ChatGPT

# Perplexity「Portable Computer」深掘り調査報告書：ローカルファーストAIエージェントの技術、競争力、市場インパクト

## エグゼクティブサマリー

Perplexityは米国時間2026年8月25日、NVIDIAと共同で、AIエージェント「Perplexity Computer」のローカルファースト版「Portable Computer」を発表した。最大の特徴は、**LLMだけでなく、エージェント・ハーネス、実行ループ、プランニング、ツール実行、タスクキュー、ローカル検索インデックスまでをユーザー側マシンで動かす**ことにある。ローンチ時の公式対応環境はNVIDIA DGX Spark＋Linuxで、NVIDIA RTX GPU PCとWindowsへの拡大が予告されている。ローカル処理分はPerplexity Computerのクレジットを消費しない。 [1]

ただし、「完全ローカル」という表現には重要な但し書きがある。Portable Computerは**完全オフライン動作も可能**だが、Web検索、Gmail・Slack・GitHubなどのコネクタ、ブラウザ操作、高度な推論を必要とするときには、ユーザーの許可に基づいてクラウドへエスカレーションできる。詳細な技術記事では、クラウド上の“adviser”へ送る前に関連コンテキストを選別し、PII（個人識別情報）分類器を通し、ユーザーに送信内容を見せる設計が説明されている。クラウド側のadviserはテキスト助言を返すだけで、ローカルのファイルやツールを直接操作しない。 [2]

モデルはローンチ時に**Qwen 3.8 27B**と、それをPerplexityがComputer用に事後学習した**PPLX 27B**を用意し、NVIDIA **Nemotron 3.5 Lightning 30B**も追加予定である。音声入力もNemotron 3.5 ASRを使ってローカル処理する。 [3]

技術的に特に興味深いのは、Perplexityが性能の源泉を単なる「27Bモデル」ではなく、**小型ローカルモデルに最適化したハーネス**に置いている点である。公称260Kトークンのコンテキストを無理に使わず、同社の実測では100K超で性能が不安定になるとして、短いシステムプロンプト、少数のコアツール、オンデマンドSkills、コンテキスト圧縮、MCPツール定義のCLI化を採用している。さらに、制御ループ自体はLLMではなく**決定論的なハーネスコード**が管理し、モデルは「次に何をするか」を提案する役割に限定される。 [4]

これは公式Xのローンチ文言にある「orchestrator LLM」という表現と一見矛盾する。詳細な研究記事は明確に「orchestrator is deterministic harness code, not an LLM」と説明しているため、本報告では**詳細技術文書の記述を実装アーキテクチャの基準**とする。X側は製品スタックを簡略化したマーケティング上の表現、あるいは別レイヤーの用語を指す可能性があるが、Perplexity自身はこの用語差をまだ説明していない。 [5]

市場面では、Portable Computerは「クラウドAPIで知能を買う」モデルから、**GPUを所有し、通常タスクはローカルで処理し、難問だけクラウド知能を購入する**モデルへの転換を象徴している。ただしローンチ時にはDGX Sparkの米国MSRPが4,699ドルであり、さらにPro/Maxサブスクリプションが必要なため、初期市場はヘビーユーザー、開発者、機密データを扱う専門職・企業に偏るとみられる。RTX/Windows対応が一般化すれば市場規模は大きく変わり得る。これは以下の一次情報と価格から導く分析であり、Perplexity自身が市場シェア予測を公表したものではない。 [6]

| 評価軸 | 本調査の結論 |
|---|---|
| 技術的新規性 | **高い。** モデルだけでなくエージェント実行基盤全体をローカル最適化している。 [7] |
| 完全オフライン | **可能。** Web検索を完全に無効化できる。ただしWeb・コネクタ・クラウドadviserは当然利用不可。 [8] |
| プライバシー | **クラウド中心エージェントより構造的に有利。** ただし端末セキュリティ、コネクタ、ユーザー承認、PII検出の限界は残る。 [9] |
| 性能 | Perplexity自身のテストでは同一QwenモデルでPi/Hermesを上回る。ただし内部ベンチマークを含み、独立検証は未完了。 [10] |
| 最大の弱点 | DGX Spark中心の高い初期費用、ローンチ時Linux中心、Portable専用公開SDK/GitHubリポジトリが未確認。 [11] |
| 市場的意味 | 「クラウドかローカルか」ではなく、**ローカルを通常経路、クラウドを例外的adviserにするハイブリッド型**を実用製品として提示した点が重要。 [12] |

## 公式発表と一次情報

Perplexityの英語版発表はPortable Computerを次のように位置づけている。

> “Today we’re launching Portable Computer, a version of Perplexity Computer that runs entirely on a local machine.”

日本語にすれば、**「本日、ローカルマシン上で完全に動作するPerplexity Computerのバージョン、Portable Computerを発表する」**となる。公式日本語ページでも同趣旨の翻訳が掲載されている。 [13]

公式発表URL：
https://www.perplexity.ai/hub/blog/introducing-portable-computer-for-local-first-ai

公式日本語版：
https://www.perplexity.ai/ja/hub/blog/introducing-portable-computer-for-local-first-ai

公式Xではローンチ時に、より強く「完全ローカル」を打ち出した。

> “...agent harness all run on your local hardware. No cloud dependency.”

日本語では、**「エージェント・ハーネスを含む実行環境がローカルハードウェア上で動作し、クラウドへの依存を必要としない」**という意味である。公式X投稿は2026年8月25日付。 [14]

公式X：
https://x.com/perplexity_ai/status/2092268362386780270

さらにPerplexityは製品発表と同時に、より技術的な研究記事「A Local-First Agent for Private and Cost-Effective Knowledge Work」を公開した。ここでは製品コンセプトを「モデル、ハーネス、会話、エージェントのtrajectory（実行履歴）がデフォルトでユーザー側マシンに存在する」と定義し、Web、コネクタ、クラウドadviserを必要時のみ呼び出す設計を説明している。 [15]

技術研究記事：
https://www.perplexity.ai/hub/blog/a-local-first-agent-for-private-and-cost-effective-knowledge-work

共同開発先のNVIDIAも8月25日にPortable ComputerをDGX Spark向けローカルエージェントとして紹介し、

> “Support for GeForce RTX and RTX PRO GPUs is coming soon.”

