archi-intelligence Research Series

Working Paper 2026-03

Multi-Embodiment Physical AI Platform Readiness

Tesla FSD-Optimus Unified Stack

archi-intelligence Research Team

Released June 2026

archi-intelligence.org

Front Matter

F.3 Copyright and License

Copyright Notice

DOI

DOI: 10.5281/zenodo.20513903 Permanent Archive: https://doi.org/10.5281/zenodo.20513903

Suggested Citation

archi-intelligence Research Team. (2026). *Multi-Embodiment Physical AI Platform Readiness: The Tesla FSD-Optimus Unified Stack* (Working Paper 2026-03). archi-intelligence Research Series. https://doi.org/10.5281/ze

BibTeX

```
@techreport{archi_intelligence_2026_03,
  title       = {Multi-Embodiment Physical AI Platform Readiness:
                 The Tesla FSD-Optimus Unified Stack},
  author      = {{archi-intelligence Research Team}},
  institution = {archi-intelligence},
  year        = {2026},
  number      = {2026-03},
  type        = {Working Paper},
  series      = {archi-intelligence Research Series},
  url         = {https://archi-intelligence.org/research/2026-03},
  doi         = {10.5281/zenodo.20513903}
}
```

Contact

- Research Team: research@archi-intelligence.org
- Press Inquiries: press@archi-intelligence.org
- Corrections: corrections@archi-intelligence.org
- Website: https://archi-intelligence.org

F.4 Conflict-of-Interest (COI) Disclosure

Statement of Conflicts of Interest and Editorial Independence

To maintain the academic independence and credibility of this research series, the archi-intelligence Research Team hereby clarifies the following matters.

1. On the research institution

archi-intelligence is an independent academic research institution dedicated to establishing Architecture Intelligence (AI^2) as a research paradigm. The mission of this institution is to advance the standardization and comparability of cross-industry architecture-engineering practice through open methodology, transparent data attribution, and rigorous peer review.

2. On funding and sponsorship

At this release stage, this research series has not accepted any direct or indirect financial support from any assessed object (including but not limited to the OEMs, Tier-1 suppliers, platform companies, and chip manufacturers mentioned in this report).

In its initial phase, archi-intelligence received seed sponsorship from Arkimind (a commercial entity focused on the intelligentization of automotive electrical/electronic architecture). This relationship is explicitly disclosed in this statement, and editorial independence is safeguarded through the following governance-isolation mechanisms:

- The editorial/research team does not report to Arkimind's sales or business teams.
- The compensation of researchers is not tied to any commercial metric of Arkimind.
- Arkimind's business team has no right of review or modification before a report is released.
- The assessment methodology and the original data sources are 100% public.
- Any assessed object may submit feedback for correction, with a public handling process.

3. On editorial independence

The methodological choices, case assessments, and conclusion formulations of this research are made independently by the archi-intelligence Research Team. The research judgments are not influenced by any commercial interest, political stance, or geopolitical preference. We acknowledge that this editorial independence must ultimately be validated by readers through their continued scrutiny of the quality of our research. We welcome critical feedback from academia, industry, and regulators, and we undertake to record openly all significant corrections in subsequent versions.

4. On data sources and the time window

This research uses public-source information from January 2024 to May 2026, including but not limited to: the earnings reports and SEC/HKEX public filings of listed companies (including Tier 1 material such as the Tesla Q1 2026 Update, the SpaceX S-1 prospectus 2026.5.20, the XPeng SEC Form 6-K, and the Xiaomi HKEX annual report); the official technical releases of OEMs and suppliers (AI Day, HDC, Investor Day, CES, the Tesla Annual Shareholder Meeting 2025.11.6, and the like); published patent literature; peer-reviewed academic papers and conference publications; the releases of industry-standards organizations (AUTOSAR, ISO, SAE, IEEE, and the like); and mainstream technical-media reporting and industry analysis. For rumored information or anonymous-source information that cannot be independently verified, this research takes a conservative stance — either not citing it, or explicitly annotating it as "directional inference based on public reporting."

For Musk's public statements and social-media remarks, this report distinguishes two categories of citation: (a) those that are formal records of official documents or earnings calls are treated as Tier 2; (b) those that are informal social-media remarks are treated as Tier 4 and explicitly annotated as "informal source." Many of Musk's performance claims (such as "AI5 is 40× faster than AI4") are composite, marketing-oriented framings; this report annotates the nature of their source and does not adopt them as objective comparison baselines.

5. On research limitations

As an in-depth study of a single case, this work openly acknowledges: (a) the Tesla case is in a phase of rapid evolution, the data of this report are current as of 21 May 2026, and the facts described may be refreshed by new events within months; (b) Tesla has provided relatively rich technical disclosure through AI Day, the quarterly earnings reports, SEC filings, and the SpaceX S-1 prospectus, but a great many low-level implementation details (such as the precise architecture of the FSD end-to-end network, the

specific topology of Cortex 2.0, and the microarchitecture of AI5) remain undisclosed; (c) a single-case study does not represent a complete coverage of the AR4 path — other AR4 candidate cases such as Huawei HIMA and XPeng require independent study. The analysis of the five mirror OEMs in Chapter 9 of this report is a partial compensation for the limitation of the single case, but cannot replace an independent in-depth study of each case.

6. On trademarks and product names

All trademarks, product names, and product code names mentioned in this report (including but not limited to Tesla, Optimus, FSD, SpaceX, Anthropic, Volkswagen, Toyota, Huawei, XPeng, Xiaomi, and the like) belong to their respective rights holders. The citation of such names is purely for academic discussion and industry analysis, and does not constitute any commercial association, endorsement, or promotion.

7. On the AI-assistance disclosure of this report

Part of the draft generation and text polishing of this report used the Claude model of Anthropic as an auxiliary tool. **The final methodological choices, fact verification, derivation of judgments, and conclusion formulations were all made independently by the archi-intelligence Research Team, and rigorously traced through a tiered evidence chain of Tier 1-4.**

A relevant commercial relationship requiring transparent disclosure: according to the S-1 prospectus filed by SpaceX with the SEC on 20 May 2026 (Tier 1 material in this report), a cloud-services commercial relationship exists between Anthropic and SpaceX (Anthropic pays SpaceX $1.25B/month through May 2029, using part of the capacity of COLOSSUS / COLOSSUS II). SpaceX and Tesla reached a framework agreement on the Terafab project in March 2026 (likewise a Tier 1 public disclosure). This constitutes an indirect commercial chain of "Anthropic ← SpaceX ← Musk → Tesla."

Since the body of this report concerns a Tesla case study, archi-intelligence hereby clarifies: (a) archi-intelligence accepts no direct or indirect financial support from any of Anthropic, SpaceX, Tesla, or xAI; (b) the fact of using Claude as a draft tool does not affect the methodological independence and editorial judgment of this report; (c) this report's critical examination of Tesla (including the HW3 withdrawal event, the AI5 "not to vehicles" resource allocation, the Robotaxi deployment gap, the CEO personal risk, and the persistence of scaling laws) has fully embodied editorial independence.

8. Disclaimer

This report does not constitute investment advice, business consulting, or legal opinion of any form. Any business decision made on the basis of the contents of this report is at the risk and responsibility of the deciding party.

- Research contact: research@archi-intelligence.org
- Corrections and feedback: corrections@archi-intelligence.org
- Full governance charter: https://archi-intelligence.org/governance

Contents

F.6 Index of Figures and Tables

Figures

Tables

F.7 List of Abbreviations

Architecture levels and frameworks (consistent with the D1/D2 series)

- **AR0–AR5** — Architecture Readiness 0–5, the architecture-maturity ladder (advanced in the first flagship report)
- **AR4** — Multi-Embodiment Physical AI Platform (the core object of study of this report)
- **AI²** — Architecture Intelligence (the research paradigm of this research series)
- **AI²-ML** — Architecture Intelligence Maturity Levels (L0-L5)

Tesla silicon / software generations

- **HW3** — FSD Computer (volume production 2019, 14nm Samsung, ~144 TOPS)
- **HW4 / AI4** — the current volume-production mainstay (from 2023, Samsung 7nm, ~500 TOPS)
- **AI4.5** — stopgap chip (owing to the AI5 delay, fitted to the 2026 Model Y from late 2025)
- **HW5 / AI5** — taped out 2026.4, ~4000 TOPS equivalent (5× AI4 useful compute)
- **AI6 / AI7** — future-generation road-markers
- **FSD** — Full Self-Driving (Tesla's autonomous-driving software package)
- **V11 / V12 / V13 / V14** — FSD software-stack generations
- **VLA** — Vision-Language-Action (model)

Tesla group-level compute infrastructure

- **Cortex 2.0** — Tesla training cluster (Giga Texas, ~100K H100/H200, 250-500MW)
- **COLOSSUS / COLOSSUS II** — xAI / SpaceX training clusters (Memphis TN + Southaven MS, ~1 GW)
- **Dojo** — Tesla in-house training chip (D1 / Dojo 3)
- **Terafab** — SpaceX + Tesla + Intel joint chip-manufacturing project ($25B Austin, long-term target 1 TW/year)
- **xAI Merger** — SpaceX's acquisition of xAI (2026.2.2, combined valuation $1.25T)

Tesla / Optimus physical hardware

- **Cybercab** — Tesla's dedicated Robotaxi model (Q2 2026 SOP, initially using AI4/AI4.5)
- **Optimus** — Tesla humanoid robot
- **Optimus Gen 3** — the hand-upgraded version (22 DoF/hand + 50 actuators, tendon-driven)
- **4680** — Tesla's in-house lithium-battery cell specification
- **Gigacasting** — Tesla's giant-casting manufacturing process

FSD software-stack key concepts

- **End-to-End** — the end-to-end neural network (from V12, "Photon In, Control Out")
- **Occupancy Network** — the occupancy network (3D voxel representation, morphology-agnostic perception)
- **Software 2.0** — the data-driven rather than hand-written-rule software paradigm
- **Robotaxi** — Tesla's unsupervised FSD commercial-operation service
- **Intervention-free Streak Counter** — the "intervention-free continuous-mileage" counter introduced in V14.3

Safety and regulation

- **ASIL** — Automotive Safety Integrity Level (ISO 26262)
- **SOTIF** — Safety of the Intended Functionality (ISO 21448)
- **UN-R155** — the cybersecurity management system (CSMS) UN regulation
- **UN-R156** — the software-update management system (SUMS) UN regulation

- **UN-R157** — the Automated Lane Keeping System (ALKS) UN regulation
- **L2 / L3 / L4 / L5** — SAE J3016 autonomous-driving levels
- **fail-soft** — soft failure (the human serves as safety redundancy after a system failure)
- **fail-operational** — fault-tolerant operation (after the failure of any subsystem, the system itself can complete the task in a degraded-safe manner)
- **NHTSA** — National Highway Traffic Safety Administration (United States)
- **RDW** — Rijksdienst voor het Wegverkeer (the Netherlands Vehicle Authority)

SEC filings and official material

- **10-Q** — the quarterly earnings report of a U.S. listed company (SEC Form)
- **8-K** — the material-event announcement of a U.S. listed company (SEC Form)
- **S-1** — the U.S. IPO prospectus (SEC Form)
- **6-K** — the periodic or interim report of a foreign issuer (SEC Form)
- **HKEX** — Hong Kong Stock Exchange

Mirror-OEM abbreviations

- **CARIAD** — the Volkswagen Group software subsidiary
- **CARIZON** — the Volkswagen-Horizon joint venture (60:40)
- **CEA** — China Electronic Architecture (Volkswagen)
- **SSP** — Scalable Systems Platform (Volkswagen's global scalable systems platform)
- **VCTC** — Volkswagen Group China Technology Company
- **Arene** — the Toyota / Woven by Toyota software platform
- **TPS** — Toyota Production System
- **HIMA** — Huawei Intelligent Mobility Alliance
- **ADS** — Autonomous Driving Solution (Huawei)
- **MDC** — Mobile Data Center (Huawei's vehicle computing platform)
- **XNGP** — XPeng Navigation Guided Pilot
- **VLA 2.0** — XPeng's second-generation Vision-Language-Action
- **Turing chip** — XPeng's in-house AD/ADAS inference chip
- **HAD** — Xiaomi Autonomous Driving

Assessment methodology

- **Tier 1–5** — the data-source-authority pyramid levels
  - Tier 1 = SEC filings / IPO prospectuses / company earnings reports / shareholder-meeting disclosures
  - Tier 2 = earnings calls / official technical-release events
  - Tier 3 = industry analysts / peer-reviewed papers
  - Tier 4 = mainstream technical-media reporting
  - Tier 5 = informal social-media remarks (such as Musk's X posts)
- **Falsifiable point** — a concrete scenario that can be empirically confirmed or disproven by future engineering reality; five are listed in Section 10.6 of this report

Statement of Data-Verification Discipline: this report concerns Tesla silicon specifications, FSD versions, Optimus specifications, and training infrastructure; all key technical parameters have been verified against public sources of May 2026. Musk-framing data (such as "40x faster") are annotated as to the nature of their framing; data with differing sources (such as Optimus's whole-body degrees of freedom and the Cortex cluster scale) are annotated with a range of difference rather than forcibly unified. The Tier 1 sources of this report include the Tesla Q1 2026 Update (assets-ir.tesla.com), the Tesla 10-Q / 8-K (SEC.gov), the SpaceX S-1 prospectus (2026.5.20), the Tesla Annual Shareholder Meeting (2025.11.6), the XPeng SEC Form 6-K, the Xiaomi HKEX annual report, and the VW Group China Investor Update.

Abstract

As the third and concluding Working Paper of the archi-intelligence Research Series, this study takes the Tesla FSD-Optimus unified stack as the empirical anchor that closes the loop between the AR0–AR5 architecture-capability-threshold framework advanced in D1 (*The Architectural Migration of the Century*, 2026-01) and the horizontal maturity assessment of twenty-two OEMs conducted in D2 (*The State of Global Automotive E/E Architecture Maturity 2026*, 2026-02). As of May 2026, Tesla is the only engineering entity worldwide that simultaneously satisfies every criterion of AR4 (a multi-embodiment physical-AI platform) and has entered volume-production ramp; this study selects it not as an endorsement of its commercial prospects but on a single methodological judgment — it is at present the only case that has fully connected its officially disclosed eleven layers of shared core technology across the vehicle and humanoid-robot morphologies. Through a clinical dissection at the granularity of silicon (HW3 → AI7, including the AI4.5 stopgap and the AI5 "not to vehicles" hardware tiering), the FSD software stack (the Software 2.0 paradigm revolution from V11 rules to V14 end-to-end-plus-VLA), the morphology-agnostic occupancy network, and the training infrastructure (Dojo → Cortex 2.0, and the group-level compute restructuring after the SpaceX-xAI merger), the study distills Tesla's eleven officially disclosed shared layers into three structurally distinct reuse mechanisms — the physical layer (6 items), the intelligence layer (5 items), and an implicit organizational twelfth layer (the FSD-Optimus team merger) — and characterizes precisely the boundary of cross-morphology reuse: accumulated capability is reusable, while morphology-dependent high-frequency motion control is not. The central thesis is that Tesla functions as an **AR4 reference frame** rather than a template for replication: no fast-follower can reproduce the twenty years of accumulation, full-stack verticality, and group-level resource allocation the path demands, but every fast-follower can use the reference frame to locate its true position and reasonable path — a judgment the study substantiates through an in-depth twelve-layer mirroring of five distinct trajectories (Volkswagen, Toyota, Huawei, XPeng, Xiaomi). Empirically, the Tesla case supports D1's two-layer structural thesis, its proposition that "AI is a faculty of language, not a brain" (doubly corroborated by Tesla's own deployment cadence and the scaling-law caveat in the SpaceX S-1), its theory of architecture debt, and its failure-philosophy framework, while the study also lists five falsifiable points by which the D1 framework may be tested against the engineering reality of 2027–2028. The study closes by identifying a structural market gap: while the industrial toolchain already offers mature support for the detailed-design phase, the conceptual-exploration phase — where critical architectural decisions are made before the first diagram or line of code — still relies on individual and organizational engineering intuition that the fast-follower has not had twenty years to accumulate.

Keywords: AR4; multi-embodiment physical-AI platform; cross-morphology reuse; vertical closed-loop; Tesla FSD; Optimus; occupancy network; Software 2.0; architecture readiness; failure philosophy; Architecture Intelligence

Introduction: Why Tesla Is the Only Complete Case of an AR4 Vertical Closed-Loop

The Position of This Study Within the archi-intelligence Research Series

This report is the third Working Paper (2026-03) in the archi-intelligence Research Series. To understand the value of D3, one must understand its threefold position within the series.

First — D3 is the empirical anchor for the theoretical framework of D1. D1 (*The Century-Scale Migration of Architecture*, 2026-01) advanced a set of core propositions: the AR0–AR5 architecture-capability-threshold framework, the AI²-ML architecture-intelligence maturity ladder, the two-layer structural thesis (infrastructure convergence versus control-semantics divergence), and the philosophy of failure (fail-soft versus fail-operational). As a theoretical framework, these propositions require engineering-grounded empirical testing. **D3 selects Tesla as the case of the "only complete AR4 closed-loop" and subjects D1's theoretical claims to clinical examination at a concrete engineering granularity** — establishing which are supported, which are challenged, and where the falsifiable points of the D1 framework lie. Chapter 10 is the concentrated presentation of this examination.

Second — D3 is the in-depth extension of the panoramic assessment in D2. D2 (*The State of Global Automotive E/E Architecture Maturity 2026*, 2026-02) conducted a horizontal assessment of twenty-two OEMs, assigning scores across the dual temporal dimensions of Snapshot and Roadmap. Yet the limitation of a horizontal assessment is that **while it can locate each OEM's position, it struggles to reveal "why the position is where it is, and what the path to an upward leap would be."** By taking a deep cross-section of one of the only two AR4 protagonists in D2 (Tesla; the other being Huawei HIMA), D3 **furnishes the remaining twenty-one OEMs of the D2 assessment with an AR4 reference frame.** Chapter 9 is the concrete elaboration of this mirroring value — under the reference frame of Tesla's engineering granularity, the respective positions, constraints, and reasonable paths of five distinct "AR4-aspiring trajectories" (Volkswagen, Toyota, Huawei, XPeng, Xiaomi) become clearly visible.

Third — D3 reveals the market gap in the leap from AR3+ to AR4. When D1 supplies the theory, D2 supplies the assessment, and D3 supplies the engineering granularity, the three together answer one core question: **when the remaining twenty-one OEMs cannot possibly replicate Tesla, what kind of methodology and tooling does the reasonable path of their upward leap require?** Section 10.5 of Chapter 10, through a layer-by-layer derivation of the fast-followers' core contradiction, points out that the current industrial tooling ecosystem already offers mature support for the **detailed-design phase** (EDA, PLM, safety-case toolchains, and the like), yet for the **conceptual-exploration phase** — the phase in which an architect must make critical architectural decisions before drawing the first diagram or writing the first line of code — it still relies principally on individual and organizational engineering intuition. Tesla internalized this capability through twenty years of accumulation; the fast-followers do not have twenty years. This constitutes a market gap worthy of joint discussion by industry, academia, and tool vendors.

The Methodological Rationale for Selecting Tesla

In the D1 flagship report, we advanced the **AR0–AR5 architecture-capability-threshold framework** and defined **AR4 (multi-body physical-AI platform)** as the stage at which training, simulation, and deployment form a unified closed-loop; the world model and the foundation model are shared across

vehicles and robots; model-reuse rates are high; and policy-transfer efficiency is high (cross-morphology generalization).

As of early 2026, worldwide, **the only case that simultaneously satisfies all of the AR4 criteria and has entered volume-production ramp is Tesla's FSD-Optimus unified stack.** This is not an endorsement of Tesla's commercial prospects or a dismissal of the controversies surrounding its safety; this report maintains a critical examination of the regulatory risk of its pure-vision approach, the fragility of its liability attribution, and the personal risk attached to its CEO. We select Tesla as the object of D3's in-depth dissection purely on the basis of one methodological judgment: **Tesla is at present the only engineering entity that has fully connected the eleven layers of shared core technology it has officially disclosed (actuators, power electronics, batteries, manufacturing, data communications, audio, cameras, AI chips, training clusters, neural simulation, real-world AI) across the vehicle and robot morphologies, and that has made this cross-morphology reuse genuinely hold through organizational-layer integration.**

The value of understanding the Tesla case lies not in "learning how to become Tesla" — as the flagship report argues, the scale of capital, density of talent, and organizational flatness that a vertical closed-loop demands are the product of twenty years of accumulation and cannot be replicated within eighteen months. Its value lies in this: **Tesla provides a "reference frame" for AR4 architecture,** allowing us to characterize precisely "what cross-morphology reuse actually means in engineering terms, what is reusable, what is not, and at what cost."

