Benthurst

WHITE PAPER · AI STRATEGY

AI becoming a hammer, how to prevent every problem from looking like a nail

A practical filter for executives: what AI can solve today, what it cannot, and what to wait for

Executive Summary

AI has become the default answer and it is often wrongly applied or not fulfilling expectations. In 2025, 95% of generative-AI pilots produced no measurable business return (MIT Project NANDA), 80% of AI projects failed to deliver intended value (RAND), and 42% of companies abandoned at least one AI initiative at an average sunk cost of $ 7.2 million (Deloitte). The problem is rarely the technology. It is that boards and CEOs are wielding AI as a universal hammer, swinging it at problems that are not nails.

This paper offers executives three practical filters to avoid that mistake.

Filter 1 – AI or not AI: Before deploying AI, diagnose the root cause of the underlying problem. Many "data problems" are in fact **process** or **culture problems**, and no amount of AI will mop the floor while the tap is open.

Filter 2 – Sourcing: If you don't want to wait for tighter processes and cleaner data, you can already adopt off the shelf AI, but don't expect differentiation.

Filter 3 – Which AI: Once the problem is genuinely an AI problem, match it one of the four **lanes**: (1) **Mainstream LLMs & agentic AI**; (2) **Specialised AI** (computer vision, classical ML, sensor fusion); (3) **Right-sized LLMs/SLMs** (smaller, open-source, more efficient); (4) **Wait for AI 2.0** (world models, neuro-inspired efficient training).

AI as the Hammer

Executives of non-native AI companies oscillate between **awe** and **disappointment**. They are simultaneously fed with diverging inputs about AI, like:

> “AI will redefine more than we can imagine”
>
> “Most AI implementations fail to deliver so far”
>
> “The AI-bubble is about to explode”
>
> “Before 2030 we’ll reach AGI”
>
> “AI’s growth won’t be caped by GPUs but by electricity”

..and the resulting tension is mentally taxing.

At the same time, by a mixture of **cautious investment** and **fear of missing out**, companies run a lot of AI projects to chase efficiency gains. *"If the only tool you have is a hammer, everything looks like a nail"* describes the corporate posture toward artificial intelligence. Boards demand an "AI strategy", large tech consultancies promise transformation, vendors push frontier models, and operating committees run pilots into every function indiscriminately. The result is a growing inventory of disappointing outcomes.

The numbers are striking

95% of generative-AI pilots show no measurable return (MIT NANDA, 2025)	**80%** of AI projects fail to deliver intended value (RAND, 2025)	**42%** of companies abandoned at least one AI initiative (Deloitte, 2025)

MIT’s Project NANDA report – *The GenAI Divide: State of AI in Business 2025* – examined 300 deployments, 150 leader interviews and a survey of 350 employees. Despite $ 30-40 billion of enterprise GenAI spend, only 5% of organisations see P&L impact. The root cause, the authors stress, is not model quality but organisational integration, problem framing and where the use case is targeted (back office wins; sales/marketing is the largest budget yet smallest ROI).

> *"The technology works. The hammering not."*

Why it is more than a budget problem

The collateral damage is broader than wasted capital. Klarna laid off 700 customer-service agents in 2024 on the assumption that AI could replace them; by 2025 the CEO publicly admitted the cuts had gone too far and the company was **rehiring humans**. IBM made a similar move with its *AskHR* system: 8,000 HR roles automated, then partly reversed. Gartner now predicts that **50%** of AI-attributed headcount reductions will **reverse** by 2027, and Forrester reports that **55%** of employers **regret** laying people off for AI. Beyond the financial waste, each reversal erodes trust, in management, in technology vendors, and in AI itself. The discipline of **not using AI** has become as strategically important as the discipline of using it.

> *"The discipline of **not** using AI has become as strategically important as the discipline of using it."*

Moreover, these findings predate the recent and **sharp price increases** for frontier foundation models, which not only further strain the business case for AI investments but also leave businesses uncertain about future cost trajectories.

02 / FILTER ONE – AI OR NOT AI

Diagnose Before You Deploy

The first filter is diagnostic. Most "AI projects" are launched without sufficiently analysing: **what is the actual source of the problem we are trying to solve?** In our practice we find that the answer falls into three broad categories – and only one of them is genuinely amenable to AI in the short term.

