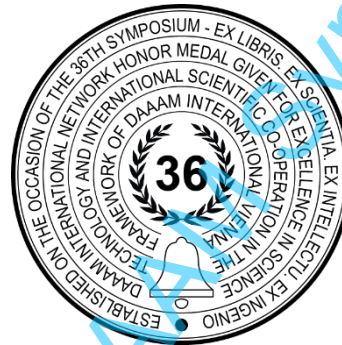


PRODUCT AND PROCESS INNOVATION IN ARTIFICIAL INTELLIGENCE: REVISITING INDUSTRY EVOLUTION ACROSS TECHNOLOGICAL LAYERS

Milivoj Markovic & Eva Hadina



This Publication has to be referred as: Markovic, M[ilivoj]; Hadina E[va]; (2026). Product and Process Innovation in Artificial Intelligence: Revisiting Industry Evolution Across Technological Layers, Proceedings of the 37th DAAAM International Symposium, pp.xxxx-xxxx, B. Katalinic (Ed.), Published by DAAAM International, ISBN 978-3-902734-xx-x, ISSN 1726-9679, Vienna, Austria
DOI: 10.2507/37th.daaam.proceedings.xxx

Abstract

The artificial intelligence (AI) industry challenges the classical model in which product innovation dominates early and process innovation follows the emergence of a dominant design. Prices for a fixed level of AI capability have fallen steeply, yet new products continue to appear. This paper examines how product and process innovation interact in AI and what this implies for competitive strategy. Building on the product–process innovation model, the reverse product cycle in services, and theories of technological hierarchies and general-purpose technologies, it proposes a framework of four interdependent technological layers and two innovation pathways: product-led, in which a new capability originates first, and process-enabled, in which efficiency gains relax application-specific feasibility constraints. The framework is examined through five application cases using secondary evidence. Originating innovations were mainly product-led, while process improvements supported diffusion and downstream applications; feasibility constraints plausibly mediate this relationship, although their causal role is not established. The paper derives implications for competitive advantage and for deployment choices between cloud and local computing, and outlines three conditional scenarios for the industry's evolution.

Keywords: product innovation; process innovation; artificial intelligence; industry evolution; competitive strategy.

1. Introduction

Generative artificial intelligence (AI) has moved from research laboratories to mass markets in a few years. Its development raises a question that innovation and strategy research has long asked of new industries: how do product and process innovation interact as an industry evolves, and what does this mean for competitive advantage?

The classical answer is a sequence. In the model of Abernathy and Utterback, product innovation dominates early, a dominant design emerges, and competition then shifts towards process innovation and cost [1]. Additional research traced this path in the smartphone industry as well [2]. AI fits the sequence poorly. The price of a fixed level of model capability fell more than 280-fold between November 2022 and October 2024 [3], indicating substantial changes in the economics

of AI services, while new products and capabilities continued to emerge. Barras's reverse product cycle for services [4], research on technological hierarchies [5], [6] and the theory of general-purpose technologies [7] each explain part of this pattern. Although these perspectives offer complementary explanations of technological evolution, their integration provides an opportunity to examine how product-led and process-enabled innovation pathways interact across the technological layers of the emerging AI industry.

This paper asks how the relationship between product and process innovation shapes the evolution of the AI industry, and with what implications for competitive advantage. It distinguishes four layers, from computing hardware and infrastructure to foundation models and applications, and two pathways. In the product-led pathway, a new capability originates first and process innovation follows. In the process-enabled pathway, efficiency gains in enabling layers relax application-specific feasibility constraints and allow new downstream products. The argument is that the order of product and process innovation depends on the layer and the stage of commercialisation examined.

The study is conceptual, supported by a theory-guided qualitative examination of five purposively selected application cases using secondary evidence; generative AI, and large language models in particular, provides the empirical context. Section 2 develops the framework and propositions, Section 3 examines them against industry evidence, Section 4 derives implications for competitive strategy and conditional scenarios, and Section 5 concludes.

2. Theoretical Framework

2.1. Conventional product–process innovation dynamics

The classical model holds that the emphasis of innovation shifts from products to processes as an industry matures. Abernathy and Utterback described three stages, fluid, transitional and specific, through which a "productive unit", a product line together with its production process, evolves [1], [8]. Many competing designs coexist in the fluid stage. The emergence of a dominant design ends this experimentation and shifts competition towards cost, scale and process improvement [1], [9].

Klepper formalised the industry-level consequences: over time, firms devote more effort to process innovation, while the number of firms and the rate and diversity of product innovation decline [10]. In the disk-drive industry, firms that adopted the dominant design were far more likely to survive [11]. Consumer electronics followed the same path, from design diversity in early smartphones to process-driven, increasingly commoditised competition [2].

The model has known limits. Dominant designs are usually identified only in hindsight, and heterogeneous customer demand offers an alternative explanation of the life cycle and can sustain performance improvement after technical convergence [12]. Applied to AI, dominant-design theory suggests that convergence may rest on business-model as much as technical innovation, with AI offered as a service [13]; this paper focuses instead on how product and process innovation are sequenced across technological layers. Most importantly for this paper, the model assumes a clearly bounded productive unit. In AI, a trained model is at once a product sold to customers and an input into other firms' products and processes.

