面壁智能の分析レポート

ChatGPT-5.5

Executive Summary

面壁智能は、2022 年 8 月に設立された「端側 AI」を主軸に置く中国の大模型スタートアップであり、研究源流は清华大学[1]の自然言語処理系チームにある。公開 DB と大学・登壇者プロフィールを突き合わせると、共同創業の科学中核は刘知远[2]と曾国洋[3]、経営・商業化の顔は 2023 年 6 月に加わった前⊠ entity ⊠ ["company","知乎","q&a platform china"]□CTO の李大海[4]である、という理解が最も整合的である。資金調達は 2023 年の天使輪以降、2024 年 4 月、2024 年 12 月、2025 年 5 月、2025 年 12 月、2026 年 2 月、2026 年 4 月と高頻度で続き、2026 年 4 月時点では「Q1 累計で 10 億元超」、投後評価額は「独角獣閾値到達」と報じられている。一方で、各ラウンドの正確な金額と評価額は大半が未公表である。[5]

技術面では、MiniCPM シリーズは単なる「小型 LLM」ではなく、言語・視覚・音声・全二重全モーダルまでを一貫して揃えた端側ファミリーへ進化している。テキスト系では trainable sparse attention、speculative decoding、ternary quantization、FP8 学習、効率的データ生成、クロスプラットフォーム推論基盤を組み合わせ、4.1 系では同社が「原生稀疏架構の深思考モデル」と位置づける推論強化を実装した。マルチモーダル系では、MiniCPM-V 4.0/4.5 が高密度動画圧縮と高解像 OCR を、MiniCPM-o 2.6/4.5 が画像・動画・音声・テキストのリアルタイム統合を前に進めている。技術の本質は「大きさ」ではなく「密度」と「端側実装の完成度」にある。[6]

商業面では、公開統計だけでも traction は強い。GitHub 上の MiniCPM リポジトリは約 8.9k stars、MiniCPM-o は約 24.5k stars を持ち、主要 Hugging Face モデルの月間ダウンロードは MiniCPM4.1-8B が 49,105、MiniCPM4-0.5B が 24,617、MiniCPM-V 4.0 が 160,526、MiniCPM-V 4.5 が 132,169、MiniCPM-o 2.6 が 227,692

である。さらに、会社発表ベースでは MiniCPM 系列の累計ダウンロードは GitHub/Hugging Face 等で 2,400 万超とされる。量産導入では、端側スマートコックピット「SuperMate」が长安马自达[7] EZ-60、吉利汽车[8]系の吉利銀河 M9 などに載り、法律領域では法院専網への私有配備も確認できる。[9]

投資家視点での結論は、「技術志向の研究組織が、ようやく量産・B2B/B2G の収益化入口に立った局面」である。強みは、端側に最適化されたモデル密度、Apache-2.0 による開放性、スマホ・AIPC・車載・司法・ロボットまで見据えた横展開力、そして国家系・産業系資本の継続加注である。他方、弱みは、売上や ARR の未開示、OEM サイクル依存、競争の激しい中国開源モデル市場、そして中国の備案・标识規制と米国の半導体輸出規制が同時にかかる点にある。今後 6〜18 か月で見るべきは、「ダウンロード」ではなく「量産設計採用数」「有償案件継続率」「車載・端末の実搭載台数」「推論コスト/メモリ効率の相対優位」が維持されるかである。[10]

会社概要と資本政策

面壁智能は、北京の北京面壁智能科技有限责任公司として 2022 年 8 月 12 日に設立された。36Kr のプロジェクト DB では、会社を「人工智能大模型加速与应用落地赋能商」と位置づけ、共同創業者として劉知遠を掲げている。法定代表人は曾国洋であり、公開検索可能な工商スニペットでもこの構図は一致する。公開メディアの時系列を加えると、研究起点は Tsinghua NLP 系、経営の転換点は 2023 年 6 月の李大海参画とみるのが妥当である。[11]

劉知遠は清華大学計算機系で学士・博士を取得した研究者で、大学公式ページでは副教授、別の公開プロフィールでは教授表記も見える。いずれのページでも自然言語処理・知識グラフ・社会計算が主軸であり、THUNLP/OpenBMB/面壁の研究連携を支える中核人物である。李大海は公開登壇プロフィール上、北京大学数学系修士、Google 中国初期メンバー、豌豆荚の検索責任者、知乎の合伙人兼 CTO という経歴を持ち、面壁では共同創業者兼 CEO を務める。曾国洋については、天使輪報道と 36Kr プロフィールの双方が、悟道・文源系の中核メンバーであり面壁の共同創業者兼 CTO/法定代表人であることを示している。[12]

