Muse Wonder Model Card

This card describes the Muse Wonder models, the evaluations we ran, and the limits of that testing.

---

1. The models

Name	Muse Wonder and Muse Wonder Full
Model Provider	Meta Platforms, Inc. or Meta Platforms Ireland Limited
Website	dev.meta.ai
Date of Release	October 12, 2026
Modalities	Input: text, image, video, audio Output: text
Languages	The model was trained on multilingual training data, with substantial representation across worldwide languages.
Context Length	1M tokens at inference
Contact	https://dev.meta.ai/support

2. Intended uses

The Muse Wonder models are intended for commercial and research use, with particular strengths in personal and professional agentic tasks. The models support a broad range of use cases, including:

1. **General-purpose assistance and reasoning** — answering questions, writing and editing, summarizing information, analyzing documents and data, and reasoning through complex problems.
2. **Coding and software engineering** — generating, explaining, debugging, testing, and modifying code, from individual programming tasks to sustained work across larger software projects.
3. **Personal and professional agents** — supporting research, planning, and multi-step workflows through tool use and interaction with applications.

Muse Wonder Model Card

This card describes the Muse Wonder models, the evaluations we ran, and the limits of that testing.

---

1. The models

Name	Muse Wonder and Muse Wonder Full
Model Provider	Meta Platforms, Inc. or Meta Platforms Ireland Limited
Website	dev.meta.ai
Date of Release	October 12, 2026
Modalities	Input: text, image, video, audio Output: text
Languages	The model was trained on multilingual training data, with substantial representation across worldwide languages.
Context Length	1M tokens at inference
Contact	https://dev.meta.ai/support

2. Intended uses

The Muse Wonder models are intended for commercial and research use, with particular strengths in personal and professional agentic tasks. The models support a broad range of use cases, including:

1. **General-purpose assistance and reasoning** — answering questions, writing and editing, summarizing information, analyzing documents and data, and reasoning through complex problems.
2. **Coding and software engineering** — generating, explaining, debugging, testing, and modifying code, from individual programming tasks to sustained work across larger software projects.
3. **Personal and professional agents** — supporting research, planning, and multi-step workflows through tool use and interaction with applications.

4. **Domain-specific knowledge work** — helping users interpret and work with specialized information in areas such as law, finance, and health, including understanding documents, synthesizing source material, and preparing drafts for review.
5. **Multimodal understanding and reasoning** — interpreting documents, screenshots, charts, images, and video alongside text.
6. **Research and development** — supporting scientific and technical research, including synthesizing information, developing and testing hypotheses, designing experiments, and analyzing results.

Known limitations are covered in Considerations and Limitations below.

3. Capabilities evaluations and eval methodology

We evaluate Muse Wonder and Muse Wonder Full on benchmarks spanning personal and professional agents (including legal, finance, and health tasks), software engineering and coding, and general intelligence.

4. **Domain-specific knowledge work** — helping users interpret and work with specialized information in areas such as law, finance, and health, including understanding documents, synthesizing source material, and preparing drafts for review.
5. **Multimodal understanding and reasoning** — interpreting documents, screenshots, charts, images, and video alongside text.
6. **Research and development** — supporting scientific and technical research, including synthesizing information, developing and testing hypotheses, designing experiments, and analyzing results.

Known limitations are covered in Considerations and Limitations below.

3. Capabilities evaluations and eval methodology

We evaluate Muse Wonder and Muse Wonder Full on benchmarks spanning personal and professional agents (including legal, finance, and health tasks), software engineering and coding, and general intelligence.

		Muse Wonder Full (max)	Muse Wonder (max)	Claude Fable 5.1 (max)	GPT-6 Astra (max)	Claude Opus 5.5 (max)
Personal and professional agents	**Personal Superintelligence Index** Internal composite index of personal and knowledge work	59.1	54.3	49.9	54.5	55.2
	Harvey LAB-AA Legal agent work	13.1	10.0	6.4	8.6	4.2
	Vals Finance Agent V2 Financial analyst research	59.1	56.6	58.9	53.5	58.6
	HealthBench Pro Open-ended health queries	62.3	60.5	62.1	65.3	65.6
	AutomationBench-AA E2E business workflows	45.4	39.1	32.1	41.6	41.2
	GDP.pdf Professional work	33.2	30.8	27.6	34.2	30.6
	JobBench Professional tool use	75.3	70.5	73.1	66.1	75.4
	OSWorld V2.1 Agentic computer use	78.4 (partial) 39.4 (binary)	72.6 (partial) 34.3 (binary)	80.7 (partial) 43.1 (binary)	75.8 (partial) 34.3 (binary)	81.8 (partial) – (binary)
	MemoryBench Remembering context	85.5	85.6	79.0	81.4	81.5
Coding	**Terminal-Bench 4.0-AA** Agentic terminal coding	57.1	52.0	52.0	59.1	59.6
	DeepSWE 1.1 Long horizon coding	79.0	78.8	67.4	74.1	74.2
	SWE-Marathon Ultra-long horizon engineering	57.5 (xhigh)	53.1	45.6	42.5	46.3
	ProgramBench From scratch programming	59.0	50.0	55.5	52.5	70.5
General	**Artificial Analysis Index**	56	52	53	53	58