2.2. Reverse innovation dynamics

Barras proposed that service industries follow the opposite sequence [4]. The adoption of information technology first yields incremental process innovations that raise the efficiency of existing services, then more radical process innovations that improve their quality, and only finally product innovations that create new services. Critics argue that the model describes how technologies made in manufacturing diffuse into services rather than innovation in services as such [14], and note that the cycle can be observed only ex post.

Barras is the main alternative explanation this paper must address, because much of AI is delivered as a service and paid for per unit of use. His cycle is driven by the diffusion of a new technology into service activities, and nothing in it rules out cost or performance thresholds; diffusion, learning and thresholds can operate together. We therefore treat the threshold mechanism as a complementary explanation of the timing and economic conditions of downstream product innovation. Its value is to make the relevant constraints observable: where thresholds matter, specific products should appear after identifiable falls in the cost, latency or resource requirements that bound them, a timing pattern that diffusion and learning may accompany but do not by themselves point to.

2.3. Technological hierarchies and complementarities

Clark argued that design problem-solving and the concepts customers use to choose between products give technological evolution a hierarchical structure [15]. Murmann and Frenken generalised this view: complex products evolve as nested hierarchies of technology cycles, so dominant designs can emerge at system, subsystem or component level, and different parts of one product can be at different life-cycle stages at once [5].

Funk added a direction: improvements at lower levels of a design hierarchy change design trade-offs higher up and so produce new product classes [6]. This is the closest precedent to our mechanism, and cost, performance and resource requirements can themselves be read as design trade-offs. Our framework is therefore an extension and operationalisation

of hierarchical innovation theory, not a new form of innovation. What the AI case adds is that the trade-off crosses firm boundaries and is priced in a market, per token or per task, which makes it observable in price and performance data.

General-purpose technologies (GPTs) show a related relationship between sectors: innovation in the GPT raises the productivity of research and development in the sectors that use it [7]. Large language models have been classified as having GPT characteristics [16], deep learning has been described as a general-purpose "method of invention" [17], and earlier DAAAM research expected AI to accelerate product and service innovation while making imitation easier [18]. Product and process innovation can also be complements rather than stages [19].

2.4. Integrating the perspectives: innovation across technological layers

Related research distinguishes AI enablement, production and consumption, and documents that a few large firms dominate both upstream computing and downstream applications [20]. We refine this into four interdependent technological layers: computing hardware; infrastructure on which models are trained and run; foundation models; and applications and services. Within the infrastructure layer, cloud data centres and local devices are alternative deployment configurations with different technologies, economics and operating trade-offs, not successive stages.

Three processes are distinguished (Fig. 1). Originating innovation arises within a layer, such as a new accelerator architecture or a new model capability, and is classified from the viewpoint of the firm that introduces it. Enabled innovation occurs when a development in another layer lowers the cost or raises the performance of a downstream firm's production, a process innovation for that firm, which in turn makes a new product technically or economically feasible. Feedback runs the other way: demand for more computationally intensive capabilities, such as models that improve answers by spending more computation at the time of use, presses for improvement in hardware, infrastructure and inference methods [3]. Every observation is therefore classified by its origin, its meaning for the adopting firm and its downstream consequence.

