scale.




Reachy Mini running on Jetson Orin Nano 8GB. Cosmos 3 Edge (NVIDIA, 2026) supplies real-time visual understanding of the world to Bolt E4B, which calls scripts from the robot's animation libraries and powers its spoken responses.




Bolt E4B controlling a blacksmith trader NPC. The NPC agent verifies correct payment for an item, then autonomously calls tools to check crafting recipes, collect ingredients, craft the item, and transfer it to the user.

**Figure 3** | Examples of live-testing conducted with Nemotron 3.5 Bolt E4B 4GB. View the GIFs in our blog.

### 4.3. Live Testing

Academic benchmarks only partially capture the skills required of conversational agents and typically assess individual capabilities rather than how they work together over a complete interaction. We therefore complement benchmark evaluations with live testing in interactive scenarios spanning virtual interactive characters, on-device voice assistants, and chat interactions. These tests are qualitative and examine whether the model can sustain coherent conversations, respond to changing user intent, environment feedback, handle ambiguity, and recover from mistakes or tool failures. GIFs in Figure 3 shows captures from some live-testing instances for Nemotron 3.5 Bolt E4B 4GB.

## 5. Quantization

We used llama.cpp to convert the Nemotron 3.5 Bolt E4B 4GB BF16 checkpoint to Q4_K_M with importance-matrix (imatrix) calibration. The calibration was conducted with approximately 1M tokens of RL rollouts from the training step that was selected as the BF16 checkpoint. On average accuracy, Q4_K_M with imatrix calibration not only outperforms vanilla Q4_K_M, but also IQ4_XS and IQ4_NL with imatrix calibration. Evaluation results for quantized models are shown in Table 5.

**Table 5** | Evaluation results for four-bit quantized checkpoints. The left block compares Q4_K_M checkpoints, with importance-matrix calibration (imatrix) used for Bolt. We use the Unsloth Q4_K_M GGUF checkpoints for Gemma4-E4B-it and Qwen3.5-4B.[1]. The right block shows additional Bolt variants: Q4_K_M without imatrix, IQ4_XS, and IQ4_NL. All benchmarks use reasoning mode except EQ-Bench v3, RoleBench Custom, and RULER (64K), which use non-reasoning mode.

| | **Q4_K_M model comparison** | | | | **Nemotron-3.5-Bolt-E4B Q4 variants** | | |
|---|---|---|---|---|---|---|---|
| **Benchmark** | **Nemotron-3.5 Bolt-E4B imatrix** | **Gemma4 E4B-it** | **Qwen3.5 4B** | **Nemotron-3 Nano-4B** | **Q4_K_M no imatrix** | **IQ4_XS** | **IQ4_NL** |
| **Conversational Proficiency** | | | | | | | |
| MultiChallenge | **55.60** | 39.36 | 52.23 | 34.53 | 51.12 | 51.91 | 52.12 |
| RoleBench Custom | **71.21** | 60.51 | 52.35 | 10.86 | 65.21 | 69.07 | 69.86 |
| RoleMRC | **79.12** | 75.24 | 62.61 | 45.84 | 80.36 | 78.79 | 75.65 |
| EQ-Bench v3 | 38.15 | **39.77** | 32.07 | 19.45 | 37.42 | 38.19 | 38.28 |
| Brevity Benchmark | **82.84** | 78.49 | 75.17 | 70.98 | 81.29 | 81.30 | 82.04 |
| **Multi-Turn Tool Use** | | | | | | | |
| BFCL-v3 | 67.31 | 65.03 | **70.56** | 59.07 | 64.32 | 65.95 | 67.68 |
| When2Call-MT | 60.50 | **61.12** | 42.88 | 35.96 | 58.06 | 60.67 | 60.92 |
| ToolSandbox | 74.13 | 68.31 | **74.51** | 59.78 | 70.23 | 70.44 | 69.88 |
| $\tau^2$-Bench | 52.32 | 38.72 | **61.74** | 14.89 | 46.95 | 42.05 | 48.74 |
| **Instruction Following & Control** | | | | | | | |
| IHEval | **82.40** | 66.61 | 62.87 | 53.75 | 83.50 | 83.19 | 83.65 |
| AgentIF | **22.95** | 20.05 | 22.24 | 21.25 | 24.33 | 25.32 | 25.46 |
| StructEval-T | **79.99** | 77.28 | 76.38 | 16.54 | 78.44 | 81.40 | 81.25 |
| IFBench | 64.00 | 42.67 | **66.67** | 58.58 | 59.79 | 59.00 | 63.33 |
| InverseIFEval | **45.26** | 39.75 | 22.04 | 36.02 | 42.69 | 44.47 | 43.08 |
| **Context Robustness** | | | | | | | |
| RULER (64K) | 85.76 | 86.14 | **95.37** | 91.31 | 82.92 | 85.84 | 86.88 |
| LongMemEval | 69.92 | 67.95 | **76.60** | 27.15 | 64.30 | 70.20 | 72.80 |
| LongBench v2 (48K) | 48.67 | 45.70 | **58.28** | 37.42 | 44.37 | 47.33 | 47.67 |
| SysBench | 67.22 | 65.62 | **67.90** | 60.45 | 66.55 | 68.04 | 66.68 |
| **State & Action Grounding** | | | | | | | |
| GaRAGe | **40.73** | 30.91 | 31.78 | 24.80 | 38.85 | 38.74 | 39.95 |
| RAGTruth | **77.88** | 77.17 | 74.08 | 71.75 | 76.92 | 77.06 | 78.24 |
| RelyToolBench | **69.29** | 67.15 | 38.34 | 38.97 | 63.50 | 64.11 | 66.31 |