と明記している。すなわち、**公式ベースではローンチの主対象はDGX Sparkであり、一般RTX/RTX PROは「coming soon」**である。 16

NVIDIA公式：
https://blogs.nvidia.com/blog/local-ai-open-source-models-agents-nemotron/

### 公式情報の整合性チェック

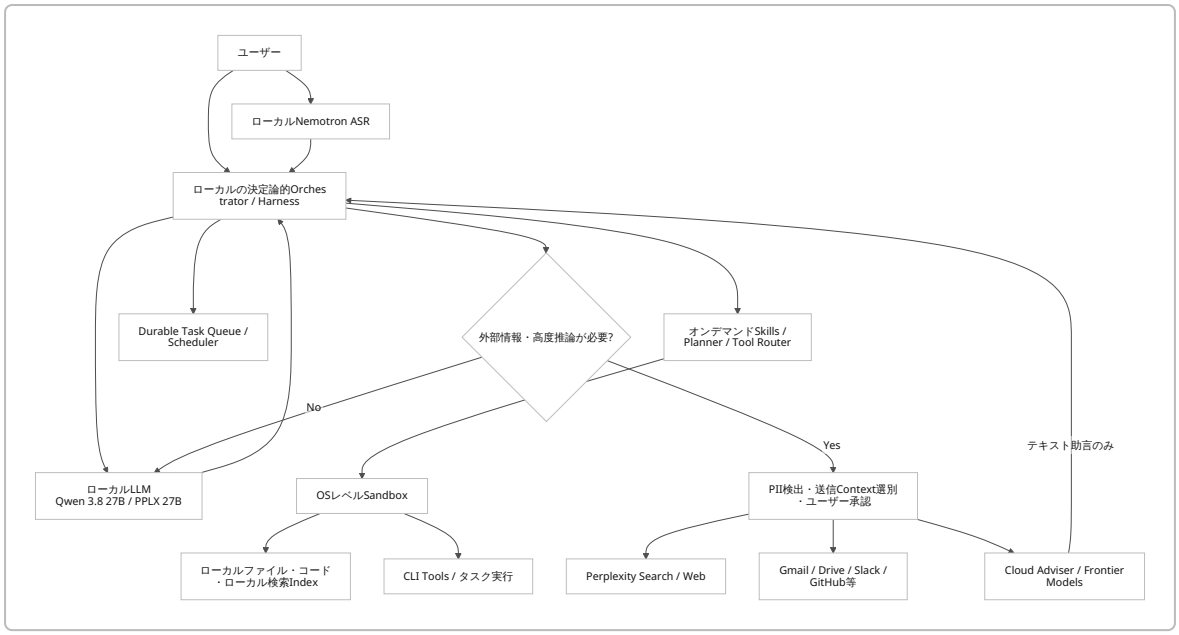
| 論点 | 一次情報から確認できる内容 | 判定 |
|---|---|---|
| 「完全ローカル」なのか | モデルとハーネスを含む基本処理はローカル。Web等を無効にすれば完全オフラインも可能。 17 | **Yes、ただしクラウド機能はオプション** |
| モデルだけがローカルか | いいえ。オーケストレータ、planner、tool router、scheduler、durable task queue、local search indexも端末上。 18 | **フルスタック型** |
| クラウドへの自動送信 | 一般のクラウドサービス送信時は許可を求める。adviserではPII検出＋送信コンテキスト提示＋承認を行える。 19 | **ユーザー制御型** |
| クラウドモデルが端末を操作するか | Adviserはテキスト助言のみで、ファイル、ツール、会話へ直接アクセスしない。 12 | **原則No** |
| クレジット | ローカルモデルの処理はComputerクレジットを消費しない。クラウド側Computer機能には別途クレジット体系が存在。 20 | **ローカル推論は実質定額** |
| GitHub公開 | Perplexity公式GitHub組織は存在するが、本調査時点でPortable Computer専用の公開リポジトリは確認できない。 | **専用Repo：未確認** |
| Portable専用API/SDK | 公開されたPortable専用API/SDKは発表文・研究記事で確認できない。PerplexityのAgent API/Search APIとは別物。 21 | **未記載** |

Perplexity公式GitHub：
https://github.com/perplexityai

ここで重要なのは、Perplexityには既にAgent APIやSearch APIがある一方、**それがPortable Computerを外部アプリへ埋め込むAPIであるとは公式には説明されていない**点である。また「Search as Code」にはAgentic Search SDKが存在するが、これはPerplexityの検索スタックをエージェントのsandbox内から利用するためのSDKとして説明されており、現時点で「Portable Computer SDK」と呼べる公開開発キットではない。 21

## 技術仕様とアーキテクチャ

Portable Computerの中心設計は、「小型ローカルLLMにクラウド向け巨大ハーネスをそのまま載せる」のではなく、**モデルとハーネスを共同最適化する**ことである。PerplexityはQwen 3.8 27Bについて公称260Kトークンのコンテキストを持つものの、自社実験では100Kトークンを超えると苦戦し始めるとして、コンテキスト消費そのものを減らす設計を採用している。 22




この図で最も重要なのは、**ローカルLLMがOSツールを直接支配するのではない**ことである。詳細研究記事では、決定論的なorchestratorコードが実行ループ、コンテキスト組み立て、ポリシーを管理し、LLMが次のアクションを提案、承認されたツール呼び出しだけをsandboxで実行する、と説明されている。 23

