AI 2025–2030: The Full-Stack Trajectory (From GPUs to Agents)

By Shaswat Gupta | LinkedIn | Email

Introduction

If 2023 was the year generative AI captured the world's imagination, 2025 is the year it becomes ubiquitous infrastructure. We stand at an inflection point: foundation models now rival human-like capabilities in language and beyond, yet we are *still* in the early stages of a massive transformative wave. AI today is simultaneously everywhere and nascent – powering copilot-style assistants and automations while hinting at far greater potential under the surface. As investor Elad Gil puts it, "AI is massively underhyped…generative AI [is] in its infancy."

Over the next five years, AI will transition from experimental novelties to mission-critical, omnipresent systems. This thesis takes a full-stack view – from compute infrastructure and data centers at the bottom, through model architectures in the middle, up to the application layer of enterprise and consumer tools – to map where we are now and where things are headed by 2030.

The stakes are high. Startup founders, venture capitalists, AI researchers, and enterprise engineers all have skin in the game. The coming half-decade will likely see AI woven into every industry workflow, new platform shifts, and possibly a couple of hype-driven busts along the way. To thrive, one must separate signal from noise.

We've seen this pattern before – the internet boom, cloud computing – but AI's trajectory may eclipse them. In the early 2000s, software ate the world; in the late 2020s, AI is poised to eat software, or at least heavily augment how it's built and used.

The State of AI in 2025: Foundations and Fault Lines

Compute & Infrastructure

The foundation of modern AI is massive compute power. Training cutting-edge models demands clusters of specialized hardware (GPUs, TPUs, AI accelerators) running in parallel across advanced data centers. A single top-tier AI training run (for models like GPT-4, Google's Gemini, or Meta's latest LLaMA) can involve **tens of thousands of GPUs and cost $500M–$1B** in hardware and electricity.

These extreme requirements have pushed cloud providers to build unprecedented infrastructure: custom AI chips, high-density cooling systems, and ultra-fast networking to keep thousands of chips in sync. Google's AI chief Amin Vahdat predicts this "tight synchronization and massive compute" will drive compute density to never-before-seen levels – essentially supercomputers that make today's look modest.

The good news? Costs are plummeting. Through competition and tech breakthroughs, the cost of AI "intelligence" is in free fall. Open-source models (like Meta's LLaMA series) now approach the quality of proprietary ones, creating price pressure, and new chips (e.g., AWS Inferentia, startups like Groq) are delivering cheaper, faster inference. Sam Altman has boldly predicted AI usage costs will drop 10× every year. Even if optimistic, the trend is clear: compute is rapidly commoditizing.

By 2030, access to potent AI compute will be far more democratized – potentially even on-device AI for many tasks. Apple is already running 3B-parameter language models locally on iPhones, hinting at a future where phones and laptops handle a big chunk of AI workloads. This shift toward the edge, plus ever-cheaper cloud cycles, means AI horsepower will be abundant – flipping compute from a bottleneck into a catalyst for widespread AI adoption.

Model Architecture Evolution

On top of this compute foundation, we have the brains – the AI models themselves. In 2025, large language models (LLMs) and their multimodal cousins are the prevailing paradigm. Frontier models like GPT-4, Anthropic's Claude, and Google's Gemini 2.0 are not only bigger and more complex than ever, but also more versatile.

They can juggle text, images, even audio and video in a single model, blurring the line between modalities. Google's latest Gemini models are **natively multimodal**, processing and generating text, images, audio, and video in one system. This means an AI can "see" and "hear" context, then respond in natural language – a recipe for far more seamless interactions. Some predict this multimodal prowess will eventually reduce our reliance on screens and keyboards entirely, as voice and vision become the primary interface.

(Does that mean AI will kill the smartphone? Unlikely outright – but it will change how we use devices. If your AI assistant can understand what you see through smart glasses and answer via an earbud, you won't pull out your phone as often.)

Model architectures are also evolving internally. While Transformers dominate today, new architectures are emerging to address their weaknesses. Research into state-space models and other alternatives promises models that are smaller and more efficient, yet still highly capable. Even within the Transformer camp, we see a trend toward specialization rather than just brute-force scaling.

One striking 2024 result: Meta's 8B-parameter LLaMA 3 model reportedly matched the performance of its predecessor LLaMA 2 at 70B parameters – a testament to training improvements, data quality, and techniques like distillation. In essence, **smarter small models are catching up to big models**. This suggests a future where we don't need 500-billion-parameter behemoths for every task – an efficient 5B or 50B model might suffice for many use cases, especially when paired with domain-specific data.

Models are also getting longer memories. Early LLMs struggled beyond a few thousand tokens of context, but by 2025 we have models boasting tens of thousands (Anthropic's 100k-token context) and even millions of tokens of context window. Such long contexts blur the line between "training" and "remembering"; a model can ingest essentially a whole corpus or a year's worth of conversation as context.