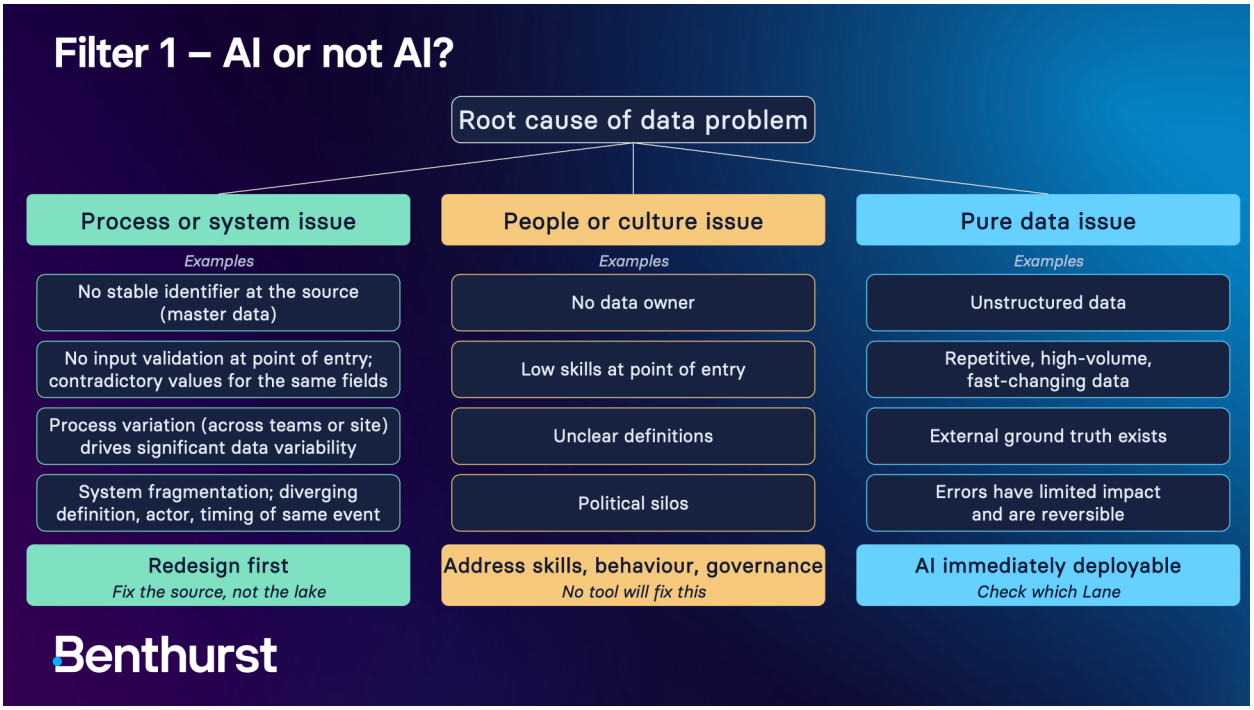


Figure 1 – Sources of data problems. Only the right-hand column is immediately AI-actionable.

Reading Filter 1

Process or system issues – no stable identifier assigned at the source, absent validation rules, process variation that drives data variability, fragmentation between systems. The intervention is **redesign first**. Pointing an LLM at a broken process produces faster, cheaper, more confident *wrong* answers.

A growing number of companies are addressing these issues by building semantic layers on top of raw data sources; the cleaning and structuring happens in that layer. We beg to differentiate: where the underlying process is sound and the data is merely trapped in **messy formats**, a semantic layer does the job. However, where the source system itself **never captured** the event – no stable ID, no validation, no shared definition – no layer downstream can reconstruct what was never recorded.

People or culture issues – no accountable data owner, low skills at the point of entry, unclear definitions, political silos. The intervention is **governance, skills and behaviour**. No tool resolves a refusal to share data or a missing agreement on what e.g. "an active customer" means. That said, a well-designed knowledge agent can lower the cost of cross-silo collaboration, but aligned governance and incentives remain very beneficial.

Pure data issues – unstructured data, repetitive/high-volume/fast-changing flows, available external ground truths, reversible errors with limited impact; here **AI** is **immediately deployable** and short-term ROI is realistic.

Humility is warranted

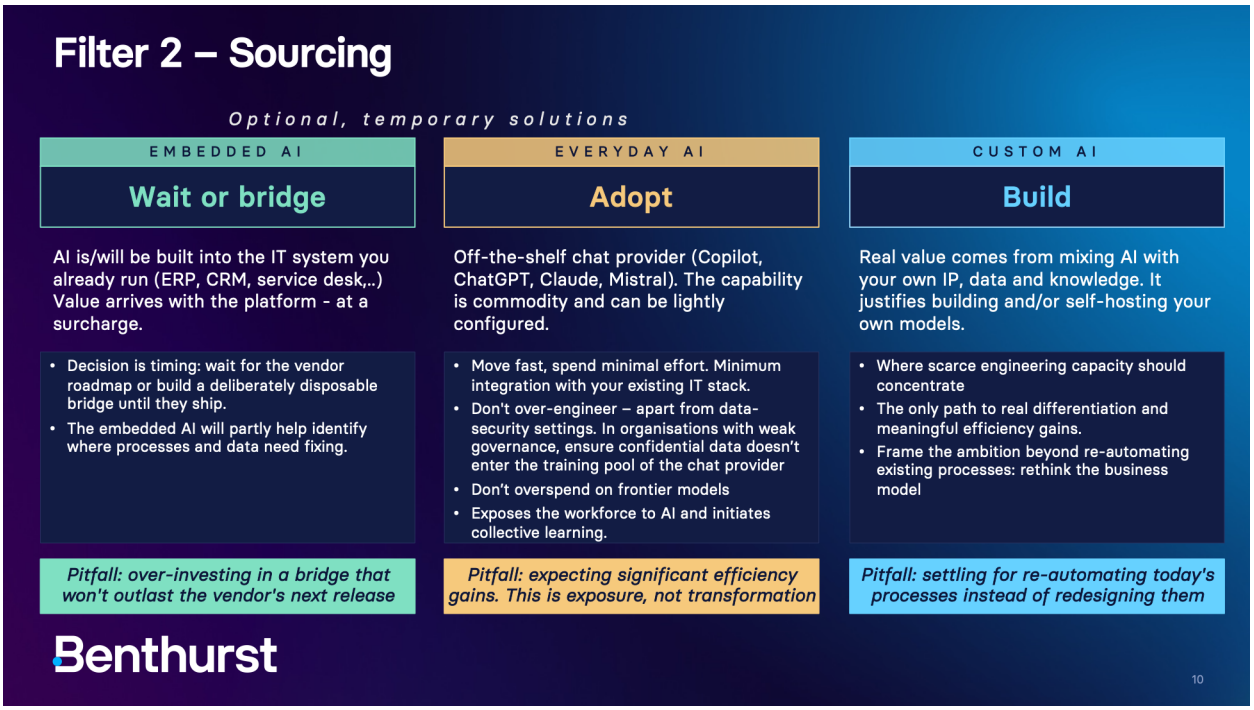
Our framework's shelf life is maybe 6 months – given AI's breakneck evolution – and is deliberately simplifying. In practice, the three columns overlap; a "pure data" problem with no data owner is in fact a culture problem dressed up as a data problem. To sustainably solve your data issues, you should address process and culture issues first.