		Muse Wonder Full (max)	Muse Wonder (max)	Claude Fable 5.1 (max)	GPT-6 Astra (max)	Claude Opus 5.5 (max)
Personal and professional agents	**Personal Superintelligence Index** Internal composite index of personal and knowledge work	59.1	54.3	49.9	54.5	55.2
	Harvey LAB-AA Legal agent work	13.1	10.0	6.4	8.6	4.2
	Vals Finance Agent V2 Financial analyst research	59.1	56.6	58.9	53.5	58.6
	HealthBench Pro Open-ended health queries	62.3	60.5	62.1	65.3	65.6
	AutomationBench-AA E2E business workflows	45.4	39.1	32.1	41.6	41.2
	GDP.pdf Professional work	33.2	30.8	27.6	34.2	30.6
	JobBench Professional tool use	75.3	70.5	73.1	66.1	75.4
	OSWorld V2.1 Agentic computer use	78.4 (partial) 39.4 (binary)	72.6 (partial) 34.3 (binary)	80.7 (partial) 43.1 (binary)	75.8 (partial) 34.3 (binary)	81.8 (partial) – (binary)
	MemoryBench Remembering context	85.5	85.6	79.0	81.4	81.5
Coding	**Terminal-Bench 4.0-AA** Agentic terminal coding	57.1	52.0	52.0	59.1	59.6
	DeepSWE 1.1 Long horizon coding	79.0	78.8	67.4	74.1	74.2
	SWE-Marathon Ultra-long horizon engineering	57.5 (xhigh)	53.1	45.6	42.5	46.3
	ProgramBench From scratch programming	59.0	50.0	55.5	52.5	70.5
General	**Artificial Analysis Index**	56	52	53	53	58

Overall Methodology

- **Models and configurations.** Muse Wonder and Muse Wonder Full results are generated with max reasoning effort. We compare with GPT-6 Astra, Claude Fable 5.1, and Claude Opus 5.5 at max reasoning effort. For controlled internal coding comparisons, every model uses the same benchmark-specific harness. External coding results retain the harness or agent product used by the reporting vendor or model provider, and that source is identified below.
- **Evaluation and score selection.** We preserve each benchmark's tasks, inputs, grading, and primary metric as closely as possible. Where a comparable third-party or provider-reported result exists, we use it; otherwise we use the highest of our internally validated run, self-reported score, or official leaderboard score. Benchmark names link to official leaderboards where available. Model refusals receive zero credit and remain in the denominator rather than being dropped. Material departures from the official setup are noted below.
- **Competitor-model results.** Internally generated competitor runs use common settings where practical but remain best-effort: prompts, tools, and runtime may not be optimally tuned for proprietary models.

Per-benchmark details

Personal and Professional Agents

- **Personal Superintelligence Index (PSI).** PSI is an internal composite of seven personal and knowledge-work evaluations: Agents' Last Exam, AutomationBench-AA, FocusBench, JobBench, GDP.pdf, VBT, and DRACO.
- **Harvey LAB-AA.** Harvey LAB-AA is Artificial Analysis's version of Harvey's Legal Agent Benchmark, in which agents produce legal work products from matter documents and an Artificial Analysis judge panel checks each deliverable against task-specific criteria. We report version 1.1, which adds hallucination checks. The result source is Artificial Analysis, and the primary metric is the share of tasks whose deliverable meets every criterion and passes the hallucination checks.
- **Vals Finance Agent V2.** Vals Finance Agent V2 evaluates financial-analyst tasks requiring research over filings and market information. The result source is Vals AI, and the primary metric is its private-test-set accuracy.
- **HealthBench Pro.** HealthBench Pro evaluates responses to 525 professional-grade health questions. Models answer without tools, and a GPT-5.4 grader scores each response against physician-written rubrics. We run the official tasks and grading in a common internal framework. The primary metric is the length-normalized weighted rubric score. For Claude Fable 5.1 and Claude Opus

Overall Methodology

- **Models and configurations.** Muse Wonder and Muse Wonder Full results are generated with max reasoning effort. We compare with GPT-6 Astra, Claude Fable 5.1, and Claude Opus 5.5 at max reasoning effort. For controlled internal coding comparisons, every model uses the same benchmark-specific harness. External coding results retain the harness or agent product used by the reporting vendor or model provider, and that source is identified below.
- **Evaluation and score selection.** We preserve each benchmark's tasks, inputs, grading, and primary metric as closely as possible. Where a comparable third-party or provider-reported result exists, we use it; otherwise we use the highest of our internally validated run, self-reported score, or official leaderboard score. Benchmark names link to official leaderboards where available. Model refusals receive zero credit and remain in the denominator rather than being dropped. Material departures from the official setup are noted below.
- **Competitor-model results.** Internally generated competitor runs use common settings where practical but remain best-effort: prompts, tools, and runtime may not be optimally tuned for proprietary models.