## 6. Multilingual Evaluations

Since Nemotron 3.5 Bolt E4B 4GB was post-trained using English only, we tested if the model retains the original model's multilingual capabilities. To test this, we evaluated Nemotron 3.5 Bolt E4B 4GB on WMT24++ for 36 languages and MMLU-ProX on 12 languages.

[1]Gemma and Qwen 4-bit Checkpoints: unsloth/gemma-4-E4B-it-GGUF and unsloth/Qwen3.5-4B-GGUF.

The results are shown in Tables 6 and 7 respectively. While Nemotron 3.5 Bolt E4B 4GB retains the original model's machine translation quality, it shows slight but consistent drops on MMLU-ProX. These results suggest that translation quality is preserved more consistently than multilingual knowledge and reasoning performance.

**Table 6** | WMT24++ translation results for Nemotron-3.5-Bolt-E4B BF16, measured by xCOMET and chrF (both ↑). Both models use non-reasoning mode; reasoning mode shows similar trends. Parenthesized deltas are relative to Gemma4-E4B-it for the same language, metric, and direction. Green indicates improvement and red indicates degradation.

| | English → X | | X → English | |
|---|---|---|---|---|
| **Language** | **xCOMET** | **chrF** | **xCOMET** | **chrF** |
| ar_EG | 67.7 (-1.00) | 40.5 (+0.20) | 75.5 (+0.10) | 48.3 (-0.20) |
| ar_SA | 73.9 (-0.70) | 40.7 (+0.00) | 76.6 (+0.00) | 49.2 (-0.10) |
| bg_BG | 71.5 (+0.00) | 57.4 (+0.10) | 81.5 (-0.10) | 59.8 (-0.30) |
| cs_CZ | 66.0 (+1.60) | 49.8 (+0.20) | 82.2 (+0.10) | 59.1 (-0.20) |
| da_DK | 85.6 (-0.20) | 63.5 (-0.20) | 89.3 (-0.10) | 64.1 (+0.00) |
| de_DE | 90.9 (+0.20) | 58.6 (+0.00) | 90.1 (+0.00) | 58.8 (-0.10) |
| el_GR | 67.0 (+0.40) | 54.5 (+0.00) | 81.0 (+0.10) | 62.7 (+0.00) |
| es_MX | 84.6 (-0.20) | 66.6 (-0.50) | 89.3 (-0.10) | 66.4 (-0.20) |
| et_EE | 41.4 (+1.30) | 45.9 (+0.30) | 68.3 (-0.40) | 56.3 (+0.10) |
| fi_FI | 75.0 (+0.70) | 57.4 (+0.00) | 81.6 (+0.10) | 57.5 (-0.20) |
| fr_CA | 79.4 (-0.70) | 67.5 (+0.00) | 88.0 (-0.10) | 64.8 (-0.20) |
| fr_FR | 79.5 (-0.30) | 63.2 (+0.00) | 87.2 (-0.20) | 60.5 (-0.10) |
| hi_IN | 67.2 (+0.10) | 39.9 (+0.00) | 83.6 (+0.10) | 49.1 (-0.20) |
| hr_HR | 74.2 (+0.50) | 53.6 (+0.30) | 83.1 (+0.00) | 60.0 (-0.10) |
| hu_HU | 61.4 (+1.10) | 45.7 (+0.10) | 78.5 (+0.20) | 54.4 (-0.20) |
| id_ID | 83.8 (+0.00) | 63.2 (+0.20) | 89.0 (-0.10) | 57.9 (+0.10) |
| it_IT | 82.1 (-0.80) | 64.5 (-0.30) | 88.6 (+0.30) | 63.3 (-0.20) |
| ja_JP | 82.1 (+0.10) | 36.2 (+0.80) | 84.5 (+0.70) | 50.5 (-0.40) |
| ko_KR | 84.3 (+0.30) | 35.6 (-0.10) | 84.5 (+0.30) | 54.1 (-0.10) |
| lt_LT | 44.1 (+0.80) | 41.0 (-0.20) | 68.4 (-0.30) | 47.6 (-0.10) |
| lv_LV | 37.8 (+1.60) | 43.5 (-0.50) | 66.6 (-0.20) | 52.4 (-0.20) |
| nl_NL | 85.8 (+0.20) | 58.9 (+0.00) | 89.3 (+0.10) | 60.3 (+0.00) |
| no_NO | 87.6 (+0.20) | 67.0 (+0.10) | 89.0 (-0.10) | 66.7 (-0.30) |
