AiNekko

AI-Assisted Semiconductor Engineering for Edge-AI Hardware

A YOLO26-based Feasibility Study with Expert Review and Implementation Feedback

Members of technical staff, AiNekko

TECHNICAL WHITE PAPER | JULY 2026

Executive Summary

This paper reports a feasibility study of an architecture-driven, AI-assisted semiconductor workflow applied to a fixed YOLO26-based edge-inference workload. Experienced engineers set architectural constraints and acceptance criteria; agents and deterministic scripts produced candidate specifications, compiler transformations, RTL, verification collateral, and implementation artifacts; executable references and EDA results determined whether those artifacts advanced.

Starting from a local quantized ONNX model and weights, the work produced 102 convolution wrappers, 102 content-addressed weight ROMs, integration RTL, and a generated structural core. Project status records report clean generated-core lint, all 102 convolution layers compared with ONNX Runtime, a later 82/82 generated-layer regression subset, selected integration tests, and unbounded equivalence for three leaf IPs; the final raw regression archive is incomplete. A later project dashboard records a 24-image, layer-by-layer RTL chain with 92 of 96 reference detections matched, 17 of 24 images fully matched, and no image classified as a collapse under the project criterion after recalibration. The per-image outputs and adopted scale artifact are not retained, so this result is not independently reproducible. Physical evidence includes 11.661 mm2 of flattened ASAP7 synthesis for the convolution layers and selected post-fix routed blocks near or above 1 GHz. The analytic throughput model reports 11,154 frames/s at an assumed 1 GHz design point.

The study demonstrates broad model-to-RTL transformation and shows specific cases in which workload validation, provenance review, synthesis, and routing feedback corrected generated artifacts. It does not establish full-core functional or physical closure, GDS, silicon, or measured productivity and cost improvement. The result is therefore evidence for an expert-governed model-to-hardware convergence loop and a defined next experiment toward weights-to-GDS, not autonomous chip design.

Focus Areas

AI-assisted semiconductor engineering; edge-AI hardware; RTL generation; functional verification; physical implementation feedback; reproducibility; YOLO26-based workload.

1. Introduction

Artificial intelligence is widely expected to become an integral component of future semiconductor development processes. Agents are rapidly demonstrating the ability to generate engineering documentation, software, hardware descriptions, verification environments, and implementation code that have traditionally required significant manual effort. Consequently, the fundamental question facing the semiconductor industry is no longer whether AI will be incorporated in chip development, but rather how engineering organizations can restructure their development processes to effectively leverage AI-assisted engineering while maintaining or improving the quality, rigor, and predictability required for commercial semiconductor products [17]-[20].

For more than four decades, semiconductor development has evolved into a highly structured engineering discipline supported by sophisticated Electronic Design Automation (EDA) tools, specialized engineering teams, and well-defined development methodologies. While this approach has enabled extraordinary advances in semiconductor design complexity and capability, it has also resulted in increasingly long development schedules, larger multidisciplinary teams, and substantial investments in proprietary software infrastructure. As AI application workloads continue to diversify and increase in system complexity, these factors have become significant challenges for development cost and time-to-market. The opportunity presented by AI-assisted engineering therefore extends well beyond automating individual engineering tasks. It creates the possibility of redesigning the semiconductor engineering process itself to improve engineering productivity, reduce schedule risk, optimize team effectiveness, better meet customer expectations, and lower the overall cost of semiconductor development. Moreover, an improved flow expands to include a "model-to-chip" development process that optimizes the product design starting with one or more underlying model architectures.

