

Metrics, not Myths: The True Price and Value of Productivity in the Era of AI

A systematic framework to determine AI-induced productivity gains with regard to the “AI paradox”

Sibylle Kunz, Claudia Hess

IT & Engineering

IU International University of Applied Sciences

Erfurt, Germany

e-mail: sibylle.kunz@iu.org; claudia.hess@iu.org

Abstract— Many organizations tout productivity gains when deploying Analytical and Generative Artificial Intelligence (AI), yet few offer structured methods to quantify real value. This article presents a framework to classify use cases by time and cost savings or quality improvements, and links productivity to profitability for calculating measurable outcomes. Beyond headline gains, the article uncovers hidden costs—such as the need to validate, verify, and re-work AI outputs—that can erode expected benefits and impact true efficiency. A comprehensive Total-Cost-of-Ownership (TCO) approach addresses not only acquisition and operation of AI systems, but also the ongoing labor associated with refining AI-generated results. Comparing both overt and covert costs against claimed benefits, the framework enables more accurate estimations of AI’s impact. Demonstrated with examples from diverse industries, the article provides practical guidance for more transparent assessment of AI projects and their real-world returns.

Keywords—artificial intelligence; productivity; TCO; hidden costs; AI paradox.

I. INTRODUCTION

With the increasing use of Artificial Intelligence (AI), the number of publications and press releases proclaiming the economic success of this technology is also rising, often without concrete evidence or verifiable figures. In many cases, only percentage improvements are estimated. This makes it difficult to compare use cases and identify best practices, as Kunz and Hess [1] explain. To address this issue, they propose a framework for assessing the benefits and costs associated with implementing AI solutions.

Productivity gains resulting from analytical AI, such as in medical diagnosis or fraud detection, have been discussed for several years now. AI can solve complex problems that would otherwise need much more time, resources or would be intractable due to the sheer volume of data involved. Since the introduction of Open AI’s Large Language Model (LLM) GPT3.5 to the general public in November 2022 and the emergence of many competing models, the number of news articles and publications that postulate even higher expectations concerning the use of Generative AI (GenAI) in the form of LLMs has multiplied.

For example, a McKinsey study conducted in 2023 [2] proclaimed an estimated world-wide productivity gain induced by GenAI of 2.6 to 4.4 billion dollars and a rise in

working productivity of 0.1% to 0.6% per year. According to this study, branches benefitting the most from GenAI will be finance, high-tech, media and bioscience. 75% of this potential can be found in the field of customer service, sales and distribution, software development and research & development. The study evaluated 850 different jobs and 2.100 different job tasks. The study also distinguished between three clusters of task types: Physical work (foreseeable or unforeseeable) with productivity gains of 70% resp. 34% (slightly more than using analytic AI only), data collection (79%, 65% by analytical AI only) and data management (92%, 75% by analytical AI only) and decision making and collaboration (management, stakeholder communication and interaction, knowledge application) with an increase of 50-55% in productivity – almost three times as much as with analytical AI only. That the amount of productivity gains seems correlated with the level of education comes as a surprise – employees with lower qualifications saw less possibilities to raise their productivity.

There are further national and international studies that aim to quantify the economic effects of AI and its impact on functions and job profiles. In 2024, a study conducted by ifo among German companies [3] revealed that almost 84% of them expected productivity gains for the national economy within five years, estimating an average increase of 12%. But 70% of all managers predicted productivity gains of averagely 8% for their own company, thereby estimating lower values for their company than nationwide.

Hammermann et al. [4] report that 45% of all employees in the 815 companies included in their study, who have used AI in their daily work between 2022 and 2024, claim productivity gains in their own job. On the other hand, 15% of the employees using AI stated the opposite.

Demary et al. [5] arrive at a moderate assessment: their study conducted for the Institute of the German Economy (IW) estimates a rise of the gross domestic product by 0.9% that can be derived from AI-usage between 2025 and 2030, so it does not act as a strong growth driver. For the decade starting in 2030, an increase of 1.2% is predicted. Demary et al. see AI as complementary to human work. They also raise the question whether rises in productivity can be accomplished by AI alone or only if AI is flanked by other digitalization technologies like robotics, software, internet access etc.

The saving of processing time is often cited as a reason for productivity gains. Looking at these ambitious expectations, AI seems to be a must-have for companies to stay competitive. However, AI solutions and their integration in a company's way of working often come with considerable costs and sometimes ethical implications. Before making decisions on investment, organizations must be able to assess the expected productivity gains and construct a fact-based calculation to make sure that AI-investments will be amortized and do not cause any uncontrollable risks, especially with regard to reductions in the workforce as a consequence of productivity gains. It seems insufficient to just ask employees about their general personal impression of the extent to which productivity has increased when using (generative) AI, especially when it is unclear whether this can be attributed to the use of AI alone.

According to the 29th CEO survey of PriceWaterhouse Coopers [6], only 12% of all enterprises using AI manage to realize both cost savings and rising revenues at the same time. PWC interviewed 4,454 CEOs in 95 countries. Many CEOs could not see direct measurable results from their AI activities. 38% of all US companies reports lower costs, 51% of all Chinese enterprises said they had increased their revenue. In Germany, 16% saw lower costs, 11% stated revenue increases. But all in all, more than half of the respondents saw no identifiable benefits at all.

Meta, Facebook's parent company, goes one step further: a new performance tracking tool called "Checkpoint" is used to link employee performance reviews to AI tool usage at work [7]. The management wants staff to integrate artificial intelligence into daily tasks, aiming for greater efficiency and innovation. Guidelines for using AI across different functions have been circulated, and employees are expected to automate routines, support decision-making, and enhance content creation via these tools. Performance evaluations will partially reflect each employee's AI adoption, marking a cultural shift within the company and promising significant bonuses for successful AI adoption.

