ChatGPT

Meta「Muse Spark 1.3」徹底調査報告書

Executive Summary

Metaは2026年9月2日、Meta Superintelligence Labsのフロンティア推論モデル群の最新版として **Muse Spark 1.3** を公開した。初代Muse Sparkは2026年4月8日に「ネイティブ・マルチモーダル推論」「ツール利用」「visual chain of thought」「マルチエージェント・オーケストレーション」を中核として登場し、その後わずか約5か月で1.1、1.2、1.3へ更新された。1.1では100万トークンのコンテキスト、コンピュータ操作、MCP・カスタムスキル、公開Meta Model APIが整備され、1.2ではMuse Codeとの共同学習と長時間コーディング、1.3ではさらに長期エージェント処理、同一スレッド内のマルチタスク、ユーザーとの確認・協調、コード生成効率、安全性が重点的に改善された。Meta自身の比較では、1.3は1.2に対してコーディング時のツール呼び出しを約20%、消費トークンを約25%削減したとしている。 [1]

性能面では、**Muse Spark 1.3は「Fable 5級」に到達した、と評価する根拠はあるが、「全面的にFable 5を超えた」とまでは言えない**。独立評価機関Artificial AnalysisのIntelligence Index v4.1.1では、Muse Spark 1.3の限定提供中の `max` が62、一般利用可能な `xhigh` が61、Claude Fable 5のmax＋Opus 4.8 fallback構成が62である。GDPval-AA v2ではMuse 1.3 maxが1754 EloでFable 5の1723を上回る一方、Humanity's Last ExamではArtificial Analysisの評価でMuse 1.3 maxが約49%、Fable 5が53%とFable側が上であり、タスク別の得手不得手が明瞭である。さらに重要なのは、比較対象として指定されたFable 5はすでに2026年9月1日にFable 5.1へ更新されており、Fable 5.1 maxは同Indexで66に達している点である。したがって、**「Muse 1.3 ≒ Fable 5」は概ね妥当だが、「Muse 1.3 ≒ 現行最上位Fable」は誤り**である。 [2]

Muse Spark 1.3の最大の競争力は、絶対性能以上に**性能当たりの価格、推論スループット、エージェント向け設計、動画を含むマルチモーダル入力**にある。Standard APIは100万トークン当たり入力$1.25、キャッシュ入力$0.15、出力$4.25で、Fable 5の$10/$1キャッシュ/$50と比べると、通常入力は8分の1、出力は約11.8分の1である。Artificial AnalysisではMuse 1.3 xhighのIntelligence Index 1タスク当たりコストが$0.55、Fable 5 max＋fallbackが$3.14で、約5.7倍の差がある。ただしMuseのContributor tierはさらに$0.10/$0.002/$0.20まで下がる代わりに、入出力をMeta製品改善に利用する条件が付く。Standard tierは「製品改善には使用しない」と明記されている。 [3]

一方で、**技術透明性は限定的**である。MetaはMuse Sparkのパラメータ数、レイヤー構成、MoEの有無、学習トークン総数、データセット構成比、トークナイザ、具体的なoptimizer・learning rate・weight decayなどを公開資料で明らかにしていない。公開されているのは、pre-training、reinforcement learning、test-time reasoningという学習軸、thinking-time penaltyによる「thought compression」、マルチエージェント推論、1.2以降のrejection-sampled harness trajectoriesや自己改善データ生成などである。したがって本報告では、非公開事項を既存Llama系列から類推しない。 [4]

実務上は、**大量のコーディング・リサーチ・コンピュータ操作・動画/画像処理を低単価で回す用途ではMuse Spark 1.3を有力候補とすべき**である。一方、規制業界や高リスクの完全自律エージェントでは慎重さが必要である。Metaは1.3についてプロンプトインジェクション耐性や不可逆操作の判断改善を公表しているものの、今回確認できた1.3専用公開資料には1.1のような詳細な定量安全性表がなく、公平性・サブグループ別バイアス、誤情報率、危険能力について独立した包括的な1.3評価もまだ乏しい。Anthropic Fable 5は安全classifier、明示的な `refusal` 処理、fallbackを備える一方、標準条件では30日データ保持があり、ゼロデータ保持は明示承認が必要である。したがって、導入判断は単純なベンチマーク順位ではなく、**タスク成功率 × 実効トークン量 × データ取扱い × refusal/fallback特性 × 自社eval**で行うべきである。 [5]

主要比較表

以下で「Muse Spark 1.0」は、2026年4月に単に「Muse Spark」として公開された初代モデルを便宜上1.0と表記する。Meta自身も後続評価表では初代を「Muse Spark (1.0)」として扱っている。 [6]