Still, simply making context windows huge isn't a panacea – it's expensive and not always efficient. This is why a hybrid approach is rising: **Retrieval-Augmented Generation (RAG)**, where the model smartly fetches relevant information from an external vector database rather than storing everything in its weights. Experts predict long context and RAG will converge, with models learning when to use stored knowledge versus when to call out to external data for optimal accuracy and speed.

By 2030, we expect AI systems that combine a modest core model with extensive external memory – effectively, an AI that knows when to "look up" facts or past interactions on the fly, akin to how a human might consult a notebook. This convergence addresses new bottlenecks: rather than model size, the **new**

challenges are memory and data pipelines – how to feed the right data to the model at the right time, and how to store knowledge over time.

Finally, a major shift in model evolution is the rise of **agentic behavior**. Beyond answering questions or generating content, models are increasingly designed to **take actions**. In 2025, AI agents can execute tasks on your behalf (e.g., browse the web, use software tools, compose and send emails). Research prototypes like AutoGPT stirred excitement by stringing together LLM calls to attempt multi-step goals.

The current state is clunky – often more of a tech demo – but big players are quickly baking tool-use and action into their models. Google's Gemini can natively use tools like search and even control a web browser to accomplish tasks. OpenAI's GPT-4 has a "Code Interpreter" and plug-ins that let it perform calculations, handle uploads, or call APIs. This trend points toward **embodied intelligence in software**: by 2030, your AI might be an ensemble of models orchestrated to perceive, reason, and act within various digital (and even physical) environments.

Application Layer – Enterprise, Consumer, and Developer Tools

If infrastructure and models are the engine, the application layer is the user interface connecting AI to real-world use. In 2025, we see **explosive growth in AI-powered applications** across domains. Every week a new "AI copilot" launches – coding copilots, writing assistants, design helpers, customer service bots.

Enterprises are piloting AI in customer support (chatbots that actually solve problems), marketing (auto-generating personalized content), HR (AI interviewers), and data analytics (natural language queries over databases). Consumers have conversational assistants in their messaging apps and productivity suites – e.g., Microsoft 365 Copilot integrates GPT-4 into Word, Excel, Outlook, essentially giving knowledge workers a semi-autonomous helper in routine tasks.

Developers are arguably the earliest adopters – coding assistants like GitHub Copilot and Amazon CodeWhisperer have become standard issue for many programmers, boosting productivity by suggesting code and catching errors. All told, these apps form a growing **AI layer on top of traditional software**.

At the same time, applications are emerging that *couldn't exist* before. Generative design tools that let users simply describe an image or UI and have it created, or video games with AI-driven characters that can hold conversations. Entirely new categories of consumer agents are emerging – AI companions that learn your tastes, manage your schedule, or help you shop. There's already talk that you might trust your personal AI agent with secrets more readily than a human confidant (after all, an AI won't judge you, though security remains a concern).

By 2030, interacting with AI in daily life could become as common as using Google search is today. Notably, **enterprise adoption, while enthusiastic in pilots, is slower in production** than the buzz suggests. Implementing AI at scale in a big company is non-trivial: data privacy concerns, integration with legacy systems, model accuracy issues, and the ever-present question of ROI all present hurdles.

As Dylan Fox (CEO of AssemblyAI) points out, the timeline for widespread enterprise adoption is likely slower than people think, due to many "last-mile" challenges that only surface when you try to deeply integrate AI into existing workflows. That said, the pressure is on – those that figure it out stand to leap ahead in efficiency, while laggards risk obsolescence.

The developer tooling ecosystem around AI is experiencing a Cambrian explosion: frameworks for prompt engineering, libraries to connect models with databases and APIs, monitoring tools to track AI outputs.

LangChain (for building agentic AI chains) is popular, and vector database services like Pinecone or Weaviate are in high demand.

However, as one CTO observed, many of these solutions are being rushed to market without proven ROI, and many will not survive. We're in a period of experimentation where the **middleware** is rapidly iterating. By 2030, expect a shakeout: a few robust frameworks will likely dominate, and many half-baked ones will disappear. *The wild west of "bring your own prompt library" will mature into established best practices.*

In summary, AI in 2025 is defined by **immense capabilities at the core**, a frenzy of **integration at the edges**, and plenty of kinks to iron out – from cost and scalability to reliability and user experience.

Shaswat Gupta is an MS Computer Science student at ETH Zurich, and a rank 1 gold medalist from IIT Bombay. With hands-on experience as an ML engineer at organizations like the World Bank, AB InBev, ISB, and startups, Shaswat specializes in machine learning, scalable engineering, and data-driven solutions. Known for his analytical rigor combined with a creative flair in music, poetry, journalism, and public speaking, Shaswat loves tackling challenging problems and welcomes exciting collaborations.