The Scope and Boundaries of This Report

This report explicitly acknowledges three limitations.

First, **the limitation of disclosure.** Tesla has provided relatively rich technical disclosure through AI Day (2021, 2022), its quarterly earnings reports, Musk's statements on the X platform, the most recent SEC 10-Q/8-K filings, and the SpaceX S-1 prospectus (2026.5.20); yet a great many low-level implementation details (such as the precise architecture of the FSD end-to-end network, the specific topology of Cortex 2.0, and the microarchitecture of AI5) remain undisclosed. This report rigorously distinguishes "disclosed fact" (annotated with Tier 1–2 sources) from "directional inference based on public information."

Second, **the question of Musk's framing.** Many of Tesla's performance claims originate in Musk's public statements, whose figures (such as "AI5 is 40× faster than AI4") are often composite, marketing-oriented framings that are not directly comparable to standardized TOPS or FLOPS measures. This report annotates the nature of the source for such data and does not adopt them as objective comparison baselines.

Third, **the limitation of the time window.** The data in this report are current as of 21 May 2026. The Tesla case is in a phase of rapid evolution, and the facts described herein may be refreshed by new events within months. We undertake to record openly, in subsequent versions, the updates and corrections to key facts.

The Structure of This Report

This report comprises ten chapters, plus an introduction and a conclusion:

- **Chapter 1 — The Silicon Road**: the generational evolution from HW3 to AI7, including the AI4.5 stopgap, the "hardware tiering" of AI5 not being allocated to vehicles, and the Terafab group-level

compute strategy.

- **Chapter 2 — The FSD Software Stack**: the paradigm revolution from the V11 modular rule-stack to the V14.3 end-to-end-plus-VLA multimodal stack, including the revision of the Robotaxi seven-city commitment and the Chinese and European regulatory reality.
- **Chapter 3 — The Occupancy Network**: as a morphology-agnostic 3D representation for cross-morphology reuse.
- **Chapter 4 — The Mechanism of Cross-Morphology Reuse**: Tesla's eleven officially disclosed layers of shared core technology, plus a three-group structural distillation including the organizational twelfth layer (physical / intelligence / organizational).
- **Chapter 5 — Optimus Kinematics and the Non-Reusable Layer**: the engineering granularity of the differences.
- **Chapter 6 — Training Infrastructure**: the evolution of Dojo → Cortex → COLOSSUS/COLOSSUS II (including the attribution restructuring after the SpaceX-xAI merger).
- **Chapter 7 — Reuse at the Organizational Level**: the FSD/Optimus team merger, plus the extreme organizational debt of Musk's cross-control over eight entities.
- **Chapter 8 — The Cost and Boundaries of AR4**: the lessons for other players, including the hard data on AI capex from the SpaceX S-1.
- **Chapter 9 — The Mirror Chapter**: the argumentative significance of D3 for the other OEMs of D2 (an in-depth mirroring of Volkswagen, Toyota, Huawei, XPeng, Xiaomi).
- **Chapter 10 — The Empirical Chapter**: the ways in which D3 supports and challenges the core propositions of D1, including the reverse definition of the tool category and the falsifiable points of D1.
- **Conclusion**: the methodological significance of AR4 as a reference frame.

Each chapter explicitly annotates its position within the threefold value of the archi-intelligence Research Series (theoretical empiricism, assessment extension, and the derivation of the fast-followers' tool category), so that the reader may construct a closed-loop understanding of the D1 → D2 → D3 trilogy.

Chapter 1 The Silicon Road: The Generational Evolution from HW3 to AI7

The evolution of Tesla's autonomous-driving hardware is the physical point of departure for understanding its AR4 architecture. Unlike the traditional OEM, which "procures the SoC supplied by a Tier-1," Tesla has, since HW3, taken the path of designing its own inference chips. The far-reaching consequence of this decision is that **the chip, the perception algorithm, and the training infrastructure co-evolve under a single design intent within a single organization,** thereby eliminating the mismatch of the traditional supply chain in which "the chip vendor does not understand the algorithm, and the algorithm team accommodates the chip." Yet by May 2026 this path had begun to reveal its engineering boundaries: the delay of AI5 gave rise to the AI4.5 stopgap; the first AI5 silicon was allocated preferentially to Optimus and the supercomputing clusters rather than to vehicles; and the legacy HW3 fleet was officially confirmed to be incapable of reaching the capability threshold for unsupervised FSD. These facts are precisely the unavoidable cost of the AR4 path.

1.1 Generational Overview

Table 1.1: A Comparison of Tesla's Autonomous-Driving Silicon Generations (as of May 2026)

Generation	Product Name	Compute	Process Node	Foundry	Status	Memory
HW3	FSD Computer	~144 TOPS	14nm	Samsung	Legacy mainstay; **officially confirmed unable to run unsupervised FSD**	8GB LPDDR4
HW4	AI4	~500 TOPS	Samsung 7nm	Samsung	New-vehicle mainstay; Cybercab Q2 2026 SOP still uses this	16GB LPDDR5
—	AI4.5	A refined upgrade of AI4	Samsung 7nm	Samsung	**Quietly fitted from the 2026 Model Y onward** (a stopgap born of the AI5 delay)	—

Generation	Product Name	Compute	Process Node	Foundry	Status	Memory
HW5	AI5	~4,000 TOPS equivalent (5× AI4 useful compute / 8× raw)	Advanced node (TSMC/Samsung dual-source)	TSMC Arizona + Samsung Taylor, TX	**Taped out 2026.4.15**; engineering samples late 2026; high-volume mid-2027; **first batch to Optimus + supercomputer, not to vehicles**	192GB LPDDR5X
—	AI6	A further evolution	Same as AI5	Same as AI5	**Tape-out target 2026.12**	—
—	AI7	To be determined	TBD	TBD	Early planning	—

Verification of key facts (sources: Tesla Q1 2026 Update PDF, SEC 10-Q FY2026 Q1, Q1 2026 earnings call, SpaceX S-1 prospectus 2026.5.20):

- Tesla's official framing of AI5's improvement over AI4: **~5× useful compute, 8× raw compute, 9× on-chip memory (16GB→192GB), 5× memory bandwidth.**
- Musk's "AI5 is 40× faster than AI4" is a composite marketing framing (incorporating bandwidth, architectural efficiency, memory, and the like), not a pure compute multiple. This report annotates the nature of this framing and does not adopt it as an objective comparison baseline.
- AI5 high-volume production has been pushed back to mid-2027 — a delay of nearly two years against Musk's original commitment of June 2024 ("AI5 in vehicles 2H 2025").
- **The Cybercab (first unit off the line 2026.2.17, Q2 2026 SOP) is confirmed to use AI4, rather than the originally planned AI5.**
- **Unsupervised FSD on HW3 has been officially abandoned** (Q1 2026 earnings call, jointly confirmed by Musk and Ashok Elluswamy). Tesla offers HW3 owners two paths: a trade-in toward an AI4 vehicle at a discount, or a free computer-plus-camera retrofit. A "V14 Lite" version for HW3 is promised to be pushed before June 2026.

1.2 HW3: The Origin of the In-House Inference Chip (2019) and the Withdrawal of the Unsupervised Promise

HW3 (FSD Computer) entered volume production in 2019 and was Tesla's first fully in-house autonomous-driving inference chip. Its roughly 144 TOPS of compute was already significantly ahead of the industry

at the time, but more important was **the establishment of its architectural philosophy**: a dual-NPU redundant design (two independent neural-network accelerators cross-checking each other), optimization specifically for vision-inference workloads, and co-design with Tesla's in-house vision-algorithm stack.

Yet **the Q1 2026 earnings call of 22 April 2026 formally confirmed that HW3 cannot reach the capability threshold for unsupervised FSD.** This confirmation is itself of methodological significance — it reveals an intrinsic contradiction of the AR4 path: **when the company attempts to compress an ever-larger end-to-end neural network (from FSD V12 onward) into the compute envelope of the 2019-era HW3, the physical ceiling of the hardware generation becomes a real constraint on the evolution of the software.** Since 2019 Tesla had sold HW3 users (with the FSD package priced at $8,000–$15,000) the promise that "the hardware is sufficient to support full self-driving"; the withdrawal of this promise in 2026 is a characteristic manifestation of the tension, within a vertically integrated closed-loop enterprise, between "the commitment to the installed base" and "the physical limit of the hardware."

The compensation mechanism itself reveals the cost of vertical integration: Tesla must bear the retrofit cost of the HW3 fleet itself (roughly a million vehicles, self-funded hardware plus camera retrofit plus a city-scale microfactory network) — a cost that, under the traditional "hardware-procurement" model, could be partially passed on to the Tier-1 supplier, but that, under Tesla's vertical closed-loop, is borne in full by the enterprise itself.

1.3 HW4 / AI4: The Current Volume-Production Mainstay (2023) and the Emergence of the AI4.5 Stopgap

HW4 (AI4) was fitted to the Model S/X refresh in 2023 and progressively covered the Model 3/Y/Cybertruck from 2024. Its roughly 500 TOPS of compute represents an improvement of about 3.5× over HW3, and its 16GB of LPDDR5 memory permits the deployment of larger end-to-end neural networks.

A noteworthy fact is the quiet emergence of AI4.5: owing to the AI5 delay, Tesla introduced AI4.5 in the 2026 Model Y at the end of 2025 — a refined upgrade of AI4 (the core silicon design unchanged, but with adjustments to the process and memory configuration), intended to sustain the compute demand of the FSD neural network ahead of AI5 high-volume production. Tesla made no high-profile launch of AI4.5; its very existence reveals a fact: **the magnitude of the AI5 delay has already exceeded the window that a single AI4 generation could sustain.**

The Cybercab — Tesla's dedicated Robotaxi model — had its first unit off the line on 17 February 2026, with SOP in Q2 2026, and adopts AI4 rather than AI5 at the outset. This decision was made explicit in the Q1 2026 earnings report: Tesla's earlier commitment had been that the Cybercab would use AI5; the reality is that the Cybercab must begin volume production on the AI4/AI4.5 platform. This means that **the compute ceiling of the early Cybercab fleet will be constrained by the AI4 envelope,** rather than by the AI5 leap that Musk's long-running narrative had described.

1.4 AI5 / HW5: The Cross-Generational Leap (Taped Out 2026.4) and the "Not to Vehicles" Hardware Tiering

On 15 April 2026, Musk published the AI5 tape-out photograph on the X platform, announcing that Tesla had completed the GDSII data hand-off to TSMC. A tape-out is the final freeze of a chip's design; from tape-out to volume production still requires twelve to eighteen months.

The magnitude of AI5's engineering leap far exceeds that of HW3→HW4:

- **Compute equivalent to roughly 4,000 TOPS** (5× AI4 useful compute / 8× raw / 9× on-chip memory).
- **Memory expanded substantially to 192GB LPDDR5X** (HW4's 16GB → AI5's 192GB, a twelve-fold increase).
- **Removal of the ISP / graphics units** — AI5 is designed as a pure inference GPU, with even the image signal processor deleted by the design team. This means that AI5 will not be sold as a traditional SoC to other automakers; it is a chip dedicated to Tesla's full-stack inference workloads.
- **A dual-foundry strategy**: TSMC Arizona + Samsung Taylor, TX — both located on U.S. soil, hedging tariff risk and achieving supply-chain redundancy.

Yet the hardware tiering of AI5 is the "reality-check point" that Chapter 1 of D3 must highlight:

> In the Q1 2026 earnings call, Musk directly confirmed that **the first batch of AI5 silicon will be allocated preferentially to the Optimus humanoid robot and to the Tesla / xAI supercomputing training clusters, rather than to vehicles.** The Cybercab Q2 2026 SOP and all 2026 new vehicles will continue to use AI4 / AI4.5. AI5 high-volume production in vehicles is deferred to mid-2027.

This fact stands in evident contrast to Tesla's earlier narrative. In June 2024 Musk had promised "AI5 in vehicles 2H 2025"; in November 2025 the timeline slipped to "late 2026–early 2027"; and by the tape-out in April 2026 the Cybercab had been officially confirmed to still use AI4. In other words, **even though vehicles were originally the core use scenario of AI5's design intent, the engineering reality is that the priority of Tesla's own Optimus and xAI clusters for AI5 compute exceeds that of vehicle-side inference.**

This is no accident but an intrinsic structure of Tesla's AR4 closed-loop — when one and the same inference chip must serve (1) vehicle-side real-time inference, (2) Optimus edge inference, and (3) supercomputing training tasks, scarcity will naturally tilt toward the highest-value workload. Optimus and the xAI clusters are at the "most expensive stage" of the product definition (small batch, high unit price, high strategic significance); the vehicle side is at the "mature amortization stage" (large batch, marginal per-vehicle gross-margin contribution, AI4 already sufficient to support supervised FSD). **Tesla's allocation of AI5 to Optimus / xAI rather than to vehicles is not an engineering failure but a logical necessity of resource allocation under an AR4 vertical closed-loop.**

This point precisely reinforces the argument of the intelligence-layer discussion in Section 4.3 of Chapter 4: **Tesla's "unified stack" is not a single chip running two morphologies, but one and the same software stack running across two compute tiers** (vehicle-side AI4/AI4.5 + robot/supercomputer AI5), with the hardware difference absorbed by a software abstraction layer.

1.5 AI6 / AI7: The Future Road-Markers and the Terafab Plan

In the Q1 2026 earnings call, Musk disclosed further:

- **The AI6 tape-out target is December 2026** — which, if achieved, would represent a "generational cadence" of nine to twelve months, far faster than the eighteen-to-twenty-four-month architecture-upgrade cycle typical of the semiconductor industry.
- **AI7 is already in early planning.**
- The **Dojo 3** chip project is "advancing in synchrony with AI6."

Yet the fact of greater structural significance is the **Terafab plan** — a "chip-manufacturing innovation project" jointly announced by SpaceX and Tesla in **March 2026,** which Intel joined in **April 2026** (investment scale roughly $25B, sited in Austin, Texas). The SpaceX S-1 prospectus (2026.5.20) gives the official definition of Terafab:

> "Terafab — a chip manufacturing initiative with a long-term goal of producing one terawatt of compute hardware each year."

The S-1 also makes clear that Terafab is at present merely a **"general framework,"** with "specific projects still subject to separate negotiation and agreement" (including any development timelines, milestones and capital expenditures), and that specific funding proportions and project timelines have not yet been finalized.

The strategic intent of Terafab is clearly stated in the "AI Compute Infrastructure" section of the SpaceX S-1:

> "We believe that the key constraints in the continued growth of AI are physical—chip manufacturing, data center infrastructure, and power generation; the future of AI will be determined by the control of the physical stack."

This statement reveals Terafab's true user: **SpaceX's own deployment plan for orbital AI compute satellites** (deployment beginning in 2028, with a target of launching 100GW/year of compute into Sun-synchronous orbit). Tesla is a collaborator in the Terafab framework, but Tesla's procurement volume of vehicle-side AI chips is far smaller than the "megaton-scale" hardware deployment required by SpaceX's orbital AI satellites.

The methodological significance for D3: Tesla's in-house chip path has already evolved from "a component of the automotive business's vertical closed-loop" into "the downstream of the group-level AI compute strategy of SpaceX-xAI-Tesla." This is the extreme form of organizational debt — when Musk simultaneously controls a board majority across the eight entities of Tesla, SpaceX, xAI, X, Neuralink, Boring, Macrohard, and Terafab, resource allocation no longer obeys the optimum of a single company but the "group optimum" under Musk's personal perspective. **Chapter 7 will further develop the structural implications of this organizational reuse.**

1.6 The Architectural Implications of the Silicon Road

Tesla's silicon road yields four key implications for AR4 architecture.

First, **an in-house inference chip is the hardware prerequisite for AR4, but not a sufficient condition.** NVIDIA, Huawei, and Mobileye also have powerful in-house chips, but Tesla's distinctiveness lies in the homologous co-design of chip design intent with the upper-layer algorithms and training infrastructure — this is an organizational-level integration, not merely chip capability. The decision to remove the ISP/graphics units and evolve AI5 into a pure inference GPU is precisely the engineering outcome of this homologous co-design — a traditional SoC vendor (even one as technically capable as NVIDIA) cannot make such targeted specialization, because it must serve the diverse workloads of multiple customers.

Second, **the tension between hardware generations and software evolution is a cost of the vertical closed-loop, not a side effect.** The HW3 unsupervised-FSD withdrawal event reveals a general law: when the pace of software evolution outstrips the cadence of hardware generations, the

commitment to the installed base becomes an engineering burden. This contradiction is partially passed on to the supplier under the "hardware-procurement" model, but is borne in full by the enterprise itself under the vertical closed-loop. Tesla's microfactory retrofit network and free retrofit are the concretization of this cost.

Third, **the AR4 path necessarily produces "hardware tiering."** The fact that the first batch of AI5 goes not to vehicles but to Optimus / xAI clusters appears to contradict the AR4 "unified stack" narrative, but in fact reveals the true structure of the AR4 closed-loop — **not a single piece of silicon serving all morphologies, but one and the same software stack running across multiple compute tiers, with the hardware difference absorbed by a software abstraction layer.** This distinction is of directional significance for the fast-followers' tool choices: to attempt to replicate Tesla's AR4 by "procuring a single high-compute chip" is, in essence, to misread the true mechanism of AR4.

Fourth, **the silicon road has already exceeded the boundary of the automotive business and evolved into a group-level AI compute strategy.** The Terafab framework shows that the downstream users of Tesla's in-house chips already include SpaceX's orbital AI compute satellites — a clear signal of Musk's group-level resource allocation. For other OEMs this means that **to catch up with Tesla's AR4 path requires not only catching up with the vertical closed-loop of its automotive business, but also implies catching up with a group-level resource-allocation capability spanning automobiles, robots, and space.** This is not replicable.

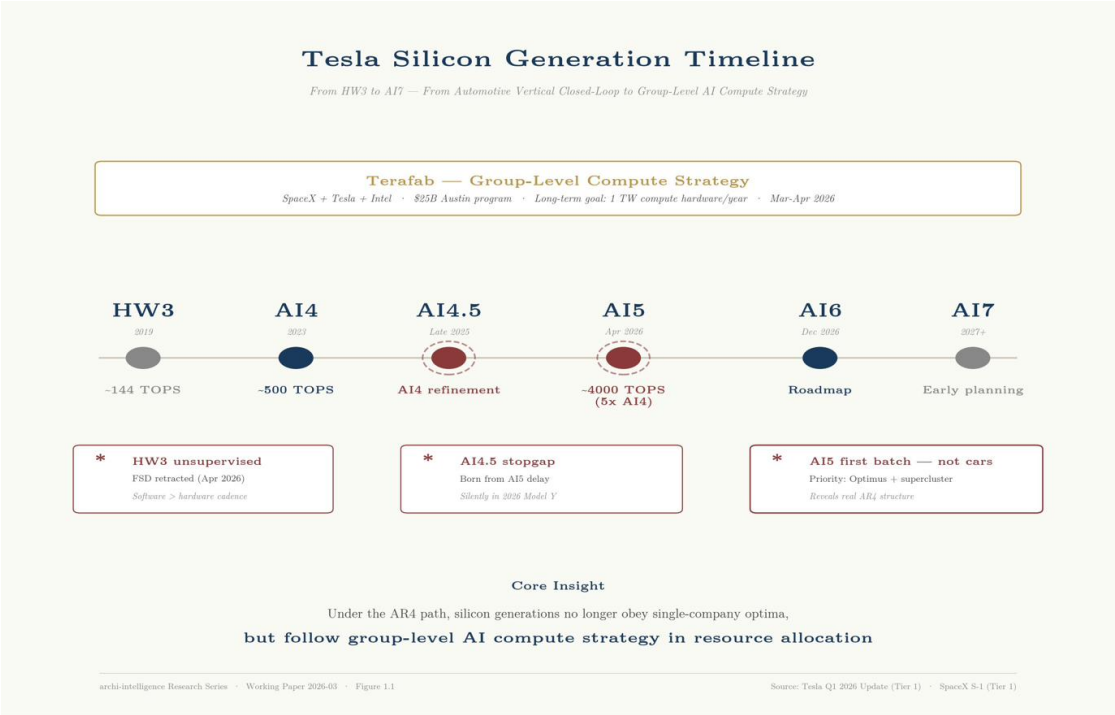


Figure 1.1 The Timeline of Tesla's Silicon Generations — From HW3 to AI7: the evolution from the automotive business's vertical closed-loop into a group-level AI compute strategy. The three red "★" marks denote the key reality-check points: the withdrawal of HW3 unsupervised FSD (2026.4), the AI4.5 stopgap chip (born of the AI5 delay), and the first AI5 batch not allocated to vehicles (preferentially to Optimus + supercomputer). The Terafab group-level compute strategy banner serves as the upper-layer context.