Per-benchmark details

Personal and Professional Agents

- **Personal Superintelligence Index (PSI).** PSI is an internal composite of seven personal and knowledge-work evaluations: Agents' Last Exam, AutomationBench-AA, FocusBench, JobBench, GDP.pdf, VBT, and DRACO.
- **Harvey LAB-AA.** Harvey LAB-AA is Artificial Analysis's version of Harvey's Legal Agent Benchmark, in which agents produce legal work products from matter documents and an Artificial Analysis judge panel checks each deliverable against task-specific criteria. We report version 1.1, which adds hallucination checks. The result source is Artificial Analysis, and the primary metric is the share of tasks whose deliverable meets every criterion and passes the hallucination checks.
- **Vals Finance Agent V2.** Vals Finance Agent V2 evaluates financial-analyst tasks requiring research over filings and market information. The result source is Vals AI, and the primary metric is its private-test-set accuracy.
- **HealthBench Pro.** HealthBench Pro evaluates responses to 525 professional-grade health questions. Models answer without tools, and a GPT-5.4 grader scores each response against physician-written rubrics. We run the official tasks and grading in a common internal framework. The primary metric is the length-normalized weighted rubric score. For Claude Fable 5.1 and Claude Opus

5.5, we use the results Anthropic reports in the Claude Fable 5.1 system card and the Claude Opus 5.5 system card.

- **AutomationBench-AA.** AutomationBench-AA evaluates realistic business workflows in a simulated application environment. Agents complete each workflow through automation tools, and deterministic end-state assertions inspect the resulting application state. The result source is Artificial Analysis, and the primary metric is its strict pass@1: the share of workflows that pass every end-state assertion.
- **Agents' Last Exam.** Agents' Last Exam evaluates demanding end-to-end scientific and professional tasks performed on Linux virtual machines. We use the 105-task Linux set. Each task is graded by its native evaluation script, which awards partial credit between 0 and 1 without an LLM judge.
- **GDP.pdf.** GDP.pdf evaluates professional work grounded in supplied PDF source material. The result source is Surge, and the primary metric is Surge's reported score.
- **OSWorld V2.1.** OSWorld V2.1 is a long-horizon computer-use benchmark of 108 real-world workflows on a full Ubuntu desktop VM. The agent controls the GUI through screenshots and actions such as clicking, typing, scrolling, and key presses. We run the updated V2.1 evaluation in a common internal framework while preserving the official task environments and programmatic graders. The primary metric is mean per-task partial reward; strict binary completion is a secondary metric. The Claude Opus 5.5 result is Anthropic's provider-reported partial score; no binary figure is available.
- **JobBench.** JobBench contains 65 professional tasks across 35 white-collar occupations. Each task asks the agent to produce a practical work product and is scored against task-specific rubrics. The primary metric is mean reward across tasks.
- **MemoryBench.** MemoryBench is an internal benchmark that tests how accurately a model can recall critical information across long, complex multi-turn conversations of 64K to 100K tokens that span several individual facts, with distractor conversations and injected turns that conflict with or later update the target details. An LLM judge scores each answer against weighted, question-specific rubric criteria.
- **Verifiable Browsing Task (VBT).** VBT is an internal agentic web-browsing evaluation in which models answer list-style information-seeking questions, written as realistic user requests, using browser search, open, and find tools.
- **Deep Research Accuracy, Completeness, and Objectivity (DRACO).** DRACO, from Perplexity, is a deep-research benchmark of 100 complex, multi-domain questions.

5.5, we use the results Anthropic reports in the Claude Fable 5.1 system card and the Claude Opus 5.5 system card.

- **AutomationBench-AA.** AutomationBench-AA evaluates realistic business workflows in a simulated application environment. Agents complete each workflow through automation tools, and deterministic end-state assertions inspect the resulting application state. The result source is Artificial Analysis, and the primary metric is its strict pass@1: the share of workflows that pass every end-state assertion.
- **Agents' Last Exam.** Agents' Last Exam evaluates demanding end-to-end scientific and professional tasks performed on Linux virtual machines. We use the 105-task Linux set. Each task is graded by its native evaluation script, which awards partial credit between 0 and 1 without an LLM judge.
- **GDP.pdf.** GDP.pdf evaluates professional work grounded in supplied PDF source material. The result source is Surge, and the primary metric is Surge's reported score.
- **OSWorld V2.1.** OSWorld V2.1 is a long-horizon computer-use benchmark of 108 real-world workflows on a full Ubuntu desktop VM. The agent controls the GUI through screenshots and actions such as clicking, typing, scrolling, and key presses. We run the updated V2.1 evaluation in a common internal framework while preserving the official task environments and programmatic graders. The primary metric is mean per-task partial reward; strict binary completion is a secondary metric. The Claude Opus 5.5 result is Anthropic's provider-reported partial score; no binary figure is available.
- **JobBench.** JobBench contains 65 professional tasks across 35 white-collar occupations. Each task asks the agent to produce a practical work product and is scored against task-specific rubrics. The primary metric is mean reward across tasks.
- **MemoryBench.** MemoryBench is an internal benchmark that tests how accurately a model can recall critical information across long, complex multi-turn conversations of 64K to 100K tokens that span several individual facts, with distractor conversations and injected turns that conflict with or later update the target details. An LLM judge scores each answer against weighted, question-specific rubric criteria.
- **Verifiable Browsing Task (VBT).** VBT is an internal agentic web-browsing evaluation in which models answer list-style information-seeking questions, written as realistic user requests, using browser search, open, and find tools.
- **Deep Research Accuracy, Completeness, and Objectivity (DRACO).** DRACO, from Perplexity, is a deep-research benchmark of 100 complex, multi-domain questions.