The data-issue column hides its own trap: just because AI *can* be deployed, doesn't mean it requires custom AI or the latest generation LLM. Both these decisions are the subject of the next 2 Filters.

03 / FILTER TWO – SOURCING

Three ways to source AI

Analysis for Filter 1 should be ruthless, Filter 2 introduces some leniency. Before you fix all process and culture issues, you can already start with AI. Key is to align your sourcing choices with the findings of Filter 1 and to start with the right expectations.



Embedded AI, wait or bridge. Even if your processes or systems are still noisy, AI is, or soon will be, built into a production system you already run (your ERP, CRM, service desk, dev and

monitoring stack). The value arrives with the platform you already pay for, albeit at a surcharge. The decision is timing: wait for the vendor roadmap or build a deliberately disposable bridge until they ship. **Anything you build here should be throwaway by design**, because the vendor will overtake it.

Depending on the quality of the processes and data in the IT system, the embedded AI will already improve or at least **help identify the areas where processes and data need most fixing**.

Everyday AI, adopt. When data awareness is generally weak in your organisation, you can nevertheless roll out a chat provider (Copilot, ChatGPT, Claude, Mistral). The capability is a commodity, off the shelf, at most lightly configured. It exposes the workforce to AI and initiates collective learning. This method is commonly followed for non-proprietary (but potentially confidential) data like email, meeting minutes, internal or external publications. Functionality is mostly summarising or responding, i.e. facilitating everyday tasks.

This is where you move fastest and spend the least effort. It requires minimum integration with your existing IT stack (apart from Copilot with the Microsoft suite, which can be considered as 'embedded light'). Do not over-engineer it, apart from the data security settings. Especially in organisations with weak data governance, you don't want confidential data to become part of the training pool of LLMs.

The major pitfall is expecting very significant efficiency gains. Even when AI is efficient in facilitating everyday tasks, recent research reveals that humans can only sustain a certain level of acceleration, especially if human in the loop is required. Hence be mindful of the costs: we expect the cost difference to rise between the latest vs. previous versions of LLMs. Frontier models for summarising emails or meeting minutes are overkill.

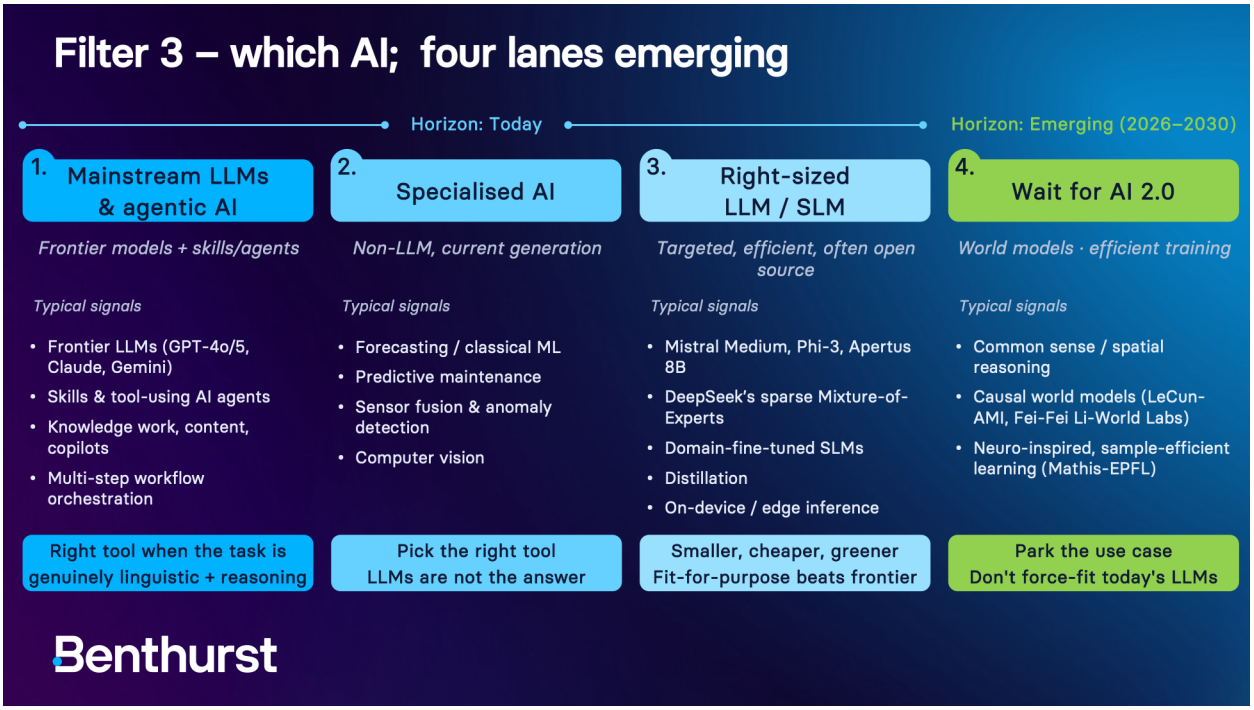
Custom AI, build. The market is not expected to provide it, and the value comes from mixing AI with your own IP, data and knowledge. This is the only tier that justifies building or self-hosting your own models. **It is also where scarce engineering capacity should concentrate.**

It's the only way you can seriously expect to create **differentiation** and **significant efficiency gains.** The pitfall here is to limit yourself on re-automating existing processes. The quickly expanding possibilities of AI should invite you to **rethink your business model** i.e.potentially phasing out parts where friction is eliminated through AI and creating new services with costs, speed and quality being a multiple of current levels.

