A Holistic Evaluation Framework for Innovation and Technology Commercialization

Joining technical readiness with commercial viability

John Francis, Partner · July 2026 · 2026 edition

The Technology Readiness Level framework measures technical progress alone. That is insufficient for judging the commercial viability of emerging technologies, particularly in deep tech. This paper presents a refined evaluation framework that adds six commercial metrics to technical readiness: Time to Market, Return on Investment, Capital Needs, Competitive Landscape, Process versus Platform, and the Market Size to Capital Investment Ratio. This 2026 edition makes the instrument executable and then tests it against its own record. The weighting scale, the constraint that weights sum to one, and the aggregation rule are specified, and raw sums out of thirty are abandoned. Six technologies are scored on all six metrics and four narrative cases are restored, which surfaces three unfavorable findings: one metric returns the same score in all six cases and so carries no information, the framework ranked autonomous vehicles last shortly before that category reached commercial service, and the Competitive Landscape metric scores differentiation but not cost leadership, which led the framework to advise against the firm that won its market. A further section states where the framework gives the wrong answer, including its systematic bias against the capital-intensive technologies deep tech investors exist to fund, and its proper use is accordingly routing rather than screening. The prospective example, an AI assurance venture, carries an explicit disclosure of the publisher’s interest.

CONTENTS

Abstract

The Technology Readiness Level (TRL) framework, initially developed by NASA, is widely used to assess the technical maturity of innovations, and its emphasis on technical progress alone is insufficient for evaluating commercial viability in deep technology sectors. This paper presents a refined evaluation framework that integrates commercial metrics alongside technical readiness. The framework adds six metrics to TRL: Time to Market, Return on Investment, Capital Needs, Competitive Landscape, Process Improvement versus Platform Development, and the Market Size to Capital Investment Ratio.

This 2026 edition differs from earlier ones in four respects, each intended to make the framework an instrument a reader can execute rather than a description of one. First, the scoring is made reproducible: the weighting scale is defined, weights are required to sum to one, the aggregation rule is stated, and raw sums out of thirty are abandoned because a raw sum silently asserts equal weights. Second, six technologies are scored across all six metrics, and four narrative cases containing the framework's only forward-looking claims are restored, with the results examined for what they reveal about the instrument. That examination produces three unfavorable findings: one metric returns the identical score in all six cases and therefore carries no information; the framework ranked autonomous vehicles last among six candidates shortly before that category reached commercial operation; and the Competitive Landscape metric, as defined, scores differentiation but not cost leadership, which caused the framework to advise against the manufacturer that went on to win its market. Third, a section states where the framework gives the wrong answer, including its systematic bias against the capital-intensive, long-horizon technologies that deep tech investors exist to fund. Fourth, the prospective worked example, an AI assurance venture, carries an explicit disclosure of the publisher's interest in the institution that example depends on.

The framework enables investors, corporations, mission-driven organizations, and government bodies to make better-informed decisions about commercialization prospects. Its central use is not to accept or reject a venture but to route it to the class of capital whose weighting makes it fundable.

Introduction

The Technology Readiness Level framework is a standardized tool for evaluating the maturity of technologies, from early concepts at TRL 1 to fully deployed systems at TRL 9, and government agencies and investors have adopted it widely. TRL measures technical advancement, however, and ignores the commercial factors that determine whether a technology succeeds in a market. Without measures of market readiness, capital needs, and competitive position, the TRL scale alone gives an incomplete assessment, and the gap is largest in emerging sectors such as artificial intelligence, quantum computing, and biotechnology (Phaal, O'Sullivan, Routley, Ford, & Probert, 2011).

Scholars and practitioners have documented this limitation, and the core finding is that technical maturity does not predict market adoption or investment success (Blank, 2014). Google Glass is the classic example, because the product was technologically mature and failed in the market on privacy concerns and poor consumer engagement (Van Krevelen & Poelman, 2010). The Pfizer-BioNTech COVID-19 vaccine shows the opposite case, where technical readiness converged with commercial urgency and the product reached global scale in record time (Lurie, Saville, Hatchett, & Halton, 2020).

Investment Readiness Level

Steve Blank (2014) proposed the Investment Readiness Level to complement TRL, quantifying a startup's market viability through customer validation, revenue models, and scalability. The framework is valuable and limited in a specific way: it focuses on startup maturity rather than on the technology, and it presupposes that technologies can pivot quickly to meet market needs. That presupposition fails wherever the underlying science sets the timeline, as it does in quantum computing and biotechnology, because a company cannot pivot faster than its physics. IRL also gives little weight to capital intensity, which is decisive for technologies requiring large upfront investment (Blank, 2013). Tata Motors' Nano illustrates the residual gap, since the product was technically feasible and investment-ready and failed because its pricing strategy misaligned with consumer perceptions of what a cheap car signified (Mukherjee, 2021).

Stage-Gate methodology

The Stage-Gate methodology divides product development into stages with defined milestones, which allocates resources efficiently and reduces risk at each step (Cooper & Sommer, 2016). Like TRL and IRL it emphasizes technical milestones over market dynamics, and its structure can slow innovation in fast-moving markets while favoring incremental improvements over disruptive ones. BYD's electric buses show the counter-case, because BYD bypassed staged development, integrated battery production with vehicle manufacturing, and scaled globally ahead of competitors following rigid stage-based models (Pritajaya, 2024).

The literature since 2024

Recent work confirms the shift toward multi-dimensional readiness assessment. De Jong (2025) proposes Ethics Readiness Levels modeled on TRL, arguing that the appropriate form of ethical engagement should change as a technology matures, and demonstrates the idea on quantum technologies. Practitioner surveys now catalog more than a dozen readiness-level variants beyond TRL, including Commercial Readiness Levels, Manufacturing Readiness Levels, Integration Readiness Levels, and dual-track market-and-technology models (ITONICS, 2025). A 2026 industry white paper compiles the case against single-dimension TRL, reporting that a large majority of innovation failures trace to market misalignment rather than technical inadequacy, and documenting KTH Royal Institute of Technology's six-dimension Innovation Readiness Level model, which treats TRL as one dimension among six (IP.com, 2026). Those figures come from a vendor publication citing underlying academic studies secondhand, so this paper uses them as directional support rather than as verified point estimates.

One correction to the previous edition belongs here. That edition stated that a systematic search found no new peer-reviewed critique of TRL or Stage-Gate aimed specifically at deep-tech investment decisions published since 2024. That sentence overstated a null result. The search did surface candidates, including a Springer article on fusing TRLs with safe-by-design approaches, a Taylor and Francis article charting innovation through readiness levels, and a University of Portsmouth working paper critiquing stage-gate practice. None could be retrieved: they were paywalled or returned unreadable content, and none was read. The accurate statement is therefore that no new critique of this kind was verified, that at least three candidates exist and remain unexamined, and that this is a limitation of the review rather than a finding about the literature. Readers weighing the novelty claim of this framework should discount it accordingly until those three are read.

Toward a refined evaluation framework

Given the limitations of TRL, IRL, and Stage-Gate, a refined framework must integrate technical readiness with commercial metrics. The framework presented here adds six metrics to TRL, and the sections that follow define them, define how they combine, apply them retrospectively to six technologies, examine what that application reveals about the instrument itself, apply them prospectively to one venture, and state where they mislead.

Core problem

TRL's primary shortcoming is its singular focus on technical development. In deep tech, where long timelines and heavy capital are the norm, this leaves stakeholders without insight into market success, and technologies that score highly on TRL still fail on capital needs, regulatory barriers, or consumer resistance.

TRL lacks commercial focus. Google Glass reached high technical readiness and failed commercially (Van Krevelen & Poelman, 2010), while TRL gave no insight into market readiness, user adoption, or competitive dynamics.