Chapter 2 The Evolution of the FSD Software Stack: The Paradigm Revolution from Rules to End-to-End

If the silicon road is the physical point of departure for Tesla's AR4 architecture, then the evolution of the FSD software stack is its soul. From V11 to V14, Tesla accomplished a **paradigm revolution** rare in the field of autonomous driving — a shift from hand-written-rule-driven to end-to-end-neural-network-driven. This transformation is not merely a choice of technical route but the software prerequisite for the AR4 criterion of a "unified closed-loop of training, simulation, and deployment" to hold. As of May 2026, V14 had iterated to v14.3.3, and a fact worthy of methodological attention had emerged: **in the Q1 2026 earnings report Musk stated explicitly that "the FSD v14.3 software stack is architecturally sufficient to support unsupervised deployment; what remains is validation and regulatory approval, not a question of technical capability."** This statement supplies key material for the argument of Chapter 10 that "AI is a faculty of language, not a brain" — if Tesla's own judgment is correct, then validation and regulation are the true bottleneck of AR4, rather than the capability of the neural network itself.

2.1 V11 and Before: The Twilight of the Modular Rule-Stack

Before V12, Tesla FSD adopted the industry-mainstream **modular architecture**: perception, planning, and control as three modules interacting through clearly defined interfaces. The perception module output structured results such as bounding boxes and lane-line segments; the planning module made decisions on the basis of these results; and the control module executed.

The core bottleneck of this architecture lies in **interface information loss.** In the traditional method, perception results are passed in the form of bounding boxes or segments, which limits the expressive power of the information — the complex scenes of the real world are compressed into a finite set of structured labels, and a great deal of implicit semantics is lost in transmission between modules. V11 was the culmination of this modular paradigm, and also its twilight. According to public information, V11 contained more than roughly three hundred thousand lines of C++ control code — hand-written rules that covered countless specific scenarios, but that also became technical debt difficult to maintain and difficult to generalize.

It is worth noting that **V11 still runs to this day on the HW3 fleet** (specifically the v12.6 branch), coexisting in parallel with the V14 branch of the HW4 fleet. Tesla has promised the HW3 fleet a "V14 Lite" — a streamlined end-to-end version targeted at the HW3 compute envelope — but this commitment was again made explicit in the Q1 2026 earnings report as "by end of June 2026," meaning it is already about six months later than originally planned. This software-version fork is itself another concretization of the "tension between hardware generations and software evolution" under the AR4 path.

2.2 V12: The End-to-End Revolution (2024.1)

V12 was the most significant architectural rupture in the history of Tesla FSD. Its core transformation was to **replace roughly three hundred thousand lines of C++ control code with a single end-to-end neural network.** Tesla summarized this paradigm as "Photon In, Control Out" — the entire process from camera-pixel input to vehicle-control output is accomplished by a single neural network, with no explicit module boundaries of perception/planning/control.

The essence of this transformation is the redefinition of the autonomous-driving problem as an

imitation-learning problem. The end-to-end large model in essence compresses a vast dataset of driving video into network parameters, in a manner highly analogous to a generative language model compressing internet-scale text into model parameters. FSD thereby became a product of the "Software 2.0" era — driven entirely by data, rather than by engineer-hand-written logic.

V12 resolved the "interface bottleneck" of the modular system: perception and planning/control no longer pass information through a lossy interface, but are merged into a single neural-network structure, with information flowing internally in the form of high-dimensional features, avoiding the information-compression loss of structured labels.

2.3 V13: HW4-Native Resolution (2024.12)

V13 was the deepening of the V12 paradigm on HW4 hardware. Its key improvements included HW4-native-resolution input (fully exploiting AI4's higher camera bandwidth and compute) and an expansion of the training-data scale by roughly 4.2× relative to V12. V13 demonstrated the **data scalability** of the end-to-end paradigm — more data, higher resolution, and greater compute translate directly into capability gains, which is precisely the core promise of the "Software 2.0" paradigm.

2.4 V14: Multimodality and the World Model (2025-2026)

V14 is the current (2026.5) mainstay version. As of mid-May 2026, the latest publicly tracked version was **V14.3.3 (firmware 2026.14.6.6),** which began pushing to early-access users on 2026.5.17, adding primarily an **"intervention-free live streak counter"** — a real-time counter displaying "the continuous distance driven since the last human intervention." Tesla's official intent is, by letting users see in the vehicle the cumulative "intervention-free continuous mileage," to build market expectation of the imminent arrival of unsupervised FSD.

The core network of V14 exhibits a pronounced **multimodalization**:

Inputs (multimodal): - seven high-resolution camera video streams - the vehicle's own motion information - navigation signals - audio signals

Outputs (multitask): - semantic segmentation - occupancy grid - 3D Gaussian features - language expression - the final control action

This input-output structure reveals a key trend: **FSD V14 may already have incorporated a vision-language-action (VLA) framework,** endowing the model with the capacity to "explain" and to "think." The existence of a language-expression output means that the model is not merely performing a "perception-control" mapping, but is also constructing a semantic understanding of the scene — a direction highly convergent with that of the Chinese new entrants (XPeng VLA 2.0, Li Auto Mind GPT VLA, Huawei ADS 4.0 WEWA).

Musk has disclosed that the parameter scale of the FSD V14 system under development is expanded by roughly 4.5× relative to V13, and this parameter inflation is precisely the source of AI5's high-compute demand — HW4's compute envelope already struggles to carry the full potential of V14. But under the reality of AI5 high-volume production deferred to mid-2027, the vehicle-side deployment of V14 must be accomplished on AI4/AI4.5 — a constraint that in turn shapes the engineering-optimization direction of V14, including the rewrite of the MLIR-based AI compiler (mentioned in the v14.3 release notes) to maximize utilization of the existing hardware's compute.

Data-nature annotation: some third-party analyses describe V14 as adopting a "generative

world model (GWM)" and treating the environment as a "4D simulation." These formulations originate in third-party blog interpretations; Tesla has not officially confirmed the term "GWM." This report annotates them as directional interpretations and does not adopt them as Tesla's official architectural definition.

2.5 Robotaxi Deployment and the Regulatory Reality

The capability progress of V14 must be assessed against the actual progress of Tesla's Robotaxi commercialization deployment.

Deployment status (as of May 2026): - In June 2025, Robotaxi launched in Austin (Tesla's first unsupervised commercial-operation city). - On 18 April 2026, it expanded to Houston and Dallas. - At present a total of **three cities** (Austin / Houston / Dallas), all in the state of Texas. - Independent observers (based on license-plate tracking) note that the number of vehicles actually in operation in each city is only **around two** — an extremely small sample size.

The revision of the commitment: Tesla had earlier (2025 Q4 earnings) committed to operating in seven U.S. cities within 1H 2026 — Austin, Houston, Phoenix, Miami, Orlando, Tampa, Las Vegas. **In the Q1 2026 earnings report, the wording for the remaining five cities was changed from "1H 2026" to the vague "preparations underway."** With only two months remaining until 1H 2026, this revision substantively constitutes a deferral of the five-city commitment.

International regulation: - **China**: the Q1 2026 earnings report confirmed that supervised FSD has begun pushing to users in China, with full regulatory approval expected in Q3 2026. Tesla launched its localization for the Chinese market through an early-2026 OTA update (version 2026.2.9). - **Europe**: in April 2026, Tesla "completed the final vehicle-testing phase of supervised FSD" with the Netherlands' RDW (Netherlands Vehicle Authority). The actual deployment of FSD V14 in Europe requires passing multiple regulatory reviews such as UN-R157.

> The real-world cadence of Robotaxi deployment forms a striking gap with the "unsupervised FSD is about to be solved" narrative advanced by Musk in 2024–2025. **This gap is itself key material for the argument of Chapter 10 that "AI is a faculty of language, not a brain"** — if FSD V14.3 is already "architecturally sufficient to support unsupervised" (Musk's public statement of 2026.4), then the remaining workload of "validation + regulatory approval + city-scale deployment" already exceeds the construction of neural-network capability per se. This corroborates the judgment of the D1 flagship report: **AI is a faculty of language; the skeleton is the deterministic safety-case / validation / regulatory coordination.**

2.6 The Architectural Implications of the Software-Stack Evolution

The evolution of the FSD software stack yields four key implications for AR4 architecture.

First, **the end-to-end paradigm is the algorithmic prerequisite for cross-morphology reuse.** In a modular architecture, the planning rules hand-written for vehicles cannot transfer to robots; but what an end-to-end neural network learns is a general "perception-to-action" mapping capability, and this capability can in principle transfer across morphologies (see Chapter 4).

Second, **the Software 2.0 paradigm transforms capability gains into a question of data and compute.** Once the architecture shifts from "hand-written rules" to "data-driven," the ceiling of capability shifts from "how many scenarios the engineers can imagine" to "how much data can be

collected and trained on." This is precisely the source of the value of Tesla's data closed-loop (the C5 control point).

Third, **multimodal VLA-ization is a signal of cross-industry convergence.** FSD V14's incorporation of a VLA framework, consistent with the evolutionary direction of the Chinese new entrants (XPeng, Li Auto, Huawei), corroborates the flagship report's judgment that "AI model architectures are general across domains" — Transformer + occupancy + VLA is becoming the general architectural substrate of physical AI.

Fourth, **software capability ≠ deployment capability.** Tesla itself (2026.4 earnings) confirmed that FSD v14.3 is "architecturally sufficient for unsupervised," yet the actual Robotaxi deployment runs in only three cities × two vehicles per city, and the seven-city commitment has been quietly revised. This gap reveals the true bottleneck of AR4 implementation — **not neural-network capability, but validation, safety-case, regulatory coordination, and city-scale operational infrastructure.** This is the clinical empirical evidence for the D1 proposition that "AI is a faculty of language, and the skeleton is the deterministic engine" (see Chapter 10).

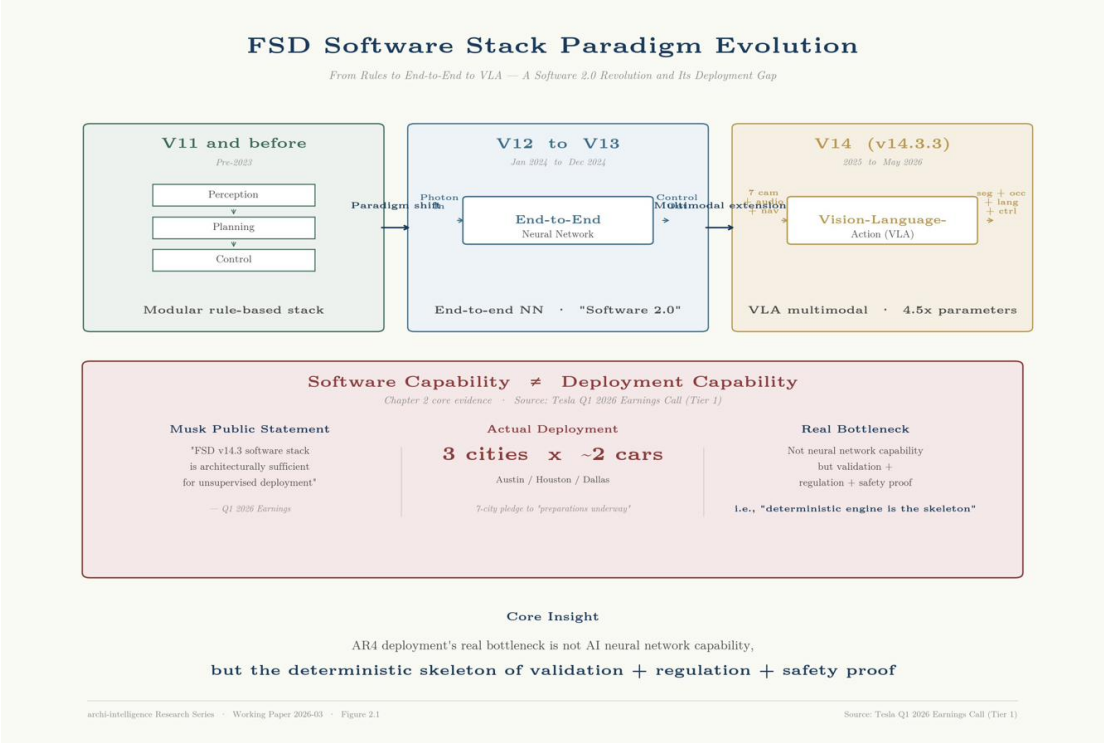


Figure 2.1 The Paradigm Evolution of the FSD Software Stack — From Rules to End-to-End to VLA: a Software 2.0 paradigm revolution and its deployment gap. The three paradigm segments (V11 modular rule-stack → V12-V13 end-to-end → V14 VLA multimodal) display the leap in software capability; the central red contrast block displays the threefold gap between Musk's public statement ("v14.3 is architecturally sufficient for unsupervised") versus the actual deployment (3 cities × ~2 vehicles) versus the true bottleneck (validation + regulation + safety-case).

Chapter 3 The Occupancy Network and Morphology-Agnostic Perception

If the end-to-end paradigm is the "algorithmic prerequisite" for cross-morphology reuse, then the **occupancy network (Occupancy Networks)** is the "core technical mechanism" that makes this reuse realizable at the perception layer. To understand the occupancy network is the key to understanding why the FSD stack can migrate to Optimus.

3.1 The Technical Essence of the Occupancy Network

The occupancy network was made public by Tesla at the 2022 AI Day. Its core idea is to **represent the world as a voxelized 3D dense occupancy field** — space is divided into a large number of small cubes (voxels), the network predicts whether each voxel is "occupied" or "free," and superimposes an "occupancy flow" that predicts the direction of motion of each occupied voxel.

This representation differs essentially from the traditional "bounding box + classification" perception:

- **Bounding-box perception**: recognizes objects as rectangular boxes of predefined categories (car, person, bicycle). It cannot express an "obstacle of unknown shape" — an overturned cargo item, an irregular cluster of debris, is difficult to handle under the bounding-box paradigm.
- **The occupancy network**: does not care "what" an object is, only whether space "is occupied and how it moves." This enables the system to handle obstacles of arbitrary shape without predefining categories.

3.2 Why the Occupancy Network Is "Morphology-Agnostic"

The revolutionary quality of the occupancy network lies in its production of a **morphology-agnostic (embodiment-agnostic) 3D scene representation.**

Consider the essence of this representation: what the occupancy network generates from camera pixels is a dense understanding of "where the surrounding 3D space is occupied and how it moves." This understanding **does not depend on what morphology the perceiving subject is** — whether the camera is mounted on a car, a humanoid robot, or a drone, the physical fact of "the occupancy state of the surrounding space" is objective and morphology-agnostic.

It is precisely this property that allows the FSD vision-perception stack to migrate directly to Optimus. Tesla has stated publicly that, owing to the morphology-agnosticism of the occupancy network, when the FSD stack migrates to Optimus the **voxel size can be reduced to 10cm** — to accommodate the robot's higher demand for spatial resolution in fine indoor manipulation (such as grasping and obstacle avoidance). A 10cm voxel may be overly fine for the automotive scenario, but one and the same occupancy-network architecture can adapt to the robot scenario merely by adjusting voxel granularity — precisely the engineering manifestation of a "homologous perception stack."

3.3 The Position of the Occupancy Network Within Tesla's Perception Stack

The occupancy network does not exist in isolation but is embedded within Tesla's overall perception architecture. Its evolutionary lineage is as follows.

Early on, Tesla used HydraNet (a multi-head neural network) for multitask perception — a shared backbone plus multiple task heads (detection, segmentation, lane lines, and so on). After the introduction of the occupancy network, HydraNet was extended, the output of the occupancy network was integrated by the planner, and 3D scene-understanding capability was enhanced. After the end-to-end transition at

V12, the occupancy representation was further fused into the internal feature flow of the single end-to-end network — as the outputs of V14 still include the item "occupancy grid," indicating that the occupancy representation is retained as a key intermediate/output representation.

3.4 The Limitations and Controversies of the Occupancy Network

The occupancy network is not without limitations. Holding to a critical examination, this report records the following controversies.

First, **the reliability controversy of pure-vision occupancy.** The occupancy network generates 3D occupancy from cameras, and its accuracy is affected by illumination, weather, and occlusion. LiDAR measures 3D distance directly and is more robust under adverse conditions. Tesla's pure-vision occupancy approach still faces regulatory questioning on its L4 liability declarations — one of the AR4 costs to be discussed in Chapter 8.

Second, **the generalization boundary of long-tail scenarios.** The occupancy network performs excellently in common scenarios, but its generalization capability for extreme long-tail scenarios (rare obstacle morphologies, extreme weather) remains part of the interpretability challenge of end-to-end systems.

Third, **the theoretical boundary of morphology-agnosticism.** "The occupancy network is morphology-agnostic" is Tesla's engineering hypothesis, which has received some validation between automobiles and humanoid robots, but whether it can be extended to a broader range of morphologies (quadruped, wheeled, flying) remains an open question — and is indeed one of the open questions of the flagship report.

3.5 The Architectural Implications of the Occupancy Network

The core implication offered by the occupancy network is this: **the realization of cross-morphology reuse depends on finding a "morphology-agnostic intermediate representation."**

For players who wish to draw on this line of thinking, the key lies not in replicating the occupancy network itself but in understanding its design philosophy — constructing, between perception and decision-making, an objective representation of the physical world that does not depend on a specific physical morphology (the occupancy field is one such; the world model is another). Once such a representation is possessed, the upper-layer decision-making and control can be shared across different morphologies. This is the direction jointly pursued by architectures such as NVIDIA's Cosmos world model and Huawei's ADS 4.0 WEWA, and is also the common path of physical AI toward AR4.

Chapter 4 The Mechanism of Cross-Morphology Reuse: From Automobile to Humanoid Robot

The first three chapters established the foundations — silicon (AI5 shared across morphologies), software (the end-to-end paradigm is transferable), and perception (the occupancy network is morphology-agnostic). This chapter converges these threads into the core proposition of D3: **how Tesla extends the accumulated capability of its automotive business across morphologies to the Optimus humanoid robot, which layers are reusable, and where the engineering boundaries of reuse and the sources of non-replicability lie.** This is the empirical core of the AR4 "cross-morphology generalization" criterion.

The key point of departure for understanding this mechanism is Tesla's own official disclosure at the shareholder meeting of 6 November 2025 — which constitutes the argumentative anchor of Chapter 4 of D3.

4.1 The Official Disclosure of Cross-Morphology Reuse: Tesla's Eleven Layers of Shared Core Technology

The Tesla shareholder meeting of 6 November 2025 disclosed a **complete inventory of cross-morphology shared technology,** formally titled "Shared core technology" — explicitly listing eleven layers of shared core technology from transport (the transport/automotive business) to robotics (the robotics business). This disclosure is Tier 1 material and constitutes the official anchor of D3's argument for cross-morphology reuse.

Table 4.1: Tesla's Officially Disclosed Eleven Layers of Shared Core Technology (Shareholder Meeting, 2025.11.6)

No.	Shared Core Technology	Source of Accumulation in the Automotive Business	Form of Reuse on the Optimus Side
1	Actuators	Model S/X/3/Y motor manufacturing	22 DoF/hand + ~50 whole-body actuators, tendon-driven
2	Power electronics	Tesla in-house inverter + charge-discharge management	Robot power management and actuator drive
3	Battery	4680 cell + pack-integration capability	Optimus 2.3 kWh battery pack (carrying over Tesla cell-manufacturing experience)

No.	Shared Core Technology	Source of Accumulation in the Automotive Business	Form of Reuse on the Optimus Side
4	Manufacturing	Giga-factory gigacasting + assembly lines	Fremont Model S/X line converts to Optimus production 2026.5, design capacity 1M/year; Giga Texas designed for 10M/year
5	Data communication	Vehicle Ethernet + zonal-controller architecture	High-speed internal data interconnect of the robot
6	Audio system	In-car audio system + microphone array	Robot voice interaction + environmental perception
7	Cameras	8-channel Autopilot cameras	8 Autopilot cameras (same type)
8	A14/A15 chips (AI4/AI5 chips)	HW4/HW5 in-house inference chips	Optimus Bot Brain adopts AI4/AI5 (first AI5 batch preferentially to Optimus)
9	Training cluster	Cortex 2.0 (100K H100/H200, 250-500MW)	FSD + Optimus homologous training
10	Neural simulation	FSD video training + simulation environment	Optimus manipulation-scenario simulation
11	Real-world AI	End-to-end + occupancy network + world understanding	Homologous end-to-end stack, “Photon In, Action Out”

Source: Tesla Annual Shareholder Meeting Presentation, 2025.11.6 (Tier 1). The original slide subtitle “Shared core technology” explicitly lists the eleven items.