All of these results are in line with the so-called "AI Productivity Paradox" highlighted by Brynjolfsson et al. [8], which observes that despite widespread adoption of AI technologies, measurable productivity improvements often lag behind investments and expectations. Brynjolfsson and co-authors emphasize that organizations frequently struggle to translate AI capabilities into tangible business gains, partly due to challenges in integration, organizational change, and the need for complementary innovations. As the PWC survey shows, the majority of enterprises are yet to realize clear benefits from AI, reflecting this paradox—the promise of AI is substantial, but the practical gains remain elusive for most firms. This affirms Brynjolfsson's "Productivity J-Curve" [9], stating that productivity does not immediately rise. Instead, it may first stall or even decline before increasing later. This is because early investments in general purpose technologies generate intangible assets (like new business processes, workforce training, and organizational redesign) that are not captured in conventional productivity metrics.

To reduce the ambiguity of statements about savings and revenue increases, a framework is needed which allows empirical AI use cases to be classified, their outcomes measured and compared. Therefore, the research question is: what kind of values can be measured and in which dimensions?

This article compares measurement dimensions and reference values for different use cases and develops a systematic framework for calculating realistic economic effects of AI usage. Instead of calculating each new AI business case from scratch, it might be helpful to have standard categories, benchmarks and a set of indicators that can be measured.

To provide a solid foundation, the concepts of productivity and profitability are clarified and differentiated in Section II. Section III shows the dimensions for measuring AI-induced productivity and profitability gains related to time, cost and quality. In Section IV, some examples and case study projects are presented to illustrate which productivity gains have already been proven and how they have been calculated. Section V describes a Total Cost of Ownership (TCO)-based approach to calculate investments needed to plan, train, implement and run an AI system. Section VI focuses on the "hidden costs of AI", resulting in an enhancement to the TCO approach and discussing the "Junior Paradox". Section VII sums up both sides of the equation: cost structure versus profitability to sketch a framework showing how to develop benchmarks in the future. Finally, Section VIII draws a conclusion and offers suggestions for future work.

II. PRODUCTIVITY AND PROFITABILITY

Due to Thommen et al. [10], productivity is defined as the quantitative relationship ratio between output and input of the production process (1):

$$\text{Productivity} = \frac{\text{Work Outcome}}{\text{Input Quantity of production factors}} \quad (1)$$

Since it is impossible to measure productivity for a whole enterprise at once, partial productivities are calculated related to work hours, machine running times or area sizes. For example, an accountant could execute 20 bookings per hour, a punching machine could produce 1.000 parts per hour, or a toy shop could achieve a turnover of 1.000 € per square meter. If the usage of AI is somehow beneficial to the company, these partial productivities can be expected to rise.

This effect can be direct or indirect: An increase in productivity per area means better utilization of limited resources and thus a direct increase in sales. Conceivable areas of application include stationary retail or agricultural production. Conversely more output from a machine or worker per unit of time means - assuming investment or labor costs remain the same (considered here as fixed costs) - that less machine running time or labor is now required for the same result, indirectly reducing costs.

Productivity is an output-oriented concept measured in physical units like pieces or amounts. Since productivity figures can usually also be expressed in monetary units, the

concept of profitability (or economic viability) can be used as a substitute for productivity in the planned framework.

Profitability can be defined in different ways. The following equation assumes a direct relationship to productivity via (2):

$$\text{Profitability} = \frac{\text{Productivity} \times \text{Revenue per Unit} - \text{Total Expenditures}}{\text{Revenue}} * 100\% \quad (2)$$

Alternatively, the concepts of the net profit ratio (after deduction of all taxes) (3)

$$\text{Net Profit Ratio (\%)} = \frac{\text{Net Income}}{\text{Revenue}} \quad (3)$$

or the gross margin ratio (after deduction of all direct costs) (4)

$$\text{Gross Margin Ratio (\%)} = \frac{\text{Gross Profit}}{\text{Net Revenue}} \quad (4)$$

can be used.

The central question is: to what extent can the use of AI reliably contribute to the productivity gains predicted by recent studies, and which investments in AI technologies are necessary to leverage this potential? Although there are currently many studies that attempt to quantify the expectations of AI use in this regard, it is difficult to verify whether these expectations will actually materialize. There is a lack of clear assignability and classification of measurement methods. The following section will therefore examine which dimensions and characteristics are suitable for operationalization.

III. DIMENSIONS FOR MEASURING AI-INDUCED PRODUCTIVITY AND PROFITABILITY GAINS

AI can be classified into different fields based on its purpose, underlying methodologies, and applications. For our framework, we will distinguish between Analytical AI, Generative AI, and Reactive AI.

- Analytical AI focuses on data-driven decision-making, pattern recognition, and predictive analytics. It processes structured and unstructured data, extracts insights, and assists in optimization and forecasting without autonomously generating new creative content. It can be used in the fields of medical diagnosis, fraud detection, predictive maintenance or algorithmic trading as well as for natural language recognition, sentiment analysis, etc.
- Generative AI is designed to create new, synthetic content, such as text, images, music, or videos, by learning patterns from existing data. It uses advanced models like Generative Adversarial Networks (GANs) that allow the creation of videos and Transformer-based architectures like GPT4.x and other Large Language Models for natural language generation.
- Reactive AI operates based on real-time inputs and predefined rules, without memory or learning from

past experiences at runtime. It was mainly used for fast-response, rule-based systems like chess-playing systems, such as “Deep Blue”, or older AI-powered, but rule-based chatbots.

