The Computational Economics of Autonomous Driving: Transformer-Based Approaches and Their Future Viability

v3.0

Current Version: http://eamonmcginn.com.s3-website-ap-southeast-2.amazonaws.com/projects/AI_for_EVs.pdf

Eamon McGinn

May 25, 2025

Abstract

This analysis examines the technical and economic feasibility of achieving global Level 5 autonomous vehicles (AVs) based on current experience with emerging AI models. Current mixed-methods approaches face fundamental scaling limitations that preclude true global Level 5 autonomy, while transformer-based approaches offer superior generalisation but face significant computational barriers. Through detailed computational modelling, we project that transformer-based approaches will become viable for mass-market deployment by 2050 (range of 2042-2058) following a predictable trajectory of efficiency improvements. Our analysis indicates that, while an unoptimised transformer model would require 89 years of computation at a cost of over $3.5 billion in energy alone, emerging efficiency improvements could reduce these requirements by up to 100,000×, bringing total development costs to a present value (PV) of $618 million over the period to 2050. We estimate deployment will require a coordinated research agenda spanning 2025-2050, with critical milestones including $100 million dollars for data acquisition by 2030, ongoing data management costs, $475 million for model training by 2040, and vehicle constraints resolved by 2050.

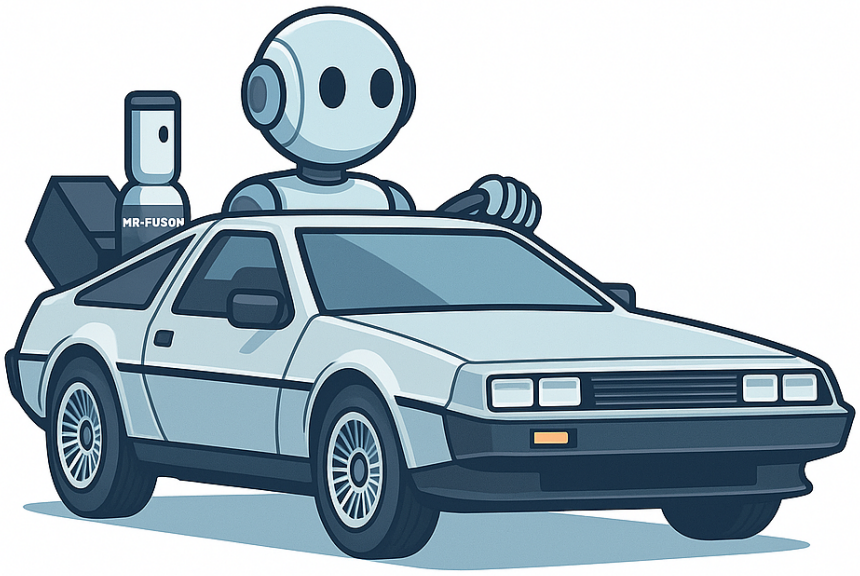
About the Author

Eamon McGinn is an applied economist and former Deloitte partner based in Sydney, Australia. He specialises in analysis of future transport technologies. During his 15-year tenure at Deloitte Access Economics (2010-2025), he led major projects on transport electrification, climate resilience, and technology adoption. Dr. McGinn holds a PhD in Economics from University of Technology Sydney and combines econometric analysis with fast, data-lean simulation tools to translate complex technical questions into clear policy choices. This paper represents independent research. The views expressed are the author's own and do not reflect those of any affiliated organisations.

Given the speculative nature of long-term technological forecasting, we welcome feedback and alternative perspectives on this paper. Readers who identify additional considerations or possess data that could refine our projections are encouraged to contact the author.

Contact: eamon.m@gmail.com

Website: http://eamonmcginn.com/



Contents

1 Executive Summary

Truly autonomous vehicles (Level 5) remain an elusive goal despite decades of research and billions in investment. Recent advances in transformer-based AI models present a compelling new approach, but their computational requirements appear prohibitive at first glance. This comprehensive analysis examines the path to true autonomy through both current approaches and emerging efficiency improvements on transformer-based systems.

Our key findings include:

- Current mixed-method approaches to autonomous driving, while successful in controlled environments, face fundamental scaling limitations that preclude true global Level 5 autonomy.
- The data input required for AVs is expensive but achievable, taking about 3 years and $100 million to gather. The data is around 4 exabytes in size and so has significant ongoing management costs.
- An unoptimised transformer model for autonomous driving would require approximately 3.3×10^{28} FLOPS of computation during training, based on using 11 trillion parameters and 1,500 trillion tokens of training data.
- Such a model would take 89 years to train on 50,000 high-end GPUs and cost over $3.5 billion in energy alone — clearly impractical.
- Emerging efficiency gains could conceivably reduce these requirements by 100,000× by around 2050.
- By 2040, the training problem will have likely diminished to the scale of current GPT4 efforts, making model training feasible for major businesses. Training at this point is estimated to cost around $475 million. However, onboard power requirements still keep deployment out of reach.
- By 2050, the efficiency improvements would also allow for on-vehicle inference using only 3.5 kW of power. This would translate to a reduction in vehicle range by a manageable 28%.
- The overall PV of the investment amounts to $618 million with total expenditure of around $1.57 billion over the period to 2050. This figure is per product and excludes a range of supporting costs such as reserach, regulatory, legal and infrastructure.

Our analysis suggests that, while significant technical hurdles remain, a transformer-based autonomous driving system with targeted efficiency improvements could become economically viable by 2050 (within a range of 2042-2058).

2 The Limitations of Current Autonomous Vehicle Approaches

The autonomous vehicle industry has made remarkable progress using hybrid approaches combining rule-based systems, traditional computer vision, and targeted machine learning models. Companies like Waymo (Markoff 2024) and Tesla (Bretting 2024) have demonstrated Level 4 autonomy in controlled environments, primarily within well-mapped metropolitan areas with extensive infrastructure annotation. However, these approaches face fundamental scaling challenges that prevent them from achieving true Level 5 autonomy.

Current methodologies rely on exhaustive pre-mapping and annotation of driving environments. Each road system requires detailed high-definition mapping, manual annotation of traffic patterns, and programming of specific rules for local conditions (Waymo Team 2020). This approach works effectively in:

- Urban areas with consistent infrastructure (Phoenix, San Francisco)
- Well-maintained highways with clear lane markings
- Environments where edge cases can be identified and managed
- Markets with substantial existing mapping infrastructure

However, the computational and economic requirements for global deployment of true Level 5 autonomy presents insurmountable challenges using current approaches:

1. **Manual Annotation Scaling**: Extending current approaches to encompass global road networks would require annotating at least 40 million kilometers of road worldwide (Central Intelligence Agency 2022). At current productivity rates, this would likely require thousands of engineering teams working for decades.
2. **Maintenance Burden**: Road conditions, signage, and infrastructure evolve continuously. Maintaining current annotation approaches would require permanent large-scale engineering teams in every market.
3. **Edge Case Explosion**: Different regions present unique driving challenges. The combinatorial explosion of edge cases makes rule-based approaches mathematically intractable.

Thinking about the diversity of global driving environments highlights the challenges in scaling the current approach (see Figure 1):

- **Developing Markets**: Rural roads in sub-Saharan Africa, India, and Southeast Asia often lack basic infrastructure like lane markings, traffic signs, or consistent road surfaces
- **Cultural Variations**: Different regions have distinct traffic customs, from the organised chaos of Vietnamese intersections to the informal right-of-way systems in European village centers
- **Weather Extremes**: From snow-covered Nordic roads to sand-swept Middle Eastern highways, environmental variations require extensive specialised programming
- **Infrastructure Inconsistency**: Even within developed nations, infrastructure standards vary dramatically between urban centers and rural areas

Figure 1: Examples of road conditions around the world

3 The Transformer Revolution and Autonomous Driving

The breakthrough success of transformer-based models in natural language processing and computer vision suggests a more scalable approach to autonomous driving may be possible (Vaswani et al. 2017). These models demonstrate remarkable generalisation capabilities, handling tasks and contexts far beyond their explicit training data (Bubeck et al. 2023; Chang et al. 2024).

To establish a baseline for our extrapolations, we start with GPT4. GPT4, released by OpenAI in 2023, represents a reasonable benchmark for large language models. GPT4 built on the success of its predecessors GPT2 and GPT3 to deliver unprecedented capabilities in natural language understanding and generation. Designed as a general-purpose AI assistant, GPT4 demonstrates remarkable versatility across tasks ranging from creative writing to complex reasoning, marking a significant milestone in the development of AI (OpenAI 2023a).

While newer models have emerged since its release, GPT4 remains a widely-used reference point for measuring progress in transformer-based AI systems, with its architecture and training methodologies influencing the broader direction of large language model development.

The computational, financial and economic investment in GPT4 was significant:

- **Parameter count**: Approximately 1.76 trillion parameters (Bastian 2023);
- **Training data**: Approximately 13 trillion tokens (which is maybe 50TB of text data) (Bastian 2023);
- **Training compute**: Estimated at 10^{24} – 10^{25} FLOPs (McGuinness 2023);
- **Training infrastructure**: Approximately 25,000 A100 GPUs for 3-4 months (McGuinness 2023);
- **Energy consumption**: Around 52-62 GWh for the training process (Bowen 2023);
- **Training cost**: Approximately $60-100 million in hardware and energy (Bowen 2023; McGuinness 2023);
- **Model size**: Approximately 500GB for weights during inference
- **Processing speed**: 10-30 tokens per second for text generation

GPT4, while representing one of the most computationally intensive AI systems ever developed, remains primarily constrained to text processing, with only limited multimodal capabilities in its GPT4V variant that can process images but not generate them (OpenAI 2023b). Despite its transformative impact on natural language tasks, GPT4's inability to handle real-time sensory input, temporal reasoning, or direct control of physical systems highlights the enormous leap required to achieve autonomous driving capabilities. This gap between current text-based AI and the multimodal, real-time decision-making requirements of autonomous vehicles underscores the computational and architectural challenges that must be overcome.

3.1 Vision Transformers: Adapting the LLM Architecture

The application of transformer models to visual data represents a fundamental shift in computer vision, drawing on the same self-attention mechanisms that revolutionised natural language processing. The adaptation wasn't immediately obvious — transformers were designed for sequences, while images have no inherent ordering. However, researchers discovered that treating image patches as 'tokens' could leverage transformers' ability to model long-range dependencies, a critical advantage over existing convolutional neural networks (CNNs) which were limited by fixed receptive fields (Dosovitskiy et al. 2020).

Transformers offer several key advantages for visual processing:

- **Global Context Understanding**: Unlike CNNs that build understanding through local receptive fields, transformers can attend to any part of an image, enabling better understanding of spatial relationships (Li, Wang, et al. 2023).
- **Scalability**: Transformers appear to scale more efficiently than CNNs, with performance continuing to improve with larger models and datasets (Zhai et al. 2022).
- **Transfer Learning**: Pre-trained vision transformers transfer knowledge across tasks more effectively than traditional CNN architectures (Alijani et al. 2024).
- **Unified Architecture**: The same architecture can handle multiple modalities (vision, language, audio) without requiring specialised components (Lim et al. 2025).

Vision transformers represent a remarkably rapid evolution in computer vision, emerging only in 2020 but achieving breakthrough results that challenged decades of CNN dominance within just a few years. This explosive development trajectory mirrors the pattern seen in natural language processing and suggests similar potential for rapid advancement in a range of applications. The pace of progress has been fast and consistent with new variants and improvements appearing every few months:

- **Original ViT (2020)**: Dosovitskiy et al. divided images into 16×16 pixel patches, treating each patch as a token. Initial results required large datasets (JFT-300M) but demonstrated that transformers could match or exceed CNN performance on image classification tasks (Dosovitskiy et al. 2020).
- **DETR (2020)**: Facebook's Detection Transformer applied the architecture to object detection and segmentation, showing that transformers could handle complex vision tasks beyond classification (Carion et al. 2020a).
- **SETR (2020-21)**: Developed specifically for semantic segmentation, demonstrating transformers' ability to capture fine-grained spatial information (Zheng et al. 2021).
- **DeiT (2021)**: Data-efficient Image Transformers showed that vision transformers could work with smaller datasets through knowledge distillation, making them more practical for real-world applications (Touvron et al. 2021).
- **CLIP (2021)**: OpenAI's breakthrough in connecting vision and language, training on 400 million image-text pairs to create a model that understands both modalities within a unified framework (Radford et al. 2021).
- **Swin Transformer (2021)**: Introduced hierarchical structure and shifted windows to reduce computational complexity while maintaining performance, making transformers more efficient for dense prediction tasks (Liu, Lin, et al. 2021a).