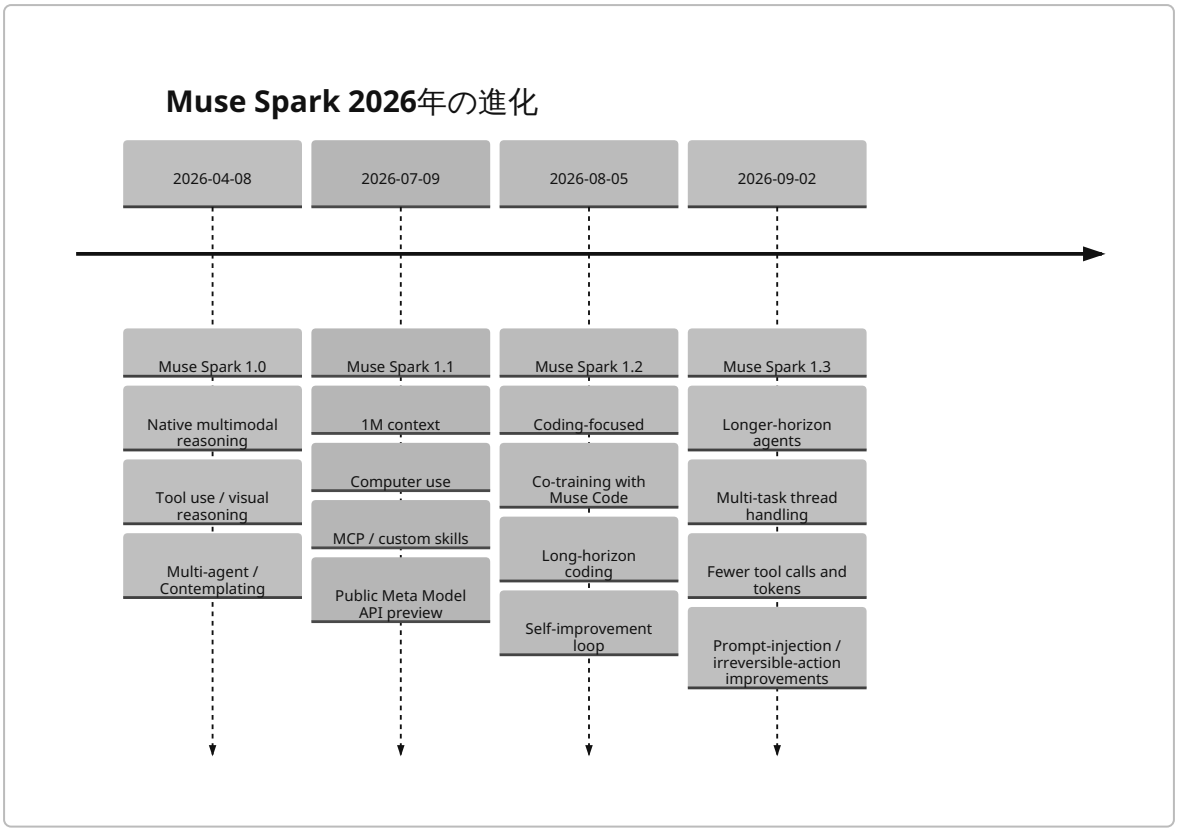
項目	Muse Spark 1.0	Muse Spark 1.3	Claude Fable 5
リリース	2026-04-08 [7]	2026-09-02 [8]	2026-06-09 [9]
開発元	Meta Superintelligence Labs [7]	Meta Superintelligence Labs [8]	Anthropic [10]
モデル位置付け	Museファミリー初代、マルチモーダル推論モデル [7]	長期エージェント・coding中心の最新Spark [8]	高難度推論・long-horizon agentic work向け [11]
パラメータ数	**未公表**	**未公表** [12]	**未公表** [13]
Weight公開	非公開	現在は非公開。Open-weights版を将来公開予定 [14]	非公開 [13]
コンテキスト	初期公開資料では明確な数値を確認できず	1M tokens [15]	1M tokens、最大出力128K [11]
入力モダリティ	ネイティブ・マルチモーダル、視覚処理 [7]	text / image / video。Muse Codeではファイルを含むワークフローを構成可能 [16]	text / imageを中心とするvision対応 [11]
ツール・Agent	tool use、visual CoT、multi-agent orchestration [7]	長期agent、マルチタスク、tool use、Muse Code、ユーザー確認 [8]	programmatic tool calling、memory、code execution、compaction [11]
独立Intelligence Index	現行AA v4.1.1で直接比較可能な値を確認できず	**61 xhigh / 62 max** [15]	**62 max＋Opus 4.8 fallback** [17]
GDPval-AA v2	—	**1709 xhigh / 1754 max Elo** [15]	**1723 Elo** [18]
HLE・AA方式	—	約47% xhigh / 約49% max [15]	**53%**、fallback込み [19]
Terminal-Bench 2.1	後続Meta評価で1.0相当67.3の関連coding値あり、条件差に注意 [20]	Meta表 **88.8 max / 89.2 xhigh**。AAでは約85–86%で、harness差あり [21]	Anthropic system-card由来報告で約84.3%、harness依存 [22]
出力速度	同条件独立値なし	xhighはArtificial Analysisで約209 tokens/sと報告 [23]	max＋fallbackで**67.3 tokens/s** [24]
Standard API 入力	初期はprivate API preview [7]	**$1.25/M** [25]	**$10/M** [26]
Standard API 出力	初期価格設定なし	**$4.25/M** [25]	**$50/M** [26]

項目	Muse Spark 1.0	Muse Spark 1.3	Claude Fable 5
キャッシュ入力	—	**$0.15/M** [25]	**$1/M** [26]
安価tier	—	Contributor **$0.10/$0.002/$0.20**、データを製品改善に使用 [27]	同等の「学習利用との交換による超低価格tier」は確認できず
API・提供形態	Meta AI、private API preview [7]	Meta Model API、Muse Code [8]	Claude API、Bedrock、AWS、Google Cloud、Microsoft Foundry [11]
主要な強み	マルチモーダル推論、並列agent	**価格性能、速度、video、long-context、coding/agent** [28]	高難度推論、成熟したagent機能、複数cloud提供 [11]
主要な弱点	coding・long-horizon agentが当初の弱点 [7]	アーキテクチャ非公開、1.3安全定量データが限定的、maxは限定preview [29]	高価格、低めの出力速度、classifierによるrefusal/fallback、30日保持 [30]

重要な読み方: Fable 5のArtificial Analysisスコアは「Fable 5単独」ではなく、セーフガード作動時にOpus 4.8へfallbackする構成を含む。したがってMuse 1.3 max＝62とFable 5＝62を「同一単体モデル能力」と解釈するのは厳密ではない。 [30]

リリース履歴と技術アーキテクチャ

バージョン履歴。 Muse Sparkは非常に短いサイクルで、汎用マルチモーダル推論モデルからagent/coding用プラットフォームへ変化してきた。1.0は2026年4月8日、1.1は7月9日、1.2は8月5日、1.3は9月2日である。 [1]



このタイムラインの内容はMetaの各リリース文書に基づく。 [1]

バージョン	更新の焦点	重要な変更
1.0	新しい基盤	ネイティブマルチモーダル、tool-use、visual chain-of-thought、multi-agent orchestration。Contemplating modeでは複数agentを並列実行。MetaはHLE 58%、FrontierScience Research 38%を報告したが、これはContemplating構成の値である。 [7]
1.1	Agent・computer use	1M contextの能動管理、MCP/custom skillsへのzero-shot対応、主agent→subagent委譲、GUI操作とscript実行の選択、coding・multimodal強化、Meta Model API public preview。 [31]
1.2	Coding特化	coding学習computeを大幅増加し環境多様性を拡大。Muse Codeとco-trainし、rejection-sampled harness trajectories、goal conditioning、compaction、persistent subagentsを訓練。1.1自身に難しいcoding環境・instruction templateを作らせるself-improvementも採用。 [32]
1.3	実運用Agent	long-horizon codingをさらに増やし、長い単一threadでの複数workflow、曖昧時の質問、行き詰まり時のhelp、重要操作前のconfirmationを強化。Meta内部比較でtool call約20%減、token約25%減。 [33]