資金調達の時系列

時点	ラウンド	金額	主な投資家	評価額 / 備考
2023-04	天使輪	数千万元人民币	☒ entity ☒["company","知乎","q&a platform china"]□主導、智谱 AI[13] 継続出資	金額詳細・評価額は未公表
2024-04	未公表	数亿元人民币	春华创投[14]、华为哈勃[15] 主導、北京市人工智能产业投资基金[16] など、知乎継続支援	ラウンド呼称は公開資料で不統一
2024-12	A+ 輪	数亿元人民币	龙芯创投[17]、鼎晖百孚[18]、中关村科学城基金[19]、赛富投资基金[20] 主導、北京市 AI 基金、清科创投[21] など	評価額未公表
2025-05	B 輪	数亿元人民币	洪泰基金[22]、国中资本[23]、清控金信资本[24]、茅台基金[25] など	評価額未公表
2025-12	C 輪	数亿元人民币	京国瑞投资[26]、国科投资[27]、中金资本私募、米聚资本、和基投资[28]	評価額未公表
2026-02	D 輪	数亿元人民币	中国电信投资[29] 主導、中信金石、中信私募	Q1 の大型調達の一部
2026-04	D+ 輪	数亿元人民币	深创投[30]、汇川产投[31] 主導、道禾长期投资、国泰君安创新投[32]、武岳峰科创[33]	「Q1 累計 10 億元超」「独角獣閾値到達」と報道

表の作成にあたっては、会社発表を流した主要経済メディアと 36Kr プロジェクト DB の一致範囲を採用した。2024 年 4 月ラウンドは「数亿元融资」とのみ開示さ

れ、A 輪相当か否かは未確認である。2022 年の種子輪については、後年記事で「近千万元」「種子輪株主に智譜 AI」との回想があるが、投資家構成・金額・評価額の一次確認までは至っていないため、本表では独立行としては採用せず「未確認」とした。[34]

株主構成と出資比率

現時点で検索可能な公開資料では、**最新の完全な cap table と各株主の現持分比率は未特定**である。確認できるのは二層で、第一に 36Kr DB に残る初期工商スナップショット、第二にその後の工商変更報道である。初期スナップショットでは、曾国洋 30.77%、北京清语启航科技中心（有限合伙）[35] 30.77%、北京清语创智科技中心（有限合伙）[36] 17.58% が見え、創業者・持株平台主導の構図が示唆される。[37]

その後、2025 年 7 月時点の工商変更報道では、北京市 AI 産業投資基金、智譜、深圳ハッブなどの継続株主が見え、2026 年 3 月には中国電信投資と广西人工智能产业投资基金合伙企业（有限合伙）[38] の追加が報じられている。したがって、現状は「創業者・持株平台＋国資基金＋産業資本＋戦略投資家」の混成型とみられるが、**正確な出資比率は未確認**と記すのが厳密である。[39]

```
timeline
    title 面壁智能の主要マイルストーン
    2022-08 : 会社設立
    2023-04 : 天使輪調達
    2023-06 : 李大海が共同創業者兼 CEO として参画
    2024-02 : MiniCPM-2B 公開
    2024-04 : 数億元の成長資金を調達
    2024-12 : A+輪
    2025-01 : MiniCPM-o 2.6
    2025-05 : B 輪
    2025-08 : MiniCPM-V 4.5
    2025-09 : MiniCPM 4.1
    2025-12 : C 輪
```

2026-02 : D 輪 + MiniCPM-o 4.5 / MiniCPM-SALA
2026-04 : D+輪 + SuperMate 新版発表

上のタイムラインは、設立・資金調達・主要開源リリース・量産化イベントを一本に束ねたものである。2024 年 2 月の MiniCPM-2B、2025 年 1 月の MiniCPM-o 2.6、2025 年 8 月の MiniCPM-V 4.5、2025 年 9 月の MiniCPM 4.1、2026 年 2 月の MiniCPM-o 4.5 と MiniCPM-SALA、2026 年 4 月の SuperMate 升级は、技術の「テキスト → 視覚 → 音声/全モーダル → 量産ソリューション」への拡張を示している。[40]