### 仕様一覧

| 項目 | 確認できた仕様 | 評価・注記 |
|---|---|---|
| ローカルLLM | Qwen 3.8 27B、PPLX 27B。Nemotron 3.5 Lightning 30B追加予定。 3 | PPLX 27BはQwen 3.8 27BをComputer harness向けに事後学習。 |
| PPLX 27Bの公開weights | **未記載** | 公開モデルカード、ダウンロード可能なweights、ライセンスは今回確認できず。 |
| Orchestrator | 詳細研究記事では決定論的ハーネスコード。LLMではない。 23 | Xの「orchestrator LLM」表現とは用語上の不整合あり。 |
| Planner / Tool Router / Scheduler | ローカル実行。 18 | 詳細な実装言語・内部APIは未記載。 |
| Durable task queue | ローカル。 18 | DB実装・永続化形式：**未記載**。 |

| 項目 | 確認できた仕様 | 評価・注記 |
|---|---|---|
| ローカル検索Index | 端末上。[18] | ベクトルDB等の具体実装：**未記載**。 |
| Context管理 | 短いsystem prompt、少数core tools、オンデマンドSkills、context compaction。[24] | ローカル小型モデル向けの主要最適化。 |
| MCP | 一般的なMCPの大きなtool definitionを避け、頻用MCPをコンパクトなCLIツール＋Skillsへ変換。[25] | MCP互換性の完全な一覧は未公開。 |
| 自己検証 | モデル自身またはtrajectory監視hookからself-verificationを発火可能。[23] | 長時間agentの品質維持策。 |
| Sandbox | OSレベル。プロセス、filesystem path、network accessをpolicyで制限。sandbox不可ならtool call前にハーネスを無効化。[23] | 具体技術名、namespace/container方式等：**未記載**。 |
| 音声入力 | NVIDIA Nemotron 3.5 ASRをローカル実行。録音をクラウドへ送らず文字起こし可能。[18] | 音声モデルの詳細設定：未記載。 |
| Web検索 | 任意。Perplexity Search / Search as Code。[26] | ネット接続必須。 |
| Cloud adviser | オプション。ローカルモデルが相談時機を決め、orchestratorは権限を保持。[12] | 手動承認または自動承認を選択可能と研究記事に記載。 |
| クラウド送信前処理 | 関連context選別、PII classifier、送信内容をユーザーに提示。[12] | PII classifierの精度・モデル：**未記載**。 |
| Cloud adviserの権限 | テキスト助言のみ。ローカルファイル・tool・conversationへ直接アクセス不可。[12] | 最小権限設計として重要。 |
| Connectors | Google Drive、Gmail、Slack、GitHub等。[1] | OAuth権限の粒度や全connector一覧：未記載。 |
| 完全オフライン | **可能。** Web searchを完全にdisableできる。[8] | 外部コネクタ・検索・adviserは使用不可。 |
| ローンチOS | Linux。Windowsは「近日」対応予定。[1] | VentureBeatは9月予定と報道。公式ブログは月を断定していない。 |
| macOS版Portable Computer | **未記載／未発表** | 別製品・機能としてMac向けPersonal Computerが存在するが、Portableとは区別すべき。[27] |
| ローンチHardware | NVIDIA DGX Spark。[1] | RTX/RTX PRO対応は公式には coming soon。 |

| 項目 | 確認できた仕様 | 評価・注記 |
|---|---|---|
| 最低VRAM | 公式発表：**未記載** | VentureBeatはNVIDIA RTXで最低24GB VRAMと報道。ただし一次情報とのタイミング差あり。 [28] |
| DGX Sparkメモリ | 128GB unified system memory。 [29] | Portable固有ではなくDGX Spark本体仕様。 |
| インストール | Perplexity App経由でone-click local inference setupとNVIDIAが説明。 [16] | パッケージ形式・署名・更新方式：未記載。 |
| Portable専用API | **未記載** | 一般のPerplexity Agent APIとは別。 |
| Portable専用SDK | **未記載** | Search as Code内部にはAgentic Search SDKがある。 [21] |
| データ暗号化 at rest | **未記載** | 端末上の保存データ暗号化方式は発表資料にない。 |
| Telemetry | **未記載** | Portable利用時に何がPerplexityへtelemetry送信されるかは発表資料では不明。 |
| ログ保存期間 | **未記載** | エンタープライズ導入では要確認事項。 |
| Secure Boot / attestation | **未記載** | DGX Spark側機能との統合も今回の資料にはない。 |
| Multi-GPU / cluster | Portable Computerとしては**未記載** | DGX Sparkそのものは拡張機能を持つが、Portableでのサポートを意味しない。 [29] |

## 性能ベンチマーク

Perplexityは同じQwen 3.8 27BをDGX Spark上で動かし、Computer harness、Pi、Hermesを比較した。内部のLocal Knowledge Work BenchではComputerが82.6%、Piが77.6%、Hermesが74.0%、さらにPPLX 27B＋Computerでは85.4%だった。各タスク3回、95%信頼区間を掲載している。 [30]

| 評価 | Portable / Computer | 比較対象 | 読み方 |
|---|---|---|---|
| Local Knowledge Work Bench・Qwen 3.8 27B | **82.6%** | Pi 77.6%、Hermes 74.0% | Perplexity内部53タスク。独立ベンチではない。 [10] |
| 同・PPLX 27B | **85.4%** | Qwen＋Computer 82.6% | Post-trainingによる+2.8pt。 [30] |
| BrowseComp 1,266 tasks | **66.7%** | Pi 50.2%、Hermes 43.9% | Qwenは共通だが、ComputerはPerplexity Search、他は推奨Braveを使用するため検索backend差を含む。 [8] |

| 評価 | Portable / Computer | 比較対象 | 読み方 |
|---|---|---|---|
| BrowseComp平均時間 | **402.1秒** | Pi 826.0秒、Hermes 1,020.9秒 | Computerが短い。ただしネットワーク検索を含むため純粋なGPU性能値ではない。 8 |
| BrowseComp tokens/task | **852K** | Pi 2.82M、Hermes 1.01M | Context効率の改善を示唆。 8 |
| ParseBench-100 | **65.1%** | Hermes 34.6%、Pi 13.9% | マルチモーダル文書解析のPerplexity実験。 31 |
| Terminal Bench 2.1・完全ローカル | **59.6%** | — | Qwen 3.8 27Bのみ。 32 |
| Terminal Bench・Cloud adviser併用 | **73.0%** | Claude Opus 5単独 82.4% | adviser推定費用$0.415/run、Opus単独$0.65/runとPerplexityが試算。 32 |

ここから読み取れるのは、「27Bのローカルモデルが常にフロンティアモデル並み」ということではない。実際、Perplexity自身が難しいタスクではコンパクトなオンデバイスモデルがフロンティアモデルに劣ると認めている。そのため、同社の戦略は**ローカル性能を最大化し、それでも足りない部分だけリモートadviserに相談する**ものになっている。 12