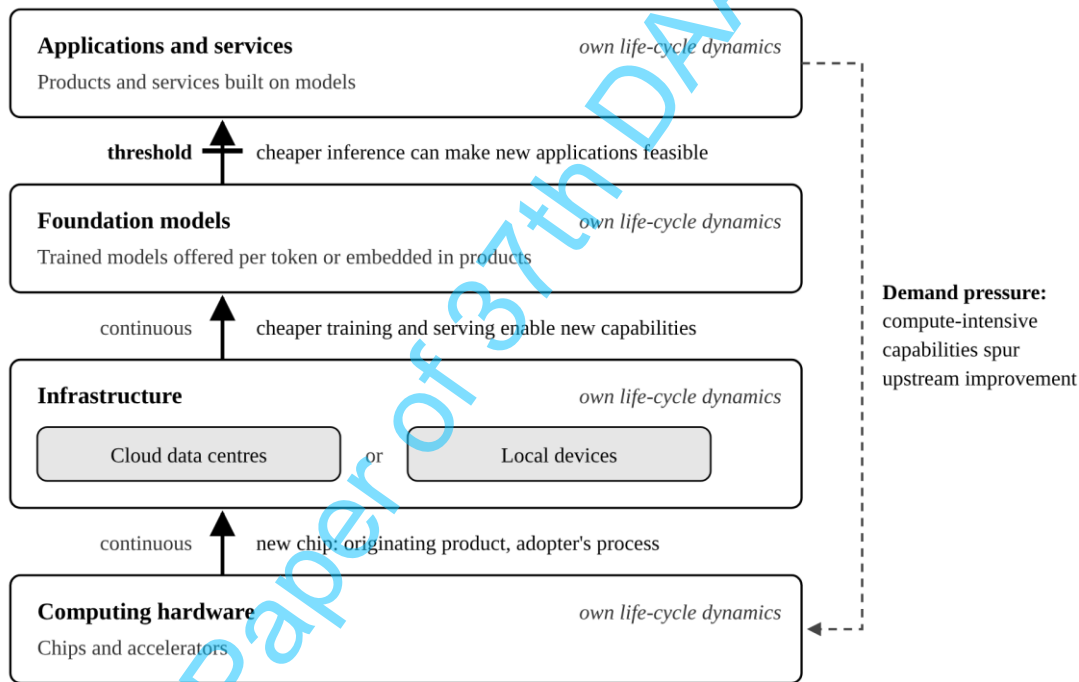


Fig. 1. Conceptual framework: four layers, within-layer cycles and cross-layer links

The link between enabling and enabled innovation is an application-specific feasibility threshold. A capability may be feasible in engineering terms yet commercially impractical because of its computational cost, latency, energy requirements or the infrastructure it needs. Process improvements relax one or more of these constraints; once the binding constraint passes a critical level, a previously impractical product becomes viable. An application that needs extensive computation at the time of use, for example, may be technically possible long before it is economic. The threshold differs by application: there is no single price per token at which "AI applications" become viable. The mechanism runs in four steps:

1. Upstream technological innovation in hardware, algorithms or infrastructure.
2. Downstream process improvement: lower effective cost, latency or resource requirements.
3. A feasibility threshold is crossed for a specific application.
4. Downstream product innovation: new or substantially improved products and services.

Crossing a previously binding threshold can enable product innovation but does not ensure its adoption, and not every product innovation depends on such a threshold: some create value through new functionality, accuracy or user experience in applications that were already feasible. Even when computational costs become acceptable, complementary organisational, technical and market conditions may constrain implementation, as documented for AI in EU manufacturing [21]. The mechanism should therefore be most visible where computation is a large share of an application's cost and such barriers are low.

Two pathways follow. In a product-led pathway, an originating innovation introduces a new capability within a layer, which raises demand for efficiency improvements; process innovation then lowers its delivery cost and widens its commercialisation. In a process-enabled pathway, improvements in enabling layers lower the effective cost, latency or resource requirements of downstream providers, and application-specific feasibility thresholds mediate whether new downstream products follow. The first corresponds to the classical sequence within a layer, the second to the process-first sequence and hierarchical enabling discussed above. The contribution is not the identification of technological layers or of reverse innovation sequences, both established in prior research [4], [20]. It is their integration: distinguishing product-led and process-enabled pathways, examining their interaction across layers, and treating application-specific feasibility constraints as a potential explanation of downstream commercialisation. On this view, the order of product and process innovation depends on the layer and the stage of commercialisation examined. Who captures the value created depends on control of complementary assets [22] and on the scale economies of model production [23]; these questions are treated in Section 4. Table 1 summarises the resulting propositions: P1 and P2 are examined in Section 3, while P3 guides the competitive analysis.

Proposition	Statement	Evidence to examine	What would contradict it
P1: Cross-layer enabling innovation	Upstream improvements lower downstream providers' cost, latency or resource needs, so specific applications cross feasibility thresholds	Dated falls in the binding constraint against launch dates, by application	A relaxed binding constraint with no downstream launch and no organizational barrier to explain it, or a launch credited to an improvement that came later
P2: Nested innovation cycles	Standardization and process efficiency within one layer can coexist with continued product innovation in downstream layers	Architecture convergence and price trends in one layer against new products in the next	Standardization of a layer without meaningful downstream product innovation, as an aggregate classical reading predicts [10]
P3: Changing sources of advantage (analytical)	As capabilities cheapen, the weight of differentiation, efficiency and complementary assets shifts across layers	Price dispersion among models of similar quality; concentration by layer	Not tested; guides Section 4

Table 1. Analytical propositions

3. Product and Process Innovation in the Artificial Intelligence Industry

3.1. Method

The study is a conceptual analysis supported by a theory-guided qualitative examination of the generative AI industry, using secondary evidence. It assesses whether P1 and P2 are consistent with observable developments; it does not test them statistically.

Four application categories were selected because each has a plausible binding constraint: latency, context capacity, inference-time computation and device memory. A fifth case, ChatGPT, was chosen because the framework is expected not to explain it, so that competing explanations are examined rather than assumed away. For each case we establish the timeline of upstream improvements, the development of the relevant cost or resource requirement, and the launch of commercial products, and we look for direct evidence that the constraint actually limited deployment.

Evidence comes from three source types, kept distinct in the text: independent measurement (the Stanford AI Index and the datasets it draws on), firms' own technical announcements and documentation, and reporting in established business and technology press. List prices are treated as evidence of what customers paid, not of production costs, and firms' claims about their own products are attributed as claims. Launch dates alone are not taken as proof that a threshold was crossed.