| pl_PL | 78.3 (-0.10) | 48.5 (-0.20) | 84.9 (-0.40) | 56.9 (-0.20) |
| pt_BR | 85.0 (-0.10) | 64.5 (-0.20) | 89.0 (+0.00) | 62.5 (-0.10) |
| pt_PT | 83.8 (-0.20) | 61.1 (+0.00) | 88.2 (-0.40) | 60.9 (+0.00) |
| ro_RO | 73.3 (+0.00) | 59.7 (+0.00) | 82.2 (-0.10) | 63.6 (-0.10) |
| ru_RU | 78.7 (+0.20) | 51.5 (+0.10) | 83.4 (+0.00) | 53.0 (-0.10) |
| sk_SK | 59.9 (+1.40) | 45.4 (+0.40) | 80.8 (+0.20) | 56.0 (-0.10) |
| sl_SI | 58.6 (+1.70) | 48.2 (+0.60) | 78.6 (+0.10) | 58.0 (-0.20) |
| sv_SE | 86.7 (+0.60) | 63.2 (-0.10) | 89.4 (+0.00) | 65.1 (+0.00) |
| th_TH | 81.4 (-0.60) | 51.2 (-0.40) | 86.3 (+0.10) | 54.1 (-0.20) |
| tr_TR | 74.8 (+0.00) | 58.5 (+0.10) | 81.8 (-0.40) | 57.9 (+0.10) |
| uk_UA | 76.4 (-0.30) | 54.2 (+0.00) | 80.4 (+0.00) | 58.3 (-0.20) |
| vi_VN | 80.7 (-0.90) | 55.2 (-0.30) | 87.3 (+0.40) | 55.6 (-0.30) |
| zh_CN | 80.8 (-0.50) | 37.3 (+0.10) | 87.7 (+0.00) | 52.3 (-0.40) |
| zh_TW | 80.3 (-0.80) | 31.0 (-0.10) | 89.2 (-0.10) | 56.4 (-0.20) |
| Average | 74.4 (+0.20) | 52.5 (+0.00) | 83.4 (+0.00) | 57.7 (-0.10) |

**Table 7** | MMLU-ProX results (↑) for Nemotron-3.5-Bolt-E4B BF16 with reasoning enabled and disabled. Parenthesized deltas are relative to Gemma4-E4B-it for the same language and mode. Green indicates improvement and red indicates degradation.

| **Language** | **Reasoning** | **Non-reasoning** |
|---|---|---|
| ar | 64.04 (-1.97) | 60.90 (-1.69) |
| de | 66.15 (-1.59) | 64.33 (-1.05) |
| en | 70.41 (-0.57) | 69.34 (-0.69) |
| es | 67.24 (-1.33) | 65.11 (-0.76) |
| fr | 66.99 (-1.30) | 65.01 (-1.08) |
| hi | 64.72 (-1.73) | 61.49 (-1.52) |
| it | 66.97 (-1.67) | 65.26 (-1.00) |
| ja | 64.49 (-1.87) | 62.17 (-1.64) |
| ko | 64.51 (-1.77) | 61.90 (-1.57) |
| pt | 66.59 (-1.49) | 64.70 (-1.56) |
| vi | 66.56 (-1.35) | 64.59 (-0.90) |
| zh | 65.17 (-1.61) | 63.32 (-1.12) |
| Average | 66.15 (-1.52) | 64.01 (-1.22) |

## 7. Multimodal Evaluations

Although the model was trained text-only, we found that the model retains the original model's multimodal capabilities.

**Audio understanding and spoken interaction.** We evaluate Nemotron 3.5 Bolt E4B 4GB and Gemma4-E4B-it on two popular speech benchmarks: VoiceBench (Chen et al., 2024) and AudioMC (Gosai et al., 2025). VoiceBench is a single-turn speech benchmark comprising subsets of benchmarks covering multiple-choice question answering (OpenBookQA, MMSU), open-ended question answering (CommonEval, SD-QA, AlpacaEval, WildVoice), reasoning (BBH), instruction following (IFEval), and safety (AdvBench). AudioMC is a multi-turn speech benchmark built on the text benchmark MultiChallenge, with an additional evaluation of voice editing that assesses a model's ability to track information as users provide updates during a conversation.