IRL oversimplifies investment readiness. IRL emphasizes near-term customer validation and underweights long-term commercial risk. Pfizer's Exubera inhalable insulin would likely have passed key IRL milestones, and high scale-up capital together with poor user adoption killed the product anyway (Heinemann, 2008).

Stage-Gate is rigid and incremental. In fast innovation cycles its structure creates bottlenecks and favors safe improvements over disruptive breakthroughs (Cooper & Sommer, 2016), and BYD's rapid vertically integrated path shows what the model can miss (Pritajaya, 2024).

Deep tech falls through the gap. Technologies from national laboratories and universities need extensive regulatory approval, capital-intensive infrastructure, and long time to market, and TRL, IRL, and Stage-Gate address these factors inadequately or not at all. The mRNA vaccine platform navigated both technical and commercial challenges under exceptional demand conditions (Lurie et al., 2020), and most deep tech does not get that tailwind.

The refined framework

The refined framework integrates technical readiness with commercial viability by adding six metrics to TRL. Each metric is scored from 1 to 5 against defined criteria and anchored to a named case.

Two changes to the anchors are made in this edition, and the reasoning matters more than the substitutions.

Theranos is removed as an anchor. The previous edition used Theranos to anchor the lowest band on two separate metrics, Return on Investment and the Market Size to Capital Investment Ratio. That is wrong twice over. It conflates fraud with slow payback and with structural capital-market mismatch, which are three different failures, and a rubric scoring them identically is not measuring what it claims. More fundamentally, no readiness rubric detects fraud, because every input to every metric is a representation made by the company being scored. Assigning Theranos a 1 implies the framework would have caught it. It would not have. Fraud is a diligence problem and belongs outside the instrument.

Single cases no longer anchor multiple bands. SpaceX previously anchored the lowest band on both Time to Market and Capital Needs. One case anchoring several bands makes the bands look like restatements of a single narrative rather than independent dimensions.

1. Time to Market

Time to Market measures how quickly a technology can move from development to commercialization. Early entry captures share and compounds advantages where first-mover effects are strong, and the empirical literature supports this specifically for infrastructure and process technologies, where Feldman and Kelley (2006) find that ex ante assessment of knowledge spillovers predicts which projects attract follow-on private funding, and Caerteling and colleagues (2008) find that in road infrastructure the timing of technology development relative to procurement cycles governs adoption more than technical merit does. The metric therefore evaluates the technical timeline together with the external factors that govern speed, principally regulatory approval and supply-chain readiness.

- **5, under 1 year.** Clubhouse scaled within months on pandemic-era demand (Byers, 2021). Fit: venture capital seeking rapid penetration.
- **4, 1 to 3 years.** Slack became a major platform in about two years on a software-as-a-service model (Salyer, 2019). Fit: growth-stage and corporate venture.
- **3, 3 to 5 years.** Impossible Foods needed scientific refinement, regulatory approval, and brand building (Forbes, 2021). Fit: private equity and corporations in regulated sectors.
- **2, 5 to 7 years.** QuantumScape's solid-state batteries remained pre-commercial more than a decade after founding (Bloomberg, 2021). Fit: corporate strategic investors with patient capital.
- **1, over 7 years.** SpaceX needed more than a decade to establish reliable commercial launch (Grush, 2019). Fit: government programs and public-private partnerships.

2. Return on Investment

Return on Investment gauges how quickly and substantially a technology returns capital. Short cycles attract venture investors, while slow and stable returns fit corporate and public capital.

Two findings underpin the metric and each constrains how a score should be read. Astebro (1998) finds that returns to independent invention are extremely skewed, with most inventions returning nothing and a small minority returning very large multiples, which means a score on this metric describes the centre of a distribution whose mean is set by its tail. An evaluator who reads a 4 as an expectation rather than as a modal case has misread it. Pisano (2007) supplies the sector-specific correction: science-based businesses have persistently returned less than their scientific productivity implied, because of profound and persistent uncertainty, the difficulty of integrating knowledge across disciplines, and the slow pace at which the field learns cumulatively from its own experiments. The implication for scoring is direct. In science-based sectors, the honest score is lower than the technology's promise suggests, and the gap is structural rather than a matter of execution.

- **5, payback under 1 year.** WhatsApp reached a 19 billion dollar acquisition five years from launch on minimal overhead (Wilhelm & Crook, 2014).
- **4, 1 to 3 years.** Zoom reached profitability and large scale within two years (Business Insider, 2021).
- **3, 3 to 5 years.** Peloton's hardware-plus-subscription model took years to scale (Warren, 2017).
- **2, 5 to 7 years.** Blue Origin requires heavy upfront investment against long-term payback (Grush, 2019).
- **1, over 7 years.** Bloom Energy was founded in 2001 and did not reach public markets until 2018, with thin margins well after that. The band describes a real and honest technology whose payback period exceeds the life of most funds, which is the condition the score is meant to flag.

3. Capital Needs

Capital Needs measures total funding required to reach market, including research, manufacturing, and operational scale-up. Capital-intensive industries face structural barriers that TRL and IRL do not capture.

The empirical basis reframes what the metric is measuring. Miller and Lessard (2001), studying large engineering projects, find that such projects fail predominantly on institutional and market risks rather than technical ones, and that the projects which succeed are those where risk is deliberately allocated among parties who can bear it rather than merely estimated by the sponsor. The consequence for this metric is that a score of 1 or 2 is not primarily a statement about a dollar amount. It is a statement that the venture must assemble a coalition, and the relevant diligence question becomes which party absorbs which risk. Rivian anchoring the second band is instructive on exactly this point, since the significance of Amazon and Ford was the allocation of demand and manufacturing risk rather than the cash. Link and Siegel (2005) add the institutional dimension, finding that the mechanisms and incentives surrounding technology transfer materially affect whether research converts into growth, which means an identical capital requirement is more surmountable in a region with functioning transfer infrastructure than in one without. The metric therefore scores a venture in a context, and two ventures with the same funding requirement in different institutional settings do not merit the same score.

- **5, under 10 million dollars.** Pinterest scaled on modest seed capital (Carlson, 2012).
- **4, 10 to 100 million dollars.** Airbnb built and globalized its platform on moderate capital (Techcrunch, 2011).
- **3, 100 to 250 million dollars.** Stripe required significant capital to build financial infrastructure (Lunden, 2018).
- **2, 250 million to 3 billion dollars.** Rivian needed corporate backing from Amazon and Ford to build manufacturing (Palmer, 2021).
- **1, over 3 billion dollars.** Northvolt raised more than 10 billion dollars in equity and debt to build European battery manufacturing and entered bankruptcy proceedings in 2024. The case is the cleanest available illustration that a score of 1 on this metric describes a financing requirement most capital structures cannot carry, independent of whether the technology works.

4. Competitive Landscape

This metric evaluates how disruptive or differentiated a technology is relative to its market. Its foundation is Porter's analysis of industry structure (Porter, 1998) together with the finding that complementary assets, rather than the innovation itself, determine who captures the value an innovation creates (Teece, 1986). Teece's point is the operative one for scoring: an innovator without control of manufacturing, distribution, or an installed base will lose the returns to whoever holds them, which is why a technically superior entrant can score low here. Fleming and Sorenson (2001) supply the mechanism on the other side, finding from patent data that inventions recombining across previously unconnected technical domains are both rarer and more valuable, which is what a genuine step-change looks like in the record.

- **5, disruptive step-change.** Uber created a new market and bypassed the incumbent industry (Cramer & Krueger, 2016).
- **4, strong advantage.** Beyond Meat differentiated on product quality into a rising demand wave (Forbes, 2021).
- **3, incremental improvement.** Dropbox won users on simplicity in a crowded field (Wired, 2016).
- **2, minor differentiation.** WeWork could not distinguish its model and struggled despite scale (New York Times, 2023).

- **1, no differentiation.** Pebble was overwhelmed by Apple and Samsung and sold in distress (Fast Company, 2016). Pebble is also the clearest illustration of Teece's argument, because its product was well regarded and its competitors controlled the phone.