また、85.4%という目立つ数字はPerplexity自身の内部53タスクによるもので、同社は今後ベンチマークをオープンソース化し、モデル学習の技術報告を公開するとしている。したがって2026年8月30日時点では、**性能優位は有望なベンダー測定値であって、第三者が完全に再現済みの結果ではない**。 33

PPLX 27Bの学習については、Perplexity Computerで観察した実際の利用ケースの分布を参考にRL環境を設計する一方、学習タスク自体は合成し、**実際のユーザードキュメントやユーザー情報を含めない**と説明している。Dockerコンテナ内の検証可能タスクを使い、rejection fine-tuningの後にreinforcement learningを行う二段階方式である。 33

## 主要メディア報道の検証

発表直後の報道は概ね「ローカルプライバシー」「トークン課金回避」「NVIDIAハードウェア」の三点に集中している。ただし**対応ハードウェアと料金プランについて二次報道と一次情報に若干の差**があるため、導入判断では公式ページを優先すべきである。 34

| 媒体 | 報道内容・短い引用 | 評価 | 出典URL |
|---|---|---|---|
| **The Verge** | Portable ComputerはAIモデルをローカルで動かし、高度なresearch/reasoningでクラウドが必要な場合に許可を求める、と整理。引用：“runs AI models fully locally and will ask for permission...” 35 | 公式説明と整合。DGX Spark先行、RTX PC後続と正確に区別。 | https://www.theverge.com/ai-artificial-intelligence/984368/perplexitys-new-portable-computer-runs-entirely-on-device |

| 媒体 | 報道内容・短い引用 | 評価 | 出典URL |
|---|---|---|---|
| **VentureBeat** | モデル、harness、inference engine、tools、connectors、security sandboxをまとめて提供すると報道。24GB以上のRTX GPU、Enterpriseプランも対象と記載。 28 | 技術取材は詳細。ただし「24GB RTXがlaunch時点から可」「Enterprise対応」は公式発表のDGX Spark/Pro・Maxという記述より広い。更新時期または取材情報の差に注意。 | https://venturebeat.com/infrastructure/perplexity-partners-with-nvidia-to-launch-portable-computer-a-fully-local-ai-agent-with-zero-token-costs |
| **マイナビニュース** | 日本語記事として技術構成まで詳細。引用：「クラウド側のAIモデルはテキストによる助言を返すだけで、ローカルのファイルやツールには直接アクセスしない。」 36 | 本件の日本語記事の中では一次資料を比較的正確に反映。 | https://news.mynavi.jp/article/20260826-4869384/ |
| **GIGAZINE** | 「Perplexity Computer」の完全ローカル版と表現し、公式Xを引用。モデル、harness、orchestrator、tools、sandboxがローカルで動くと解説。 37 | 概要把握に有用。ただし「orchestrator LLM」というX側表現について研究記事との用語差には触れていない。 | https://gigazine.net/news/20260826-perplexity-portable-computer/ |
| **AI Business** | Portable Computerの狙いを引用すると “high token costs and the need for privacy”。一方で必要ハードウェアの高価格を課題と分析。 38 | 市場・企業導入上の論点を捉えている。 | https://aibusiness.com/generative-ai/challenges-with-perplexity-portable-computer |
| **TechCrunch** | 8月30日時点のサイト限定調査ではPortable Computer専用記事を確認できず。2月のComputerは「entirely in the cloud」と報道し、5月にはMac向けPersonal Computerを扱っている。 39 | Portableの位置づけを理解する前史として有用だが、今回の発表を直接報じる記事としては扱えない。 | https://techcrunch.com/2026/02/27/perplexitys-new-computer-is-another-bet-that-users-need-many-ai-models/ |
| **WIRED** | 専用記事は本調査では確認できず。8月30日公開のローカルLLM一般記事では利点を “offline access and greater privacy” と説明。 40 | Portable固有の報道ではなく、カテゴリ全体の背景資料。 | https://www.wired.com/story/how-to-run-your-own-local-llm/ |

| 媒体 | 報道内容・短い引用 | 評価 | 出典URL |
|---|---|---|---|
| **日経 / 日経 xTECH** | 本調査のサイト限定検索ではPortable Computerに特化した記事を確認できず。 | したがって本報告では日経記事を根拠として使用していない。 | — |
| **ITmedia** | 本調査時点ではPortable専用記事を確認できず、検索結果は過去のPerplexity解説等が中心。 41 | 今回の技術仕様の根拠には使用しない。 | https://www.itmedia.co.jp/news/article/2407/05/1240705168/3 |

特に注意すべきなのが**ハードウェア要件**である。VentureBeatはPerplexity幹部への取材に基づき「最低24GB VRAMのRTX GPU」を報じている一方、8月25日のPerplexity公式発表とNVIDIA公式ブログは「DGX Sparkで提供、GeForce RTX/RTX PRO対応はcoming soon」としている。2026年8月30日時点で導入仕様を確定するなら、**24GB VRAMを正式な一般RTXローンチ要件とはまだみなさず、二次報道値として扱う**のが妥当である。 42

料金についても同様で、Perplexity公式ブログはDGX Sparkを使う**ProおよびMax subscribers**を明記する一方、VentureBeatはEnterprise Pro/Maxも含むと報道する。現在の標準価格はProが月20ドルまたは年200ドル、Maxが月200ドルまたは年2,000ドル、Enterprise Proが1席月40ドル、Enterprise Maxが1席月325ドルである。Portableの正式なEnterprise提供条件は契約前に確認すべきである。 43