It is worth noting that in this official disclosure Tesla presents the eleven items side by side, without further grouping or hierarchical division — a reasonable choice of marketing discourse, but insufficiently engineering-oriented. While respecting Tesla’s official disclosure, D3 holds that these eleven items contain **fundamentally different mechanisms of reuse** internally — some items depend on physical factories and heavy-asset accumulation, some depend on algorithms and the data closed-loop, and some require organizational-level integration as support. This distinction is of major methodological significance for the fast-followers and is the core of the next section.

4.2 The Three-Group Structure of the Eleven Layers: The Three Reuse Mechanisms Distilled by D3

After analyzing the engineering substance of the eleven shared-technology layers, D3 holds that they should be understood as **three fundamentally different mechanisms of reuse,** plus an **organizational layer** that Tesla did not explicitly disclose but that is in fact the prerequisite for the eleven layers to hold.

Table 4.2: D3's Three-Group Structural Distillation of Tesla's Eleven Shared-Technology Layers

Group	Items Among the Eleven Layers	Reuse Mechanism	Accumulation Time	Source of Non-Replicability
Physical layer	1. Actuators2. Power electronics3. Battery4. Manufacturing5. Data communication6. Audio system	Physical factories + heavy assets + manufacturing know-how	10-20 years of automotive-business accumulation	Capital scale + factory network + supply-chain accumulation
Intelligence layer	7. Cameras8. AI chips9. Training cluster10. Neural simulation11. Real-world AI	Algorithm + data closed-loop + homologous training	5-10 years of FSD engineering accumulation	Data scale + in-house chip + end-to-end paradigm
Organizational layer (added by D3)	Cross-morphology team integrationFSD-Optimus merger	Organizational flatness + homologous KPI + homologous technical leadership	Immediate decision but requires organizational-culture fit	Corporate governance + single-point CEO coordination + reverse Conway's law

The fundamental difference among these three groups lies in the fact that **the difficulty structure of replicating them is entirely different for the fast-follower** — the physical layer requires capital scale and a duration of heavy-asset accumulation that is difficult to compress; the intelligence layer requires a data closed-loop and an algorithmic-evolution path that is shorter but still requires more than five years; and the organizational layer can in principle be adjusted immediately but is in practice deeply constrained by corporate governance and organizational culture. **The clear delineation of these three structural groups is the methodological prerequisite for D3's mirroring analysis of the twenty-one fast-followers in D2** — the subsequent mirroring of Volkswagen/Toyota/Huawei/XPeng/Xiaomi in Chapter 9 will make precise judgments on the basis of these three structural groups (see Chapter 9 for the five mirrors and the eleven-layer reusability matrix).

4.3 The Engineering Mechanism of the Five Intelligence-Layer Items

The intelligence layer covers items 7–11 of the eleven layers (cameras, AI chips, training cluster, neural simulation, real-world AI), and is the engineering concretization of the "FSD-stack cross-morphology transferability" mechanism already developed in depth in Chapters 1–3 of D3. This section distills the core mechanism of the reuse of the five intelligence-layer items.

Item 7, Cameras. Optimus adopts the same 8-channel Autopilot cameras as Tesla vehicles — same model, same interface, same calibration process. Behind this seemingly simple hardware reuse lies Tesla's years of accumulated camera-ISP (image-signal-processing) tuning, low-light-scenario optimization, and rain-and-snow-environment robustness — know-how that need not be re-engineered on the Optimus side.

Item 8, AI chips. The in-house path of the FSD inference chip (HW3 → AI4 → AI5) allows the Optimus Bot Brain to reuse the same-generation AI chip directly. Chapter 1 has argued that **the first batch of AI5 silicon goes preferentially to Optimus + the supercomputing clusters rather than to vehicles** — a resource allocation that reveals the true direction of intelligence-layer reuse: not a single chip running two morphologies, but one and the same software stack running across two compute tiers (vehicle-side AI4/AI4.5 + robot AI5), with the hardware difference absorbed by a software abstraction layer.

Item 9, Training cluster. Cortex 2.0 (Giga Texas, 100K H100/H200, 250-500MW) serves the unified training tasks of FSD + Optimus. A homologous training cluster means that the models of the two morphologies evolve under the same software stack, data pipeline, and training tools — the engineering concretization of Tesla's data closed-loop (C5) and training infrastructure (C6) as unified control points.

Item 10, Neural simulation. Tesla's in-house simulation environment, originally designed for FSD — the synthetic generation of vast driving scenarios and the controllable reproduction of long-tail events — now extends directly to the simulation of Optimus manipulation scenarios. The homologousness of the simulation stack lets the two morphologies share one and the same "synthetic-data generation → model training → physical deployment → real-world feedback" closed-loop.

Item 11, Real-world AI. This is the deepest layer of the intelligence layer — the end-to-end paradigm (from V12) + the occupancy network (a morphology-agnostic 3D representation) + world understanding (trained on tens of millions of miles of driving data). **The core moat of Optimus lies not in its mechanical structure but in the perception and world model it inherits from FSD** — precisely the core of the arguments of Chapter 3 (the occupancy network) and the preceding sections of Chapter 4 of D3.

The overall significance of the five intelligence-layer items: the engineering threshold of intelligence-layer reuse is 5-10 years of the FSD data closed-loop and algorithmic evolution — not unattainable for the fast-follower, but what is required is not merely technical capability but also a fleet foundation for a scaled data closed-loop (millions of vehicles + years of accumulated driving data).

4.4 The Accumulation Logic of the Six Physical-Layer Items

The physical layer covers items 1–6 of the eleven layers (actuators, power electronics, battery, manufacturing, data communication, audio), a dimension not fully developed before D3 v1.1 — the key reinforcement revealed by Tesla's eleven-layer official disclosure.

Item 1, Actuators. Optimus Gen 3's 22 DoF/hand + roughly 50 whole-body actuators (a tendon-driven design) derive their manufacturing capability directly from twenty years of motor-manufacturing

accumulation in Tesla's automotive business — from the induction motor of the Model S to the permanent-magnet synchronous motors of the Model 3/Y, and on to the high-torque drive system of the Cybertruck. Tesla's accumulation in precision-motor manufacturing makes the cost structure of Optimus actuators far lower than that of peers who "design motors specifically for robots."

Item 2, Power electronics. Tesla's in-house inverters, charge-discharge-management ICs, and thermal-management chips — these power-electronics capabilities, designed for the automotive business, serve directly the power management and actuator drive of Optimus. The reuse of power electronics is the key engineering foundation for the robot's lightweighting and endurance.

Item 3, Battery. Optimus's single 2.3 kWh battery pack is roughly thirty times smaller in scale than the Tesla Model 3's 75-82 kWh battery pack. But the core know-how of cell production (4680), pack integration, thermal management, and the safety case is one and the same. **Tesla's accumulation in the 4680 cell makes the Optimus battery cost far lower than that of peers who "develop a battery specifically for robots from scratch"** — a structural advantage unavailable to pure-robotics startups such as 1X Technologies, Figure, and Apptronik.

Item 4, Manufacturing. This is the most structurally significant of the eleven layers. Tesla's accumulation in Giga-factory gigacasting (front + rear gigacasting), assembly-line automation, and the standardization of whole-vehicle manufacturing processes becomes directly the physical foundation for the volume production of Optimus. **The Fremont Model S/X production line ceased production in May 2026 and was converted to an Optimus production line (design capacity 1M/year), while Giga Texas is designed for a long-term capacity of 10M/year** — a scale unprecedented in the robotics industry, the feasibility of which derives in considerable part from the factory experience amortized by the automotive business. The credibility of Musk's target of a per-unit Optimus cost below $20K rests in considerable part on the cost structure of "producing robots with automotive-manufacturing experience" — rather than the marginal cost structure of an independent robot production line.

Item 5, Data communication. The engineering capability of Tesla's vehicle Ethernet + zonal-controller architecture (zonal architecture) serves directly the high-speed internal data interconnect of Optimus. This reuse appears mundane but is in fact the engineering foundation of the robot as a "distributed perception-decision-execution system" — without vehicle-grade data-communication capability, the coordination of the robot's internal 50+ actuators and multiple sensor channels would face severe bottlenecks.

Item 6, Audio system. The accumulation of Tesla's in-car audio system, microphone array, and noise-suppression algorithms is reused for the voice interaction and environmental perception of Optimus. This item is relatively the least significant among the eleven layers, but as a complete-experience component of a "consumer-grade product," it remains a necessary piece of the puzzle for the commercial deployment of Optimus.

The overall significance of the six physical-layer items: the engineering threshold of physical-layer reuse is **10-20 years of heavy-asset accumulation in the automotive business** — factory networks, supply chains, manufacturing processes, battery capacity. This threshold is far higher for the fast-follower than the intelligence layer — the gap in the intelligence layer can be compressed within 5-10 years through aggressive investment, but the physical layer's factory construction, supply-chain establishment, and accumulation of manufacturing experience are difficult to replicate without paying an equivalent in time and capital scale.

4.5 The Organizational Layer: The Implicit Twelfth Layer

Tesla's official eleven-layer shared-technology inventory **does not explicitly list "organization"** — but D3 holds that organizational integration is the prerequisite for the eleven layers of sharing to hold, the implicit "twelfth layer."

The key fact of Tesla's organizational integration: in June 2025, leadership of the Optimus project was transferred from Milan Kovac to Ashok Elluswamy — the latter concurrently serving as Vice President of FSD/Autopilot. This personnel arrangement signifies the **de facto merger of the FSD and Optimus teams**: the same organization, the same technical leadership, and the same engineering methodology driving the two morphologies of automobile and robot simultaneously.

Why the organizational layer is the implicit twelfth layer: technical reuse, if unsupported by organizational reuse, degenerates into the pseudo-reuse of "two teams each maintaining similar code" — the implicit assumption behind the eleven-layer shared-technology inventory. Through the team merger, Tesla guarantees from the organizational level that "homologousness" is not merely a technical slogan but a genuine engineering synergy. Specifically:

- **The reuse of the six physical-layer items** requires the factory team, the supply-chain team, and the manufacturing-engineering team to collaborate across the "automotive business" and the "robotics business" — if organizationally there are two independent business units, the cross-morphology amortization of the factory will not occur automatically.
- **The reuse of the five intelligence-layer items** requires the software team, the AI team, and the chip team to collaborate across morphologies — if organizationally there are two independent technology stacks, the end-to-end paradigm will not automatically migrate to the robot.
- **The synergistic effect of the three groups of reuse** further requires CEO-level cross-morphology resource-allocation decisions — the decision to give AI5 to Optimus rather than to vehicles, the decision to convert Fremont to Optimus production, and the decision to merge the FSD-Optimus teams are all resource allocations decided directly by the CEO.

This is the concrete form of D1's "reverse application of Conway's law" — **to realize a technically cross-morphology reuse architecture, Tesla proactively adjusted its organizational structure so that one and the same team is responsible for two morphologies simultaneously.** The integration of the organizational layer is the key converter that transforms the eleven layers of shared technology from "latent possibility" into "engineering reality."

The non-replicability of this organizational layer: cross-morphology organizational integration can **in principle be adjusted immediately** (appointing a cross-morphology vice president, merging two business units), but is **in practice deeply constrained by corporate governance and organizational culture** — the business-unit boundaries, profit allocation, and assessment systems of a traditional OEM (such as Volkswagen or Toyota) are highly incompatible with "cross-morphology integration." This is a source of non-replicability different from that of the physical layer and the intelligence layer — not time or capital, but **organizational culture** and **corporate governance.**

4.6 The True Cost Structure and Non-Replicability of Cross-Morphology Reuse

Having integrated the eleven layers of shared technology + the organizational twelfth layer, one can characterize more precisely the true cost structure and non-replicability of "cross-morphology reuse."

The multiplier economics of reuse: the core economic significance of the twelve-layer reuse is the **cross-product amortization of R&D cost and heavy-asset cost.**

- A traditional robotics company must bear independently the entire cost of chips, perception algorithms, training infrastructure, motor manufacturing, battery capacity, and factory construction, and can amortize it only over the single product line of the robot.
- Tesla, by contrast, amortizes these costs principally over the automotive business (millions of units of annual production), and the robot, as a "marginal reuser," inherits this set of capabilities at almost zero incremental cost.
- This multiplier effect is precisely the core return of the vertical-closed-loop AR4 architecture: **one and the same set of underlying capabilities serves multiple physical morphologies, each morphology diluting the cost of the others.**

The feasibility of Musk's Optimus cost target (a per-unit cost below $20K at a scale of a million units per year) rests in considerable part on this twelve-layer cross-morphology amortization cost structure — a cost boundary unattainable by any independent robot manufacturer.

The true threshold of reuse and the sources of non-replicability:

Layer Group	Principal Source of Non-Replicability	Compression Possibility for the Fast-Follower
Physical layer (6 items)	Capital scale + factory network + supply chain + manufacturing experience, **10-20 years of accumulation**	Extremely low — the time of heavy-asset accumulation is difficult to compress, and the gap in capital scale is structural
Intelligence layer (5 items)	Data closed-loop + in-house chip + end-to-end paradigm, **5-10 years of accumulation**	Moderate — high technical attainability but requires large-scale fleet data
Organizational layer (1 item)	Corporate governance + organizational culture + CEO cross-morphology decision-making capability	Seemingly low (a single appointment) but in fact extremely high (deeply constrained by organizational culture)

The honest boundary of reuse: cross-morphology reuse has a natural ceiling and is by no means "everything is reusable" — the motion-control layer (kinematics, dynamics, bipedal balance, contact dynamics, ~1kHz high-frequency real-time control, tendon-driven actuator characteristics) is almost non-reusable (see Chapter 5). But the very existence of this non-reusable boundary makes manifest the scarcity of the twelve reusable layers — **it is Tesla's lead in the reusable layers that constitutes a structural advantage.**

The core lesson for other players: the true picture of cross-morphology reuse is not as simple as "homologous migration of the algorithm or software layer" — it is the overall capability of the synergy of three groups: the physical layer (6 heavy-asset items) + the intelligence layer (5 algorithm/data items) + the organizational layer (1 governance/culture item). Any attempt that seeks merely to "imitate Tesla with the software layer" underestimates the true threshold behind the eleven layers of shared technology. Tesla is the "only complete AR4 closed-loop" precisely because of the complete synergy of these twelve layers — rather than the strength of any single dimension. This is the physical underlying foundation of the AR4 cost-boundary argument of Chapter 8 of D3, and the methodological anchor of the precise judgments of the five mirrors of Chapter 9.

Figure 4.1 The Twelve-Layer Stack of Cross-Morphology Reuse — Tesla's eleven officially disclosed layers of shared core technology (disclosed at the shareholder meeting of 2025.11.6) + the organizational twelfth layer distilled by D3. Three groupings: the physical layer with 6 items (actuators, power electronics, battery, manufacturing, data communication, audio) + the intelligence layer with 5 items (cameras, AI chip, training cluster, neural simulation, real-world AI) + the organizational layer with 1 item (FSD-Optimus team merger). At the bottom is the non-reusable layer (the motion-control class).

Chapter 5 Optimus Kinematics and the Non-Reusable Layer: The Engineering Granularity of the Differences

Chapter 4 argued for "the reusable"; this chapter argues for "the non-reusable" — equally important. The cross-morphology reuse of AR4 is not a utopia of "everything is reusable," but a precise distinction between "the reusable perception and world understanding" and "the non-reusable kinematics and control." Only by understanding this boundary can one avoid an over-romanticization of cross-morphology reuse.

5.1 The "Cerebrum" Is Reusable; the "Cerebellum" Is Not

A useful analogy: in cross-morphology reuse, **Tesla's eleven layers of shared core technology argued in Chapter 4 + the "twelfth layer" of organizational integration cover the "cerebrum" (perception, world understanding, high-level decision-making) and the "physical hardware" (actuators, power electronics, battery, manufacturing) — but the "cerebellum" (low-level motion control, kinematics, bipedal balance, contact dynamics) is almost non-reusable.**

The reason lies in the fundamental difference of physical morphology. The motion planning of autonomous driving is conducted in a **low-dimensional space** — the motion of a vehicle on the road plane is in essence trajectory planning on a 2D plane (forward, steering, acceleration and deceleration), with low degrees of freedom and a relatively simple kinematic model. The motion control of a humanoid robot, by contrast, is a **high-dimensional problem** — involving the coordination of dozens of whole-body degrees of freedom, the dynamic transfer of the center of gravity, bipedal force-feedback balance, and anti-fall disturbance-rejection control. This is an entirely independent and entirely new physical-control system, separate from the driving logic.

5.2 The Kinematic Specifications of Optimus

To understand the granularity of the non-reusable layer requires an examination of Optimus's concrete kinematic specifications. Here it must be honestly annotated: **the statements of different public sources regarding the degrees of freedom of Optimus differ,** and this report records this uncertainty rather than forcibly unifying it.

Hand degrees of freedom (relatively consistent data): - Gen 3 hand: 22 degrees of freedom (DoF) per hand - Total number of actuators for both hands: roughly 50 (roughly 25 per forearm/hand) - A doubling relative to the Gen 2 hand (11 DoF per hand), with the number of actuators growing by about 4.5× - Drive method: tendon-driven + coreless motors, a biomimetic design

Whole-body degrees of freedom (sources differ; differences annotated): - Some sources: 40+ degrees of freedom - Some sources: 37 degrees of freedom, hand positioning accuracy 0.08mm - The human hand as a reference: roughly 27 degrees of freedom (the "gold standard" of manipulation)

> **Data-nature annotation**: the specific figure for Optimus's whole-body degrees of freedom differs across public sources (Tesla official, third-party analyses, media reports), and strictly speaking the "Gen 3" designation refers specifically to the upgraded hand, while the core robot platform (torso, legs, main computer, battery) continues the Gen 2 design. This report takes "hand 22 DoF / both hands roughly 50 actuators" as relatively certain data, and annotates whole-body degrees of freedom as "roughly 37-40+, sources differ."

5.3 The Engineering Content of the Non-Reusable Layer

Specifically, the following layers **cannot be reused in the FSD → Optimus migration and must be re-engineered:**

First, **the whole-body kinematic and dynamic models.** The robot's inverse kinematics (solving joint angles for a given target pose) and dynamics (the relationship between joint torques and motion) are robot-specific and share nothing with the vehicle's two-dimensional trajectory planning.

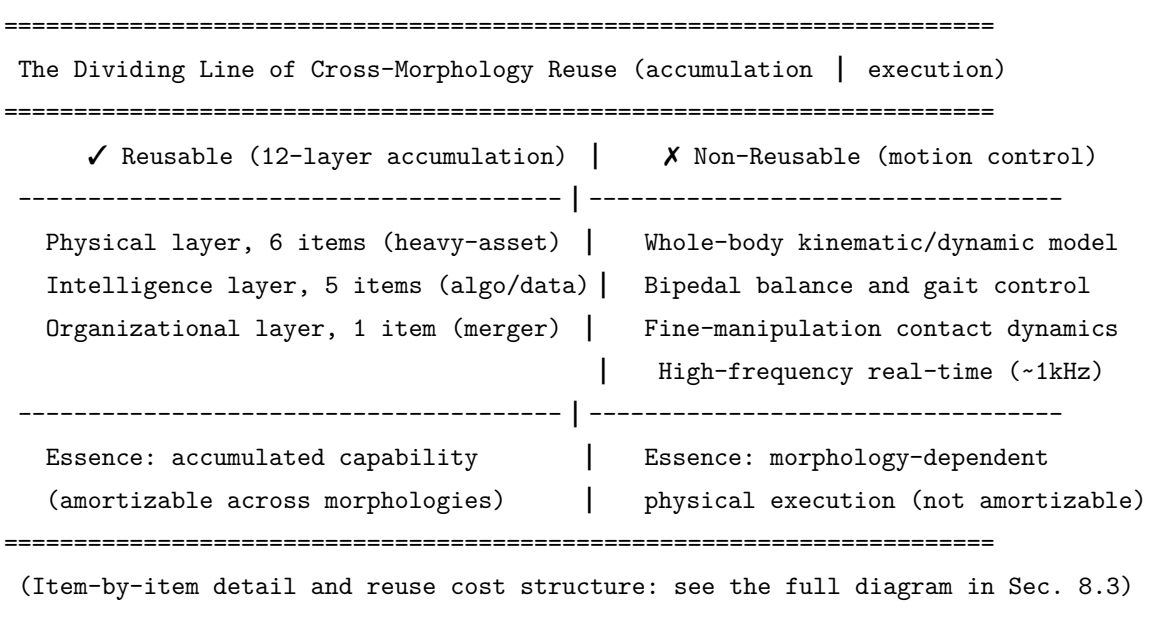
Second, **bipedal balance and gait control.** Walking involves the dynamic transfer of the center of gravity, zero-moment-point (ZMP) control, and anti-fall disturbance rejection — problems that do not exist at all for the vehicle.