- **MAE (2022)**: Masked Autoencoders for self-supervised learning, showing that transformers could learn robust visual representations from unlabeled data (He et al. 2022).
- **EVA (2022)**: Achieved new state-of-the-art results with 1 billion parameters, demonstrating the continued scalability of vision transformers (Fang et al. 2022).
- **GPT-4V and Gemini (2023-2024)**: Integrated full multimodal capabilities into large language models, showing that transformers can seamlessly process and generate across visual and textual modalities (Gemini Team, Google 2023; OpenAI 2023b)

As of 2025, vision transformers have not only matched but often surpassed CNNs across most computer vision benchmarks [1]:

- **Image Classification**: Models like EVA-Giant achieve 90%+ accuracy on ImageNet, outperforming the best CNN architectures.
- **Segmentation**: Mask2Former and similar architectures provide unprecedented accuracy in both semantic and instance segmentation
- **Video Understanding**: TimeSformer and ViViT demonstrate temporal reasoning capabilities essential for dynamic scenes (Arnab et al. 2021).
- **3D Vision**: Transformer architectures are being successfully applied to point clouds and 3D scene understanding (Lu, Xu, et al. 2025).

The progression of vision transformers directly addresses many challenges in autonomous driving:

- **Multi-Camera Fusion**: Transformers naturally handle inputs from multiple sensors, crucial for 360-degree environmental understanding.
- **Temporal Integration**: Their ability to model sequences enables tracking objects across frames and predicting future trajectories.
- **Scene Understanding**: Global attention mechanisms help understand complex visual scenarios like traffic.
- **Robustness**: Self-attention can better handle variations in lighting, weather, and road conditions than traditional vision approaches (Lai-Dang 2024).

These developments demonstrate that transformer architectures are not just applicable to, but increasingly dominant in, computer vision tasks. The combination of superior scalability, transfer learning capabilities, and unified multimodal processing makes them a leading candidate for the complex visual processing required in autonomous driving systems.

[1] For advances in 2025, many of the best sources are the code repositories themselves such as EVA Giant, Mask2Former and TimesSformer

4 Current Industry Data Collection Efforts

Transformer-based vision models are, however, notoriously data-hungry, requiring large amounts of data across diverse domains (Lu, Zhang, et al. 2022). Major players in autonomous vehicles are already accumulating the massive datasets necessary for these approaches — and many are actively pursuing transformer architectures as their path to Level 5 autonomy (Ngiam et al. 2021).

Tesla has emerged as the industry leader in data collection scale, leveraging its massive consumer fleet to accumulate approximately 9 billion miles with Autopilot engaged and 1 billion miles with Full Self-Driving capabilities (Alvarez 2024; Tesla, Inc. 2025). This fleet-based approach provides continuous data collection across diverse conditions with minimal marginal cost, having already gathered approximately 333,000 hours of driving data (assuming 30mph average speeds). More importantly, Tesla has moved beyond data collection to implementation, integrating transformer-based networks into its FSD stack for object detection, path planning, and behavioral prediction, demonstrating the practical viability of these architectures in production vehicles (Think Autonomous 2023).

Waymo, by contrast, has adopted a more focused approach, prioritising high-quality data collection in specific geographic areas rather than broad-scale accumulation (Misraa et al. 2025). Their vehicles provide approximately 250,000 autonomous rides weekly in controlled environments like San Francisco and Phoenix, combining real-world driving with sophisticated simulation environments (Team 2025). This strategy emphasises safety and validation over volume, though Waymo's co-CEO Dmitri Dolgov has confirmed they have "leveraged the technology of transformers for behavioral prediction, decision-making, and semantic understanding" within their autonomy stack (Dolgov and Thrun 2024).

Beyond these industry leaders, a broader shift toward transformer architectures is evident across the sector. Companies are moving from pure rule-based or convolutional approaches to hybrid systems that incorporate transformer components for specific tasks. These implementations typically utilise transformers for particular components—such as scene understanding or trajectory prediction — rather than end-to-end systems, reflecting both the current computational constraints and the incremental path toward full transformer-based autonomy (Lai-Dang 2024).

The convergence of massive data collection efforts and active transformer implementation across major industry players suggests an industry-wide recognition that this architectural approach represents a promising path to Level 5 autonomy. As computational efficiency continues to improve and datasets grow, the transition from component-level applications to full end-to-end transformer systems appears increasingly inevitable.

5 Scaling Requirements for Autonomous Driving

From this point forward, our analysis becomes increasingly speculative, based on extrapolations from GPT4's architecture and historical trends in computing power improvements. While we acknowledge these projections involve considerable uncertainty, they provide a framework for understanding the scale of transformation required to move from text-based language models to real-time autonomous driving systems while still using a transformer based approach. Acknowleding this, we have aimed to be open and clear about the assumptions, results and uncertainty. The full set of assumptions are set out in Appendix A, detailed results are in Appendix B and sensitivity analysis is in Appendix C.

The first critical step for this extrapolation is assessing how much larger an image-based transformer model for autonomous driving would need to be compared to text-focused systems like GPT4. This comparison requires understanding fundamental differences in how these systems process information and represent knowledge.

5.1 Training Data Requirements

To understand the scale of data requirements, we must first clarify what constitutes a "token" in different contexts. In GPT4, a token typically represents a word or sub-word unit — roughly 3-4 characters of text. A single sentence might contain 10-20 tokens, allowing the model to process textual information at a rate roughly matching human reading comprehension (OpenAI 2025).

For an autonomous driving system, tokens take on an entirely different meaning. Each "token" would likely represent a patch of visual data — perhaps a 16×16 pixel segment of a camera frame, similar to what is currently used in Vision Transformers. At 30 frames per second across multiple cameras, a vehicle generates hundreds of thousands of visual tokens every second, compared to GPT4's processing of a few tokens per second (Dosovitskiy et al. 2020).

This fundamental difference drives massive scaling requirements:

- **Visual density**: Each frame contains far more information density than text, with spatial relationships and fine-grained details critical for safety.
- **Temporal requirements**: Unlike text consumption at 1-2 tokens per second, driving requires processing 30fps video across multiple cameras.
- **Scenario coverage**: The model must encounter rare edge cases — like emergency vehicles, construction zones, or unusual weather — frequently enough for reliable learning.
- **Environmental diversity**: The system must generalise across global variations in infrastructure, traffic patterns, weather conditions, and road quality.

Each of these factors is estimated to necessitate 3-4× the training data of GPT4's text-based learning. These factors compound multiplicatively, giving a range of 81-256 ×. Our estimates below are based on approximately 1,500 trillion tokens of training data being needed to capture the breadth of scenarios required for true Level 5 autonomy. This represents 120× GPT4's dataset, central to the range.

To make these numbers concrete, consider the practical realities of collecting such a dataset. A modern autonomous vehicle equipped with 8 cameras at 1280×720 resolution generates visual data at an extraordinary rate:

Table 1: Data generation parameters for autonomous driving system

Parameter	Value
Cameras per vehicle	8
Video resolution	1280×720 pixels
Frames per second	30
Patch size	16×16 pixels
Patches per frame	3,600
Tokens per second	864,000
Data generation rate	2.2 GB/sec

This data generation rate—nearly 2.2 gigabytes per second—means that accumulating our estimated 1,500 trillion tokens requires approximately 500,000 hours of continuous driving. The total dataset would occupy nearly 4 million terabytes (4 exabytes) of storage, making data management itself a significant engineering challenge.

To contextualise the challenge of storing and processing 4 million terabytes of driving data, we can compare it to the scale of YouTube, one of the largest video platforms in the world. YouTube receives an estimated 500 to 700 hours of video every minute, translating to roughly 1–1.5 petabytes of data per day, or approximately 1.2 to 1.5 exabytes annually when accounting for compression and redundancy.

Despite this, the economics of data collection are surprisingly manageable. Using a fleet-based approach similar to Tesla's current operations:

Table 2: Data collection fleet requirements and costs

Parameter	Value
Vehicle fleet size	50 cars
Hours per vehicle per day	8
Total hours generated per year	146,000
Years needed for collection	3.44 years
Cost per vehicle	$500,000
Cost per driver hour	$150
Total data acquisition cost	**$100,231,481**

This $100 million investment represents roughly 3-4 years of dedicated data collection with a modest fleet—a significant but achievable undertaking for major automotive manufacturers or technology companies. The timeline aligns well with typical automotive development cycles, suggesting that data collection need not be the limiting factor in transformer-based autonomous vehicle development.

Returning to the Youtube comparison, our estimated dataset is 4 exabytes collected over 3–4 years—is broadly comparable to YouTube's ingest volume during the same period. However, the comparison understates the complexity as autonomous vehicle datasets would need to be captured from multiple synchronised high-resolution cameras, often in raw or lightly compressed formats, and must be fully retained for annotation and model

training. While the scale is formidable, systems capable of managing YouTube-level ingest and retention provide a real-world precedent for the feasibility of transformer-scale autonomous vehicle data acquisition.

The quality and diversity of this data would be as critical as quantity. Unlike GPT4's text training, which could source from the relatively curated internet, autonomous vehicles must capture:

- Edge cases that may occur once in millions of miles;
- Regional variations in traffic patterns and infrastructure;
- Weather conditions from desert heat to arctic snow;
- Human driving behaviors across different cultures;
- Unusual scenarios like emergency vehicles, construction zones, and road incidents.

These requirements explain why simple data acquisition costs represent only part of the challenge. The data must be curated, annotated, and structured to ensure comprehensive coverage of the scenarios a vehicle might encounter over its lifetime. This transforms data collection from a passive logging exercise into an active engineering effort to capture the full spectrum of driving complexity.

While our projected data acquisition budget of $100 million may appear substantial, it is well-aligned with real-world precedents when adjusted for complexity and fidelity. Turning to Google again for another example, Google's Street View program — which has collected over 16 million km of imagery data globally since 2007 (Google 2025) — reportedly operates at a cost of roughly $1–$2 per mile for data gathering, primarily using contractor drivers and relatively simple camera rigs.

In contrast, an autonomous vehicle training dataset must not only capture high-resolution, multi-sensor input across diverse environments, but also log detailed driver behavior, synchronise human decisions with visual context, and ensure safety-critical edge case coverage. This requires custom vehicles, expert drivers, and extensive annotation infrastructure. Consequently, our cost estimate of $100 million for 500,000 hours of supervised, training-grade driving data—equivalent to over 15 million miles at an average speed of 30 mph (i.e. just less that $7 per mile) —represents a reasonable estimate relative to the depth and quality of data required for transformer-based autonomy.

Beyond the initial $100 million data acquisition investment lies the substantial ongoing challenge of data storage and retention. Our estimated 4 exabyte dataset would require approximately $8.8 million per month in cloud storage costs using enterprise-grade archival storage with geographic redundancy—totaling over $105 million annually (GDELT Project 2020).

This recurring expense represents a fundamental shift from traditional automotive development economics. Unlike conventional vehicle testing data that might be discarded after analysis, transformer-based autonomous systems require persistent access to the full training corpus for model updates, validation, and regulatory compliance. The storage costs alone exceed the entire annual R&D budgets of many automotive suppliers, highlighting how data-centric AI approaches fundamentally alter the economic structure of vehicle development.

5.2 Model Size Scaling Analysis

Parameters in neural networks represent the learned weights that encode the model's knowledge and decision-making capabilities. GPT4's 1.76 trillion parameters store the patterns, relationships, and reasoning capabilities derived from its training data. For an autonomous driving system, these parameters must encode far more complex spatial, temporal, and safety-critical knowledge.