Connect with me:

- LinkedIn: Shaswat Gupta
- Email: shagupta@ethz.ch

Key Inflection Points Ahead (2025–2030)

Several **critical shifts** will determine how AI evolves over the next five years:

Falling Compute Costs, Rising Accessibility

The cost of AI compute is dropping **exponentially** – some estimate an order of magnitude per year – due to competition, new chips, and better efficiency. This means what was once only feasible for tech giants will become accessible to startups and individuals.

As AI becomes cheaper to run, expect a proliferation of AI services everywhere. We're entering an era where "intelligence could become nearly free to use." Cloud providers will offer more generous free tiers, and on-premise hardware will become more affordable (or unnecessary, as many tasks move to edge devices).

Implication: If you're building, don't assume current high costs will last – plan for abundance. If you're investing, consider that today's moat of "only we can afford to do X with AI" will erode quickly. The flip side is that demand for compute will also increase (people will find new uses when it's cheap), so companies that provide AI infrastructure still have a growth story, albeit with tightening margins.

Specialization of Models & Contextual Intelligence

We've likely reached diminishing returns on simply "making models bigger" for general use. The next five years will see **more specialized AI systems** – domain-specific models (for medicine, law, biology), smaller models fine-tuned for specific tasks, and architectures optimized for particular modalities.

Retrieval-augmented and long-context models will work in tandem: instead of one monolithic model trying to know everything, AI systems will route queries intelligently – sometimes using long context to absorb a large document, other times retrieving facts from a vector database. Models will learn when to use long context versus when to use external retrieval to optimize for both accuracy and efficiency.

In essence, **AI is becoming more modular** – a core reasoning engine with supplemental modules for factual lookup, images, code, etc. This specialization extends to hardware too: custom chips for specific AI workloads (training, inference, RAG queries).

Implication: The "one-size-fits-all" AI service might give way to composite AI solutions. Startups can exploit this by focusing on niches – an LLM tailored only to finance that outperforms general models, or a plugin that gives any model state-of-the-art chemistry knowledge via retrieval. For users, AI will feel more personalized because it will effectively draw on specialist knowledge when needed.

Strategic Implications and Actionable Insights

What do these trends mean for various stakeholders in the AI ecosystem? Let's translate the trajectory into tactics for startups, investors, researchers, and enterprise technology leaders:

For Startup Founders: Positioning in an AI-Everywhere World

It's a thrilling time to be an AI startup – and a treacherous one. The landscape is crowded and fast-moving, but clear strategies are emerging for how to build enduring companies.

1. **Build Moats Beyond the Model:** A startup whose main "tech" is calling an open API is sitting on quicksand. Focus on what *only you* can build – proprietary datasets, unique interfaces, or novel model tweaks. Consider going vertical: many general AI capabilities are commoditized, but applying them skillfully in a specific sector with domain expertise creates a defensible position. The most useful data is often tied to specific real-world use cases, leaving room for new entrants who understand those niches.

2. **"AI-First" Means Solving Real Problems:** In 2023, many products were essentially tech demos. That won't cut it anymore. Start with real pain points and figure out how AI can solve them better. If your product is 90% AI magic without fitting a business need, you may get early users but not lasting customers. Approach AI as a tool, not the product itself. Design your user experience so users can accomplish goals without wrestling with prompts or babysitting the AI. Hide the complexity; make the AI's role feel natural within a larger application. Successful AI startups often won't market themselves as "AI companies" at all – they'll be solutions providers with AI under the hood.

3. **Stay Agile and Infrastructure-Agnostic:** Design your architecture to be modular and model-agnostic, allowing you to plug in new models or switch providers with minimal pain. Build a layer in your code to abstract your model choice, so tomorrow you can swap it for a cheaper or better one without rewiring everything. Avoid over-reliance on any one platform. Account for AI's rapid rate of change in your roadmap. Make calculated bets on what you develop in-house versus what you expect the ecosystem to solve.