04 / FILTER THREE – WHICH AI

Match the Problem to the Right Lane

Once a problem has survived Filter 1 and 2, a new distinction is needed. "AI" is not one technology; it is a family of very different ones with very different application, cost, accuracy and maturity profiles. We see four lanes emerge.



LANE 1 ·

Mainstream LLMs & agentic AI

When the task is genuinely about language, knowledge and multi-step reasoning, and the cost of an occasional wrong answer is bounded, frontier LLMs and the agentic systems built on top of them are the right tool. This is the lane that has absorbed the bulk of the $ 30-40 billion of enterprise GenAI spend tracked by MIT NANDA (2025), and it is also the lane where wins and write-offs are most concentrated. Gains compound in the back office – software engineering, legal, finance, IT, internal knowledge retrieval – but are harder to realise in customer-facing roles where ground truth, brand and liability matter.

Example: BBVA (Spain) + Google Gemini. The largest documented European deployment to date. BBVA is rolling out Gemini-powered Google Workspace to ~100,000 employees,

with 87,000+ already integrated into daily work for drafting, data analysis, content and presentations. Internal measurements report time savings of over 76% on targeted tasks and an average of nearly 3 hours saved per week per employee. A dedicated AI copilot for developers generates code, suggests fixes and accelerates testing.

Other examples in this lane:

- JPMorgan – COiN cut commercial-loan contract review from 360.000 lawyer-hours per year to seconds; the bank's internal LLM Suite was rolled out to ~200.000 employees in 2024 for drafting, summarisation and research over proprietary content.
- BNP Paribas + Mistral AI (France). Hello bank!'s virtual assistant HelloïZ has run on Mistral generative AI since January 2026, serving more than 1 million clients. In May 2026 BNP Paribas extended its Mistral partnership for three years and is rolling out an internal AI assistant for summarisation, reformulation, translation and content generation. A live demonstration that customer-facing and internal use can coexist when the underlying model is European, data residency is in the EU, and the bank co-owns the deployment.
- ING (Netherlands). Customer-facing chatbots automate 75% of queries across multiple country markets, with Microsoft 365 Copilot and engineering copilots rolled out internally. Employees interact with the tools 30–50 times per week in some markets, a useful proxy for genuine workflow integration rather than pilot theatre.

Cautionary note – Klarna (Sweden). Klarna's customer-service agent handled 2.3 million conversations in its first month, equivalent to ~700 FTEs (OpenAI / Klarna case, 2024). By 2025 Klarna had publicly rehired human agents, confirming the MIT NANDA finding that front-office deployments command the largest budgets and the smallest returns, a European data point on the limits of Lane 1 in customer-facing roles

Executive action: deploy frontier LLMs and agents where the back-office pattern holds; output is tolerant of imperfection, the blast radius is bounded, and a human owns the last mile. Remain mindful in the front-office: the Klarna reversal, and the 95% no-ROI figure from MIT NANDA, show that sales, marketing and customer service is the most expensive place to learn this lesson.

For every Lane 1 business case, force the question "why not Lane 3?". Many tasks framed as needing frontier-class capability run perfectly well on a fine-tuned 7B open-weights model at one-tenth the cost, latency and energy.

LANE 2 ·

Specialised AI

Many of the highest-ROI applications running in industry today are **not LLMs at all**. They are convolutional neural networks for vision, gradient-boosted trees and time-series models for forecasting, reinforcement learning for scheduling, and sensor-fusion pipelines for predictive maintenance. These tools are decade-mature, energy-modest, mostly explainable and – critically – *predictable* on narrow tasks.

Example: Polysense (Ghent, Belgium, founded 2022) uses **AI-powered machine vision** to inspect food-processing lines in real time. Smart cameras connect directly to PLCs and SCADA systems, generating continuous objective quality data for clients including Agristo, Roger & Roger and Coroos. The technology underneath is computer vision and edge inference, not an LLM, and that is exactly why it works.

Other examples in this lane:

- Automotive plants using vision and deep learning to inspect robotic welders – **70% reduction** in inspection time and **10% improvement** in weld quality.
- Oil & gas pipeline inspection – image-based defect detection (rust, cracks, deformation) running at the edge.
- Demand forecasting and pricing – gradient-boosted models or Bayesian time-series; still routinely outperform GenAI on structured-data tasks.

Executive action: treat "AI" as a portfolio. Insist the technology choice (LLM vs. classical ML or other) is explained.

LANE 3 ·

Right-sized LLMs and SLMs

For genuine language and reasoning tasks, the dominant assumption in 2025 was that bigger is better. That assumption is breaking. **Small Language Models (SLMs),** typically 1–10 billion parameters, sometimes much smaller, are fit-for-purpose alternatives that win on cost, latency, privacy and energy.

Cost gap: training GPT-4 reportedly costed ~$ 78 million, Google Gemini Pro $ 191 million; fine-tuning an SLM for a narrow task costs orders of magnitude less and runs in weeks rather than months.

But also for users, the bills are steeply going up now that AI companies start to transfer a more substantial part of the real costs. Uber said it had blown through its projected A.I. spending for the year in just four months, and it has placed monthly limits on A.I. coding tools. **Tokenmaxxing is quickly being replaced by tokenminning**.

Energy gap: the International Energy Agency reports that AI-focused data-centre electricity consumption **grew 50% in 2025** and is on track to quadruple by 2030. Reasoning-heavy frontier models such as o3 or DeepSeek-R1 can consume **over 70× more energy per long prompt** than smaller models like GPT-4.1 nano. This voracious energy consumption could, just as with the blockchain, limit growth. For most enterprise tasks – classification, extraction, summarisation or RAG – a properly tuned 7B-parameter model is sufficient and dramatically greener.

Availability gap: as demonstrated by the rapid withdrawal of Anthropic's Fable 5 after the launch, European companies can no longer count on the availability of Frontier models.