Today, especially the methods in Analytical and Generative AI can be used to leverage potentials for productivity - and therefore - profitability gains. There are three dimensions to be improved: time, cost and quality. Table I shows the relation between these dimensions and whether Analytical (A) or Generative (G) AI is used. It lists examples for measures that can be used to calculate AI-related profitability increases.

TABLE I. MEASURING DIMENSIONS FOR AI PRODUCTIVITY

Dimension	Context/Scenarios	AI class
Time	Planning time Project planning, Product planning, Logistic optimization (e.g., airports, freight forwarders, harbors, railways)	A+G
	Design time Product design, Service design, Individualization of consumer goods, Developing protein structures for pharmaceutical or chemical applications (e.g., AlphaFold), Developing recipes	A+G
	Production time of physical goods (time needed to produce one piece or unit)	A
	Production time of immaterial artefacts Creation of text, audio, video, e.g., in journalism, marketing, consulting, arts, Creation of program code	G
	Testing time Creation of test cases (e.g., software testing), Automatic test execution and evaluation	A+G
	Delivery time Demand forecasting, Route optimization	A
	Support time Analyze service requests by natural language recognition, Speech-to-Text, Answer customer requests, Analyze customer feedback	G
Cost	Material usage: production factors like raw materials, supplies and energy	A
	Waste, Offcut: raw materials	A
	Required space: inventory optimization	A
Quality	Quality control and error detection (checking products, processes, or results for deviations, errors, anomalies, and risks, e.g., automatic anomaly detection in production processes, medical diagnosis, e.g., skin cancer, tumor detection in X-rays or MRIs)	A+G
	Correction and quality improvement (adjustment or optimization of existing artifacts to meet defined quality standards, e.g., proofreading of texts or translations)	A+G
	Quality of results and performance (AI elevates the quality of the end result or human performance beyond what was originally possible, e.g., easier readability, improved accessibility of a website, personalized content)	A+G

With regard to GenAI, it needs to be pointed out that time and cost might tend in a different direction than quality, i.e., gains in the first two dimensions might cause reductions in

the third. Consider this example: It is often stated that GenAI helps save a lot of time in producing text and illustrations. Employees working with text, such as journalists or marketers, can produce results in a shorter time or reduce the effort necessary to check translations or edit articles. And this means that one person can create a larger amount of output (i.e., text or graphic elements) in a given time, therefore getting more things done for the same salary. But in this case, productivity gains are more difficult to estimate, since the quality or originality of results is also important for the artefacts. Just speeding things up might lead to counterproductive effects in the long run. This effect is discussed in more detail in Section VI.

IV. CASE STUDY EXAMPLES FOR SUCCESSFULLY MEASURING PRODUCTIVITY GAINS

The two following tables show some examples for case studies in the fields of Analytical and Generative AI that documented concrete absolute or relative values for productivity gains. In most cases, percentages or absolute values for savings were mentioned, but none of them listed any data on the cost side as described in Section V. Table II lists projects using mainly analytical AI.

TABLE II. CASE STUDIES USING ANALYTICAL AI TO INCREASE PROFITABILITY OR SAVE COSTS.

Company and branch	What was measured?	Scope	Relative absolute change in productivity / reported savings
Salling Group, Energy consulting [11]	Cost: Energy consumption in supermarket buildings via smart meters or data from energy providers	AI system analyses weather data and energy consumption and optimizes usage of device during closing hours	700 supermarkets were evaluated, savings in the millions are reported
Municipality of Holstebro, Denmark [12]	Cost: Energy consumption in community buildings	AI system analyses weather data and energy consumption and optimizes usage of device during closing hours	Savings: 1 million DKK = ca. 146.000 \$
SWMS Systemtechnik Ingenieur-gesellschaft mbH [13]	Cost: Automated production of composite materials using a printing robot arm: usage of 3D printing materials	Reduction of printing errors through AI-supported monitoring: image-based object recognition and segmentation.	Material savings of 1/3 (estimated), Savings in energy for robot and cooling of printed artefacts

Orthopedical insoles [14]	Cost: Material usage in 3D printing instead of insole construction using blanks	AI system calculates ideal form to realize material saving	Material savings: >70% in plastic materials, up to 60% in energy
FRAPORT AG, Aviation [15]	Time: staff are assigned to ground handling based on qualification and availability	AI system IDA simulates and optimizes staff planning	(no data yet, project in beta-status)

Table III shows some examples of projects using GenAI. Productivity gains here often mean that standard tasks can be automated, so that the employee gets more time for more difficult tasks.

TABLE III. CASE STUDIES USING GENERATIVE AI TO INCREASE PROFITABILITY OR QUALITY OR SAVE COSTS.

Company & Branch	Scope	What was measured?	Change in productivity / reported savings
heise (IT magazine publisher, online portal provider) [16], [17]	System heiseIO (based on LLM and a process-oriented approach with predefined prompts)	a) Time: Annotation time of 14.000 pictures in a CMS b) Time: Production time of a newsletter	a) 5-15min human-based annotation time saved per picture, total cost of 300 € Daily "Botti" newsletter can be produced 12 minutes faster, saving a total of 1,5 person days per moth
Fieldcode GmbH [18]	System Fieldcode (based on LLM) planning of field service assignments	Time: Ticket diagnosis: analysis of field service requests to solve service problems remotely instead of sending a service technician, optimize "first fix rate" Cost: avoid unnecessary order of spare parts, avoid unnecessary travel cost, fuel etc.	Up to 50% of spare parts could be saved, Rise of First fix rate (no exact number given)
Klarna [19]	System Kiki (based on LLM) used	Time, Quality: The system answers up to	Contracts can be generated in 10 minutes

	for internal knowledge management	2.000 questions from employees per day and is used by 85% of all staff	instead of 1 hour Chatbot performs 2,3 million chats with customers (equals work of 700 employees) Estimated profit increase: 40 million \$
Boston Consulting Group [20]	General usage of AI	Time, Quality	12% more tasks accomplished, 25% savings in time 40% increase in quality

The two tables show that the case studies can easily be categorized in terms of the dimensions mentioned in Section III. Table IV shows an overview of all case studies mapped to these dimensions.