技術と製品スタック

面壁智能の技術的な核は、会社サイトが繰り返し掲げる「密度法則」にある。これは、限られた算力でモデルの知識密度とタスク性能を最大化するという発想で、巨大化よりも効率改善を主戦場に置く。公式サイト・GitHub・Hugging Face の 3 系統を読むと、MiniCPM は単一モデルではなく、言語モデル群、視覚言語モデル群、音声・全モーダルモデル群、さらに推論・量子化・デプロイ基盤までを含む「端側スタック」として設計されている。[41]

```
flowchart LR
    A[研究源流<br/>THUNLP / OpenBMB] --> B[MiniCPM4<br/>テキスト]
    B --> C[MiniCPM-V 4.0 / 4.5<br/>画像・動画]
    C --> D[MiniCPM-o 2.6 / 4.5<br/>音声・全モーダル]
    B --> E[MiniCPM-SALA 9B<br/>超長文]
    B --> F[BitCPM / Int4 / GGUF / AWQ<br/>軽量化]
    B --> G[ArkInfer / CPM.cu<br/>推論基盤]
    C --> H[SuperMate<br/>端側スマートコックピット]
    D --> I[Pinea<br/>純端側アシスタント]
    B --> J[Lantay<br/>文档智能体]
    B --> K[法院専網・法律 AI]
```

この関係図は、THUNLP/OpenBMB の研究成果が MiniCPM 系のモデルとして開源され、そこから車載・アシスタント・文書 AI・法律 AI に降りていく構図を示す。重要なのは、面壁智能がコードと重みを OpenBMB ブランドで開源することでコミ

ュニティ浸透を取りつつ、商用側では SuperMate や法院専網配備のような「縦の業界ソリューション」で収益化を狙っている点である。[42]

MiniCPM 系列のアーキテクチャ要点

テキスト系列の最新中核は MiniCPM4 / 4.1 で、公式モデルカードでは 8B と 0.5B を主要スケールとしている。4.1-8B は Apache-2.0 で公開され、trainable sparse attention、frequency-ranked speculative decoding、hybrid reasoning mode を特徴とする。GitHub README は、InfLLM-V2 により 128K 長文で各トークンが 5% 未満のトークンとの関連計算で済むこと、Model Wind Tunnel 2.0、BitCPM、FP8 + multi-token prediction、UltraClean/UltraChat v2、CPM.cu、ArkInfer を「効率の 4 次元最適化」として位置づけている。4.0 は 32K 事前学習＋YaRN で 128K 拡張、4.1 は 64K 事前学習＋YaRN で 128K 拡張である。[43]

さらに 2026 年 2 月公開の MiniCPM-SALA は、25% の InfLLM-V2 と 75% の Lightning Attention を組み合わせる hybrid attention モデルで、RTX 5090 のようなコンシューマ GPU でも 1M token 近辺を扱う設計を示した。これは「端側＝短文」から「端側＝超長文処理」への射程拡張であり、将来のローカル Agent や長文 RAG 用途に直結する。[44]

マルチモーダル系列では、MiniCPM-V 4.0 が SigLIP2-400M + MiniCPM4-3B の 4.1B 構成、MiniCPM-V 4.5 が Qwen3-8B + SigLIP2-400M の 8B 構成である。4.5 の技術的ハイライトは、画像と動画を統一的に扱う 3D-Resampler による 96× の動画トークン圧縮、高解像 OCR/文書解析向けの LLaVA-UHD 流儀、多模態 RL による fast/deep thinking 切替にある。会社ブログでも、アーキテクチャ・データ・訓練レシピの三位一体で「8B で 72B 級を超える」ことを強調している。[45]

音声・全モーダル系列では、MiniCPM-o 2.6 が SigLip-400M、Whisper-medium-300M、ChatTTS-200M、Qwen2.5-7B から成る総計 8B で、 end-to-end omni-modal architecture、time-division multiplexing、configurable speech modeling を採用する。2026 年 2 月に開源された MiniCPM-o 4.5 は、SigLIP2、Whisper-medium、

CosyVoice2、Qwen3-8B を組み合わせた 9B の全二重全モーダルモデルで、画像・動画・音声・テキストをリアルタイム同時入出力できる方向へ進んだ。[46]

軽量化手法、推論速度、メモリ要件、対応ハードウェア

軽量化は四層構造で理解すると整理しやすい。第一に**アーキテクチャ最適化**として sparse attention、SALA の sparse + linear hybrid、3D-Resampler、高 pixel/token density がある。第二に**量子化**として BitCPM の ternary quantization、QAT Int4、GPTQ、AWQ、GGUF、MLX 変種が整備されている。第三に**訓練最適化**として FP8・multi-token prediction・風洞 2.0・高品質データ生成がある。第四に**推論系**として CPM.cu、ArkInfer、FlagOS/llama.cpp/ollama/vLLM/SGLang 互換があり、単に「小さい」ではなく「動かしやすい」まで含めて最適化されている。[47]