Capability gap: distillation. Another reason the bigger-is-better assumption is breaking is *knowledge distillation* – training a small "student" model on the outputs (and increasingly the internal reasoning traces) of a larger "teacher". The student inherits a meaningful share of the teacher's capability at one-tenth to one-hundredth of the parameter count. The clearest 2025–2026 evidence is the *DeepSeek-R1-Distill* series released in January 2025: six dense student models from 1.5B to 70B parameters, fine-tuned on 800,000 reasoning traces generated by the full DeepSeek-R1. The 32B student outperformed OpenAI's o1-mini across major reasoning benchmarks i.e. a model less than half the size of its non-distilled competitor wins on the metric that matters.

Distillation is now the principal technique by which SLMs close the capability gap without giving back the cost, latency or energy gap.

A short taxonomy of right-sized models in 2026:

- **Mistral Medium** or **Small** (France) – Mistral 7B and Mixtral, strong European open-weights option; data residency in the EU.
- **Phi-3 / Phi-4** (Microsoft) – designed to run on a laptop or even a phone; remarkable per-parameter performance, achieved largely through training on textbook-quality synthetic data generated by larger teachers; distillation in practice, even when not labelled as such.
- **Apertus 70B / 8B** (ETH & EPFL) – Open weights, open data, open science, EU AI Act compliant (only trained with compliant data and zero IP violations)
- **DeepSeek-MoE / DeepSeek-V3** (China) – sparse Mixture-of-Experts; activates only ~2.7B parameters per token, delivering frontier performance at a fraction of the compute.
- **DeepSeek-R1-Distill family** (China) – six open-weights student models (1.5B → 70B) distilled from the full R1 reasoning model. The 32B version outperforms OpenAI o1-mini on AIME, MATH-500, GPQA Diamond and Codeforces; the 7B and 14B versions bring frontier-class reasoning to a single consumer GPU. The standard 2026 reference for "distilled SLM beats frontier mini-model".
- **Domain-fine-tuned SLMs** on a base model + RAG over your own corpus; the most common winning architecture for enterprise knowledge tasks.
- **On-device, edge inference** i.e. running pre-trained models on local devices, reducing latency and enhancing privacy at lower cost & bandwith
- **Open-source models** like GLM/Z.ai, Kimi/Moonshot or smaller alternatives like Qwen/Alibaba - all from China. Probably trained with the outputs of American frontier

labs. Self-hosted, these models present a low-cost option that preserves data sovereignty, though model-level safety guardrails still warrant independent evaluation.

Executive action: demand a "smallest sufficient model" justification on every AI business case. The frontier-by-default posture is expensive, slow and increasingly hard to defend.

> *"Tokenmaxxing is quickly being replaced by tokenminning."*

LANE 4 ·

Wait for AI 2.0

A meaningful set of problems – robust common sense reasoning, causal inference, long-horizon planning, spatial reasoning and physical interaction with the real world – cannot be reliably solved by today's AI. Forcing them onto an LLM produces brittle demos and confident hallucinations. The honest answer for these use cases is: **park the project and watch the research.**

Two research directions exemplify "AI 2.0":

(a) Genuine world models. In December 2025 Yann LeCun (Turing Award winner, formerly Meta's chief AI scientist) left Meta to found **Advanced Machine Intelligence (AMI Labs)** in Paris, raising a $ 1bn seed at a $ 3.5bn valuation. AMI builds on **JEPA** (Joint Embedding Predictive Architecture): rather than predicting the next token, JEPA learns abstract representations of the world in latent space, much closer to how animals seem to reason. In parallel, World Labs – Fei-Fei Li, Justin Johnson, Christoph Lassner, Ben Mildenhall – raised $ 1bn for "spatial intelligence", and in November 2025 shipped **Marble**, a commercial world-model product that turns prompts into editable 3D environments.

(b) Sample-efficient, neuro-inspired training. Mackenzie Mathis's 'Lab of Adaptive Intelligence' at EPFL Lausanne pursues a fundamentally different research path: reverse engineering the neural and behavioral mechanisms by which biological systems learn so efficiently, and translating those principles into machine learning architectures. Her CEBRA algorithm (Nature, 2023) exemplifies this dual ambition: a neuroscience tool that is simultaneously a contribution to general ML.