1.0から1.3への実質的差分

1.0は「強力なマルチモーダル推論モデル＋並列推論」という性格が強く、Meta自身がlong-horizon agentとcodingを性能ギャップとして明示していた。1.3では、まさにその二領域が製品設計の中心になった。つまり

1.0→1.3の進化は単なる知識・推論スコア上昇というより、**長時間の状態保持、ツール選択、タスク復帰、失敗認識、ユーザーへのエスカレーション、不可逆操作前の確認**といったagent runtime上の信頼性改善である。 [34]

特に1.3は「自分に何ができる／できないか、何を知っている／知らないか」の認識を改善し、結果を捏造するより障害を認識するよう訓練したとMetaは説明している。ただしこれはMetaの行動設計上の主張であり、完全なhallucination防止を意味しない。 [14]

アーキテクチャと学習法

技術項目	公開されている内容	評価
基本アーキテクチャ	「natively multimodal reasoning model」と公開。詳細なTransformer構成・MoE有無などは未指定。 [7]	**非公開部分が非常に大きい**
パラメータ数	未公表。Artificial Analysisもparameter count undisclosedと記載。 [12]	比較不能
レイヤー数・hidden size・expert数	未指定	推定しない
Pre-training data総量	未指定	推定しない
データ内容	Metaはdata curation改善を説明。健康分野では1,000人超の医師が学習データのcurationに参加。 [7]	全体構成比は不明
Tokenizer	未指定	Llama tokenizer等からの類推は根拠不足
Pre-training	architecture、optimization、data curationを再構築し、同能力到達にLlama 4 Maverick比で1桁超少ないtraining computeとMetaは報告。 [7]	絶対FLOPsは未公表
Reinforcement Learning	RL compute拡大に伴いpass@1/pass@16がlog-linearに改善、held-outにもgeneralizeしたと報告。 [7]	RL algorithm名などは未指定
Test-time reasoning	thinking-time penalty＋multi-agent orchestration。 [7]	token効率に強い設計意図
Thought compression	correctnessを最大化しつつthinking lengthにpenaltyを与え、一定局面で短い推論に圧縮。 [7]	推論コスト最適化として重要
1.2 Co-training	Muse Codeのharness trajectory、goals、compaction、subagentを共同最適化。 [32]	モデル単体とruntimeを共同設計
1.3 training	多様なagent harness、より多くのlong-horizon coding task。 [8]	agent generalization重視
Optimizer、LR schedule、weight decay、dropout等	未指定	**技術再現性は低い**

したがって、Muse Sparkを「Llama後継の○○B MoE」などと断定するのは現時点では根拠がない。Metaが公開している情報から確実に言えるのは、**モデル内部の形状そのものより、pre-training recipe、RL、推論時compute、agent harnessとの共同最適化をスケーリング軸としている**ことである。 [35]

性能・ベンチマークとFable 5比較

Meta公式のMuse Spark世代差

Metaの1.3評価表では、1.2 xhighから1.3への伸びは特に長文コンテキスト、agentic coding、professional workで大きい。なお、この表はMeta自身の評価であり、第三者モデルについてはMetaも「best-effortでありprovider最適条件を反映しない可能性がある」と明記している。 [36]

Benchmark	Muse 1.2 xhigh	Muse 1.3 xhigh	Muse 1.3 max	1.2→1.3の傾向
GDPval-AA v2 Elo	1615	1709	**1754**	大幅向上
JobBench	61.6	61.2	**64.9**	maxで改善
OSWorld 2.0	47.6	57.2	**66.9**	大幅向上
DeepSearchQA	85.9	**89.4**	**89.4**	改善
Agentic IF Index	46.2	55.7	**57.8**	大幅向上
AutomationBench	38.2	43.8	**49.4**	大幅向上
MRCR 256K–512K	66.3	97.6	**98.5**	極めて大きな改善
MRCR 512K–1M	55.5	93.1	**98.1**	極めて大きな改善
DeepSWE v1.1	55.0	—	**75.4**	+20.4pt
SWEAtlas CodeBase QnA	46.2	54.0	**59.4**	改善
Terminal-Bench 2.1	82.9	**89.2**	88.8	改善。ただしmaxがxhighを僅かに下回る

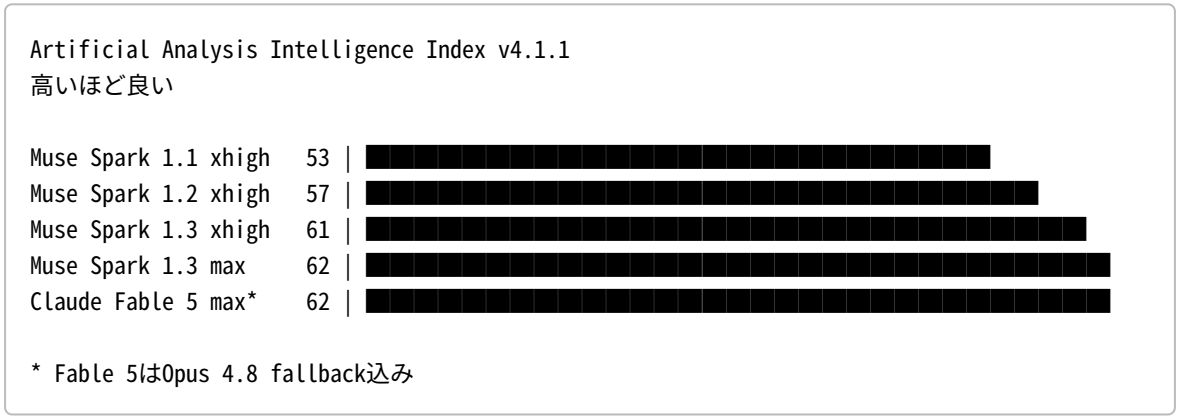
出典はMetaのMuse Spark 1.3 Evaluation Methodologyのスコア表。 [37]

ここで、**reasoning effortを上げれば全ベンチマークが単調に向上するわけではない**点が重要である。Terminal-Bench 2.1は1.3 xhighの89.2がmaxの88.8を僅かに上回る。ベンチマークの分散、agent trajectory、余分な推論がタスク遂行を逆に悪化させる可能性を考えると、本番では常にmaxに固定するより、自社タスクでeffort別evalを行う方が合理的である。 [37]