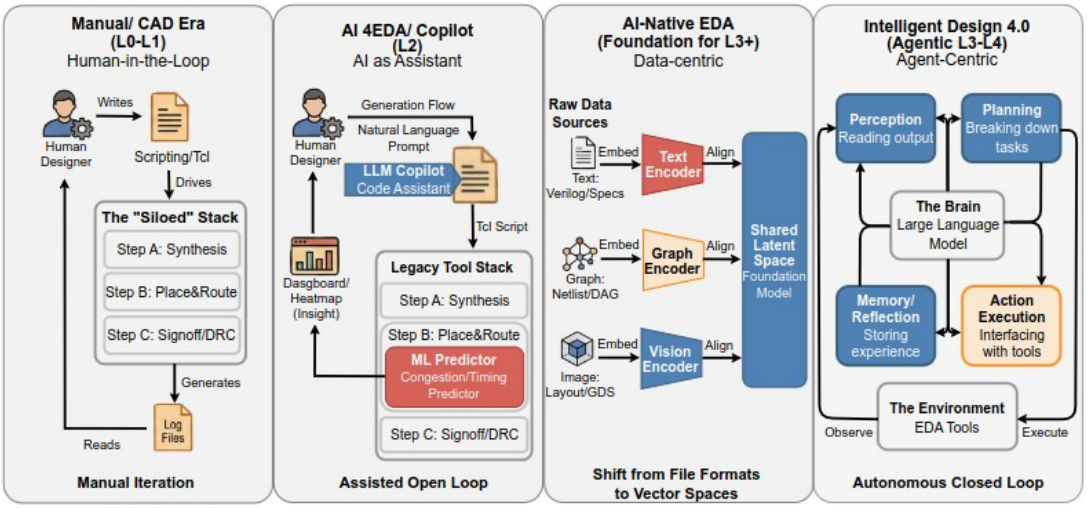

Figure 1. Evolution of the semiconductor design flow, adapted from [17].

Recent publications have highlighted the potential for agent-assisted design methodologies [17]-[20]. These works share several themes with this project, including the use of specialized agents, iterative tool feedback, and human supervision. They also report limitations that indicate the current maturity of agents for design automation. Earlier work on natural-language-to-RTL, LLM-assisted hardware generation, and high-level synthesis provides additional context for the present study [2]-[8].

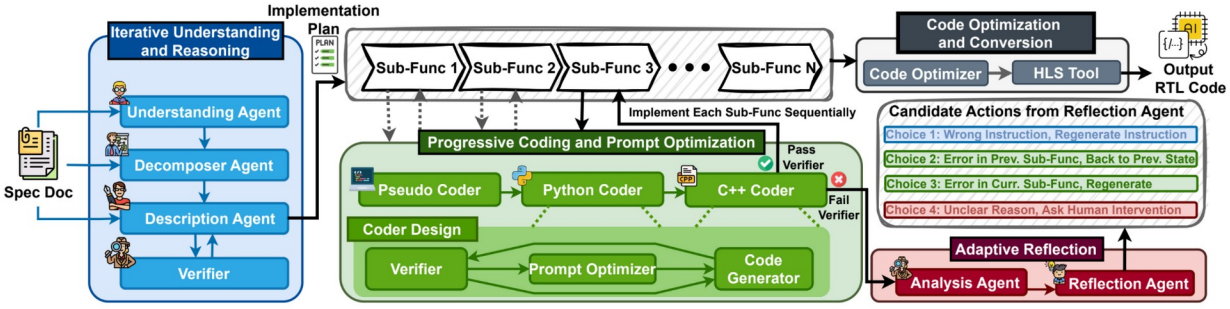

Figure 2. Spec2RTL-Agent process proposed by NVIDIA and collaborators [20].

This project was initially framed as a focused two-week engineering feasibility study and ultimately extended closer to one month. Rather than assessing the capabilities of individual AI models or attempting to demonstrate fully autonomous chip design, the project investigated whether a semiconductor development methodology centered on interactive collaboration between specialized AI agents and experienced semiconductor architects could substantially improve the efficiency of engineering execution. The objective was to determine whether AI-generated engineering content, continuous architectural guidance, and rapid iterative refinement could accelerate semiconductor development while preserving the engineering quality and outcomes required for commercial implementation.

