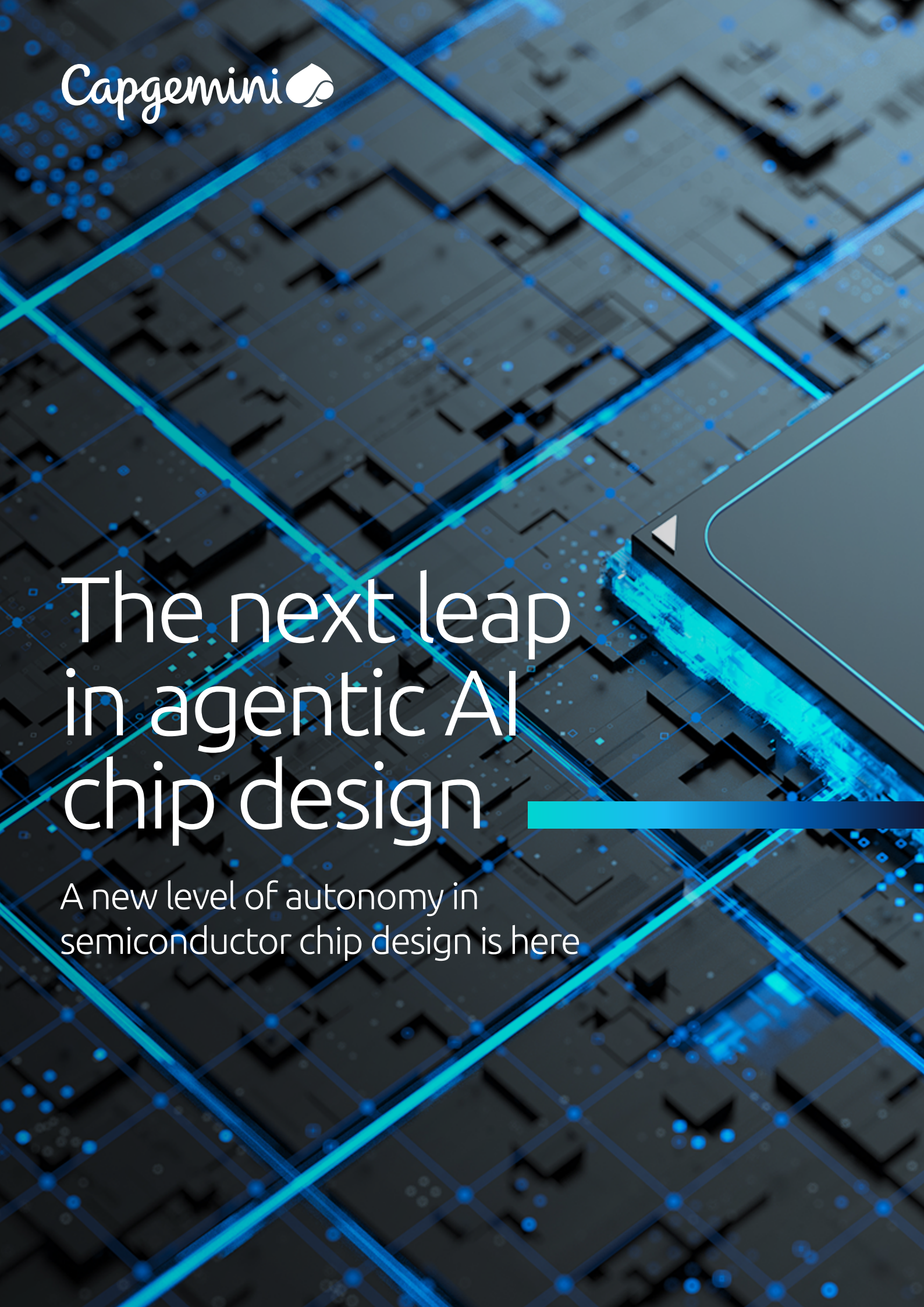
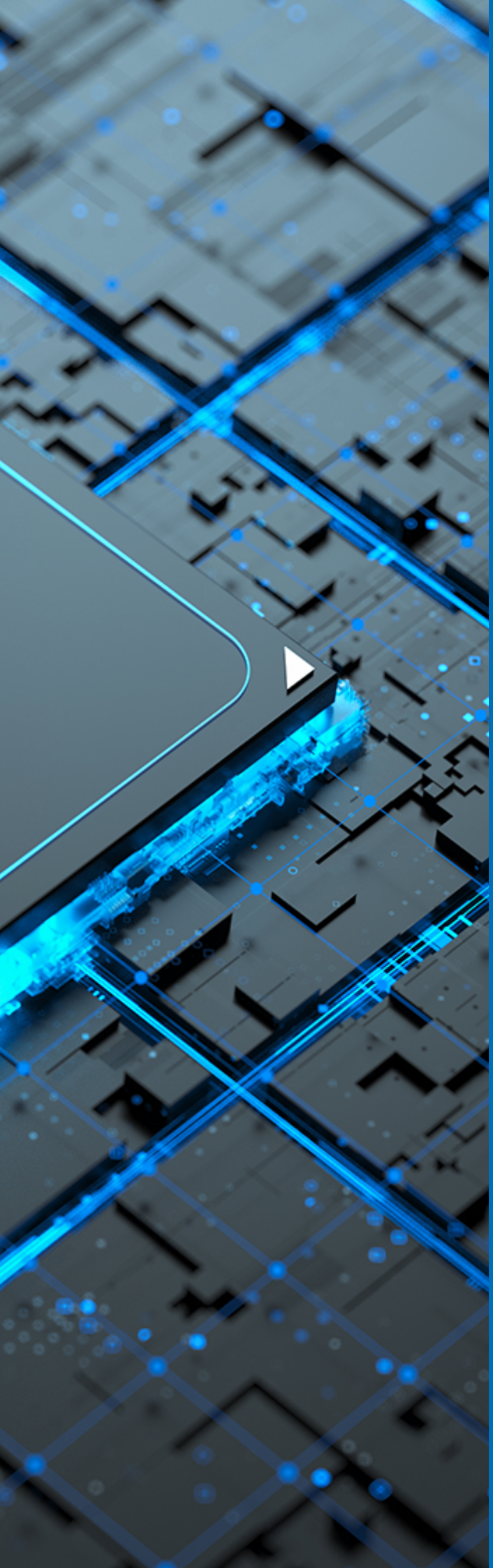