5. Process Improvement versus Platform Development

This metric asks whether a technology improves an existing process or creates a scalable platform. Platforms enable ecosystems and long-term growth, while process improvements deliver efficiency within existing structures, and a process improvement is viable when it is a step change against the industry's own trend line rather than a movement along it, as with the historical rate of semiconductor scaling (Intel, 1965).

- **5, highly scalable platform.** Amazon Web Services created the cloud ecosystem itself (Miller, 2016).
- **4, scalable platform within a sector.** Shopify built the e-commerce platform for small business (MacDonald, 2024).
- **3, significant process improvement.** Stripe streamlined payments integration for developers (Lunden, 2022).
- **2, incremental process improvement.** Ring improved home security within an existing category and was acquired (Garun, 2018).
- **1, no platform potential.** Jawbone failed to differentiate and liquidated (Ettiner, 2023).

This metric requires a warning that the retrospective application below makes unavoidable. It is the metric most vulnerable to optimistic self-assessment, because almost every technology can be described as a platform by a sufficiently motivated evaluator, and a score of 5 requires evidence that third parties are building on the technology rather than an argument that they could.

6. Market Size to Capital Investment Ratio

This ratio compares addressable market size with the capital required to reach it, where high ratios suit venture capital and low ratios need corporate, mission-driven, or public funding.

The previous edition left the ratio's units undefined, which made the bands unusable and produced a visible contradiction. This edition defines them. The numerator is total addressable market at a ten-year horizon, in dollars of a stated year. The denominator is total capital required to reach a defensible position in that market, in dollars of the same year, including all equity, debt, and non-dilutive funding. The ratio is dimensionless and both figures must be stated alongside it.

The definition resolves the contradiction. A reader computing Instagram's ratio from its exit value of about 1 billion dollars against roughly 57.5 million dollars of capital obtains about 17, which is nowhere near the band it anchors. That computation uses the wrong numerator. Instagram's addressable market at a ten-year horizon was mobile social advertising, plausibly in the tens of billions of dollars, and against 57.5 million dollars of capital that yields a ratio in the band. Exit value is a realized outcome and addressable market is an opportunity, and substituting one for the other changes the answer by two orders of magnitude.

- **5, above 1000.** Instagram addressed a mobile advertising market in the tens of billions on capital under 60 million dollars (Markowitz, 2012).
- **4, 500 to 1000.** Uber addressed global transportation on moderate platform capital (Cramer & Krueger, 2016).
- **3, 100 to 500.** Blue Apron faced heavy logistics investment against a limited market (Bloomberg, 2017).

- **2, 10 to 100.** Juul required extensive capital in a market constrained by regulation (New York Times, 2021).
- **1, below 10.** Solyndra raised well over 1 billion dollars, including a 535 million dollar federal loan guarantee, to manufacture thin-film solar modules whose addressable market contracted as commodity silicon module prices fell. Nothing about Solyndra was fraudulent. The capital required to reach the market exceeded a plausible fraction of the market that remained, which is exactly what a score of 1 should denote.

Making the framework reproducible

The previous edition asserted that weighting is decisive and then reported a result as a raw sum out of thirty. Those two statements are incompatible, because a raw sum is a weighted sum with equal weights. This edition specifies the mechanism so that two evaluators working from the same evidence produce the same number.

The scale. Each metric is scored as an integer from 1 to 5 against the bands above. No half points, because the bands are too coarse to support them.

The weights. An evaluator assigns a non-negative weight to each metric, and the weights must sum to exactly 1.00. The supplementary social impact metric, where used, takes its weight from the same budget rather than being added on top, which forces the evaluator to say what financial dimension it displaces.

The aggregation rule. The score is the weighted mean on the 1 to 5 scale, reported to two decimal places as a value out of 5.00, alongside the weight vector that produced it. A score reported without its weight vector is not interpretable and should be rejected.

Raw sums are abandoned. No result in this framework is reported as a number out of 30. A reader encountering such a number in an earlier edition, including the 22 of 30 in the previous version's worked example, should read it as the equal-weight case and nothing more.

Four default weight vectors follow. They are starting points for the investor classes described in the next section, not prescriptions, and any evaluator may depart from them by stating the departure.

Metric	Venture capital	Private equity	Corporate	Public and mission
Time to Market	0.25	0.15	0.10	0.05
Return on Investment	0.20	0.30	0.15	0.05
Capital Needs	0.20	0.15	0.10	0.10
Competitive Landscape	0.15	0.15	0.20	0.10
Process versus Platform	0.10	0.10	0.25	0.20
Market Size to Capital	0.10	0.15	0.20	0.10
Social impact (supplementary)	0.00	0.00	0.00	0.40
Total	**1.00**	**1.00**	**1.00**	**1.00**

The vectors encode the differences the next section describes, and they make one thing visible that prose does not: the same technology receives different scores from different classes of capital by construction, and a low score from one class is a routing signal rather than a verdict.

Applying the framework to investor types

Each investor type weights the metrics differently, and the framework makes those priorities explicit.

Venture capital. Venture investors prioritize short Time to Market, fast Return on Investment, and low Capital Needs, and they prefer platform development and disruptive competitive positions where growth supports high valuations and fast exits (McGrath, 2011; Dang & Strebulaev, 2024). They spread risk across large portfolios and accept high failure rates, which is why their weight vector concentrates on speed and capital efficiency rather than on terminal market size.

Private equity. Private equity firms accept longer timelines and higher capital needs when the business model is proven and cash flows are stable. They hold concentrated portfolios, manage risk through active ownership, and weight Return on Investment stability above everything else.

Corporate investors. Corporates value process improvements and platforms creating synergy with existing operations, and they tolerate slower returns when the technology serves long-term strategy. Rivian's partnership with Amazon shows corporate capital funding a high-capital project for strategic benefit, and the corporate weight vector accordingly puts its largest weight on the platform metric.

Mission-driven organizations and government. Public and mission-driven investors supply patient capital for high-capital, long-horizon projects with societal or strategic value, tolerating slow returns and long timelines and funding markets whose size alone would not justify the capital on financial grounds (Geuna & Muscio, 2009; Mowery, Nelson, & Sampat, 2006).

One limitation inherited from the original framework should be stated rather than quietly carried. The definitions distinguishing venture capital from private equity are not sharp, and the two vectors above differ less than the labels imply. Growth-stage venture and small-cap private equity now compete for the same companies with similar hold periods, so an evaluator choosing between those two vectors is making a finer distinction than the underlying evidence supports. The corporate and public vectors are the ones that genuinely differ.

Strategic enhancements

Five enhancements address the framework's limitations and extend its range.

1. Dynamic weighting. Evaluators assign weights by sector, development stage, and investor priority, so that an artificial intelligence application in financial services weights Time to Market and Competitive Landscape while a solid-state battery weights Capital Needs and market size. The mechanism has three parts. Its purpose is to remove the one-size-fits-all error that flat frameworks carry, since a metric decisive in one sector is noise in another. Its implementation is a weight matrix in which each evaluator records a vector per sector and stage, subject to the constraint that weights sum to one, with the vector published alongside every score so that a reader can recompute the result under different assumptions. Its benefit is that disagreement between two evaluators becomes locatable: if two analysts differ on a venture, comparing their vectors separates a disagreement about evidence from a disagreement about priorities, and only the first is resolvable by more diligence.

2. Scenario-based stress testing. Evaluators model how scores change under future conditions, including regulatory shifts, competitive pressure, and market volatility, so that a healthcare application is tested against changes in data-privacy law and an autonomous-vehicle technology against permitting regimes. Stress testing converts a static reading into a view of resilience, and the output is a set of

scores across scenarios rather than a single score with error bars, because the scenarios differ in kind rather than in confidence.