An equally important objective was to explore how the engineering process itself must evolve when AI agents become active engineering contributors. Conventional semiconductor development assumes that engineers are responsible for creating nearly all of the content, either through direct code design or through use of commercial tools. In contrast, the methodology evaluated in this study positions AI agents as collaborative engineering participants that generate candidate specifications, architectural alternatives, RTL implementations, software components, verification environments, documentation, and implementation packages. Domain experts remain central to the process, providing technical leadership, defining engineering objectives, evaluating tradeoffs, integrating the outputs of multiple AI agents, resolving inconsistencies, and determining when engineering content has achieved sufficient maturity to advance. The resulting methodology is therefore not an automated design flow, but an interactive engineering process in which AI accelerates execution while experienced engineers retain responsibility for system architecture, engineering quality, and final technical decisions.

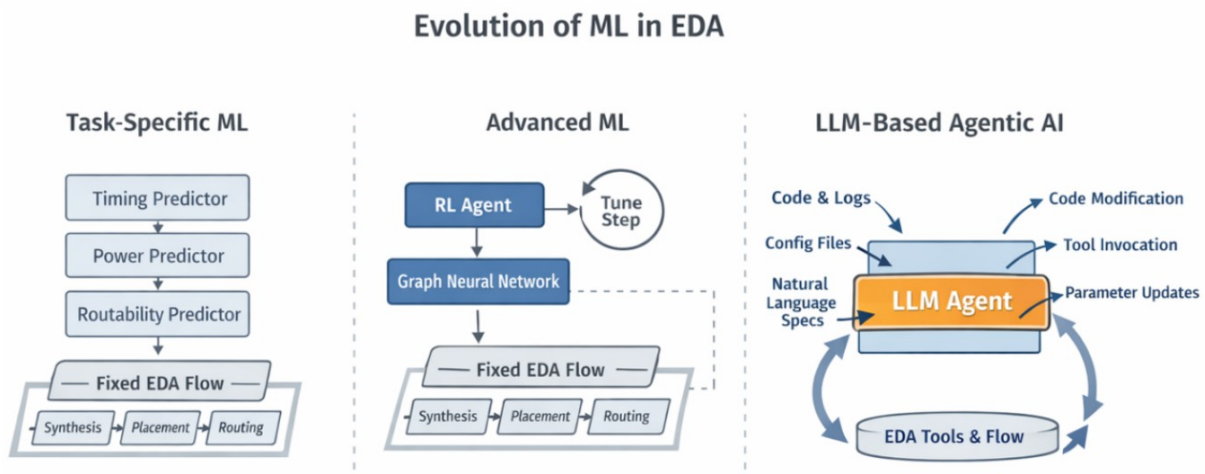

Figure 3. Evolution of machine learning in the chip-design flow, adapted from [19].

The project intentionally focused on a hypothetical chip design using a modified design process that relied heavily on agents to generate design elements. A representative AI workload based on YOLO26 was selected as the starting point because it provides computational, memory-bandwidth, and data-movement characteristics associated with modern edge-AI systems. Beginning with a fixed application workload allowed the team to concentrate on evaluating AI-assisted specification development, architectural refinement, hardware/software partitioning, RTL generation, verification, and place-and-route. The work also used external workload and algorithm context from COCO, QOI, and FlashAttention-related material [14]-[16].

In a commercial semiconductor program, however, architecture development begins considerably earlier. The starting point is a comprehensive set of customer requirements that define the intended application, deployment environment, performance objectives, power and thermal constraints, manufacturing cost targets, software ecosystem, safety and security requirements, and long-term product roadmap. These constraints become the foundation for Design Space Exploration (DSE), during which architects evaluate alternative processor organizations, accelerator architectures, memory hierarchies, interconnect topologies, software execution models, and implementation technologies before detailed hardware design begins. Although this upstream process was intentionally omitted in the present study, it represents a necessary extension of the methodology and will be a focus of subsequent process restructuring, including similar agent-assisted approaches.