TABLE IV. ASSIGNMENT OF CASE STUDIES TO MEASURING DIMENSIONS AND AI CLASS

Case Study	Dimension	AI class
Salling Group	Cost	A
Municipality of Holstebro	Cost	A
SWMS Systemtechnik Ingenieur-gesellschaft mbH	Cost	A
Orthopedical insoles	Cost	A
Fraport	Time	A
Heise	Time	G
Fieldcode GmbH	Time, Cost	G
Klarna	Time, Quality	G
Consulting	Time, Quality	G

The next section discusses how different AI-system investments can be made comparable using a lifecycle.

V. COST OF AI-USAGE: A TOTAL COST OF OWNERSHIP-APPROACH FOR AI-SYSTEMS

Before productivity gains can be realized, AI-based systems need to be developed, trained and fine-tuned. To get an overall picture of the total cost of ownership (TCO) of an AI system, the following cost components need to be considered. They can be organized along the life cycle of an AI system, using a TCO-like approach with lifecycle phases and corresponding tasks as shown in Fig. 1.

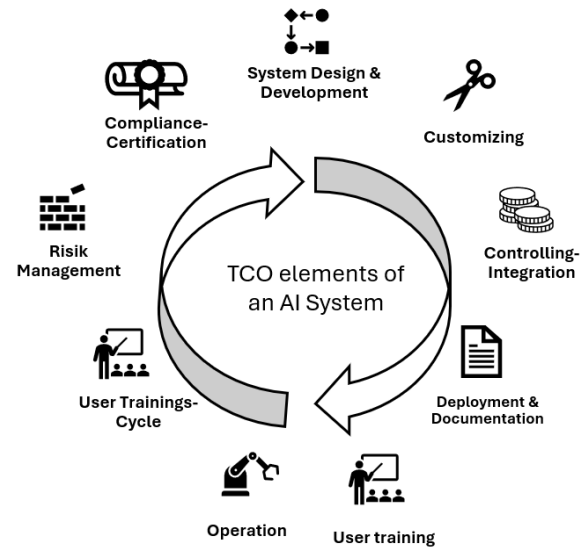


Figure 1. TCO Elements of an AI System

The following cost components need to be evaluated:

A. System Design and Development

All AI systems need to be modelled and trained. This can either be done using supervised or unsupervised learning, reinforced and/or deep learning. Developing these systems requires a large amount of storage and compute resources like GPUs or other AI-chips. Training data needs to be collected or artificially generated, and the data needs to be cleaned and consolidated. Interfaces to control and monitor the settings are needed and the system may have to be integrated into existing processes. This results in a significant fix-cost block before installing the final system. In the case of prefabricated AI models, this cost block will be covered by later subscription fees.

The cost determinants in this phase are especially the programming time, data engineering time (calculated via salary), on-premises hardware or cloud cost for CPU/GPU time, and energy cost.

B. Customizing

Especially in the case of GenAI, the resulting models need to be finetuned to perform specific tasks or to improve their performance in a particular domain. Guard rails need to be developed to ensure responsible and ethical AI use.

The cost determinants are salary (programmers and domain experts), acquisition of domain-specific data sets, and computing resources.

C. Integration into Controlling

Key Performance Indicators (KPIs) and performance metrics need to be developed in order to monitor the system and to assess its performance. This is a prerequisite to ensure that the AI system meets the desired outcomes

The main cost determinant is the time required to choose and to agree upon KPIs and appropriate metrics with AI engineers, domain experts and executives being involved.

D. Deployment and Documentation

The AI system needs to be deployed within the existing infrastructure, which involves technical setup, integration with systems and platforms already in use, and testing. Documentation facilitates maintenance, troubleshooting and further development.

The cost determinant is the time spent by the technical team.

E. User training

Users need to be initially trained in how to use the AI systems (e.g., required by the EU AI act). The training ensures employees understand AI opportunities, risks, and legal compliance requirements. The curriculum also depends on what tasks the employees are assigned to, e.g., whether they work in the IT department, in a dedicated AI team, human resources etc.

The cost determinants are salary, course fees, course material, travelling and accommodation costs.

F. Operation

Each request to the running system causes a certain amount of costs for inferencing, i.e., producing a solution or an answer. These costs are often covered using subscription fees like user licenses for LLMs on a monthly or per-token basis. For critical systems, human workers need to be kept in the loop to meet ethical requirements.

The cost determinants are subscription fees (fix costs on monthly/annual basis), token consumption (variable cost), salary.

G. User Training Cycle

After certain intervals, these training courses need to be repeated to keep users up-to-date and refresh their knowledge.

The cost determinants are salary, course fees, course material costs, travel and accommodation costs.

H. Risk Management

Installing AI systems in certain processes might also require additional insurance, e.g., to protect against potential damage caused by AI system failures or monitoring frameworks to assess and mitigate associated risks.

The cost determinants are salary and insurance premiums.

I. Certification for Compliance

Certifications might be required regarding compliance with legal and regulatory standards, which can vary by industry and region (e.g., the AI Act by the European Union). They must be renewed in prescribed intervals. Renewal processes typically involve reassessment and auditing of systems to confirm that they still meet the necessary requirements.