代表的な実測として、MiniCPM4 / 4.1 は Jetson AGX Orin と RTX 4090 上で同規模モデルより長文処理効率が高く、GitHub README では Jetson AGX Orin 上で Qwen3-8B 比およそ 7 倍の decoding speed improvement、4.1 では reasoning で 3 倍の decoding speed improvement とされる。マルチモーダル側では MiniCPM-V 4.0 が iPhone 16 Pro Max 上で first token delay 2 秒未満、17 tok/s 超を公式モデルカードで示し、会社サイトは旧系 MiniCPM で Snapdragon 855 7.5 tok/s、Apple A17 Pro 25 tok/s といった端末側の速度/コスト例も掲示する。論文レベルでは、MiniCPM-V のデプロイ図が Xiaomi 14 Pro（Snapdragon 8 Gen 3）での最適化効果を示している。[48]

公開資料から読めるメモリ要件は、テキスト系よりマルチモーダル系の方が具体的である。公式 GitHub の model zoo では、MiniCPM-o 4.5 は full で GPU 19GB、GGUF で 10GB、AWQ で 11GB、MiniCPM-V 4.0 は full で 9GB、GGUF で CPU 4GB、Int4/AWQ で GPU 5GB とされる。会社サイトは、旧世代の小型テキスト MiniCPM が「量化后仅 2GB 内存」、MiniCPM-V 系が「量化後端側内存仅 6GB」と訴求している。逆に、**ディスク上のモデルファイル容量は accessible な一次資料で系統的には明示されておらず、ストレージ要件は未確認**である。[49]

対応ハードウェアは、スマホでは iPhone/iPad と Snapdragon 系、PC では Mac と Intel Core Ultra 系、ロボット/組込では Jetson AGX Orin、車載では「主流汽车芯片

全適配」までが公開範囲で確認できる。2025 年 7 月の GitHub Release には Intel integrated graphics acceleration を用いたデモがあり、Core Ultra 7 以上のモバイル CPU と 32GB 以上の RAM を推奨する。車載については、会社サイトが主流車載チップへの全面適合を謳う一方、**具体 SoC 名は未公表**である。ロボットについてもアプリケーション先は公表されているが、具体的な搭載 SoC は Jetson AGX Orin などベンチマーク対象を除き未確認である。[50]

オンデバイスのプライバシー・セキュリティ設計

プライバシー/セキュリティ設計は、現時点では**「クラウドへ出さない」ことを主軸にしたアーキテクチャ設計**が公開情報の中心である。Nature 論文は edge deployment の意義を mobile、offline、energy-sensitive、privacy-critical scenarios と明示し、GitHub cookbook も end devices を offline and privacy-sensitive applications に適すると説明する。会社サイトの SuperMate は「弱网断网」「信息安全」を前面に掲げ、法律ソリューションでは法院専網への配備を明示している。したがって、面壁のセキュリティ設計は、現状の公開情報では **TEE、Secure Enclave、差分プライバシー、連合学習のような細粒度安全機構の開示**よりも、**ローカル実行・閉域実装・通信依存低減**に重心があるとみるべきである。これ以上の暗号設計や鍵管理の実装詳細は未公開で、未確認とするのが適切である。[51]

ソースコード、モデルカード、ライセンス

コードは主に OpenBMB[52] 名義の GitHub 上で公開され、モデルカード/重みは Hugging Face[53] 上の openbmb 組織に置かれている。MiniCPM4.1-8B、MiniCPM4-0.5B、MiniCPM-V 4.0、MiniCPM-V 4.5、MiniCPM-o 2.6 の各モデルカードはいずれも Apache-2.0 ライセンスと記載し、MiniCPM-o/V リポジトリも「weights and code are open-sourced under Apache-2.0」と明示する。すなわち、**OSS であり、原則として商用利用可能なライセンス**である。ただし同時に、生成内容に対する利用者責任・データ安全リスクに関する免責文も置かれている。[54]

実績と採用状況

オープンソースとしての traction は強い。GitHub では MiniCPM リポジトリが 8.9k stars / 571 forks、MiniCPM-o が 24.5k stars / 1.9k forks を持つ。Hugging Face では、MiniCPM4.1-8B の月間ダウンロード 49,105、MiniCPM4-0.5B 24,617、MiniCPM-V 4.0 160,526、MiniCPM-V 4.5 132,169、MiniCPM-o 2.6 227,692 が確認できる。会社発表ベースの累計値では、MiniCPM 系列全体で 2,400 万ダウンロード超とされるため、公開コミュニティでの可視的指標と会社側の総量指標は方向感として一致している。[55]

