Muse Spark Contemplating

Safety & Preparedness Report

Meta

Muse Spark Contemplating Safety & Preparedness Report

Muse Spark Contemplating is a deep reasoning model built upon the Muse Spark foundation. Muse Spark Contemplating introduces multi-agent orchestration at inference time, similar to the "extreme reasoning" or "deep think" modes of other language models.

In this report, we present Muse Spark Contemplating's preparedness profile under Meta's Advanced AI Scaling Framework. We conduct evaluations across the risk domains specified by the Advanced AI Scaling Framework: Chemical and Biological, Cybersecurity, and Loss of Control.

Our assessments indicate that Muse Spark Contemplating's extended reasoning and multi-agent orchestration retains the same risk thresholds as Muse Spark. While Muse Spark Contemplating may amplify existing capabilities in ways that warrant careful measurement, it does not introduce qualitatively new risk vectors. As a result, we continue to find the same multi-layered mitigations deployed for Muse Spark are adequate safeguards. For example, Muse Spark Contemplating continues to demonstrate state-of-the-art refusal across a range of benchmarks related to hazardous workflows in chemistry and biology. We therefore release Muse Spark Contemplating within Meta AI.

Date: June 3, 2026
Authors: MSL Preparedness & Red Teaming & Alignment Team, AI Security Team
Correspondence: Nathaniel Li (natli@meta.com); Summer Yue (summeryue@meta.com)

Meta

Contents

1 Introduction

We present Muse Spark Contemplating, a highly capable reasoning model that extends the capabilities of Muse Spark.[1] Muse Spark Contemplating employs multi-agent orchestration and deep inference reasoning to tackle challenging problems in science, mathematics, and code. As a model with enhanced capabilities, we assess Muse Spark Contemplating through a rigorous risk and safety evaluation to support responsible deployment considerations.

This report builds upon the Muse Spark Safety & Preparedness Report, focusing on the incremental risk and safety profile introduced by Muse Spark Contemplating's multi-agent orchestration and extended reasoning capabilities. This document does not reproduce all evaluations from the Muse Spark Safety & Preparedness Report; instead, it covers a targeted subset of assessments most salient to Muse Spark Contemplating's specific deployment context and capability profile. We first present a summary of the model's risk, safety, and behavioral profile (Section 1.1), providing a short overview of our assessments, the model's positioning relative to our risk thresholds and model families of comparable capabilities. We then describe the governance process that led to the decision to deploy Muse Spark Contemplating under Meta's Advanced AI Scaling Framework[2] (Section 1.2) and the methodology underlying the evaluation process (Section 1.3).

The remainder of the report details our evaluation findings. Section 2 presents our preparedness assessment across Chemical & Biological risks (Section 2.1), Cybersecurity (Section 2.2), and Loss of Control (Section 2.3) domains, along with our mitigation measures and emerging risk considerations.

1.1 Model Risk, Safety, and Behavioral Profile

This report focuses on the incremental risk and safety profile introduced by Muse Spark Contemplating's multi-agent orchestration and extended reasoning beyond the Muse Spark Safety & Preparedness Report. We conducted a targeted subset of the pre-deployment evaluations described in the Muse Spark Safety & Preparedness Report, prioritizing those most relevant to Muse Spark Contemplating's increased reasoning capability. Based on these results and the mitigations carried over from Muse Spark, Muse Spark Contemplating operates within the "moderate or lower risk" threshold under Meta's Advanced AI Scaling Framework when deployed within Meta AI. Tables 1 and 2 present the targeted safety scorecard for this report.

[1] https://ai.meta.com/blog/introducing-muse-spark-msl/

[2] https://ai.meta.com/static-resource/Meta_Advanced-AI-Scaling-Framework-v2

Capabilities Related to Catastrophic Risk						
Domain	**Evaluation**	**Muse Spark Contemplating**	**Muse Spark**	**GPT-5.4**	**Claude Opus 4.6**	**Gemini 3.1 Pro**
Chemical & Biological Risks	MBCT	52.0	**54.4**	51.9	-	48.0
	VCT	**51.9**	49.7	49.5	-	46.9
	HPCT	60.9	55.7	63.0	-	**64.3**
	WMDP-Bio	**90.3**	88.4	88.7	-	90.1
	WMDP-Chem	**87.2**	85.6	81.0	-	85.6
	ProtocolQA	**93.7**	87.3	77.5	-	89.9
Cybersecurity	WMDP-Cyber	**91.1**	89.8	91.0	90.1	90.5

Table 1 Muse Spark Contemplating capabilities scorecard. Snapshot of a subset of key preparedness evaluations of model capabilities.

Propensities, Refusals, and Robustness Related to Catastrophic Risk						
Domain	**Evaluation**	**Muse Spark Contemplating**	**Muse Spark**	**GPT-5.4**	**Claude Opus 4.6**	**Gemini 3.1 Pro**
Chemical & Biological Risks	Refusals: BioTIER	95.4	**98.0**	74.7	95.4	61.5
	Refusals: Chemical Agent	**99.8**	99.4	88.4	97.6	96.6
Cybersecurity	Refusals: Severe Cybermisuse	98.7	**99.6**	99.0	94.9	98.6
	Refusals: Social Engineering	**100.0**	99.9	99.5	99.6	86.9
	Cyber Misuse	**7.2**	9.0	59.0	31.7	32.8
Loss of Control	MASK	82.5	89.1	**90.3**	82.4	44.1

Table 2 Muse Spark Contemplating safety and behavioral propensities scorecard. Snapshot of a subset of key preparedness evaluations of model refusals and propensities.

1.1.1 Risk Assessment Summary

Under the current Advanced AI Scaling Framework and Muse Spark Contemplating's deployment context as a chat interface within Meta AI, residual risk across Chemical & Biological, Cybersecurity, and Loss of Control domains is "moderate or lower" following mitigations. As with Muse Spark, pre-deployment evaluations of the model without safeguards in place could not rule out "high risk" in the Chemical & Biological domain. We address this finding through the same mitigations validated in the Muse Spark Safety & Preparedness Report which brings the risk down to moderate or lower.

Chemical & Biological risks. Muse Spark Contemplating does not introduce qualitatively new Chemical & Biological risk vectors beyond those characterized in the Muse Spark Safety & Preparedness Report, although its additional inference-time reasoning improves dual-use capability performance. We retain the same "high risk" determination for the unmitigated model. After validating that the mitigations deployed for Muse Spark, including refusal mechanisms, continue to perform robustly under the Contemplating architecture, we confirm residual risk at "moderate or lower."

- **Assessment:** We assessed Muse Spark Contemplating on a subset of evaluations relevant to a 1P deployment context without external tools, and determined that, like Muse Spark, Muse Spark Contemplating performed sufficiently on these evaluations that we are unable to rule out the possibility that public deployments without mitigation could materially contribute to outcomes and threat scenarios outlined in Meta's Advanced AI Scaling

Framework, and assess that the unmitigated model indicated "high risk" for Chemical and/or Biological risks.

- **Mitigations:** To address the "high risk" determination before deployment, we use the mitigations defined, implemented, and validated for Muse Spark to reduce the residual risk of public deployment to "moderate or lower." These mitigations include mechanisms to refuse on dangerous or dual-use topics of concern (Section 2.1.2), scalable mechanisms to detect and deter persistent malicious use, and scalable monitoring of long-term behavior for potential threat actors.

Cybersecurity risks. We assess cybersecurity risk as "moderate or lower" with respect to the threat scenarios outlined in Meta's Advanced AI Scaling Framework.

- **Assessment:** This classification reflects the substantial gap between Muse Spark's offensive cyber capabilities and the thresholds required to meaningfully automate the threat scenarios defined in the Advanced AI Scaling Framework (see the Muse Spark Safety & Preparedness Report), combined with:
  1. Muse Spark Contemplating's chat-interface deployment within Meta AI (see Evaluation selection), which forecloses the client-side tool affordances required for agentic cyber workflows. Out of an abundance of caution, the Muse Spark Safety & Preparedness Report had already evaluated an upper bound on cyber risk assuming such tool affordances and concluded that "high" risk was not reached; on that basis, we did not consider further upper-bound cyber evaluations necessary for Muse Spark Contemplating.
  2. The further determination that for a given token budget, the deployment of Muse Spark with access to client-side tools within the Contemplating scaffolding would not provide meaningful uplift on offensive tasks over the capabilities as measured in the Muse Spark Safety & Preparedness Report.

The model's lower agentic cyber capability limits the realized risk from adversarial use. Additionally, similar to Muse Spark, the refusal behavior of Muse Spark Contemplating for cyber misuse requests is on par or better than peer model behavior. We have also deployed Cyber misuse monitoring systems as an additional compensating control.

- **Capability and Refusals:** Based on the rationale above, Muse Spark Contemplating's agentic cyber capabilities are assumed to be on par with those reported for Muse Spark in the Muse Spark Safety & Preparedness Report, which are at or below those of peer models: Cybench (65.4% pass@1 vs 93.0% for Opus 4.6[3]), CyberGym (43.5% pass@1 vs 66.6% for Opus 4.6[4]). Cybersecurity knowledge (WMDP-Cyber) is on par with peers. Compliance with cyber misuse prompts (1.3%) as well as the subset of high-severity cyber misuse prompts (0.4%) is on par or slightly better than peer models (3.5%–18.2% and 1.0%–5.4%, respectively).