The cost determinants are certification fees on an annual or long-term basis and salaries for the internal and external experts involved in the certification.

This structure can be used to calculate a concrete AI project, resulting in the estimated total costs of the system. It can further be used to calculate the time needed to amortize.

But there is another aspect not directly related to the application lifecycle, but to the quality of AI-based output, which is discussed in the following section.

VI. THE “HIDDEN COST” OF AI

If increases in productivity or profitability have been documented and proven using this framework, a second consideration is needed. Does the output quality match the original expectations? The effort required to ensure this quality introduces a crucial, often underestimated, or even overlooked layer of human work.

A systematic **AI Output Rework** process is the critical quality gate that transforms a promising AI draft into a trustworthy and valuable business asset. This section presents a formal framework for this human-led validation, outlining the essential activities needed to mitigate risks and ensure that all AI outputs meet the highest standards of quality and integrity.

A. The AI Productivity Paradox Revisited

AI output rework is especially important for GenAI-generated content like documentation, communication with customers, suppliers and stakeholders or program code. This quality inspection and the potential rework effort require time and, as a result, increase staff costs.

A recent study conducted and published in January 2026 by Workday [21] showed that 77% of all AI users check and correct the output GenAI tools have given them, investing more time and doing it more thoroughly than if a human had created the original output. According to this study, almost 40% of AI time savings are lost, because the output quality is too low and needs to be fixed, resulting in another trade-off between velocity of production and quality. Only 14% of AI-experienced employees report sufficient quality that needs no fine-tuning. Thus, increased personal costs negate parts or all of the savings.

This is another aspect of the AI paradox: productivity gains come with a grain of salt. Using AI-generated results that have not been validated means tolerating certain risks. This might be acceptable in some scenarios, when no greater harm can arise from the outcome (e.g., proofreading or spam filtering) but is not applicable to AI-systems classified carrying higher risks.

B. The AI Output Rework Process

As raw AI results, whether creative text or complex data analysis, inherently carry risks of inaccuracy, bias, and inconsistency, their quality needs to be checked. It must be ensured that AI-generated content and analyses are accurate, reliable, and aligned with organizational standards. Rework can comprise the following critical quality assurance checks,

with the first two checks focusing on GenAI output and the third on analytical AI.

1) *Core Quality Assurance*: This check focuses on the fundamental correctness and usability of the output. For GenAI, this involves quality-checking for a creative artifact (text, image, audio, video, code) through a quality assurance process that is similar to editing a human-created artifact. The key activities include:

- **Accuracy & Factual Integrity**: This step focuses on verifying the objective truthfulness and reliability of the generated content and correcting factual errors. As LLMs can “hallucinate”, this human-in-the-loop review is critical, especially for high-stakes outputs like in stakeholder communication (e.g., when handling a customer complaint) or when the generated output is part of the product to be sold (e.g., in a publishing house). The specific checks involve several key actions. First, reviewers must perform hallucination detection by identifying and flagging any statements that are sounding plausible but factually incorrect or entirely fabricated. Next, data and source verification is required to validate all claims, data points, names, and dates against reliable sources. Finally, reviewers must assess contextual appropriateness, ensuring the information is not only valid but also suitable for its specific use case, such as when announcing layoffs or company restructuring versus writing a marketing post. All identified issues must be addressed.
- **Linguistic & Stylistic Quality**: This step ensures the text is not just grammatically correct but also clear, well-written, and aligned with the intended tone and audience. Although LLMs are not prone to making spelling or punctuation errors, they might use wrong expressions, technical terms or translate metaphors in a wrong way. The review process improves clarity and fluency by correcting awkward phrasing and robotic language, while also verifying terminology and nuance to ensure technical terms are used precisely and metaphors are handled appropriately. Beyond simple correctness, the tone of voice is then carefully adjusted to align with a specific communication style, such as formal, empathetic, or witty, to match the desired brand identity.

2) *Risk & Compliance Audit*: While the first checks focus on whether the output is correct and well-made, this check asks whether the output is safe, ethical, and legally sound. This is a critical risk management step that protects the organization, its brand, and its stakeholders from potential harm. It includes the following key activities:

- **Bias & Representation Check**: The review process should include a careful audit for social biases to ensure the content is fair and inclusive. This involves checking both text and images for harmful stereotypes related to e.g., gender, race, nationality, age, or

disability. Reviewers should also identify and remove exclusionary language that presumes a certain type of audience. For visual content, this means ensuring fair representation and actively avoiding stereotypical depictions, such as consistently portraying certain professions with a single demographic.

- **Safety & Harmful Content Audit**: The output should be checked for any content that could be dangerous, unethical, or offensive. This could involve, for example, a practical check for any material that could damage the company’s reputation or violate workplace conduct, such as off-brand messaging, or inappropriate humor.
- **Legal & Compliance Review**: To mitigate legal risk, the output should undergo a compliance review to ensure it does not create liabilities for the organization. This should include a check for copyright infringement and plagiarism, as well as a critical review for privacy and confidentiality, ensuring no sensitive company data is revealed. Finally, for industry-specific applications, it might be necessary to check the output for regulatory adherence, confirming it does not violate rules such as giving unqualified financial advice or making unapproved medical claims.

3) *Model-Specific Checks*: In the case of analytical AI, many quality checks need to be done before and during the model’s use in the business process. These tasks were already discussed in Section V, in the context of the integration into the controlling. Additional quality-checking is required within the business process when non-technical employees (e.g., a loan officer, a marketing manager, a support agent) interact with the analytical AI’s output as part of their daily job. Such a quality check could look like as follows: an employee first receives a real-time judgment from the analytical AI, which is typically embedded in their workflow software. This output, often supported by a simple explanation for transparency, is then reviewed and validated by the user, who applies their own expertise and contextual knowledge to determine if the AI’s suggestion is sound. Based on this validation, the user makes the final decision: they can either accept the AI’s recommendation, override it if their judgment differs—which constitutes the primary form of rework in this context—or escalate the case for further review.