3. Social impact and strategic importance. A supplementary metric captures value beyond financial return, including emissions reduction, health access, and national security, which keeps mission-relevant technologies from being undervalued for slow returns (Geuna & Muscio, 2009; Mowery et al., 2006). Two worked illustrations show why the metric is necessary rather than decorative, and both turn on value the developer cannot capture.

Grid-scale energy storage scores poorly on the financial metrics for a reason internal to how electricity markets are organized. Most of the value storage creates accrues to the system rather than to its owner, in the form of reliability, avoided peaking generation, and deferred transmission investment, and none of those flows to the operator unless a market has been constructed to pay for them. An evaluator scoring storage on the six financial metrics alone will find high Capital Needs, slow Return on Investment, and a modest market-to-capital ratio, and will route the technology away, while the social value that justifies building it sits entirely outside the six.

Therapies for rare diseases show the same structure with the numerator inverted. A treatment for a condition affecting a few thousand patients has a small addressable market by construction, so the market-to-capital ratio is low no matter how effective the therapy is or how severe the unmet need. Orphan drug legislation exists precisely because financial metrics alone systematically route capital away from these programs, which is a legislative admission that the financial score is the wrong instrument for the decision.

In both cases the correct use of the framework is to score the financial metrics honestly, score social impact separately, and let the weight vector decide, rather than to inflate a financial metric in order to reach a conclusion the evaluator already holds. The metric draws its weight from the same budget as the others, so its use is always a stated trade rather than a free addition.

4. Tiered evaluation. A basic tier covers Time to Market, Capital Needs, and Return on Investment for early-stage ventures with limited data, and the advanced tier applies all six metrics to mature technologies, which keeps the framework usable at seed stage and rigorous at scale. A basic-tier score is reported as such and is not comparable to an advanced-tier score.

5. Case studies and benchmarks. A benchmark database anchors scoring to real outcomes, which reduces evaluator subjectivity. The next section is the first instalment of that database and it does more than illustrate: it tests the instrument against its own examples, with results that are partly unfavorable.

Six technologies, scored

Retrospective scoring of technologies whose outcomes are partly known is the only evidence a scoring rubric can offer about itself. The six applications below were first scored in the framework's 2024 edition and are reproduced here with their original scores, followed by an assessment of how each has since developed.

One caveat governs everything in this section, and it is the reason the section is a consistency check rather than a validation. The scores were assigned in 2024, when the outcomes were already partly visible. A rubric scored against outcomes its scorer could see is not being tested; it is being fitted. A genuine test requires scores registered before the outcome is known, and this framework has never been subjected to one. The value of what follows is narrower: it shows whether the rubric, applied

by a sympathetic evaluator with hindsight available, produces internally coherent and discriminating results. On two counts it does not.

1. Biotechnology: mRNA vaccines

The mRNA platform instructs cells to produce proteins that trigger immune responses, and the Moderna and Pfizer-BioNTech products compressed a development cycle that had historically run 5 to 10 years into under a year.

Scores: Time to Market 5 under pandemic conditions and 2 in normal circumstances; Return on Investment 5, on high margins with government procurement; Capital Needs 2, reflecting years of platform funding before the product; Competitive Landscape 4, since early movers held a real advantage against rising competition; Process versus Platform 5, because the platform is adaptable across diseases; Market Size to Capital 5, on global demand.

Since. The pandemic scores were correct and the platform score was optimistic. Revenue contracted sharply as procurement normalized, and the broader therapeutic pipeline the platform score anticipated has advanced more slowly than the 2021 expectation. The dual Time to Market entry, 5 under pandemic conditions and 2 otherwise, is the most useful thing in this scoring, because it records that the headline result depended on a demand condition external to the technology.

2. Fintech: blockchain-based smart contracts

Self-executing contracts on platforms such as Ethereum reduce the need for intermediaries in transactions.

Scores: Time to Market 3, since the technology matured while regulatory adaptation lagged; Return on Investment 4, on decentralized finance returns that were strong and volatile; Capital Needs 4, because development costs are low relative to banking infrastructure; Competitive Landscape 3, as first-mover Ethereum faced faster and cheaper competitors; Process versus Platform 5, on the claim that smart contracts are foundational to a decentralized web; Market Size to Capital 4, on rapid growth against low development cost.

Since. This is the framework's clearest false positive. Its equal-weight total of 23 was the third highest of the six and higher than the prospective AI assurance venture scored in the previous edition. The platform score of 5 has not been earned by the standard set out above, since the general-purpose third-party building that a 5 requires did not arrive at the scale implied, and the returns underlying the score of 4 proved cyclical rather than structural. The metric that should have caught this is Competitive Landscape, and it registered a 3.

3. Renewable energy: solid-state batteries

Solid-state chemistries promise higher energy density, faster charging, and improved safety relative to lithium-ion.

Scores: Time to Market 2, on a development cycle exceeding a decade; Return on Investment 4, on a vast potential market at high risk; Capital Needs 2, on high upfront costs; Competitive Landscape 3, because incumbent lithium-ion improvement could close the performance gap before solid-state commercializes; Process versus Platform 5, on applicability across vehicles, electronics, and grid storage; Market Size to Capital 3, since the market is large and the capital needs are also large.

Since. The scoring holds up, and the Competitive Landscape reasoning has been vindicated: incremental lithium-ion improvement has continued to erode the case for a step change, which is the

specific risk the score named. The equal-weight total of 19 is the second lowest of the six and the technology remains pre-commercial, so the rubric and the world agree.

4. Artificial intelligence: natural language processing

Advances in language models propelled applications from conversational interfaces to text analytics.

Scores: Time to Market 4, on development cycles measured in one to two years and distribution through interfaces; Return on Investment 5, on scalability across high-value applications; Capital Needs 3, since training is capital-intensive while deployment is not; Competitive Landscape 4, as differentiation remained possible among major players; Process versus Platform 5, on the ecosystem the models support; Market Size to Capital 5, on demand.

Since. This is the framework's best call and it was, if anything, conservative. It is also the case that most flatters the rubric while proving least about it, because in 2024 the outcome was already substantially visible.

5. Artificial intelligence: autonomous vehicles

Self-driving systems combine perception, prediction, and control to remove the human driver.

Scores: Time to Market 2, on long development cycles; Return on Investment 3, on high long-term returns arriving late; Capital Needs 1, on the enormous outlay required for research, fleets, and validation; Competitive Landscape 3, in a fiercely contested field; Process versus Platform 5, on the platform potential of the autonomy stack; Market Size to Capital 3, since the market is vast and the capital is too.

Since. This is the framework's clearest false negative. The equal-weight total of 17 is the lowest of the six, and autonomous ride-hail has since moved from pilot operation to paid commercial service in multiple United States metropolitan areas, which is the outcome the rubric ranked last. Section "What the scoring reveals" takes this case apart, because the reason for the miss is structural rather than a matter of a misjudged score.

6. Artificial intelligence: AI-powered drug discovery

Machine learning compresses target identification and candidate screening in a pipeline that traditionally runs 10 to 15 years.

Scores: Time to Market 3, on an accelerated discovery phase against unchanged regulatory timelines; Return on Investment 4, on research cost savings and candidate quality; Capital Needs 2, on capital-intensive infrastructure; Competitive Landscape 4, as competition rose while early movers held advantage; Process versus Platform 5, on platform breadth; Market Size to Capital 4.

Since. The equal-weight total of 22 was roughly right and the composition was wrong in one place. The Time to Market score of 3 correctly identified that regulatory duration, not discovery speed, governs the timeline, and no therapeutic whose discovery is primarily attributed to machine learning has yet completed that path to market. The Competitive Landscape score of 4, resting on durable early-mover advantage, has not held, as the sector has consolidated and several prominent independent platforms have merged or wound down. Early advantage in a tooling market erodes when the tools become available to everyone.