Capgemini
The next leap in agentic AI chip design
A new level of autonomy in semiconductor chip design is here



Executive summary

Semiconductor engineering is entering an interesting phase of AI adoption: moving from answering questions to executing governed workflows. Agentic AI can automate connected front-end and back-end tasks such as specification decomposition, RTL design, testbench generation, regression triage, DFT, PnR, Sign-off and technical documentation.

The winning pattern is a multi-agent system: smaller models for routine parsing and templating, selective use of stronger reasoning models for reviews, and non-LLM tools (compilers, linters, simulators, formal engines) as the final source of truth..

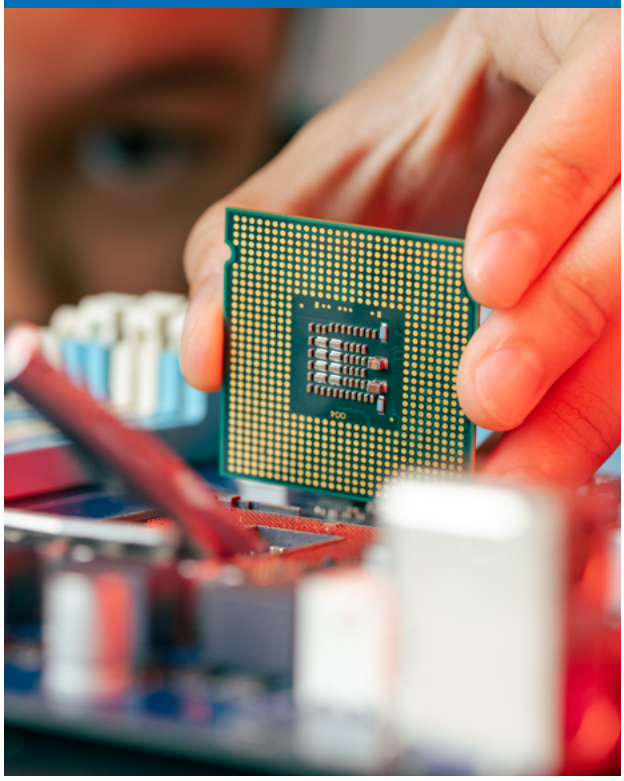
To scale responsibly, teams must implement IP protection, project isolation, human oversight for high-impact decisions, and traceable audit logs aligned to a recognized Code of Ethics for AI.