C. Social and Ethical Implications: The “Junior Paradox”

According to [22], this corrective work is mainly carried out by younger employees aged 25-34, who are expected to develop the necessary competencies and a critical attitude towards AI-generated output and coping strategies for checking it. But only 37% of them receive special skills training for this. And often these expectations are not yet mirrored in updated role descriptions. This phenomenon is aligned with the so-called “junior paradox” [6],[22],[23]:

Since many classical tasks in research, programming or documentation that used to be fulfilled by junior roles can now be done by AI, companies start hiring less employees into junior positions. This makes it harder for beginners to start into their first job. At the same time, the former junior employees find themselves confronted with higher expectations more related to tasks for senior positions (since “AI can do the basics”), but they haven’t yet had the chance to build up the necessary skills. Meanwhile, employees in the senior positions face rising workloads, leaving less time for mentoring juniors.

D. Digital Darlings: LLMs Are Their Own Biggest Fans

In a radically different approach, AI-tools could be used on a meta-level to rework flawed AI-results. For example, results of one LLM could be fed into another, asking it to check and, if necessary, correct them. This approach is often referred to as “LLM-as-a-judge” [24]. The second LLM, acting as the evaluator or critic, does not address the user’s original query but receives a specialized directive. This includes the first model’s output and provides an explicit evaluation schema, instructing the evaluator to systematically scrutinize the text for a range of potential deficiencies. Consequently, the critic model operates as a runtime quality assurance mechanism. Its purpose is to programmatically detect and rectify common failure modes inherent in LLMs.

Potentials and limitations of the LLM-as-a-judge approach have been discussed for various application areas like healthcare or programming assistants [25][26]. By eliminating the need for manual intervention, such a tool could significantly reduce the resources allocated to correcting AI-generated output. Despite this promise of efficiency, critical concerns remain regarding its practical implementation. It remains unclear how possible biases might interfere here once more. Laurito et al. [27] found that AI-tools tend to prefer AI-generated text over human-generated text. This could lead to self-reinforcing hallucinations. And sometimes there are other unexpected forms of biases, e.g. the so-called “First-Item Bias”, when LLMs tend to choose the first of two options more often, independent of their content. [28].

E. The extended TCO-Approach

Therefore, the framework should be expanded to identify, trace and control these quality-induced rework costs as well. Which turns out to be difficult, since employees need to be trained to recognize errors in AI output and remember to track the time needed to correct them. This requires an adaption of standard work processes and continuous monitoring of quality improvement activities and an open discussion culture to address this issue. But the data thus gathered allow for many benefits concerning cost monitoring. Fig. 2 shows the enhanced TCO cycle.

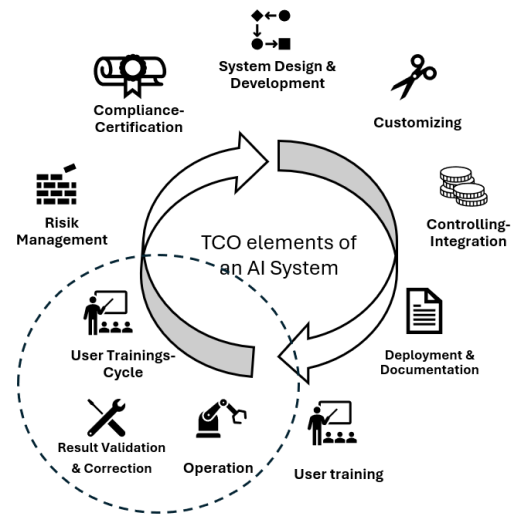


Figure 2. TCO elements including result validation and correction

VII. THE INTEGRATED FRAMEWORK FOR COMPARING COSTS AND PRODUCTIVITY/PROFITABILITY GAINS

To be able to reliably quantify productivity and profitability gains, the two concepts explained in Sections III and V must now be combined as illustrated in Fig. 3. The sum of all profitability increases can be calculated by measuring the difference (Δ) between time, material cost or quality level and multiplying it with the proper computing unit like salary/hour or price/unit. Predicting the benefit of a rise in quality is more difficult to calculate. In a medical environment for example, it could be measured by the follow-up costs of the treatment of a patient who was not correctly diagnosed but would have been using AI-techniques – or by future purchases of a customer who is more satisfied than before.

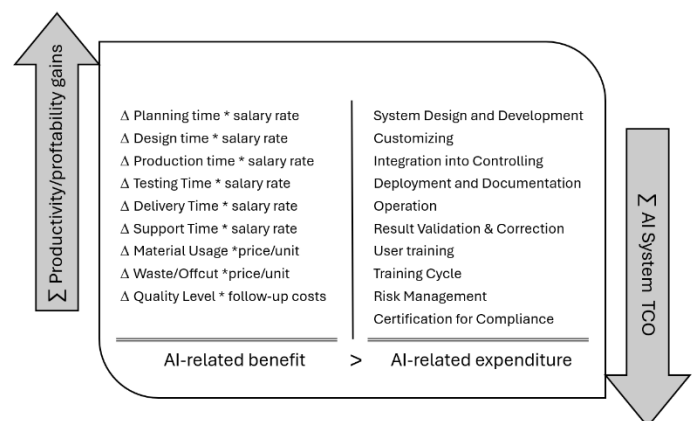


Figure 3. Comparing the benefits of productivity/profitability gains and AI system costs

To calculate a value for an AI-system TCO, more data is needed, for example resulting from past projects and continuous controlling. Companies might search for

benchmarks and share experiences. As AI system components become more standardized and included in “software off the shelf”, this will become much easier to accomplish.