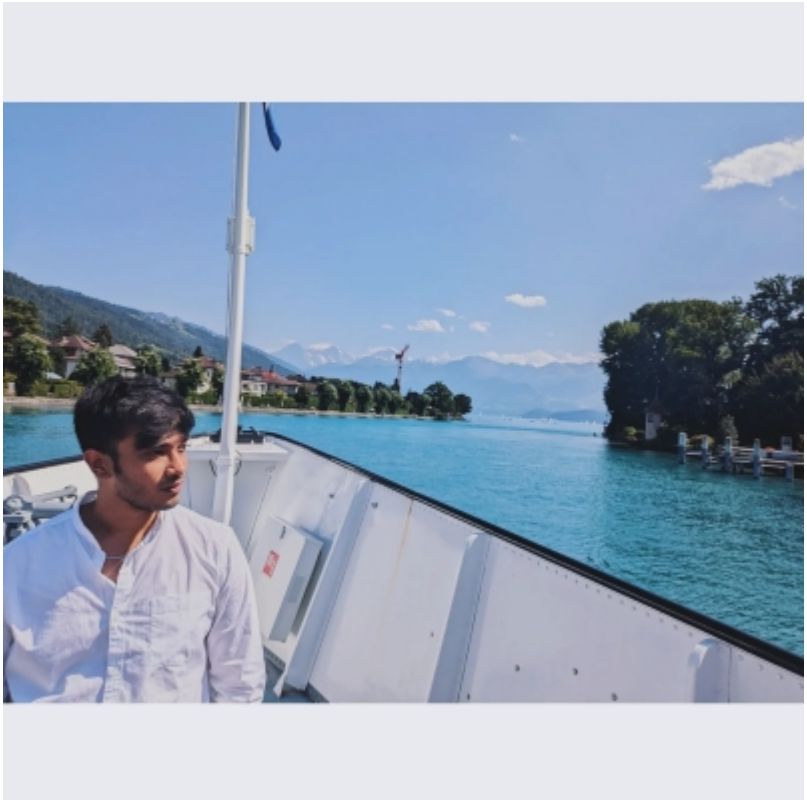
4. **Don't Obsess Over Short-Term ROI – Invest in Learning:** An over-focus on immediate returns can be limiting. AI is still evolving rapidly and true breakthroughs may not have an obvious short-term ROI. Now is not the time to demand hard ROI on every AI project; instead, use any cost savings AI gives you to reinvest in ambitious AI-driven innovations. The real promise of AI is not just doing the same for less cost, it's doing *new things that were never possible before*. Foster an environment of experimentation – some AI ideas will flop, but you're learning and staying ahead of the pack.

5. **Ensure AI Fluency Across the Team:** Every team member should feel comfortable using AI tools. Train your staff on prompt engineering basics, interpreting model output, and understanding limitations and ethical pitfalls. Encourage engineers to play with the latest APIs and open-source models. Create an *AI-native culture* so integrating AI isn't an afterthought – it's built into how you brainstorm solutions. When recruiting, test for AI aptitude. In 2025, being "fluent" with AI is akin to being internet-savvy in 1999; by 2030 it will be assumed, so get there ahead of time.

The bottom line for founders: combine *strategic focus* (solving a real problem with a moat around your solution) with *tactical flexibility* (adapting to the fast-changing tech), while nurturing a team that "speaks AI" as a second language.

AI Startup Strategy	Description & Example	Key Advantages	Risks & Challenges
General-Purpose AI Wrapper	Use a public model API with a nice UI (e.g., a GPT-powered chatbot for everyone)	Fast to market; broad initial appeal	**Highly commoditized** – easy to replicate; little proprietary tech. Low moat, vulnerable to API price changes or model updates
Vertical AI SaaS	AI tailored to a specific industry or function (e.g., AI contract analysis for legal firms)	Deep domain integration; leverage proprietary data; clear ROI for target users	Harder to expand beyond niche; requires domain expertise and sales cycle. But less competition if executed well

AI Startup Strategy	Description & Example	Key Advantages	Risks & Challenges
Infrastructure/Picks & Shovels	Provide tools or infrastructure for others to build AI (e.g., a vector database, custom AI chip, MLOps platform)	Technical moat if done well; can capture wide user base. "Arms dealer" strategy in gold rush	Capital intensive; must compete with big cloud providers or open-source alternatives. Risk of being squeezed if giants offer similar in-house solutions
Developer Tooling & Middleware	Build frameworks or APIs to make AI development easier (e.g., orchestration library, prompt optimization tool)	Can become a standard if widely adopted; network effects. Opportunity to ride the wave of every new model	Middle-layer **pressure** – if base models add those features natively, your value shrinks. Many tools will be free or open-source, hard to monetize unless truly superior
Data-Rich Solutions	Acquire or utilize unique datasets to power AI solutions (e.g., agriculture AI using proprietary satellite and sensor data)	**Data moat** – hard for others to replicate your training data. High value insights and performance in your domain	Obtaining and maintaining exclusive data is challenging; must navigate privacy and compliance. Need strategy to continually enrich the data advantage



Shaswat Gupta is an MS Computer Science student at ETH Zurich, and a rank 1 gold medalist from IIT Bombay. With hands-on experience as an ML engineer at organizations like the World Bank, AB InBev, ISB, and startups, Shaswat specializes in machine learning, scalable engineering, and data-driven solutions. Known

for his analytical rigor combined with a creative flair in music, poetry, journalism, and public speaking, Shaswat loves tackling challenging problems and welcomes exciting collaborations.