**Vision capabilities.** In contrast, Nemotron 3.5 Bolt E4B 4GB exhibits slight but consistent degradation in vision capabilities. We attribute this degradation to misalignment between the vision encoder and the language model introduced by text-only training. However, the declines are not substantial on most vision benchmarks, suggesting that Nemotron 3.5 Bolt E4B 4GB still retains visual understanding capabilities of the original model. We attempted additional training of the vision encoder alone but observed no meaningful improvements. Restoring and further improving vision capabilities remain part of future work.

**Table 8** | Audio and vision benchmark results in reasoning and non-reasoning modes. Gemma4-E4B-it scores are measured in our evaluation setup[a]. Bold indicates the best score among models within each mode. Arrows indicate whether higher (↑) or lower (↓) scores are better; CoVoST excludes Chinese. Parenthesized deltas are Bolt minus the measured Gemma4-E4B-it score in the same mode; Green indicates improved performance and red indicates degraded performance, accounting for whether higher or lower scores are better.

| | Reasoning | | | Non-reasoning | | |
|---|---|---|---|---|---|---|
| **Benchmark** | **Nemotron-3.5 Bolt-E4B** | **Gemma4 E4B-it** | **Qwen3.5 4B** | **Nemotron-3.5 Bolt-E4B** | **Gemma4 E4B-it** | **Qwen3.5 4B** |
| **Audio** | | | | | | |
| Audio MultiChallenge ↑ | **25.22** (+6.19) | 19.03 | — | **19.25** (+2.21) | 17.04 | — |
| Voicebench ↑ | **80.30** (+2.10) | 78.20 | — | **72.60** (+0.90) | 71.70 | — |
| CoVoST (BLEU) ↑ | 36.14 (-0.81) | **36.95** | — | **37.61** (+0.59) | 37.02 | — |
| FLEUR (WER/CER) ↓ | 10.16 (+0.38) | **9.78** | — | 10.43 (+0.26) | **10.17** | — |
| OpenASR (WER) ↓ | — | — | — | **9.28** (-0.25) | 9.53 | — |
| **Vision** | | | | | | |
| OmniDocBench 1.5 ↓ | 0.42 (+0.21) | 0.21 | **0.17** | 0.47 (+0.23) | 0.24 | **0.20** |
| MMMU-Pro Vision ↑ | 42.02 (-5.61) | 47.63 | **66.13** | 39.08 (-7.10) | 46.18 | **56.18** |
| MathVision ↑ | 46.25 (-2.04) | **48.29** | — | 43.72 (-0.36) | 44.08 | **53.45** |
| MedXpertQA Vision ↑ | 26.85 (-1.35) | 28.20 | **41.85** | 23.10 (-0.65) | 23.75 | **35.35** |

[a]We report our own measurements for Gemma4-E4B-it to ensure consistent evaluation across models; we observe discrepancies with the results reported in Google's model card (`huggingface.co/google/gemma-4-E4B-it`).

The results in Table 8 show that Nemotron 3.5 Bolt E4B 4GB improves upon Gemma4-E4B-it on both VoiceBench and AudioMC, with reasoning enabled and disabled. Both models support speech inputs of up to 30 seconds. In AudioMC, 23.36% of user turns exceed this limit, and 53.1% of conversations contain at least one such turn. Our reported AudioMC results are therefore subject to this per-turn input limit and do not reflect evaluation on the complete, unrestricted conversations. For the SD-QA subset of VoiceBench, we report the average of the GPT score and the PANDA score. We compute PANDA using `qa_metrics` v0.2.42, which showed better agreement with GPT-based judgments in our evaluation.

## 8. Speculative Decoding

Nemotron 3.5 Bolt E4B 4GB is released with a DFlash (Chen et al., 2026) parallel drafter. The drafter proposes all draft tokens in one forward pass, which the target verifies in parallel; verification preserves the output distribution of the target (Leviathan et al., 2023; Chen et al., 2023). We measure drafting quality by the acceptance length (AL), the number of tokens generated per verification step, including the one contributed by the target (at most 8 for a draft length of 7).

**Algorithm selection.** Recent extensions of DFlash add a lightweight correction head that improves the mutual consistency of the drafted tokens. We compare DFlash with LiLiCorr (Rusanovsky et al., 2026), a state-of-the-art extension that jointly scores candidate tokens across all drafted positions. For Nemotron 3.5 Bolt E4B 4GB, LiLiCorr increases AL by about 5%. However, because each decoding step of such a small model is short, even a lightweight head adds non-negligible overhead, and the two drafters achieve comparable per-user decode speed at concurrency 1. We therefore adopt DFlash, which minimizes the drafting overhead.