We run it in a common internal framework, where models answer using browser search, open, and find tools plus code execution in a sandbox.

- **FocusBench.** FocusBench is an internal evaluation of personal agentic tasks across a variety of domains like housing, travel, IT, and finance. It evaluates how well agents can persistently work through real-world blockers across multiple interactions and adapt to dynamically changing environments.

Software Engineering and Coding

- **Terminal-Bench 4.0.** Terminal-Bench 4.0 evaluates harder tasks completed in isolated terminal environments, with executable verifiers grading the final container state. The result source is Artificial Analysis, and the primary metric is its reported accuracy.
- **DeepSWE v1.1.** DeepSWE is a long-horizon software-engineering benchmark with 113 tasks across 91 repositories and five languages. Handwritten functional and regression tests grade each task, and the primary metric is pass@1. The scorecard uses Datacurve's third-party results for Muse Wonder, Muse Wonder Full, and Claude Opus 5.5, and OpenAI's provider-reported result for GPT-6 Astra. For Claude Fable 5.1, we use the result reported in Anthropic's Claude Fable 5.1 system card.
- **SWE-Marathon.** SWE-Marathon is an ultra-long-horizon software-engineering benchmark of substantial projects graded by executing project tests. The primary metric is pass@1. Abundant AI is the result source for all models. Its leaderboard results for GPT-6 Astra and Claude Fable 5.1 use Codex and Claude Code, respectively, so this row is not a controlled common-harness comparison. Muse Wonder Full was reported with xhigh reasoning effort.
- **ProgramBench.** ProgramBench evaluates whether an agent can reconstruct a complete program in a clean-room environment. Each submission is compiled and graded by a hidden behavioral test suite, with internet access disabled after setup. All compared models use mini-swe-agent. The primary metric is the percentage of rollouts passing at least 95% of tests.

General

- **AA Intelligence Index v4.3.** AA Intelligence Index v4.3 is Artificial Analysis's composite index. It combines evaluations Artificial Analysis runs independently, including GDPval-AA, AA-Briefcase, AutomationBench-AA, GDP.pdf, Terminal-Bench 4.0, SciCode, AA-LCR, AA-Omniscience, CritPt, and Humanity's Last Exam, into a single score. The result source is Artificial Analysis, and the primary metric is the index score.

We run it in a common internal framework, where models answer using browser search, open, and find tools plus code execution in a sandbox.

- **FocusBench.** FocusBench is an internal evaluation of personal agentic tasks across a variety of domains like housing, travel, IT, and finance. It evaluates how well agents can persistently work through real-world blockers across multiple interactions and adapt to dynamically changing environments.

Software Engineering and Coding

- **Terminal-Bench 4.0.** Terminal-Bench 4.0 evaluates harder tasks completed in isolated terminal environments, with executable verifiers grading the final container state. The result source is Artificial Analysis, and the primary metric is its reported accuracy.
- **DeepSWE v1.1.** DeepSWE is a long-horizon software-engineering benchmark with 113 tasks across 91 repositories and five languages. Handwritten functional and regression tests grade each task, and the primary metric is pass@1. The scorecard uses Datacurve's third-party results for Muse Wonder, Muse Wonder Full, and Claude Opus 5.5, and OpenAI's provider-reported result for GPT-6 Astra. For Claude Fable 5.1, we use the result reported in Anthropic's Claude Fable 5.1 system card.
- **SWE-Marathon.** SWE-Marathon is an ultra-long-horizon software-engineering benchmark of substantial projects graded by executing project tests. The primary metric is pass@1. Abundant AI is the result source for all models. Its leaderboard results for GPT-6 Astra and Claude Fable 5.1 use Codex and Claude Code, respectively, so this row is not a controlled common-harness comparison. Muse Wonder Full was reported with xhigh reasoning effort.
- **ProgramBench.** ProgramBench evaluates whether an agent can reconstruct a complete program in a clean-room environment. Each submission is compiled and graded by a hidden behavioral test suite, with internet access disabled after setup. All compared models use mini-swe-agent. The primary metric is the percentage of rollouts passing at least 95% of tests.