Connect with me:

- LinkedIn: Shaswat Gupta
- Email: shagupta@ethz.ch

For Venture Capitalists: Where to Double-Down

VCs are pouring money into AI – but where will it yield outsized returns versus getting washed away in hype? Here are some strategic thoughts for investors:

- **Pick Shovel Sellers (Carefully):** The classic "picks and shovels" play makes sense in AI, but with a caveat: Big Tech is also building shovels. The sweet spot for investment might be tools that are *platform-agnostic or truly best-in-class* such that even the big clouds adopt them. Think AI security, auditing and compliance, or hyper-optimized model training tools. Developer communities can signal promising picks/shovels – if a tool rapidly gains adoption, that's a flag to investigate. Does it solve a persistent pain point that won't vanish as models improve? Enterprises will *always* care about search over their internal data – so vector search tech has a strong foundation. In contrast, a product for "prompt versioning" might become moot if prompting paradigms shift.

- **Data Moats and Vertical AI:** Some of the biggest near-term wins may come from applying AI in *data-rich domains that have been underserved*. Sectors like **media, SaaS, and biology** are ripe for AI transformation. In biology, "genome language models" can design new proteins or predict DNA sequences – an area where having the right data is key. A startup with exclusive partnerships or datasets could become extremely valuable. Similarly, in traditional SaaS, incumbents are slow; a startup that rebuilds a vertical SaaS product with generative AI woven throughout could rapidly gain traction. Also consider less sexy industries: construction, logistics, agriculture. They produce lots of data and have labor-intensive decision processes – perfect ground for AI copilots. Ask "does this team have a data or distribution advantage in this vertical, and is the AI truly adding value?" Incumbents often don't leverage their data well, and startups can seize that gap.

- **AI-Driven Biology and Robotics:** Beyond software, AI is starting to crack open the physical world. Foundation models for robotics, manufacturing, and healthcare are set to leap forward. Investing in the intersection of AI and the real world could be like investing in the early internet of things, but turbocharged. These often need more capital (hardware is involved), but the moats can be deep. A robotics AI that collects proprietary data from thousands of warehouse deployments creates a virtuous cycle hard for a pure-software competitor to emulate. Similarly in drug discovery – AI that understands chemistry or genetics can unlock new IP extremely valuable. We're seeing the early innings of AI moving from bits to atoms. VCs should not ignore "hard tech" AI startups thinking they're too far from revenue – the timing is getting right for these to make commercial sense.

- **Be Wary of the Trendy Middle Layer:** Exercise caution with investments whose core value could be subsumed by an update from OpenAI, Google, or open-source communities. Scrutinize whether middleware is adding enough value beyond what base models will soon provide. Is it an independent developer community *or* big enterprise contracts that give it staying power? If the startup's pitch is "you don't have to know how to prompt ChatGPT, our UI does it," that's likely not investable now. On

the other hand, "we offer fine-grained control and monitoring for model outputs in a regulated industry" addresses a serious need less likely to be nullified by model improvements.

- **Patience for New Business Models:** The final business models for AI haven't been figured out yet. We're seeing usage-based pricing, seat licenses, infrastructure reselling – but there might be completely new ways AI value is monetized. VCs might need flexibility in evaluating companies. Some may not have typical SaaS metrics if they're usage-heavy; others might be open core or services-heavy as they figure things out. It will take time to understand what "good" margins or LTV/CAC look like in certain AI businesses. Value might accumulate in different layers than expected – perhaps in consulting and integration, or in usage tolls on foundation models. A balanced portfolio might hedge across these possibilities. Stay close to research – breakthroughs in labs can spawn new startups or make others obsolete. For every deal ask: *If AI capability becomes X times cheaper/faster, does this company become more valuable or less?* Invest in those positioned to ride the wave, not be drowned by it.

For AI Researchers and Lab Leaders: Pushing the Frontier

Researchers in academia or corporate AI labs (DeepMind, OpenAI, FAIR, etc.), while not directly driven by market strategy, still influence and are influenced by these trends. Here's what the trajectory means for the R&D side:

- **Beyond Scale: New Architectures and Methods:** Pure scaling of Transformers has given us a lot, but we know it's not the end game. Research should double down on **alternative approaches** that could leapfrog current models in efficiency or capability. This includes investigating *state-space models, recurrent architectures, neurosymbolic hybrids*, and more. Companies like Cartesia are working on state-space models that outperform Transformers of similar size – these hints should be pursued vigorously. By 2030, it's quite plausible the winning AI designs will look different internally from 2025's Transformers. Techniques like *fine-tuning via distillation, low-rank adaptation (LoRA), and advanced pruning/quantization* deserve focus – making models smaller and faster without losing power is incredibly high-impact (it's how we got an 8B model matching a 70B model). The research community has a mandate to **keep improving the "quality per compute" of AI**. If today we need 100B parameters and 1000 GPUs to solve a task, can we do it with 10B and 100 GPUs?