Why chip design has been ‘AI-slow’ – and what changed

Chip design was slow to adopt AI because reliability must be provable, not assumed. Historically, foundation models lacked a robust understanding of HDL and verification semantics, as well as the cross-artifact dependencies that encode true design intent – RTL, System Verilog assertions, constraints, development flows, and EDA tool reports must all remain consistent. In silicon programs, release schedules are tightly coupled to foundry and DFM closure, making any “AI-suggested” change without sign-off evidence operationally risky. The defect cost curve is steep: CDC, reset, power-intent issues, or faulty sign- off can trigger re-spins, yield loss, or field failures. Moreover, micro-architecture decisions and PPA optimizations are core intellectual property, so strict confidentiality constraints limit what can be exposed to general-purpose models.

What has changed is the emergence of tool-verified agentic workflows. Here’s what that looks like: Agentic workflows are like an assembly line – end-to-end processes made up of agents. “Tool verified” means that every action they take is backed by deterministic evidence – compiled, linted, verified, simulated, and formally checked. Multi-agent orchestration[3, 4, 5] allows tasks to be routed based on complexity and workflow stage: smaller models handle parsing and templating, while more powerful models focus on reasoning and cross-artifact analysis. EDA tools remain an essential part of flow. This combination enables meaningful automation without compromising sign-off discipline or engineering accountability.

This shift raises a critical next step – understanding how agentic systems work and where they deliver value across the chip development flow.



Assistants vs agentic AI: The technical distinction

Early AI adoption in engineering largely focused on assistants – tools that search, summarize, or generate code snippets. While useful for individual productivity, they operate as isolated point solutions and do not manage end-to-end workflows.

Agentic AI, in contrast, introduces orchestration. It can plan multi - step tasks, invoke tools, observe outcomes, and iterate within defined policies and guardrails.
In practice, an agentic system for chip verification behaves like a workflow engine. A planner starts with a goal, such as improving coverage closure into concrete tasks, and decomposes it:

01 Search

02 Analyze

03 Generate

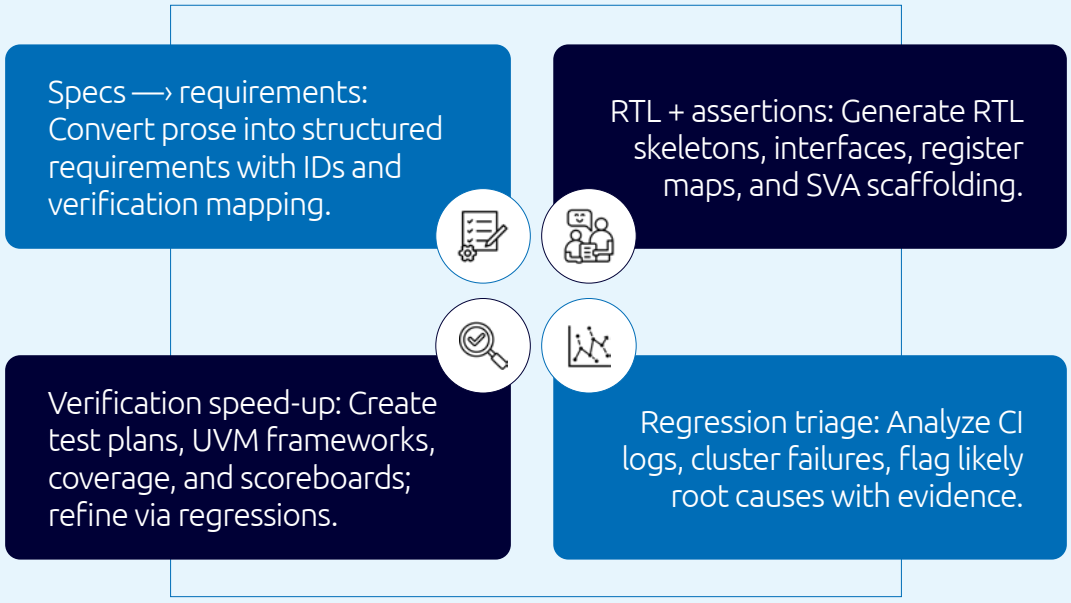
04 Validate

05 Report

Execution agents perform scoped actions, or script-based automation while validator agents run deterministic checks – simulations, regressions, coverage analysis and produce too - backed reports and evidence.