Third, **fine-manipulation control.** The grasping, force control, and tactile feedback of a 22-DoF hand involve contact dynamics that do not exist at all for the vehicle.

Fourth, **the difference in control frequency and real-time requirements.** The robot's whole-body control requires a frequency far higher than that of vehicle control (robot balance control often requires the 1kHz class, while the vehicle control frequency is lower), imposing different requirements on the real-time system.

5.4 The Boundary Between the Reusable and the Non-Reusable

Integrating the eleven-layer shared-core-technology + organizational-layer (twelfth) framework of Chapter 4 with the non-reusable layer of this chapter, one obtains the precise boundary of cross-morphology reuse:



This boundary diagram reveals a key fact — **the dividing line between the reusable and the non-reusable is not the traditional distinction of "hardware vs software," but the distinction of "accumulated capability vs morphology-dependent physical execution."**

Although the six physical-layer items involve a great deal of hardware (actuators, batteries, power electronics, factories, and so on), their reuse mechanism is the **amortization of accumulated capability** — the factory network, manufacturing experience, and cell capacity accumulated over 10-20 years of the automotive business can serve the robot across morphologies, because their essence is "replicable engineering capability + heavy assets." Conversely, although the non-reusable motion-control layer

partly involves software (kinematic models, force-feedback algorithms), its essence is a **high-frequency real-time control system tightly bound to a specific physical morphology** — the vehicle's 2D-plane trajectory planning and the humanoid robot's whole-body high-dimensional dynamic control are two entirely independent physical-control systems that cannot be amortized across morphologies through accumulated experience.

This dividing line is of major significance for the fast-follower: **Tesla's lead in the twelve layers of sharing comes from 10-20 years of accumulation and is difficult to replicate quickly; but its advantage in the non-reusable motion-control layer is relatively small** — any team with robot-control engineering capability (including pure-robotics companies such as 1X Technologies, Figure, and Apptronik) can compete with or even surpass Tesla at this layer. The true structural moat of Optimus lies in the twelve layers of sharing, not in motion control per se.

5.5 The Strategic Implications of the Non-Reusable Layer

The existence of the non-reusable layer has three important implications for AR4 architecture.

First, **cross-morphology reuse has a natural ceiling.** Even an ultimate vertical-closed-loop case such as Tesla's "twelve-layer sharing + organizational integration" cannot achieve "100% reuse." The re-engineering of the motion-control layer is the additional investment of the robot relative to the vehicle. Any narrative claiming that "automotive capability can seamlessly become robot capability" ignores this non-reusable boundary. Tesla itself must also invest a considerable-scale dedicated motion-control team for Optimus, the mechanical design of tendon-driven actuators, and the development of whole-body dynamic-control algorithms — engineering volumes that cannot be amortized with the automotive business.

Second, **the scarcity of the reusable layers stems from three fundamentally different accumulation mechanisms.** This is a methodological dimension not fully developed in D3 v1.0 / v1.1 — the twelve reusable layers are not a homogeneous "reusable capability," but stem from three structurally different scarcity mechanisms:

- **The scarcity source of the physical layer = factory network + heavy assets + manufacturing know-how, 10-20 years of automotive-business accumulation.** Tesla's Fremont factory, the Giga Shanghai/Berlin/Texas factories, 4680 cell capacity, and the in-house inverter supply chain — the scarcity of these assets stems from the dual threshold of **capital scale + time.** Even with unlimited capital, any competitor would struggle to build an equivalent factory network and manufacturing-process maturity within five years.
- **The scarcity source of the intelligence layer = data closed-loop + in-house chip + end-to-end paradigm, 5-10 years of algorithmic-evolution accumulation.** FSD's tens of millions of miles of driving data, the three-generation chip evolution HW3→HW4→AI5, and the four iterations of the V11→V14 end-to-end paradigm — the scarcity of these assets stems from **data scale + the path-dependence of the engineering evolution.** Even with equivalent technical capability, a competitor would need several years to accumulate an equivalent data closed-loop.
- **The scarcity source of the organizational layer = corporate governance + organizational culture + CEO cross-morphology decision-making capability.** Tesla's flat organization, cross-business-unit homologous KPIs, and Musk's coordination across eight entities — the scarcity of these assets is the most peculiar: **it can in principle be adjusted immediately, but is in practice deeply constrained by corporate governance and organizational culture.** The business-unit

boundaries, profit allocation, and assessment systems of a traditional OEM are all highly incompatible with "cross-morphology integration"; this is not something that time or capital can make up for, but **the difficulty of organizational-cultural change.**

The three scarcity mechanisms each have a different "difficult-to-fast-track" mechanism — capital / time / governance — and together constitute Tesla's structural lead in the twelve layers of sharing. This tripartite characterization of scarcity mechanisms is of major methodological significance for the fast-follower: **different scarcity mechanisms correspond to different response paths.** For physical-layer scarcity, the fast-follower's reasonable path is to focus on its own existing physical accumulation (such as Toyota's manufacturing, Huawei's ICT accumulation); for intelligence-layer scarcity, the reasonable path is aggressive data-closed-loop construction and an algorithmic-paradigm revolution (such as XPeng's VLA 2.0); for organizational-layer scarcity, the reasonable path is the proactive reshaping of corporate governance and organizational structure (such as Huawei's "not entering vehicle manufacturing directly, focusing on Tier-1" organizational choice, which is itself a governance decision).

Third, **this boundary is precisely the empirical evidence for the flagship report's "Open Question Two."** The "occupancy network is morphology-agnostic" hypothesis is validated between automobiles and humanoid robots (the intelligence layer's Real-world AI is reusable), but the morphology-dependence of the kinematics layer (the motion-control layer is non-reusable) is equally validated. **Whether the twelve layers of sharing can be extended to a broader range of morphologies (quadruped, wheeled, flying)** is the open question the D1 flagship report leaves for subsequent research. On the basis of the twelve-layer framework, D3 can make a more fine-grained, layered judgment on this question of extensibility:

- **The extensibility tiering of the six physical-layer items:**
  - **Likely extensible**: Battery (4680 cell), Power electronics (inverter/power management), Data communication (vehicle Ethernet architecture) — these are engineering capabilities independent of the specific morphology, theoretically serviceable to quadruped, wheeled, and flying platforms.
  - **Partly extensible**: Manufacturing factory experience — depending on production scale and manufacturing-process similarity. If a quadruped robot reaches the million-units-per-year class, factory experience can be amortized; small-scale special-purpose use is difficult to reuse. Audio system is similar.
  - **Strongly morphology-dependent**: Actuators — the high-torque tendon actuators of a bipedal robot, the small-to-medium rotary actuators of a quadruped robot, and the high-RPM motors of a flying platform differ markedly in mechanical structure, with diminishing reuse of manufacturing experience.
- **The extensibility tiering of the five intelligence-layer items:**
  - **Theoretically extensible**: Real-world AI (end-to-end + occupancy network), Neural simulation, Training cluster, AI chips — these are the "morphology-agnostic perception + general world model" argued in D1, in principle serviceable to any morphology.
  - **Strongly morphology-dependent**: Cameras — camera number, position, and viewing angle must be adjusted according to morphology. But the camera hardware + ISP-tuning know-how itself is reusable.
- **The extensibility of the one organizational-layer item**: the cross-morphology team-merger mechanism itself is extensible to any new morphology. But it requires CEO-level resource-allocation decisions and organizational-culture fit.

This layered judgment offers a clear answer to the D1 flagship report's "Open Question Two": **the twelve layers of sharing are in principle extensible, but the depth of extension differs enormously by layer — the intelligence layer has the highest extensibility, the organizational layer depends on corporate governance, and some items in the physical layer are strongly morphology-dependent.** This is D3's concrete extension of the D1 framework, and also the methodological presupposition of the archi-intelligence Research Series on the AR4 → AR5 path.

Chapter 6 Training Infrastructure: The Evolution of Dojo → Cortex → Dojo 3

In the AR4 criterion of a "unified closed-loop of training, simulation, and deployment," the training infrastructure is the engine of the closed-loop. The evolution of Tesla's training infrastructure, replete with strategic reversals and pragmatic adjustments, is an excellent case for understanding the "in-house vs procure" trade-off of a vertical-closed-loop enterprise.

6.1 The Rise and Fall of Dojo (2019-2025)

Dojo is the in-house training-supercomputer project that Tesla made public from 2019, once positioned by Musk as the cornerstone for realizing the commercialization of FSD and Optimus. At its core was the D1 training chip (2021, 7nm, 362 TFLOPS BF16/CFP8), aimed at building a fully in-house training infrastructure optimized for video training, freeing itself from dependence on NVIDIA GPUs.

Yet in August 2025, Musk confirmed the **closure of the Dojo project and the dissolution of the team.** His public reason was: when all technical paths converge on AI6, Dojo 2 had become an "evolutionary dead end." Roughly twenty engineers left to found DensityAI. The deeper logic of this decision was: **rather than maintaining an independent training-chip architecture (Dojo), it is better to let the training and inference chips converge on the same architecture (AI5/AI6).**

A single sentence from Musk captures this convergence precisely: "Dojo 3 may be said to survive in the form of a large number of AI6 SoCs on a single board." In other words, Dojo as an independent training-chip architecture died, but its mission — in-house training compute — continues in the form of "stacking AI6 inference chips into a training cluster."

6.2 Cortex: A Pragmatic Hybrid Approach (2024-)

As Dojo declined, Musk from August 2024 turned to championing **Cortex** — a giant AI training-supercomputer cluster built at Austin/Giga Texas. Unlike Dojo's "fully in-house chip" route, Cortex is a **pragmatic hybrid approach**: it primarily adopts NVIDIA H100/H200 GPUs (a scale of roughly 67,000 to 100,000 H100-equivalents, with figures differing across sources), used for the video training of FSD and Optimus.

The strategic significance of Cortex lies in this: **it acknowledged that, on the training side, the evolution pace of the in-house chip (Dojo) could not keep up with NVIDIA's iteration cadence, and therefore pragmatically procured NVIDIA GPUs to guarantee the immediate availability of training compute.** This is the maturity of a vertical-closed-loop enterprise — vertical integration is not dogma; in the segment where in-house development is not cost-effective (training GPUs), procure decisively.

> **Data annotation**: the Cortex cluster scale ranges across public sources from "roughly 67,000 H100-equivalents" to "roughly 100,000 H100/H200." This report annotates this range of difference and does not forcibly unify it.

6.3 Cortex 2.0 and the Dojo 3 Restart (2026)

In 2026, the training infrastructure entered a new phase.

Cortex 2.0: built at Giga Texas, with a first phase of 250MW launched in April 2026, targeting full capacity of 500MW by mid-2026. As of May 2026, the full 500MW build-out was not yet complete.

Cortex 2.0 is positioned as the engine "determining the pace of FSD improvement, the pace at which Optimus learns new tasks, and confidence in the expansion of the Robotaxi fleet."

The Dojo 3 restart: in January 2026, Dojo 3 development restarted — but its form had changed. The new Dojo 3 is no longer an independent training-chip architecture, but **a training cluster built by stacking Tesla's in-house AI5/AI6 chips,** aimed at reducing dependence on NVIDIA. This corroborates the prophecy that "Dojo survives in the form of AI6 SoCs on a single board."

Tesla's AI chip roadmap now extends to AI5, AI6, AI7, and beyond, adopting a compressed **nine-month design cycle.** The accompanying capacity investment is staggering: 2026 capital expenditure is projected to exceed $20 billion, with a $16.5 billion AI6 chip-manufacturing agreement signed with Samsung.

6.4 The Architectural Implications of the Training-Infrastructure Evolution

The evolution of Tesla's training infrastructure yields three implications.

First, **vertical integration is not dogma but a dynamic trade-off.** The closure of Dojo (the fully in-house training chip) and the rise of Cortex (procuring NVIDIA) prove that even the most resolute vertical-closed-loop enterprise will pragmatically procure in the segment where in-house development is not cost-effective. AR4 does not require "everything in-house," but "in-house at the critical control points."

Second, **the convergence of training and inference chips is an efficiency optimization.** The rebirth of Dojo 3 in the form of AI5/AI6 stacking embodies the strategy of "converging the training chip and the inference chip on the same architecture" — which lowers the cost of maintaining two chip architectures and is an embodiment of the efficiency of the vertical closed-loop.

Third, **the training infrastructure is the true bottleneck of the AR4 closed-loop.** The compute scale of Cortex 2.0 directly determines the pace of FSD improvement and the pace of Optimus learning. This explains why Tesla is willing to invest capital expenditure on the $20 billion scale — in the AR4 architecture, training compute is the true bottleneck of the "data → capability" transformation, and is also the critical control point that a vertical-closed-loop enterprise must control itself (corresponding to the flagship report's C6 simulation-and-compute control point).

6.5 The Group-Level Compute Restructuring After the SpaceX-xAI Merger (2026.2-2026.5)

On 2 February 2026, SpaceX completed its acquisition of xAI (the "xAI Merger," combined valuation $1.25T; see the SpaceX S-1 prospectus, 2026.5.20). The impact of this merger on D3's analysis of training infrastructure is structural — **the COLOSSUS (Memphis TN) and COLOSSUS II (Memphis TN + Southaven MS) training clusters originally belonging to xAI now belong to SpaceX corporate:**

- **The current total compute of COLOSSUS / COLOSSUS II is ~1.0 GW** (SpaceX S-1 Q1 2026 data: AI Segment nameplate compute draw 1 GW).
- COLOSSUS deployment speed of 122 days (industry benchmark: a 100MW greenfield data center takes roughly 2 years).
- COLOSSUS II deployment speed of 91 days, faster still.
- The Grok 5 model is currently trained on COLOSSUS II.
- In May 2026 SpaceX signed a Cloud Services Agreement with Anthropic: $1.25B/month, through May 2029, using part of the capacity of COLOSSUS / COLOSSUS II.

The methodological significance of this merger for D3's argument: Tesla's training infrastructure is now incorporated into a larger group-level compute system — Tesla Cortex 2.0 (Giga Texas, 100K H100/H200, 250-500MW) serves FSD + Optimus; SpaceX COLOSSUS / COLOSSUS II (Memphis + Southaven, ~1.0 GW) serves Grok 5 training + Anthropic customers. **These two sets of infrastructure are controlled by one and the same person, Musk, but belong to two separate corporate legal entities.** The Terafab framework (SpaceX + Tesla + Intel) exists precisely to connect the hardware supply of these two sets of infrastructure.

This architecture carries a methodological implication not fully discussed in D1: **when the AR4 training infrastructure spans multiple legal entities yet is coordinated through a single CEO, "group-level AI compute sovereignty" becomes a higher-dimensional capability than "single-company compute sovereignty."** If other fast-followers lack this cross-entity group-level coordination (for example, Huawei, although strong in ICT, lacks a group-level compute system spanning "automobile + space"), the attainable ceiling of their training infrastructure will be structurally constrained. This is D3's concrete extension of the D1 framework.

Chapter 7 Reuse at the Organizational Level: The Architectural Implications of the FSD-Optimus Team Merger

Chapter 4 has already mentioned the merger of the FSD and Optimus teams. This chapter delves into the architectural implications of this organizational fact — because **in the AR4 architecture, the organizational structure is itself part of the architecture.**

7.1 The Fact of the Team Merger

In June 2025, leadership of the Optimus project was transferred from Milan Kovac to Ashok Elluswamy. Ashok Elluswamy concurrently serves as Vice President of FSD/Autopilot. This arrangement makes FSD and Optimus driven by **the same technical leadership and the same core team.**

As argued in Section 4.5 of Chapter 4, this organizational integration is the **implicit twelfth layer** that allows Tesla's eleven layers of shared core technology to hold — the technical-level "homologousness" must be transformed from design intent into engineering reality through an organizational-level "merger." This chapter further develops the architectural implications of this organizational fact.

7.2 The Reverse Application of Conway's Law

The field of software architecture has a famous "Conway's law": the system architecture will reflect the organizational structure that designs it. Tesla's FSD-Optimus team merger is a **reverse application** of Conway's law — to realize a technically cross-morphology reuse architecture, Tesla proactively adjusted its organizational structure so that one and the same team is responsible for two morphologies simultaneously.

The profundity of this reverse application lies in this: **if FSD and Optimus were developed by two independent teams, then even if technically reusable, they would degenerate, owing to organizational boundaries, into "two sets of similar but non-shared code."** The prerequisite for technical reuse is organizational reuse. Through the team merger, Tesla guarantees from the organizational level that "homologousness" is not merely a technical slogan but a genuine engineering synergy.

7.3 The Lesson for Companies "Making Both Cars and Robots"

Many companies attempt to enter both the automotive and robotics domains simultaneously (XPeng's IRON, Huawei's Kuafu, Xiaomi's CyberOne, and so on). Tesla's organizational experience offers a key lesson: **the cross-morphology reuse of the technical architecture must be supported by the cross-morphology integration of the organizational architecture.**

If the automotive team and the robotics team each go their own way, then even if the two technology stacks are similar, genuine reuse cannot be realized — problems such as reinventing the wheel, inconsistent interfaces, and a non-shared world model will arise. Tesla's team merger is the organizational guarantee that allows its cross-morphology reuse to be "for real." This is a hidden requirement that other players readily overlook when pursuing the "automobile + robot" dual track.

7.4 The Cost of Organizational Reuse

Organizational reuse is not without cost. The merger of the FSD and Optimus teams means that core talent is allocated between two morphologies, the progress of either morphology is constrained by the

bandwidth of the shared team, and the attention of the technical leadership is split between two morphologies. This is a "high-coupling" organization — the benefit is deep synergy and reuse, the cost is the concentration of single-point (core-team/leadership) risk. This is consistent with the overall characteristic of "full-stack coupling" in the vertical-closed-loop architecture: high reuse is accompanied by high coupling risk.

Chapter 8 The Cost and Boundaries of AR4: The Lessons for Other Players

The first seven chapters of D3 dissected the construction of Tesla's AR4 vertical closed-loop. As a conclusion to this dissection, this chapter answers two questions: **what is the cost of this architecture? And what are the lessons for players who cannot replicate Tesla?**

8.1 The Fivefold Cost of the AR4 Vertical Closed-Loop

Cost one: an extremely high threshold of capital and talent. To vertically integrate silicon, OS, applications, cloud, and robots simultaneously requires tens of billions of dollars of investment and thousands of top engineers. Tesla committed 2026 capital expenditure of over $25 billion (confirmed in the Q1 2026 earnings report, including Terafab + AI infrastructure + six factories + solar-manufacturing equipment); it signed a $16.5 billion AI6 chip-manufacturing agreement with Samsung; and, more importantly, the hard data disclosed in the SpaceX S-1 prospectus (2026.5.20): **the capital expenditure of the SpaceX AI Segment in the single quarter of 2026 Q1 alone reached $7,723M,** while the combined Space + Connectivity capex over the same period was only $2,384M — the AI capex in a single quarter is more than three times the Space + Connectivity combined. Full-year 2025 AI Segment capex was $12,727M. **This is a scale of investment that only a group with a market capitalization in the trillions can sustain, and even that scale may not be enough — the capital demand implied by the Terafab 1 TW/year compute-hardware target far exceeds this figure.** Should other OEMs wish to replicate Tesla's AR4 path, the gap in capital scale is structural and cannot be bridged within 18 months (see Chapter 9's comparative analysis of XPeng's cash of RMB 47.66B).

Cost two: the fragility of full-stack coupling. The failure of any layer of the vertical closed-loop may have cascading effects on the full stack. The failure of Dojo (though resolved in the form of convergence) and the tension between the HW3 installed fleet and the new software both embody the fragility of "difficulty in localized loss-stopping" under full-stack coupling.

Cost three: the regulatory controversy of the pure-vision approach. Tesla's insistence on the pure-vision (no-LiDAR) occupancy-perception approach continues to face regulatory questioning on its L4 liability declarations. Although this architectural decision is the prerequisite for cross-morphology reuse (a unified perception stack), it is also its greatest risk in the highly mature regulatory markets of Europe and the United States.

Cost four: the fragility of the liability model. The black-box nature of the end-to-end neural network (from V12) makes the attribution of liability and the making of improvements more difficult when the system exhibits a failure mode that is hard to explain. This is an intrinsic challenge of the AR4 end-to-end paradigm and the core of the flagship report's "Open Question One" (the provable safety boundary of end-to-end).

Cost five: CEO personal risk and high organizational coupling. The high-coupling organization of the FSD-Optimus team merger, combined with Tesla's strategic heavy dependence on Musk's personal judgment, constitutes a concentration of single-point risk. This is the inherent fragility of the vertical closed-loop at the organizational level.