- **Long-Term Memory and Autonomy:** Another frontier is giving AI something akin to a memory or the ability to continually learn. Currently, LLMs are mostly stuck with their fixed training and a bit of context. Research into **long-term memory** (via retrieval, external knowledge graphs, or dynamic training updates) is crucial. An interesting concept is an AI that *self-updates* – not quite AGI self-improvement, but a system that can learn from new data on the fly without a full retrain. Additionally, if we want reliable agents, we need research on *planning algorithms, goal decomposition, and safety in autonomous decision-making*. By 2030, we'd like agents that don't just have one-shot prompt->response behavior, but can carry a *thread of thought* safely over hours or days.

- **Evaluation, Transparency, and Trustworthiness:** As AI integrates into sensitive domains, the demand for *explainability* and robust evaluation skyrockets. Research must provide better ways to understand what these models know and don't know, and to systematically test them. The era of "prompt and pray" is ending – instead of treating an LLM like an oracle, we need to architect **AI systems with predictable reliability**. One approach is the *orchestration of multiple models/tools* for cross-checking. Research into such **compositional AI systems** (and how to formally verify parts of their behavior) will be highly valuable. By 2030, society will likely require that AI decisions, especially in regulated areas, come with an

audit trail. There's also a role for **Web3 and decentralization** in trust – ideas like using blockchain for auditability of AI agent actions. While "AI + blockchain" has seen hype, *specific* benefits like transparent logging might find practical use.

- **Multimodal and Embodied AI:** The next five years will also reward research that breaks the AI out of the text box. Multimodal models that truly understand *visual cues, audio inflection, spatial and physical context* are key to AI becoming more human-like in interaction. In robotics, combining vision, language, and motor control is a grand challenge – and progress here will literally bring AI into the real world. By 2030, a household robot that can tidily arrange your living room based on a verbal request might be within reach. *Emotions and personalization* are another aspect: can AI detect human emotions and adjust its responses? The aim is a more **human-centric AI**, not just in ethical terms but in interaction quality.

- **Collaboration and Interdisciplinarity:** Finally, researchers should maintain close feedback loops with real-world deployments. It's a fascinating time where academic ideas become products in months. Engaging with the open-source community can keep research grounded. It might also be wise to collaborate with social scientists, policy experts, and designers – AI is no longer just a comp sci problem. In sum, the path to advanced AI that is *useful, safe, and ubiquitous* will be paved by research that emphasizes efficiency, autonomy with guardrails, transparency, and embodiment.

For Enterprise Tech Leaders & Engineers: Adapting and Thriving

If you're leading technology (CTO, CIO, or an engineering manager) in a non-AI-focused enterprise, you might feel equal parts excitement and anxiety about AI. The mandate is clear: figure out how to leverage AI for your business, or risk being outpaced. Here's how to stay ahead:

- **Start with Low-Hanging Fruit (But Aim Higher):** In 2025, many enterprises have done the basics – maybe added a chatbot on the website, or started using a SaaS AI tool for summarizing documents. Identify where AI can quickly add value in your operations: customer service, internal knowledge management, and software development are common areas. Implementing a pilot can deliver quick wins and build organizational confidence. However, don't stop there. Make a roadmap for more transformative AI projects over the next 3–5 years. Could your company have an *AI-powered analytics agent* that proactively finds insights in your data lakes? The key is to **avoid complacency**: enterprise AI adoption will take time but it *will* happen, and those investing early will have solved the kinks by the time others start. You don't want to be scrambling in 2029 to catch up.

- **Invest in Data Readiness:** Enterprises often underestimate how much prep is needed to effectively use AI. If your CRM, ERP, or document repositories are full of messy, siloed, or unlabeled data, even the best AI won't magically fix that. A significant portion of your AI budget should go into **data engineering**: consolidating datasets, cleaning and labeling where possible, and setting up pipelines that feed fresh data to AI models. Consider building a **feature store or vector database** that captures your company's knowledge in a form AIs can use. *Making internal knowledge accessible is a killer app*. Address data governance: mark which data is sensitive and decide whether you'll use cloud AI or require on-prem solutions. By solving data issues early, you set a strong foundation for any AI initiative.