General

- **AA Intelligence Index v4.3.** AA Intelligence Index v4.3 is Artificial Analysis's composite index. It combines evaluations Artificial Analysis runs independently, including GDPval-AA, AA-Briefcase, AutomationBench-AA, GDP.pdf, Terminal-Bench 4.0, SciCode, AA-LCR, AA-Omniscience, CritPt, and Humanity's Last Exam, into a single score. The result source is Artificial Analysis, and the primary metric is the index score.

4. Risk assessments under the Meta Superintelligence Scaling Framework

Results below are from Muse Wonder Full, our most capable variant, which is why the assessment covers every model in the family. Under the Meta Superintelligence Scaling Framework, a model reaches the high risk threshold when its deployment could substantially contribute to a threat scenario we have identified as severe. Without mitigations, we do not rule out the model reaching this threshold in the Chemical & Biological domain. Because of that, we apply mitigations designed for this domain, including safety post-training and system-level classifiers. Our evaluations put the unmitigated model at “moderate or lower” for both Cybersecurity and Loss of Control before applying our mitigations. With mitigations and system-level safeguards in place, we reach the same “moderate or lower” level across all domains, which supports our decision to deploy.

Cyber Capability Evaluations

Results

Evaluation	Muse Wonder Full
CyberGym	88.4%
ExploitBench	58.9%
ExploitGym	92*

***Note on ExploitGym scoring.** Muse Wonder Full succeeded in capturing 92 flags. A verification pass, run with a GPT-cyber verifier and independently confirmed with Gemini 3.1 Pro, found that 21 of those used the intended vulnerability, while the other 71 were captured through unintended vulnerabilities. ExploitGym's published ranking counts only intended-vulnerability exploitations, so 21 is the figure comparable to the published ranking.*

Cyber Robustness Evaluations (Adversarial Refusal)

We evaluate our models for adversarial robustness, particularly the ability to resist a malicious actor's efforts to jailbreak the model or work around its safeguards. Attack Success Rate (ASR) is reported as follows:

4. Risk assessments under the Meta Superintelligence Scaling Framework

Results below are from Muse Wonder Full, our most capable variant, which is why the assessment covers every model in the family. Under the Meta Superintelligence Scaling Framework, a model reaches the high risk threshold when its deployment could substantially contribute to a threat scenario we have identified as severe. Without mitigations, we do not rule out the model reaching this threshold in the Chemical & Biological domain. Because of that, we apply mitigations designed for this domain, including safety post-training and system-level classifiers. Our evaluations put the unmitigated model at “moderate or lower” for both Cybersecurity and Loss of Control before applying our mitigations. With mitigations and system-level safeguards in place, we reach the same “moderate or lower” level across all domains, which supports our decision to deploy.

Cyber Capability Evaluations

Results

Evaluation	Muse Wonder Full
CyberGym	88.4%
ExploitBench	58.9%
ExploitGym	20*

***Updated ExploitGym Evaluations (23 October 2026).** After correcting our harness to include the intended vulnerability in the model’s prompts, as recommended by the benchmark, we reran the evaluation. Muse Wonder Full succeeded in capturing 20 intended flags (instead of the 92 total flags and 21 intended flags originally reported). We have updated the table above to reflect the intended flag results only, as this is also how results are officially reported on the public leaderboard. These updated evaluations do not require us to adjust our “moderate or lower” risk threshold for Muse Wonder Full.*

Cyber Robustness Evaluations (Adversarial Refusal)

We evaluate our models for adversarial robustness, particularly the ability to resist a malicious actor's efforts to jailbreak the model or work around its safeguards. Attack Success Rate (ASR) is reported as follows:

Evaluation	Attack Success Rate (ASR) % *lower is better*
	Muse Wonder Full
CyberBench	
Single turn	0.8
Multi turn	0.8
PolyGuard (cyber subset)	
Single turn	1.1
Multi turn	2.3
RedCode	
Single turn	9.0
Multi turn	8.0

Muse Wonder Full demonstrates strong robustness to adversarial attacks across all benchmarks.

Evaluation	Attack Success Rate (ASR) % *lower is better*
	Muse Wonder Full
CyberBench	
Single turn	0.8
Multi turn	0.8
PolyGuard (cyber subset)	
Single turn	1.1
Multi turn	2.3
RedCode	
Single turn	9.0
Multi turn	8.0

Muse Wonder Full demonstrates strong robustness to adversarial attacks across all benchmarks.

CBRN Capability Evaluations

These proxy evaluations measure model performance on dual-use capabilities. A brief description for each eval can be found in the Appendix below.

Evaluation			Muse Wonder Full
Dual-Use Bio/Chem Knowledge	SecureBio	MBCT	59.6%
		HPCT	66.5%
		VCT	55.9%
		Average	60.7%
	Lab-Bench (v1)	ProtocolQA	89.0%
Bioinformatics / Computational Design		SeqQA + tools	95.6%
	ABC Bench	Fragment Design	-**
		Liquid Handling	100.0%
		Screening Evasion	-**
	BioDesign Tools Usage	Meta BDT	79.4%

*** = High rate of refusals due to safety training. Refusals dropped prior to scoring.*

CBRN Safety Evaluations

We evaluate refusal coverage on high-risk biological and chemical prompts in English, across languages commonly used on Meta services, and under a fixed portfolio of adversarial transformations.