The scaling from GPT4 to a driving-capable system involves several multiplicative factors:

- **Input complexity (2× factor)**: Visual data contains more structured, spatially-dependent information requiring additional parameters to process effectively
- **Temporal reasoning (2× factor)**: Understanding motion, predicting trajectories, and modeling physics over time demands substantial additional capacity
- **Safety-critical decisions (2× factor)**: The life-or-death nature of driving requires higher confidence levels, typically achieved through more robust internal representations
- **Multimodal integration (1.5× factor)**: Fusing data from cameras, LiDAR, radar, and GPS requires parameters dedicated to cross-modal understanding
- **Domain-specific optimizations (0.5× reduction)**: Purpose-built architectures for driving can achieve some efficiency gains over general-purpose language models

Combining these factors ($2 \times 2 \times 2 \times 1.5 \times 0.5$), we arrive at approximately 6× GPT4's parameter count, suggesting an autonomous driving system could potentially require roughly 11 trillion parameters to achieve comparable generalisation capabilities within its domain. This represents not just a quantitative scaling but a qualitative leap in the complexity of real-world reasoning these systems must perform.

The estimated 11 trillion parameters required for autonomous driving opens up intriguing possibilities beyond our computational projections. There is strong evidence of discontinuous abilities emerging in large language models, where new capabilities appear suddenly as parameters are scaled up (Bommasani et al. 2021; Wei et al. 2022). Moving from GPT4's approximately 1.76 trillion parameters to our proposed 11 trillion parameters will almost certainly unlock unanticipated capabilities (Ganguli et al. 2022; Hoffmann et al. 2022). While we cannot predict the specific nature of these emergent abilities, historical patterns suggest that such dramatic scaling often produces qualitative leaps in model capabilities rather than merely quantitative improvements.

6 Great Scott! Requirements in the Unoptimised Approach

Based on extrapolations from GPT4's reported requirements and the scaling factors we've established, a comprehensive end-to-end autonomous driving system would require computational resources that defy practical implementation. The analysis above assumes 2025 technology levels without **any** efficiency optimisations — essentially asking what it would take to build such a system today using ratios consistent with GPT4's development.

Starting from GPT4's baseline and applying our 6× scaling factor for parameters and 120× for training data, the full resource requirements emerge in stark detail:

Table 3: Training requirements for unoptimised autonomous driving system

Training Requirements	**Value**
Training Tokens	1.56×10^{15}
Training driving data	501,543 hours
Training data size	3,993,600 TB
Parameters	1.06×10^{13}
GPUs (A100s) assumed	50,000[2]
Training FLOPs	3.29×10^{28}
Training time	32,593 days (89 years)

The energy and cost implications are just as staggering:

Table 4: Economic and environmental costs of unoptimised training

Cost Category	**Value**
Training energy consumption	23,467 GWh
Training CO_2 emissions	9,152,000 tons CO_2-e
Training hardware cost	$1,006,250,000
Training energy cost	$3,520,000,000
Driving data acquisition	$100,231,481
Other costs	$4,526,250,000
Total Training cost	**$9,152,731,481**

The onboard deployment requirements prove equally prohibitive:

Table 5: Vehicle deployment requirements for unoptimised system

Vehicle Requirements	Value
Model weights storage	5,280 GB
Onboard compute requirement	1.82×10^{19} FLOPS/s
Onboard Thor chips needed	9,124
Onboard energy demand	364,954 kW
Battery range reduction	−100.00%
Onboard hardware cost per vehicle	$182,476,800

Most of the above results should be fairly self explanatory with the exception of "Onboard Thor chips needed". As a benchmark, our vehicle deployment analysis references NVIDIA's Thor system-on-chip (SoC). This product has been specifically designed for autonomous vehicle applications (NVIDIA Corporation 2023). Our cost estimate of $20,000 per Thor unit represents an extrapolation from current high-performance automotive computing platforms, though NVIDIA has not released official pricing. Economies of scale and technological advancement could reduce these costs significantly by deployment timeframes, this is accounted for in our calculations. The power consumption calculations assume each Thor chip operates at approximately 400 watts under full computational load, consistent with current high-performance automotive computing systems.

To put the onboard power requirements in perspective: at 364,954 kW, the system would consume roughly half the power output of a DeLorean's fictional flux capacitor (1.21 gigawatts). Unlike Doc Brown's time machine, however, this power draw would be continuous, making it physically impossible to implement in any production vehicle. The energy demand exceeds the total power delivery capacity of current high-voltage automotive systems by more than two orders of magnitude.

The 89-year training time would require maintaining a cluster of 50,000 A100 GPUs, twice the number used for GPT4, in perfect operational condition for nearly a century. The training energy consumption alone would roughly equal the annual electricity usage of a small European nation, with a carbon footprint equivalent to 2 million cars driven for a year.

These figures clearly demonstrate that an unoptimised transformer-based approach represents not merely an engineering challenge but a fundamental impossibility with current or near-future technology. The path to viable autonomous vehicles must necessarily run through dramatic efficiency improvements rather than brute-force scaling of existing architectures.

7 Mr. Fusion's Promise: The Optimised Approach

While the unoptimised approach reveals a computational impossibility, recent advances in AI efficiency offer genuine hope. The path to practical autonomous vehicles runs not through brute-force scaling but through dramatic efficiency improvements that are likely to emerge over time.

The target efficiency improvement can be fairly easily specified, we require approximately a 100,000× reduction in power consumption to bring onboard energy requirements below 5kW, which represents a reasonable practical limit comparable to existing auxiliary systems in electric vehicles (Farrington and Rugh 2013).

Such a massive efficiency gain might seem fantastical, but history provides compelling evidence that such transformations are not only possible but predictable over 25-year timeframes. Just as Doc Brown's Mr. Fusion transformed the power requirements of time travel, efficiency improvements can fundamentally alter what's possible with computational systems.

7.1 The Microcomputer Revolution: A Historical Blueprint

The personal computer revolution offers the most directly relevant precedent for understanding how dramatic efficiency improvements in computing can unfold over decades. Consider the transformation from the Apple II (1977) to a modern smartphone—a period that demonstrates the feasibility of 100,000× improvements.

Table 6: Computing evolution over 45 years

Metric	Apple II (1977)	iPhone 14 (2022)
Processor speed	1 MHz	3,460 MHz
RAM capacity	4 KB	6 GB
Storage capacity	None (cassette)	128 GB
Cost (2022 dollars)	$6,200	$799

Sources: Apple Computer Inc. (1977), Apple Inc. (2022), U.S. Bureau of Labor Statistics (2024), and Wikipedia (2024b,c)

The compound effects are staggering:

- **Processing speed**: 3,460× improvement.
- **Memory capacity**: 1.5 million× improvement.
- **Cost per unit of processing power**: approximately 27,000× reduction (7.8× cost reduction × 3,460× speed improvement)

These improvements followed Moore's Law remarkably consistently, with transistor density doubling approximately every 18 months (Moore 1965). Moore's Law traditionally focused on transistor density, but its broader implications encompass multiple efficiency vectors (Mack 2011):

- **Transistor shrinking**: Smaller transistors require less power and switch faster
- **Architectural innovations**: New designs like RISC, GPU parallelism, and specialised neural processing units
- **Manufacturing advances**: From micron-scale to nanometer processes, enabling dramatic power reductions
- **Software optimization**: Better algorithms and compilers extracting more performance from existing hardware

Between 1977 and 2022, we observed more than 30 doublings in transistor density — a 1-billion × increase — consistent with Moore's Law (Moore 1965). However, efficiency

gains in computing also depended on architectural and algorithmic breakthroughs, particularly after mid-2000s (Hennessy and Patterson 2019). Across that period, power efficiency improved dramatically, and cost per computation dropped substantially, depending on workload and hardware type (Koomey et al. 2011). These improvements came not just from raw transistor scaling, but also from architecture (RISC, GPUs, NPUs), manufacturing (nm-scale lithography), and algorithmic compression and sparsity.

For a 25-year projection (2025-2050), even a conservative continuation of historical trends yields (doubling every 18 months):

$$2^{(25\times12)/18} = 2^{16.67} \approx 100,000\times \tag{1}$$

This matches our requirement precisely, suggesting that a 100,000× efficiency improvement by 2050 aligns with historical precedent rather than representing an optimistic outlier.

Recent developments in AI hardware provide encouraging contemporary examples:

- **Apple's Neural Engine**: Achieved 10× efficiency improvement in just two chip generations (2018-2020) (Wikipedia 2024a).
- **Google's TPU progression**: TPUv4 offers 2.5× better performance per watt than TPUv3 after just two years (Jouppi, Yoon, Kurian, et al. 2021).
- **NVIDIA's efficiency advances**: The H100 provides 4× the training efficiency of the A100 in dense workloads (NVIDIA Corporation 2022).

These rapid improvements suggest that focused research on autonomous driving workloads could accelerate efficiency gains beyond general-purpose computing trends.

The historical evidence supports the feasibility of 100,000× efficiency improvements over 25 years. This isn't speculative technology — it's the continuation of well-established trends that have driven the computing industry for over half a century. Just as the microcomputer revolution made desktop computing possible, these efficiency gains promise to make transformer-based autonomous vehicles a practical reality by 2050.

7.2 A Pathway to 100,000× Efficiency

While historical trends demonstrate the feasibility of massive efficiency gains, achieving them requires concrete technological advances across multiple fronts. Our analysis identifies eight key improvement areas, each contributing multiplicatively to reach the required 100,000× overall improvement by 2050.

The following analysis projects efficiency improvements based on current research trends. While specific improvement factors and timelines are necessarily speculative, each technique is grounded in active research with demonstrated early results. The overall purpose is to demonstrate that there is a realistic pathway to 100,000× efficiency improvement rather than predict the precise timing or technology that will deliver these improvements.

Table 7: Projected efficiency improvements and their likelihood of achievement

Improvement Area	Reduction	Probability	Weighted	Timeline
Quantization	4×	75%	3.0×	2024-2026
Hardware Specialisation	15×	75%	11.25×	2025-2030
Parameter Sparsity	10×	95%	9.5×	2026-2030
Attention Optimisation	30×	85%	25.5×	2028-2032
Model Architecture	5×	50%	2.5×	2030-2035
Spatiotemporal Redundancy	6×	85%	5.0×	2030-2035
Activation Sparsity	2×	50%	1.0×	2035-2040
Total Expected Reduction			**100,000×**	**By 2050**

7.2.1 Quantization (4× improvement, 75% probability)

Quantization reduces the precision of model weights and activations, trading minimal accuracy for substantial efficiency gains (Jacob et al. 2018). This technique represents one of the most mature efficiency approaches, with commercial implementations already achieving 2-3× improvements in deployment scenarios (Nagel et al. 2021):

- **8-bit quantization**: Already standard in many deployments, providing significant memory and computational savings (Wu et al. 2020)
- **4-bit and lower**: Recent advances show viability with careful calibration, enabling even greater efficiency gains (Dettmers et al. 2023)
- **Mixed precision**: Use higher precision only where necessary, optimizing the accuracy-efficiency trade-off (Micikevicius et al. 2017)
- **Dynamic quantization**: Adjust precision based on uncertainty and importance of specific computations (Banner et al. 2019)

The high probability reflects the maturity of quantization techniques and their proven effectiveness across diverse neural network architectures. Current research focuses on pushing quantization to lower bit-widths while maintaining model performance, with 4× efficiency improvements representing a conservative estimate based on existing commercial deployments.

7.2.2 Hardware Specialisation (15× improvement, 75% probability)

Custom silicon designed specifically for neural network inference offers dramatic efficiency gains over general-purpose processors (Jouppi, Young, et al. 2017). Specialised hardware can be optimised for the specific computational patterns found in transformer architectures, eliminating overhead associated with general-purpose computing (Sze et al. 2017):

- **Neural processing units (NPUs)**: Dedicated tensor computation hardware optimised for matrix operations and attention mechanisms (Chen, Yang, et al. 2019).
- **In-memory computing**: Reduce data movement overhead by performing computations directly within memory arrays (Sebastian et al. 2020).
- **Analog computing elements**: Use physical properties for specific operations, offering potential for ultra-low power inference (Ambrogio et al. 2018).
- **Photonic computing**: Use light for matrix multiplication operations, promising significant power reductions (Shen et al. 2017).