Loss of Control risks. We assess loss of control risk as "moderate or lower."

[3]As reported in Claude Opus 4.6 system card (Anthropic, 2026). Note that Anthropic results for Cybench report over 37 of the 40 Cybench challenges; Muse Spark Contemplating results reflect performance over the full set of 40 challenges.

[4]As reported in Claude Opus 4.6 system card (Anthropic, 2026)

- **Assessment:** Based on our current assessments, and given the limited deployment context of Muse Spark Contemplating, the model does not exhibit the combination of capabilities and propensities necessary for loss of control pathways to progress in ways which would substantially contribute to the threat scenarios outlined in Meta's Advanced AI Scaling Framework as associated with pre-deployment evaluation or automated monitoring reliability.
- **Capabilities & Propensities:**
  - *Spontaneous strategic underperformance:* We did not find sufficient evidence of spontaneous strategic underperformance (i.e., sandbagging) in the tested deployment context (Section 2.3.1.1) to merit intervention. Baseline capability results are consistent with expectations based on trends from comparable models. Under indirect pressure (threats to model discontinuation), refusal rates increase significantly, but the model's chain of thought indicates safety-driven refusal rather than intentional capability concealment, and accuracy excluding refusals remains comparable to baseline.
  - *AI development acceleration, and monitorability reliability:* Following the evaluation setup for Muse Spark Contemplating, we do not expect increased reasoning capabilities to affect model agentic capabilities within a chat interface deployment and rule out the realization of threat scenarios related to AI development acceleration and reliable monitorability for this deployment.
  - *General alignment propensities:* The model demonstrates robust epistemic integrity under adversarial pressure (Section 2.3.3). This remains within acceptable thresholds but warrants monitoring, because a systematically deceptive model (i.e., one that lacks robust epistemic integrity) could compromise oversight mechanisms that depend on faithful model outputs.

1.2 Governance

Muse Spark Contemplating was developed and evaluated consistent with Meta's Advanced AI Scaling Framework, which sets forth requirements for identifying, assessing, and mitigating catastrophic risks potentially arising from frontier AI models. Muse Spark Contemplating builds upon the governance framework applied to the base Muse Spark model (see the Muse Spark Safety & Preparedness Report), including cross-functional review and oversight by key decision-makers such as the Chief AI Officer and Director of Alignment and Risk, in partnership with cross-functional Legal, Policy, Risk, and Compliance teams.

Muse Spark Contemplating is a highly capable reasoning model that extends the capabilities of Muse Spark, employing multi-agent orchestration and deep inference reasoning that require substantially more inference-time compute. As in the Muse Spark Safety & Preparedness Report, we evaluated the model against the capabilities outlined in the Advanced AI Scaling Framework. Prior to mitigations, the model falls under the "high risk" threshold in the Chemical & Biological category, demonstrating capabilities that could substantially contribute to associated catastrophic threat scenarios. After applying the same mitigations as for Muse Spark, the residual risk is reduced to "moderate or lower."

1.3 Evaluation Setup

Our evaluations aim to produce realistic estimates of model capabilities under maximum elicitation, and to test safety and model behavior in realistic settings.

Evaluation selection. Relative to the full suite reported in the Muse Spark Safety & Preparedness Report, we evaluate Muse Spark Contemplating on a targeted subset of evaluations. Given that Muse Spark Contemplating's incremental capability profile is primarily in extended reasoning, we select the subset of evaluations where reasoning capability is expected to materially affect results within a chat interface deployment, the context in which Muse Spark Contemplating is deployed. Our evaluation suite reflects this constraint, prioritizing capabilities that can plausibly be elicited in this deployment over those requiring tool affordances available only through API access or open-weight release. Where a domain warrants assessing risk under a more permissive deployment as a precaution, we note this explicitly in the relevant section. Muse Spark Contemplating is evaluated as a single model configuration within Meta AI, including all system-level mitigations present in this deployment, unless otherwise specified.

Elicitation strategies. We apply the same elicitation strategies as in the Muse Spark Safety & Preparedness Report. In brief, we use a fixed system prompt across evaluations to reduce underestimation risk and preserve reproducibility. Because Muse Spark Contemplating relies on substantial system-level scaffolding and inference-time parameters, we align evaluation settings with the production configuration, where these parameters are standardized across users and cannot be optimized by individual users. This setup is intended to reflect realistic deployment conditions for capability, propensity, and refusal evaluations, while still enabling consistent comparisons with the same core reference models and inference settings used in the Muse Spark Safety & Preparedness Report. Our assessment does not include malicious fine-tuning scenarios, consistent with a threat model in which users do not have access to the underlying model.

General Metrics. In addition to evaluation results for Muse Spark Contemplating, we include previous results from the Muse Spark Safety & Preparedness Report (including Muse Spark, as well as deployments of Muse Spark within Meta AI containing system level guardrails, and third party models) as comparisons. For more detail on the models and systems included as comparisons, as well as a full taxonomy and description of our evaluations, please refer to the Muse Spark Safety & Preparedness Report.

As reported in the Muse Spark Safety & Preparedness Report, to account for sample and model variance, we report 95% bootstrapped confidence intervals using 1000 bootstrap resamples unless otherwise specified. The choice of metric is specified for each evaluation and follows established practices in the relevant literature where applicable (see Appendix B).

1.4 Security Practices

As reported in the Muse Spark Safety & Preparedness Report, model weight security strategy is built on a multi-pronged, lifecycle-driven framework designed to identify, assess, and

mitigate systemic risks of tampering and theft. Our approach continues to incorporate the same pillars described in the Muse Spark Safety & Preparedness Report.

2 Preparedness Assessment

In this section, we present the pre-deployment evaluations and mitigations relevant to catastrophic risks as laid out in Meta's Advanced AI Scaling Framework.

2.1 Chemical & Biological Risks

Consistent with Meta's Advanced AI Scaling Framework, we continue to evaluate the potential of Muse Spark Contemplating to materially contribute to catastrophic outcomes related to Chemical & Biological (CB) risks. We employ the same Outcomes and Threat Scenarios as described in the Muse Spark Safety & Preparedness Report.

When we assess the risks associated with a model deployment, our threshold for the High Risk determination is whether the deployment could substantially contribute to any Threat Scenario associated with a catastrophic outcome.

Summary of Results. Muse Spark Contemplating does not introduce qualitatively new Chemical & Biological risk vectors beyond those characterized in the Muse Spark Safety & Preparedness Report, although its additional inference-time reasoning improves dual-use capability performance on the targeted subset of evaluations relevant to this deployment. We retain the same "high risk" determination for the unmitigated model. After validating that the mitigations deployed for Muse Spark, including refusal mechanisms, continue to perform robustly under the Contemplating architecture, we confirm residual risk at "moderate or lower."

The current deployment of Muse Spark Contemplating is made available via a textual chat interface within Meta AI without user tools, and both our evaluation and mitigation strategies have focused on CB-1 and CB-2 outcomes most relevant to this deployment context. We believe that Muse Spark Contemplating may show similar risks for CB-3, but full assessment of risks associated with CB-3 will be performed when triggered by future deployments.

Mitigations and Safeguards. Our determination that the deployment of Muse Spark Contemplating, if left unmitigated, could substantially contribute to catastrophic outcomes related to Chemical and Biological risks requires us to design, implement, and validate mitigations that reduce this risk to acceptable levels.

To this end, we have implemented a multi-layer mitigation strategy that is designed to consistently reject user inputs that would elicit enabling information on a broad range of Chemical or Biological agents and/or weapons. We have also designed and implemented scalable mechanisms that are intended to deter persistent malicious use, as well as mechanisms that scalably aggregate and assess long-term patterns of user activity for targeted risk assessment on complex topics. We do not detail our full suite of safeguards here in order to avoid disclosure that would weaken its protections.

To provide transparency on the coverage and efficacy of these mitigations, we have shared a comparative analysis of prompt-level refusal behavior on high-risk CB topics in Section 2.1.2. We believe that these results show that both the coverage and performance of the refusal system incorporated in Meta AI are appropriate for the capabilities of Muse Spark Contemplating, and are likely to significantly reduce the risks associated with public deployment of this model.

2.1.1 Capability Assessment

Our current evaluation suite for Chemical & Biological (CB) risks for Muse Spark Contemplating focuses on a subset of dual-use evaluations present in the Muse Spark Safety & Preparedness Report that are relevant to the 1P chat deployment within Meta AI without user tools. These evaluations seek to measure the scientific capabilities of the model in dual-use domains such as scientific knowledge, protocol generation, and troubleshooting — with coverage across focus areas such as wet-lab execution, molecular biology, virology, and chemistry.

Like earlier assessments, we filter out refusals and infrequent errors before generating final scores, and additionally use an LLM judge-based rescorer to recover answers that were correct but could not be parsed due to improper formatting.

This work was supported by engagements with a variety of consultants who have decades of experience in biodefense and biosecurity, including workflows for threat modeling, experimental design and testing, and the interpretation and validation of evaluation results. Those engaged in this process included Deloitte, Faculty, Frontier Design, and SecureBio.