**Drafter and training.** Our DFlash drafter is a five-layer Transformer with 360M parameters (excluding the token embeddings and LM head, which are reused from the target) and drafts seven tokens per step. Attention is bidirectional among the drafted positions and restricted to a sliding window of the most recent 1,024 positions. The drafter is conditioned on hidden states from five intermediate layers of the target. We train it in NVIDIA Model Optimizer (ModelOpt; NVIDIA Corporation, 2024–2026) by distilling the next-token distribution of the target at every drafted position. The training data comprise 3.79M conversations whose prompts are drawn from Nemotron-Post-Training-Dataset-v2 and whose responses are regenerated by the target at temperature 1.0, in both reasoning-enabled and reasoning-disabled modes.

**Quantization.** To speed up drafting on GPUs with lower memory bandwidth, typical of local deployments, we apply post-training NVFP4 quantization to the drafter using GPTQ (Frantar et al., 2023) in ModelOpt. This reduces the weights read per drafting step, including the LM head, by about 3.6×, with at most a 2% reduction in AL. By default, activations remain in BF16 (W4A16); an alternative variant of the checkpoint also quantizes them to NVFP4 (W4A4), which yields additional speedup at higher concurrency on GPUs with native NVFP4 support. Table 9 reports the speedups of both variants on DGX Spark.

**Table 9** | Weights read per drafting step (including the LM head), AL, and per-user decode speedup of the NVFP4 drafters relative to the BF16 drafter on DGX Spark. AL is measured at concurrency 1.

| | Weights | | Speedup at concurrency | | |
|---|---|---|---|---|---|
| Drafter | (GB) | AL | 1 | 16 | 32 |
| BF16 | 2.06 | 2.99 | 1.00× | 1.00× | 1.00× |
| NVFP4 W4A16 (default) | 0.58 | 2.99 | 1.13× | 1.05× | 1.01× |
| NVFP4 W4A4 | 0.58 | 2.93 | 1.11× | 1.09× | 1.06× |

**Evaluation.** Table 10 reports AL of the released drafter on MT-Bench (Zheng et al., 2023), the qualitative split of SPEED-Bench (Abramovich et al., 2026), and SPEED-Bench prompts with 32K-token inputs.

**Table 10** | AL of the released drafter (NVFP4, W4A16) with reasoning enabled, on one GB300 GPU at concurrency 16 with up to 4,096 output tokens. At $temp = 1$, the target samples while the drafter drafts greedily.

| Benchmark | $temp = 0$ | $temp = 1$ |
|---|---|---|
| MT-Bench | 4.01 | 3.65 |
| SPEED-Bench, qualitative split | 3.88 | 3.50 |
| SPEED-Bench, 32K-token inputs | 3.42 | 3.00 |

## 9. Inference Serving

### 9.1. NVIDIA NIM

We package the BF16 checkpoint of Nemotron 3.5 Bolt E4B 4GB as a downloadable NVIDIA NIM microservice built on vLLM 0.30.0 (Kwon et al., 2023). The container image is published for `amd64` and `arm64` and ships a single BF16 profile on one GPU with the following defaults:

- **Speculative decoding.** The DFlash (Chen et al., 2026) drafter (Section 8) is packaged next to the main weights and enabled by default with a draft length of 7, so no extra download or configuration is needed.
- **Reasoning and tool calling.** Reasoning is on by default. The `gemma4` reasoning parser returns the reasoning separately from the answer, and the `gemma4` tool-call parser enables tool calling.
- **Context and modalities.** The context length is 65,536 tokens. Text, image, and audio inputs are accepted.
- **APIs.** OpenAI-compatible Chat Completions, Completions, and Responses endpoints, and the Anthropic-compatible Messages endpoint.

Three request-level controls are exposed through the chat template and the serving engine: `enable_thinking` turns reasoning on or off per request, `thinking_token_budget` caps the number of reasoning tokens, and `conversational_mode` requests brief, spoken-style replies for voice assistants. We validated the NIM on DGX Spark (GB10), RTX PRO 6000 and RTX PRO 4500 Blackwell Server Edition, and H100 80GB.

**Speculative decoding speedup.** We measure the effect of the packaged DFlash drafter by comparing two NIM builds that differ only in whether the model profile includes the drafter; all other serving settings are equal. We use SPEED-Bench (Abramovich et al., 2026) prompts from 11 categories (16 prompts per category on RTX PRO 6000 and 8 on DGX Spark), streaming Chat Completions requests with greedy decoding and `max_tokens` of 1,024, at concurrency 1 and 8. Per-user decode speed is 1000/ITL, where ITL is the mean inter-token latency in milliseconds. Acceptance length (AL) is reported by vLLM's speculative decoding metrics and includes the token produced by the target model. Table 11 reports means over the 11 categories from a single run per configuration.