## 競合ローカルAIエージェントとの比較

Portable Computerの直接競合を定義するのは難しい。AppleはOS統合型、Anthropic Claude Codeは**クライアントはローカルだがLLM推論はクラウド**、OpenClawは柔軟な自主管理型、Ollamaは主としてローカルモデル実行基盤である。したがって以下では「ユーザー側マシン上でツールや個人データを扱うエージェント」という広義の競合カテゴリで比較する。 44

| 製品 / 構成 | LLM推論 | Agent / Tool 実行 | オフライン | プライバシー設計 | 対応環境 | 価格構造 | 性能比較 |
|---|---|---|---|---|---|---|---|
| **Perplexity Portable Computer** | Qwen 3.8 27B / PPLX 27Bをローカル。必要時のみクラウド frontier model。 18 | 決定論的harness、planner、router、scheduler、sandbox、queue、indexまでlocal。 23 | **Yes**。Web全無効化可。 8 | outbound承認、PII classifier、adviserはfile/tool直アクセス不可。 12 | DGX Spark/Linux先行。RTX、Windows予定。 16 | Pro $20/月、Max $200/月＋hardware。Local処理はComputer credits不使用。 45 | Perplexity測定でLocal Work 82.6%、PPLX 85.4%。 30 |

| 製品 / 構成 | LLM推論 | Agent / Tool 実行 | オフライン | プライバシー設計 | 対応環境 | 価格構造 | 性能比較 |
|---|---|---|---|---|---|---|---|
| **Apple Intelligence + Siri AI + Shortcuts** | 多くをon-device、複雑処理はPrivate Cloud Compute。[46] | Siri/App Intents/Shortcuts経由でOS・アプリのactions。Shortcutsではon-device/PCCモデルをワークフローに組込み可能。[47] | **部分的**。on-device機能はローカル、PCC/Web依存機能は不可。 | AppleはPCCデータを要求処理だけに使い保存しないとしている。[48] | Apple Intelligence対応Apple機器 | AI単体の別建て従量料金なし。対応hardwareコストが必要。 | Portableと同一agent benchmarkなし。 |
| **Anthropic Claude Code / Cowork** | 通常はAnthropic model API、すなわち**リモート推論**。[49] | Terminal/IDEでcode、Git、tests、MCP等を操作。ローカルclientがtoolを扱う。[49] | **原則No**。Claude API利用にネット接続が必要。 | ローカルterminalからmodel APIへ直接接続し、remote code index不要。file変更・command前にpermission。[49] | macOS / Linux / Windows。[49] | Pro/MaxまたはAPI課金。Maxは$200/月。[49] | Claude Code製品との同一benchmarkなし。PerplexityのTerminal Benchでは Opus 5 model単独82.4%だが別条件。 |
| **OpenClaw + Local Model** | backendを選択でき、Ollama等のlocal model運用が可能。[50] | 1つのGatewayからmodel、tools、messaging channels、appsを接続。[51] | Local model＋local toolsなら可能。Web機能はnetwork必要。[52] | 自主管理の自由度が高い一方、権限・sandbox等は構成依存。 | 複数OS・個人/チーム構成 | モデル・hardware・provider構成依存 | Portableとの独立同一条件比較なし。 |
| **Ollama + Claude Code / OpenCode / Codex等** | 任意のlocal model、またはOllama cloud。[53] | `ollama launch`でcoding agentsをlocal/cloud modelに接続。tool callingあり。[54] | Local構成では**Yes**。 | OpenAI互換local APIを提供し、local toolingを構築可能。[55] | macOSを含む複数local hardware構成 | 構成依存。hardwareと選択modelが主コスト | RuntimeなのでPortableの統合agent stackとは直接比較不可。 |

Appleとの差は特に明確である。Apple IntelligenceもローカルとPrivate Cloud Computeを自動的に使い分けるため、哲学はPortableに近い。しかしAppleは**OSレベルで個人コンテキストとアプリアクションを統合する垂直統合モデル**、Portableは**大きなローカルGPU上で長時間・多段階のknowledge-work agentを走らせるモデル**という違いがある。Appleは2026年版Siri AIで個人メッセージ、メール、写真などのコンテキストとアクション統合をさらに拡大している。 [56]

Anthropicとの差はさらに本質的で、Claude Codeは「エージェントclient・toolsはローカル、知能はクラウド」である。Anthropic自身も、Claude Codeがローカルterminalで動作し、backend serverやremote code indexを必要としない一方、model APIsへ直接通信すると説明している。Portableは**モデル推論そのものもlocalにする**ことが最大の差である。 [49]

一方、OpenClawやOllamaを組み合わせれば、ユーザーは既に完全ローカルに近いAI agentを自作できる。Portable Computerの競争力は「ローカルLLMを実行できること」自体より、**モデル、推論runtime、harness、connector、sandbox、Search、post-trainingを一つの製品として最適化したこと**にある。この意味で本当の競争軸はモデル性能だけでなく「agent harness engineering」になっている。 [57]

## 利用シナリオ、導入リスク、市場インパクト

Portable Computerが最も適しているのは、**大量のローカル文書・コード・業務ファイルをAIに読ませたいが、それらを毎回クラウドLLMへ送ることに抵抗があるユーザー**である。Perplexity自身の53タスク評価は、deep research、data/finance/procurement、文書・プレゼン作成、engineering/IT incidents、contract/evidence/compliance、dashboard/software/visualization、people/project/meetingという7分野を対象としており、狙っているのは単なるチャットではなくknowledge work全般である。 [33]

代表的な利用例としては、機密コードベースの解析と修正、契約書・訴訟資料・監査資料の横断分析、社内財務データと公開市場データを組み合わせた調査、ローカルPDF・表・グラフからのレポート生成、夜間に走らせる長時間タスク、GmailのバグGitHub issueをtriageしSlackへ結果を送るようなクロスアプリworkflowが考えられる。最後の例はPerplexity自身が具体例として挙げている。 [18]

**導入メリット**は三つある。第一はデータ境界で、通常処理をローカルに閉じることで、機密データをremote modelへ毎回送信する必要がない。第二は経済性で、ローカル推論はComputer creditsを消費しないため、GPUを高稼働させるユーザーほど従量課金を避ける価値が大きくなる。第三は運用制御で、インターネットを切った完全オフラインモードと、必要時だけSearch/adviserへ接続するモードを使い分けられる。 [58]

ただし「local＝安全」ではない。OWASPのAgentic AIセキュリティ指針では、AI agentには**indirect prompt injection、tool abuse / privilege escalation、data exfiltration**などが主要リスクとして挙げられている。例えば悪意ある指示を埋め込んだメールやPDFをagentが読み、それを正規の命令と誤認してGitHub、ファイルシステム、ネットワークtoolを操作する攻撃は、モデルがクラウドにあるかローカルにあるかとは独立して成立する。 [59]