- **Empower and Educate Your Workforce:** Introducing AI into workflows can be as much a people challenge as a tech challenge. As a tech leader, you should champion a program of **AI upskilling** and tool adoption. The goal is to **build AI fluency enterprise-wide**. When staff see AI as a *partner* rather than a threat, adoption soars. Rethink your org structure and roles: you might establish new positions

like "AI Product Manager" or "Prompt Engineer" to help different departments implement AI solutions. Encourage teams to incorporate AI in their projects – maybe set KPIs like "by Q4, propose at least one AI augmentation in your process." And yes, some roles will evolve – a marketer might become more of a curator/editor of AI-generated content. Plan for retraining where needed. Companies that successfully navigate this will unlock *topline growth* – doing new things that drive revenue, not just cost-cutting.

- **Architect for Integration and Flexibility:** Add AI capabilities in a way that's modular and future-proof. Set up a proper **AI platform or middleware** in your organization: a layer that can interface between various AI services and your internal systems. Monitor usage and outcomes of AI – logging is crucial for debugging when the AI makes a mistake. You'll want a feedback loop where employees can flag incorrect outputs, triggering a review or improvement. Given AI's rapid evolution, your tech stack should be ready to incorporate new models or methods. Being tied completely to one vendor's AI could become a strategic risk, so maintain flexibility. Consider hybrid approaches: cloud AI for some tasks, smaller on-prem models for sensitive data. Build your infra to handle both. This ties into cost management: an internal platform can route requests – high-priority ones to the best (expensive) model, low-priority ones to a cheaper model – optimizing spend.

- **Governance, Ethics, and Oversight:** Enterprise leaders must ensure AI is used responsibly. Put policies in place: guidelines on employees using public AI services (to prevent data leaks), or rules on when AI decisions need human review. Establish an **AI ethics committee or review board** if you're deploying AI that affects customers or employees significantly. Test your models on different demographics to ensure no unintended discrimination. Get legal/compliance teams involved early – regulators worldwide are formulating AI regulations. In highly regulated contexts, using more interpretable models that you can fully audit might be wiser than an opaque giant model. Start AI usage in an assistive capacity (human-in-the-loop), gather data on its behavior, then gradually automate more as confidence grows. Be **transparent with users** when AI is involved and have recourse if things go wrong. Remember, *AI can amplify your company's capabilities, but also its mistakes*.

In short, enterprise engineers and tech execs should treat AI as both a new tool and a new team member. It needs onboarding (data and integration), training (fine-tuning and feedback), supervision (governance), and inclusion in the team (upskilling colleagues). Done right, it's like adding a whole cadre of ultra-efficient interns that, over five years, may grow into invaluable lieutenants in your organization.

From Copilots to Co-Creators: The Road Ahead to 2030

As we cast our eyes to the end of the decade, a few **long-term shifts** stand out:

- **Compound Agent Workflows Will Transform Work:** Today we have one AI assistant for coding, one for writing, maybe another in our email. By 2030, these may converge into **compound agents** – multi-capability AI entities that handle complex projects by orchestrating multiple skills. You might simply assign a high-level goal ("Launch a marketing campaign for product X"), and a fleet of coordinated AI sub-agents will handle research, copywriting, generating images, scheduling posts, buying ads, etc., consulting you only for approval or creative direction. This is an extension of the "agentic AI" trend, but at a higher abstraction: entire workflows automated, not just tasks. It will feel less like using a tool and more like collaborating with an autonomous team. Human workers will focus more on defining objectives, providing feedback, and handling nuanced decisions or relationships. This shift could dramatically boost productivity, but it will also require rethinking job roles and developing trust in these agents. *Co-pilots will evolve into co-creators and co-executors*. Smart enterprises will prototype these

workflows early – those who master compound workflows will outpace competitors still using siloed tools.