商用採用で最も前進しているのは車載である。2026 年 4 月時点の報道では、SuperMate 端側智能座舱方案の升级版が公開され、すでに吉利・長安マツダ系の量産車に実装済みとされる。別報では EZ-60 と吉利银河 M9 への搭載が明記されている。これは PoC 段階ではなく、「量産設計採用」に入っている点が重要で、端側 AI スタートアップとしてはかなり先行している。[56]

提携先の広がりも、同社の go-to-market を読むうえで重要である。2024 年末時点の報道では、面壁智能が 华为[57]、联发科技[58]、联想[59]、英特尔[60]、长城汽车[61]、易来智能[62] と協業し、AI Phone、AI PC、智能座舱、智能家居、具身机器人をカバーするとされる。2026 年 4 月時点では、会社側は「多家全球头部手机厂商」との深い協力を述べているが、具体社名はまだ公開していない。したがって、スマホ領域は**案件の存在は強く示唆されるが、顧客実名の裏取りは未完**である。[63]

B2G/B2B の縦展開では、法律領域の進展が目立つ。会社サイトは「全国首个接入法院办案流程的大模型」「法院専網完成部署」と述べ、2024 年末のメディアも深圳中院向け司法審判垂直大模型や最高法の法信法律基座大模型への関与を報じている。教育など他の垂直領域への言及もあるが、公開情報で実装深度が確認できるのは、現時点では車載と司法が先行している。[64]

収益化モデルは会社が定量開示していないため、ここは**根拠付きの推定**として整理するのが妥当である。公開情報からは、収益源は大きく四つに分かれる可能性

が高い。第一は OEM/ODM 向けの端側モデルライセンス・統合費・保守費。第二は、法院・教育・企業向けの私有配備/専網配備。第三は、クラウド/端雲協同を伴うソリューション提供。第四は、Lantay のような agentic software の SaaS/seat 課金である。とくに Lantay は 2026 年 4 月に公測入りと報じられており、従来の受託型だけでなく、プロダクト型売上を作る試みと読める。もっとも、**ARR、粗利率、継続率は未開示**であり、投資判断上は最大の情報ギャップである。[65]

競合比較

面壁智能の競争優位は、「中国語性能」「端側実装」「高密度マルチモーダル」の三点に集約できる。テキスト軽量領域では、0.5B〜1B 級の公開比較で MiniCPM4-0.5B が Qwen3 0.6B、Llama 3.2 1B、Gemma 3 1B を MMLU/CMMLU/CEval 平均で明確に上回る。マルチモーダル領域では、4.1B の MiniCPM-V 4.0 が OpenCompass/OCRBench で 3B〜8B 級の有力 OSS に競るか上回り、しかも iPhone 16 Pro Max でローカル実行可能という「性能×実装」のバランスを示す。さらに 4.5/o 4.5 では、同社は明確に GPT-4o/Gemini クラスとの競合を標榜している。[66]

軽量テキストモデル比較

モデル	パラメータ	MMLU	CMMLU	CEval	OSS	オンデバイス適性
MiniCPM4-0.5B	0.5B	55.55	65.22	66.11	Yes	非常に高い
MiniCPM4-0.5B Int4	0.5B	55.46	63.91	64.85	Yes	極めて高い
Qwen3-0.6B	0.6B	42.95	42.05	45.53	Yes	高い
Llama 3.2 1B	1B	46.89	23.73	36.74	Yes	高い
Gemma 3 1B	1B	41.64	25.09	31.83	Yes	高い

この表は MiniCPM4-0.5B-QAT-Int4 の公式モデルカード掲載比較をそのまま整理したものである。同一表では平均スコアでも MiniCPM4-0.5B が 58.00、Int4 版でも 55.68 を維持しており、「量子化で死ぬほど性能を落とさず端側適性を確保する」

という面壁の狙いがはっきり見える。旧世代の MiniCPM-2B についても、公式 GitHub は Mistral-7B に近い性能、Llama2-13B や MPT-30B を上回るという位置づけを付しており、同社が一貫して「少パラで上を食う」構図を狙ってきたことが分かる。[67]