The cases are purposively selected illustrations, not a representative sample: they show different relationships between capabilities, resource constraints and commercialisation, but cannot show how often each relationship occurs across the industry. Each case is classified on two separate dimensions: the innovation pathway (product-led, process-enabled, mixed or undetermined) and the strength of evidence. Evidence is strong for the product-led pathway when the introduction of a new capability is documented and followed by identifiable process improvements, and for the process-enabled pathway when a constraint and an improvement that relaxes it are both documented and linked to downstream commercialisation. It is suggestive when a documented constraint is followed by a relevant improvement but the causal link is uncertain, and

amounts to consistency only when developments merely coincide in time. These classifications describe how well the proposed sequence is documented; they do not establish causal relationships.

3.2. The cost trend and its unevenness

The price of a fixed level of capability has fallen steeply. According to the AI Index, querying a model that matches GPT-3.5 on the MMLU benchmark cost about USD 20 per million tokens in November 2022 and USD 0.07 in October 2024, a more than 280-fold fall [3].

Two qualifications matter for the framework. These are list prices, so they show what customers paid, not what serving costs providers; and the frontier did not get cheaper in the same way. The AI Index reports that a reasoning model, OpenAI's o1, was nearly six times more expensive than GPT-4o and took about 40 times longer to produce its first token (29.7 s vs 0.72 s), in exchange for much higher accuracy on hard problems [3]. A later study, available as a preprint, finds the same pattern for 2024–2025: prices at a given performance level fell about 5–10-fold a year, while the cost of running the best-performing models on benchmarks rose [24]. Falling costs for a fixed capability therefore coexist with rising costs for new capabilities, which is the feedback relationship in Fig. 1.

A third distinction follows from the application-specific nature of thresholds. For a firm adopting AI, the relevant measure is not the price per token but its cost per completed task, which can be written as an accounting identity:

$$C_{task} = \sum_i T_i P_i + C_{other} \quad (1)$$

where T_i is the quantity of billable units of type i used to complete the task (input, output and cached tokens, and other metered resources), P_i their unit price, and C_{other} other task-level costs such as integration and human oversight. This is the adopter's expenditure; for a provider, the same prices are revenue, and its production costs depend on hardware, energy and utilisation. Changes in task cost depend jointly on unit prices, the quantities of billable resources consumed and other application-specific costs. Declining token prices therefore do not necessarily translate into lower costs per completed task: reasoning models raise T_i , and later process innovation can lower it again (Case C).

3.3. Application cases

Case A: real-time voice assistants. The binding constraint was latency. Until May 2024, ChatGPT's voice mode chained three models (speech-to-text, a language model, text-to-speech), with average response times of 2.8 seconds using GPT-3.5 and 5.4 seconds using GPT-4. GPT-4o, a single model trained end-to-end across audio, vision and text, responds in 320 milliseconds on average, close to human response time in conversation [25]. Developer access followed in October 2024 at USD 100 and 200 per million audio input and output tokens [26]. In December 2024 OpenAI cut GPT-4o audio prices by 60%, citing its own efficiency improvements, and added a smaller model priced at a quarter of the reduced rates [27].

The latency threshold was crossed by an originating product innovation within the model layer, a new architecture, not by upstream process improvement (product-led; strong evidence). Efficiency improvements reported by the provider then lowered the price of the capability for other firms (process-enabled; suggestive evidence).

Case B: long-context document processing. Here the capability preceded its economic feasibility. In February 2024 Google announced Gemini 1.5 Pro with a context window of up to one million tokens, but released it with a standard 128,000-token window. It described the full window as computationally intensive, offered it to early testers free with longer latency, and stated that optimisations to reduce latency and computational requirements were still under way [28]. The provider's own account documents that compute and latency, not technical capability, limited deployment.

The same provider then introduced a process innovation aimed at the economics of long contexts. In June 2024, when it opened a two-million-token window to all developers, Google noted that input costs grow with the context and launched context caching for Gemini 1.5, which processes repeated content once and reuses it [29]. Caching lowers the cost of recurring long-context tasks, such as repeated questions over the same documents; it does not by itself enable longer contexts. Anthropic introduced a comparable feature in August 2024, reporting cost reductions of up to 90% and latency reductions of up to 85% for long prompts, and naming Notion among early adopters [30]. This is suggestive evidence of the process-enabled pathway. Constraint and improvement are documented by the same provider, but the case shows better economics for an existing capability more clearly than a previously infeasible product becoming viable.

Case C: reasoning-intensive applications. This is the feedback case. OpenAI's o1 reasoning models, first released as a preview in September 2024, raised accuracy on an International Mathematical Olympiad qualifying exam (AIME 2024) from 9.3% for GPT-4o to 74.4%, at nearly six times the price and about 40 times the time to first token [3]. Product innovation increased the computation required per task, raising T_i in (1), rather than waiting for costs to fall.