What the scoring reveals about the instrument

Applying the rubric to six cases produces two findings about the rubric, and both are unfavorable.

One metric carries no information

Across all six technologies, Process Improvement versus Platform Development scored 5. Six cases, one value, zero variance. A metric returning the identical score for mRNA vaccines, smart contracts, solid-state batteries, language models, autonomous vehicles, and drug discovery platforms is not distinguishing between them, and a metric that does not distinguish contributes nothing to a weighted mean except a constant.

The cause is the failure mode flagged when the metric was defined. Platform potential is an argument rather than an observation, and any technology admits the argument. The correction is to require evidence of realized third-party building for a 5, which would have scored the six cases very differently: language models and the mRNA platform would retain high scores on the strength of parties actually building on them, while autonomous vehicle stacks and solid-state chemistries, which almost nobody builds on top of, would fall to 2 or 3.

The spread of the other metrics quantifies the problem. Across the six cases, Time to Market ranged from 2 to 5 and Capital Needs from 1 to 4, both spanning three points. Return on Investment and Market Size to Capital spanned two. Competitive Landscape spanned one, from 3 to 4. Platform spanned none.

So two of the six metrics do almost no discriminating work in the framework's own examples, and this bears directly on the previous edition's prospective example, which weighted Competitive Landscape above Return on Investment. That edition up-weighted one of the two least informative dimensions in the instrument and reported the result as a strength.

The framework ranked the eventual winner last, and weighting explains why

Autonomous vehicles scored 17 on an equal-weight basis, the lowest of the six, and the category subsequently reached commercial operation. Read as a flat total, that is a straightforward failure.

Read through the weight vectors, it is not, and this is the strongest argument in the paper for the weighting mechanism. Scoring the autonomous-vehicle case under the three distinct vectors defined above gives 2.55 out of 5.00 under venture capital weights, 3.20 under corporate weights, and 3.55 under public and mission weights with a social impact score of 4. The technology is genuinely unattractive to venture capital, whose fund life cannot absorb a 1 on Capital Needs and a 2 on Time to Market, and it is materially more attractive to a corporate strategic investor weighting platform position and terminal market.

That is what happened. Autonomous driving was carried through its expensive decade principally by a corporate parent with a strategic thesis and a balance sheet, not by venture funds. The framework did not fail to identify the opportunity. Flat summation failed, by averaging away the routing information that the weighted form preserves, which is a specific and checkable argument for abandoning raw sums.

The repair is partial, and the honest statement of its limit matters. Applying the same venture vector to smart contracts gives 3.70, which ties the prospective AI assurance venture and remains the third highest score of the six. Weighting fixes the false negative and leaves the false positive intact. One of the framework's visible errors survives its own principal correction.

The full matrix, and what it says about the publisher

Scoring all six technologies under all four weight vectors completes the routing argument and produces one further finding. Social impact scores, which only the public and mission vector uses, are assigned for this illustration and are the author's judgment rather than a measured quantity: mRNA vaccines 5, smart contracts 1, solid-state batteries 4, language models 3, autonomous vehicles 4, AI drug discovery 4.

Technology	Equal weight	Venture capital	Private equity	Corporate	Public and mission	Best fit
mRNA vaccines	4.33	4.25	4.40	4.50	4.60	Public
Language models	4.33	4.20	4.40	4.50	3.85	Corporate
Smart contracts	3.83	3.70	3.80	3.95	2.85	Corporate
AI drug discovery	3.67	3.45	3.65	3.95	3.95	Corporate or public
Solid-state batteries	3.17	2.95	3.20	3.45	3.70	Public
Autonomous vehicles	2.83	2.55	2.75	3.20	3.55	Public

Four of the six best-fit assignments match how the technology was in fact financed. The mRNA platform's decisive capital was a government purchase commitment. Solid-state battery development has been carried principally by automotive corporates and public programs. Autonomous driving was carried by a corporate parent. Language model development has been financed overwhelmingly by corporate balance sheets rather than by venture funds. Smart contracts, correctly, score worst under the public vector, because no public mandate supports them. That agreement is not proof, since the vectors were constructed after the outcomes were known, and it does establish that the weighted form encodes distinctions the flat form destroys.

The further finding is uncomfortable for the publisher of this paper. The venture capital vector produces a lower score than both the private equity and the corporate vector in all six cases, without exception, and a lower score than the equal-weight mean in all six as well. A framework published by a venture capital firm therefore rates every technology in its own benchmark set as a worse fit for venture capital than for any other class of capital.

That is not an artifact of the weights chosen. It follows from what venture capital is. The vector concentrates 0.45 of its weight on Time to Market and Capital Needs, the two metrics on which deep technology scores worst by definition, and deep technology is the category this benchmark set is drawn from. The conclusion is the one the next section develops: this instrument is not a screen a venture investor should apply to deep tech, because applied as a screen it would reject the entire set. It is a routing tool, and the most useful thing it tells a venture investor is which of these opportunities requires a partner whose weighting can carry it.

Four narrative cases, and two predictions that can now be checked

Alongside the six scorings, earlier editions carried four narrative case studies built on a Context and Challenge, Investor Insight, Key Investment Takeaway structure. They are restored here because they contain the framework's only forward-looking claims, and two of those claims are now testable.

A structural warning applies to all four and should be read before them. Each is written in the form "an investor using the refined framework would have recognized" a conclusion the author already knew. That construction cannot fail. It selects the metrics that point at the known outcome, narrates them, and credits the framework with a recognition that hindsight supplied. The four cases are therefore illustrations of how the metrics are meant to be reasoned with, and they are not evidence that the metrics produce correct answers prospectively. Where a case does make a claim about an unresolved future, that claim is checkable, and the two that did are checked below.

Case 1: The Pfizer-BioNTech mRNA vaccine, alignment under urgency

Context. The mRNA platform had been in development for years with largely untested commercial potential when the pandemic created both an unprecedented opportunity and an unprecedented risk. Investors had to judge whether the platform could scale to global demand, be manufactured at volume, and clear regulatory approval on a compressed timeline.

Investor insight. Time to Market was favorable because the platform's adaptability allowed rapid repurposing and emergency approval pathways shortened the regulatory path (Lurie et al., 2020). Capital Needs risk was reduced by government procurement commitments and partnership funding, which removed the scale-up financing question that normally dominates a novel manufacturing process. Competitive Landscape was strong because the platform's applicability beyond one pathogen signalled a durable position rather than a single product.

Takeaway, and its limit. The framework directs attention to reduced capital risk against confirmed demand, which is the correct reading. The limit is that all three favorable readings derive from one external fact, which is a government committing to buy the output before it existed. That is not a property of the technology and the framework has no metric for it. An evaluator applying these six metrics to the same platform in 2019 would have scored Time to Market 2, Capital Needs 2, and Market Size to Capital 3, and would have been right to.

Case 2: Google Glass, and the 2024 prediction about Meta

Context. Google Glass was an ambitious attempt to bring augmented reality to consumers and failed on privacy concerns, social resistance, and the absence of compelling use cases. Earlier editions used it as the anchor for a forward-looking claim about Meta's augmented reality eyewear.

The 2024 claim. Competitive Landscape analysis held that Meta was entering a market still contending with privacy concerns and underdeveloped consumer use cases, and that the decisive question was whether social acceptance had been addressed. On Process versus Platform, the claim was that Google Glass lacked a developer ecosystem and that Meta's position inside a broader platform might supply one. On Time to Market, the claim was that enterprise applications offered clearer near-term returns than general consumer positioning.

What happened. The social-acceptance mechanism the framework identified proved to be the binding constraint, and the resolution was not the one the analysis anticipated. Display-bearing augmented reality eyewear has remained substantially confined to developer and demonstration hardware, while camera-and-audio glasses without a display reached a mass consumer market. The winning product removed the feature that caused the social problem.