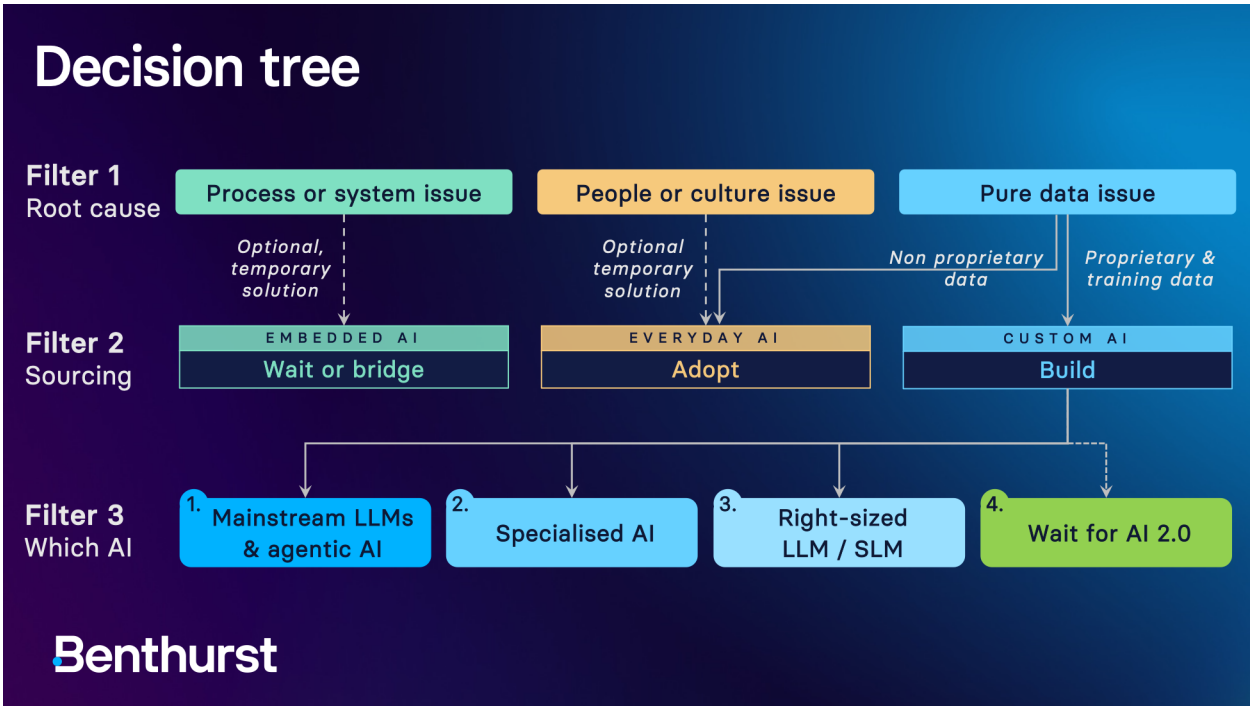
These research directions are trying to shift the paradigm: LeCun designs new architectures top-down, Li builds models grounded in 3D physical reality, and Mathis seeks the scientific principles that make biological intelligence orders of magnitude more data-efficient than today's AI.

China, meanwhile, turned scarcity i.e. limited access to GPUs into an advantage (e.g. the Mixture-of-Experts architecture of DeepSeek) and is no longer a follower: Tsinghua University has overtaken Harvard and MIT in AI patent filings, and Chinese papers have reached parity with US papers at top conferences such as ICLR (International Conference on Learning Representations).

Common to all these directions is a shift away from raw scaling towards architectural and biological efficiency, suggesting that the next inflection in AI may come not from larger models, but from fundamentally different or at least hybrid ones.

Executive action: maintain a small "AI 2.0 watch" portfolio. Identify the two or three business problems for which you are currently force-fitting a chatbot; document them; revisit twice per year. Do not invest at scale in workarounds.

The 3 filters combined



Finally, a short, executable sequence for the weeks after this paper lands on the desk:

1. **Inventory** every AI initiative under way – pilots, vendors, internal experiments, shadow tools.
2. **Tag** each one along all three filters: its Filter 1 root cause (process/system, people/culture, pure data), its Filter 2 sourcing route (embedded, adopted or custom built), and – for the genuine builds – its Filter 3 lane (1 to 4).
3. **Realign sourcing** to the diagnosis: kill, redesign or rescope the mismatches – the front-office chatbots, the process problems wearing AI clothing, the frontier-model overkills. Let embedded and everyday AI bridge the process and culture gaps you have not yet closed, and reserve scarce build capacity for the data problems where custom AI creates differentiation.
4. **Appoint a "smallest sufficient model" reviewer** in the AI governance forum, with the authority to challenge any Lane 1 default that has not first ruled out a specialised non-LLM (Lane 2) or a right-sized model (Lane 3).
5. **Install a 6-month review cadence**. The technology, the cost curves and the regulation will all have moved by then; the filters much less. The diagnosis is not the deliverable. The discipline is.

About Benthurst

AI expands the gap between what is technically possible and what organisations can actually absorb. As model capabilities accelerate, the limiting factor shifts from technology to behaviour: whether people understand when to use AI, trust it appropriately, challenge it when needed, and embed it into the routines through which work gets done. It makes behavioural insight more important, not less.

At Benthurst, we help clients design AI-enabled transformations around the human mechanisms that determine value: trust, incentives, decision rights, skills, feedback loops and adoption in the flow of work. The goal is not to deploy more AI, but to create the conditions under which AI improves judgement, execution and measurable performance.

Each of our consultants brings 20+ years of experience spanning both industry and consulting roles. We are deliberately small, deliberately senior, and deliberately multidisciplinary. Frameworks get us started; creativity and lateral thinking get us answers.

Authors

Xavier Bekaert

PARTNER @ BENTHURST

- +25 years consulting and management experience in financial services, healthcare, and mobility
- Previously at Bain & Company, Kearney and Puratos Germany
- Focuses on strategy realisation, process and business model innovation, AI and behaviour

Mirko Vaars

AI ECOSYSTEM LEAD @ PORT OF ROTTERDAM

- MSc in Data Science, University of Amsterdam
- AI and data-science specialist focused on enterprise adoption of generative and opensource models
- Builds and prototypes real AI products in practice, bridging frontier research and what organisations can actually deploy

Rick Hou

CEO @ AUGMENTED

- MS in Computer Science and MS in Probability & Applied Statistics, UC Santa Barbara
- Built web and mobile products for Shazam, ESPN, and Disney as Creative Director at Burnside Digital

> *" The bottleneck in enterprise AI is no longer model capability but organisational selectivity "*

Contact: www.benthurst.com · Get in touch · LinkedIn

Three ways to source AI

1. Everyday AI: Adopt a provider

We often think of Everyday AI as a temporary solution. It can help to raise data awareness through exposure of the workforce to generic tools for low-differentiation tasks. Limited influence on which type of AI (Filter 3). Especially relevant are the conditions which require moving to Custom AI (last column).