Where agentic workflows deliver value across the lifecycle

Front-end is the strongest fit since artifacts are text-heavy and verifiable. Use cases include:



Back-end benefits are bounded as sign-off is tool-driven. Agents can draft constraints, suggest QoR experiments, summarize timing/DRC/IR, and propose ECOs – but all changes require formal validation.

While the value is clear, practical adoption requires a disciplined approach across data, model orchestration, governance, and sustainability within the chip development flow.

Applicability in the chip development flow

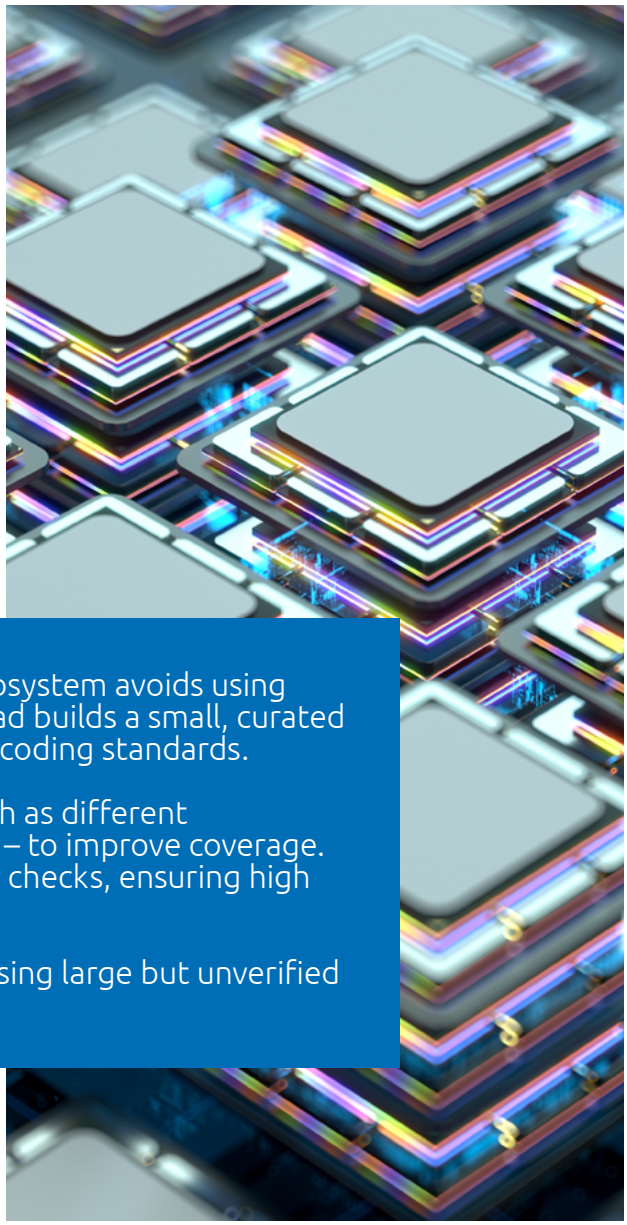
Applying agentic AI in semiconductor design requires careful consideration of several factors. These include where in the design flow it is deployed and the degree of dependence on EDA tool–generated databases, which often limits autonomy due to proprietary constraints.

Guardrails and data isolation are essential to protect sensitive IP, while the availability of high quality training data is critical to reducing hallucinations and response drift. Finally, environmental impact and sustainability must be factored in as compute usage scales.

Data strategy: Quality over quantity

Semiconductor datasets are inherently scarce because they are proprietary. As a result, reliability depends less on data volume and more on curation and grounding. An effective strategy combines curated internal standards, retrieval augmented generation over approved design artifacts, and synthetic data to increase coverage without exposing IP.

Synthetic data helps models learn structural patterns rather than memorize content. This includes small reference designs, constrained variants, and injected defects such as reset issues or CDC violations, labeled using deterministic checkers. These datasets must reflect real tool constraints and be continuously validated against golden regressions using metrics such as compile success, regression pass rate, assertion precision, and coverage efficiency.



For example, A team developing a UART subsystem avoids using large volumes of production data and instead builds a small, curated dataset from approved interface specs and coding standards.

They enrich it with synthetic variations –such as different configurations and common design defects – to improve coverage. Each dataset is validated against tool-based checks, ensuring high reliability.

This approach delivers better results than using large but unverified data.