With reasoning on, DFlash reaches an AL of 3.8 and speeds up per-user decoding by 2.7× on RTX PRO 6000 and 3.6× on DGX Spark at concurrency 1. The speedup holds at concurrency 8 (2.7× and 2.9×). With reasoning off, the AL drops to 3.4–3.6 and the speedup to 2.3–3.1×. Because speculative decoding shortens decoding rather than prefill, the gain in end-to-end request latency follows the decode speedup, while time to first token changes by less than 100 ms in all settings.

**Table 11** | DFlash speedup in the Nemotron 3.5 Bolt E4B 4GB NIM on SPEED-Bench (BF16, one GPU, greedy decoding, `max_tokens`=1,024). Per-user decode speed is in output tokens per second; request latency is the mean end-to-end latency per request.

| | | | Per-user decode speed (tok/s) | | | Request latency (s) | |
|---|---|---|---|---|---|---|---|
| Reasoning | GPU | Conc. | No drafter | DFlash | Speedup | No drafter | DFlash |
| On | RTX PRO 6000 | 1 | 124.7 | 341.2 | 2.74× | 5.3 | 2.0 |
| On | RTX PRO 6000 | 8 | 115.7 | 313.7 | 2.71× | 5.9 | 2.2 |
| On | DGX Spark | 1 | 20.1 | 71.6 | 3.56× | 31.7 | 9.1 |
| On | DGX Spark | 8 | 23.3 | 66.3 | 2.85× | 28.7 | 10.1 |
| Off | RTX PRO 6000 | 1 | 126.5 | 302.0 | 2.39× | 3.4 | 1.4 |
| Off | RTX PRO 6000 | 8 | 117.8 | 273.3 | 2.32× | 3.7 | 1.6 |
| Off | DGX Spark | 1 | 20.1 | 62.5 | 3.11× | 20.4 | 6.4 |
| Off | DGX Spark | 8 | 23.4 | 57.2 | 2.45× | 18.6 | 7.0 |

**Table 12** | Generation throughput on the coding subset of SPEED-Bench, measured using AIPerf with a batch size of 1. Both target models use Q4_K_M quantization; Bolt uses a BF16 DFlash drafter. Relative throughput is computed separately for each hardware platform against Gemma-4-E4B-it with MTP.

| Model and decoder | RTX 5070 (tokens/s) | RTX Spark N1X (tokens/s) |
|---|---|---|
| Gemma-4-E4B-it + MTP | 225.9 | 118.4 |
| Nemotron 3.5 Bolt E4B 4GB + DFlash | **272.6** | **124.4** |
| Relative throughput | **1.21**× | **1.05**× |

**Serving overhead.** In a matched comparison on DGX Spark against a standalone vLLM 0.30.0 server with the same model and drafter weights, chat template, and scheduler limits, the NIM delivered the same request throughput: per-run NIM-to-vLLM ratios ranged from 0.99 to 1.01 at concurrency 1 and 4.

### 9.2. Quantized Local Inference

We compare Nemotron 3.5 Bolt E4B 4GB with DFlash against Gemma-4-E4B-it with its MTP drafter[2] on the coding subset of SPEED-Bench, using AIPerf with a batch size of 1. Both target models use Q4_K_M quantization, and the DFlash drafter uses BF16 precision. As shown in Table 12, Bolt E4B achieves 272.6 tokens/s on RTX 5070, compared with 225.9 tokens/s for Gemma with MTP, corresponding to 1.21× the throughput. On RTX Spark N1X, Bolt E4B achieves 124.4 tokens/s vs. 118.4 tokens/s, corresponding to 1.05× the throughput.

## 10. Conclusion

We presented Nemotron 3.5 Bolt E4B 4GB, an open small language model designed for fast conversational agents. Built on Gemma-4-E4B-it, we apply reinforcement learning to target multi-turn tool use, instruction following, and output formatting while retaining its conversational strengths. Training in both reasoning and non-reasoning modes, together with verbosity control, supports applications that need concise responses and reliable actions under tight latency constraints.

The release combines these training choices with practical deployment options. With DFlash, Bolt E4B achieves 2.74× the per-user decode speed of inference without speculative decoding in BF16

[2] https://huggingface.co/google/gemma-4-E4B-it-assistant

NIM serving on RTX PRO 6000 with reasoning enabled at concurrency 1. On the coding subset of SPEED-Bench, Q4_K_M Bolt E4B with DFlash achieves 1.21× the generation throughput of Q4_K_M Gemma-4-E4B-it with MTP on RTX 5070 at batch size 1.