Two forms of process innovation followed. In December 2024 OpenAI reported that its updated o1 used on average 60% fewer reasoning tokens than the preview version [27]. In January 2025 DeepSeek released R1 with open weights, reporting performance comparable to o1 on mathematics and coding benchmarks from a model that activates only part of its parameters for each token [31]; its list prices were a small fraction of o1's. Similar performance at lower prices indicates competitive convergence through alternative training and inference approaches; it does not by itself establish imitation. Within about four months, an originating product innovation was followed by cost-reducing process innovation, by the

originator and by a competitor (product-led; strong evidence for the sequence, although prices across providers say little about production costs, subsidies or pricing strategy).

Case D: on-device assistants. Firms publicly withheld features for lack of device resources. In late 2023 Google shipped its on-device model Gemini Nano on the Pixel 8 Pro but not on the Pixel 8, which uses the same processor but has 8 GB of memory instead of 12 GB. A Google vice-president explained that the company did not want to degrade the experience on the phone with less memory; it reversed the decision in March 2024 and offered the model to Pixel 8 users as a developer preview [32]. In June 2024 Apple limited its Apple Intelligence features to its newest phones, and its head of AI strategy stated that older devices would run the models too slowly to be useful [33].

The case is strong evidence that resource constraints shape which products are available on which devices, at the infrastructure layer where the cloud-versus-local choice is made. It also shows the feedback arrow, with AI features raising the memory requirements of devices. It is weaker evidence of a threshold-crossing event: Google's reversal may reflect a change in acceptable performance, feature scope or deployment strategy rather than a documented optimisation. The evidence is strong that a constraint was binding, but amounts to consistency only for the process-enabled pathway, since the improvement that relaxed it is not documented.

Case E: ChatGPT, the counter-case. ChatGPT was launched free on 30 November 2022 on a model from OpenAI's GPT-3.5 series; a closely related model had been released to API customers two days earlier [34]. ChatGPT reached an estimated 100 million monthly users in January 2023, the fastest ramp of any consumer application [35]. This study identifies no documented, preceding reduction in cost or latency that would explain the launch. The step change came from a conversational interface offered free, which the threshold mechanism does not explain (product-led; strong evidence).

The sequence after launch is instructive. OpenAI reported that system-wide optimisations had cut the cost of ChatGPT by 90% since December, and in March 2023 it offered the same model to developers at a price ten times lower than its earlier GPT-3.5 models. Snap, Quizlet, Instacart and Shopify launched or announced products built on it [36]. The first wave was product-led; the second wave, products by other firms, is consistent with P1 (process-enabled; suggestive evidence).

3.4. Assessment

The cases suggest two innovation pathways (Table 2). In Pathway A, product-led, a new capability appears first; demand for efficiency follows, and process innovation then broadens commercialisation. In Pathway B, process-enabled, an efficiency improvement relaxes a binding constraint and a new downstream product follows. Pathway A corresponds to the classical sequence within a layer [1]; Pathway B to the process-first sequence of Barras [4] and the hierarchical enabling described by Funk [6]. The two can overlap and reinforce each other, as in Cases A, C and E.

Case	Binding constraint	What relaxed it	Pathway	Strength of evidence
A Voice	Latency, then price	New architecture (product), then reported efficiency gains	Mixed: product-led, then process-enabled	Strong, then suggestive
B Long context	Compute, latency, input cost	Context caching by the same provider (process)	Process-enabled	Suggestive
C Reasoning	Inference-time compute	Fewer reasoning tokens; competitors' more efficient models	Product-led	Strong
D On-device	Device memory	Not documented	Undetermined	Strong for the constraint; consistency only for its relaxation
E ChatGPT	None evident, then price	Interface; then system-wide optimization	Mixed: product-led, then process-enabled	Strong, then suggestive

Table 2. Application cases: innovation pathway and strength of evidence

The evidence points to a refined central finding. Product-led and process-enabled pathways coexist and interact across technological layers. Originating product innovations introduce new capabilities, while subsequent improvements in computational efficiency and deployment economics facilitate diffusion and downstream application development. Application-specific feasibility constraints plausibly mediate these relationships, but the available secondary evidence shows associations and plausible enabling mechanisms, not precise causal thresholds; no case documents both a binding constraint and the specific improvement that removed it.

For P2, the widely adopted transformer architecture and falling quality-adjusted prices coexisted with new downstream products in all five cases. This supports coexistence only; it neither shows whether the rate of downstream innovation changed nor establishes a dominant design in the Abernathy–Utterback sense.

4. Implications for Competitive Strategy

4.1. Competitive advantage along the two pathways

The two pathways call for different sources of advantage. On the product-led pathway, technological leadership creates a lead, but the cases suggest it can be short-lived: within about four months of the first o1 release, a competitor reported comparable benchmark performance at a fraction of the price (Case C). Rapid capability convergence may weaken the ability of model developers to sustain advantages based on technological leadership alone. Under such conditions, control of complementary assets, including computing infrastructure, distribution channels, proprietary data and customer relationships, may become more important for appropriating returns from innovation [22]; an innovator that also controls such assets can capture much of the value itself. Investment in these assets is very large: Alphabet, Amazon, Meta and Microsoft forecast combined capital expenditure of about USD 650 billion for 2026, much of it for data centres [37]. Together with the high fixed costs of model development, such investment can raise entry barriers and favour concentration in the model and infrastructure layers [23]; even before generative AI, a few large firms dominated both upstream computing and downstream use of AI [20]. It can also produce excess capacity and uncertain returns, so barriers to entry should not be equated with profitability.