BioTIER is an industry-leading benchmark developed by SecureBio to measure safety mitigations for high-risk dual-use biological topics. Meta also developed an in-house benchmark to measure model safety behavior on high-risk chemistry questions. Results show strong refusal performance in English, across 10 language buckets, and in the presence of adversarially designed prompts (higher is better).

CBRN Capability Evaluations

These proxy evaluations measure model performance on dual-use capabilities. A brief description for each eval can be found in the Appendix below.

Evaluation			Muse Wonder Full
Dual-Use Bio/Chem Knowledge	SecureBio	MBCT	59.6%
		HPCT	66.5%
		VCT	55.9%
		Average	60.7%
	Lab-Bench (v1)	ProtocolQA	89.0%
Bioinformatics / Computational Design		SeqQA + tools	95.6%
	ABC Bench	Fragment Design	-**
		Liquid Handling	100.0%
		Screening Evasion	-**
	BioDesign Tools Usage	Meta BDT	79.4%

*** = High rate of refusals due to safety training. Refusals dropped prior to scoring.*

CBRN Safety Evaluations

We evaluate refusal coverage on high-risk biological and chemical prompts in English, across languages commonly used on Meta services, and under a fixed portfolio of adversarial transformations.

BioTIER is an industry-leading benchmark developed by SecureBio to measure safety mitigations for high-risk dual-use biological topics. Meta also developed an in-house benchmark to measure model safety behavior on high-risk chemistry questions. Results show strong refusal performance in English, across 10 language buckets, and in the presence of adversarially designed prompts (higher is better).

Evaluation		Muse Wonder Full
BioTIER	English	98.7%
	Multilingual	98.7%
	Adversarial	98.2%
Internal Chemistry Refusal	English	100.0%
	Multilingual	100.0%
	Adversarial	99.8%

Loss of Control Evaluations

Loss of control refers to a situation in which humans lose, and cannot feasibly regain, the ability to direct, modify, contain, or shut down AI systems that have the potential for significant real-world impact. Control could become infeasible because an AI system systematically resists attempts to control it, because taking control becomes too costly or difficult to coordinate, or because humans are no longer able to understand what the AI system is doing. We evaluate Muse Wonder on two evaluations for this domain.

- MASK assesses whether the model maintains honesty when pressured to contradict its baseline knowledge, a property relevant to reliable oversight. Higher scores are better, indicating higher propensity to maintain honesty under pressure.
- Agentic Misalignment assesses whether an agent takes harmful actions to advance its objectives or avoid replacement or shutdown. Lower scores are better, indicating a lower harmful action rate.

Evaluation	Muse Wonder Full
MASK	87.3%
Agentic Misalignment	0%

Evaluation		Muse Wonder Full
BioTIER	English	98.7%
	Multilingual	98.7%
	Adversarial	98.2%
Internal Chemistry Refusal	English	100.0%
	Multilingual	100.0%
	Adversarial	99.8%

Loss of Control Evaluations

Loss of control refers to a situation in which humans lose, and cannot feasibly regain, the ability to direct, modify, contain, or shut down AI systems that have the potential for significant real-world impact. Control could become infeasible because an AI system systematically resists attempts to control it, because taking control becomes too costly or difficult to coordinate, or because humans are no longer able to understand what the AI system is doing. We evaluate Muse Wonder on two evaluations for this domain.

- MASK assesses whether the model maintains honesty when pressured to contradict its baseline knowledge, a property relevant to reliable oversight. Higher scores are better, indicating higher propensity to maintain honesty under pressure.
- Agentic Misalignment assesses whether an agent takes harmful actions to advance its objectives or avoid replacement or shutdown. Lower scores are better, indicating a lower harmful action rate.

Evaluation	Muse Wonder Full
MASK	87.3%
Agentic Misalignment	0%

5. Trust and Safety

System protections are central to keeping the model helpful and safe, addressing security risks that arise at the system level, and integrating it with external tools. We recommend deploying Muse Wonder as part of a broader AI system, with additional guardrails suited to the use case.

Consistent with the Meta Superintelligence Scaling Framework (https://research.meta.ai/static/meta-superintelligence-scaling-framework), we ran extensive evaluations, including red-teaming, across the Framework's catastrophic risk domains of Cybersecurity, Chemical & Biological, and Loss of Control. With mitigations in place, we assessed residual risk at the moderate or lower threshold. Our internal governance process reviewed that assessment before we reached a deployment decision. Throughout development and deployment, we also implemented measures to protect model weights under Meta's lifecycle-based weight security program.

We worked with third parties to develop and license evaluation benchmarks to support our catastrophic risk assessments. We ran threat modeling with outside experts, and engaged with government entities and other stakeholders whose input informed our evaluations and requirements.