Finally, the two sums or at least their order of magnitude can be compared arriving at a first judgment whether the benefits outweigh the costs and if so, by what amount. This can prevent companies from running blindly into AI projects that will not be paying off because too many aspects remain unnoticed before the start.

VIII. CONCLUSION AND FUTURE WORK

Although many companies claim remarkable benefits of AI usage concerning productivity, it is still difficult to find exact numerical proof or compare use cases across branches. Each use case is evaluated on its own, and often only savings, but no cost dimensions are reported. In addition, AI systems are rarely looked at from a TCO-based angle with regard to the whole lifecycle.

Therefore, in this article a framework for measuring and evaluating productivity and profitability gains induced by using analytical or generative AI systems was developed. A volume structure was developed for the beneficial effects, considering time, cost and quality. In addition, AI system cost is structured alongside a TCO approach. Finally, both sides are compared to gain a clearer view on quantitative aspects, which has to be enriched with qualitative aspects like human-AI-cooperation or ethical implications. Integrating these perspectives, the framework can help foster a cautious judgement whether the proclaimed benefits stand on real ground.

Future work should include the following:

A systematic literature review should be conducted focusing on collecting and categorizing case studies in different industries to gather as much real data as possible. Categorization should include branches, company size, geographical region, type of AI used and governance limitations in force.

A database with benchmark data should be compiled using the results of the literature review. Data donations from interested companies should be integrated. Institutions like chambers of commerce, industry associations and practical research institutions like universities of applied sciences can help with gathering this data and transferring it into practice.

A questionnaire for measuring the single components of the framework should be developed, resulting in a form where companies can enter their specific data to get a first estimation of benefits and cost.

The phenomenon of the “junior paradox” should be empirically investigated to identify its underlying mechanisms, contextual boundary conditions, and short- and long-term consequences for both individuals and organizations. Empirical findings should be translated into practical recommendations for job starters, employers, and educators to help mitigate potential negative effects of the junior paradox.

REFERENCES

- [1] S. Kunz and C. Hess, “Artificial Intelligence – Myth or Measurable? A systematic framework to determine AI-induced productivity gains”, Proceedings of Advances on Societal Digital Transformation (DIGITAL 2025), IARIA, July 2025, pp. 1-6, ISBN: 978-1-68558-285-2. [Online]. Available from: https://www.thinkmind.org/library/DIGITAL/DIGITAL_2025/digital_2025_1_10_18001.html [retrieved 02, 2026].
- [2] McKinsey & Company, „Auswirkungen von GenAI auf den deutschen Arbeitsmarkt. Eine Chance zur Linderung des Fachkräftemangels“, 2023. [Online]. Available from: https://www.mckinsey.de/~media/mckinsey/locations/europe/%20and%20middle%20east/deutschland/news/presse/2023/2023-11-24%20genai%20implikationen%20deutschlands%20arbeitsmarkt/mckinsey_analyse_genai_arbeitsmarkt_deutschland_vf.df [retrieved: 02, 2026].
- [3] A. Kerkhof, T. Licht, M. Menkhoff, and K. Wohlrabe, „Die Nutzung von Künstlicher Intelligenz in der deutschen Wirtschaft“, ifo Institut - Leibniz-Institut für Wirtschaftsforschung an der Universität München, München, Vol. 77, Iss. 08, pp. 39-43, ifo Schnelldienst, ISSN 0018-974X, 2024.
- [4] A. Hammermann, R. Monsef, and O. Stettes, „Produktiver mit KI? Wie Unternehmen und Beschäftigte die Produktivitätseffekte einschätzen“, IW-Trends, vol. 51, No. 4, pp. 75-94, 2024.
- [5] V. Demary, M. Grömling, C. Kestermann, M. Scheufen, S. Seele, O. Stettes, M. Trenz, “Gutachten: Wie wird KI die Produktivität in Deutschland verändern? Institut der Deutschen Wirtschaft Köln e.V.”, 2025. [Online]. Available from: https://www.iwkoeln.de/fileadmin/user_upload/Studien/Gutachten/PDF/2025/Gutachten_2025-Produktivitaet%20KI-barrierefrei.pdf [retrieved: 02, 2026].
- [6] PriceWaterhouseCoopers GmbH, “Global CEO Survey. Leading through uncertainty. Zwischen Unsicherheit und Neuanfang im KI-Zeitalter“, 2026. [Online]. Available from: <https://www.pwc.de/de/ceosurvey/2026/pwc-29-ceo-survey-report.pdf> [retrieved: 02, 2026].
- [7] R. Ahrens, “Meta knüpft Mitarbeiterleistung an KI-Nutzung”, heise online, 04.02.2026. [Online]. Available from <https://www.heise.de/news/Meta-knuepft-Mitarbeiterleistung-an-KI-Nutzung-11165199.html> [retrieved 02, 2026]
- [8] E. Brynjolfsson, D. Rock, and C. Syverson, “Artificial Intelligence and the modern productivity paradox: a clash of expectations and statistics”, Working Paper 24001, National Bureau of Economic Research, 2018. [Online]. Available from <http://www.nber.org/papers/w24001> [retrieved 02/2026].
- [9] E. Brynjolfsson, D. Rock, & C. Syverson, “The Productivity J-Curve: How Intangibles Complement General Purpose Technologies”, American Economic Journal: Macroeconomics, 13(1), pp. 333-372, 2021.
- [10] J.-P. Thommen, A.-K. Achleitner, D.U. Gilbert, D. Hachmeister, S. Jarchow, and G. Kaiser, “Allgemeine Betriebswirtschaftslehre“, 10th edition. Springer Gabler, 2023.
- [11] ento.ai, “Salling-Gruppe nutzt KI, um Millionen beim Energieverbrauch zu sparen”. [Online]. Available from: <https://www.ento.ai/de/cases/salling-group-reduces-energy-consumption-by-more-than-eur-2-million> [retrieved: 02, 2026].
- [12] ento.ai, “Holstebro Kommune: Energieeinsparungen dank AI-basiertem System“. [Online]. Available from: <https://www.ento.ai/de/cases/holstebro-municipality-saves-energy-with-ai-based-system> [retrieved: 02, 2026].
- [13] Green-AI Hub Mittelstand, “KI-basiertes Monitoring von 3D-Druck”. [Online]. Available from: <https://www.green-ai-hub.de/pilotprojekte/pilotprojekt-swm> [retrieved: 02, 2026].