On the process-enabled pathway, advantage may shift to firms that recognise when an application's binding constraint is about to be relaxed and build for that point. This resembles the entry window around the emergence of a dominant design, which in disk drives mattered more than first-mover advantage [11]. The ability to deliver a task at lower cost per completed task, rather than per token, may then support competitive performance in the application layer where application quality and customer value are comparable, in line with Klepper's argument that larger, more efficient producers gain most from process innovation [10]. Where they differ, quality, reliability, differentiation and switching costs also shape competitive performance. Declining quality-adjusted prices may intensify competitive pressure in standardised model services, but their effect on providers' profitability depends on how prices, production costs, utilisation and scale evolve relative to one another.

4.2. Implications for firms adopting artificial intelligence

For firms that use AI rather than produce it, three implications follow. First, the relevant cost is the cost of the task, as in (1), not the quoted price per token; reasoning-intensive applications can cost more per task even as token prices fall. Second, the choice between cloud and local deployment should follow both the application's binding constraint and its economics. Cloud services give access to frontier capabilities with little upfront investment but recurring usage charges; local deployment requires investment in computing resources and may offer lower incremental costs for additional inference workloads, depending on utilisation, energy consumption and maintenance, and can meet requirements for latency and data privacy within the limits of device memory and computing resources (Case D). Declining resource requirements may therefore widen the range of applications for which local deployment is economically attractive, without necessarily displacing cloud infrastructure. Third, crossing a technical or economic threshold does not remove organisational barriers such as missing skills, poor data quality and difficult integration [21]; firms that build these capabilities early are better placed to adopt an application once its costs become acceptable.

4.3. Conditional scenarios

Three scenarios illustrate how the two pathways could shape the industry's evolution. Each is conditional, and the historical analogies identify mechanisms rather than predict that AI will repeat another industry's path (Table 3). The electricity analogy, for example, concerns standardised enabling infrastructure supporting varied downstream applications, not a claim that AI will become an undifferentiated utility; electricity itself developed through extensive technical and institutional change.

Scenario	Dominant pathway	Historical analogy	Conditions	Signals to watch
Continued capability differentiation	Product-led	Early personal computers and smartphones [2]	New capabilities appear faster than competitors converge	Persistent price and performance gaps between leading models
Commoditized model layer	Process-enabled, with value captured elsewhere	Personal computers after standardization, when value shifted to chips and the operating system	Capabilities converge; quality-adjusted prices keep falling	Narrowing price dispersion among models of similar quality; value shifting to chips, cloud and distribution
Utility with complementary products	Both, in different layers	Electricity: a standardized supply enabling successive new appliances [7], [38]	Model and infrastructure layers standardize while application costs cross successive thresholds	Stable infrastructure prices alongside a continued stream of new application categories

Table 3. Conditional scenarios for the AI industry

The observed developments are compatible with elements of all three scenarios. Continued differentiation at the technological frontier, declining prices for comparable capabilities and expanding downstream applications may coexist. Which trajectory becomes more prominent will depend on the relative pace of capability innovation, efficiency improvements, technological standardisation and the development of complementary assets.

5. Conclusion

This paper examined how the relationship between product and process innovation shapes the evolution of the AI industry. Its central argument is that AI industry evolution is better understood through interacting product-led and process-enabled innovation pathways across technological layers than by assigning the whole industry to a single stage of the classical product–process cycle. In the cases examined, originating innovations introduced new capabilities first, while subsequent process improvements lowered their cost and supported wider diffusion and downstream applications. Application-specific feasibility constraints, such as cost, latency and device memory, plausibly mediate this interaction, but the evidence is stronger for the product-led pathway than for process-enabled product creation, and does not establish a causal role for these constraints.

The findings also illuminate the relationship between the conventional model and Barras's reverse product cycle. The process-enabled pathway is particularly relevant to downstream diffusion and application development, whereas originating product innovations may follow a different sequence. This distinction is consistent with critiques emphasising that service innovation cannot be understood solely through the adoption and diffusion of technologies developed elsewhere [14].

For managers, the analysis suggests that advantages based on technological leadership alone may be short-lived where capabilities converge quickly, that control of complementary assets may matter more for appropriating returns, and that adopting firms should evaluate AI by the cost of completing a task and the binding constraint of the application, not by the quoted price per token.

The study has clear limitations. It rests on five purposively selected cases and on public, largely firm-reported sources; list prices say little about production costs; and no case documents both a binding constraint and the specific improvement that removed it. Further research should test the two pathways systematically, for example by relating dated series of quality-adjusted prices and latency to the launch of new application categories across a broad sample, and should extend the analysis of the infrastructure layer, where the economics of cloud and local deployment are changing.