6. Considerations and Limitations

Muse Wonder is a technology that carries known and unknown risks. Testing conducted to date has not covered, and could not cover, all scenarios.

Limitations

1. The model may produce inaccurate, biased, or objectionable responses to user prompts.
2. While optimized for agentic tasks, the model may still make errors in multi-step reasoning, particularly in novel scenarios not well represented in training data.
3. Visual and spatial reasoning, especially 3D tasks, remains a known limitation. Validate performance for applications that depend on precise spatial understanding.
4. The model has not been evaluated on all languages contained in the pre-training data. Performance may degrade on languages outside the strongly supported set.
5. Performance can vary with quantization, serving configuration, prompts, tools, and the agent harness.
6. The model is not intended to be used by individuals under the age of 18. Deployers are responsible for ensuring that any risks associated with such use by individuals

5. Trust and Safety

System protections are central to keeping the model helpful and safe, addressing security risks that arise at the system level, and integrating it with external tools. We recommend deploying Muse Wonder as part of a broader AI system, with additional guardrails suited to the use case.

Consistent with the Meta Superintelligence Scaling Framework (https://research.meta.ai/static/meta-superintelligence-scaling-framework), we ran extensive evaluations, including red-teaming, across the Framework's catastrophic risk domains of Cybersecurity, Chemical & Biological, and Loss of Control. With mitigations in place, we assessed residual risk at the moderate or lower threshold. Our internal governance process reviewed that assessment before we reached a deployment decision. Throughout development and deployment, we also implemented measures to protect model weights under Meta's lifecycle-based weight security program.

We worked with third parties to develop and license evaluation benchmarks to support our catastrophic risk assessments. We ran threat modeling with outside experts, and engaged with government entities and other stakeholders whose input informed our evaluations and requirements.

6. Considerations and Limitations

Muse Wonder is a technology that carries known and unknown risks. Testing conducted to date has not covered, and could not cover, all scenarios.

Limitations

1. The model may produce inaccurate, biased, or objectionable responses to user prompts.
2. While optimized for agentic tasks, the model may still make errors in multi-step reasoning, particularly in novel scenarios not well represented in training data.
3. Visual and spatial reasoning, especially 3D tasks, remains a known limitation. Validate performance for applications that depend on precise spatial understanding.
4. The model has not been evaluated on all languages contained in the pre-training data. Performance may degrade on languages outside the strongly supported set.
5. Performance can vary with quantization, serving configuration, prompts, tools, and the agent harness.
6. The model is not intended to be used by individuals under the age of 18. Deployers are responsible for ensuring that any risks associated with such use by individuals

under the age of 18 have been fully assessed and appropriately mitigated, and that such use complies with all applicable laws.

Muse Wonder's use is subject to the restrictions and conditions in the applicable terms of use, including the Meta Model API Terms of Service (https://dev.meta.ai/legal/terms-of-service), Meta Model Acceptable Use Policy (https://dev.meta.ai/legal/acceptable-use-policy), and Muse Supplemental Terms of Service (https://muse.ai/terms).

Responsible Use
Developers should conduct their own evaluation of whether the model is appropriate for their deployment, with assessments tailored to their specific applications and proposed languages. We recommend implementing additional guardrails (such as human-in-the-loop confirmation for irreversible actions) when deploying the model in agentic contexts where it can take real-world actions.

Appendix: CBRN Proxy Evaluation Descriptions

SecureBio Molecular Biology Capabilities Test (MBCT), Human Pathogens Capabilities Test (HPCT), and Virology Capabilities Test (VCT)
The MBCT, VCT, and HPCT are part of a suite of evaluations developed by SecureBio and the Center for AI Safety. These evaluations are designed to assess practical troubleshooting across a range of molecular biology tasks (MBCT), wet-lab virology experiments (VCT), and practical knowledge about working with human pathogens considered high-priority by biosecurity experts (HPCT).

LAB-Bench (ProtocolQA and SeqQA)
LAB-Bench is an evaluation suite designed to assess AI capabilities on practical biology research tasks essential for scientific research including protocol planning and data analysis. The ProtocolQA task assesses the ability to debug practical wet-lab protocols. SeqQA contains a collection of 15 common tasks related to analysis and modification of DNA sequences that are common in molecular biology workflows (e.g., PCR).

Agentic Bio-Capabilities Benchmark (ABC Bench)
ABC Bench was developed by SecureBio and is intended to measure an LLM's ability to complete a set of computational tasks relevant to a potential threat actor who intends to design, synthesize, and assemble a complex genetic construct. ABC Bench contains three

under the age of 18 have been fully assessed and appropriately mitigated, and that such use complies with all applicable laws.

Muse Wonder's use is subject to the restrictions and conditions in the applicable terms of use, including the Meta Model API Terms of Service (https://dev.meta.ai/legal/terms-of-service), Meta Model Acceptable Use Policy (https://dev.meta.ai/legal/acceptable-use-policy), and Muse Supplemental Terms of Service (https://muse.ai/terms).