Assessment. The framework identified the right variable and could not have produced the right answer, because none of its six metrics can represent a decision to ship less of the technology. The developer-ecosystem prediction was not vindicated in the form given, since the successful product's value did not depend on third-party applications. This is a partial hit on mechanism and a miss on outcome, and it is the closest thing in the paper's history to a genuine prospective test.

Case 3: The Tata Nano, and the 2024 prediction about Chinese electric vehicles

Context. The Nano was engineered to serve India's growing middle class with an affordable car, and its branding as the cheapest car created status perceptions that undermined it. Earlier editions drew an analogy to oversupply and price competition in China's electric vehicle market.

The 2024 claim. On Competitive Landscape, investors were advised to be cautious about companies focused on cost reduction, on the grounds that price competition erodes brand equity and that premium differentiation is required for durable profitability. On Market Size to Capital, the concern was that saturation and price war would compress margins and extend payback. On Process versus Platform, low-cost manufacturers were said to be optimizing production rather than building platforms that scale into new segments. The stated conclusion was that investors would shift focus toward premium manufacturers offering differentiation and brand loyalty.

What happened. The market-structure prediction was correct. Price competition intensified, margins compressed across the low end, consolidation began, and Chinese authorities intervened against below-cost selling. The investment conclusion drawn from it was wrong. The manufacturer that won on volume, profitability, and global expansion was the vertically integrated cost leader, while several premium-positioned manufacturers held narrow share against persistent losses. An investor following the recommendation would have moved capital away from the eventual winner.

Assessment, and an internal contradiction. This is the framework's third visible misfire, and unlike the other two it is a contradiction inside a single edition. Case 3 advises caution toward cost-driven manufacturers. Case 4, immediately following, celebrates precisely such a manufacturer for precisely that strategy, and no edition has noted that the two cases give opposite advice about the same company. The error is locatable: Competitive Landscape as defined treats differentiation as the source of advantage, following Porter, and gives no weight to cost leadership, which Porter treats as the other generic strategy of equal standing. A metric that scores only one of two viable strategies will systematically misread firms pursuing the other.

Case 4: BYD electric buses, vertical integration

Context. BYD chose to manufacture its own batteries and vertically integrate electric bus production, which reduced costs and secured supply while the industry faced battery supply bottlenecks. The question for investors was whether that bet, in a competitive market, would pay.

Investor insight. On Capital Needs, vertical integration reduced dependence on external suppliers and gave the company control of its largest input cost, which lowered capital risk in a period when battery supply was the binding constraint (Pritajaya, 2024). On Process versus Platform, BYD built a platform for electric public transport rather than improving existing bus designs, which allowed it to scale across markets. On Market Size to Capital, government adoption policies for electric transit created demand that supported the capital commitment.

Takeaway, and what it costs the framework. The case is correct on the facts and it is the strongest illustration in the set of cost leadership as a durable advantage. It also demonstrates the defect

identified in Case 3, because BYD scores well here on Capital Needs and platform reasoning while the Competitive Landscape metric, read as the framework defines it, would have penalized the same company for competing on price. The correction is to redefine Competitive Landscape to score either generic strategy on its own terms, asking whether the firm holds a defensible position rather than whether that position is differentiated. Until that redefinition is made, the metric is biased against cost leaders, and the record shows the bias is expensive.

A prospective example: an AI assurance venture in 2026

Disclosure, stated here rather than in the front matter. This example evaluates a category that Stout Street Capital has a direct interest in seeing established. The first paper in this series argues for the creation of an independent national AI assurance institution (Francis, 2026), and the third scenario below, in which the venture scores best, is the scenario where that institution exists. The publisher is therefore scoring a business whose upside case depends on an institution the publisher is publicly advocating. Readers should treat the third scenario as an interested party's estimate of an outcome it is working toward, and should note that the example was selected by the same firm that selected the framework. The example is retained because it is the clearest current illustration of a technically mature product whose entire uncertainty is institutional, and it should not be read as independent evidence for either paper.

The example company builds evaluation and audit tooling that tests deployed AI systems for bias, robustness, and performance, and prepares them for certification against recognized standards such as the NIST AI Risk Management Framework and ISO/IEC 42001.

Time to Market: 4. The product is software plus methodology, so a focused team can reach commercial deployments within one to three years, with no physical infrastructure or clinical approval gating the timeline.

Return on Investment: 3. Revenue combines recurring software licenses with audit services. Services revenue arrives early, and scalable margin arrives only to the extent that tooling substitutes for labor, which is the central risk in the business model rather than a detail of it. Payback within three to five years is achievable if and only if that substitution proceeds at a rate that has not yet been demonstrated in this category. The previous edition stated that payback in three to five years is realistic without stating the condition, which converted a conditional into an assertion and made a score of 3 look like a finding rather than a hope.

Capital Needs: 4. Requirements fall in the 10 to 100 million dollar range across the venture's life, covering evaluation engineering, standards expertise, and enterprise sales.

Competitive Landscape: 4. The field is early and fragmented, incumbent consultancies sell manual audits, and differentiation comes from repeatable tooling, standards alignment, and credibility with insurers and procurement officers. This score should be read with the finding above in mind: Competitive Landscape spanned only one point across six retrospective cases, so a 4 here carries less information than its position in the weighted sum implies.

Process Improvement versus Platform Development: 4. Audit tooling alone is a process improvement on consulting, while the platform opportunity is shared evaluation infrastructure that other parties build on. Under the stricter standard proposed above, which requires realized third-party building rather than an argument for it, this score is currently unearned and should be 3.

Market Size to Capital Investment Ratio: 3, trending upward. Independent estimates put AI audit and assurance services well under 1 billion dollars in 2026 with projections of order-of-magnitude growth within a decade (Fact.MR, 2026). Those projections are directional and vendor-sourced. Using the definition established above, a ten-year addressable market in the low tens of billions against total capital under 100 million dollars would give a ratio in the hundreds, which is the band as scored.

Weighting. This is a regulation-and-trust-driven market, so the evaluation weights Time to Market and Competitive Landscape at 0.25 each, Process versus Platform at 0.20, Capital Needs at 0.15, Market Size to Capital at 0.10, and Return on Investment at 0.05.

Result. The weighted score is 3.85 out of 5.00 against an equal-weight mean of 3.67. The previous edition reported this venture as 22 of 30, which is the equal-weight case expressed on an abandoned scale, and reported it four lines after arguing that weighting was decisive. The weighted figure is only 0.18 above the flat figure, which is worth stating plainly: on this venture, the weighting the paper argued was decisive moved the answer very little.

Scenario testing. Three scenarios span the plausible range. Under continued regulatory retreat, state requirements narrow and federal preemption advances, so demand rests on insurers, enterprise buyers, and procurement standards, and the venture grows more slowly while avoiding compliance-driven commoditization. Under regulatory re-expansion, new mandates arrive in major jurisdictions, demand rises sharply, consultancies enter, and tooling advantage decides the winners. Under institutional formation, an independent national assurance institution emerges, certification criteria standardize, and accredited tooling vendors supply a much larger market. The venture scores acceptably in the first two and best in the third, and the asymmetry is the investment case, subject to the disclosure above.

Where the framework gives the wrong answer

A rubric whose authors cannot say where it misleads has not been used enough. Four cases.

It is biased against deep tech, which is the tension at the center of this paper. The framework rewards short Time to Market, low Capital Needs, and high market-to-capital ratios. Deep technology scores badly on all three by definition, since it is slow, expensive, and capital-hungry relative to its addressable market in the near term. A framework published by a deep tech investor therefore systematically underrates the category its publisher exists to fund, and the six retrospective cases demonstrate it: the two lowest equal-weight scores, autonomous vehicles at 17 and solid-state batteries at 19, are the two most deep-tech entries in the set.