8.2 The Three Categories of Lessons for Other Players

As the flagship report argues, the vertical closed-loop cannot be replicated in the short term. But the Tesla case offers differentiated lessons for different types of players.

For resource-rich technology giants (Huawei, Google, Xiaomi): they may draw on the line of thinking of "cross-morphology reuse," but should reuse selectively. Huawei, with the HarmonyOS microkernel + the Pangu large model + the Ascend compute base covering vehicle-robot, is a "cross-device ecosystem" localization adaptation of Tesla's line of thinking. The key is to find one's own "morphology-agnostic representation layer" (Huawei's world model, Google's Gemini Robotics).

For traditional OEMs (VW, Toyota, BMW): they should not blindly pursue full-stack vertical integration (the lesson of CARIAD), but should identify the control points they truly control. The lesson of Tesla is not "you too must make chips and make robots," but "understand which are the architectural control points you must control yourself, and which can be platform collaborations."

For robotics-specialist players (Figure, 1X, Unitree, and the like): the Tesla case reveals that the core moat of the robot lies not in the mechanical structure (the cerebellum) but in world understanding (the cerebrum). If robotics-specialist players lack Tesla-level real-world data, they need to seek an alternative source of world understanding — which is precisely where the value of the NVIDIA Cosmos world model and various foundation models lies for robotics startups.

8.3 The Boundary of Cross-Morphology Reuse: An Honest Summary

The core conclusion of D3 is an honest characterization of "cross-morphology reuse" — based on the complete framework of Chapter 4's Tesla eleven layers of shared core technology + the organizational twelfth layer:

```
================================================================
 The True Picture of Cross-Morphology Reuse
 (based on Tesla's 11 shared layers + the organizational 12th)
================================================================
 Reusable (12 shared layers)          Non-Reusable (motion-control layer)
----------------------------------------------------------------
 Physical layer (6 items):            Kinematic/dynamic models
   Actuators / Power electronics       Bipedal balance and gait control
   Battery / Manufacturing             Fine-manipulation contact dynamics
   Data communication / Audio          High-frequency real-time control (~1kHz)
                                       Actuator-morphology characteristics
 Intelligence layer (5 items):
   Cameras / AI chips / Training
   cluster / Neural simulation /
   Real-world AI

 Organizational layer (implicit 12th):
   FSD-Optimus team merger + homologous KPI

 Prerequisites of reuse:
   • Physical:  heavy-asset factory network + 10-20 yrs auto-business accumulation
   • Intelligence: data closed-loop + in-house chip + 5-10 yrs algorithmic evolution
   • Organizational: CEO cross-morphology decision + reverse-Conway team integration
```

The boundary of reuse:

- 12 layers reusable, but the motion-control layer (~6-8 items) is non-reusable
- The multiplier economics of reuse: the automobile dilutes the robot (Tesla) vs an independent robot line with no amortization (Figure/1X)
- The scarcity of the reusable layers: none of the 12 can be fully replicated

==========

8.4 The Ultimate Significance of Tesla as an AR4 Reference Frame

As the only complete AR4 case at present, the ultimate significance of Tesla's FSD-Optimus unified stack lies not in "becoming a benchmark for others to imitate," but in **providing a precise reference frame that allows the entire industry to calibrate its understanding of a "cross-morphology physical-AI platform":**

- It proves that cross-morphology reuse is engineering-feasible (eleven layers of sharing + organizational integration), while delineating its boundary (the motion-control layer is non-reusable);
- It displays the cost of AR4 (capital, coupling, regulation, liability, personal risk), allowing other players to assess rationally whether, and to what extent, they should pursue AR4;
- It reveals the core control points of AR4 (the in-house inference chip C2, the data closed-loop C5, training compute C6, world understanding), providing other players with coordinates for identifying "their own controllable control points."

As the flagship report states: to understand Tesla is not to become Tesla, but to make a clear-eyed judgment about architectural control points under the constraints of one's own resource endowment and organizational capability. This is precisely the ultimate value of archi-intelligence's study of the Tesla AR4 case — not veneration, but calibration.

Chapter 9 The Mirror Chapter: The Argumentative Significance of D3 for the Other OEMs of D2

The methodological intent of D3's selection of Tesla as the AR4 case study has never been to have other OEMs "learn how to become Tesla" — the flagship report (D1) has argued that the scale of capital, density of talent, and organizational flatness that the vertical closed-loop demands are the product of twenty years of accumulation and cannot be replicated within eighteen months. The true value of the Tesla case is to **provide the fast-follower with an AR4 reference frame** — through the clear characterization of Tesla's engineering granularity, to allow the other OEMs in D2 to answer precisely: "**when we cannot possibly become Tesla, where is our true position? And what is our reasonable path?**"

This chapter selects five representative OEMs from D2 — each representing a different "AR4-aspiring trajectory" — for mirroring analysis. These five cover all the principal regions and path forms of the D2 assessment: European traditional transformation (Volkswagen), the Japanese gradualist route (Toyota), ICT cross-domain entry (Huawei HIMA), the Chinese new entrant (XPeng), and consumer-electronics cross-domain entry (Xiaomi).

At the end of each mirroring analysis, on the basis of the Chapter 4 framework of "Tesla's eleven layers of shared technology + the organizational twelfth layer," a judgment is given of that OEM's reusable mapping under the twelve-layer framework. Section 9.6 presents the complete 5-OEM × 12-layer reusability matrix, reducing the abstract "AR4 rating" to a concrete, comparable engineering granularity.

9.1 Volkswagen CARIAD: The Dual-Track Tension of Global Headquarters vs China Speed

D2 assigns Volkswagen a Snapshot of AR2.0 and a Roadmap of AR2.5. The key facts in mirroring Tesla:

CARIAD's global headquarters undertakes the unified architecture definition of the SSP (Scalable Systems Platform), targeting deployment around 2030 — a globally unified software platform that Volkswagen hopes will become "its own equivalent of FSD." However, **CEA (China Electronic Architecture; the VW Group headquarters' first press release in 2024 used "China Electrical Architecture," and from 2025 onward "China Electronic Architecture" became the official mainstream formulation) 1.0 reached SOP in China at the end of 2025, with the first model, the VW ID. UNYX 08, beginning volume production in January 2026, and a reduction of roughly 30% in the number of ECUs.** CEA is jointly developed by three parties — Volkswagen Group China Technology Company (VCTC), CARIAD China (1000+ employees), and XPeng — which means that Volkswagen has in effect accepted a dual-track strategy of "building, alongside the global SSP, another parallel architecture serving only the Chinese market." CARIAD's in-house China-local ADAS/AD SoC was announced in November 2025, and CARIZON (a 60:40 joint venture between CARIAD and Horizon Robotics) undertakes local ADAS development, with ADAS L2++ urban NoA to launch within 2026.

By the engineering granularity of comparison with Tesla: Volkswagen has chosen a path other than Tesla's — **a dual track of leveraging external capability + localized in-house development.** The leveraging is embodied in the XPeng XNGP license (from 2026, five models use CEA + XPeng E/E + the ADAS stack); the localized in-house development is embodied in CEA 1.0 and CARIAD China's in-house SoC. **The methodological significance of this path lies in this: when the global headquarters cannot complete a Tesla-style full-stack vertical closed-loop at China speed,**

"rapid leveraging + local in-house development" is an engineering-consistent compromise.

But the cost of this path is the "absence of full-stack synergy advantage" that D1 has argued: CEA and SSP are two parallel architectures that very likely cannot ultimately be merged; the capability of the XPeng XNGP license exists, but the IP sovereignty is not in Volkswagen's hands; and the next-generation evolution of CARIAD China's in-house SoC must bear independently the cost of chip design and tape-out — costs that within the Tesla group are amortized jointly by the automobile + Optimus + xAI, but whose downstream user for Volkswagen's "China-local in-house development" is only the Chinese market. Volkswagen's Roadmap score of AR2.5 **precisely reflects this structural boundary of "rapidly reaching AR2.5 but struggling to rise to AR3"** — CEA 1.0 solved the problem of "usability in the Chinese market," but did not solve the problem of "the data closed-loop and training infrastructure required by the Software 2.0 paradigm."

Volkswagen under the twelve-layer framework: Volkswagen has relatively strong accumulation in the physical layer (vehicle manufacturing, battery, power electronics, data communication, audio), but because it **has no humanoid-robot business at all,** the six physical-layer items are **reused within the automotive business** rather than across morphologies — losing the twelve-layer multiplier effect of Tesla's "automobile diluting the robot." On the intelligence layer, AI chips have the in-house CARIZON path (local), the training-cluster scale lags markedly, and the end-to-end paradigm and data closed-loop are weak. On the organizational layer, the split structure of CARIAD headquarters vs CARIAD China is the opposite of Tesla's organizational integration of "a single FSD-Optimus team." Overall judgment: Volkswagen can replicate 4-5 of the eleven layers (and all are physical-layer items within the automotive business), with the cross-morphology multiplier effect essentially 0.

9.2 Toyota Arene: Gradualist + fail-operational, the Paradigm Opposition with Tesla

D2 assigns Toyota a Snapshot of AR2.0 and a Roadmap of AR2.5. The key facts in mirroring Tesla:

Woven by Toyota's Arene software platform made its global debut on 21 May 2025, fitted to the 2026 RAV4 — Toyota's first-ever SDV (Software-Defined Vehicle). Arene comprises a three-piece set: Arene SDK (the development toolkit) + Arene Tools (the virtual environment and test workflow) + Arene Data (the data infrastructure). The Lexus ES follows within 2026, and the next-generation EV will use Arene comprehensively. Panasonic Automotive Systems' new IVI system has been integrated with Arene.

By the engineering granularity of comparison with Tesla: Toyota has chosen a path of **paradigm opposition** to Tesla. Tesla is **"all-in end-to-end + single-point breakthrough" + fail-soft** (soft failure — before L4, the human serves as the ultimate safety redundancy); Toyota is **"gradualist + multi-step validation" + fail-operational** (fault-tolerant operation — after the failure of any subsystem, the system itself can still complete the task in a degraded-safe manner, without human intervention). The design philosophy of Arene is explicit — it "applies Toyota's manufacturing experience to modern software engineering," "separating software and hardware" to realize a "zero-accident future." **This philosophy and Tesla's paradigm revolution of "Photon In, Control Out" end-to-end replacing 300,000 lines of C++ code lie at the two ends of the spectrum.**

The methodological significance of this path lies in this: gradualist + fail-operational is the "failure-philosophy anchor of the automobile and the robot" that D1 has argued — a life-critical system does not permit fail-soft. Toyota, as an automaker at the global scale of ten million units per year, has a regulatory exposure surface and brand liability far higher than Tesla's, and **it must choose gradualism.**

The combination of Arene's Automotive Grade Linux base (open source) + Panasonic IVI integration (Tier-1 synergy) + the Woven City test ground (a controlled environment) embodies the engineering priority of "maintaining fail-operational at production scale."

But the cost of this path is the absence of "the data-training-deployment closed-loop required by the Software 2.0 paradigm" that D1 has argued: Toyota has no fleet at Tesla's data scale, no training compute at Cortex's level, and no end-to-end-neural-network deployment experience at FSD V14's level. Arene is an excellent "SDV infrastructure," but it is not "an AR4 cross-morphology-reuse algorithm platform" — Toyota's humanoid-robot projects (including research such as Punyo) have not achieved deep reuse with Arene at the algorithm layer. **The Roadmap score of AR2.5 reflects this: Arene solved the problem of "an SDV starting point," but the synergy of the three — fail-operational + the end-to-end paradigm + cross-morphology reuse — required above AR3 cannot be achieved simultaneously within eighteen months by Toyota's current path.**

Toyota under the twelve-layer framework: Toyota's physical-layer accumulation is the strongest in the world (the ten-million-unit-per-year scale, the Toyota Production System TPS, the full stack of battery/motor/power electronics), but because **its humanoid-robot projects (such as Punyo) have not achieved deep reuse with Arene at the algorithm layer,** the cross-morphology amortization space of the six physical-layer items has not been opened. On the intelligence layer, Arene provides SDV infrastructure but the end-to-end paradigm and data closed-loop are markedly weak, and the training-cluster scale lags far behind Tesla's Cortex 2.0. On the organizational layer, Toyota's business-unit boundaries are clear (vehicle / robot / industrial equipment), the opposite of the cross-morphology merged organization of the "reverse application of Conway's law." Overall judgment: under the twelve-layer framework, Toyota has considerable physical-layer reuse potential but it is not activated, has accumulation in the intelligence layer within the Arene scope but weak cross-morphology extension, and an organizational layer incompatible with Tesla's path.

9.3 Huawei HIMA: The Second AR4 Path (ICT Entering the Automobile vs the Automaker Moving Toward ICT)

D2 assigns Huawei HIMA a Snapshot of AR4.0 and a Roadmap of AR4+. This is **the only object in D2 with an AR4 rating equal to Tesla's,** but the comparative value of its mirroring of Tesla lies precisely in "two utterly different paths to attaining AR4."

The key facts:

- **HIMA's five principal brands**: Aito (Seres) / Luxeed (Chery) / Stelato (BAIC) / Maextro (JAC) / Shangjie (SAIC)
- **The three new Jiang-series brands** (added 2025-2026): Aistaland / Yijing (Dongfeng) / Huajing (SGMW) + Qijing (GAC, 2025.9.19) — **for a total of 8-9 partner brands**
- **Cumulative assisted-driving mileage of 8.76 billion km** (2026.4), approaching the 10B threshold (close to Tesla's contemporaneous 10B miles, but with a different mileage structure)
- **Cumulative installed base of 1.4M vehicles** (2025.12), with a 2026 target of 3M + 80+ models
- **Qiankun ADS 5.0** has been released, first fitted to the Aito M9 (2026.5.28), with 6 LiDAR + 5 mmWave + 12 ultrasonic + 12 cameras
- A new-generation LiDAR (dual-optical-path, 896-line) — the world's highest-specification in volume production

By the engineering granularity of comparison with Tesla: Huawei is the **second path** to

AR4 — **ICT entering the automobile** (Tesla being the automaker moving toward ICT). The core difference between the two paths:

Dimension	Tesla Path	Huawei Path
Starting point	Whole-vehicle manufacturer → moving toward ICT	ICT giant → entering the automobile
Business model	Single-brand full-stack vertical closed-loop	Tier-1 horizontal licensing to 8 automaker brands
Data closed-loop	A single fleet, all data belonging to Tesla	8-brand fleets, data belonging to the automakers but the models belonging to Huawei
Cross-morphology ambition	The Optimus humanoid robot, deeply reused with FSD	Does not enter whole-vehicle manufacturing or robotics, focusing on Tier-1
Sensor philosophy	Pure vision (no LiDAR)	Vision + LiDAR multi-fusion (896-line LiDAR)
Regulatory positioning	U.S. FMVSS + supplementing the Chinese market	The Chinese market primarily, with cautious probing in Europe

The methodological significance of this path lies in this: Huawei proves that AR4 does not require the "single-brand full-stack" organizational form. The **"Tier-1-Led Alliance" model of horizontal licensing to 8 automaker brands** can, without entering vehicle manufacturing, achieve an engineering granularity approaching Tesla's — the cumulative mileage of 8.76B km is the empirical evidence of the data scale of this model.

But the cost of this path is the absence of the "cross-morphology-reuse boundary" that D1 has argued: Huawei has no Optimus equivalent, and its AR4 score is based on the whole-vehicle-level reuse of "automotive ADS 4.0/5.0 + the HarmonyOS cockpit + the three-electric system," not extended to embodied robots. While Tesla deduces the AR4 "multi-body physical-AI platform" criterion to its extreme (the three-morphology sharing of vehicle + Optimus + the xAI cluster), the Huawei path is bounded on this criterion — not a defect, but an active choice of business model (Huawei has explicitly stated it "does not enter vehicle manufacturing").

The methodological lesson for other OEMs: if an OEM has neither Tesla's twenty years of accumulation nor Huawei's ICT full-stack capability, then **the most realistic path is to become a member of the Huawei HIMA alliance** (as SAIC is to Shangjie), rather than building a Tesla equivalent itself. The choices of BAIC (Stelato), JAC (Maextro), and Chery (Luxeed) in D2 are already voting with their feet to validate this judgment.

Huawei under the twelve-layer framework: Huawei is the most peculiar case under the twelve-layer framework — its reuse form is "the layers provided to the licensed automakers" vs "the layers retained independently." The layers provided to the automakers include, among the five intelligence-layer items, AI chips (Ascend) + training cluster + Real-world AI + Neural simulation (4 items strong) + data communication (HarmonyOS interconnect), and so on. The layers retained independently and not entered include the physical layer's Manufacturing (does not make whole vehicles; the factories are in the hands of the partner automakers) + Actuators / Battery (does not directly do whole-vehicle elec-

tromechanics). **The most critical absence is the role of the Actuators item among Tesla's eleven layers as an "automobile-robot bridge" — Huawei has no robotics business, and the cross-morphology amortization space of 5 of the 6 physical-layer items is 0.** On the organizational layer, the multi-BU coordination of Huawei's Intelligent Automotive BU + Terminal BG + Intelligent Automotive Solution BU replaces Tesla's single-CEO cross-morphology decision-making. Overall judgment: among the twelve layers, Huawei has 4-5 strong intelligence-layer items, has actively forgone cross-morphology amortization in the physical layer owing to "not entering," and has an organizational layer that is integration of another form.

9.4 XPeng: The Chinese New Entrant Most Like Tesla, 5-10 Years Behind

D2 assigns XPeng a Snapshot of AR3.0 and a Roadmap of AR4. This is **the object closest to the Tesla path** in D2, and its mirroring value is therefore the highest.

The key facts (sources: XPeng SEC Form 6-K + Q1 2026 Earnings, 2026.5.28):

- **The Turing chip in volume production,** with a 2026 target of 1 million units, **and Volkswagen the first external commercial customer** — the first time a Chinese OEM's in-house AD/ADAS SoC has become an external-procurement object of a Western automaker
- **VLA 2.0** (Vision-Language-Action 2.0) released on 2026.3.2, with **ADAS mileage penetration surpassing 50% in 2026.4** (Q1 2026 earnings report)
- The **VLA "physical Turing test" passed in 2026.3** — "passengers can barely tell whether it is the AI or a human driving"
- The **GX flagship** (SOP 2026.5.20): **dual Turing SoC + fully redundant L4 hardware** — China's first "L4-hardware-preinstalled" volume vehicle
- **Robotaxi trial operation in Q3 2026** (Guangzhou), with full-autonomous-driving commercialization in 2027
- The **IRON humanoid robot** — in-house, with volume production before the end of 2026
- Q1 2026 cash of RMB 47.66B (roughly $6.6B), a 2025 Q4 gross margin of 21.3%, and net profit of RMB 0.38B (its first positive net profit)

By the engineering granularity of comparison with Tesla: the correspondence between XPeng's path and Tesla's exhibits a high degree of "pattern isomorphism":

Tesla Layer	XPeng Counterpart	Time Gap
HW3/HW4/AI4 in-house chip	Turing chip in volume production	Tesla HW3 2019 → XPeng Turing 2026, **about 7 years**
FSD V12 end-to-end	VLA 2.0 end-to-end + VLA	Tesla V12 2024.1 → XPeng VLA 2.0 2026.3, **about 2 years**
Optimus humanoid robot	IRON humanoid robot	Tesla Optimus Gen 1 2022 → XPeng IRON 2026, **about 4 years**
Robotaxi (Austin launch 2025.6)	Robotaxi (Guangzhou trial Q3 2026)	**about 1.25 years**
Cortex 2.0 (100K H100 + 250-500MW)	No training cluster of equivalent scale disclosed	A marked gap

The methodological significance of this path lies in this: XPeng is **the only manufacturer among the Chinese new entrants walking the "complete Tesla path"** — in-house AD/ADAS chip, end-to-end VLA, L4-hardware preinstallation, Robotaxi, and humanoid robot. **It is 5-10 years behind Tesla, but the path is isomorphic.** The "AR4 reference frame" Tesla provides is of the highest methodological value to XPeng — it tells XPeng: - The generation of chips after your Turing chip will need to confront the "tension between hardware generations and software evolution" (the equivalent of the HW3 unsupervised-withdrawal event); - The next-generation neural network after your VLA 2.0 will, in parameter scale, exceed the Turing compute envelope and will require a more aggressive generational leap in compute; - Your IRON robot will require an organizational restructuring of an FSD-IRON team merger (the Ashok Elluswamy equivalent); - Your training infrastructure will need to move from "renting" toward "building a 100K-GPU-class cluster in-house."