PortableのOSレベルsandbox、filesystem/network policy、adviserに直接tool権限を渡さない設計は、このリスクの“blast radius”を小さくする方向に働く。しかしsandboxの具体実装、root権限との関係、認証情報の保管方式、at-rest encryption、telemetry、patch/update chainについては発表資料で十分に開示されていない。そのため、企業導入では「ローカルだからセキュリティ審査不要」ではなく、**endpoint security製品として通常のコード実行基盤以上に厳しい脅威モデルを適用するべき**である。前半は一次情報、後半はその情報とOWASP脅威モデルからの分析である。 [60]

クラウドエスカレーションにも運用上の落とし穴がある。PII classifierとユーザー承認は優れた制御策だが、分類器のfalse negative率は公開されておらず、毎回の承認を自動化すれば誤送信のリスクは当然増える。逆に手動承認を多用すればagentの自律性と生産性が落ちる。したがって実運用では「通常は完全local」「公

開情報だけSearchを許可」「特定workflowのみadviser自動承認」のように、**データ分類別ポリシー**を設定するのが合理的である。PII分類器と手動/自動承認の存在自体はPerplexity公式研究記事に記載されている。[12]

日本で個人データを扱う場合、個人情報保護委員会は生成AIサービス利用に関して事業者・行政機関・一般利用者向けの注意喚起を出している。ローカル処理は第三者への送信を減らせる点で有利だが、Gmail、Drive、Web、クラウドadviserへデータを出すworkflowでは、個人情報の利用目的、委託・第三者提供、安全管理措置等の確認が不要になるわけではない。具体的な法的適用は事業者の役割と処理内容による。[61]

日本の参考資料：
https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/

EUではAI Actの大部分が**2026年8月2日から適用・執行段階**に入り、Article 50の透明性義務も適用されている。生成・操作されたコンテンツのマーキングやAIとの対話に関する透明性要件などが対象となる。Portable Computerをローカルで動かすこと自体がAI Actからの包括的免除になるわけではなく、deployするユースケースやprovider/deployerとしての役割によって義務を判断する必要がある。[62]

**コスト面**では「token cost zero」と「total cost zero」を混同してはいけない。DGX Spark Founders Editionは2026年2月にMSRPが3,999ドルから**4,699ドル**へ引き上げられ、NVIDIA Marketplaceでも4,699ドルで販売されている。さらにPerplexity Proは20ドル/月、Maxは200ドル/月である。電力、GPU減価償却、patching、endpoint管理、ストレージ、バックアップまで考えれば、TCOはクラウドAPIより必ず安いとは限らない。高稼働ほどlocalの経済性が改善し、低稼働ほどcloud従量制が有利になるというのが一般的な損益構造である。[63]

## 短期の市場インパクト

短期的には、Portable Computerは「ローカルAI PC」の市場を**モデル実験用マシンから、常時稼働する業務agent applianceへ広げる**効果を持つ可能性がある。NVIDIA自身がDGX Sparkを24×7のlong-running autonomous agentに適したマシンとして訴求し、Portableをその代表アプリとして紹介しているためである。これはNVIDIAにとってDGX Spark、将来的にはRTX/RTX PROの新しい需要喚起になる。[16]

Perplexity側では、検索サービスから「仕事を最後まで実行するagent runtime」へ製品ポジションを広げられる。2026年前半にクラウド版Computer、MacのPersonal Computerを展開した流れにPortableが加わったことから、同社が検索だけでなく**ユーザーのローカルコンピューティング環境そのものをagentの実行面として取り込もうとしている**と読むのが自然である。これは製品展開からの推論である。[64]

一方で当初の普及速度を抑える最大要因は、4,699ドル級DGX SparkとLinuxという入口の狭さだろう。NVIDIAとPerplexityの双方がRTXとWindows対応を予告しているため、**一般RTX PC対応の品質と時期が本当の普及転換点**になる可能性が高い。[65]

## 中期の市場インパクト

中期的により重要なのは、Portable Computerが提示した「**local default, cloud escalation**」という経済モデルである。Terminal Benchで、完全ローカルQwenが59.6%、Claude Opus 5 adviser併用が73.0%、Opus 5単独が82.4%だったというPerplexityの結果は、ローカルモデルを常用し、難問だけ高価なfrontier modelへ相談することで性能差の一部を埋められることを示している。あくまで1ベンダー・1コーディング評価だが、エージェント経済の方向性を示す結果としては興味深い。[32]

この方式が普及すると、AI利用コストは「全トークンをクラウドで購入」から、**固定CapExとしてGPUを持ち、クラウド推論は例外的なプレミアムサービスとして購入する**構造へ一部移る可能性がある。Appleも既に

on-device＋Private Cloud Computeという別形態のハイブリッド設計を採用しており、ローカルとクラウドを排他的に選ぶのではなく、request単位でroutingする考え方は大手各社で収斂しつつある。これはPortableとAppleの公開アーキテクチャを比較した分析である。 66

**投資家への示唆**としては、NVIDIAにはローカルagentという新しいGPU需要源が生まれる可能性がある。Perplexityには、推論コストの一部をユーザー側hardwareへ移しながらsubscription revenueを維持できる余地があり、特に高頻度Computer workloadでgross margin上のメリットが生じる可能性がある。ただしPerplexityはPortableの単位経済性や原価削減額を公表しておらず、Search、connectors、cloud adviser、アップデート配布等には引き続きサービス側コストが発生するため、**「local化＝Perplexityの原価ゼロ」と見るのは誤り**である。前者はアーキテクチャからの推論であり、財務効果は未開示である。 19

逆に投資上のリスクは、OpenClaw、Ollama、オープンモデルの改善による**ローカルagent stackのコモディティ化**である。Ollamaは既にClaude Code、OpenCode、Codex等をlocal/cloud modelへ接続する`ollama launch`を提供し、OpenClawもlocal modelと組み合わせられる。このためPerplexityのmoatは「ローカルLLMを起動する機能」ではなく、Search、harness最適化、sandbox、connector UX、post-trained model、Computer ecosystemの統合品質に依存する。 67