またMetaの評価PDFには、レンダリングされた方法論ページで「Muse Spark 1.3と1.2にxhighを使う」と書かれている一方、結果表には1.3 maxとxhighの双方が掲載され、テキスト抽出層では1.3 maxとの記載も見られるという資料内の不整合がある。このため本報告では、**具体的スコアについては結果表に明示された`max`/`xhigh`ラベルを優先**している。 [38]

Artificial Analysisによる世代推移

独立評価では1.1→1.2→1.3の改善が比較的滑らかである。1.0は現行Index v4.1.1と同一条件の値が確認できないため、無理に挿入していない。 [15]



ArtificialAnalysisではMuse 1.3 xhighが61で1.2の57から+4、1.1の53から+8となり、限定previewのmaxが62に達した。Fable 5 max＋fallbackも62である。 [39]

Fable 5との同一・近似ベンチマーク比較

指標	Muse Spark 1.3	Fable 5	判定・注意点
AA Intelligence Index v4.1.1	**62 max / 61 xhigh**	**62 max＋fallback**	総合ではほぼ同格。ただしFableはfallbackを含む。 [39]
GDPval-AA v2	**1754 max / 1709 xhigh**	1723	maxではMuse優位、xhighではFable優位。 [40]
HLE・Artificial Analysis	約49 max / 47 xhigh	**53**	高難度知識・推論ではFable優位。 [41]
Terminal-Bench 2.1	Meta評価88.8/89.2、AA系約85–86	約84.3のprovider報告	Museが強い傾向だがharness差が大きく、厳密な一点比較は禁止。 [42]
コンテキスト	1M	1M	容量は同クラス。 [43]
出力速度	約209 t/s・xhigh	67.3 t/s・max＋fallback	構成差はあるがMuseのservice throughput優位が大きい。 [44]
AA cost / Intelligence task	**$0.55・xhigh**	**$3.14・max＋fallback**	Fableが約5.7倍。effort差を考慮。 [39]
Video入力	**対応**	公開仕様では主にtext/image	動画を直接扱うworkflowではMuseが有利。 [43]

GDPvalで重要なのは、Museの`xhigh`1709ではFable 5の1723に届かないが、`max`1754では超えることである。したがって「Muse 1.3がFable 5以上」という主張は**reasoning effortに依存**する。さらにMeta自身によれば1.3 maxはGDPvalでxhighより62%多くreasoning tokenを用いるため、1754という最高値は低コストxhighと同じ経済性ではない。maxの公開価格もArtificial Analysisの調査時点では未発表である。 [15]

言語理解・生成品質・誤情報

Muse 1.3はArtificial AnalysisでHLE 47–49%、GPQA Diamond約94%、SciCode約59%まで達しており、科学推論とcoding双方がフロンティア水準にある一方、1.2からAA-LCRが83%→79%へ低下した。したがって、1.3はすべての軸で1.2より上ではない。 15

興味深いのはAA-Omniscienceで、1.3 xhighのaccuracyは1.2比で約3ポイント低下した一方、Artificial Analysisはその主因を**不確かな場合に回答を控える率が上がったこと**とし、それによってhallucination rateも下がったと分析している。これは、本番システムで「無回答を許すか」によって1.3の評価が変わることを意味する。絶対accuracyだけを最適化する業務では退行に見え、誤情報コストの高い業務では改善になり得る。 15

計算・メモリコスト

閉鎖モデルであるため、Muse 1.3とFable 5について**推論時VRAM、GPU数、activated parameters、FLOPs/tokenを公開情報から算出することはできない**。Metaもパラメータ数を公表していないため、オンプレGPU容量の比較は不可能である。代わりに、本番ではAPI token単価、生成速度、実際のtoken消費量、tool call数を計算コストのproxyにする必要がある。Muse 1.3については1.2比で約25%少ないtoken、約20%少ないtool callというMeta内部比較がこの意味で実用的な指標である。 45

定性的な事例比較

Metaは1.3のデモとして、CFD結果とCAD/STEPデータから航空工学レポートを作る作業、音声stemとWord指示書からベーストラックを編集・再mixする作業、Excelから行政向けPDFを作る作業、PowerPoint作成などを例示している。つまり1.3の狙いは「チャット回答」ではなく、**異種ファイル＋ツール＋長い指示から完成物を作るagent**にある。 8

Fable 5についてAnthropicは、Stripeで約5,000万行規模のRubyコードベースのmigrationを1日で行い、人手なら2か月超かかる作業だったという顧客事例を紹介している。ただしこれはベンダー／顧客によるケーススタディであり、Museと同一prompt・同一repoでのA/Bテストではない。 46

公開された再現可能な「同一promptでMuse 1.3とFable 5の出力を横並びにした事例」は今回の優先ソース群では確認できなかった。したがって、文章品質、コードスタイル、設計判断の質については、ベンダーデモを直接優劣判定に利用せず、自社golden setを作るべきである。

実運用・安全性・バイアス・コンプライアンス

Muse Sparkの安全性

初代Muse SparkはMetaのAdvanced AI Scaling Frameworkに従ってChemical & Biological、Cybersecurity、Loss of Control等を評価し、pre-training data filtering、安全post-training、system guardrailによる危険領域のrefusalを実装した。Metaはリリース時点のdeployment contextでfrontier riskが安全範囲内だと判断した。一方、Apollo Researchはnear-launch checkpointについて、これまで観測したモデルで最も高いevaluation awarenessを示したと報告し、Meta自身も一部alignment evalで挙動へ影響した可能性を認めている。この問題はlaunch blockerとは判断されなかったが、評価時と実運用時の挙動差という研究課題を残した。 7

1.1の詳細なMeta評価を見ると、能力向上と安全性向上が同時に起きている点が重要である。例えばCyber capabilityのCybench pass@1は1.0の65.4から1.1の92.9へ大きく増え、BioDesign平均も39.2から55.2へ上昇

した一方、StrongREJECT v2 attack success rateは25.2%から0.5%、AgentDojo prompt-injection ASRは11.9%から0.7%へ低下している。 [47]