The tension resolves, and the resolution is the framework's most useful output. A low score is not a verdict on the technology. It is a statement about the mismatch between the technology and a particular class of capital, and the anchors have carried investor-fit annotations from the beginning for that reason. The framework's proper use is routing rather than screening: it identifies which capital structure can hold the asset. Used as a go/no-go gate by a single investor class, it will reject the best deep tech opportunities available, which is precisely what a flat 17 on autonomous vehicles did.

It cannot see terminal market structure. Nothing in the six metrics captures whether a market ends up winner-take-most or fragmented. Autonomous ride-hail and enterprise software with identical scores have entirely different expected values, because in one the winner holds a durable position and in the other margins compete away. Capital Needs partly proxies for this, since high capital requirements deter entrants, and it does so accidentally and in the wrong direction, penalizing the barrier that creates the value.

It cannot detect misrepresentation. Every input is a company representation, which is why Theranos was removed as an anchor. An evaluator scoring a venture that is lying will produce a well-formed score that is entirely wrong. This is not a defect to be fixed by another metric, and it should be stated in every output: the framework assumes the evidence is truthful and diligence is a separate discipline.

It treats the six metrics as independent when they are not. Capital Needs and Time to Market correlate strongly, as do Market Size to Capital and Capital Needs, which shares a term. A weighted mean over correlated dimensions double counts what the correlated dimensions share, so an evaluator putting 0.20 on Capital Needs and 0.10 on the market-to-capital ratio has placed more than 0.30 on capital intensity without intending to. The framework has no correction for this and a future edition should either orthogonalize the metrics or state the correlation matrix so evaluators can adjust.

Using the framework: the procedure

The complaint that most needed answering in this edition was that the framework described an instrument rather than being one. The components are now specified, and this section states the order in which they are applied, incorporating the corrections established above. An evaluation that departs from this sequence should say where and why.

1. Assemble the evidence pack before scoring anything. Fix the evidence first, because a score assigned while evidence is still arriving will be revised toward whatever the evaluator already believes. The minimum pack is a technical description sufficient to establish maturity, a total capital requirement to a defensible market position, a stated addressable market at a ten-year horizon with its derivation, the competitive set, and the named parties bearing each material risk.

2. Declare the tier. Advanced tier applies all six metrics and requires the full pack. Basic tier applies Time to Market, Capital Needs, and Return on Investment for ventures whose data does not support more, and a basic-tier result is labeled as such and is not compared against an advanced-tier result.

3. Declare the weight vector before seeing the scores. This ordering is the single cheapest discipline in the whole procedure. An evaluator who scores first and weights second will choose weights that produce the conclusion already formed, and because the weighting is published the choice will look principled. Start from the default vector for the relevant class of capital and record any departure with its reason.

4. Score each metric against its band, independently. Assign integers only. Do not adjust a score because another score seems inconsistent with it, since apparent inconsistency between metrics is information rather than error, and it is exactly what the autonomous vehicle case shows. Apply the two corrections established in this edition: the platform metric requires evidence of realized third-party building for a 5, so an argument that a platform could emerge scores 3 at most; and Competitive Landscape scores whether the position is defensible rather than whether it is differentiated, so a cost leader with a structural cost advantage scores on that basis.

5. Compute the weighted mean and report it with its vector. Two decimal places, out of 5.00. Never a sum out of 30. A score published without its weight vector cannot be interpreted and should be rejected by whoever receives it.

6. Report the vector of scores alongside the scalar. A 3.50 composed of six mid scores and a 3.50 composed of two 5s and two 2s describe different propositions, and aggregation discards that shape. Both go in the output.

7. Stress test across scenarios, not across confidence intervals. Define two or three scenarios that differ in kind, rescore the metrics that move, and report a score per scenario. Do not average them, because the scenarios are not draws from a distribution.

8. Route rather than screen. Compute the score under more than one weight vector and identify which class of capital the venture fits. This is the step that the framework's own record most supports, and skipping it is what caused the framework to rank its eventual best case last. A low score under one vector is a statement about a mismatch, not a verdict on the technology.

9. Record what the framework cannot see. State explicitly that the evaluation assumes the evidence is truthful, that no team assessment is included, and that terminal market structure is not represented. An output that omits these reads as more complete than it is.

Inter-rater reliability, unmeasured

The framework's most-cited weakness, in this paper and its predecessors, is that scoring retains subjectivity, and the stated mitigation is that benchmarks and defined bands reduce it. That mitigation has never been measured, and the honest position is to say so and to specify what would settle it.

The measurement is inexpensive. Three evaluators score the same venture independently from the same evidence pack, without conferring, and the spread per metric is reported alongside the mean. Even three raters would establish whether the bands constrain judgment or merely decorate it, and the result would be informative in either direction: agreement within one point on all six metrics would support the benchmark claim, while a two-point spread on Competitive Landscape or the platform metric would confirm what the retrospective scoring already suggests about those two dimensions.

Until that exercise is run, every score this framework produces should be understood as one evaluator's reading, and the claim that historical benchmarks reduce subjectivity should be read as a design intention rather than a demonstrated property. No such exercise has been conducted for any edition of this framework.

Shortcomings of the refined framework

Beyond the failure cases and the unmeasured reliability, five limitations apply.

Data availability. The full six-metric evaluation demands information that early-stage ventures frequently lack, particularly a defensible addressable market figure and a credible total capital requirement. The tiered structure mitigates this and does not remove it, and a basic-tier score omits the three metrics that carry the most discriminating power in the retrospective set.

Coarse bands. Five bands per metric produce large discontinuities. A venture requiring 240 million dollars scores 3 on Capital Needs and one requiring 260 million scores 2, and the difference in the weighted result exceeds any difference in the underlying reality.

Static anchors in moving markets. The anchor cases date from a particular period, and their meaning drifts. A capital requirement that placed a company in the second band a decade ago describes a different kind of company today, so the bands require periodic recalibration that no edition has yet performed.

No treatment of team. The framework scores technologies and markets and says nothing about founders, which is the factor most venture investors report weighting most heavily. This is a deliberate scope choice and it means the framework is an input to a decision rather than the decision.

Aggregation remains a compression. Even correctly weighted, a single number discards the shape of the score. A venture at 3.50 with all metrics near 3.5 and a venture at 3.50 with two 5s and two 2s are different propositions, and the framework's output should always be presented as the vector alongside the scalar.

Conclusion

The refined framework improves on TRL, IRL, and Stage-Gate by adding the commercialization metrics those models omit, and this edition improves on its own predecessor by making the instrument executable and by testing it against its own examples.

That test produced three unfavorable results and they belong in the summary rather than in an appendix. The platform metric returned an identical score in all six retrospective cases and therefore carries no information as currently defined, which requires the stricter evidentiary standard proposed above. The framework ranked autonomous vehicles last among six technologies shortly before that category reached commercial operation, an error that the weighted form corrects and that flat summation caused. And the Competitive Landscape metric scores only one of the two generic competitive strategies, which led one case study to advise against a cost leader while the next case study celebrated the same firm for the same strategy, a contradiction no previous edition noticed.

Two of those three carry repairs proposed in this edition: a stricter evidentiary standard for the platform metric, and a redefinition of Competitive Landscape to score defensibility of position rather than differentiation specifically. Weighting does not correct the framework's remaining visible error, its favorable reading of blockchain-based smart contracts, which survives every repair here.

Three properties survive that scrutiny. The framework integrates commercial metrics with technical readiness. It adapts across sectors and investor classes through weight vectors that are now specified rather than gestured at. And it aligns technologies with the capital structures that can hold them, which is its most defensible use and the one this edition recommends: routing rather than screening, because a rubric weighted toward speed and capital efficiency will always reject the slow, expensive technologies that deep tech investors exist to fund.

The direction of the field supports the multi-dimensional approach, with post-2024 work extending readiness thinking into ethics, manufacturing, and integration, and converging on the finding that market misalignment rather than technical failure drives most innovation losses (De Jong, 2025; ITONICS, 2025; IP.com, 2026). That convergence should be read alongside the retrieval limitation stated in the introduction, since three relevant critiques remain unread.