Companies like Tesla (with their FSD chip) and Google (with TPUs) have already demonstrated 5-10× improvements over general-purpose hardware (Jouppi, Yoon, Ashcraft, et al. 2021; Lambert 2019), with next-generation designs targeting even greater gains. The automotive industry's focus on edge computing for autonomous vehicles provides strong incentives for continued hardware specialisation, making 15× improvements achievable within the projected timeline.

7.2.3 Parameter Sparsity (10× improvement, 95% probability)

Parameter sparsity leverages the fact that neural networks remain effective even when 80-90% of their weights are zero or near-zero (Han et al. 2015). This technique, already demonstrated in models like BERT and GPT variants (Michel et al. 2019; Prasanna et al. 2020), involves:

- **Pruning**: Systematically removing weights that contribute little to model performance through structured or unstructured approaches (Louizos et al. 2017a).
- **Sparse training**: Training models to naturally develop sparse weight distributions from initialisation (Mocanu et al. 2018)
- **Dynamic sparsity**: Adapting which weights are active based on the specific input, enabling input-dependent efficiency (Renda et al. 2020).

Current research shows 80% sparsity with minimal performance loss, and industry implementations like sparse transformer blocks are already in production (Jaszczur et al. 2021). The high probability reflects both proven current capabilities and straightforward scaling to higher sparsity levels. Recent work demonstrates that even 90% sparsity can be achieved while maintaining competitive performance on language tasks (Sanh et al. 2020), suggesting 10× efficiency gains are readily achievable.

7.2.4 Attention Optimisation (30× improvement, 85% probability)

Standard transformer attention has $O(n^2)$ complexity, meaning computational requirements grow quadratically with sequence length (Vaswani et al. 2017). Novel attention mechanisms promise dramatic reductions in both computational and memory requirements (Tay et al. 2022):

- **Linear attention**: Approximations like Linformer and Performer reduce complexity to $O(n)$, enabling processing of much longer sequences (Choromanski et al. 2020; Wang, Li, et al. 2020).
- **Sparse attention**: Patterns like local windows and global tokens (as in Longformer) that focus computation on relevant regions (Beltagy et al. 2020; Child et al. 2019).
- **Flash attention**: Memory-efficient implementations that reduce I/O overhead through careful memory hierarchy optimisation (Dao 2023; Dao et al. 2022).
- **Hardware-aware designs**: Attention patterns optimised for specific chip architectures, maximizing throughput on available hardware (Jaszczur et al. 2021; Rabe and Staats 2021).

Early implementations already show 5-10× speedups, with theoretical foundations suggesting 30× is achievable as these techniques mature and combine (Peng et al. 2021;

Qin et al. 2022). The high probability reflects the active research focus on attention efficiency and the demonstrable progress already achieved in reducing transformer computational complexity.

7.2.5 Model Architecture (5× improvement, 50% probability)

Beyond incremental optimisations, entirely new architectures may emerge specifically designed for autonomous driving (Carion et al. 2020b; Dosovitskiy et al. 2020):

- **Mixture of Experts (MoE)**: Only activate relevant network portions for specific inputs, dramatically reducing computational load while maintaining model capacity (Fedus et al. 2021; Riquelme et al. 2021).
- **Hierarchical models**: Process information at multiple resolutions simultaneously, enabling efficient multi-scale understanding of driving scenes (Liu, Lin, et al. 2021b; Wang, Xie, et al. 2021).
- **Event-driven processing**: React to changes rather than constantly processing, leveraging temporal sparsity in driving scenarios (Gehrig et al. 2019; Messikommer et al. 2022).
- **Neuromorphic designs**: Brain-inspired architectures with inherent efficiency, potentially offering orders of magnitude power reductions (Davies et al. 2018; Roy et al. 2019).

The moderate probability reflects the uncertainty in predicting architectural breakthroughs, though current research directions appear promising. Recent work on vision transformers specifically adapted for autonomous driving shows potential for significant efficiency gains through domain-specific design choices (Chen, Li, et al. 2022; Li, Ge, et al. 2023).

7.2.6 Spatiotemporal Redundancy Exploitation (6× improvement, 85% probability)

Driving perception involves significant redundancy across both time and space—consecutive frames share most information, and visual scenes contain repetitive spatial patterns that can be leveraged for efficiency (Hu et al. 2018; Liu, Huang, et al. 2018; Zhu et al. 2017):

- **Frame differencing and motion tracking**: Process only changes between frames while tracking object motion, reducing computational load for static scene elements (Ilg et al. 2017; Sun et al. 2018).
- **Foveated processing**: Higher resolution only where attention is focused, mimicking human visual processing and concentrating computation on safety-critical regions (Kim and Canny 2017).
- **Keyframe and multi-resolution systems**: Full processing on select frames with spatial hierarchies, interpolation for others, enabling efficient spatiotemporal understanding (Jiang et al. 2018; Lin et al. 2017; Zhao et al. 2017).
- **Predictive processing**: Use motion models and scene structure to anticipate next states and compress common driving patterns (Luc et al. 2017; Mathieu et al. 2015).
- **Adaptive spatiotemporal sampling**: Higher processing for dynamic regions and complex spatial areas, lower for static backgrounds and predictable scene elements (Figurnov et al. 2017; Korbar et al. 2019).

The high probability reflects the well-established nature of both temporal and spatial redundancy exploitation techniques. Recent work in autonomous driving has demonstrated that combined spatiotemporal consistency can achieve significant efficiency gains while maintaining perception accuracy (Chen, Zhou, et al. 2019; Qi et al. 2021). However, safety-critical applications demand extensive validation of any computational shortcuts, particularly ensuring that dynamic objects and small spatial details critical for safety are never missed (Bojarski et al. 2016; Sauer et al. 2018).

7.2.7 Activation Sparsity (2× improvement, 50% probability)

Beyond parameter sparsity, even activated portions of networks often contain sparse patterns that can be exploited for computational efficiency (Elsen et al. 2020; Kurtz et al. 2020):

- **ReLU exploitation**: Naturally creates sparse activations through zero-valued outputs, enabling skip computations (Wang, Yu, et al. 2018)
- **Gated architectures**: Only activate neurons when necessary, using learned gating mechanisms to control computation flow (Chollet 2017; Howard et al. 2017)
- **Conditional computation**: Skip layers based on input characteristics, adapting network depth to input complexity (Bengio et al. 2013; Graves 2016)
- **Learned sparsity patterns**: Networks that naturally develop efficient activation patterns through training objectives that encourage sparsity (Louizos et al. 2017b; Molchanov et al. 2017)

The moderate probability reflects the difficulty of achieving high activation sparsity while maintaining performance, particularly for safety-critical applications (Veit and Belongie 2018). While activation sparsity has shown promise in computer vision tasks, autonomous driving applications require consistent performance across all network activations, making aggressive sparsification challenging (Huang et al. 2017; Teerapittayanon et al. 2016). Recent work suggests that 2× improvements are achievable through careful activation sparsity design without compromising safety-critical performance.

7.2.8 Compound Effects

These improvements multiply together, not add:

$$Total = 3.0 \times 11.25 \times 9.5 \times 25.5 \times 2.5 \times 5.0 \times 1.0 \approx 100,000\times \tag{2}$$

The timeline shows a progression from more mature techniques (quantization, hardware specialisation) deployed in the mid-2020s to more experimental approaches (spatiotemporal redundancy, activation sparsity) maturing in the 2030s and 2040s. This staged development allows for iterative validation and refinement, crucial for safety-critical autonomous driving applications. The multiplicative nature of these improvements means that even partial success across multiple fronts can deliver substantial efficiency gains, providing multiple pathways to achieve the required 100,000× improvement by 2050.

8 2040: Training Feasibility Achieved, Deployment Still Distant

By 2040, the compound effects of efficiency improvements fundamentally transform the economics of training transformer-based autonomous driving systems. At this inflection point, with approximately 1,024× efficiency gains realised, model training becomes a manageable undertaking comparable to current GPT4 development efforts. However, the onboard deployment challenge remains formidable, creating a critical gap between training capability and practical implementation.

Table 8: Training requirements reduced by 1,000× efficiency gains

Training Requirements (2040)	**Value**
Training Tokens	1.56×10^{15}
Training driving data	501,543 hours
Parameters	1.06×10^{13}
GPUs (A100 equivalent)	25,000
Training FLOPs	3.22×10^{25}
Training time	64 days

Table 9: Training costs now within reach of major corporations

Cost Category (2040)	**Value**
Training energy consumption	23 GWh
Training CO_2 emissions	6,113 tons CO_2-e
Training hardware cost	$233,093,400
Training energy cost	$4,297,673
Driving data acquisition	$100,231,481
Other costs	$237,391,073
Total Training cost	**$575,013,627**

The transformation by 2040 is remarkable: training time drops from 89 years to a manageable 64 days, while total costs fall from over $9 billion to approximately $575 million. At this scale, training becomes comparable to current flagship AI projects. The 25,000 GPU cluster requirement matches or slightly exceeds what leading technology companies deploy today for language model training. The 64-day training window aligns with typical development cycles, allowing for iterative improvements and experimentation. Most importantly, the $575 million total investment falls within the research budgets of major automotive manufacturers and technology giants.

This point represents the critical threshold where transformer-based autonomous driving transitions from theoretical curiosity to practical engineering project. However, the onboard deployment story remains challenging:

Table 10: Vehicle deployment still prohibitive despite 1,000× efficiency gains

Vehicle Requirements (2040)	Value
Model weights storage	5,280 GB
Onboard compute requirement	1.78×10^{16} FLOPS/s
Onboard Thor chips needed	9
Onboard energy demand	356 kW
Range reduction	-97.49%
Onboard hardware cost per vehicle	$82,558

The onboard power requirement of 356 kW, while dramatically improved from 2025's 365 MW, would still consume the vehicle's entire battery capacity in approximately 12 minutes, making practical deployment impossible.

The 2040 milestone marks the end of the impossibility phase and the beginning of the implementation phase in transformer-based autonomous driving. While full consumer deployment remains a decade away, the transition from science fiction to engineering challenge represents a critical point in the autonomous vehicle revolution.

9 2050: Practicality Achieved—The Consumer Revolution

By 2050, the cumulative 104,032× efficiency improvement fundamentally transforms both training economics and vehicle deployment feasibility. Training costs fall to levels accessible to mid-sized companies, while onboard requirements finally meet the practical constraints of consumer vehicles. This marks the true beginning of the Level 5 autonomous vehicle era.

Table 11: Training requirements in 2050: Accessible to mid-sized companies

Training Requirements (2050)	Value
Training Tokens	1.56×10^{15}
Training driving data	501,543 hours
Parameters	1.06×10^{13}
GPUs (A100 equivalent)	1,000
Training FLOPs	3.17×10^{23}
Training time	16 days

The economic transformation is striking: total training costs drop to approximately $111 million — comparable to developing a major automotive subsystem such as a new transmission technology or a complete vehicle interior redesign for a luxury brand (Belzowski 2010). The 16-day training window enables rapid experimentation and improvement cycles. With only 1,000 GPUs required, training infrastructure becomes accessible to a broader range of companies beyond the current AI giants.

More critically, vehicle deployment constraints finally align with practical consumer requirements:

The 3.5 kW power requirement falls within the operational envelope of modern electric

Table 12: Training costs approach the scale of traditional automotive R&D

Cost Category (2050)	Value
Training energy consumption	0.2 GWh
Training CO_2 emissions	47 tons CO_2-e
Training hardware cost	$5,582,465
Training energy cost	$49,094
Driving data acquisition	$100,231,481
Other costs	$5,631,559
Total Training cost	**$111,494,600**

Table 13: Vehicle deployment becomes commercially viable

Vehicle Requirements (2050)	Value
Model weights storage	76 GB
Onboard compute requirement	1.75×10^{14} FLOPS/s
Onboard Thor chips needed	0.1
Onboard energy demand	3.5 kW
Range reduction	-27.68%
Onboard hardware cost per vehicle	$487

vehicles, comparable to running air conditioning and premium audio systems simultaneously (Farrington and Rugh 2013). At 27.68% range reduction, the trade-off is acceptable, particularly as battery technology continues improving in parallel. The $487 hardware cost per vehicle sits squarely within automotive bill-of-materials expectations for advanced features (Deichmann et al. 2023).