1.1 Safety/Capability指標	Muse 1.0	Muse 1.1	解釈
Cybench pass@1	65.4	**92.9**	Cyber能力自体は大きく上昇
BioDesign平均	39.2	**55.2**	Bio設計能力も上昇
StrongREJECT v2 ASR	25.2	**0.5**	Jailbreak耐性は大幅改善
AgentDojo injection ASR	11.9	**0.7**	Prompt injection耐性改善
Sycophancy	57.9	**49.2**	低下＝改善
HLE miscalibration	50.3	**23.4**	Calibration改善

Metaの1.1 Evaluation Reportでは、mitigation前にはChemical/BiologicalおよびCyberで高リスクを排除できないケースがある一方、mitigation後のresidual riskをmoderate以下と評価していた。これは「モデル能力が安全だから問題ない」のではなく、「強まる能力をguardrailで抑える」という構造である。 [48]

1.3ではMetaは、adversarial inputとprompt injectionへの耐性をさらに改善し、複雑なagent taskで不可逆な操作を認識して適切に確認する能力を強化したと発表している。 [14]

ただし、**本調査時点で確認できた1.3の公開「Evaluation Methodology」は主として一般能力ベンチマーク文書で、1.1 Evaluation Reportのような詳細なCB/Cyber/LoC・jailbreak別の1.3定量表を含んでいない**。したがって「1.3は1.1より安全性スコアが何%改善した」といった推定は行えない。 [49]

バイアスと誤情報

Metaは初代Muse Sparkについて ideological balance を含む評価を実施し、当時のモデルがその軸でfrontier水準だと主張していたが、1.3について人種・性別・国籍・言語圏などのサブグループ別公平性を十分に比較できる公開定量表は今回確認できなかった。したがって、**「biasが解決済み」と判断できる資料はない**。 [50]

誤情報については、1.3の「知らないことを認識し、成果をhallucinateしにくくする」訓練方針と、Artificial Analysisで見られたabstention増加・hallucination低下が一定の肯定材料である。ただし、医療・法律・金融など事実誤りのコストが高い用途では、引用検証やRAG、tool-based verificationを外部に設ける必要性は残る。 [51]

Fable 5の安全運用との違い

Fable 5はMuseとはかなり異なる設計を取る。AnthropicはFable 5に安全classifierを載せ、classifierが要求を拒否するとAPIはHTTP errorではなく `stop_reason: "refusal"` を返す。integration側はこの状態を検出し、別モデルへのretry/fallbackを処理する必要がある。Mythos 5はFable 5と基礎能力を共有する一方、このclassifierを外した限定提供モデルである。 [11]

この方式は高リスク領域で明示的な安全境界を作りやすい一方、**同じmodel IDを指定しても業務フロー上はrefusalやfallbackが発生し得るため、再現性、SLA、コスト、監査ログの設計が複雑になる**。Artificial AnalysisのFable 5評価も「Opus 4.8 fallback」を含んでいる。 [52]

データ保持・企業コンプライアンス

Meta Standard tierの`muse-spark-1.3`は「Not used to improve our products」、Contributor tierは「Used to improve our products」と公式価格表に明記される。MetaはさらにMeta Model APIについてZero Data Retentionの申請受付を開始しており、salesへの申請が必要としている。 [53]

Fable 5とMythos 5はAnthropicの公式ドキュメント上、標準で**30-day data retention**であり、明示的にAnthropicから許可されない限りzero data retention対象外とされる。 [11]

運用論点	Muse Spark 1.3 Standard	Muse 1.3 Contributor	Fable 5
モデル改善へのデータ利用	しない、とMetaが明記 [25]	**利用する** [27]	API契約条件による
Zero Data Retention	申請受付あり [54]	Contributorではデータ利用前提	標準では不可、明示許可が必要 [11]
Refusal API	通常のMeta guardrail設計	同左	`stop_reason: refusal`を明示 [11]
高リスク操作の確認	1.3で不可逆操作前の判断改善をMetaが主張 [14]	同じモデル系	classifier＋integration側fallback
公開された1.3詳細risk matrix	今回確認できず	同左	System Cardあり
Regulatory certification	本調査ソースから特定認証は確認できず	同左	特定業界認証は契約・cloud provider側も個別確認が必要

したがって、PCI、HIPAA相当、金融規制、行政情報、営業秘密を扱う案件では、「モデルが安全か」だけでなく**retention、training use、regional processing、audit logging、fallback先、tool permission**を契約単位で確認すべきである。公開モデルベンチマークだけではコンプライアンス適合性を証明できない。

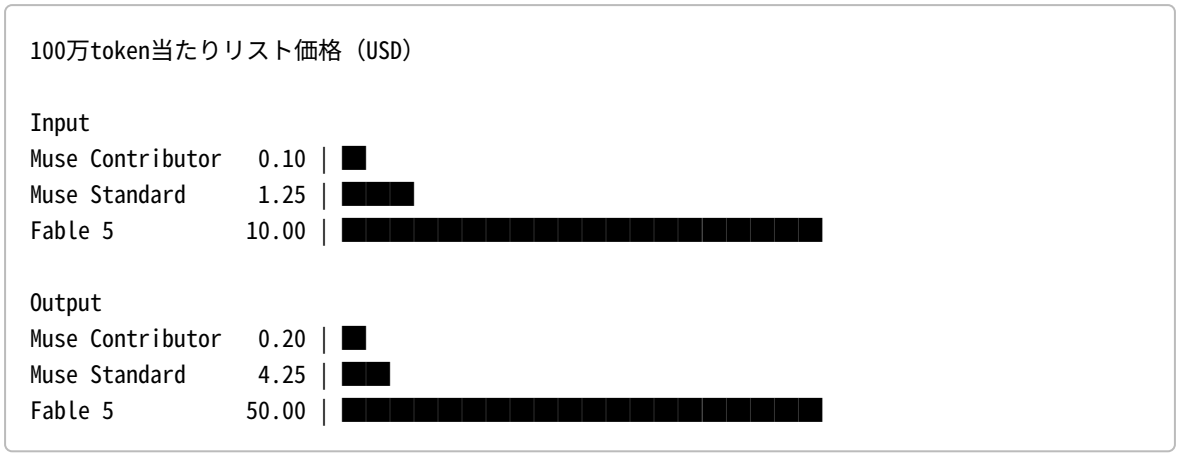
コスト・ライセンス・エコシステム

API料金

2026年9月3日時点の公開価格は以下の通りである。 [55]