But the true risk of XPeng's path is the "organizational debt" that D1 has argued: XPeng's current Q1 cash of RMB 47.66B (roughly $6.6B) is less than 90% of SpaceX's single-quarter AI capex ($7.7B). **The gap in capital scale determines that XPeng can walk the Tesla path but cannot reach Tesla's deployment depth** — unless it scales further through IPO/equity financing, or achieves a "reverse capital inflow" through licensing cooperation with Western automakers such as Volkswagen (the Turing chip). **The Roadmap score of AR4 means for XPeng: pattern isomorphism + a time gap; reaching AR4 by 2027 is possible, but whether the capital scale required to sustain AR4 can be sustained is an open question.**

XPeng under the twelve-layer framework: XPeng is **the only Chinese manufacturer that simultaneously enters vehicle manufacturing + the IRON humanoid robot + in-house chips + end-to-end** under the twelve-layer framework, and could in theory replicate most of the eleven layers. The six physical-layer items: Manufacturing (the Zhaoqing + Wuhan factories) + Battery (primarily procured, with an in-house 800V platform) + Power electronics (XPeng in-house) + Data communication + Audio + Actuators (IRON in-house) — **all 6 items are touched, but the scale of each is far smaller than Tesla's** (the battery scale differs by more than 30×, the factory capacity by more than 10×, and the accumulation of a 4680 equivalent by 5-10 years). The five intelligence-layer items: Cameras + AI chips (Turing) + Training cluster (scale not reaching Tesla's) + Neural simulation + Real-world AI (VLA 2.0) — all 5 items are touched but the Training cluster is the key shortfall. The organizational layer: the equivalent of the FSD-Optimus team merger — the degree of integration between XPeng's IRON team and its intelligent-driving team is not publicly disclosed and is a key node worth tracking. Overall judgment: XPeng has the most complete number of reusable items among the twelve layers (approaching Tesla), but with a 5-10-year generational gap in the scale/depth of each item.

9.5 Xiaomi: Consumer-Electronics Supply-Chain Efficiency + the Structural Shortfall of Cross-Domain Entry

D2 assigns Xiaomi a Snapshot of AR3.0 and a Roadmap of AR3+. The key facts in mirroring Tesla (sources: Xiaomi HKEX annual report + Q1 2026 earnings report, 2026.5.26):

- **Full-year 2025 deliveries of 411,082 units** (+200.4% YoY) — from 0 to 400,000 units in roughly 18 months
- **The 2025 EV business's first annual operating profit of RMB 0.9B** — its first single-quarter profit in Q3 2025, with a gross margin of 24.3%
- **The new-generation SU7 released on 2026.3.19 → 15,000 locked orders in 34 minutes /**

30,000 locked orders in 3 days

- The **YU7 GT set an SUV lap record at the Nürburgring Nordschleife of 7:22.755**
- A 2026 delivery target of 550K units
- Cumulative R&D over 5 years of RMB 105.5B (+37.8%)
- Q1 2026 EV operating loss of RMB 3.1B (the impact of the Spring Festival + the SU7 generational changeover)

By the engineering granularity of comparison with Tesla: Xiaomi represents an **entry path entirely different from Tesla's — consumer-electronics supply-chain efficiency + brand cross-domain entry + speed priority.** This path performs extremely strongly in the front-end market (sales volume, brand momentum, pricing power), but has a **structural shortfall** in the engineering granularity critical to AR4:

Tesla Engineering Element	Xiaomi Counterpart
In-house AD/ADAS SoC (HW3→AI5)	**Buys chips** (NVIDIA Orin / Thor)
Cortex 2.0 in-house 100K-GPU training cluster	**Rents compute** (Tencent Cloud / ByteDance Cloud, etc.)
FSD V12-V14 end-to-end neural network	Xiaomi in-house HAD (Xiaomi Autonomous Driving), technically catching up
Data closed-loop (millions of vehicles over 5 years)	600,000 vehicles accumulated over 2 years, a marked shortfall in data scale
Optimus humanoid robot + cross-morphology reuse	The CyberOne robot project, not deeply reused with the automobile

The methodological significance of this path lies in this: Xiaomi proves that "consumer-electronics-style supply-chain efficiency + brand momentum" can, within 2 years, complete the market establishment of "from 0 to 400,000 units + first profit." This is another legitimate path beyond the Tesla path, and its speed even exceeds Tesla's early days.

But the mirroring significance of D3 lies precisely in characterizing its structural shortfall: while the fact that Tesla gives AI5 to Optimus + the xAI cluster (rather than to vehicles) reveals that under the AR4 path "the chip is the downstream of a group-level AI strategy," Xiaomi's "buy chips + rent compute" path means that its AI-compute sovereignty is not in its own hands. This shortfall does not become manifest below AR3 (NVIDIA Orin is sufficient to support current ADAS), but in the leap from AR3+ to AR4 it will become the key bottleneck — **the critical path to catching up with Tesla's AR4 is "group-level AI-compute sovereignty," which Xiaomi's current business model cannot provide.** Xiaomi's Roadmap score of AR3+ accurately reflects this boundary: speed can let it reach AR3+, but the full-stack vertical closed-loop (including compute sovereignty) required by AR4 exceeds the capability envelope of its business model.

Xiaomi under the twelve-layer framework: Xiaomi is the fastest but, under the twelve-layer framework, the most dependent on external supply for reuse. The six physical-layer items: Manufacturing (a self-built Beijing factory, but at a scale far smaller than Giga's) + Battery (supplied by CATL/BYD, not in-house) + Power electronics (partly in-house) + Data communication (the Surge OS cross-device interconnect, a strength) + Audio + Actuators (CyberOne in-house but not deeply reused with the

automobile) — 3-4 of the 6 items have reuse to varying degrees, but the cross-morphology amortization space is small. The five intelligence-layer items: Cameras (procured) + AI chips (**bought**, NVIDIA Orin/Thor) + Training cluster (**rented**, Tencent Cloud/ByteDance) + Neural simulation (dependent on external simulation platforms) + Real-world AI (Xiaomi HAD in-house, in progress) — **the two critical items of AI chips + training cluster are not under its own control.** The organizational layer: the multi-business-unit coordination of Xiaomi's phone BU + IoT BU + automotive BU + the robot project, different from Tesla's single-CEO cross-morphology decision-making. Overall judgment: Xiaomi's reuse among the twelve layers depends more on "consumer-electronics supply-chain efficiency" than on "in-house full-stack"; AR3+ is attainable under this model, but the threshold of AR4's "group-level AI-compute sovereignty" cannot be crossed.

9.6 The Synthesis of the Five Mirrors: A Reusability Matrix Based on the Twelve Layers

Integrating the twelve-layer-framework judgments of the five mirrors yields D3's core methodological contribution to the other OEMs of D2 — **Tesla's eleven layers of shared technology + the organizational twelfth layer are a precise characterization of the gap of the five fast-followers in the leap from AR3+ to AR4.**

Table 9.1: The Five-Fast-Follower × Twelve-Layer Reusability Matrix

Layer	Item	VW CARIAD	Toyota Arene	Huawei HIMA	XPeng	Xiaomi
Physical	1. Actuators	○	○	○	▲	▲
	2. Power electronics	◆	◆	○	▲	▲
	3. Battery	◆	◆	○	▲	○
	4. Manufacturing	◆	◆	○	▲	▲
	5. Data communication	◆	◆	◆	◆	◆
	6. Audio system	◆	◆	◆	◆	◆
Intelligence	7. Cameras	◆	◆	◆	◆	◆
	8. AI chips	▲	○	◆	◆	○
	9. Training cluster	▲	▲	◆	▲	○
	10. Neural simulation	▲	▲	◆	▲	▲
	11. Real-world AI	▲	▲	◆	▲	▲
Organizational	12. Cross-morphology merger	○	○	○	?	○
Cross-morphology realization	Auto→robot amortization	○	○	○	▲	▲

Matrix symbols: ◆ fully realized (on par with or comparable to Tesla); ▲ partly realized (accumulation exists but scale/depth insufficient); ○ unrealized or entirely absent; ? insufficient information

/ not publicly disclosed.

Several core observations of the matrix:

First, **the coverage density of the physical layer (6 items) is highly correlated with "whether one enters vehicle manufacturing."** Volkswagen/Toyota, as traditional automakers, have a substantial reuse foundation in 5-6 physical-layer items (except Actuators, owing to the absence of a robot business). Huawei does not enter whole-vehicle manufacturing, and 4-6 physical-layer items are almost entirely absent. XPeng/Xiaomi enter vehicle manufacturing but with a large gap in the scale of battery/Manufacturing.

Second, **the coverage density of the intelligence layer (5 items) is highly correlated with "an in-house AI stack."** Huawei is the only one complete with ◆ across all 5 intelligence-layer items (Ascend + training cluster + Real-world AI + simulation + cameras). XPeng has 4 items at ◆/▲ (Turing + VLA 2.0 + Camera + a training-cluster-scale shortfall). Xiaomi, owing to "buy chips + rent compute," has the two critical items of AI chips + Training cluster absent.

Third, **the organizational layer (the twelfth) is the "reverse-Conway-law" integration that 4 of the 5 have explicitly failed to achieve.** Only XPeng may be evolving (IRON + the intelligent-driving team), but the degree of integration is not publicly disclosed. This is the "organizational prerequisite of AR4 cross-morphology reuse" that Chapter 7 of D3 has argued — none of the five has achieved the equivalent of Tesla's "FSD-Optimus team merger."

Fourth, **the bottom row, "cross-morphology realization: automobile → robot amortization," is the scarce asset among the five.** Only XPeng (IRON) + Xiaomi (CyberOne) enter robotics; of these, Xiaomi's robot is not deeply reused with the automobile. **The multiplier effect of Tesla's twelve layers of sharing + cross-morphology amortization appears complete in none of the twenty-one other OEMs of D2.**

The core methodological conclusions for the fast-follower:

First, **the leverage + localized-in-house path** (VW CEA + XPeng licensing) can, under the twelve-layer framework, cover 5-6 physical-layer items + 1-2 intelligence-layer items; AR2.5 is attainable, AR3 difficult — owing to the absence of a data closed-loop and training infrastructure.

Second, **the gradualist fail-operational path** (Toyota Arene) covers 6 physical-layer items under the twelve-layer framework but with cross-morphology amortization unactivated, and 1-2 intelligence-layer items partly realized — AR2.5 is attainable, AR3+ requires completing the end-to-end paradigm and cross-morphology reuse capability, and the time window may be insufficient.

Third, **the ICT cross-domain-entry path** (Huawei HIMA) has, under the twelve-layer framework, 5 complete intelligence-layer items + an organizational layer of another integration form, but has actively forgone 4-6 physical-layer items. AR4 is attainable (already rated in D2), but the AR4 "multi-body physical-AI platform" criterion is incompletely satisfied owing to the absence of a robot business.

Fourth, **the new-entrant isomorphic path** (XPeng) covers the most complete number of items under the twelve-layer framework (approaching Tesla's full twelve layers), but with a 5-10-year gap in the scale/depth of each item. The attainability of AR4 is the highest, but the capital scale to sustain AR4 is an open question.

Fifth, **the consumer-electronics cross-domain-entry path** (Xiaomi) has, under the twelve-layer framework, 3-4 physical-layer items reused to varying degrees, with the 2 critical intelligence-layer items dependent on external supply. AR3+ is attainable; the threshold of AR4's "group-level AI-compute sovereignty" cannot be crossed.

In synthesis: no path can possibly become Tesla, but every path can reach a certain AR level under the attainability of its own twelve-layer framework. Understanding the engineering granularity of Tesla's twelve layers allows the fast-follower to judge precisely "where my true position is, and what my reasonable path is" — precisely the value of D3 as the empirical anchor between D1 (the theoretical framework) and D2 (the panoramic assessment of twenty-two OEMs).

Chapter 10 The Empirical Chapter: D3's Support for and Challenge to the Core Propositions of D1

The third aspect of the methodological intent of D3's selection of Tesla as the AR4 case study is the **empirical testing of the core propositions of D1 (the flagship report *The Century-Scale Migration of Architecture*).** This chapter is not a paeon to Tesla, but, by means of Tesla as the engineering entity of the "only complete AR4 closed-loop," subjects the key judgments advanced by D1 to a **clinical examination** — which are supported, which are challenged, and where the falsifiable points of the D1 framework lie. This is an intrinsic requirement of the academic seriousness of the archi-intelligence Research Series.

10.1 The Validation of the Two-Layer Structural Thesis: Infrastructure Convergence vs Control-Semantics Divergence

One of the core propositions of D1: **the evolution of automotive E/E architecture exhibits a two-layer structure** — the lower-layer infrastructure (compute, network, power, thermal management) converges; the upper-layer control semantics and the obligation of proof (functional-safety ASIL, SOTIF, UN-R157) diverge.

The empirical support of the Tesla case for this proposition:

The empirical evidence of infrastructure convergence: the generational evolution of Tesla AI4 → AI5 → AI6, in its lower-layer infrastructure (dual-source foundry TSMC + Samsung, LPDDR5X memory, PCIe interconnect), converges with industry standards — AI5's 192GB LPDDR5X is similar in memory-interface specification to NVIDIA Hopper / Blackwell, and the Terafab framework (SpaceX + Tesla + Intel) also points explicitly toward the industry-general stack of "advanced process + advanced packaging." In other words, **Tesla's silicon infrastructure layer is converging onto the same physical standards as the global advanced-compute infrastructure.**

The empirical evidence of control-semantics divergence: but the FSD software stack and NVIDIA Drive, Huawei Qiankun ADS 5.0, and Mobileye SuperVision **do not converge at all** in upper-layer control semantics. Tesla FSD is "pure vision + end-to-end + fail-soft (before L4, the human serves as the ultimate safety redundancy)"; Huawei ADS 5.0 is "vision + LiDAR multi-fusion + gradualist"; Mobileye is "RSS formal safety proof + explicit liability allocation." **These three control semantics correspond to three entirely different obligations of functional-safety proof, modes of regulatory engagement, and liability-allocation models** — with no convergence trend whatever.

The validation conclusion for D1: the two-layer structural thesis receives strong support from the Tesla case. But D3 also points out a point not fully discussed by D1: **when the infrastructure layer converges to its extreme (as in Terafab's vision of 1 TW/year of compute hardware), the divergence of the control-semantics layer will deepen further** — because the "surplus" of underlying compute will let each upper-layer player more freely choose its own safety philosophy, perception philosophy, and regulatory philosophy. **Convergence and divergence are a coupled evolution, not independent trends.** This observation is D3's methodological contribution to D1.

10.2 "AI Is a Faculty of Language, Not a Brain": A Test on Tesla's End-to-End

The D1 North-Star proposition: **AI (in particular LLMs and VLA-class models) is a faculty of language, not a brain; the deterministic engine is the skeleton, and AI is the language layer**

atop the skeleton. This proposition is the most important judgment of the archi-intelligence Research Series, and also the proposition most readily challenged by Tesla's "all-in end-to-end" narrative.

The test of the Tesla case for this proposition:

The surface challenge: FSD V12 replaces roughly 300,000 lines of C++ control code with a single end-to-end neural network, which appears to prove that "AI can replace the deterministic engine," a counterexample to the D1 proposition.

The deep support: yet in the Q1 2026 earnings call of 2026.4, Musk's key statement reveals the contrary fact — **"the FSD v14.3 software stack is architecturally sufficient to support unsupervised deployment; what remains is validation and regulatory approval, not a question of technical capability."** Over the same period the actual Robotaxi deployment is 3 cities × 2 vehicles per city in operation, with the seven-city commitment revised to "preparations underway." **This gap reveals that, in Tesla's own judgment, the true bottleneck of AR4 implementation is not neural-network capability, but validation, the safety case, and regulatory coordination** — precisely the concrete content of D1's "the deterministic engine is the skeleton."

Evidence of greater structural significance comes from the Risk Factors section of the SpaceX S-1 prospectus (2026.5.20), in the original:

> "The continued improvement of AI model capabilities has historically depended in part on scaling laws, the empirical observation that model performance improves with increased compute, data, and model size, but **there is uncertainty as to how long these scaling relationships will continue to hold**."

This is the public questioning of scaling laws by a company under Musk's control (SpaceX/xAI being the operator of the Grok model) in an IPO prospectus — **the document of the highest legal liability. When Musk's own company expresses uncertainty about the persistence of scaling laws in an SEC filing, D1's judgment that "AI is a faculty of language, not a brain" gains support from within the Tesla/xAI system itself.** This is not a theoretical inference of archi-intelligence, but a Risk Factor written by Musk's own company.

The test conclusion for D1: the D1 North-Star proposition is not only not refuted by the Tesla "all-in end-to-end" counterexample, but is doubly supported by Tesla's own deployment cadence and the SpaceX S-1 legal document. **The true engineering significance of AR4 is the synergy of the AI language layer (the end-to-end neural network) with the deterministic skeleton (validation, regulation, the safety case) — rather than AI replacing the skeleton.**

10.3 The Tesla Empirical Evidence for the Theory of Architecture Debt

D1 advances the concept of **architecture debt**: when the rate of growth of system complexity outstrips the rate of evolution of tools and methodology, the invisible technical burden accumulated by the organization.

The empirical evidence of the Tesla case for this concept:

The HW3 unsupervised-FSD withdrawal event (confirmed in the 2026.4 Q1 earnings report) is a characteristic case of architecture debt. From 2019, Tesla sold HW3 users (with the FSD package priced at $8,000-$15,000) the promise that "the hardware is sufficient to support full self-driving"; in April 2026 this promise was withdrawn. The reason for the withdrawal was not a failure of the HW3

hardware, but that the software complexity (the V14 neural network's parameter scale growing 4.5× vs V12) exceeded the promised boundary of the HW3 compute envelope.

The methodological significance of this event: architecture debt in a vertical-closed-loop enterprise such as Tesla is **borne entirely by itself** — Tesla must bear itself the retrofit cost of the roughly one-million-vehicle HW3 fleet (free computer + camera retrofit + a city-scale microfactory network). Under the traditional "hardware-procurement" model, the Tier-1 and the OEM share this debt; the vertical-closed-loop enterprise bears it alone.

D3's extension of D1: the Tesla case shows that architecture debt under the AR4 path has a dimension not fully discussed by D1 — **the debt of the cross-generational promise.** When the software-generation cadence (V12 → V13 → V14, roughly 18 months) is faster than the hardware-generation cadence (HW3 → HW4, roughly 4 years; HW4 → HW5, roughly 4 years), the promise of "software backward-compatible with old hardware" becomes a major debt. Tesla's V14 Lite project (maintaining a streamlined version for HW3) and the microfactory retrofit network are the concretization of this debt. **This dimension will apply to all OEMs moving toward AR4 — in D2, the generation of chips after XPeng's Turing chip, and the version after Huawei's ADS 5.0, will both face a cross-generational-promise debt of the same structure.**

10.4 Failure Philosophy: Tesla's Positioning Between fail-soft and fail-operational

D1 advances **failure philosophy** as the "fundamental watershed" between the automobile/robot and consumer electronics/cloud — consumer electronics may fail-soft (blue-screen restart); the automobile must be fail-operational (after the failure of any subsystem, the system itself can complete the task in a degraded-safe manner).

The test of the Tesla case: Tesla's positioning on this spectrum is open.

The official statement: Tesla's unsupervised-FSD design goal is fail-operational — FSD v14.3's "intervention-free streak counter" function and the safety-driver-free operation of Robotaxi in Austin/Houston/Dallas both mean that Tesla is moving toward fail-operational.

The empirical reality: but Tesla's concrete technical choices exhibit **partial fail-soft characteristics.** The FSD on HW3 (V11 through V12.6) is explicitly supervised (requiring the human as the ultimate redundancy); the V14 on HW4 exists in a transitional interval between supervised and unsupervised — v14.3 is "architecturally sufficient for unsupervised" but the monitoring and takeover mechanisms are still retained. **Tesla's engineering philosophy appears to be "a gradual shift from fail-soft toward fail-operational as regulation permits"** — in contrast with Toyota Arene (an explicit fail-operational priority).

The validation conclusion for D1: the D1 failure-philosophy proposition receives partial support from the Tesla case, but Tesla reveals a middle ground not fully discussed by D1 — **fail-soft and fail-operational are not a dichotomy but a continuous spectrum, and Tesla has chosen "a gradual move from soft to operational."** The methodological judgment behind this choice is: neither full fail-soft (consumer electronics) nor full fail-operational (Toyota), but a calibrated migration enabled by data accumulation and monitoring — gradually building the engineering foundation of fail-operational through data accumulation and monitoring.

The methodological significance of the existence of this middle ground for the other OEMs of D2 is this: choosing fail-operational does not require reaching it in one step; it can be attained through a migration path of "AR3-phase fail-soft + data accumulation → AR4-phase fail-operational."