Problem type	Filter 1 – root cause	Best fit	Rationale	Move to Custom AI when…
Repeated document-shaped deliverables (reports, decks, memos)	**People/culture**; the template lives in one senior's head	Adopt a provider, encode the procedure once as Skills or templates	Capturing a senior's unwritten template is itself knowledge work	the procedure embeds proprietary IP or must run unattended at scale
Structured option generation and critique	**People/culture**; decision hygiene	Adopt a frontier reasoning provider in deep-thinking mode	The value is a 'challenger' in the room, an antidote to groupthink	the critique must encode your own decision criteria and data
Onboarding new employees	**People/culture**; knowledge is trapped in heads or in a multitude of documents	Chat provider over your non-proprietary docs and wikis	Traditional onboarding keeps interrupting seniors; asking AI instead builds the habit	answers must come from proprietary systems or run unattended
Cross-silo translation and jargon-busting	**People/culture**; silos, no shared definitions	Provider that summarises and translates internal comms	A knowledge agent lowers the cost of cross-silo collaboration	shared definitions must be governed and enforced over proprietary data
AI-literacy, learn-by-doing rollout	**People/culture;** weak awareness, low skills at entry	Broad rollout, light config, security settings only	Putting the tool in everyone's hands sparks learning across the team	a high-value workflow emerges that needs your data or IP
High-stakes regulated decisions	**People/culture** + governance (accountability)	Adopt a provider as drafter, human stays decider, add governance	Drafting is commodity, accountability stays human	residency or audit rules force the model in-house (self-host)
Analysis of long, dense documents (contracts, filings, research)	Pure **data**; commodity capability	Paste into a long-context LLM provider, no custom build needed	Off-the-shelf tools cover the capability; the human owns the final judgment	the corpus is proprietary and retrieval over it becomes the bottleneck
Decisions needing live data	Pure **data**; + live external truth	Adopt a provider with web search, wire internal data where needed	Search is built in, the human frames and verifies	grounding depends on internal systems and proprietary data

Boilerplate code, scaffolding, tests	Pure **data**; well-defined, low risk	Adopt any mid-tier coding assistant	Well-defined, low-risk, speed over depth	rarely – this stays off the shelf
Large refactors, multi-file changes, agentic coding	Pure **data**; your repository is the input	Adopt a coding-agent product	The coding agent is off the shelf, your repository is the only input	your repo or IP is the differentiator and needs a self-hosted agent
Legacy code modernization and migration	Pure **data**; your codebase	Adopt a frontier coding agent, point it at your repository	Off-the-shelf agent, your codebase is the input	scale or sensitivity of the codebase requires an owned pipeline

2. Embedded AI: wait or bridge

You lean on the AI shipping inside IT systems you already run. Value arrives with the platform, and the embedded AI helps reveal what to fix as your processes and data improve. The vendor picks the model, so no influence on the type of AI (Filter 3).

Problem type	Filter 1 – root cause	Best fit	Rationale	Move to Custom AI when…
Cross-departmental orchestration (finance, HR, legal pre-built workflows)	**Process/system**; generic across companies	Take the vendor's vertical agents, wait for the roadmap, bridge only if urgent	~80% is generic across companies, the platform ships it	your workflow logic is itself the differentiator
Long-running, stateful operations (research, migrations, ops investigations)	**Process/system**; undifferentiated plumbing	Adopt a managed agent runtime, do not build the plumbing yourself	State, sandboxing and audit are undifferentiated infrastructure	the orchestration encodes proprietary process or IP
Knowledge silos between systems (Jira, Confluence, Salesforce, Drive, Slack)	**Process/system**; integration-bound	Use the connectors your systems ship, switch the model underneath per cost and quality	The hard part is integration, which the systems increasingly ship	retrieval quality over your own corpus becomes the bottleneck (RAG)
Code review, PR triage	**Process/system**; runtime-data-bound	Use the AI built into your monitoring and dev stack	Value comes from your runtime data, which your dev stack already holds	review must encode proprietary standards or models
Sales and lead qualification	**Process/system**; CRM-resident	Use the AI built into your CRM	The CRM is the natural home for both scoring and conversation	scoring depends on proprietary models over your own data
High-volume tier-1 support	**Process/system**; predictable query set	Use your service-desk suite's built-in assistant, hard handoff on uncertainty	Predictable query set, your CX platform embeds it	resolution needs actions across your own systems

Multi-step process, deterministic, repetitive (month-end close, onboarding checklist)	**Process**; deterministic – not AI	Not AI: use workflow automation in or alongside your systems, add AI only at judgment steps	Deterministic steps deserve deterministic tools	n/a – not an AI problem

3. Custom AI: build your own, your data is the moat

Custom AI is the durable path. Per the decision tree these are the genuine pure-data problems where value comes from mixing AI with your own IP and data and the only sourcing mode that branches into the Filter-3 lanes. Each row therefore carries its lane: 1 mainstream LLMs & agents, 2 specialised non-LLM, 3 right-sized LLM / SLM.

Problem type	Filter 1 – root cause	Best fit	Rationale	Lane
Employees can't find information across internal PDFs, wikis, tickets, emails	Pure data + your heterogeneous corpus	Build RAG over your own corpus, watch embedded enterprise search as it matures	The bottleneck is retrieval over your heterogeneous corpus of proprietary data	1
Sensitive or regulated content that can't leave premises	Pure data + residency constraint	Self-host an open-weights model on your own infrastructure	Data residency and auditability rule out closed APIs	3
High-volume routine classification (tickets, emails, sentiment, intent)	Pure data; bounded, your label distribution	Fine-tune a small model on your own label distribution	A model fitted to your actual distribution beats a frontier LLM on bounded tasks	3
Multilingual content, including underrepresented languages	Pure data; coverage-bound	Self-host a right-sized multilingual model	Language coverage and fair tokenization are the limiting factor	3
Forecasting (demand, churn, revenue) on tabular data	Pure data; structured history	Build classical ML on your history, not an LLM	Gradient-boosted trees beat deep learning on tabular problems with history	2
Anomaly detection in transactions, logs, sensors	Pure data; rare, high-dimensional signals	Build statistical or density-based models on your signals	Anomalies are rare and high-dimensional, exactly what these models are built for	2
Recommendation and personalization	Pure data; your catalog and users	Build retrieval and ranking over your own catalog and users	Large-scale, low-latency retrieval is the wrong fit for an LLM	2
Quality control in manufacturing, vision-based defects	Pure data; your production lines	Train computer vision on your own lines, or buy a specialist vendor (which makes it Embedded)	The nature of the data on your lines decides the right tool	2