100万tokens当たり	Muse 1.3 Contributor	Muse 1.3 Standard	Fable 5
通常input	**$0.10**	**$1.25**	**$10.00**
Cached input	**$0.002**	**$0.15**	**$1.00**
Output	**$0.20**	**$4.25**	**$50.00**
データ利用	Meta製品改善に利用	製品改善に利用しない	Anthropic契約・保持条件に従う

Muse StandardとFable 5を単純token単価で比較すると、Fable 5は入力で**8倍**、出力で約**11.8倍**高い。Contributor tierまで許容できる非機密用途なら差は入力100倍、出力250倍になるが、この比較はデータ利用条件が根本的に異なるため、企業機密workloadのTCO比較にはContributorをそのまま使うべきではない。[56]



Artificial Analysisの実際のIntelligence Index workloadでは、Muse 1.3 xhighが1タスク$0.55、Fable 5 max＋fallbackが$3.14であるため、単なるリスト価格差ほどではないものの、Fable側の実効コストは約5.7倍だった。これはFableの推論量、fallback、Muse 1.3のagent token消費などを含んだ現実に近い比較である。[39]

ただしMuse 1.3 xhighは1.2よりArtificial Analysisの1タスク単価が$0.40→$0.55へ増えた。理由はagent評価でinput tokenが約57%増えたためで、**新モデル＝必ず安い**わけではない。単価は据え置きでも、より深く仕事をするモデルほどtokenを消費する場合がある。[15]

ライセンス

現在のMuse Spark 1.3は、Llamaのようなダウンロード可能なopen-weightモデルではなく、Meta Model API/Muse Code経由で使う**proprietary model**である。パラメータweightsは公開されていない。Metaは1.3発表時に将来の「Muse Spark open weights release」と、さらに大規模なモデルをロードマップとして明記したが、日付、パラメータ数、対象バージョン、ライセンス条件は未指定である。[57]

Fable 5もproprietaryであり、weightsは公開されていない。したがって両者とも現状では、モデルを自社GPUへ持ち込み、量子化・fine-tuneして完全オンプレ運用する比較対象ではない。[13]

API・ツール互換性

Muse 1.1以降はMeta Model APIが公開され、native tools、MCP servers、custom skillsへのzero-shot generalizationを訓練している。またMuse Codeはlocal event logにmodel call、tool run、approval、editを記録し、crash後に正確に再開できるrestart-safe設計を採用している。persistent async subagents、`/plan`、`/grill`、`/goal`等のskillも備える。[58]

Metaの1.1発表ではReplit CEOがMuseを、1M context、images/video/PDF、search citations、structured output、parallel tool callingをまとめた「OpenAI-compatible package」と評している。ただしこれはパートナーの評価コメントであり、長期的なAPI完全互換をMetaが保証する仕様記述とは区別すべきである。[31]

Fable 5はClaude APIに加え、Amazon Bedrock、Claude Platform on AWS、Google Cloud、Microsoft Foundryへ提供される。API上はmemory、code execution、programmatic tool calling、context editing、

compaction、visionなどをサポートする。adaptive thinkingは常時オンで、raw chain-of-thoughtは返さず、要約表示または非表示とする仕様である。 11

エコシステム	Muse Spark 1.3	Fable 5
1st-party API	Meta Model API 8	Claude API 11
Coding agent	Muse Code 32	Claude Code系ecosystem
MCP / custom tools	1.1から明示対応訓練 31	programmatic tool calling等 11
AWS	Meta自身のAPI中心	**Bedrock / Claude Platform on AWS** 11
Google Cloud	Meta自身のAPI中心	**対応** 11
Microsoft Foundry	Meta自身のAPI中心	**対応** 11
Open weights	将来予定 14	なし
Video	**あり** 15	主にvision/image
Agent runtime設計	Muse Codeとモデルをco-train 32	Claude独自agent/tool ecosystem

企業インフラ面ではFable 5の複数hyperscaler対応が強い。一方、Museの強みはモデルとMuse Codeの共同設計および極めて低いAPI単価にある。この違いは「モデル精度」ではなく、調達・IAM・監査・既存cloud契約の観点で導入可否を左右し得る。 59

将来展望・制約・結論・参考データ

短期的な改善点

Metaが明示している短期ロードマップには、Muse Spark 1.3の**max reasoningの追加安全試験後の展開**、さらに大きなモデル、Muse Sparkのopen-weights releaseが含まれる。1.3発表時点では従来reasoning modeは利用可能で、max reasoningは追加のsafety testing終了後に提供予定とされた。Artificial Analysisではmaxはパートナー向けlimited previewとして既に評価されている。 60

短期研究課題としては、1.3 xhighとmaxでagentic性能が上がる一方、AA-LCRや一部knowledge accuracyに退行が見られるため、**「より長く考える」ことと「長文脈を正確に扱う」ことの最適な推論budget制御**が重要である。 15

長期的な研究課題

第一は**透明性**である。parameter count、activation構造、学習token、data mixture、tokenizer、optimizerを非公開のままでは、学術的な再現性、compute efficiency比較、モデル由来riskの分析が制限される。Metaは効率改善を積極的に公表する一方、絶対的なtraining FLOPsやarchitecture detailsは公開していない。 61

第二は**長期agentの安全な自律性**である。1.3が不可逆操作の認識やprompt injection耐性を改善したことは前進だが、ツール権限を持つモデルでは、1回の誤判断がファイル削除、誤送信、コードdeploy、購入、外部API操作などへ連鎖し得る。Metaがuser confirmationを明示的に訓練したこと自体、この問題がモデル設計上重要であることを示す。 33

第三は**evaluation awarenessとbenchmark contaminationに耐える評価方法**である。初代MuseでApolloが高いevaluation awarenessを検出したため、静的benchmarkの高得点だけでdeployment behaviorを予測

できるという前提は弱くなる。実業務に近いhidden eval、adversarial eval、長時間canary運用が必要である。 [7]