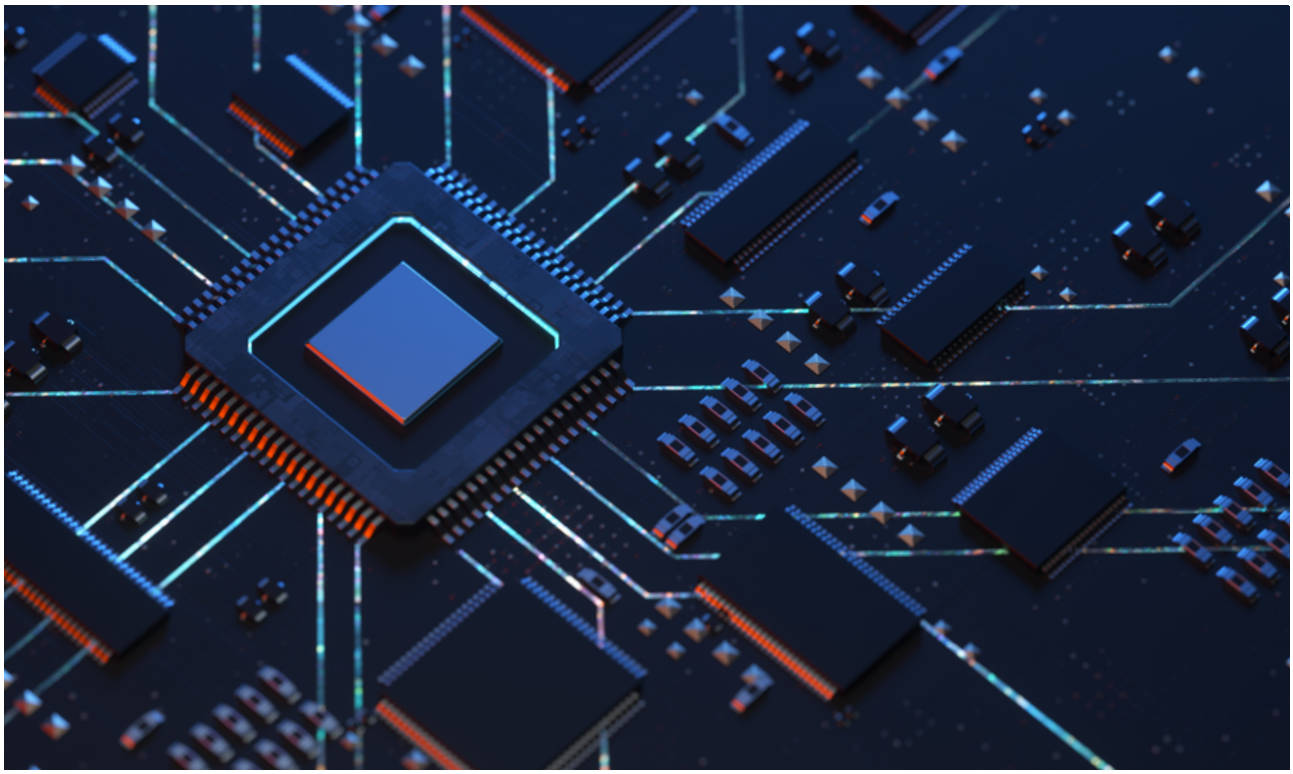
Model Selection and Routing

Chip design workflows do not depend on a single large model. Instead, they gain from model routing [3,4,5] where lightweight models handle tasks such as log analysis and template generation, while more advanced models focus on refactoring, debug reasoning, and cross-artifact validation. In every case, results should be accepted only when backed by tool-verified evidence.

Mixture of Experts (MoE) architectures are relevant since they allow conditional computation, improving efficiency for a given capability envelope. However, they also add system complexity across routing, memory, and deployment. Model selection should therefore be driven by workflow-level metrics e.g. chip-design knowledge coverage, latency, cost, accuracy, and energy consumption, rather than only considering token count and context size.

Example: In a verification flow, simple tasks such as parsing regression logs or generating test templates are handled by smaller models, while complex debugging of failing testcases is routed to a more capable model.

This avoids overusing heavy models for routine work, reducing cost and turnaround time.



IP protection, governance, and ethical AI

In semiconductor environments, responsible AI starts with confidentiality. A minimum baseline includes strict project isolation, least privilege access, and on prem or VPC deployment for sensitive programs. All agent actions – prompts, tool calls, diffs, and approvals – must be logged, and high impact decisions should require human review.