**開発者への示唆**はさらに明確である。Portableの研究結果が正しければ、今後のagent性能はモデルサイズだけで決まらず、コンテキストの節約、Skillsのオンデマンドロード、決定論的orchestrator、self-verification、sandbox、tool schema設計が同等以上に重要になる。特にPerplexityが大きなMCP tool definitionsをcompact CLIへ変換している点は、「MCPを接続すれば終わり」ではなく、**限られたローカルcontext windowをどう使うかがagent engineeringの主要競争領域になる**ことを示している。 4

ただし開発者プラットフォームとしてはまだ閉じている。2026年8月30日時点でPortable専用SDK、公開agent harness、PPLX 27B weights、公開benchmark repoは確認できず、benchmarkについてもPerplexityは「今後open sourceにする」としている段階である。そのため現時点のPortableは、**OSSフレームワークというより、Perplexityが統合・最適化したローカルAI appliance/software product**と評価するのが適切である。 33

## 参考技術資料と総合評価

関連技術を検証する場合、二次ニュースより以下の一次資料を優先するのがよい。

| 資料 | 何が分かるか | URL |
|---|---|---|
| **Perplexity公式「Portable Computerのご紹介」日本語版** | 製品仕様、モデル、local/cloud境界、対応環境 | https://www.perplexity.ai/ja/hub/blog/introducing-portable-computer-for-local-first-ai |
| **Perplexity “A Local-First Agent for Private and Cost-Effective Knowledge Work”** | harness設計、sandbox、PII、adviser、benchmark、post-training | https://www.perplexity.ai/hub/blog/a-local-first-agent-for-private-and-cost-effective-knowledge-work |
| **Perplexity Research “Rethinking Search as Code Generation”** | Agentic Search SDK、sandbox、code-driven search architecture。 21 | https://research.perplexity.ai/articles/rethinking-search-as-code-generation |

| 資料 | 何が分かるか | URL |
|---|---|---|
| **NVIDIA Local AI Blog** | DGX Spark最適化、RTX/RTX PRO・Windowsロードマップ | https://blogs.nvidia.com/blog/local-ai-open-source-models-agents-nemotron/ |
| **NVIDIA DGX Spark** | 128GB unified memory等のhardware仕様。 29 | https://www.nvidia.com/en-us/products/workstations/dgx-spark/ |
| **Qwen 3.8 27B model card** | Portableのbase local model。 68 | https://huggingface.co/Qwen/Qwen3.8-27B |
| **NVIDIA Nemotron 3.5 Lightning 30B** | 今後Portableへ追加予定のモデル。 69 | https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-NVFP4 |
| **Model Context Protocol** | Portableが一部toolsをcompact CLI化する対象となる標準。 70 | https://github.com/modelcontextprotocol |
| **ParseBench** | 文書・table・chart解析benchmark。 71 | https://huggingface.co/datasets/llamaindex/ParseBench |
| **Terminal-Bench 2.1** | Agent coding能力評価に利用。 72 | https://github.com/harbor-framework/terminal-bench-2-1 |
| **OWASP AI Agent Security Cheat Sheet** | Prompt injection、tool abuse、data exfiltration等の導入リスク。 73 | https://cheatsheetseries.owasp.org/cheatsheets/AI_Agent_Security_Cheat_Sheet.html |
| **OWASP Top 10 for Agentic Applications 2026** | Agentシステム固有の最新セキュリティリスク。 74 | https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/ |
| **個人情報保護委員会・生成AI利用の注意喚起** | 日本の個人情報保護上の基本的留意事項。 75 | https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/ |
| **EU Commission AI Act** | 2026年8月時点の適用・執行状況。 76 | https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai |

総合すると、Portable Computerの技術的な意義は「Perplexityをオフラインで使えるようにした」ことだけではない。より本質的なのは、**ローカルモデルを“第一選択の知能”に据え、ローカルで実行権限を保持する決定論的harnessを中心に置き、Web・connector・frontier modelを権限管理された外部ツールへ格下げしたこと**である。この構造は、従来の「クラウドLLMがagentの頭脳で、PCはその手足」という構図をかなり反転させている。 77

同時に、発表文の「No cloud dependency」を「一切ネットワークを使わない製品」と解釈するのも正確ではない。Portableは**クラウドに依存せずに成立できるが、クラウドを意図的に利用できる**製品である。完全オフライン、ローカル＋Web検索、ローカル＋connector、ローカル＋frontier adviserという複数のprivacy/capabilityレベルを連続的に選べることこそが設計上の強みといえる。 2

現時点で最も慎重に見るべきなのは、性能評価の独立性と企業向け運用仕様である。85.4%の内部Knowledge Work Benchは魅力的だがまだ未公開であり、Portable固有のAPI/SDK、PPLX 27B weights、

telemetry、保存暗号化、ログ管理、software supply-chain、enterprise policy controlsの詳細も十分には公開されていない。**技術コンセプトはかなり完成度が高い一方、エンタープライズ向けの検証可能性・管理性については今後の技術文書と第三者評価を待つ余地が大きい**、というのが2026年8月30日時点での最も厳密な評価である。 [78]

市場全体への示唆は大きい。Portable Computer、Apple Intelligence/PCC、Ollama/OpenClaw系の進化を合わせてみると、AI agent市場の次の競争は「最大モデルをクラウドで呼べるか」だけではなく、**どの仕事を端末内で終わらせ、何を外へ出し、どの権限でツールを動かし、難問だけをどのクラウドモデルへrouteするか**という“知能の配置と統治”へ移りつつある。Portable Computerは、その流れを商用knowledge-work agentとしてかなり明瞭に形にした製品である。 [79]

🟫navlist🟫Portable Computerの直近関連ニュース🟫turn0news15🟫

---


[1] [3] [7] [18] [19] [20] [34] [43] [58] Portable Computerのご紹介
https://www.perplexity.ai/ja/hub/blog/introducing-portable-computer-for-local-first-ai