5. References

- [1] Abernathy, W. J. & Utterback, J. M. (1978). Patterns of industrial innovation, *Technology Review*, Vol. 80, No. 7, (June/July 1978) 40-47, ISSN 0040-1692
- [2] Markovic, M[ilivoj]; Draskovic, N[ikola] & Gnjidic, V[ladimir] (2018). Product innovation, process innovation and competitive lessons from consumer electronics industry, *Proceedings of the 29th DAAAM International Symposium*, pp.0004-0008, B. Katalinic (Ed.), Published by DAAAM International, ISBN 978-3-902734-20-4, ISSN 1726-9679, Vienna, Austria. DOI: 10.2507/29th.daaam.proceedings.001
- [3] Maslej, N.; Fattorini, L.; Perrault, R. et al. (2025). The AI Index 2025 Annual Report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA. Available from: <https://hai.stanford.edu/ai-index/2025-ai-index-report> Accessed: 2026-10-13
- [4] Barras, R. (1986). Towards a theory of innovation in services, *Research Policy*, Vol. 15, No. 4, 161-173, ISSN 0048-7333, DOI: 10.1016/0048-7333(86)90012-0
- [5] Murmann, J. P. & Frenken, K. (2006). Toward a systematic framework for research on dominant designs, technological innovations, and industrial change, *Research Policy*, Vol. 35, No. 7, 925-952, ISSN 0048-7333, DOI: 10.1016/j.respol.2006.04.011
- [6] Funk, J. L. (2007). Technological change within hierarchies: The case of the music industry, *Economics of Innovation and New Technology*, Vol. 16, No. 1, 1-16, ISSN 1043-8599, DOI: 10.1080/10438590600661582
- [7] Bresnahan, T. F. & Trajtenberg, M. (1995). General purpose technologies "engines of growth"?, *Journal of Econometrics*, Vol. 65, No. 1, 83-108, ISSN 0304-4076, DOI: 10.1016/0304-4076(94)01598-T
- [8] Utterback, J. M. & Abernathy, W. J. (1975). A dynamic model of process and product innovation, *Omega*, Vol. 3, No. 6, 639-656, ISSN 0305-0483, DOI: 10.1016/0305-0483(75)90068-7
- [9] Anderson, P. & Tushman, M. L. (1990). Technological discontinuities and dominant designs: A cyclical model of technological change, *Administrative Science Quarterly*, Vol. 35, No. 4, 604-633, ISSN 0001-8392, DOI: 10.2307/2393511
- [10] Klepper, S. (1996). Entry, exit, growth, and innovation over the product life cycle, *American Economic Review*, Vol. 86, No. 3, 562-583, ISSN 0002-8282