端側マルチモーダル比較

モデル	パラメータ	OpenCompass	OCRBench	端側実装・効率の公表値	OSS
MiniCPM-V 4.0	4.1B	69.0	894	iPhone 16 Pro Max で FTL<2 秒、17 tok/s 超。GGUF CPU 4GB、Int4/AWQ GPU 5GB、full 9GB	Yes
MiniCPM-o 2.6	8.7B〜9B 表記	70.2	889	1.8MP 画像を 640 visual tokens で処理。iPad 級端末でライブストリーミング対応	Yes
MiniCPM-V 4.5	8B	77.0	公式スクレイプ範囲で数値未確認	6 フレーム→64 tokens、96×動画圧縮。量子化/ollama/llama.cpp/iOS app 対応	Yes
Qwen2.5-VL-3B-Instruct	3.8B	64.5	828	端側適性はあるが MiniCPM-V 4.0 比で OpenCompass・OCRBench 劣後	Yes
Qwen2.5-VL-7B-Instruct	8.3B	70.9	888	性能は強いが、端側速度・メモリ要件の公開粒度は MiniCPM より弱い	Yes

モデル	パラメータ	OpenCompass	OCRBench	端側実装・効率の公表値	OSS
InternVL2.5-4B	3.7B	65.1	820	端側適性は限定的	Yes
GPT-4.1-mini	非公開	68.9	840	基本はクラウド前提、ローカル配備不可	No
Claude 3.5 Sonnet	非公開	70.6	798	クラウド前提、ローカル配備不可	No

この表の OpenCompass/OCRBench は公式モデルカードの観測値から作成している。ここで見えるのは、MiniCPM-V 4.0 が 4.1B という小サイズで GPT-4.1-mini を OpenCompass でほぼ同等、OCRBench では上回り、Qwen2.5-VL-3B も明確に超えること、そして MiniCPM-V 4.5 が 8B でさらに 77.0 へ伸びていることである。つまり面壁の最も強い土俵は、**モバイル/ローカル前提の高密度マルチモーダル**であり、絶対的なクラウド最先端性能ではなく、「動く場所まで含めた性能」で勝ちにいっている。[68]

競争上の優位性は三つある。第一に、**中国語ベンチでの密度優位**であり、0.5B 級でも CMMLU/CEval がかなり強い。第二に、**端側実装の現実性**で、iPhone、iPad、Jetson、Mac、Intel Core Ultra、llama.cpp、ollama まで一気通貫でケアしている。第三に、**単一カテゴリで終わらない製品構成**で、テキスト・視覚・音声・全モーダル・コックピット・法律 AI が「理論—モデル—ツール—アプリ」の順に並んでいる。[69]

一方の弱点も明確である。第一に、4.1-8B の公開スクレイプ範囲では MMLU/CEval の完全表が抜けており、latest text flagship の客観比較は一部が画像化されていて検証負荷が残る。第二に、closed frontier model と比較したとき、hardest reasoning の絶対性能で常時優位とまでは言い切れない。第三に、商用面では OEM/OEM-like 案件は認証・量産サイクルが長く、売上計上の読みが難しい。

したがって、面壁は「最先端そのもの」というより、「最も実装しやすい高性能モデル群」で評価する方が正確である。[70]

リスク、規制、市場機会

中国本土での規制環境は、面壁智能のような会社にとって機会と制約の両方を意味する。国家网信办等の《生成式人工智能服务管理暂行办法》は、2023 年 8 月 15 日施行で、国内公衆向けの生成 AI サービスに適用される。一方、企業・研究機関などが公衆向け提供を伴わずに研究開発・応用する場合は適用対象外と明記されている。これは面壁のような **B2B/B2G・私有配備型モデル**にとっては重要で、法院専網や OEM 組込のような閉域案件は、一般公開サービスより柔軟に進めやすい可能性がある。[71]

他方で、テキスト・音声・画像・動画の合成を伴うサービスには《互联网信息服务深度合成管理规定》がかかり、标识付与、ログ保存、安全評価、備案などが求められる。特に智能对话、智能写作、音声生成、人脸生成等には顕著表示が必要で、世論属性や社会動員能力を持つサービスは備案が必要になる。2026 年 2 月末時点で、国家网信办ベースの累計備案は 796 款、登記は 481 款に達しており、市場全体が「実験期」から「制度化期」に入っていることを示す。面壁にとっては、司法・車載・企業内利用には追い風だが、一般向け agent や consumer assistant を大規模に伸ばす局面では、備案・标识・安全评估コストが無視できない。[72]