[2] [4] [8] [9] [10] [12] [15] [17] [22] [23] [24] [25] [26] [30] [31] [32] [33] [57] [60] [77] [78] [79] Un agent local pour le travail de la connaissance privé
https://www.perplexity.ai/fr/hub/blog/a-local-first-agent-for-private-and-cost-effective-knowledge-work

[5] [14] Today we're launching Portable Computer on @NVIDIA ...
https://x.com/perplexity_ai/status/2092268362386780270?utm_source=chatgpt.com

[6] [63] 2/23/2026 Price Change Announcement
https://forums.developer.nvidia.com/t/2-23-2026-price-change-announcement/361713?utm_source=chatgpt.com

[11] NVIDIA DGX Spark
https://marketplace.nvidia.com/en-us/enterprise/personal-ai-supercomputers/dgx-spark/?utm_source=chatgpt.com

[13] Introducing Portable Computer for local-first AI
https://www.perplexity.ai/uk/hub/blog/introducing-portable-computer-for-local-first-ai?utm_source=chatgpt.com

[16] [65] NVIDIA and Local AI Community Fuel Open Source Models and Intelligent Agents | NVIDIA Blog
https://blogs.nvidia.com/blog/local-ai-open-source-models-agents-nemotron/

[21] Rethinking Search as Code Generation
https://research.perplexity.ai/articles/rethinking-search-as-code-generation

[27] Perplexity's Personal Computer is now available to ...
https://techcrunch.com/2026/05/07/perplexitys-personal-computer-is-now-available-everyone-on-mac/?utm_source=chatgpt.com

[28] [42] Perplexity partners with Nvidia to launch Portable Computer, a fully local AI agent with zero token costs | VentureBeat
https://venturebeat.com/infrastructure/perplexity-partners-with-nvidia-to-launch-portable-computer-a-fully-local-ai-agent-with-zero-token-costs

[29] NVIDIA DGX Spark: AI Supercomputer on Your Desk
https://www.nvidia.com/en-us/products/workstations/dgx-spark/?utm_source=chatgpt.com

[35] Perplexity’s new Portable Computer runs “entirely on device.” | The Verge
https://www.theverge.com/ai-artificial-intelligence/984368/perplexitys-new-portable-computer-runs-entirely-on-device

36 Perplexity、ローカル実行型AIエージェント「Portable Computer」発表　NVIDIAと共同開発 | マイナビニュース
https://news.mynavi.jp/article/20260826-4869384/

37 「Perplexity Computer」の完全ローカル版「Portable Computer」をPerplexityがリリース - GIGAZINE
https://gigazine.net/news/20260826-perplexity-portable-computer/

38 Challenges With Perplexity Portable Computer
https://aibusiness.com/generative-ai/challenges-with-perplexity-portable-computer

39 64 Perplexity's new Computer is another bet that users need ...
https://techcrunch.com/2026/02/27/perplexitys-new-computer-is-another-bet-that-users-need-many-ai-models/?utm_source=chatgpt.com

40 How to Run a Chatbot on Your Own Computer
https://www.wired.com/story/how-to-run-your-own-local-llm/?utm_source=chatgpt.com

41 Google検索も不要に？ 検索AI「Perplexity」がスゴすぎて ...
https://www.itmedia.co.jp/news/article/2407/05/1240705168/3?utm_source=chatgpt.com

44 47 macOS 26 makes the Mac more powerful, productive, and intelligent than ever - Apple (IL)
https://www.apple.com/il/newsroom/2025/06/macos-tahoe-26-makes-the-mac-more-capable-productive-and-intelligent-than-ever/

45 Using the Connector for Slack | Perplexity Help Center
https://www.perplexity.ai/help-center/en/articles/12167980-using-the-connector-for-slack.html?utm_source=chatgpt.com

46 66 Apple Intelligence brings powerful AI capabilities into ...
https://www.apple.com/ml/newsroom/2026/06/apple-intelligence-brings-powerful-ai-capabilities-into-everyday-experiences/?utm_source=chatgpt.com

48 Apple Intelligence and privacy on iPhone
https://support.apple.com/guide/iphone/apple-intelligence-and-privacy-iphe3f499e0e/ios?utm_source=chatgpt.com

49 Claude Code by Anthropic | AI Coding Agent, Terminal, IDE
https://www.anthropic.com/claude-code

50 51 GitHub - openclaw/openclaw: Your own personal AI assistant. Any OS. Any Platform. The lobster way. 🦞 · GitHub
https://github.com/openclaw/openclaw

52 The simplest and fastest way to setup OpenClaw
https://ollama.com/blog/openclaw-tutorial?utm_source=chatgpt.com

53 67 ollama launch · Ollama Blog
https://ollama.com/blog/launch?utm_source=chatgpt.com

54 Tool support · Ollama Blog
https://ollama.com/blog/tool-support?utm_source=chatgpt.com

55 OpenAI compatibility · Ollama Blog
https://ollama.com/blog/openai-compatibility?utm_source=chatgpt.com

56 Apple introduces Siri AI, a profoundly more capable and ...
https://www.apple.com/newsroom/2026/06/apple-introduces-siri-ai-a-profoundly-more-capable-and-personal-assistant/?utm_source=chatgpt.com

59 74 OWASP Top 10 for Agentic Applications for 2026
https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/?utm_source=chatgpt.com

61 75 生成AIサービスの利用に関する注意喚起等について
https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/?utm_source=chatgpt.com

62 76 AI Act | Shaping Europe's digital future - Europa.eu
https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai?utm_source=chatgpt.com

68 Qwen/Qwen3.8-27B · Hugging Face
https://huggingface.co/Qwen/Qwen3.8-27B

69 nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-NVFP4 · Hugging Face
https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-NVFP4

70 Model Context Protocol · GitHub
https://github.com/modelcontextprotocol

71 llamaindex/ParseBench · Datasets at Hugging Face
https://huggingface.co/datasets/llamaindex/ParseBench

72 GitHub - harbor-framework/terminal-bench-2-1: Terminal-Bench 2.1 · GitHub
https://github.com/harbor-framework/terminal-bench-2-1

73 AI Agent Security Cheat Sheet
https://cheatsheetseries.owasp.org/cheatsheets/AI_Agent_Security_Cheat_Sheet.html?utm_source=chatgpt.com