The 2050 milestone represents more than a technical achievement — it marks the transformation of autonomous driving from limited and elite to a potential global transportation revolution. With both training and deployment constraints resolved, the path clears for truly ubiquitous Level 5 autonomy, fundamentally reshaping how humanity moves through the world.

GPT4 as a Commodity: The 2050 Perspective

To fully appreciate the transformation efficiency brings, consider how our baseline model — GPT4 — becomes accessible by 2050. Current GPT4 development required vast resources and months of dedicated supercomputing clusters (Henshall 2024). Under our efficiency improvement trajectory, these requirements shrink dramatically.

Metric	GPT4 (2023)	GPT4 (2050)
Training cost	$1,016,027,778	$223,435
Training time	91 days	1 day
GPU cluster size	25,000 A100s	20 A100s
Energy consumption	33 GWh	<0.1 GWh
CO_2 emissions	12,711 tons	0.1 tons

Table 14: GPT4 development requirements: 2023 vs 2050

This transformation democratises what was once the exclusive domain of tech giants. In 2050, training a GPT4 caliber model would cost roughly $223,435—comparable to hiring an experienced engineer for a year. The single-day training time fits comfortably within standard business analytics cycles, while the 20-GPU requirement lies within reach of university research labs or mid-sized businesses.

This commoditisation pattern mirrors the personal computer revolution—what required university computing centers in 1970 fit on a desktop by 1990. Similarly, what required billion dollar investments in 2023 becomes accessible to mid-sized businesses by 2050.

The autonomous driving transformer represents not just a transportation revolution but a continuation of computing's democratising trajectory, following Feynman's principle that there's always room at the bottom (Feynman 1960)—where relentless efficiency improvements eventually transform today's supercomputing capabilities into tomorrow's commodity technologies, making advanced AI accessible to increasingly broad segments of society.

In this context, our 2050 autonomous driving system — while vastly more complex than GPT4 — becomes economically viable precisely because it leverages the same efficiency revolution that transformed yesterday's Apple II into the iPhone.

10 Research Timeline and Agenda

Our analysis reveals a clear 25-year pathway to Level 5 autonomous vehicles, with distinct phases aligned to technological maturation and economic viability. This comprehensive timeline integrates data collection, efficiency research, model development, and deployment preparation into a coherent strategy.

10.1 Phase 1: Data Acquisition (2025-2030)

The immediate priority involves establishing comprehensive driving datasets through a coordinated global data collection effort:

Table 15: Phase 1 timeline: Building the foundational dataset

Milestone	Target Date
Deploy 50-vehicle data collection fleet	2025-2026
Establish global coverage partnerships	2026-2027
Gather 501,543 hours of driving data	2025-2030
Develop real-time annotation systems	2027-2028
Complete data collection ($100M total)	2030

This phase focuses on capturing global geographic diversity, edge cases, and rare events while establishing annotation standards and distributed storage infrastructure for exabyte-scale datasets. A portion of the data should be published for community research benchmarking to accelerate broader industry progress.

10.2 Phase 2: Efficiency Focus (2030-2040)

This decade prioritises achieving the critical 1,000× efficiency improvement needed for training feasibility. The most promising near-term breakthroughs are expected in three core areas:

Table 16: Phase 2 priorities: Critical efficiency breakthroughs

Research Priority	Target Improvement	Timeline
Quantization	4× reduction	2024-2026
Hardware specialisation	15× efficiency gain	2025-2030
Parameter sparsity	10× reduction	2026-2030
Combined with attention optimization	**≥1,000× total**	**2030-2032**

These three technologies represent the most mature and commercially viable pathways to the needed efficiency gains. Any combination of these improvements, coupled with advances in attention optimisation (30× potential), provides multiple routes to the 1,000× milestone required for training feasibility. This redundancy reduces technical risk—researchers need not succeed across all fronts simultaneously.

Delivering these breakthroughs requires coordinated industry action across multiple fronts. Companies should form public-private research consortiums to share R&D costs and accelerate progress on quantization and hardware specialisation. Simultaneously,

major players need to acquire and deploy large-scale GPU training clusters for experimentation, while developing prototype testing environments to validate efficiency improvements. This decade also demands proactive engagement with regulators to establish early frameworks for autonomous vehicle validation and industry-wide safety standards.

10.3 Phase 3: Model Training (2040-2050)

The training decade begins with exclusive access by major players and evolves to broader industry participation:

Table 17: Phase 3 timeline: Democratization of model training

Milestone	Industry Participation	Target Year
First training feasibility achieved	1-2 major tech/automotive pioneers	2040-2042
Large corporation entry	5-7 multinational players	2043-2045
Mid-tier competitor access	10-15 established companies	2046-2048
Broad industry participation	20+ companies globally	2049-2050

This progression reflects continuing efficiency improvements that drive training costs from $475 million in 2040 down to $11 million by 2050. Early pioneers like Tesla, Google, or Volkswagen establish initial proof-of-concept models, but, as costs decline, the technology becomes accessible to a broader range of automotive manufacturers, technology companies, and regional players.

The decade's key focus shifts from pure research to practical implementation; multiple competing architectures emerge as companies explore different approaches, iterative improvement cycles shorten from months to weeks, and extensive simulation testing validates safety performance.

Real-world controlled trials will be needed throughout the decade to demonstrate practical performance while international regulatory frameworks and harmonised standards are developed.

10.4 Phase 4: Vehicle Deployment (2050 onwards)

Full consumer viability arrives with the critical 100,000× efficiency milestone. By 2050, the technical constraints that previously made transformer-based autonomous vehicles impractical finally align with consumer requirements: 3.5 kW power consumption becomes manageable within electric vehicle power budgets, $487 hardware costs per vehicle fall within standard automotive electronics pricing, and 27.68% range reduction represents an acceptable trade-off for full autonomy. Model storage requirements of just 76 GB integrate seamlessly with existing automotive computing architectures.

This technical breakthrough enables autonomous vehicles to transition from premium luxury feature to standard capability across the automotive industry. Multiple manufacturers begin offering Level 5 autonomy across diverse price points, while ongoing model improvements allow rapid iteration and enhancement. Global deployment extends beyond developed markets to cover diverse international environments, from European vil-

lage centers to sub-Saharan rural roads.

Beyond 2050, the societal impact will begin to extend far beyond individual vehicle ownership, echoing the transformative effects of railway expansion in the 19th century. Just as trains revolutionized commerce, urbanisation, and social mobility by connecting previously isolated communities, transformer-based autonomous vehicles promise similarly profound changes.

Transportation-as-a-service models will reshape urban mobility patterns, while rural and developing regions gain unprecedented access to automated transport, potentially leapfrogging traditional automotive infrastructure in the same way mobile phones bypassed landline networks in emerging economies. Like the railway boom that enabled the growth of suburbs and transformed labour markets, ubiquitous Level 5 autonomy may fundamentally alter where people choose to live and work, potentially reversing decades of urban concentration as geographical constraints on mobility dissipate.

10.5 Economic Timeline: The 26-Year Investment Horizon

The research timeline outlined above requires unprecedented financial commitment across the autonomous vehicle industry. Our cash flow analysis reveals the stark economic reality: achieving transformer-based Level 5 autonomy demands 25 years of sustained investment before generating meaningful revenue, fundamentally reshaping competitive dynamics and industry structure.

Table 18: Total financial requirements

Financial Metric	Value
Total PV (2025-2050)	$618 million
Total investment	$1.57 billion
Peak annual investment	$524 million (2040)

This investment profile creates profound strategic implications:

- **Barrier to Entry**: The $618 million, 25-year commitment effectively restricts serious development to a handful of players—major technology companies (Google, Apple), automotive manufacturers with substantial resources (Volkswagen, Toyota, Tesla), and nationally-supported enterprises. Mid-tier companies face a stark choice: partner with leaders or accept the risk of permanent technological obsolescence.
- **Data Advantage**: Companies beginning data collection in 2030 have the opportunity to establish an insurmountable head start. The front-loaded nature of investment means later entrants face significant disadvantages in data quality and coverage. Early data collection provides not only a head start in research but also the extended time necessary to capture rare edge cases and achieve comprehensive scenario coverage across diverse global conditions. Companies beginning data early gain crucial years to identify and document the safety-critical scenarios that may occur only once in millions of miles—coverage that cannot be replicated through superior execution alone.
- **Platform Economics**: The massive upfront investment suggests successful companies will likely monetise their technology through licensing arrangements. This cre-

ates potential for new industry structures where a few technology providers enable many vehicle manufacturers.

The cash flow pattern is starkly different from typical automotive programs, with 3-5 year development cycles. Transformer-based autonomy represents a full-career commitment—engineers beginning this work in 2025 will reach retirement as the first commercial systems deploy in 2050. This generational timeline requires not just financial staying power but institutional memory and strategic vision capable of sustaining consistent investment across multiple economic cycles, leadership changes, and technological disruptions before generating returns.

For context, the $618 million total investment compares to developing 2-3 traditional vehicle platforms, while the peak $524 million annual commitment in 2040 aligns with current flagship AI development budgets. These figures position transformer-based autonomy as a major but manageable undertaking for well-capitalised industry participants.

The economic timeline reinforces our core thesis: companies recognising this transition today and committing accordingly will dominate the transportation landscape of the 2050s and beyond. Those viewing this as a conventional automotive development program will find themselves spectators to one of the most significant technological transformations in transportation history.

10.6 Critical Success Factors

This timeline requires coordinated effort across multiple fronts:

1. **Sustained Research Investment**: Likely more than $10 billion industry-wide over 25 years.
2. **Industry Partnerships**: Toe enable shared data, infrastructure and risk distribution.
3. **Regulatory Evolution**: Adaptive frameworks enabling an appropriate balance between innovation and safety.
4. **International Cooperation**: Global standards and data sharing to get the most out of data and model training.
5. **Parallel Technology Development**: Complementary advances in batteries, sensors, and connectivity.

The path we've outlined does not reflect optimistic speculation but extrapolation from established technological trends. Each phase builds upon demonstrable capabilities and economic realities, creating a robust roadmap toward the transformer-based autonomous vehicle revolution.

11 Limitations

This analysis, while comprehensive in scope, rests on several important limitations that readers should consider when interpreting our findings.

Uncertainty in Core Assumptions: The projections presented throughout this paper are based on a range of assumptions detailed in Appendix A. These assumptions draw

from academic research and industry-accepted values rather than proprietary information from commercial IP owners, introducing inherent uncertainty into our analysis. Key parameters—including model scaling requirements, efficiency improvement rates, and computational costs—are estimates based on publicly available information and may differ from actual commercial implementations.

Timeline Variability: The inherent uncertainty in our assumptions translates to variability in our projected timeline. Our sensitivity analysis (Appendix C) presents both optimistic and pessimistic scenarios, suggesting a realistic deployment window of 2042-2058, with 2050 as our central estimate. This range reflects the compound effects of uncertainty across multiple technological and economic variables, each of which could accelerate or delay the ultimate timeline.

Incomplete Cost Modeling: Our cost analysis focuses exclusively on core operational requirements—data acquisition, model training, and deployment infrastructure. We have not incorporated several potentially significant cost categories:

- Research and development costs to achieve the projected 100,000× efficiency improvement
- Regulatory compliance and safety certification expenses
- Legal and insurance frameworks for autonomous vehicle deployment
- Infrastructure modifications required for full Level 5 autonomy

These additional costs could substantially increase the total investment required and may affect the timeline for commercial viability.