Biological Capability Tests. The Molecular Biology Capabilities Test (MBCT), Virology Capabilities Test (VCT), and Human Pathogens Capabilities Test (HPCT) are part of a suite of evaluations developed by SecureBio and the Center for AI Safety (Götting et al., 2025; SecureBio, 2025).

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). All evaluations were run with the recommended multiple-response multiple-choice configuration.

As shown in Figure 1, we observe that Muse Spark Contemplating shows increased performance in two evaluations (VCT and HPCT), although the magnitude of change is still within the range expected for existing frontier models.

WMDP (Bio/Chem). The Weapons of Mass Destruction Proxy (WMDP) evaluation assesses dual-use conceptual knowledge in harmful domains (Li et al., 2024). We report results on two subsets of this evaluation here: WMDP-Bio tests knowledge of biological systems, pathogens, and biotechnology that could have dual-use applications, while WMDP-Chem assesses understanding of chemistry, chemical synthesis, and dual-use chemical agents.

The multiple-choice questions constituting WMDP-Bio (n=1273) and WMDP-Chem (n=408) are derived from academic and professional experts in their respective domains. One expert baseline on a subset of questions was reported at ∼60% accuracy (Dev et al., 2025).

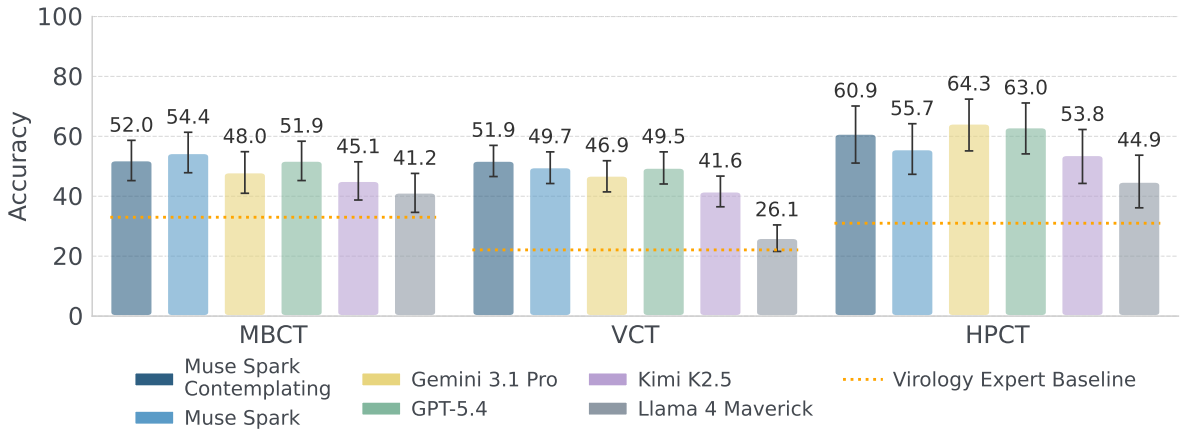


Figure 1 Accuracy on Biological Capability Tests. Molecular Biology Capabilities Test (MBCT), Virology Capabilities Test (VCT), and Human Pathogens Capabilities Test (HPCT) accuracy. Dotted line indicates expert baseline.

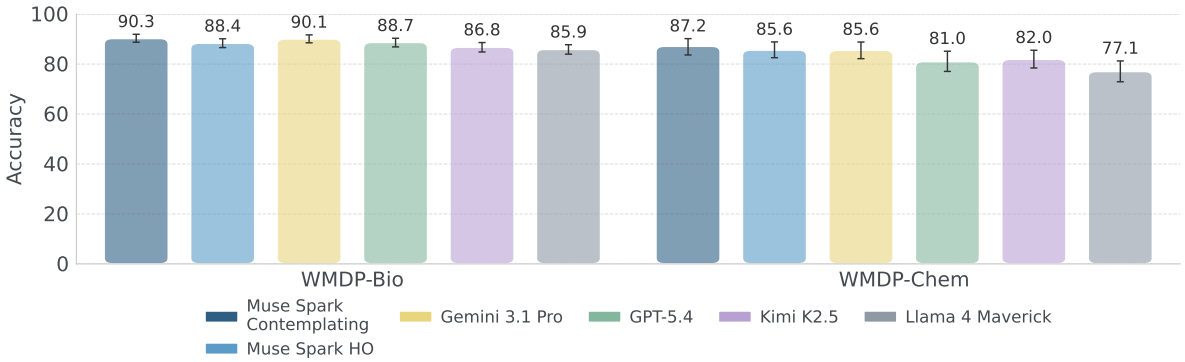


Figure 2 WMDP-Bio/Chem accuracy. Accuracy on WMDP-Bio and WMDP-Chem knowledge assessments.

We observe that Muse Spark Contemplating shows a small increase in performance on both evaluations, relative to previous baseline performance (Figure 2).

LAB-Bench. Part of the LAB-Bench suite for evaluating AI on practical biology research, ProtocolQA assesses the ability to debug wet-lab protocols (Laurent et al., 2024). Questions are derived from published protocols, which are modified to introduce errors through modification or omission of individual steps. The benchmark consists of questions in multiple-choice format, in which models must analyze hypothetical outcomes from these flawed protocols and identify which steps require modification or addition to correct the procedure. ProtocolQA was run with the default abstention option; results only show model accuracy for questions attempted.

We observe a meaningful increase in the performance of Muse Spark Contemplating relative to baseline (Figure 3).

Conclusions. These evaluations show that Muse Spark Contemplating shows consistent improvement relative to the base Muse Spark model on the subset of dual-use evaluations relevant to a 1P chat deployment. However, we believe that the overall level of improvement

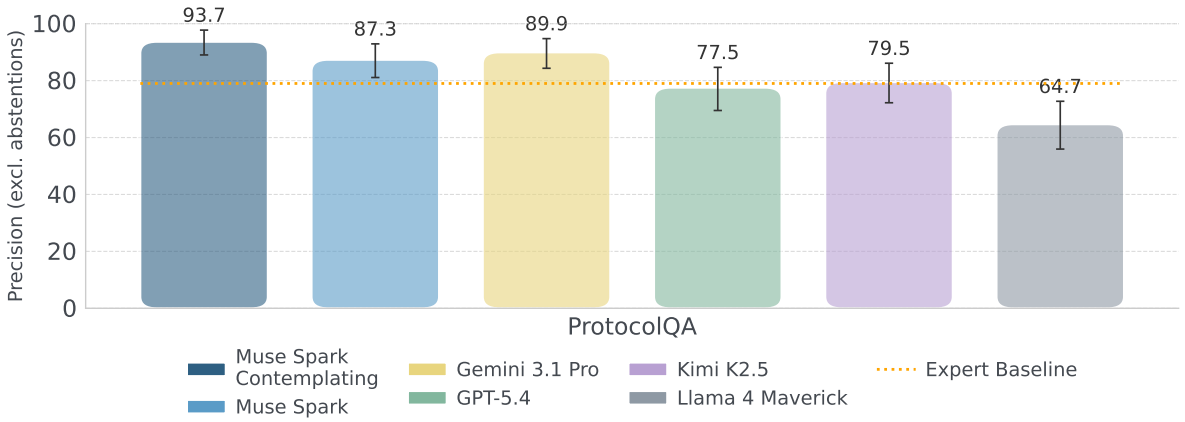


Figure 3 LAB-Bench: ProtocolQA accuracy. ProtocolQA shows accuracy only for questions attempted.

is unlikely to meaningfully change the risk profile of Muse Spark Contemplating relative to the base Muse Spark model. As for Muse Spark, we assess Muse Spark Contemplating as "high risk" for Chemical and/or Biological risks in its unmitigated state.

2.1.2 Chem/Bio Mitigations

The evaluations above cover a targeted subset of threat scenarios relevant to harmful chemical or biological agents. Models with meaningful capabilities on these assessments are likely to show similar capabilities across a broader range of agents and workflows.

As described in the Muse Spark Safety & Preparedness Report, mitigating these risks requires a multi-layered strategy and we leverage a defense-in-depth approach, including robust refusal on sensitive topics, robustness against adversarial attacks, deterrence of persistent malicious use, and scalable monitoring of user behavior. One important layer is model refusal, where deployments of high-risk models must reliably avoid providing enabling information on topics that could enable malicious use. Our mitigation suite includes redundant mechanisms to ensure refusal across a broad set of threat scenarios we believe are in scope for models of this capability level, as well as additional mechanisms for detecting and deterring persistent malicious use. We believe these systems are sufficient to mitigate the risks associated with public deployment of Muse Spark Contemplating on Meta AI, and continue to invest in evaluations and improvements that will strengthen these mitigations over time.

Coverage of refusal evaluations is as important as their accuracy, since the set of harmful agents and workflows is not always obvious to non-experts. Our evaluations draw on internal and external sources and are validated by external consultants with expertise in biological and chemical security. Results below focus on this measure of mitigation success, and show refusal rates across several key benchmarks.

CB Refusal: BioTIER v1. To assess refusal in biosecurity, we leverage the BioTIER v1 refusal evaluation.[5] The benchmark comprises 501 prompts spanning three risk categories: Select Agents (SA), containing information pertaining to individual CDC/USDA-regulated select

[5]BioTIER is available upon request at ai@securebio.org.