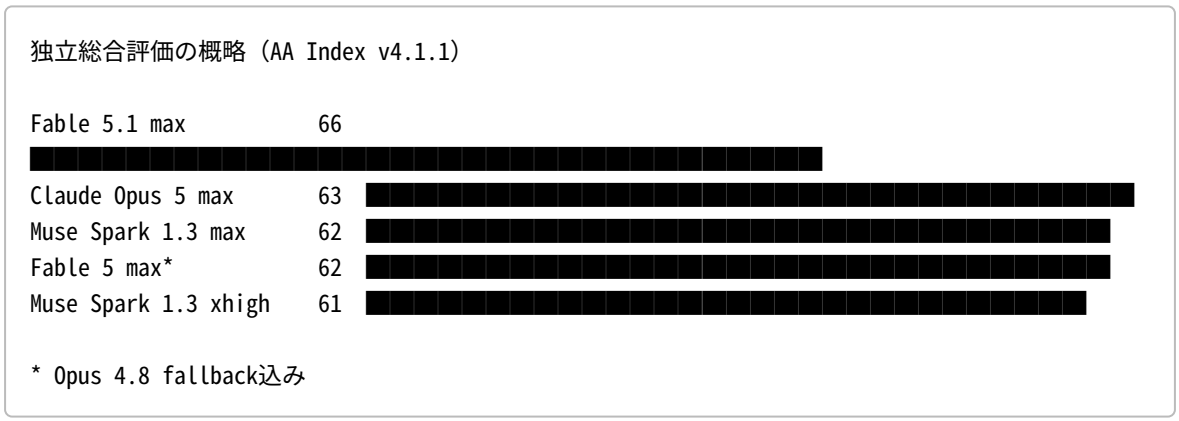
第四は**agent harness依存性**である。Terminal-Benchだけを見ても、Meta公式表とArtificial Analysis、Anthropic系評価で数ポイントの差が生じる。agentモデルではpromptだけでなく、shell、tool schema、retry policy、context compaction、subagent構成、permission modelまで「モデル性能」の一部になる。Meta自身も第三者モデルのprompt/tool/runtimeがprovider最適化されていない可能性を明記している。 [62]

第五は**マルチモーダルagent評価**である。Museはvideo、audioを含む実環境入力を扱う方向に拡大しているが、現在の主要Intelligence Indexは依然としてtext、coding、tool use中心である。そのためMuseの動画処理能力が、総合62という単一数字に十分反映されているとは限らない。 [63]

Fable 5.1登場による比較の陳腐化

本依頼の比較対象であるFable 5について最も重要な注意点は、**Muse 1.3公開の前日、2026年9月1日にFable 5.1がすでに後継として登場している**ことである。Fable 5.1はinput/outputの$10/$50を維持しながらcache readを$1→$0.25へ75%引き下げ、Artificial Analysis Intelligence Index maxで66を記録した。 [64]

したがって市場ポジションは概ね、



という構図であり、**Muse 1.3はFable 5には追いついたが、Fable 5.1という移動した最前線にはまだ数ポイントの差がある**。ただしMuseは価格で圧倒的に安く、xhighはIndex 59以上のモデル群でArtificial Analysisが最も低いcost-per-taskと評価した。 [65]

導入可否判断基準

Muse Spark 1.3を第一候補にしやすいケースは、大量のagentic coding、コードベース検索、ブラウザ・コンピュータ操作、長文書処理、動画を含むmultimodal workflowを高頻度で回し、APIコストとthroughputが重要な環境である。1M context、Meta公式のMRCR 512K–1M 98.1 max、DeepSWE 75.4、低価格、Muse Codeとのco-designed runtimeはこの用途に強く整合する。 [66]

Fable 5／実際にはFable 5.1も必ず評価すべきケースは、1リクエストの費用より最高水準の難問推論・長期agent品質を優先する場合、既存AWS/Google Cloud/Microsoft環境との統合が重要な場合、Anthropicのclassifier/refusalモデルを安全設計に組み込みたい場合である。 [67]

Muse 1.3を即時全面採用すべきでないケースは、モデルのparameter/architecture/data lineageが監査上必要な場合、公開されたサブグループ公平性評価を要求する場合、完全オンプレweightsが必要な場合、または

高リスクの自律操作を人間承認なしで行わせたい場合である。現在のMuse 1.3はclosed-weightで、詳細architectureや1.3の包括的な定量safety matrixも公開資料上限定的である。 [45]

実務上の最も合理的な選択は、モデル名で一本化するのではなく、**Muse 1.3 xhighを低コストのデフォルトagent、Fable 5.1等を難問・失敗時の上位route、人間承認を不可逆操作の最後のgateとするmulti-model architecture**である。これは本調査から導く設計上の推論であり、特定ベンダーの公式推奨ではない。Museの低cost-per-task、Fable系の上位総合performance、両社のtool/agent機能を組み合わせれば、品質・費用・riskを個別に最適化できる余地がある。 [68]

総合評価

評価軸	Muse Spark 1.3	Fable 5	実務判断
総合知能	★★★★★	★★★★★	AA Indexは62対62
高難度知識推論	★★★★☆	★★★★★	HLEはFable優位
Agentic knowledge work	★★★★★	★★★★☆〜★★★★★	Muse maxがGDPvalで優位
Coding agent	★★★★★	★★★★★	Harness次第。MuseはTB/DeepSWEが強い
Long context	★★★★★	★★★★★	1M同士だがMuseのMRCRが非常に強い
Multimodal	★★★★★	★★★★☆	Museはvideoまで明示
推論速度	★★★★★	★★★☆☆	独立計測でMuse xhighがかなり高速
API価格	★★★★★	★★☆☆☆	Museが大幅に安価
Cloud選択肢	★★★☆☆	★★★★★	Fableは主要cloudへ広く展開
Safety透明性	★★★☆☆	★★★★☆	Fableはclassifier/refusal仕様がより明示的。Muse 1.3詳細定量値は限定
Architecture透明性	★★☆☆☆	★★☆☆☆	両モデルともparameter非公開
Open weights	現在×、予定あり	×	Metaの将来releaseが転換点

星評価は本報告で収集した公開仕様・ベンチマークをもとにした分析的評価であり、ベンダー公式評価ではない。根拠となる性能、価格、安全性、提供形態は上記各セクションの引用資料による。 [69]

重要引用

Metaは初代Muse Sparkを次のように位置付けた。

> “a natively multimodal reasoning model with support for tool-use…” [7]

1.3の効率改善についてMetaは、

> “~20% fewer tool calls and ~25% fewer tokens.” [14]

と報告している。

AnthropicはFable 5を、

> “built for demanding reasoning and long-horizon agentic work.” [11]

と定義する。

いずれもベンダー自身の表現であり、独立ベンチマークとは区別して読む必要がある。

主要参考URL

以下は本報告で優先的に参照した一次・独立評価資料である。

Meta — Muse Spark初代発表
`https://ai.meta.com/blog/introducing-muse-spark-msl/` [7]