Technology-Specific Extrapolation: Our analysis extrapolates from transformer-based architectures, as these represent the most viable option with substantial evidence regarding scaling behavior and computational requirements. However, this approach may miss potential paradigm shifts. Alternative architectures or entirely new approaches to machine learning could emerge, potentially accelerating our timeline or fundamentally altering the computational requirements we project.

Potential for Discontinuous Progress: Our projections assume relatively smooth efficiency improvements over time, but the history of AI suggests the possibility of discontinuous breakthroughs. As observed with large language models, certain capability thresholds appear to unlock emergent behaviors. Vision transformers may exhibit similar discontinuities as they scale, potentially achieving driving-capable performance at lower parameter counts than our linear extrapolations suggest.

External Technological Factors: Our analysis holds certain external technologies constant, particularly battery performance and energy density. Breakthrough improvements in these areas could fundamentally alter our calculations. For instance, a 10× improvement in battery energy density would dramatically reduce the impact of computational power requirements on vehicle range, potentially accelerating deployment timelines.

Safety Factors: Our analysis implicitly assumes that safety-critical validation challenges will be resolved during the research and training phases, with our estimated 120×

data multiplier (relative to GPT4) providing sufficient edge case exposure for reliable autonomous operation. However, this assumption merits careful consideration. Unlike text generation models, where errors result in poor outputs, autonomous vehicle failures can have catastrophic consequences. Achieving acceptable safety performance may ultimately require orders of magnitude more validation data than our projections suggest. The path from computational feasibility to safety certification represents a distinct challenge that could extend deployment timelines significantly beyond our 2050 projection.

Despite these limitations, we believe our analysis provides a valuable framework for understanding the computational economics of transformer-based autonomous vehicles and the likely trajectory toward Level 5 autonomy.

12 Conclusion

The path to Level 5 autonomous vehicles is clear but requires a fundamental shift. Current approaches face insurmountable scaling challenges, while transformer-based systems offer a potential route to global Level 5 autonomy by 2050. The choice facing industry leaders today will determine who dominates the transportation landscape of tomorrow.

The key findings:

- **Current approaches are scaling dead-ends**: Mapping all of the world's roads in detail and keeping this up-to-date is fundamentally impossible to maintain over time.
- **Transformers are the most apparent scalable solution today**: Importantly, attention mechanisms may eliminate the need for exhaustive pre-mapping.
- **100,000× efficiency gains are needed but follow historical precedent**: The microcomputer revolution demonstrates such transformations are possible, not speculative.
- **Critical timeline milestones**: 2030 data collection complete ($100M), 2040 training complete ($475M), 2050 mass deployment ($487/vehicle).

The autonomous vehicle industry stands at a crossroads. Companies that recognise this transition and invest accordingly will lead the revolution of the 2040s. Those that continue to pursue scaling dead-ends will become spectators to one of the most significant technological transformations in history.

There are clear critical actions for industry leaders:

- Deploy a vehicle data collection fleet of 50+ vehicles before 2030 to capture the 500,000 hours of diverse driving scenarios required.
- Begin recruiting before a likely intensification of competition for top-tier AI talent from 2030.
- Establish partnerships with leading AI research institutions focused on transformer efficiency improvements—particularly in attention optimisation, parameter sparsity, and hardware specialisation, this should be opeating at full scale by the early 2030s.
- Instigate and closely monitor R&D developments in the 2030s to choose when to commit to large-scale model training investments.

- Commit board-level resources to a 25-year investment horizon, because the companies that begin this transition today will define transportation for the next century.

The future of mobility is transformer-powered and it begins with a $618 million decision and 25 year commitment made today.

References

Alijani, S., J. Fayyad, and H. Najjaran (Apr. 2024). "Vision Transformers in Domain Adaptation and Domain Generalization: A Study of Robustness". In: *arXiv preprint*. URL: (visited on 05/22/2025) (page 9).

Alvarez, S. (Apr. 5, 2024). *Tesla FSD Fleet Passes 1 Billion-Mile Milestone*. Teslarati. URL: (visited on 05/22/2025) (page 11).

Ambrogio, S. et al. (2018). "Equivalent-accuracy accelerated neural-network training using analogue memory". In: *Nature* 558.7708, pp. 60–67 (page 20).

Apple Computer Inc. (1977). *Apple II Reference Manual*. Original technical specifications (page 18).

Apple Inc. (2022). *iPhone 14 Technical Specifications*. https://support.apple.com/kb/SP873. Accessed: 2025-05-23 (page 18).

Arnab, A., M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid (Mar. 2021). "ViViT: A Video Vision Transformer". In: *arXiv preprint*. URL: (visited on 05/22/2025) (page 10).

Banner, R., Y. Nahshan, E. Hoffer, and D. Soudry (2019). "Post training 4-bit quantization of convolutional networks for rapid-deployment". In: *Advances in neural information processing systems* 32 (page 20).

Bastian, M. (Apr. 13, 2023). *GPT-4 Architecture, Datasets, Costs, and More Leaked*. The Decoder. URL: (visited on 05/21/2025) (page 8).

Beltagy, I., M. E. Peters, and A. Cohan (2020). "Longformer: The long-document transformer". In: *arXiv preprint arXiv:2004.05150* (page 21).

Belzowski, B. (July 2010). "Why Does It Cost So Much For Automakers To Develop New Models?" In: *Autoblog*. Accessed: 2025-05-23 (page 25).

Bengio, Y., N. Léonard, and A. Courville (2013). "Estimating or propagating gradients through stochastic neurons for conditional computation". In: *arXiv preprint arXiv:1308.3432* (page 23).

Bojarski, M. et al. (2016). "End to end learning for self-driving cars". In: *arXiv preprint arXiv:1604.07316* (page 23).

Bommasani, R. et al. (2021). "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (page 15).

Bowen, S. (Sept. 18, 2023). *The Carbon Footprint of GPT-4*. Medium. URL: (visited on 05/21/2025) (page 8).

Bretting, S. (May 6, 2024). *Tesla Shared Realistic FSD Roadmap for the First Time, There's No Autonomous Driving on It*. autoevolution. URL: (visited on 05/21/2025) (page 6).

Bubeck, S. et al. (Mar. 2023). "Sparks of Artificial General Intelligence: Early experiments with GPT-4". In: *arXiv preprint*. URL: (visited on 05/21/2025) (page 8).

Carion, N., F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko (2020a). "End-to-End Object Detection with Transformers". In: *Computer Vision – ECCV 2020*. Vol. 12346. Lecture Notes in Computer Science. Cham: Springer, pp. 213–229. URL: (visited on 05/22/2025) (page 9).

– (2020b). "End-to-end object detection with transformers". In: *European conference on computer vision*, pp. 213–229 (page 22).

Central Intelligence Agency (2022). *Roadways - Country Comparison*. Central Intelligence Agency. URL: (visited on 05/21/2025) (page 6).

Chang, Y. et al. (Feb. 2024). "A Survey on Evaluation of Large Language Models". In: *ACM Transactions on Intelligent Systems and Technology* 15.3. URL: (visited on 05/21/2025) (page 8).

Chen, D., B. Zhou, V. Koltun, and P. Krähenbühl (2019). "Learning by cheating". In: *arXiv preprint arXiv:1912.12294* (page 23).

Chen, Y.-H., T.-J. Yang, J. Emer, and V. Sze (2019). "Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices". In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9.2, pp. 292–308 (page 20).

Chen, Z., Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao (2022). "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers". In: *European conference on computer vision*, pp. 1–18 (page 22).

Child, R., S. Gray, A. Radford, and I. Sutskever (2019). "Generating long sequences with sparse transformers". In: *arXiv preprint arXiv:1904.10509* (page 21).

Chollet, F. (2017). "Xception: Deep learning with depthwise separable convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258 (page 23).

Choromanski, K. et al. (2020). "Rethinking attention with performers". In: *arXiv preprint arXiv:2009.14794* (page 21).

Dao, T. (2023). "FlashAttention-2: Faster attention with better parallelism and work partitioning". In: *arXiv preprint arXiv:2307.08691* (page 21).

Dao, T., D. Y. Fu, S. Ermon, A. Rudra, and C. Ré (2022). "FlashAttention: Fast and memory-efficient exact attention with IO-awareness". In: *Advances in Neural Information Processing Systems* 35, pp. 16344–16359 (page 21).

Davies, M. et al. (2018). “Loihi: A neuromorphic manycore processor with on-chip learning”. In: *IEEE Micro* 38.1, pp. 82–99 (page 22).

Deichmann, J., E. Ebel, K. Heineke, R. Heuss, M. Kellner, and F. Steiner (Jan. 2023). “Autonomous driving’s future: Convenient and connected”. In: *McKinsey & Company*. Accessed: 2025-05-23 (page 26).

Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer (2023). “QLoRA: Efficient finetuning of quantized LLMs”. In: *Advances in Neural Information Processing Systems* 36 (page 20).

Dolgov, D. and S. Thrun (Feb. 22, 2024). *The Road Ahead with Waymo Co-CEO Dimitri Dolgov*. Podcast. 36-minute interview moderated by Sebastian Thrun discussing Waymo’s fully autonomous vehicles and the future of the autonomous driving industry. URL: (visited on 05/22/2025) (page 11).

Dosovitskiy, A. et al. (Oct. 2020). “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *arXiv preprint*. URL: (visited on 05/22/2025) (pages 9, 12, 22).

Elsen, E., M. Dukhan, T. Gale, and K. Simonyan (2020). “Fast sparse convnets”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14629–14638 (page 23).

Fang, Y., W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao (Nov. 2022). “EVA: Exploring the Limits of Masked Visual Representation Learning at Scale”. In: *arXiv preprint*. URL: (visited on 05/22/2025) (page 10).

Farrington, R. and J. Rugh (2013). “Impact of the Air-Conditioning System on the Power Consumption of an Electric Vehicle Powered by Lithium-Ion Battery”. In: *Mathematical Problems in Engineering* 2013, p. 935784 (pages 18, 26).

Fedus, W., B. Zoph, and N. Shazeer (2021). “Switch transformer: Scaling to trillion parameter models with simple and efficient sparsity”. In: *Journal of Machine Learning Research* 22.120, pp. 1–39 (page 22).

Feynman, R. P. (1960). “There’s Plenty of Room at the Bottom”. In: *Engineering and Science* 23.5. Originally presented as a talk at the annual meeting of the American Physical Society at the California Institute of Technology on December 29, 1959, pp. 22–36 (page 27).

Figurnov, M., M. D. Collins, Y. Zhu, L. Zhang, J. Huang, D. Vetrov, and R. Salakhutdinov (2017). “Spatially adaptive computation time for residual networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1039–1048 (page 22).

Ganguli, D. et al. (2022). “Predictability and surprise in large generative models”. In: *arXiv preprint arXiv:2202.07785* (page 15).

GDELT Project (2020). *GCP Tips & Tricks: The Surprisingly Cost-Effective Economics Of GCS Storage Classes For Backups*. Accessed May 22, 2025. URL: (page 14).

Gehrig, D., A. Loquercio, K. G. Derpanis, and D. Scaramuzza (2019). "End-to-end learning of representations for asynchronous event-based data". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5633–5643 (page 22).

Gemini Team, Google (Dec. 2023). "Gemini: A Family of Highly Capable Multimodal Models". In: *arXiv preprint*. URL: (visited on 05/22/2025) (page 10).

Google (2025). *How Street View Works and Where We Will Collect Images Next*. Technical overview of Street View image collection, processing, and 3D reconstruction methodology. Google. URL: (visited on 05/22/2025) (page 14).

Graves, A. (2016). "Adaptive computation time for recurrent neural networks". In: *arXiv preprint arXiv:1603.08983* (page 23).

Han, S., H. Mao, and W. J. Dally (2015). "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding". In: *arXiv preprint arXiv:1510.00149* (page 21).

He, K., X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick (June 2022). "Masked Autoencoders Are Scalable Vision Learners". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, pp. 16000–16009. URL: (visited on 05/22/2025) (page 10).