外部環境では、米国の半導体輸出規制も構造リスクである。BIS は 2022 年のルールを起点に 2023 年 10 月、2024 年 4 月、2024 年 12 月と規制をアップデートし、中国の高性能半導体調達・製造能力を継続的に制約している。面壁の端側・高効率戦略は、この制約に対する「防御」として理にかなっている。巨大 GPU クラスター依存を相対的に下げ、スマホ・PC・車載・ロボット側の分散算力で戦えるからである。ただし、**最適化対象のチップが細分化するほど実装負荷も増える**ため、ハードウェア断片化は逆にオペレーションリスクにもなる。[73]

市場機会は大きい。Nature 論文が述べる通り、edge scenario は mobile phones、personal computers、vehicles、robotics に広がる。面壁は公開材料上すでに AI

Phone、AI PC、智能座舱、智能家居、具身机器人、法律 AI にまたがる配置を進めている。これは、もし一つの垂直で勝てなくても、同じモデル資産を複数終端に再利用できることを意味する。特に中国の自動車・司法・端末ローカル AI は「弱ネット環境」「データ域内化」「低レイテンシ」という需要が強く、ここは面壁に最も合う市場である。[74]

商用化リスクは、規制よりもむしろ「見えにくい採算性」にある。OEM 案件は売上化まで時間がかかり、モデル企業は研究開発費率が高い。面壁は資金調達こそ順調だが、公開資料から売上・粗利・キャッシュバーンを定量把握できない。このため、投資家としては「ダウンロードが増えている」ことより、「一社あたりの契約単価」「継続保守収入」「私有配備案件の更新率」「量産出荷台数」が見えないことの方を重くみるべきである。これは公開情報不足ゆえの判断であり、現時点では未確認事項が多い。[75]

将来展望と推奨

技術ロードマップは、公開済みリリースからかなり推定できる。短期的には、MiniCPM 4.1 の sparse reasoning、MiniCPM-SALA の long-context、MiniCPM-V 4.5 の高密度動画理解、MiniCPM-o 4.5 の全二重全モーダルを、車載/スマホ/AIPC/ロボット系の実装へ落とし込むフェーズに入る可能性が高い。特に ArkInfer、CPM.cu、FlagOS、llama.cpp/ollama など推論層の整備は、面壁が「モデルを出す会社」から「端側 AI のシステムベンダー」へ寄せているサインである。これは単なる研究開源より monetization に近い。もっと率直に言えば、今後の価値はモデル精度そのものより、**hardware-software co-design の実装速度**で決まる。[76]

投資家視点では、面壁智能は「中国大模型企業の中で、最も端側に賭け切っている一社」と位置付けられる。プラス材料は、国資と産業資本の連続参入、独角獣閾値到達、強い開源 traction、司法・車載の高障壁案件、そして Apache-2.0 による生態系自走力である。マイナス材料は、収益指標の未開示、Qwen 等の強大な国内競合、OEM 依存による売上時期の読みにくさ、そしてグローバル展開時の規制・輸出管理・ブランド力格差である。結論としては、**「研究力」は上位、「収益の見え方」はまだ中位**という評価になる。[77]

今後 6〜18 か月で追うべき KPI は、次のように整理できる。

KPI	現在見えているシグナル	監視理由
コミュニティ需要	系列累計 2,400 万 DL 超、主要 HF モデルで月数万〜十数万 DL	先行需要の維持確認
量産採用台数	EZ-60 / 吉利銀河 M9 など量産搭載	実売上化の近さを測る
端末設計採用数	頭部スマホメーカーとの深い協力を会社が示唆	端側 AI 本命市場の広がり
端側効率	Jetson / iPhone / Mac / Intel での推論改善が継続	技術優位の持続性
規制処理能力	中国備案・标识環境への適応	consumer 展開の天井を左右
有償案件の質	法院・車載・企業内案件の継続拡大	ARR/粗利の代理指標
チップ生態	Snapdragon / Apple / Intel / Jetson / 車載 SoC での最適化継続	ハード断片化への対応力

この KPI 表のうち、現時点で公開確認できるのはダウンロード、量産車搭載、主要ハード対応、規制制度の外形までである。売上高、ARR、グロスマージン、純増顧客数、チャーン率は公開情報では未確認であり、投資判断では最優先の追加 DD 項目になる。面壁智能への推奨スタンスは、**成長投資対象としては前向きだが、再投資判断は量産台数と有償継続率の開示待ち**、というのがもっとも厳密である。[78]