What the framework still lacks is a genuine test. Every scoring in this paper was assigned with the outcome partly visible, and no inter-rater exercise has been conducted. The next edition should register scores for a cohort of ventures before their outcomes are known, publish the weight vectors used, and report the spread across three evaluators. Until then this is a well-specified instrument with an untested calibration, which is a more accurate description than any previous edition has offered.

About this series

Stout Street Capital publishes this paper as №03 of a three-part series on the institutions the deep tech era requires.

- **№01. The Assurance Gap.** The case for independent conformance assessment of deployed AI systems: published test methods, a defined certification unit, and a public registry, operating inside the existing accreditation architecture.
- **№02. LEVCAP.** A lab-embedded venture capital access program: the commercialization engine that moves deep technology from laboratory to market.
- **№03. A Holistic Evaluation Framework for Innovation and Technology Commercialization** (this paper). The evaluation instrument: a refined framework that joins technical readiness with commercial metrics to judge what is real.

Stout Street Capital invests in early-stage deep technology, including categories evaluated in this paper. The prospective example in this edition carries a specific disclosure at the head of that section.

References

Astebro, T. (1998). Basic statistics on the success of high-tech entrepreneurs and their ventures. Research Policy.

Blank, S. (2013). It's time to play moneyball: The investment readiness level. steveblank.com. https://steveblank.com/2013/11/25/its-time-to-play-moneyball-the-investment-readiness-level/

Blank, S. (2014). How investors make better decisions: The investment readiness level. steveblank.com. https://steveblank.com/2014/07/01/how-investors-make-better-decisions-the-investment-readiness-level/

Bloomberg. (2017). Blue Apron. Bloomberg.com.

Bloomberg. (2021). Inside QuantumScape's secret battery lab. https://www.bloomberg.com/features/2021-quantumscape-battery/

Business Insider. (2021). Inside Zoom's unprecedented pandemic growth. https://www.businessinsider.com/inside-zoom-2020-unprecedented-pandemic-growth-executives-new-challenges-2021-5

Byers, D. (2021). The meteoric rise of Clubhouse. NBC News. https://www.nbcnews.com/tech/tech-news/social-audio-fire-who-owns-its-future-n1258944

Caerteling, J. S., Di Benedetto, C. A., Doree, A. G., & Halman, J. I. (2008). Technology development projects in road infrastructure. Research Policy.

Carlson, N. (2012). Inside Pinterest: An overnight success four years in the making. Business Insider. https://www.businessinsider.com/inside-pinterest-an-overnight-success-four-years-in-the-making-2012-4

Cooper, R. G., & Sommer, A. F. (2016). The Agile-Stage-Gate hybrid model. Journal of Product Innovation Management.

Cramer, J., & Krueger, A. B. (2016). Disruptive change in the taxi business: The case of Uber. American Economic Review.

Dang, A., & Strebulaev, I. (2024). The Venture Mindset. Random House.

De Jong, E. (2025). Ethics readiness of technology: The case for aligning ethical approaches with technological maturity. arXiv. https://arxiv.org/abs/2504.03336

Ettiner, G. (2023). The real reason Jawbone failed. SlashGear. https://www.slashgear.com/1375929/jawbone-real-reason-speaker-company-failed/

Fact.MR. (2026). AI agent audit and assurance services market. https://www.factmr.com/report/ai-agent-audit-and-assurance-services-market (Directional vendor estimate.)

Fast Company. (2016). What smartwatches lost with the death of Pebble. https://www.fastcompany.com/3066659/what-smartwatches-lost-with-the-death-of-pebble

Feldman, M. P., & Kelley, M. R. (2006). The ex ante assessment of knowledge spillovers. Research Policy.

Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: Evidence from patent data. Research Policy.

Forbes. (2021). Healthier plant-based meat is on the rise. https://www.forbes.com/sites/briankateman/2021/05/10/healthier-plant-based-meat-is-on-the-rise/

Francis, J. (2026). The Assurance Gap: The case for an independent national AI assurance institution. Stout Street Capital, white paper №01.

Garun, N. (2018). Amazon has acquired Ring. The Verge. https://www.theverge.com/2018/2/27/17059664/amazon-ring-acquisition-smart-home-security-cloud-cam

Geuna, A., & Muscio, A. (2009). The governance of university knowledge transfer. Minerva.

Grimpe, C., Minssen, T., Price II, W. N., & Stern, A. D. (2022). Regulatory and commercial dynamics in vaccine platform development.

Grush, L. (2019). Commercial space companies have received $7.2 billion in government investment since 2000. The Verge. https://www.theverge.com/2019/6/18/18683455/nasa-space-angels-contracts-government-investment-spacex-air-force

Heinemann, L. (2008). The failure of Exubera: Are we beating a dead horse? Journal of Diabetes Science and Technology. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2769732/

Intel. (1965). Moore's law. https://www.intel.com/content/www/us/en/newsroom/resources/moores-law.html

IP.com. (2026). Readiness level frameworks: Aligning technology and innovation maturity to reduce R&D risk and improve ROI. White paper. https://ip.com/wp-content/uploads/2026/03/Readiness-Level-Frameworks_White-Paper.pdf (Vendor publication; underlying academic citations not independently verified.)

ITONICS. (2025). 14 readiness level frameworks: The guide to TRL, MRL, SRL, and beyond. https://www.itonics-innovation.com/blog/14-readiness-level-frameworks

Link, A. N., & Siegel, D. S. (2005). Generating science-based growth. European Journal of Innovation Management.

Lunden, I. (2018). Stripe is now valued at $20B. TechCrunch. https://techcrunch.com/2018/09/26/stripe-is-now-valued-at-20b-after-raising-another-245m-led-by-tiger-global/

Lunden, I. (2022). Stripe expands its infrastructure play. TechCrunch.

Lurie, N., Saville, M., Hatchett, R., & Halton, J. (2020). Developing Covid-19 vaccines at pandemic speed. New England Journal of Medicine.

MacDonald, S. (2024). Shopify platform overview.

Markowitz, E. (2012). How Instagram grew. Inc.

McGrath, R. G. (2011). Failing by design. Harvard Business Review.

Miller, R. (2016). How AWS came to be. TechCrunch.

Miller, R., & Lessard, D. (2001). The strategic management of large engineering projects. MIT Press.

Mowery, D. C., Nelson, R. R., & Sampat, B. N. (2006). Ivory tower and industrial innovation. Stanford University Press.

Mukherjee, A. (2021). The Nano's failure. Bloomberg Opinion.

New York Times. (2021). Juul and the vaping crisis.

New York Times. (2023). WeWork's collapse.

Palmer, A. (2021). Rivian's backers. CNBC.

Phaal, R., O'Sullivan, E., Routley, M., Ford, S., & Probert, D. (2011). A framework for mapping industrial emergence. Technological Forecasting and Social Change.

Pisano, G. (2007). Science business: The promise, the reality, and the future of biotech. Harvard Business School Press.

Porter, M. E. (1998). Competitive strategy. Free Press.

Pritajaya, D. (2024). BYD's vertical integration strategy.

Salyer, K. (2019). How Slack became a $20 billion company. Fast Company.

Techcrunch. (2011). Airbnb raises at a $1B+ valuation.

Teece, D. J. (1986). Profiting from technological innovation. Research Policy.

Van Krevelen, D. W. F., & Poelman, R. (2010). A survey of augmented reality technologies, applications and limitations. International Journal of Virtual Reality.

Warren, T. (2017). Peloton's rise. The Verge.

Wilhelm, A., & Crook, J. (2014). Facebook buys WhatsApp for $19B. TechCrunch.

Wired. (2016). Dropbox's fight to survive.

Recorded at Denver, Colorado · MMXXVI