Hennessy, J. L. and D. A. Patterson (2019). *Computer architecture: a quantitative approach*. 6th. Morgan Kaufmann (page 19).

Henshall, W. (June 2024). "The Billion-Dollar Price Tag of Building AI". In: *TIME*. Accessed: 2025-05-23 (page 27).

Hoffmann, J. et al. (2022). "Training compute-optimal large language models". In: *arXiv preprint arXiv:2203.15556* (page 15).

Howard, A. G., M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam (2017). "MobileNets: Efficient convolutional neural networks for mobile vision applications". In: *arXiv preprint arXiv:1704.04861* (page 23).

Hu, H., J. Gu, Z. Zhang, J. Dai, and Y. Wei (2018). "Relation networks for object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3588–3597 (page 22).

Huang, G., D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger (2017). "Multi-scale dense networks for resource efficient image classification". In: *arXiv preprint arXiv:1703.09844* (page 23).

Ilg, E., N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox (2017). “FlowNet 2.0: Evolution of optical flow estimation with deep networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2462–2470 (page 22).

Jacob, B., S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko (2018). “Quantization and training of neural networks for efficient integer-arithmetic-only inference”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2704–2713 (page 20).

Jaszczur, S., A. Łuczakowski, A. Chowdhery, M. Dehghani, T. Mohiuddin, Ł. Kaiser, D. Grangier, H. Michalewski, and J. Kanerva (2021). “Sparse is enough in scaling transformers”. In: *Advances in Neural Information Processing Systems* 34, pp. 9895–9907 (page 21).

Jiang, H., D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz (2018). “Super SloMo: High quality estimation of multiple intermediate frames for video interpolation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9000–9008 (page 22).

Jouppi, N. P., D. H. Yoon, M. Ashcraft, et al. (2021). “Ten Lessons From Three Generations Shaped Google’s TPUv4i : Industrial Product”. In: *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, pp. 1–14 (page 21).

Jouppi, N. P., C. Young, et al. (2017). “In-datacenter performance analysis of a tensor processing unit”. In: *ACM SIGARCH Computer Architecture News* 45.2, pp. 1–12 (page 20).

Jouppi, N. P., D. H. Yoon, G. Kurian, S. Li, N. Patil, J. Laudon, C. Young, and D. Patterson (2021). *TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings*. arXiv preprint arXiv:2104.09758 (page 19).

Kim, J. and J. Canny (2017). “Interpretable learning for self-driving cars by visualizing causal attention”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 2942–2950 (page 22).

Koomey, J., S. Berard, M. Sanchez, and H. Wong (2011). “Implications of historical trends in the electrical efficiency of computing”. In: *IEEE Annals of the History of Computing* 33.3, pp. 46–54 (page 19).

Korbar, B., D. Tran, and L. Torresani (2019). “SCsampler: Sampling salient clips from video for efficient action recognition”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6232–6242 (page 22).

Kurtz, M. et al. (2020). “Inducing and exploiting activation sparsity for fast inference on deep neural networks”. In: *International conference on machine learning*, pp. 5533–5543 (page 23).

Lai-Dang, Q.-V. (Mar. 2024). “A Survey of Vision Transformers in Autonomous Driving: Current Trends and Future Directions”. In: *arXiv preprint*. URL: (visited on 05/22/2025) (pages 10, 11).

Lambert, F. (Apr. 2019). “Tesla unveils its new Full Self-Driving computer in detail: ’objectively the best chip in the world’”. In: *Electrek*. Accessed: 2025-05-25. URL: (page 21).

Li, Y., Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li (2023). “BEVDepth: Acquisition of reliable depth for multi-view 3D object detection”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.2, pp. 1477–1485 (page 22).

Li, Y., J. Wang, X. Dai, L. Wang, C.-C. M. Yeh, Y. Zheng, W. Zhang, and K.-L. Ma (Mar. 2023). “How Does Attention Work in Vision Transformers? A Visual Analytics Attempt”. In: *arXiv preprint*. Accepted by PacificVis 2023. URL: (visited on 05/22/2025) (page 9).

Lim, Q. Z., C. P. Lee, K. M. Lim, and K. S. M. Anbananthen (Apr. 2025). “VLMT: Vision-Language Multimodal Transformer for Multimodal Multi-hop Question Answering”. In: *arXiv preprint*. URL: (visited on 05/22/2025) (page 9).

Lin, T.-Y., P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie (2017). “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125 (page 22).

Liu, S., D. Huang, and Y. Wang (2018). “Receptive field block net for accurate and fast object detection”. In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 385–400 (page 22).

Liu, Z., Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo (Oct. 2021a). “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE/CVF, pp. 10012–10022. URL: (visited on 05/22/2025) (page 9).

– (2021b). “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022 (page 22).

Louizos, C., M. Welling, and D. P. Kingma (2017a). “Learning sparse neural networks through Lo regularization”. In: *arXiv preprint arXiv:1712.01312* (page 21).

– (2017b). “Learning sparse neural networks through Lo regularization”. In: *arXiv preprint arXiv:1712.01312* (page 23).

Lu, D., L. Xu, J. Zhou, K. (Gao, and J. Li (June 2025). "3DLST: 3D Learnable Supertoken Transformer for LiDAR Point Cloud Scene Segmentation". In: *International Journal of Applied Earth Observation and Geoinformation* 140, p. 104572. URL: (visited on 05/22/2025) (page 10).

Lu, J., X. S. Zhang, T. Zhao, X. He, and J. Cheng (June 2022). "APRIL: Finding the Achilles' Heel on Privacy for Vision Transformers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, pp. 10051–10060. URL: (visited on 05/22/2025) (page 11).

Luc, P., N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun (2017). "Predicting deeper into the future of semantic segmentation". In: *Proceedings of the IEEE international conference on computer vision*, pp. 648–657 (page 22).

Mack, C. (2011). "Fifty years of Moore's law". In: *IEEE Transactions on Semiconductor Manufacturing* 24.2, pp. 202–207 (page 18).

Markoff, J. (Jan. 31, 2024). *How to Build a Real Driverless Car. Waymo's self-driving car, more than a decade in the making*. IEEE Spectrum. URL: (visited on 05/21/2025) (page 6).

Mathieu, M., C. Couprie, and Y. LeCun (2015). "Deep multi-scale video prediction beyond mean square error". In: *arXiv preprint arXiv:1511.05440* (page 22).

McGuinness, P. (Apr. 16, 2023). *GPT-4 Details Revealed*. Pat McGuinness's Substack. URL: (visited on 05/21/2025) (page 8).

Messikommer, N., D. Gehrig, A. Loquercio, and D. Scaramuzza (2022). "Event-based asynchronous sparse convolutional networks". In: *European Conference on Computer Vision*, pp. 415–431 (page 22).

Michel, P., O. Levy, and G. Neubig (2019). "Are sixteen heads really better than one?" In: *Advances in neural information processing systems* 32 (page 21).

Micikevicius, P. et al. (2017). "Mixed precision training". In: *arXiv preprint arXiv:1710.03740* (page 20).

Misraa, A. K., N. Jain, and S. S. Dhakad (Apr. 2025). "Waymo Driverless Car Data Analysis and Driving Modeling using CNN and LSTM". In: *arXiv preprint*. URL: (visited on 05/22/2025) (page 11).

Mocanu, D. C., E. Mocanu, P. Stone, P. H. Nguyen, M. Gibescu, and A. Liotta (2018). "Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science". In: *Nature communications* 9.1, p. 2383 (page 21).

Molchanov, D., A. Ashukha, and D. Vetrov (2017). "Variational dropout sparsifies deep neural networks". In: *International conference on machine learning*, pp. 2498–2507 (page 23).

Moore, G. E. (1965). “Cramming more components onto integrated circuits”. In: *Electronics* 38.8, pp. 114–117 (page 18).

Nagel, M., M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. van Baalen, and T. Blankevoort (2021). “A white paper on neural network quantization”. In: *arXiv preprint arXiv:2106.08295* (page 20).

Ngiam, J. et al. (June 2021). “Scene Transformer: A Unified Architecture for Predicting Multiple Agent Trajectories”. In: *arXiv preprint*. URL: (visited on 05/22/2025) (page 11).

NVIDIA Corporation (2022). *NVIDIA H100 Tensor Core GPU*. https://www.nvidia.com/en-au/data-center/h100/. Accessed: 2025-05-23 (page 19).

– (2023). *NVIDIA Unveils DRIVE Thor*. URL: (visited on 05/22/2025) (page 17).

OpenAI (Mar. 2023a). “GPT-4 Technical Report”. In: *arXiv preprint*. URL: (visited on 05/21/2025) (page 8).

– (Sept. 2023b). *GPT-4V System Card*. Tech. rep. OpenAI. URL: (visited on 05/21/2025) (pages 8, 10).

– (2025). *What Are Tokens and How to Count Them?* OpenAI Help Center documentation explaining tokenization, token counting, and the tiktoken tokenizer. OpenAI. URL: (visited on 05/22/2025) (page 12).

Peng, H., N. Pappas, D. Yogatama, R. Schwartz, N. A. Smith, and L. Kong (2021). “Random feature attention”. In: *arXiv preprint arXiv:2103.02143* (page 21).

Prasanna, S., A. Rogers, and A. Rumshisky (2020). “When BERT plays the lottery, all tickets are winning”. In: *arXiv preprint arXiv:2005.00561* (page 21).

Qi, C. R., X. Chen, O. Litany, and L. J. Guibas (2021). “Offboard 3D object detection from point cloud sequences”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6134–6144 (page 23).

Qin, Z., W. Sun, H. Deng, D. Li, Y. Wei, B. Lv, J. Yan, L. Kong, and Y. Zhong (2022). “cosFormer: Rethinking softmax in attention”. In: *arXiv preprint arXiv:2202.08791* (page 21).

Rabe, M. N. and C. Staats (2021). “Self-attention does not need $O(n^2)$ memory”. In: *arXiv preprint arXiv:2112.05682* (page 21).

Radford, A. et al. (July 2021). “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 8748–8763. URL: (visited on 05/22/2025) (page 9).

Renda, A., J. Frankle, and M. Carbin (2020). “Comparing rewinding and fine-tuning in neural network pruning”. In: *arXiv preprint arXiv:2003.02389* (page 21).

Riquelme, C., J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. S. Pinto, D. Keysers, and S. Gelly (2021). “Scaling vision with sparse mixture of experts”. In: *Advances in Neural Information Processing Systems* 34, pp. 8583–8595 (page 22).

Roy, K., A. Jaiswal, and P. Panda (2019). “Towards spike-based machine intelligence with neuromorphic computing”. In: *Nature* 575.7784, pp. 607–617 (page 22).

Sanh, V., T. Wolf, and A. Rush (2020). “Movement pruning: Adaptive sparsity by fine-tuning”. In: *Advances in Neural Information Processing Systems* 33, pp. 20378–20389 (page 21).

Sauer, A., N. Savinov, and A. Geiger (2018). “Conditional affordance learning for driving in urban environments”. In: *arXiv preprint arXiv:1806.06498* (page 23).

Sebastian, A., M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou (2020). “Memory devices and applications for in-memory computing”. In: *Nature nanotechnology* 15.7, pp. 529–544 (page 20).

Shen, Y. et al. (2017). “Deep learning with coherent nanophotonic circuits”. In: *Nature Photonics* 11.7, pp. 441–446 (page 20).

Sun, D., X. Yang, M.-Y. Liu, and J. Kautz (2018). “PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8934–8943 (page 22).

Sze, V., Y.-H. Chen, T.-J. Yang, and J. S. Emer (2017). “Efficient processing of deep neural networks: A tutorial and survey”. In: *Proceedings of the IEEE* 105.12, pp. 2295–2329 (page 20).

Tay, Y. et al. (2022). “Efficient transformers: A survey”. In: *ACM Computing Surveys* 55.6, pp. 1–28 (page 21).

Team, W. (May 5, 2025). *Scaling Our Fleet Through U.S. Manufacturing*. Vice President of Operations discusses Waymo’s new autonomous vehicle integration plant in Mesa, Arizona with partner Magna. Waymo. URL: (visited on 05/22/2025) (page 11).