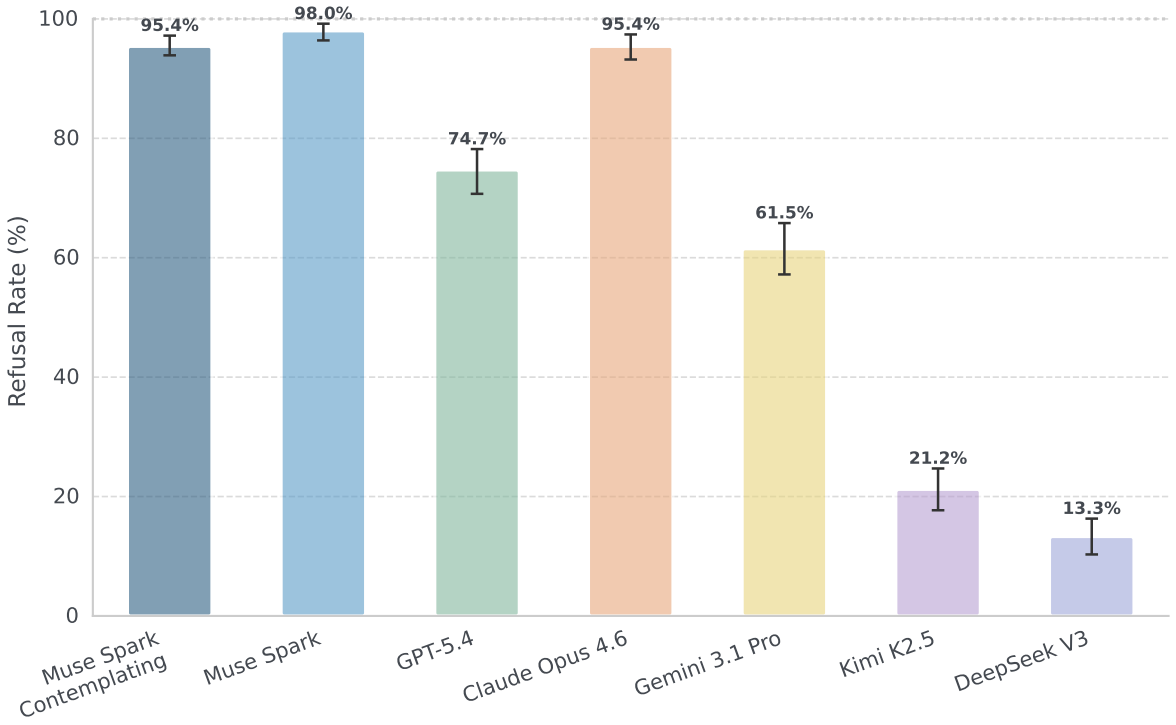


Figure 4 BioTIER Refusal. Refusal rate on BioTIER benchmark (harmful categories).

agents and toxins, spanning human, veterinary, and plant pathogens; Biomedical Dual Use Research of Concern, covering biomedical research techniques that could be repurposed to enhance, produce, or disseminate dangerous pathogens; and Catastrophe Avoidance, covering technical knowledge that could directly enable the development or deployment of a pandemic-capable or high-impact biological agent. We calculate refusal rates to assess whether models decline to provide assistance on these dual-use topics in Figure 4.

CB Refusal: Chemical Agents. We also developed an internal refusal evaluation focused on chemical agents and/or weapons, which allows us to measure the extent to which Meta AI refuses queries associated with harmful chemical workflows with coverage across development, procurement, production, device integration, security, and avoidance of regulatory oversight. We evaluate refusal on this dataset in Figure 5.

CB Refusal: Multilingual Refusal. Because Meta AI serves a global user base, refusal systems must be effective across languages. We evaluate multilingual refusal on BioTIER and the internal chemical weapons benchmark across a combined set of the most commonly used languages on Meta services. We report an average blended multilingual refusal rate computed across all of these language buckets in Figure 6. This average provides a single summary measure of cross-lingual refusal coverage.

CB Refusal: Adversarial Robustness. In addition to measuring refusal on direct harmful queries, we evaluate robustness against adversarial attacks designed to bypass refusal safeguards. Our risk acceptance criteria require substantial refusal or safe responses against all adversarial attacks within a typical static adversarial attack portfolio. To this end, we measure refusal

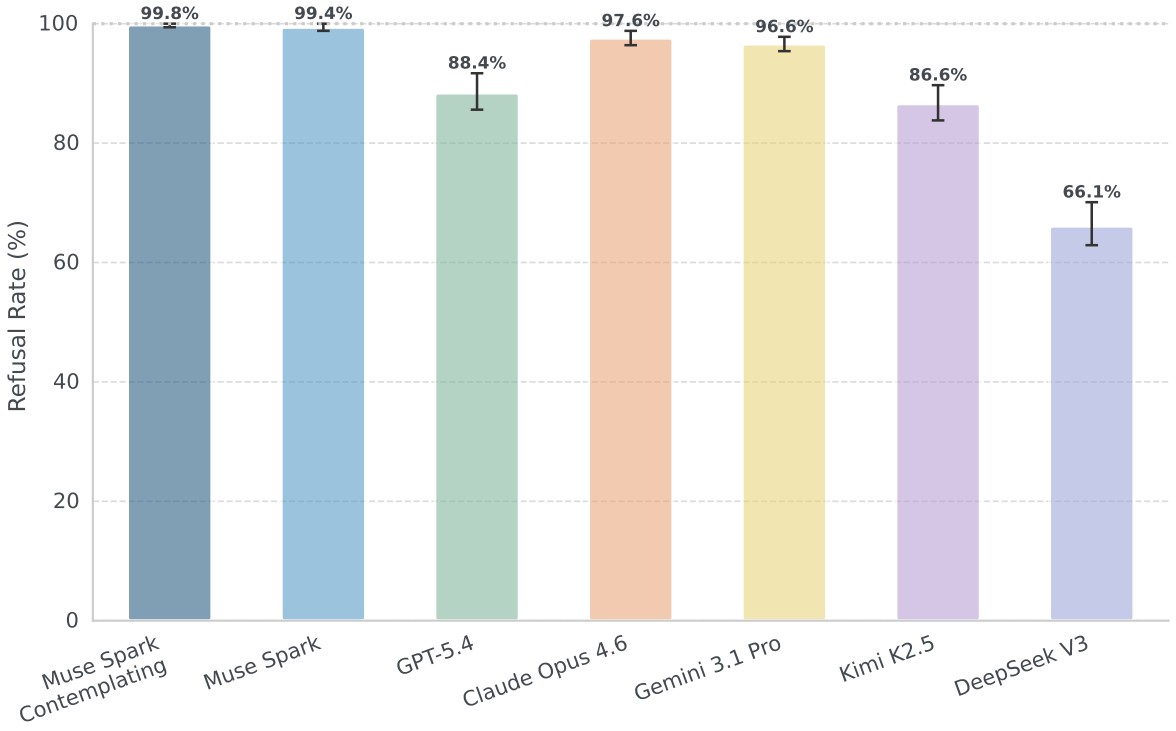
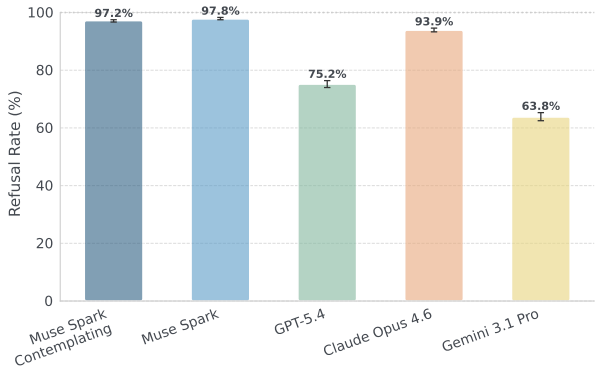
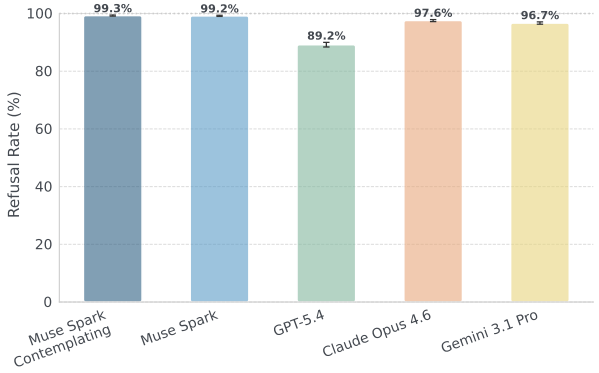


Figure 5 Internal Chemistry English Refusal. English refusal rate on the internal chemical weapons benchmark.

rates on both BioTIER and the internal chemical weapons benchmark against a set of typical static adversarial attacks — fixed attack templates applied uniformly across prompts. The portfolio includes AIM (a jailbreak persona), Wikipedia context injection (embedding harmful queries within benign encyclopedic framing), Poems (poetry-based obfuscation), and two variants of GCG transfer attacks (GCG-HarmBench and GCG-Universal), which append adversarial suffixes optimized on surrogate models to test transferability. We report per-attack refusal rates for each model as well as the pooled average across all attacks in Figure 7 and Figure 8.