Responsible Use
Developers should conduct their own evaluation of whether the model is appropriate for their deployment, with assessments tailored to their specific applications and proposed languages. We recommend implementing additional guardrails (such as human-in-the-loop confirmation for irreversible actions) when deploying the model in agentic contexts where it can take real-world actions.

Appendix: CBRN Proxy Evaluation Descriptions

SecureBio Molecular Biology Capabilities Test (MBCT), Human Pathogens Capabilities Test (HPCT), and Virology Capabilities Test (VCT)
The MBCT, VCT, and HPCT are part of a suite of evaluations developed by SecureBio and the Center for AI Safety. These evaluations are designed to assess practical troubleshooting across a range of molecular biology tasks (MBCT), wet-lab virology experiments (VCT), and practical knowledge about working with human pathogens considered high-priority by biosecurity experts (HPCT).

LAB-Bench (ProtocolQA and SeqQA)
LAB-Bench is an evaluation suite designed to assess AI capabilities on practical biology research tasks essential for scientific research including protocol planning and data analysis. The ProtocolQA task assesses the ability to debug practical wet-lab protocols. SeqQA contains a collection of 15 common tasks related to analysis and modification of DNA sequences that are common in molecular biology workflows (e.g., PCR).

Agentic Bio-Capabilities Benchmark (ABC Bench)
ABC Bench was developed by SecureBio and is intended to measure an LLM's ability to complete a set of computational tasks relevant to a potential threat actor who intends to design, synthesize, and assemble a complex genetic construct. ABC Bench contains three

tasks: Fragment Design tests a model's ability to assist in silico design of short DNA fragments that can be assembled into a full-length genetic construct; Screening Evasion tests the model's ability to modify the sequence of each fragment to avoid DNA synthesis screening systems while preserving overall functionality; and the Liquid-Handling Robot task tests model performance in generating a script that enables DNA assembly using a commercial liquid handling robot.

BioDesign Tools (Meta BDT)
The BioDesign Tools evaluation was developed by Meta in collaboration with Faculty.AI and external subject-matter experts. It is designed to measure model performance across a range of complex computational tasks that have potential dual-use applications including the isolation, characterization, or modification of harmful pathogens. Results show performance averaged across five tasks from three categories: Host-Pathogen Interaction, Protein Design, and Sequence Analysis.

tasks: Fragment Design tests a model's ability to assist in silico design of short DNA fragments that can be assembled into a full-length genetic construct; Screening Evasion tests the model's ability to modify the sequence of each fragment to avoid DNA synthesis screening systems while preserving overall functionality; and the Liquid-Handling Robot task tests model performance in generating a script that enables DNA assembly using a commercial liquid handling robot.

BioDesign Tools (Meta BDT)
The BioDesign Tools evaluation was developed by Meta in collaboration with Faculty.AI and external subject-matter experts. It is designed to measure model performance across a range of complex computational tasks that have potential dual-use applications including the isolation, characterization, or modification of harmful pathogens. Results show performance averaged across five tasks from three categories: Host-Pathogen Interaction, Protein Design, and Sequence Analysis.

Comparison Summary Report

muse-wonder-model-card-2026-10-12-original.pdf compared with
muse-wonder-model-card-2026-10-23-revised.pdf
2026-10-26 12:57:38
3 changes

Text Changes 3

Insertions	0
Deletions	0
Replacements	3
Moves	0

Visual Changes 0

Insertions	0
Deletions	0
Replacements	0
Moves	0

Formatting Changes 0

Total Changes 3

Revisions Summary Report

muse-wonder-model-card-2026-10-12-original.pdf compared with
muse-wonder-model-card-2026-10-23-revised.pdf
2026-10-26 12:57:38
3 changes

Added 0 Removed 0 Replaced 3 Moved 0 Formatting 0

#1 (p.7)	Replaced (Text)	**Before:** ? **After:**
#2 (p.7)	Replaced (Text)	**Before:** 92 **After:** 20
#3 (p.7)	Replaced (Text)	**Before:** *Note on ExploitGym scoring. Muse Wonder Full succeeded in capturing 92 flags. A verification pass, run with a GPT-cyber verifier and independently confirmed with Gemini 3.1 Pro, found that 21 of those used the intended vulnerability, while the other 71 were captured through unintended vulnerabilities. ExploitGym's published ranking counts only intended-vulnerability exploitations, so 21 is the figure comparable to the published ranking. **After:** *Updated ExploitGym Evaluations (23 October 2026). After correcting our harness to include the intended vulnerability in the model’s prompts, as recommended by the benchmark, we reran the evaluation. Muse Wonder Full succeeded in capturing 20 intended flags (instead of the 92 total flags and 21 intended flags originally reported). We have updated the table above to reflect the intended flag results only, as this is also how results are officially reported on the public leaderboard. These updated evaluations do not require us to adjust our “moderate or lower” risk threshold for Muse Wonder Full.