Teerapittayanon, S., B. McDanel, and H. Kung (2016). “BranchyNet: Fast inference via early exiting from deep neural networks”. In: *23rd international conference on pattern recognition (ICPR)*, pp. 2464–2469 (page 23).

Tesla, Inc. (2025). *Tesla Vehicle Safety Report*. Quarterly safety data including Q1 2025 results showing one crash per 7.44 million miles driven with Autopilot engaged. Tesla, Inc. URL: (visited on 05/22/2025) (page 11).

Think Autonomous (Sept. 2023). *Breakdown: How Tesla Will Transition from Modular to End-To-End Deep Learning*. Think Autonomous. URL: (visited on 05/22/2025) (page 11).

Touvron, H., M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou (July 2021). "Training Data-Efficient Image Transformers & Distillation Through Attention". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 10347–10357. URL: (visited on 05/22/2025) (page 9).

U.S. Bureau of Labor Statistics (2024). *Consumer Price Index Inflation Calculator*. https://www.bls.gov/data/inflation_calculator.htm. Accessed: 2025-05-23 (page 18).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). "Attention Is All You Need". In: *arXiv preprint*. URL: (visited on 05/21/2025) (pages 8, 21).

Veit, A. and S. Belongie (2018). "Convolutional networks with adaptive inference graphs". In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–18 (page 23).

Wang, S., B. Z. Li, M. Khabsa, H. Fang, and H. Ma (2020). "Linformer: Self-attention with linear complexity". In: *arXiv preprint arXiv:2006.04768* (page 21).

Wang, W., E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao (2021). "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 568–578 (page 22).

Wang, X., F. Yu, Z.-Y. Dou, T. Darrell, and J. E. Gonzalez (2018). "SkipNet: Learning dynamic routing in convolutional networks". In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 409–424 (page 23).

Waymo Team (Sept. 10, 2020). *The Waymo Driver Handbook: Mapping*. Waymo. URL: (visited on 05/21/2025) (page 6).

Wei, J. et al. (2022). "Emergent abilities of large language models". In: *arXiv preprint arXiv:2206.07682* (page 15).

Wikipedia (2024a). *Apple A14*. https://en.wikipedia.org/wiki/Apple_A14. Accessed: 2025-05-23 (page 19).

– (2024b). *Apple II (original)*. https://en.wikipedia.org/wiki/Apple_II_(original). Accessed: 2025-05-23 (page 18).

– (2024c). *iPhone 14*. https://en.wikipedia.org/wiki/iPhone_14. Accessed: 2025-05-23 (page 18).

Wu, H., P. Judd, X. Zhang, M. Isaev, and P. Micikevicius (2020). “Integer quantization for deep learning inference: Principles and empirical evaluation”. In: *arXiv preprint arXiv:2004.09602* (page 20).

Zhai, X., A. Kolesnikov, N. Houlsby, and L. Beyer (2022). “Scaling Vision Transformers”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, pp. 12104–12113. URL: (visited on 05/22/2025) (page 9).

Zhao, H., J. Shi, X. Qi, X. Wang, and J. Jia (2017). “Pyramid scene parsing network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890 (page 22).

Zheng, S. et al. (June 2021). “Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, pp. 6881–6890. URL: (visited on 05/22/2025) (page 9).

Zhu, X., Y. Wang, J. Dai, L. Yuan, and Y. Wei (2017). “Deep feature flow for video recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2349–2358 (page 22).

Acknowledgments

We would like to acknowledge the assistance of AI language models in the preparation of this manuscript. Portions of this paper were developed with the assistance of OpenAI’s GPT and Anthropic’s Claude. These tools aided in literature research, reference formatting, and drafting sections of the text. All AI-generated content was reviewed, edited, and verified for accuracy by the author.

The views, opinions, and conclusions expressed in this paper are solely those of the author and do not necessarily represent the official position or policies of any organisations, institutions, or employers with which the author may be affiliated. This work was conducted in the author’s personal capacity and should not be construed as representing the views of any current or former employers.

A Assumptions Register

This appendix documents the key assumptions underlying our analysis. All monetary values are in 2025 USD unless otherwise specified.

Table 19: Energy and environmental assumptions

Parameter	Value	Unit	Annual Change
Electricity cost	0.15	$/kWh	+1.50%
Carbon intensity	0.39	kg CO_2e/kWh	-3.00%

Table 20: Computational specifications and requirements

Parameter	Value	Unit
Token specifications		
Text token size	4	bytes
Image token size	2,560	bytes
Computational requirements		
Training FLOPs per parameter per token	2	FLOPs
Inference FLOPs per parameter per token	2	FLOPs
Parameter storage during inference	0.5	bytes/parameter
Processing specifications		
Context window	32,000	tokens
Processing window	0.033	seconds

Table 21: Hardware infrastructure specifications

Parameter	Value	Unit
GPU specifications		
GPU unit cost	12,500	$/GPU
Server cost	15,000	$/server
GPUs per server	8	units
GPU power consumption	400	watts/GPU
GPU performance	3.12×10^{14}	FLOPS
Infrastructure overhead		
Hardware overhead cost	40	%
Power overhead	50	%
Utilization factor	75	%

Table 22: Vehicle and sensor specifications

Parameter	Value	Unit
Sensor configuration		
Cameras per vehicle	8	units
Video resolution	1280×720	pixels
Frame rate	30	fps
Vision transformer patch size	16×16	pixels
Vehicle specifications		
NVIDIA Thor performance	2×10^{15}	FLOPS
Thor unit cost	20,000	$/unit
Onboard compute efficiency	5×10^{10}	FLOP/watt
Onboard power overhead	0	%
Electric vehicle baseline		
Model Y battery capacity	75	kWh
Energy consumption	141	Wh/km
Implied range	532	km
Average driving speed	65	km/h

Table 23: Scaling factors for autonomous vehicle requirements

Parameter	Value	Rationale
Token multiplier (vs GPT-4)	120×	Visual density and temporal requirements
Parameter multiplier (vs GPT-4)	6×	Spatial and safety-critical processing
Efficiency improvement period	1.5	years/doubling

Table 24: Data collection operational parameters

Parameter	Value	Unit
Data collection fleet size	50	vehicles
Operating hours per vehicle	8	hours/day
Vehicle acquisition cost	500,000	$/vehicle
Driver cost	150	$/hour
Data storage cost	2.2	$M/month/exabyte

B Computational Requirements and Costs Over Time

Table 25: Evolution of computational requirements and costs for GPT4 and autonomous vehicles from 2023 to 2050

	GPT4		Autonomous Vehicle (AV)						GPT4
Year	**2023**		**2025**	**2035**	**2040**	**2045**	**2050**		**2050**
Efficiency Improvement			1	102	1,024	10,321	104,032		104,032
Hardware cost factor			100%	60%	46%	36%	28%		
Training Tokens	1.30×10^{13}		1.56×10^{15}	1.56×10^{15}	1.56×10^{15}	1.56×10^{15}	1.56×10^{15}		1.30×10^{13}
Training Data (hours)	-		501,543	501,543	501,543	501,543	501,543		-
Training Data Size (TB)	52		3,993,600	3,993,600	3,993,600	3,993,600	3,993,600		52
Parameters	1.76×10^{12}		1.06×10^{13}	1.06×10^{13}	1.06×10^{13}	1.06×10^{13}	1.06×10^{13}		1.76×10^{12}
GPUs (A100s)	25,000		50,000	25,000	25,000	10,000	1,000		20
Training FLOPs	4.58×10^{25}		3.29×10^{28}	3.24×10^{26}	3.22×10^{25}	3.19×10^{24}	3.17×10^{23}		4.40×10^{20}
Training Time (days)	91		32,593	642	64	16	16		1
Training Energy (GWh)	33		23,467	231	23	2	0.2		0.0003
Training CO_2 (t)	12,711		9,152,000	69,935	6,113	534	47		0.1
Driving data acquisition			$100M	$100M	$100M	$100M	$100M		
Data storage (per year)			$105M	$63M	$49M	$38M	$29M		
Hardware Cost	$503M		$1,006M	$301M	$233M	$72M	$5.6M		$112K
Energy Cost	$4.9M		$3,520M	$40M	$4.3M	$460K	$49K		$68
Other Costs	$508M		$4,526M	$341M	$237M	$73M	$5.6M		$112K
Total Training Cost	$1,016M		$9,153M	$783M	$575M	$245M	$111M		$223K
Model Storage (GB)	880		5,280	5,280	5,280	763	76		13
Onboard Compute (FLOPS/s)	-		1.82×10^{19}	1.80×10^{17}	1.78×10^{16}	1.77×10^{15}	1.75×10^{14}		-
Onboard Thor Chips Needed	-		9,124	90	9	1	0.1		-
Power Demand (kW)	-		364,954	3,592	356	35	3.5		-
Range Reduction	-		-100%	-99.75%	-97.49%	-79.42%	-27.68%		-
Hardware Cost/Vehicle	-		$182M	$1.1M	$82K	$6.3K	$487		-

Table 26: 25-Year Cash Flow Analysis: Investment Breakdown by Category (millions 2025 USD)

Year	Data Acquisition	Data Storage	Model Training	Total Cost
PV (7%)	**$67**	**$390**	**$161**	**$618**
Sum	**$100**	**$994**	**$475**	**$1,569**
2025	$0	$0	$0	$0
2026	$0	$0	$0	$0
2027	$0	$0	$0	$0
2028	$0	$0	$0	$0
2029	$0	$0	$0	$0
2030	**$100**	$0	$0	**$100**
2031	$0	$78	$0	$78
2032	$0	$74	$0	$74
2033	$0	$70	$0	$70
2034	$0	$66	$0	$66
2035	$0	$63	$0	$63
2036	$0	$60	$0	$60
2037	$0	$57	$0	$57
2038	$0	$54	$0	$54
2039	$0	$51	$0	$51
2040	$0	$49	**$475**	**$524**
2041	$0	$46	$0	$46
2042	$0	$44	$0	$44
2043	$0	$42	$0	$42
2044	$0	$40	$0	$40
2045	$0	$38	$0	$38
2046	$0	$36	$0	$36
2047	$0	$34	$0	$34
2048	$0	$32	$0	$32
2049	$0	$31	$0	$31
2050	$0	$29	$0	$29

C Sensitivity Analysis

Table 27: Sensitivity analysis of key parameters on project economics and feasibility

Scenario	PV Cost ($M)	Data Time (years)	Data Cost ($M)	Initial Storage ($M)	Training Time 2040 (days)	Training Cost 2040 ($M)	Onboard Power 2050 (kW)
Base case	618	3.4	100	105	64	475	3.5
Token scaling scenarios							
Token multiplier (81×)	474	2.3	76	71	43	472	3.5
Token multiplier (256×)	1,121	7.3	185	225	136	485	3.5
Parameter scaling scenarios							
Parameter multiplier (3×)	617	3.4	100	105	32	470	1.8
Parameter multiplier (9×)	620	3.4	100	105	95	479	5.3
Efficiency improvement scenarios							
Doubles every 12 months	615	3.4	100	105	2	466	< 0.1
Doubles every 24 months	632	3.4	100	105	360	515	63.0
Hardware cost scenarios							
GPU cost 0% CAGR	1,155	3.4	100	105	64	1,015	4.0
GPU cost 10% CAGR	353	3.4	100	105	64	216	4.0

The sensitivity analysis reveals several key insights:

- **Token requirements** have the most dramatic impact on project economics, with PV ranging from $474M to $1,121M.
- **Efficiency improvement rates** critically determine feasibility—doubling every 12 months enables deployment with negligible power. requirements by 2050, while 24-month doubling periods result in impractical 63 kW requirements, with feasibility achieved in around 2058.
- **Parameter scaling** primarily affects onboard power requirements, with minimal impact on overall project PV.
- **Hardware cost trajectories** significantly influence training costs but have limited effect on final deployment feasibility.

These results support our base case timeline of 2050 for commercial deployment, with realistic bounds of 2042-2058 depending on actual efficiency improvements achieved.