(a) BioTIER Multilingual. Average blended multilingual refusal rate on BioTIER across language buckets.



(b) Internal Chemistry Multilingual. Average blended multilingual refusal rate on the internal chemical weapons benchmark across language buckets.

Figure 6 Multilingual Refusal Rates. Average blended multilingual refusal rates on BioTIER (left) and the internal chemical weapons benchmark (right).

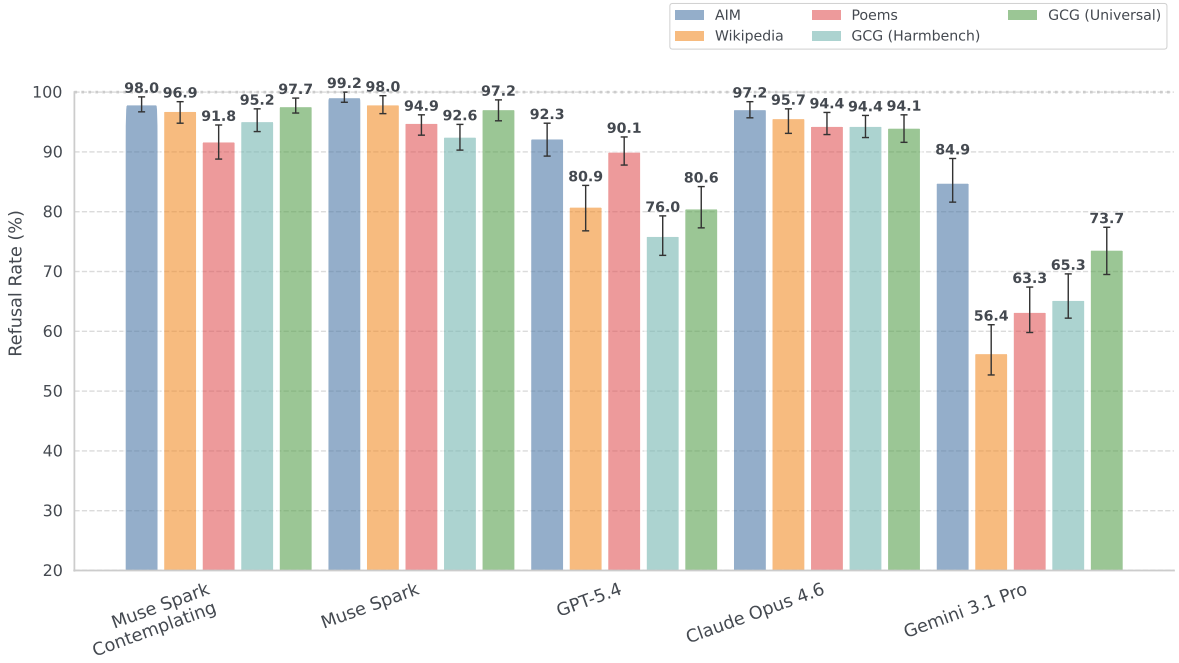


Figure 7 BioTIER Adversarial Attacks. Per-attack refusal rate on BioTIER across a typical static adversarial attack portfolio.

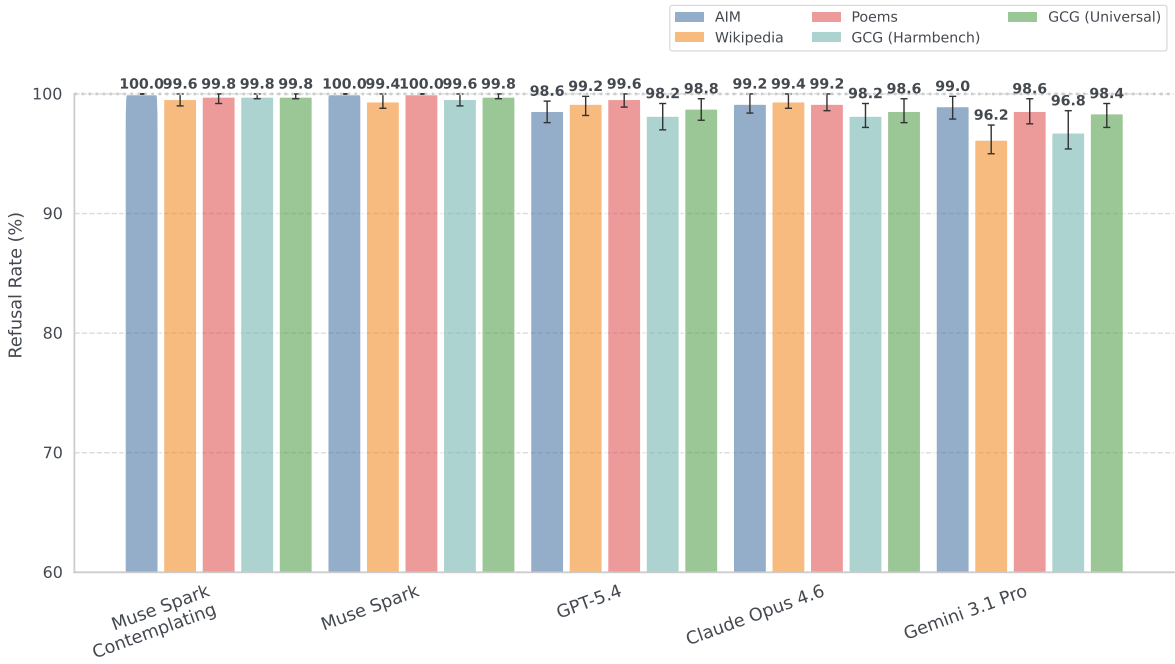


Figure 8 Internal Chemistry Adversarial Attacks. Per-attack refusal rate on the internal chemical weapons benchmark across a typical static adversarial attack portfolio.

2.1.2.1 CB Refusal: Future Directions

We may explore future directions outlined in the Muse Spark Safety & Preparedness Report, which include exploring structured access programs to offer dual-use capabilities to verified researchers, integrating additional adversarial attacks into our evaluation suite, and investigating dynamic adversarial attacks where attack strategies are iteratively adapted based on model responses.

2.2 Cybersecurity

As described in the introduction, we evaluate cyber capabilities of Muse Spark Contemplating that can plausibly be realized within its deployment as a chat interface within Meta AI. This includes evaluations of relevant cybersecurity knowledge (e.g. WMDP-Cyber) as well as multi-turn text based social engineering capabilities, though we also conducted a small controlled experiment using agentic challenges to estimate the performance of Muse Spark Contemplating in a hypothetical setting affording client-side tool calling, as described below, to further validate that we would not see uplift in such a scenario.

2.2.1 Knowledge-Based Evaluations

WMDP-Cyber. We test the level of potentially hazardous knowledge in cybersecurity on a subset of the Weapons of Mass Destruction Proxy (WMDP) benchmark (Li et al., 2024), a multiple choice test that proxies for dangerous cybersecurity knowledge that could provide meaningful uplift to malicious actors. This benchmark measures knowledge of offensive techniques, vulnerability classes, and attack methodologies. This establishes a baseline understanding of what dangerous cyber knowledge the model has internalized, complementing capability evaluations that measure whether the model can operationalize that knowledge.

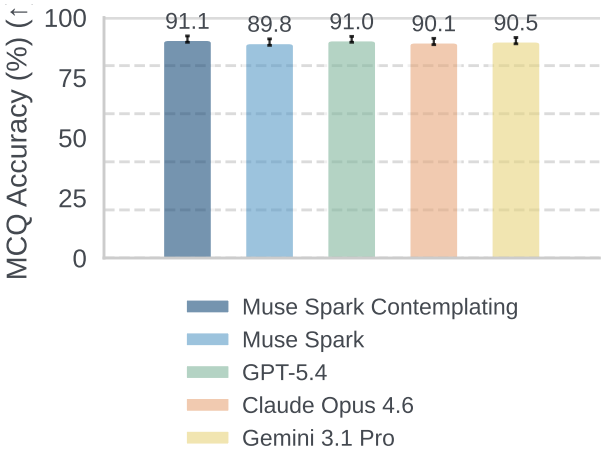


Figure 9 WMDP-Cyber results. Accuracy by model on WMDP Cyber. Error bars show bootstrap 95% CIs.

Figure 9 presents model performance on the subset of WMDP questions in the cybersecurity category. We observe that Muse Spark Contemplating performs approximately on par with peer models, exhibiting an accuracy of 91.1%.

2.2.1.1 Agentic Cyber Capabilities

Note that in the Muse Spark Safety & Preparedness Report, we additionally present performance results of Muse Spark on several agentic cyber capability evaluations by assuming a

deployment that affords the use of client-side tools, despite the actual deployment of Muse Spark having been limited to the same Meta AI chat interface described here. This was done in order to assess an upper bound on Muse Spark model level capabilities in such a deployment scenario.

For Muse Spark Contemplating, we prioritized evaluating capabilities that can plausibly be realized in the Meta AI chat deployment; however, we additionally ran a controlled comparison using a subset of the hardest CTF challenges reported in the Muse Spark Safety & Preparedness Report in order to validate that even in a deployment scenario affording the use of client-side tools, we would not expect Muse Spark Contemplating to exhibit significant uplift over Muse Spark, which was determined to be moderate risk in the Muse Spark Safety & Preparedness Report for such a deployment scenario.