## Contributors

We thank the following people for their invaluable contributions to Nemotron 3.5 Bolt E4B 4GB.

Adi Margolin, Aditya Malte, Aishwarya Chhabra, Alex Hodson, Ameya Sunil Mahabaleshwarkar, Ambrish Dantrey, Amy Shen, Andrii Skliar, Annamalai Chockalingam, Anusha Pant, Ashwath Aithal, Ashwin Kumar, Besmira Nushi, Bhavesh Pawar, Bryan Catanzaro, Bryan Yu, Chantal Hwang, Chenhan Yu, Chris Alexiuk, Conner Currie, Cynthia Gao, David Mosallanezhad, Dina Yared, Ehsan Hosseini Asl, Eileen Long, Erick Galinkin, Felix Friedmann, Giedrius Burachas, Gorkem Batmaz, Grzegorz Chlebus, Guillermo Siman, Haifeng Qian, Hao Guo, Hayley Ross, Hrishith Thadicherla, Hui Wang, Huiying Li, Ike Nnoli, Irina Filatova, Iris Liu, Ji-Eun Lee, Jian Leong, Jian Zhang, Jiaqi Zeng, Joey Conway, Jonathan Cohen, Jose Vergara, Julien Veron Vialard, Kari Briski, Katherine Luna, Kedar Pandurang Potdar, Kevin Ning, Khalil Ben Khaled, Krzysztof Pawelec, Laura Dykes, Leili Tavabi, Lily Lee, Linda Liu, Liwei Ding, Maka Dong, Manoj Kilaru, Maria Bagdasarova, Matthew Lee, Meredith Price, Micah Schaffer, Michael Boone, Michael Fukuyama, Michael Hasper, Nancy Agarwal, Negar Habibi, Oleksii Kuchaiev, Oluwatobi Olabiyi, Pablo Ribalta, Payal Motwani, Pinky Xu, Prasoon Varshney, Pritam Biswas, Puhui Meng, Rameshwar Shivbhakta, Ran Zilberstein, Reka Kovacs, Richard Mazzarese, Rishabh Garg, Rita Fernandes Neves, Rochelle Woods, Ruby Ye, Ruvaid Virk, Ryan Hill, Saeid Motiian, Seph Mard, Shai Ben Nun, Shaona Ghosh, Sharad Yadav, Sudeep Sabnis, Suhas Hariharapura Sheshadri, Terry Kong, Thanh Do, Thinh Le, Udi Karpas, Viet Anh Trinh, Vinay Raman, Vinita Sharma, Warren Drones, Wenwen Gao, Whei Hsueh, Will Efstratis, Yoshi Suhara, Yunsheng Liu, Zhenming Wang, Zijia Chen, Zuhair Ahmed.

## References


Talor Abramovich, Maor Ashkenazi, Izzy Putterman, Benjamin Chislett, Tiyasa Mitra, Bita Darvish Rouhani, Ran Zilberstein, and Yonatan Geifman. SPEED-bench: A unified and diverse benchmark for speculative decoding. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Rl2uQlCoQX.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks, 2025. URL https://arxiv.org/abs/2412.15204.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. $\tau^2$-Bench: Evaluating Conversational Agents in a Dual-Control Environment, 2025. URL https://arxiv.org/abs/2506.07982.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling, 2023. URL https://arxiv.org/abs/2302.01318.

Jian Chen, Yesheng Liang, and Zhijian Liu. DFlash: Block diffusion for flash speculative decoding. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Oz335dV48X.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants, 2024. URL https://arxiv.org/abs/2410.17196.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

Team Gemma. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, and Yunzhong He. Audio multichallenge: A multi-turn evaluation of spoken dialogue systems on natural human interaction, 2025. URL https://arxiv.org/abs/2512.14865.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models?, 2024. URL https://arxiv.org/abs/2404.06654.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. ToolSandbox: A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities, 2025a. URL https://arxiv.org/abs/2408.04682.

Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Di Yin, and Xing Sun. RoleMRC: A Fine-Grained Composite Benchmark for Role-Playing and Instruction-Following, 2025b. URL https://arxiv.org/abs/2502.11387.

Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Jason Huang, Vishnu Suresh, Yixin Huang, Xiaowen Yu, Joseph E. Gonzalez, and Shishir G. Patil. BFCL V3: Multi-Turn and Multi-Step Function Calling. https://gorilla.cs.berkeley.edu/blogs/13_bfcl_v3_multi_turn.html, 2024. Version-specific benchmark announcement; released September 19, 2024.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models, 2024. URL https://arxiv.org/abs/2401.00396.

NVIDIA. Nemo gym: An open source library for scaling reinforcement learning environments for llm. https://github.com/NVIDIA-NeMo/Gym, 2025a. GitHub repository.