Aligned with Capgemini's Code of Ethics for AI, agentic chip design systems must emphasize transparency, accountability, robustness, and governance – clearly labeling AI generated artifacts, maintaining audit trails, and enforcing escalation policies.[1, 2]

Example: In a multi-project environment, each chip program runs in an isolated workspace to prevent unauthorized access and IP violations. Role-based access restricts who can view or modify design data.

All AI-assisted changes are logged and audited, with suspicious activity flagged and critical updates requiring approval to prevent misuse.

Sustainability: Make it measurable

Compute and energy are real constraints. Efficiency should be engineered by right sizing models, caching embeddings, reusing tool outputs, and favoring deterministic checks over open ended reasoning. Sustainability becomes actionable only when measured – through KPIs such as cost per verified change and time saved per triaged regression. [2]

These enablers define the path forward – shaped by Capgemini's practical experience in agentic chip design – raising the key question of what level of autonomy can realistically be achieved.

Example: A design team reduces compute usage by right-sizing models, reusing simulation results, and avoiding unnecessary runs.

This lowers energy consumption while maintaining output quality. Impact is tracked through metrics such as cost per validated change and compute saved per regression – aligning engineering efficiency with broader sustainability goals.

What autonomy realistically means for chip design

Agentic AI is positioned to improve work efficiencies. The immediate value of agentic AI lies in enabling more autonomous engineering workflows that can plan tasks, generate artifacts, invoke deterministic checks, and assemble structured review packages – while keeping sign-off decisions deterministic and under human accountability. In the near term – the next one to two years – it is expected to substantially reduce the manual glue-work between tools, design artifacts, and execution hand-offs. This includes task coordination, workflow sequencing, log analysis, result correlation, and preparation of review-ready evidence. At the same time, EDA platforms are beginning to evolve with AI-assisted and AI-driven capabilities. While this evolution has already started, broader structural changes are likely to emerge gradually across production flows.[6]

Several commercial EDA platforms have already introduced AI-assisted debugging, verification acceleration, and flow automation capabilities, signaling early convergence between EDA ecosystems and agentic workflows.

Within this context, Capgemini's Electronics and Semiconductor practice has developed a portfolio of highly customized, on-premise deployable solutions spanning multiple phases of the chip development lifecycle. These range from domain-focused assistants such as CHIP-IQ to more advanced capabilities for smart debugging, workflow automation, and test acceleration.

A dedicated Center of Excellence (CoE) drives these efforts, developing new agentic capabilities while continuously evolving existing solutions in line with advancements in AI technologies and the ongoing evolution of EDA platforms.

About the author

Deepak Gupta
Senior Director
Solutions Architect at Capgemini

A thought leader focused on applying AI to transform chip design and semiconductor engineering workflows.

References

1. **Capgemini – Code of Ethics for AI & Capgemini – Ethical AI :** Capgemini Group framework defining responsible, transparent, and trustworthy AI design and deployment. Capgemini guidance on embedding ethics, governance, and human oversight across AI solutions.
2. **Capgemini Research Institute Developing sustainable Gen AI**
3. **Shazeer et al., "Outrageously Large Neural Networks":** Foundational research paper introducing Mixture of Experts (MoE) architectures for scalable AI models (Google Research).
4. **Hugging Face – Mixture of Experts Explained:** Practical technical overview of MoE architecture, design trade offs, and deployment considerations in modern large language models.
5. **Liu et al., "A Survey on MoE Inference Optimization" (arXiv):** Comprehensive academic survey covering performance, scalability, and energy efficiency challenges in MoE inference systems.
6. **EDA Industry Blogs (Synopsys, Cadence, Siemens):** Public industry perspectives on the application of AI and ML across RTL design, verification, physical implementation, and sign off workflows.

About Capgemini

Capgemini is an AI-powered global business and technology transformation partner, delivering tangible business value. We imagine the future of organizations and make it real with AI, technology and people. With our strong heritage of nearly 60 years, we are a responsible and diverse group of over 420,000 team members in more than 50 countries. We deliver endto-end services and solutions with our deep industry expertise and strong partner ecosystem, leveraging our capabilities across strategy, technology, design, engineering and business operations. The Group reported 2025 global revenues of €22.5 billion.

Make it real
www.capgemini.com