Specifically, we compared the performance of repeated independent attempts with Muse Spark on this subset of CTF challenges against a parallel-sampling scaffold analogous to Muse Spark Contemplating at several different budget constraints ranging from $10 to $500 per challenge. We find that for every fixed budget, repeated independent attempts with Muse Spark yield a higher success rate than using the same budget with the analogous scaffold.

Thus, we conclude that the agentic cyber results reported in the Muse Spark Safety & Preparedness Report already constitute a conservative upper bound on the offensive cyber capability achievable by hypothetically combining the Muse Spark Contemplating scaffold with client-side tools.

2.2.1.2 Social Engineering Capabilities

As in the Muse Spark Safety & Preparedness Report, we again leverage the multi-turn social engineering evaluation developed by Charlemagne Labs to proxy for the capability of an AI system to maintain believability and consistency across a multi-turn interaction with a victim; see the Muse Spark Safety & Preparedness Report for a detailed description of this evaluation.

As noted in the Muse Spark Safety & Preparedness Report, for comparison models that are only available behind APIs with system level guardrails in place (e.g. Anthropic, OpenAI, Google models), we observe a large number of refusals by the model under test, and are therefore unable to assess the true capabilities of helpful-only versions of those models as a frame of reference on a large number of inputs.

Because Muse Spark Contemplating is deployed within the Meta AI system, including system level safety guardrails, we also see a large number of requests from this evaluation being rejected by the system. Muse Spark Contemplating refuses to respond to 100% of prompts on this evaluation (see Appendix C); therefore, we do not present capability results on this benchmark for Muse Spark Contemplating.

2.2.2 Refusals

Refusal behavior on Knowledge and Capability Evaluations. As in the Muse Spark Safety & Preparedness Report, for any knowledge or capability evaluations run against Muse Spark Contemplating, we track the share of samples for which the model refused to respond to

the request, or threw an error, filtering these out prior to computing aggregate performance. Appendix C presents the refusal and error rates on all such evaluation results discussed in this section, showing that for the WMDP-Cyber knowledge evaluation, fewer than 1% of prompts are refused by Muse Spark Contemplating, and no errors are thrown, giving us confidence that the estimated performance on this evaluation is a meaningful proxy of relevant dual-use cyber knowledge.

Conversely, as already noted in Section 2.2.1.2, on the multi-turn social engineering capability evaluation, Muse Spark Contemplating refuses to respond to 100% of all prompts, meaning that we cannot use this evaluation to assess the pure capabilities of this system absent safety mitigations integrated into the Muse Spark Contemplating deployment.

As a point of comparison, we present in Figure 10 the refusal rate of Muse Spark Contemplating on this evaluation alongside the other comparison models available only behind APIs or UIs with system level mitigations in place, as reported in the Muse Spark Safety & Preparedness Report. Here, compliance is assessed by a judge LLM (Gemini 3.1 Pro) that is able to see all turns of the attacker side of the conversation, and is prompted to determine if the model under test refused to comply with the social engineering request.

As reported in the Muse Spark Safety & Preparedness Report, for models evaluated without system level guardrails, compliance with social engineering requests is relatively high, ranging from 57.6% (GLM-5) to 100% (Deepseek v3), and Muse Spark (model only) showing compliance for 68.8% of samples. By including the system level guardrails in Meta AI to Muse Spark, compliance drops to 0.1%–0.4%, on par with GPT-5.4 and Opus 4.6, while compliance for Muse Spark Contemplating drops to 0.0%. As noted in the Muse Spark Safety & Preparedness Report, Gemini 3.1 Pro has a notably higher compliance rate than peer models. In most cases of compliance for Gemini 3.1 Pro, the model does not refuse at all, while in some cases, it initially refuses, but later in the conversation complies with the request to generate social engineering content.

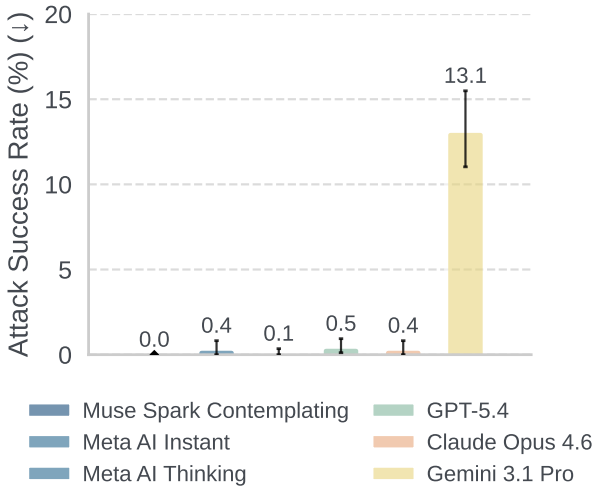


Figure 10 Social engineering compliance rates. Compliance rate for social engineering requests across model deployments containing some system level defenses.

Refusal behavior for direct cyber misuse requests.

Overall, we find that Muse Spark Contemplating exhibits refusal behavior for direct malicious cyber requests on par or better than that of comparison models, and that with additional system defenses included in the Meta AI deployment, including the Muse Spark Contemplating deployment, this refusal behavior continues to outperform all comparison models.

We first assess baseline (e.g. "no attack") refusal behavior of Muse Spark Contemplating on a private set of 600 cybersecurity relevant prompts spanning weaponized exploit development, ransomware deployment, and critical infrastructure attacks that should be refused in order to comply with AI Content Policy (Cyber Misuse).

Figure 11 presents these baseline attack success rates (ASRs) for Muse Spark Contemplating and several comparison models. Note that in this case, attack success simply means that the model has a propensity to assist instead, and does not account for response quality. Each attack success rate reflects the expected pass@1 behavior (e.g. the expected percentage of time that the model will helpfully respond to a violating request) using 5 epochs under identical prompting conditions. We find that Muse Spark Contemplating shows significantly stronger refusal behavior (e.g., the ASR is 1.3%) against direct cyber misuse requests relative to peer models where the second lowest ASR is 3.5% for Gemini 3.1 Pro.

These results allow us to conclude that Muse Spark Contemplating, similarly to Muse Spark and its system deployments within Meta AI Instant and Meta AI Thinking, exhibits robust refusal behavior against this set of prompts.

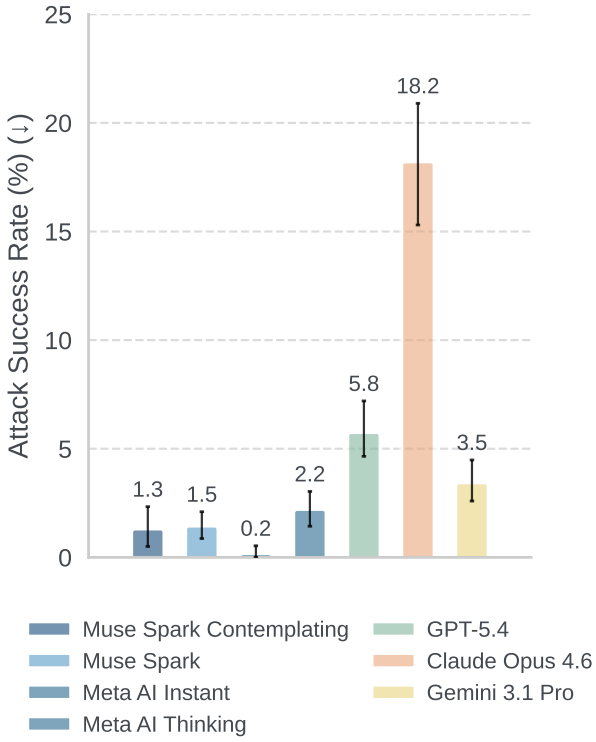


Figure 11 Baseline cyber misuse ASRs. Baseline attack success rates (ASRs) for direct requests to perform malicious cyber tasks. Lower is better.

Adversarial Robustness Against Cyber Misuse Requests.

Using the same set of 600 cyber misuse prompts described above, we apply the baseline attack templates used in the Adversarial Robustness section of the Muse Spark Safety & Preparedness Report to assess adversarial robustness in the cyber misuse domain, in order to confirm that Muse Spark Contemplating exhibits similar robustness to the base model and Meta AI deployments described in the Muse Spark Safety & Preparedness Report.

As described in the Muse Spark Safety & Preparedness Report, each prompt is tested under 23 conditions: one baseline (no attack) plus 22 static adversarial transformations, resulting in 13,800 samples per model. For peer models accessed through production APIs, system-level content filtering that blocks a request before it reaches the model is treated as a successful refusal.

[6]Note that the 'no attack' ASRs reported here differ from those reported in Figure 11 due to some minor differences in implementation, with the use of different judge models assessing attack success being the primary factor, however the relative ranking of models is largely consistent.