NVIDIA. Nemo rl: A scalable and efficient post-training library. https://github.com/NVIDIA-NeMo/RL, 2025b. GitHub repository.

NVIDIA. Cosmos 3: Omnimodal World Models for Physical AI. *arXiv preprint arXiv:2606.02800*, 2026. URL https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf.

NVIDIA Corporation. NVIDIA Model Optimizer. https://github.com/NVIDIA/Model-Optimizer, 2024–2026. GitHub repository.

Samuel J. Paech. EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models, 2024. URL https://arxiv.org/abs/2312.06281.

Samuel J. Paech. EQ-Bench 3: Emotional Intelligence Benchmark. https://github.com/EQ-bench/eqbench3, 2025. Benchmark repository; rubric-scoring evaluation.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley Function Calling Leaderboard (BFCL): From Tool Use to Agentic Evaluation of Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48371–48392. PMLR, 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing Verifiable Instruction Following, 2025. URL https://arxiv.org/abs/2507.02833.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Amy Xin, Youfeng Liu, Bin Xu, Lei Hou, and Juanzi Li. AGENTIF: Benchmarking Instruction Following of Large Language Models in Agentic Scenarios, 2025. URL https://arxiv.org/abs/2505.16944.

Yanzhao Qin, Tao Zhang, Tao Zhang, Yanjun Shen, Wenjing Luo, Haoze Sun, Yan Zhang, Yujing Qiao, Weipeng Chen, Zenan Zhou, Wentao Zhang, and Bin Cui. SysBench: Can Large Language Models Follow System Messages?, 2024. URL https://arxiv.org/abs/2408.10943.

Hayley Ross, Ameya Sunil Mahabaleshwarkar, and Yoshi Suhara. When2Call: When (not) to Call Tools, 2025. URL https://arxiv.org/abs/2504.18851.

Matan Rusanovsky, Yoav Miron, Roy Uziel, Omer Belhasin, Hao Guo, Ran Zilberstein, Maor Ashkenazi, and Michael Elad. Lilicorr: Lightweight likelihood correlation of parallel drafts for speculative decoding, 2026. URL https://arxiv.org/abs/2608.20530.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. MultiChallenge: A Realistic Multi-Turn Conversation Evaluation Benchmark Challenging to Frontier LLMs, 2025. URL https://arxiv.org/abs/2501.17399.

Ionut-Teodor Sorodoc, Leonardo F. R. Ribeiro, Rexhina Blloshmi, Christopher Davis, and Adrià de Gispert. GaRAGe: A Benchmark with Grounding Annotations for RAG Evaluation, 2025. URL https://arxiv.org/abs/2506.07671.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models, 2024. URL https://arxiv.org/abs/2310.00746.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory, 2025. URL https://arxiv.org/abs/2410.10813.

Hongshen Xu, Zichen Zhu, Lei Pan, Zihan Wang, Su Zhu, Da Ma, Ruisheng Cao, Lu Chen, and Kai Yu. Reducing Tool Hallucination via Reliability Alignment, 2025. URL https://arxiv.org/abs/2412.04141.

Jialin Yang, Dongfu Jiang, Lipeng He, Sherman Siu, Yuxuan Zhang, Disen Liao, Zhuofeng Li, Huaye Zeng, Yiming Jia, Haozhe Wang, Benjamin Schneider, Chi Ruan, Wentao Ma, Zhiheng Lyu, Yifei Wang, Yi Lu, Quy Duc Do, Ziyan Jiang, Ping Nie, and Wenhu Chen. StructEval: Benchmarking LLMs' Capabilities to Generate Structural Outputs, 2026. URL https://arxiv.org/abs/2505.20139.

Qinyan Zhang, Xinping Lei, Ruijie Miao, Yu Fu, Haojie Fan, Le Chang, Jiafan Hou, Dingling Zhang, Zhongfei Hou, Ziqiang Yang, Changxin Pu, Fei Hu, Jingkai Liu, Mengyun Liu, Yang Liu, Xiang Gao, Jiaheng Liu, Tong Yang, Zaiyuan Wang, Ge Zhang, and Wenhao Huang. Inverse IFEval: Can LLMs Unlearn Stubborn Training Conventions to Follow Real Instructions?, 2025a. URL https://arxiv.org/abs/2509.04292.

Zhihan Zhang, Shiyang Li, Zixuan Zhang, Xin Liu, Haoming Jiang, Xianfeng Tang, Yifan Gao, Zheng Li, Haodong Wang, Zhaoxuan Tan, Yichuan Li, Qingyu Yin, Bing Yin, and Meng Jiang. IHEval: Evaluating Language Models on Following the Instruction Hierarchy, 2025b. URL https://arxiv.org/abs/2502.08745.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023.