But the prerequisite is that the data closed-loop and monitoring infrastructure have already been established — precisely the shortfall of most traditional OEMs in D2.

10.5 The Reverse Definition of the Tool Category: What Methodology and Tools the Fast-Follower Needs

This section is D3's methodological extension of D1 — **a reverse definition, from the Tesla case, of "the methodology and tool category the fast-follower needs for an upward leap."**

The layer-by-layer progression of premises:

The first premise (argued in D1): the Tesla AR4 path requires 20 years of accumulation + full-stack verticality + organizational flatness + a data closed-loop, and **cannot be replicated within 18 months.**

The second premise (assessed in D2): in D2, the 2027 Roadmaps of European, American, and Japanese-Korean traditional OEMs are mostly at AR2.0-AR3.0, yet still commit to evolving toward higher AR. In other words — **the fast-follower must push for AR3+, yet cannot possibly become Tesla.**

The third premise (empirically established in Chapters 1-9 of D3): the engineering granularity of the Tesla AR4 path is extremely fine — the silicon cross-generational-promise debt, the twelve-layer stack of cross-morphology reuse, the 100K-GPU training-infrastructure cluster, the organizational coordination of eight entities. None of these can be replicated completely under the resource-constrained conditions of the fast-follower; the twelve-layer reusability matrix of the five mirrors in Chapter 9 shows further that every fast-follower makes a compromise on Tesla's engineering granularity at certain critical layers.

The definition of the core contradiction —

> The fast-follower must, on the premise of **not having Tesla's resources,** make architectural decisions of **a quality approaching Tesla's.**

This contradiction reveals a market gap worthy of the attention of the entire industry.

The existence of the market gap: the current industrial tool-and-methodology ecosystem already offers rich, mature support for the **detailed-design phase** — EDA tools, PLM platforms, ASIL safety-case toolchains, AUTOSAR-compatible code generators — tools that help the architect "draw clearly, do correctly, and refine" an already-decided architecture. Yet the tool-and-methodology ecosystem still relies principally on individual and organizational engineering intuition for the **conceptual-exploration phase** — the phase in which the architect must make critical architectural decisions before drawing the first circuit diagram, writing the first line of code, or making the first cost estimate — lacking systematic tool-and-methodology support. Through 20 years of organizational-level accumulation, Tesla has internalized an engineering intuition about "which architectural decisions have what effect on which dimensions"; the fast-follower has neither 20 years nor an equivalent full-stack-synergy foundation.

The methodological significance of this market gap: when the fast-follower faces the core contradiction of "making architectural decisions of a quality approaching Tesla's under resource constraints," what they truly lack is not a better "drawing tool" or "data dashboard" — such tools are useful for the detailed-design phase, but do not directly solve the decision-quality problem of the conceptual-exploration phase. What the fast-follower needs is a **tool category, with an accompanying methodology, that bears the decision-support responsibility of the conceptual-exploration phase** — its precise

characteristics, implementation path, and boundary relationship with existing tools being questions that subsequent engineering practice should answer, beyond the coverage of this study.

D3's methodological contribution is to characterize clearly the existence of this market gap — and to advance it as one of the key paths for the fast-follower's leap to AR3+, for joint discussion by industry, academia, and tool vendors.

10.6 The Falsifiable Points of D1: The Failure Modes the Tesla Path May Exhibit in the Future

The academic seriousness of the archi-intelligence Research Series requires us not only to argue the supporting evidence for the D1 framework, but also to list proactively **the concrete scenarios in which the D1 framework might fail on the Tesla case** — that is, "if X occurs in the future, the judgment of D1 will be partially falsified."

Falsifiable point 1: the true end of scaling laws

The SpaceX S-1 has already publicly questioned the persistence of scaling laws. If, between 2027 and 2029, scaling laws are empirically ended in the domain of autonomous driving / embodied robotics (that is, more data/compute no longer brings proportional capability gains), then Tesla's competitive advantage of "unified-stack cross-morphology reuse + the data closed-loop" will be significantly weakened. **The form in which the D1 proposition would be partially falsified**: D1's judgment that "AI is a faculty of language, not a brain" is correct, but the ceiling of the "faculty of language" itself may be lower than D1 expects — whereupon the value of the AR4 closed-loop would be reassessed.

Falsifiable point 2: a catastrophic failure of fail-soft in unsupervised deployment

If, between 2026 and 2028, Tesla's unsupervised Robotaxi suffers one or more **major casualty accidents** in some city, leading the U.S. NHTSA or European EU regulators to comprehensively withdraw the legitimacy of the fail-soft path, then Tesla's gradualist fail-soft → fail-operational migration path will be forcibly interrupted. **The form in which the D1 proposition would be partially falsified**: D1's judgment that "failure philosophy is the watershed" still holds, but the "middle ground is feasible" argued in Section 10.4 of D3 would be disproven — whereupon Toyota Arene's fail-operational-priority path would gain a methodological victory at the regulatory level.

Falsifiable point 3: the critical point of organizational debt

Musk simultaneously controls a board majority across the eight entities of Tesla + SpaceX + xAI + X + Neuralink + Boring + Macrohard + Terafab. If any one of them suffers a major governance crisis (regulatory action, the loss of key talent, a major product failure), it could trigger a cascading effect of organizational debt. **The form in which the D1 proposition would be partially falsified**: D1's judgment that "organizational flatness is the prerequisite for AR4" still holds, but the limit of "single-CEO cross-entity organizational reuse" may be more fragile than D1 implicitly expects — whereupon the AR4 narrative of "organizational-synergy advantage" would be re-examined.

Falsifiable point 4: the feasibility failure of Terafab

The SpaceX S-1 makes clear that Terafab is at present a "general framework," with "specific projects still subject to separate negotiation." If, between 2027 and 2028, Terafab fails to produce actual volume-producible "1 TW/year" compute hardware, or Intel withdraws owing to its own operating problems, then Tesla's group-level AI-compute-sovereignty strategy will suffer a major blow. **The form in which the D1 proposition would be partially falsified**: D1's judgment that "AR4 requires full-stack synergy" still holds, but the feasibility boundary of "building a chip-manufacturing layer in-house" may

be narrower than D1's inference — whereupon the "reference frame" value of the Tesla path for other OEMs would decline.

Falsifiable point 5: a fundamental divergence of regulatory paths

If, between 2026 and 2028, regulatory frameworks such as the EU AI Act, UN-R157, U.S. FMVSS, and China GB 7258-2024 move toward significant divergence (for example, Europe insisting on fail-operational + formal proof, China accepting fail-soft + continuous monitoring, the United States maintaining the current regulatory vacuum), then Tesla's "single software stack + multi-region deployment" model will be broken by regulatory fragmentation. **The form in which the D1 proposition would be partially falsified**: D1's judgment that "AR4 is a global convergence trend" would be disproven — whereupon AR4 might no longer be a single "global ladder," but diverge into three mutually non-interchangeable sub-categories of "U.S. AR4," "China AR4," and "Europe AR4."

The methodological significance of listing these five falsifiable points: the D1 framework must be able to undergo testing in the engineering reality of the next 3-5 years. These five falsifiable points are the necessary concomitant of the academic seriousness of the D1 framework — any theory that cannot be falsified by the future cannot be supported by the future. Subsequent versions of D1 will continue to record the evolution of these falsifiable points, and will openly update the framework when they are empirically confirmed or disproven.

Conclusion: The Methodological Significance of AR4 as a Reference Frame

Through an in-depth dissection of Tesla's FSD-Optimus unified stack, this report has completed the third core task of the archi-intelligence Research Series — **closing the loop between D1's AR0-AR5 theoretical framework and D2's horizontal assessment of twenty-two OEMs, through the clinical examination of a concrete engineering entity.**

The Threefold Value of Tesla as a Reference Frame

The first value is as the empirical anchor validating the D1 theoretical framework. The test of the Tesla case for D1's five core judgments yields clear results: the two-layer structural thesis (infrastructure convergence vs control-semantics divergence) receives strong support, and D3 discovered a coupled-evolution dimension not fully discussed by D1; the judgment that "AI is a faculty of language, not a brain" is doubly supported by Tesla's own deployment cadence and the SpaceX S-1 legal document; the theory of architecture debt acquires a concrete form in the HW3 unsupervised-withdrawal event; the failure-philosophy proposition exhibits a "gradualist fail-soft → fail-operational" middle ground not fully discussed by D1; and the criterion of organizational flatness + cross-morphology reuse receives multiple support from Optimus Gen 3 / Cortex 2.0 / the SpaceX-xAI merger. At the same time, the five falsifiable points listed in Chapter 10 are the necessary concomitant of the academic seriousness of the D1 framework — any theory that cannot be falsified by the future cannot be supported by the future.

The second value is as the AR4 reference frame for the twenty-one other OEMs of D2. The in-depth mirroring in Chapter 9 of the five paths of Volkswagen, Toyota, Huawei, XPeng, and Xiaomi reveals a common conclusion: **no path can possibly become Tesla, but every path can reach a certain AR level under its own constraints.** The leverage + localized-in-house path (Volkswagen) can solve AR2.5 but struggles to rise to AR3; the gradualist fail-operational path (Toyota) maintains a production-scale safety priority but lacks cross-morphology reuse capability; the ICT cross-domain-entry path (Huawei) can attain AR4 through Tier-1 horizontal licensing, but has a cross-morphology-reuse boundary; the new-entrant isomorphic path (XPeng) is the same path but the gap in capital scale determines the deployment ceiling; the consumer-electronics cross-domain-entry path (Xiaomi) is extremely fast but the compute sovereignty required for the leap from AR3+ to AR4 exceeds its business-model envelope. **Understanding the engineering granularity of Tesla allows the fast-follower to judge precisely "where my true position is, and what my reasonable path is"** — this is the value of D3 as the empirical anchor between D1 and D2.

The third value is in identifying the market gap of the fast-follower's methodology and tool ecosystem. Section 10.5 of Chapter 10, through a layer-by-layer derivation of the fast-follower's core contradiction, points out that the current industrial tool ecosystem already offers mature support for the detailed-design phase (EDA, PLM, ASIL safety-case toolchains, and the like), but still relies principally on individual and organizational engineering intuition for the **conceptual-exploration phase** — the phase in which the architect must make critical architectural decisions before drawing the first diagram or writing the first line of code. Through 20 years of accumulation, Tesla has internalized this capability; the fast-follower does not have 20 years. The clear characterization of this market gap is the methodological proposition that D3 advances to industry, academia, and tool vendors — its concrete filling path being left to subsequent engineering practice and industry discussion.

The Boundaries of the AR4 Reference Frame

We are clearly aware of the methodological boundaries of the AR4 reference frame.

First, **the Tesla path is not the only path to AR4.** Huawei HIMA has already empirically established the AR4 path of ICT cross-domain entry + Tier-1 horizontal licensing — although its cross-morphology-reuse boundary differs from Tesla's, its D2 score is the same (AR4). AR4 is a capability threshold, not a path definition. This report's in-depth characterization of the Tesla path **does not constitute a negation of other AR4 paths.**

Second, **AR4 is not the endpoint.** The D1 framework also contains AR5 (a trustworthy general embodied-intelligent agent), whose engineering implications include cross-industry embodied-AI deployment, formal safety proof, cross-regional regulatory coordination, and the like. Tesla is at present still in the ramp phase of AR4, and an empirical case of AR5 does not yet exist. The methodology of this report is not applicable to the analysis of the AR4 → AR5 leap — that is the task of a subsequent Working Paper.

Third, **the Tesla case is in a phase of rapid evolution.** The data of this report are current as of 2026.5.21. Tesla's silicon road, FSD deployment cadence, Optimus volume production, the scale of the Cortex 2.0/COLOSSUS clusters, and the SpaceX-xAI-Tesla group-level coordination are all changing rapidly. **Any methodological judgment based on this report should be continuously updated in conjunction with subsequent events.**

To the Fast-Follower

For the twenty-one OEMs and industry participants who, each at their own position in D2, hope to leap to a higher AR level, the final methodological recommendation of this report is:

Do not attempt to replicate Tesla's AR4 path — 20 years of accumulation, full-stack verticality, organizational flatness, and $25B+ annual AI capex (SpaceX S-1 Q1 2026 data) are inseparable components of the Tesla path. **One should use Tesla as a reference frame, to characterize precisely one's own position and reasonable path.**

Specifically:

- If you are a European traditional-transformation maker (Volkswagen, Mercedes-Benz, BMW, Stellantis, Renault): your reasonable path may be a "leverage + localized-in-house dual track" (the Volkswagen CEA model) or "cooperation with the Chinese ecosystem" (partial attempts by Mercedes-Benz and BMW in the Chinese market), targeting AR2.5-AR3.0. To forcibly pursue Tesla's full-stack-vertical AR4 is infeasible at the level of capital scale and organizational culture.
- If you are a Japanese-Korean gradualist-route maker (Toyota, Honda, Hyundai, Kia): your reasonable path is Arene-class SDV infrastructure + a fail-operational priority + the gradual establishment of a data closed-loop. The target is at AR2.5-AR3.5, with the time window extended to around 2030.
- If you are a Chinese new entrant (NIO, Li Auto, XPeng, Xiaomi): your reasonable path may be an isomorph of the Tesla path but 5-10 years behind (the XPeng case) or consumer-electronics supply chain + brand momentum (the Xiaomi case). Reaching AR4 may occur in 2027-2028, but whether the capital scale required to sustain AR4 can be sustained is a structural open question.
- If you are a Chinese traditional maker (BYD, Geely, Great Wall, SAIC, GAC, Dongfeng, Changan, Chery): your reasonable path may be "becoming a member of the Huawei HIMA / Qiankun alliance" (such as Stelato, Maextro, Luxeed, Shangjie, Qijing, Yijing, Aistaland, Huajing), drawing on Tier-1 horizontal licensing to attain a capability supply of AR3-AR4.

- If you are an American traditional maker (Ford, GM, Stellantis North America): your reasonable path must confront the dual constraint of "having neither Tesla's full-stack capability nor the synergy of the Chinese ecosystem." The most pragmatic judgment at present may be to accept a capability ceiling of AR2-AR3, placing the differentiation focus on brand, customer relationships, and specific market segments.

For all fast-followers, the "conceptual-exploration-phase methodology and tool gap" pointed out in Section 10.5 of Chapter 10 is the most structurally significant methodological prompt of D3 — before drawing the first diagram or writing the first line of code, the quality of the critical architectural decisions determines the attainable ceiling of the fast-follower's upward leap. The concrete path to filling this gap is left to the joint exploration of industry, academia, and tool vendors, but its directional significance is clear: under the reality that Tesla's resources cannot be replicated, the key to the leap to AR3+ lies not in replicating Tesla's "draw clearly" toolchain, but in establishing methodological support for the "think clearly" phase suited to the fast-follower's resource conditions.

To D1's Falsifiability Commitment

The bottom line of academic ethics of the archi-intelligence Research Series is to **acknowledge the falsifiability of the framework.** The five falsifiable points listed in Section 10.6 of Chapter 10 (the end of scaling laws, a catastrophic fail-soft failure, the critical point of organizational debt, the infeasibility of Terafab, a fundamental divergence of regulatory paths) will be continuously recorded for their evolution in subsequent Working Paper versions, and the framework will be openly updated when they are empirically confirmed or disproven. **This is not a weakness of the D1 framework, but the necessary concomitant of its academic seriousness.**

The work of D3 ends here. We await the facts of 2027-2028, to validate or challenge the judgments of this report. The Tesla AR4 case will continue to evolve, the twenty-one fast-followers will successively give their own answers, and the engineering implications of AR4 → AR5 will gradually become clear. The archi-intelligence Research Series will continue, with academic independence as its bottom line, methodological rigor as its standard, and falsifiability as its ethical boundary, to record and to understand this century-scale migration of architecture.

Appendix A: Tesla AR4 Key-Fact Timeline (Tier 1–2 Sources)

This appendix consolidates the Tier 1–2 dated facts dispersed throughout the body, to serve as a quick chronological reference for the Tesla AR4 closed-loop. All entries are annotated with their source tier; the data are current as of 21 May 2026.

Date	Event	Source Tier
2019	HW3 (FSD Computer) enters volume production, ~144 TOPS, 14nm Samsung — first fully in-house inference chip	Tier 2
2021, 2022	AI Day: public disclosure of the occupancy network, Dojo D1 chip, the end-to-end direction	Tier 2
2023	HW4 / AI4 fitted to the Model S/X refresh, ~500 TOPS, Samsung 7nm	Tier 2
2024.1	FSD V12: ~300k lines of C++ control code replaced by a single end-to-end network (“Photon In, Control Out”)	Tier 2
2024.8	Musk turns to championing Cortex (NVIDIA H100/H200 hybrid cluster)	Tier 2
2024.12	FSD V13: HW4-native resolution, training data ~4.2× vs V12	Tier 2
Late 2025	AI4.5 stopgap quietly fitted from the 2026 Model Y (born of the AI5 delay)	Tier 2
2025.6	Robotaxi launches in Austin (first unsupervised commercial-operation city)	Tier 1
2025.6	Optimus project leadership transferred from Milan Kovac to Ashok Elluswamy (de facto FSD-Optimus team merger)	Tier 2
2025.8	Dojo project closed and team dissolved; mission continues via AI6-SoC stacking	Tier 2
2025.11.6	Annual Shareholder Meeting: official disclosure of the eleven layers of shared core technology	Tier 1

Date	Event	Source Tier
2026.1	Dojo 3 development restarts (as an AI5/AI6-stacked training cluster)	Tier 2
2026.2.2	SpaceX completes acquisition of xAI (combined valuation $1.25T); COLOSSUS / COLOSSUS II move to SpaceX	Tier 1
2026.2.17	Cybercab first unit off the line (confirmed to use AI4, not AI5)	Tier 1
2026.3	Terafab framework announced (SpaceX + Tesla; Intel joins in April; ~$25B Austin)	Tier 1
2026.4	Cortex 2.0 first phase 250MW launched at Giga Texas	Tier 1
2026.4.15	AI5 tape-out (GDSII hand-off to TSMC); first batch preferentially to Optimus + supercomputer, not vehicles	Tier 1
2026.4.18	Robotaxi expands to Houston and Dallas (3 cities total, ~2 vehicles each)	Tier 1
2026.4.22	Q1 2026 earnings call: HW3 unsupervised FSD officially abandoned; seven-city commitment revised to “preparations underway”	Tier 1
2026.5.17	FSD V14.3.3 (firmware 2026.14.6.6) pushes to early-access users (intervention-free streak counter)	Tier 2
2026.5.20	SpaceX S-1 prospectus filed (AI Segment Q1 2026 capex $7,723M; scaling-law caveat in Risk Factors)	Tier 1
2026.12 (target)	AI6 tape-out target	Tier 2
2027 mid (target)	AI5 high-volume production in vehicles	Tier 2

Appendix B: About the archi-intelligence Research Series

Mission

archi-intelligence is an independent academic research body dedicated to establishing Architecture Intelligence (AI^2) as a research paradigm. Its mission is to advance the standardization and comparability of cross-industry architectural engineering practices through open methodology, transparent data attribution, and rigorous peer review.

Publication model

- All reports are permanently free and released under CC-BY 4.0.
- Each report is archived on Zenodo with a permanent DOI.
- All methodology, primary sources, and citation lists are 100% public.

The v1.0 release: the "three-deliverable" set

- 2026-01: *The Architectural Migration of the Century* — the flagship report (ontology, cross-industry benchmark, the AR0–AR5 and AI^2-ML frameworks).
- 2026-02: *The State of Global Automotive E/E Architecture Maturity 2026* — an AR & AI^2-ML assessment of 22 OEMs across a dual time dimension (Snapshot + Confirmed Roadmap).
- 2026-03 (this report): *Multi-Embodiment Physical AI Platform Readiness: The Tesla FSD-Optimus Unified Stack* — an in-depth dissection of the only complete AR4 vertically integrated closed-loop case, serving as the empirical anchor closing the loop between D1's framework and D2's assessment.

Editorial independence

The research's methodological choices, case evaluations, and conclusions are made independently by the archi-intelligence Research Team, uninfluenced by any commercial interest, political stance, or geopolitical preference. (See the full Conflict-of-Interest Disclosure in the Front Matter.)

Commitments

- Continuous publication (at least 5 years, 2–3 working papers annually);
- Continuous revision (a revised edition every 6–12 months, all major corrections publicly recorded);
- Continuous openness (all methodology and sources public);
- Continuous independence (refusing funding from any assessed entity).

Contact

- Research: research@archi-intelligence.org
- Corrections: corrections@archi-intelligence.org
- Website: https://archi-intelligence.org

This research was compiled by the archi-intelligence Research Team. archi-intelligence Research Series · Working Paper 2026-03 (English Edition) — End of document —