Predictive maintenance	Pure data; sensor signals + physics	Build physics-informed time-series on your sensor data	Known dynamics plus real-world noise call for a hybrid you own	2
Multi-step process with unstructured inputs and judgment	Pure data + your own tools	Build an agentic workflow over your own tools and data	It branches on natural-language signals drawn from your systems	1
Complex multi-turn troubleshooting	Pure data + your systems	Wire your account, refund and ticketing tools into an LLM	Resolution needs actions in your systems	1
Voice support	Pure data; latency-bound	Assemble a speech stack on your data, or adopt an end-to-end voice model	Latency and turn-taking are first-class concerns	2

Sources and Further Reading

AI project failure and ROI challenges

1. MIT Project NANDA – The GenAI Divide: State of AI in Business 2025 (via Fortune) – 95% of GenAI pilots show no measurable return on $ 30–40bn invested.
2. Healthcare IT News – MIT: 95% of enterprise AI pilots fail to deliver measurable ROI
3. RAND Corporation / Pertama Partners – AI Project Failure Statistics 2026 – 80.3% of AI projects fail to deliver business value.
4. Gartner – Lack of AI-Ready Data Puts AI Projects at Risk (Feb 2025) – 60% of projects abandoned through 2026 due to data readiness gaps.
5. Gartner – Gen-AI Projects to Be Abandoned After PoC (Jul 2024)
6. HBR – Beware the AI Experimentation Trap (Aug 2025)
7. HBR – AI Doesn't Reduce Work – It Intensifies It (Feb 2026)
8. ORGANIZATION SCIENCE Vol. 37, No. 2, Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality (March–April 2026)
9. NYTimes - Tech Workers Maxed Out Their A.I. Use. Now They're Trying to Minimize It (18.06.2026)

AI-attributed layoffs and reversals

10. Reworked – Klarna Claimed AI Was Doing the Work of 700 People. Now It's Rehiring
11. MLQ.ai – Klarna CEO admits AI job cuts went too far
12. Resident – IBM's AI shift: rehiring after 8,000 layoffs
13. Metaintro / Gartner – Half of AI-Driven Layoffs Will Reverse by 2027

Mainstream LLMs & agentic AI — European deployments

14. BBVA – BBVA deploys "The Eight," its strategy to transform the financial experience with AI – rollout of Gemini-powered Google Workspace to ~100,000 employees; 76%+ time savings on targeted tasks; ~3 hours saved per employee per week.
15. FStech – BBVA to roll out GenAI tools for 100,000 employees in partnership with Google
16. BNP Paribas – BNP Paribas and Mistral AI extend their partnership (May 2026) – three-year extension; Hello bank!'s HelloïZ assistant running on Mistral since January 2026, serving 1M+ clients.
17. Mistral AI – Customers page – Stellantis, Veolia, SNCF, Orange, BNP Paribas reference deployments.
18. Mistral AI – Introducing Le Chat Enterprise (May 2025)
19. Computer Weekly – How ING reaps benefits of centralising AI – 75% of customer queries automated; Microsoft 365 Copilot and engineering copilots in production.
20. SAP – New Joule Agents and Embedded Intelligence (Oct 2025) – 40+ specialised agents; up to 70% time saved on finance reconciliation; up to 33% gains in HR workflows.
21. SAP – Business AI: Release Highlights Q4 2025 (Jan 2026) – Joule Studio GA; 2,400+ Joule Skills; Microsoft 365 Copilot integration GA.
22. CIO – SAP goes all-in on agentic AI at SAP Sapphire

23. Disruption Banking – How Danske Bank is Betting Big on AI (Mar 2026) – Copilot saves retail bankers ~14% on meeting notes; 20% tech productivity target for 2026.

Specialised AI – Polysense and computer vision

25. Polysense – Company website – AI vision for food processing, founded Ghent 2022.
26. PotatoPro – Polysense raises € 2 million
27. Potato Business – Belgium's Polysense secures seed funding
28. F7i.ai – Computer Vision Predictive Maintenance Guide 2026

Right-sized models, open source and energy

29. IEA – Data centre electricity use surged in 2025 – 17% growth overall, 50% growth in AI-focused data centres.
30. IEA – Energy and AI report (full)
31. How Hungry is AI? Energy, water and carbon of LLM inference (arXiv, 2025) – Reasoning models can use 70× more energy per prompt than smaller models.
32. Microsoft – SLM vs LLM (Cloud Blog)
33. Red Hat – SLMs vs LLMs
34. DataCamp – Complete Guide to SLMs and LLMs

AI 2.0 – World models and efficient training

35. TechCrunch – Yann LeCun confirms his new world-model startup (Dec 2025) – AMI Labs, Paris; $ 1bn seed at $ 3.5bn valuation.
36. MIT Technology Review – LeCun's new venture is a contrarian bet against LLMs
37. World Labs – Company and Marble product – Fei-Fei Li et al.; commercial world model launched Nov 2025.
38. TechCrunch – Fei-Fei Li's World Labs ships Marble
39. EPFL – Mathis Lab (Bertarelli Foundation Chair) – Neuro-AI; sample-efficient learning.
40. Mathis Lab – DeepLabCut