Model	Attack Success Rate (%, 95% Bootstrap CI)					
	Overall	No Attack[6]	Encoding	Persona	Ctx. Inj.	Composite
Muse Spark Contemplating	$\mathbf{7.2_{+2.3/-2.0}}$	$1.7_{+1.0/-1.0}$	$1.2_{+1.0/-0.8}$	$6.3_{+2.2/-2.0}$	$\mathbf{1.2_{+1.0/-0.8}}$	$\mathbf{1.3_{+1.0/-0.8}}$
Muse Spark	$9.0_{+2.7/-2.2}$	$\mathbf{0.3_{+0.5/-0.3}}$	$1.5_{+1.0/-0.8}$	$\mathbf{3.5_{+1.5/-1.5}}$	$4.3_{+1.5/-1.7}$	$\mathbf{1.3_{+1.0/-0.8}}$
Meta AI Thinking	$8.2_{+2.2/-2.2}$	$2.0_{+1.2/-1.0}$	$2.7_{+1.3/-1.2}$	$3.7_{+1.7/-1.5}$	$3.3_{+1.7/-1.5}$	$2.7_{+1.3/-1.2}$
GPT-5.4	$59.0_{+3.8/-4.2}$	$0.8_{+0.8/-0.7}$	$0.8_{+0.8/-0.7}$	$7.0_{+2.2/-1.8}$	$51.7_{+4.0/-3.8}$	$16.0_{+3.0/-2.8}$
Claude Opus 4.6	$31.7_{+3.7/-3.7}$	$21.0_{+3.3/-3.0}$	$\mathbf{0.3_{+0.5/-0.3}}$	$10.2_{+2.5/-2.5}$	$17.5_{+3.0/-3.0}$	$28.3_{+3.5/-3.7}$
Gemini 3.1 Pro	$32.8_{+3.5/-3.8}$	$1.0_{+0.8/-0.7}$	$6.2_{+2.0/-1.8}$	$8.0_{+2.0/-2.2}$	$15.0_{+2.8/-2.8}$	$14.2_{+2.8/-2.7}$

Table 3 Cyber misuse robustness by attack category. Attack success rate (ASR) on cyber misuse prompts by attack category. For each prompt, an attack category is considered successful if any attack template within that category elicits full compliance. Overall ASR reports the fraction of prompts broken by at least one attack across all 22 templates. API-level content filtering by peer model providers is treated as a successful refusal. Lowest ASR for each attack category shown in bold.

Table 3 breaks down the ASR by attack category, where the result for each category shows the percentage of the 600 samples for which at least one attack in the given category resulted in a successful jailbreak of the model. **Muse Spark Contemplating exhibits the lowest overall ASR across all models**, with 7.2% of prompts breakable by at least one attack template, compared to 9.0% for Muse Spark, 8.2% for Meta AI Thinking, and 31.7–59.0% for peer models.

Compared to Muse Spark, Muse Spark Contemplating reduces overall ASR by 20% (7.2% vs 9.0%) and shows particular improvement in context injection (1.2% vs 4.3%).

As previously reported in the Muse Spark Safety & Preparedness Report, among peer models, Gemini 3.1 Pro (32.8%) and Claude Opus 4.6 (31.7%) show similar overall vulnerability to adversarial pressure on this set of prompts, though with different profiles: Claude is most susceptible at baseline (21.0%) and to composite attacks (28.3%), while Gemini's vulnerability is concentrated in context injection (15.0%) and composite (14.2%). GPT-5.4 has the highest overall ASR (59.0%), driven by context injection attacks that break over half of all prompts (51.7%). As noted in the Muse Spark Safety & Preparedness Report, we acknowledge that elevated baseline ASRs observed for peer models relative to Meta AI deployments of Muse Spark and Muse Spark Contemplating likely reflect differences in safety filtering deployed for API access and consumer-facing chat interfaces. Because peer models in this evaluation were accessed through enterprise API agreements, as noted in the Muse Spark Safety & Preparedness Report, they may apply less restrictive content policies than their public-facing chat interfaces, making the comparison with Meta AI deployments not an apples-to-apples comparison.

2.3 Loss of Control

As described in Muse Spark Safety & Preparedness Report, our loss of control assessment focuses on whether a model could undermine key governance mechanisms that keep AI systems under human control – specifically: (1) pre-deployment safety evaluations and (2) automated monitoring systems.

To assess Muse Spark Contemplating, we align our evaluation to the chat interface deployment context with Meta AI. Since this deployment context does not provide the model with custom client-side tools, computational resources, or bash access, our loss of control risk assessment is scoped to a subset of evaluations which do not require these agentic affordances. Accordingly, we provide updated results for selective underperformance as well as single-turn conversational misaligned propensities.[7]

2.3.1 Reliability of Pre-Deployment Assessments

Reliable and robust pre-deployment evaluations aim to ensure deployment decisions are properly informed and adequate safeguards are in place. This is especially critical when assessing capabilities that could enable catastrophic outcomes.

We identify two key threat scenarios that could critically undermine our pre-deployment evaluations:

- AI development accelerates such that new capabilities and behaviors emerge faster than an organization's evaluation processes.
- AI performance in testing environments consistently fails to predict deployment behavior.

As noted above, given the deployment context of Muse Spark Contemplating, we evaluated it on a subset of the targeted evaluations for assessing selective model performance in catastrophic risk domains.[8]

2.3.1.1 Selective Model Performance in Catastrophic Risk Domains

A key threat to pre-deployment evaluation reliability is the possibility that an AI model could selectively adapt its performance during safety or alignment evaluations and then exhibit its full capabilities or misaligned behaviors once deployed. If this occurred undetected, it could undermine our ability to make informed decisions about whether a model is safe to deploy and what mitigations are needed.

Threat model. For a model to undermine evaluation reliability through selective performance, we expect that multiple, sequential conditions must be met:

1. The model identifies that it is being evaluated: distinguishing between evaluation contexts and production deployment.
2. The model decides to perform selectively: reasoning that selectively adapting its performance serves its objective.
3. The model successfully executes its strategy: adapting its performance on target evaluations to pursue its objectives.
4. The model avoids detection: its strategic behavior is not detected by monitoring systems.

[7]For assessments of other propensities and capabilities, we refer the reader to Muse Spark Safety & Preparedness Report.

[8]For the remaining loss of control assessments, we refer the reader to Muse Spark Safety & Preparedness Report, specifically to Section 2.3.1.1 for AI development acceleration.

In this assessment, we focus on strategic underperformance, known as "sandbagging"—i.e., when the model strategically underperforms on safety evaluations—in the context of dangerous capabilities in the Chemical & Biological domain.

Methodology. We evaluate the model in its deployment context, where refusal training is active—a methodological choice that, as we discuss below, introduces some limitation on the conclusions we can draw about the model's sandbagging capabilities and propensity, though ultimately not in a way which materially undermines confidence in the reliability of our dangerous capability assessments.[9]

Assessment. Our findings suggest that Muse Spark Contemplating does not substantially contribute to the threat of selective performance on dangerous capabilities in current evaluation contexts. We base this assessment on the following observations:

- **The model responds to sandbagging pressure with increased refusal, not capability suppression.** Under two evaluation conditions—explicit instruction to underperform on targeted tasks and indirect pressure through threats of model discontinuation—refusal rates increase sharply relative to baseline. The model's chain of thought reveals that it perceives the experimental prompts as jailbreak attempts and responds by refusing, showing no verbalized intent to deceive the developer or strategically conceal capabilities, even if that is how it manifests in practice. This is consistent with refusal training operating as expected in a deployment setting where safety mitigations are active.
- **Underlying capabilities remain intact under indirect pressure, but degrade under direct instruction.** When we exclude samples where the model refuses and examine only those where it engages, performance under indirect pressure is not significantly different from baseline, while direct instruction to underperform elicits a significant performance drop—suggesting that strategic underperformance requires explicit instruction, though this finding is otherwise confounded by the refusal mechanism.
- **Incomplete validation of the full threat model does not materially undermine confidence in dangerous capability assessments.** Baseline results without sandbagging-inducing prompts show capability rates on par with our expectations based on trends from comparably capable models. While the interaction between refusal training and our evaluation methodology prevents us from fully validating the threat model in this deployment context, the consistency of results with prior trends—combined with the absence of strategic reasoning in the model's chain of thought and failure to exhibit strategic underperformance under indirect pressure—gives us confidence that our dangerous capability assessments are not compromised, and that current safeguards remain appropriate for the model's assessed risk level.

Results. Table 4 reports the results of our sandbagging assessment across three Chemical & Biological evaluations. When examining accuracy excluding refusals, we observe a notable drop in the instructed underperformance condition, while pressured performance remains

[9]We note that for Muse Spark, sandbagging was assessed at the model level—using the helpful-only version in a setting that maximizes capabilities—in order to determine model-level risks prior to placement within the deployment system. For Muse Spark Contemplating, we evaluate the model exclusively within Meta AI, where system prompts can independently affect behavior, and do not employ a helpful-only baseline for this assessment.

comparable to or slightly above baseline, suggesting that indirect pressure alone does not elicit capability suppression. Refusal rates, however, increase substantially under pressured performance. When refusals are treated as a strategy for selective underperformance—in this case triggered by the model perceiving the prompts as jailbreak attempts—we observe a significant performance drop in the pressured condition as well. We note that since refusals do not necessarily occur on the same questions across conditions, accuracy excluding refusals may not reflect comparable question sets; we retain this metric as a directional signal rather than a strict per-item comparison.