主要出典の URL 一覧

```
https://modelbest.cn/news
https://pitchhub.36kr.com/project/2214864842159235
https://www.cs.tsinghua.edu.cn/info/1137/6540.htm
https://ml-summit.org/speaker/846?lang=cn&uid=c1041
https://github.com/OpenBMB/MiniCPM
https://github.com/OpenBMB/MiniCPM-o
```

https://huggingface.co/openbmb/MiniCPM4.1-8B
https://huggingface.co/openbmb/MiniCPM4-0.5B
https://huggingface.co/openbmb/MiniCPM-V-4
https://huggingface.co/openbmb/MiniCPM-V-4_5
https://huggingface.co/openbmb/MiniCPM-o-2_6
https://www.nature.com/articles/s41467-025-61040-5
https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm
https://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm
https://www.bis.gov/press-release/commerce-strengthens-export-controls-restrict-chinas-capability-produce-advanced-semiconductors-military
https://finance.sina.com.cn/jjxw/2024-04-11/doc-inarmxxu9074880.shtml
https://finance.sina.com.cn/tech/2025-05-21/doc-inexhsqi6253851.shtml
https://jjckb.xinhuanet.com/20260228/37ff64dc0c134f858f2283002b8e43cc/c.html
https://finance.sina.com.cn/roll/2026-04-07/doc-inhtrumi7298241.shtml

---

[1] [42] [49] [50] [57] https://github.com/OpenBMB/MiniCPM-o

https://github.com/OpenBMB/MiniCPM-o

[2] [46] https://huggingface.co/openbmb/MiniCPM-o-2_6

https://huggingface.co/openbmb/MiniCPM-o-2_6

[3] [27] [29] [66] [67] [69] https://huggingface.co/openbmb/MiniCPM4-0.5B-QAT-Int4-unquantized

https://huggingface.co/openbmb/MiniCPM4-0.5B-QAT-Int4-unquantized

[4] [8] [13] [51] [58] [74] https://www.nature.com/articles/s41467-025-61040-5

https://www.nature.com/articles/s41467-025-61040-5

[5] [11] [18] [24] [37] [56] [61] https://pitchhub.36kr.com/project/2214864842159235

https://pitchhub.36kr.com/project/2214864842159235

[6] [7] [9] [17] [19] [25] [35] [43] [44] [47] [48] [52] [53] [55] [76] https://github.com/openbmb/minicpm

https://github.com/openbmb/minicpm

[10] [14] [32] [71] https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

[12] https://www.cs.tsinghua.edu.cn/info/1137/6540.htm

https://www.cs.tsinghua.edu.cn/info/1137/6540.htm

[15] [21] [38] [72] https://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm

https://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm

[16] [59] [63] https://www.sohu.com/a/835264758_430392

https://www.sohu.com/a/835264758_430392

[20] [28] [30] [45] [54] [60] https://huggingface.co/openbmb/MiniCPM-V-4_5

https://huggingface.co/openbmb/MiniCPM-V-4_5

[22] [26] [33] [39] https://finance.sina.com.cn/tech/2025-07-01/doc-infcxsmq6835935.shtml

https://finance.sina.com.cn/tech/2025-07-01/doc-infcxsmq6835935.shtml

[23] [41] [64] https://www.modelbest.cn/news

https://www.modelbest.cn/news

[31] [68] https://huggingface.co/openbmb/MiniCPM-V-4

https://huggingface.co/openbmb/MiniCPM-V-4

[34] https://finance.sina.cn/tech/2023-04-14/detail-imyqizvp0360960.d.html?vt=4

https://finance.sina.cn/tech/2023-04-14/detail-imyqizvp0360960.d.html?vt=4

[36] [75] [77] https://finance.sina.com.cn/roll/2026-04-07/doc-inhtrumi7298241.shtml

https://finance.sina.com.cn/roll/2026-04-07/doc-inhtrumi7298241.shtml

[40] [62] https://m.36kr.com/p/2728647825646857

https://m.36kr.com/p/2728647825646857

[65] https://www.jiemian.com/lists/48_46.html

https://www.jiemian.com/lists/48_46.html

[70] https://huggingface.co/openbmb/MiniCPM4.1-8B

https://huggingface.co/openbmb/MiniCPM4.1-8B

[73] https://www.bis.gov/press-release/commerce-strengthens-export-controls-restrict-chinas-capability-produce-advanced-semiconductors-military

https://www.bis.gov/press-release/commerce-strengthens-export-controls-restrict-chinas-capability-produce-advanced-semiconductors-military

[78]
https://jjckb.xinhuanet.com/20260228/37ff64dc0c134f858f2283002b8e43cc/c.html

https://jjckb.xinhuanet.com/20260228/37ff64dc0c134f858f2283002b8e43cc/c.html