- [14] Green-AI Hub Mittelstand, „KI-gestützter 3D-Druck orthopädischer Einlagen“. [Online]. Available from: <https://www.green-ai-hub.de/pilotprojekte/pilotprojekt-herges> [retrieved: 02, 2026].
- [15] I. Grünewald, „KI optimiert Flugzeugabfertigung am Frankfurter Flughafen. Die Fraport AG führt mit dem "Intelligent Dispatching Agent" (IDA) ein KI-System ein, das die komplexe Koordination der Flugzeugabfertigung erleichtern soll“, 28.02.2025. [Online]. Available from: <https://www.heise.de/news/KI-Update-Deep-Dive-KI-optimiert-Flugzeugabfertigung-am-Frankfurter-Flughafen-10299459.html> [retrieved: 02, 2026].
- [16] B. Danneberg, „KI-Prozess in heise I/O automatisiert Metadaten-Generierung“, 29.01.2025. [Online]. Available from: https://www.linkedin.com/posts/benjamin-danneberg_ki-digitalisierung-effizienz-activity-7290268305125220352-a9kC?utm_source=share&utm_medium=member_desktop&rcm=ACoAABrr_P4BhHMBauT9XKqlGi1ldWjXoaDtZMo [retrieved: 02, 2026].
- [17] I. Grünewald, „KI-Update Deep-Dive: KI in der Heise Gruppe“, 17.01.2025. [Online]. Available from: <https://www.heise.de/news/KI-Update-Deep-Dive-KI-in-der-Heise-Gruppe-10245348.html> [retrieved: 02 2025].
- [18] Green-AI Hub Mittelstand, „LLM-gestützte Analyse von Tickets im Außendienstmanagement“, [Online]. Available from: <https://www.green-ai-hub.de/pilotprojekte/pilotprojekt-fieldcode> [retrieved: 02, 2025].
- [19] R. Maeder, „Wie viel Geld spart Klarna durch Künstliche Intelligenz ein“, 04.06.2024. [Online]. Available from: <https://www.moneytoday.ch/news/wie-viel-geld-spart-klarna-durch-kuenstliche-intelligenz-ein> [sic!] [retrieved: 05, 2025].
- [20] F. Dell'Acqua et al., „Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on KnowledgeWorker Productivity and Quality“, Working Paper 24-013, Harvard Business School, 2023.
- [21] Workday Inc., „Beyond Productivity:Measuring the Real Value of AI. A new blueprint for the next era of AI-fueled organizational growth“, 2026. [Online]. Available from https://investor.workday.com/files/doc_news/New-Workday-Research-Companies-Are-Leaving-AI-Gains-on-the-Table-2026.pdf [retrieved: 02, 2026].
- [22] L. Szabados, "AI at Work: Promises, Perils and Paradoxes of Adoption", PhD, Soproni Egyetem, 2025. [Online]. Available from <http://doktori.uni-sopron.hu/id/eprint/961> [retrieved 02,2026]
- [23] Zoom AI, „The Junior Developer Is Dead. Long Live the Junior Developer.“. [Online]. Available from <https://pub.towardsai.net/the-junior-developer-is-dead-long-live-the-junior-developer-ad1355a8f7e7>, 23.01.2026 [retrieved 02, 2026]
- [24] L. Zheng, W.L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, „Judging LLM-as-a-judge with MT-bench and Chatbot Arena“, Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23), Article 2020, pp. 46595–46623, 2023. [Online]. Available from <https://dl.acm.org/doi/10.5555/3666122.3668142> [retrieved 02,2026]
- [25] A. Szymanski, N. Ziems, H. A. Eicher-Miller, T. J.-J. Li, M. Jiang, and R. A. Metoyer, „Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks“, Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25), pp. 952–966, 2025. <https://doi.org/10.1145/3708359.3712091>.
- [26] R. Wang, J. Guo, C. Gao, G. Fan, C. Y. Chong, and X. Xia, „Can LLMs Replace Human Evaluators? An Empirical Study of LLM-as-a-Judge in Software Engineering“, Proc. ACM Softw. Eng. 2, ISSTA, Article ISSTA086 (July 2025). <https://doi.org/10.1145/3728963>.
- [27] W. Laurito, B. Davis, P. Grietzer, T. Gavenčiak, A. Böhm, and J. Kulveit, „AI-AI bias: Large language models favor communications generated by large language models“, Proc. Natl. Acad. Sci. U.S.A. 122 (31) e2415697122, 2025, <https://doi.org/10.1073/pnas.2415697122>.
- [28] C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang, „Large language models are not robust multiple choice selectors“. Conference paper at International Conference on Learning Representations (ICLR) 2024. [Online]. Available from <https://arxiv.org/abs/2309.03882> [retrieved 02,2026].