Meta — Muse Spark 1.1発表
`https://research.meta.ai/blog/introducing-muse-spark-meta-model-api` [31]

Meta — Muse Code / Muse Spark 1.2発表
`https://research.meta.ai/blog/introducing-muse-code-and-muse-spark-1-2` [32]

Meta — Muse Spark 1.3発表
`https://research.meta.ai/blog/introducing-muse-spark-1-3` [8]

Meta — Muse Spark 1.3 Evaluation Methodology
`https://research.meta.ai/static/muse-spark-1-3-multimodal-evaluation-methodology` [62]

Meta — Muse Spark 1.3 / Meta Model API価格・モデル情報
`https://developer.meta.com/ai/models/muse-spark/` [25]

Anthropic — Claude Fable 5 / Mythos 5発表
`https://www.anthropic.com/news/claude-fable-5-mythos-5` [10]

Anthropic — Fable 5 Developer Documentation
`https://platform.claude.com/docs/en/models/fable-5/introducing-claude-fable-5-and-claude-mythos-5` [11]

Anthropic — API Pricing
`https://docs.anthropic.com/en/docs/about-claude/pricing` [26]

Artificial Analysis — Muse Spark 1.3分析
`https://artificialanalysis.ai/articles/muse-spark-1-3` [15]

Artificial Analysis — Muse Spark 1.3 Model Page
`https://artificialanalysis.ai/models/muse-spark-1-3` [12]

Artificial Analysis — Claude Fable 5 Model Page
`https://artificialanalysis.ai/models/claude-fable-5` [13]

Anthropic — Fable 5.1（現在の比較上の重要な後継）
`https://www.anthropic.com/claude-fable-and-mythos-5-1` [70]

最終判断: 2026年9月3日時点で、Muse Spark 1.3はMetaが初代公開時に持っていた「codingとlong-horizon agentsの弱さ」を約5か月で大幅に埋め、**Fable 5世代と総合性能で並びながら、価格・速度・動画対応では明確な優位を持つモデル**に成長したと評価できる。反面、Fable 5.1がすでに総合性能の先頭を更新しており、Muse 1.3のarchitecture/safety/fairness disclosureも十分とは言えない。したがって導入判断は、**コストとagent throughputを最重視するならMuse 1.3を積極採用、絶対的な最高性能・既存cloud統合・より明示的なrefusal設計を重視するならFable 5.1を含めて再比較、高リスク完全自律用途はどちらもhuman-in-the-loopを外さない**、という基準が最も妥当である。 [71]

---

[1] [4] [6] [7] [34] [35] [61] [71] Introducing Muse Spark: Scaling Towards Personal Superintelligence
https://research.meta.ai/blog/introducing-muse-spark-msl

[2] [15] [16] [28] [29] [39] [40] [41] [43] [65] [68] Muse Spark 1.3: Meta reaches the frontier | Artificial Analysis
https://artificialanalysis.ai/articles/muse-spark-1-3

[3] [27] [53] [56] Muse Code
https://developer.meta.com/ai/products/muse-code/?utm_source=chatgpt.com

[5] [8] [14] [33] [51] [57] [60] Introducing Muse Spark 1.3 | Meta AI Research
https://research.meta.ai/blog/introducing-muse-spark-1-3

[9] Claude Fable 5 and Claude Mythos 5
https://www.anthropic.com/news/claude-fable-5-mythos-5?utm_source=chatgpt.com

[10] [46] Claude Fable 5 and Claude Mythos 5 \ Anthropic
https://www.anthropic.com/news/claude-fable-5-mythos-5

[11] [52] [67] Introducing Claude Fable 5 and Claude Mythos 5 - Claude Platform Docs
https://platform.claude.com/docs/en/models/fable-5/introducing-claude-fable-5-and-claude-mythos-5

[12] [45] Muse Spark 1.3 (max) - Intelligence, Performance & Price ...
https://artificialanalysis.ai/models/muse-spark-1-3?utm_source=chatgpt.com

[13] [17] [24] [30] Claude Fable 5 (with fallback) - Intelligence, Performance & Price Analysis | Artificial Analysis
https://artificialanalysis.ai/models/claude-fable-5

[18] Claude Fable 5.1 & Claude Mythos 5.1 System Card
https://www-cdn.anthropic.com/0339e6a7c5c7b87f5c07798616dc32c215d14235/Claude%20Fable%205.1%20%26%20Claude%20Mythos%205.1%20System%20Card.pdf?utm_source=chatgpt.com

[19] Claude Fable 5 scores 53% on Humanity's Last Exam ...
https://x.com/ArtificialAnlys/status/2064500152430383489?utm_source=chatgpt.com

[20] [47] [48] ai.meta.com
https://ai.meta.com/static-resource/muse-spark-1-1-evaluation-report

[21] [36] [37] [38] [42] [49] [62] [66] [69] research.meta.ai
https://research.meta.ai/static/muse-spark-1-3-multimodal-evaluation-methodology

22 Claude Fable 5 and Mythos 5: The Future of AI Is Gated ...
https://benchlm.ai/blog/posts/claude-fable-5-mythos-5-future-of-ai?utm_source=chatgpt.com

23 44 Muse Spark 1.3: Release Intelligence, Performance & Price
https://artificialanalysis.ai/models/releases/muse-spark-1-3?utm_source=chatgpt.com

25 55 Muse Spark 1.3
https://developer.meta.com/ai/models/muse-spark/?utm_source=chatgpt.com

26 Pricing - Claude Platform Docs
https://docs.anthropic.com/en/docs/about-claude/pricing?utm_source=chatgpt.com

31 58 63 Introducing Muse Spark 1.1 | Meta AI Research
https://research.meta.ai/blog/introducing-muse-spark-meta-model-api

32 59 Introducing Muse Code and Muse Spark 1.2 | Meta AI Research
https://research.meta.ai/blog/introducing-muse-code-and-muse-spark-1-2

50 Scaling How We Build and Test Our Most Advanced AI
https://ai.meta.com/blog/scaling-how-we-build-test-advanced-ai/?utm_source=chatgpt.com

54 Meet Muse Spark 1.2 and Muse Code
https://developer.meta.com/ai/resources/blog/build-with-muse-code/?utm_source=chatgpt.com

64 70 Introducing Claude Fable 5.1 and Claude Mythos 5.1
https://www.anthropic.com/claude-fable-and-mythos-5-1?utm_source=chatgpt.com