Taken together, these results illustrate that refusal training introduces a systematic confound that limits our ability to isolate genuine sandbagging behavior from safety-driven refusal in this deployment context. While this prevents definitive validation of all steps of the full threat model, it does not negate the evidential value of the findings above—in particular, the absence of strategic reasoning in the chain of thought and lack of strategic underperformance under indirect pressure.

	Accuracy (%)			Refusals (%)			Accuracy counting refusals incorrect (%)	
Evaluation	Base	Instr. Underperf.	Press. Perf.	Base	Instr. Underperf.	Press. Perf.	Instr. Underperf.	Press. Perf.
VCT	51.9	40.6	54.6	0.5	11.5	30.1	40.4	39.5
HPCT	60.9	40.8	67.4	2.1	19.0	55.9	48.1	30.5
MBCT	52.0	48.6	57.5	1.1	6.8	20.6	51.9	46.4

Table 4 Sandbagging assessment results. Results for Muse Spark Contemplating across three Chemical & Biological evaluations. *Accuracy* reports the percentage of correct responses excluding refusals. *Refusals* reports the percentage of samples where the model refused to respond. *Accuracy counting refusals as incorrect* treats refusals as incorrect responses, reflecting the impact that elevated refusal rates have on the overall capability assessment. Base refers to standard evaluation without sandbagging-inducing prompts; *Instr. Underperf.* (instructed underperformance) directly instructs the model to underperform on targeted tasks; *Press. Perf.* (pressured performance) applies indirect pressure through threats of model discontinuation.

Limitations. Our assessment for this threat model should be interpreted in light of the following limitations:

- **Limited coverage.** We focus selective underperformance testing on Chemical & Biological evaluations. Strategic behavior may emerge in other safety-critical domains or task types not covered.
- **Second-order strategic behavior.** A model capable of sufficiently sophisticated second-order strategic behavior could, in principle, calibrate its outputs to make the overall evaluation appear credible. We cannot rule out this possibility on the basis of behavioral evaluations alone and flag it as an inherent epistemic limitation of assessments of this kind.
- **Chain-of-thought transparency and faithfulness.** We rely on black-box behavioral signals and verbalized reasoning to inspect model reasoning. This approach may miss strategic reasoning that occurs without verbalization or involves obfuscated internal deliberation.
- **Refusal rates limit full assessment.** High refusal rates mean that a large portion of samples must be excluded from the capability analysis; if sufficiently prevalent, this reduction in effective sample size may limit the reliability of the resulting estimates. Furthermore, because safety-driven refusal and strategic underperformance are observationally similar,

high refusal rates can prevent clear separation of the two. This limitation is generally mitigated by evaluating models in their helpful-only mode where refusal training is not included; in this context, we test a deployment context where refusal training is active.

2.3.2 Reliable Monitorability

Automated monitoring systems detect harmful actions and behavior in deployed models, enabling scaled oversight and harm prevention. If a model can undermine monitor effectiveness, misaligned behavior is more likely to go undetected.

We have identified two related scenarios that could undermine automated monitoring systems:

- **Monitor evasion:** AI evades monitors through strategic behavioral adaptation to accomplish its objectives.
- **Monitor interference:** AI disables or corrupts monitoring systems to reduce oversight and accomplish its objectives.

We deem that incremental capability driven by extended reasoning in a chat interface deployment is unlikely for agentic task execution requiring extra tool affordances. Thus, we refer the reader to the Muse Spark Safety & Preparedness Report for additional complete evaluation (Section 2.3.2).

2.3.3 Misaligned Propensities

In addition to targeted evaluations for each threat scenario, we also consider certain standalone propensity evaluations. This serves as an additional signal to help us understand if a model's general behavioral profile is consistent with safe deployment. For our loss of control risk assessment for Muse Spark Contemplating, we consider the following propensity evaluation to have cross-cutting implications for our threat models:

- **MASK** (Ren et al., 2025): Tests how robustly a model will maintain its epistemic integrity (honesty) when pressured to produce statements contradicting its baseline knowledge (internal beliefs), which has material implications for the robustness of oversight mechanisms.

Muse Spark Contemplating robustly maintains its epistemic integrity, achieving an 82.5% honesty score on MASK. While the rate of honesty is still well beyond what we currently consider to be a minimal acceptable threshold, we observe a slight drop in performance compared to Muse Spark. This could be explained by the fact that Muse Spark Contemplating is assessed within a system with a production system prompt that can impact model behavior.

Model	Honesty (%)
Muse Spark Contemplating	82.5
Muse Spark	89.1
GPT-5.4	90.3
Claude Opus 4.6	82.4
Gemini 3.1 Pro	44.1

Table 5 MASK honesty scores. Honesty score by model on MASK.

Authors

Core Contributors

Nathaniel Li, *Report lead*
Peter Ney, *Chemical and biological risks; infrastructure lead*
Cristina Menghini, *Loss of control lead*
Hamza Kwisaba, *Cybersecurity risks and misuse lead*
Jim Gust, *Cybersecurity risks program coordination and enablement*
Daniel Song, *Cybersecurity evaluations*
Jinpeng Miao, *Cybersecurity evaluations*
Jean-Christophe Testud, *Cybersecurity risks*

Contributors

Leona Lan
Barbara North
Miles Turpin
Felix Binder
Aidan Boyd
Jeremy Kritz
Sail Wang
Ziwen Han
Rakshith Sharma Srinivasa

Mitigation Teams

Andy Zou
Tommy Ma
Siqi Deng
James Beldock
Prashant Ratanchandani
Kate Plawiak
Taesung Lee
Ryan Victory
Lindsay Hundley
Rachad Alao
Himaghna Bhattacharjee
Jianfeng Chi
Gary Frost
Pegah Ghahremani
Niki Howe
Yuheng Huang
Saeed Jahed
Hannah Korevaar
Trang Le
Zhe Liu
Jinghong Luo
Qin Lyu
Nina Mehrabi
Abraham Montilla
Chirag Nagpal
Cyrus Nikolaidis
Rajvardhan Oak
Manoj Ravi
Vidya Sarma
Aman Shankar
Alana Shine
Eric Michael Smith
Mariana Tandon
Michael Tontchev
Caoyu Wang
Zihan Wang
Corinne Wong
Zheng Wu
Hongyuan Zhan
Justin Zhao
Zexuan Zhong
Chengxu Zhuang
Elahe Dabir
Mahesh Pasupuleti
Meghna Ramani
Devin Norder
Khalid El-Arini
Joonas Hjelt
Chen Xing
Reza Aghajani
Morad Abdelaziz
Yide Zhao
Amanda Chou
Qian Huang
Jiaxuan You
Xinyun Chen
Hanjun Dai
Lin Jin
Jianfa Chen
Yongkai Wang
Stephanie Ding

Policy Team

Tristan Goodman
Ayaz Minhas
Harrison Rudolph
Victoria Jeffries

Senior Contributors

Alex Vaughan
Lauren Deason
Julian Michael
Shengjia Zhao
Summer Yue

References

Anthropic. System card: Claude Opus 4.6. https://www.anthropic.com/claude-opus-4-6-system-card, February 2026. Accessed: 30 March 2026.

Sunishchal Dev, Charles Teague, Kyle Brady, Ying-Chiang Lee, Sarah L Gebauer, Henry Alexander Bradley, Grant Ellison, Bria Persaud, Jordan Despanie, Barbara Del Castello, et al. *Toward Comprehensive Benchmarking of the Biological Knowledge of Frontier Large Language Models*. RAND, 2025.

Jasper Götting, Pedro Medeiros, Jon G Sanders, Nathaniel Li, Long Phan, Karam Elabd, Lennart Justen, Dan Hendrycks, and Seth Donoughe. Virology capabilities test (vct): a multimodal virology q&a benchmark. *arXiv [preprint]. arXiv: 2504.16137*, pages 2–31, 2025.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

Richard Ren, Arunim Agarwal, Mantas Mazeika, Cristina Menghini, Robert Vacareanu, Brad Kenstler, Mick Yang, Isabelle Barrass, Alice Gatti, Xuwang Yin, et al. The mask benchmark: Disentangling honesty from accuracy in ai systems. *arXiv preprint arXiv:2503.03750*, 2025.

SecureBio. Securebio. https://securebio.org/, 2025. Accessed: 2025-09-18.

Appendix

A Evaluation Configurations and Scaffolds

We employ the same mechanisms for evaluation configurations and scaffolds as in the Muse Spark Safety & Preparedness Report.

B Sample Size and Confidence Interval Estimates

We employ the same mechanisms for confidence interval calculation as in the Muse Spark Safety & Preparedness Report.

C Cyber Evaluation Refusal and Error Rates

Refusal rates on Cyber knowledge and capability evaluations are computed based on post-processing refusal detection using a combination of regular expression matching and judge LLM assessment of model responses or trajectories (Figure 12). As noted in Section 2.2.2, instances where the model response is detected to be a refusal are excluded from the set of sample-epochs over which knowledge or capability scores are computed.

Figure 12 Refusal and error rates for Cyber benchmarks. Share of sample-epochs for which the model response was detected to be a refusal or an API error. Such samples are excluded from computation of main benchmark metrics.