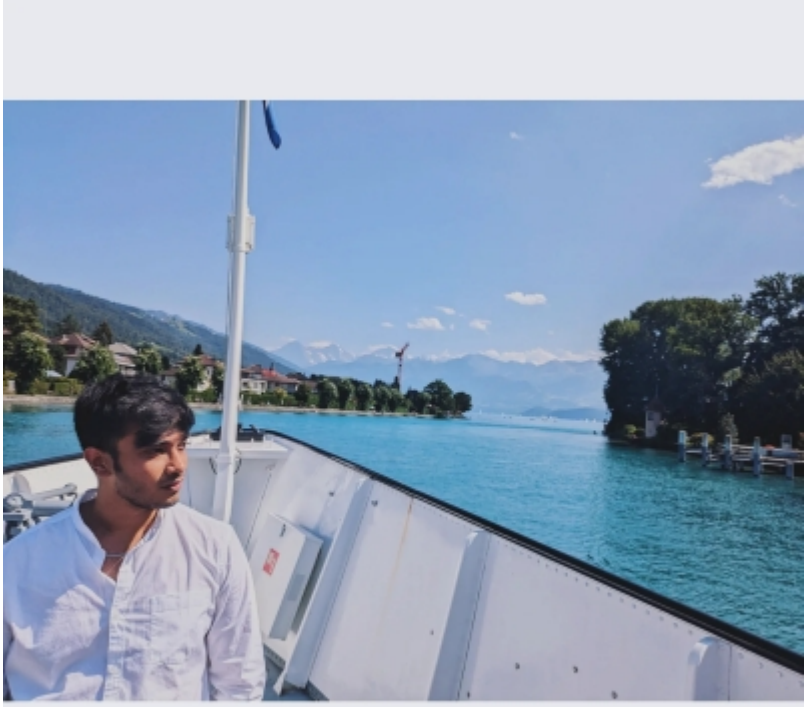
- **From Prompt Engineering to UX Orchestration:** The early days of using LLMs involved crafty prompt engineering – that era is fading. The "prompt and pray" approach is giving way to robust AI systems design. By 2030, interacting with AI will be less about writing clever prompts and more about designing *experiences* where AI seamlessly fits in. The emphasis will shift to **UX orchestration**: how to structure multi-turn interactions, how to present AI suggestions, when to have AI take initiative vs. wait, and how to recover from errors. The *user experience* around AI becomes the competitive differentiator, not the raw model output. We can expect new tools and design paradigms that let creators prototype AI interactions visually or with minimal code. "Prompt engineering" may disappear into the backend, much like assembly coding gave way to high-level programming. The best AI-driven products will be ones that users hardly realize have AI under the hood – they'll just solve problems naturally. This means **interdisciplinary collaboration**: designers, psychologists, and AI developers working together to craft AI behaviors that feel intuitive and trustworthy. Companies will also invest in *orchestrating multiple AIs* together for reliability – e.g., one AI generates an answer, another checks it, a third summarizes it in a friendly tone, all behind a single interface. The skillset emphasis shifts: less on knowing model quirks, more on system design and user empathy.

- **AI and the Commoditization of Software Development:** As AI takes on more coding and design tasks, the barrier to creating software drops precipitously. By 2030, a solo entrepreneur might build a complex app by chiefly describing what they want to an AI. This will lead to a flood of software – for every niche problem, there could be a tailored app because development costs will be so low. Consequently, **software itself becomes more of a commodity**. The unique value will not lie in basic features – those can be spun up by AI in days – but in community, data network effects, continuous learning, and brand trust. For businesses, this means the moats shift: owning a user base or proprietary dataset will be more defensible than having fancy features. We might see "*disposable software*" – apps spun up for a one-time use or short-term project, then discarded. This could challenge traditional software business models. Perhaps more software becomes free or open source, with companies monetizing usage of their AI models or the data collected. It may accelerate the trend of **software as a service** (SaaS) turning into **software as a commodity utility** – basic applications essentially bundled with other services or provided at very low cost, with revenue coming from AI insights on top. The role of human software engineers will evolve to more of *architects, integrators, and overseers*. There will still be a need for experts to design complex systems, ensure security, and maintain quality, but much coding could be automated. This democratization means more innovation everywhere, but also means incumbents must innovate faster since newcomers can replicate their products with AI help. Companies will focus on the *creative and human-centric aspects* of software – understanding user needs deeply and creating delightful experiences.

In essence, by 2030 AI will be less visible as a standalone "wow" factor and more ingrained in every digital interaction. We'll talk less about AI itself and more about what it enables – new business models, new ways of working, perhaps even new societal structures. For all stakeholders – startups, investors, researchers, enterprise leaders – the throughline is: **adaptability and vision**. Those who ride the wave of AI's evolution, updating their tactics as the landscape shifts, will find tremendous opportunity. Those who stick to old paradigms risk irrelevance in a world where AI capabilities compound year over year.

To conclude: In the next five years, AI will go from *predicting the next word* to *shaping the next world*. The journey from here to 2030 will be one of turning science fiction into mundane reality. The early days of

"prompt and pray" are over – the era of **"plan and prosper"** with AI has begun. Whether you're building a startup, investing millions, researching novel models, or implementing AI in a Fortune 500, the mandate is the same: **embrace the full-stack AI revolution**. The companies and individuals who combine technical savvy with strategic insight will not only navigate the coming shifts – they'll drive them. And perhaps when we look back from 2030, we'll marvel at how far we've come, noting that the wild ideas we pondered in 2025 are now simply *business as usual*.



Shaswat Gupta is an MS Computer Science student at ETH Zurich, and a rank 1 gold medalist from IIT Bombay. With hands-on experience as an ML engineer at organizations like the World Bank, AB InBev, ISB, and startups, Shaswat specializes in machine learning, scalable engineering, and data-driven solutions. Known for his analytical rigor combined with a creative flair in music, poetry, journalism, and public speaking, Shaswat loves tackling challenging problems and welcomes exciting collaborations.

Connect with me:

- LinkedIn: Shaswat Gupta
- Email: shagupta@ethz.ch