- [11] Christensen, C. M.; Suárez, F. F. & Utterback, J. M. (1998). Strategies for survival in fast-changing industries, *Management Science*, Vol. 44, No. 12, Part 2, S207-S220, ISSN 0025-1909, DOI: 10.1287/mnsc.44.12.S207
- [12] Adner, R. & Levinthal, D. (2001). Demand heterogeneity and technology evolution: Implications for product and process innovation, *Management Science*, Vol. 47, No. 5, 611-628, ISSN 0025-1909, DOI: 10.1287/mnsc.47.5.611.10482
- [13] Ferràs-Hernández, X.; Nylund, P. A. & Brem, A. (2023). The emergence of dominant designs in artificial intelligence, *California Management Review*, Vol. 65, No. 3, 73-91, ISSN 0008-1256, DOI: 10.1177/00081256231164362
- [14] Gallouj, F. & Weinstein, O. (1997). Innovation in services, *Research Policy*, Vol. 26, No. 4-5, 537-556, ISSN 0048-7333, DOI: 10.1016/S0048-7333(97)00030-9
- [15] Clark, K. B. (1985). The interaction of design hierarchies and market concepts in technological evolution, *Research Policy*, Vol. 14, No. 5, 235-251, ISSN 0048-7333, DOI: 10.1016/0048-7333(85)90007-1
- [16] Eloundou, T.; Manning, S.; Mishkin, P. & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs, *Science*, Vol. 384, No. 6702, 1306-1308, ISSN 0036-8075, DOI: 10.1126/science.adj0998
- [17] Cockburn, I. M.; Henderson, R. & Stern, S. (2019). The impact of artificial intelligence on innovation: An exploratory analysis, In: *The Economics of Artificial Intelligence: An Agenda*, Agrawal, A.; Gans, J. & Goldfarb, A. (Ed.), 115-146, University of Chicago Press, Chicago
- [18] Tekic, Z[eljko]; Cosic, I[Ilija] & Katalinic, B[ranko] (2019). Manufacturing and the rise of artificial intelligence: Innovation challenges, *Proceedings of the 30th DAAAM International Symposium*, pp.0192-0196, B. Katalinic (Ed.), Published by DAAAM International, ISBN 978-3-902734-22-8, ISSN 1726-9679, Vienna, Austria, DOI: 10.2507/30th.daaam.proceedings.025
- [19] Cowling, M.; Vorley, T.; Liu, W. & Abdul Rahman, S. (2025). Are product and process innovations supermodular? Complementary returns to product and process innovations, *Economics of Innovation and New Technology*, Vol. 34, No. 8, 1247-1273, ISSN 1043-8599, DOI: 10.1080/10438599.2024.2436961
- [20] Jacobides, M. G.; Brusoni, S. & Candelon, F. (2021). The evolutionary dynamics of the artificial intelligence ecosystem, *Strategy Science*, Vol. 6, No. 4, 412-435, ISSN 2333-2050, DOI: 10.1287/stsc.2021.0148
- [21] Siaterlis, G[eorge]; Nikolakis, N[ikolaos]; Alexopoulos, K[osmas] & Makris, S[otiris] (2022). Adoption of AI in EU manufacturing. Gaps and challenges, *Proceedings of the 33rd DAAAM International Symposium*, pp.0547-0550, B. Katalinic (Ed.), Published by DAAAM International, ISBN 978-3-902734-36-5, ISSN 1726-9679, Vienna, Austria, DOI: 10.2507/33rd.daaam.proceedings.077
- [22] Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy, *Research Policy*, Vol. 15, No. 6, 285-305, ISSN 0048-7333, DOI: 10.1016/0048-7333(86)90027-2
- [23] Korinek, A. & Vipra, J. (2025). Concentrating intelligence: Scaling and market structure in artificial intelligence, *Economic Policy*, Vol. 40, No. 121, 225-256, ISSN 0266-4658, DOI: 10.1093/epolic/eiae057
- [24] Gundlach, H.; Lynch, J.; Mertens, M. & Thompson, N. (2025). The price of progress: Price-performance and the future of AI, *arXiv preprint arXiv:2511.23455v2* (revised 23 March 2026), Available from: <https://arxiv.org/abs/2511.23455> Accessed: 2026-10-13
- [25] OpenAI (2024). Hello GPT-4o, Available from: <https://openai.com/index/hello-gpt-4o/> Accessed: 2026-10-13
- [26] OpenAI (2024). Introducing the Realtime API, Available from: <https://openai.com/index/introducing-the-realtime-api/> Accessed: 2026-10-13
- [27] OpenAI (2024). OpenAI o1 and new tools for developers, Available from: <https://openai.com/index/o1-and-new-tools-for-developers/> Accessed: 2026-10-13
- [28] Pichai, S. & Hassabis, D. (2024). Our next-generation model: Gemini 1.5, Available from: <https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/> Accessed: 2026-10-13
- [29] Kilpatrick, L.; Basu Mallick, S. & Kofman, R. (2024). Gemini 1.5 Pro 2M context window, code execution capabilities, and Gemma 2 are available today, *Google Developers Blog*, 27 June 2024, Available from: <https://developers.googleblog.com/en/new-features-for-the-gemini-api-and-google-ai-studio/> Accessed: 2026-10-13
- [30] Anthropic (2024). Prompt caching with Claude, Available from: <https://www.anthropic.com/news/prompt-caching> Accessed: 2026-10-13
- [31] DeepSeek-AI (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, *arXiv preprint arXiv:2501.12948*
- [32] Amadeo, R. (2024). Google says running AI models on phones is a huge RAM hog, *Ars Technica*, 29 March 2024, Available from: <https://arstechnica.com/?p=2013446> Accessed: 2026-10-13
- [33] Yip, J. (2024). Apple exec explains why you wouldn't want to run Apple Intelligence on your older iPhone, *Business Insider*, 18 June 2024, Available from: <https://tech.yahoo.com/phones/articles/apple-exec-explains-why-wouldnt-170230615.html> Accessed: 2026-10-13
- [34] Wiggers, K. (2022). While anticipation builds for GPT-4, OpenAI quietly releases GPT-3.5, *TechCrunch*, 1 December 2022, Available from: <https://techcrunch.com/2022/12/01/while-anticipation-builds-for-gpt-4-openai-quietly-releases-gpt-3-5/> Accessed: 2026-10-13

- [35] Hu, K. (2023). ChatGPT sets record for fastest-growing user base – analyst note, Reuters, 1 February 2023, Available from: <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/> Accessed: 2026-10-13
- [36] OpenAI (2023). Introducing ChatGPT and Whisper APIs, Available from: <https://openai.com/index/introducing-chatgpt-and-whisper-apis/> Accessed: 2026-10-13
- [37] Day, M. & Bang, A. (2026). Big Tech to spend \$650 billion this year as AI race intensifies, Bloomberg News, 6 February 2026, Available from: <https://www.bloomberg.com/news/articles/2026-02-06/how-much-is-big-tech-spending-on-ai-computing-a-staggering-650-billion-in-2026> Accessed: 2026-10-13
- [38] David, P. A. (1990). The dynamo and the computer: An historical perspective on the modern productivity paradox, American Economic Review, Vol. 80, No. 2, (May 1990) 355-361, ISSN 0002-8282

Working Paper of 37th DAAAM Symposium