A second objective of the project was to investigate whether this engineering methodology could be implemented within an open-source ecosystem. The project team considered open-source software, open hardware intellectual property, RISC-V processor architectures, and community-supported development tools to reduce dependence on expensive proprietary toolchains during the early phases of development [9]-[13]. The objective was not necessarily to replace commercial EDA environments, which continue to provide critical capabilities for advanced implementation and production signoff, but to identify where open technologies could provide technically effective and economically attractive alternatives during architecture development, specification generation, hardware/software co-design, verification, and implementation.

The main conclusion of this study is that the greatest opportunity presented by AI is not simply faster generation of design content, but the creation of a fundamentally different development process. By combining specialized AI agents, continuous expert guidance, concurrent hardware/software development, continuous verification, and a strategically selected open engineering ecosystem, the entire development process can be restructured. Success would be measured by improvements in engineering schedule, reductions in engineering effort, optimization of team size and expertise, reduced dependence on proprietary tool infrastructure, and improved overall project cost. These outcomes were not quantitatively measured in the present feasibility study.

2. Research Questions and Evaluation Method

2.1 Research Questions

The study evaluates a workflow, not an individual language model. Four research questions connect the architecture-driven thesis to evidence that can be inspected in the repository.

Table 2-1. Research Questions and Acceptance Evidence

Research Question	Evidence Examined	Acceptance Boundary
RQ1: Can a fixed model artifact be transformed into broad structural RTL under architect direction?	Generator source, ONNX-derived topology and dimensions, wrapper and ROM manifests, integration RTL, and generated-core lint	Establishes structural coverage and elaboration, not complete function
RQ2: What functional evidence supports the generated hardware?	Leaf and integration DV, ONNX Runtime layer comparisons, formal equivalence, lint, and image-oriented chained checks	Each result applies only to the tested block, layer, property, or workload path
RQ3: Did implementation feedback produce physically meaningful evidence?	Analytic cycle model, flattened ASAP7 synthesis, selected OpenROAD routes, timing fixes, memory disposition, and reported commercial observations	Establishes feedback on represented logic and selected blocks, not full-core PPA or signoff
RQ4: Which controls caused the workflow to reject or revise plausible artifacts?	Target revision, calibration failure, invented-IP correction, memory restructuring, and resource-limited runs	Supports causal case observations, not defect rates or productivity deltas

2.2 Evidence Hierarchy

The evaluation distinguishes five evidence levels. Artifact presence establishes that a file or report exists. Structural evidence establishes parse, elaboration, shape, or connectivity properties. Functional evidence establishes agreement for defined inputs, properties, or references. Physical evidence establishes synthesis or routed behavior under stated tools and constraints. Signoff evidence would establish full-design closure and release readiness. A result at one level is not promoted to the next without the required experiment.

This hierarchy is important because generated hardware can pass syntax and local tests while remaining numerically wrong, physically poor, or incomplete at system level. Conversely, an unsuccessful full-scale run does not erase valid leaf evidence; it limits the scope to which that evidence may be generalized.

2.3 Case-Study Records

The evaluated records are the current repository artifacts and the project accounts from the model-to-RTL and physical-implementation workstreams. Quantitative claims are reported only where a named source records a value: generated manifests and RTL for structure, `TASKS.md` and module documentation for DV, `formal/README.md` for proof scope, `TIMING.md` for the cycle model, and the `pnr/` reports for ASAP7 synthesis and routing. `ARTIFACT_MANIFEST.md` retains file identities and known release gaps.

The study was initially framed as two weeks and extended closer to one month, but no complete person-hour, agent-run, prompt, review, or rework ledger exists. No parallel conventional implementation was run. The method can therefore answer the four technical questions above at bounded evidence levels; it cannot test schedule, cost, staffing, or productivity advantage.

3. Conventional Baseline and Comparison Limit

Commercial semiconductor development converts customer requirements into architecture, hardware and software specifications, RTL, verification evidence, physical implementation, signoff, manufacturing data, and post-silicon validation. Specialized teams and milestone reviews provide the rigor required to manage this chain. The architecture-driven workflow studied here is intended to preserve those authorities while changing how intermediate artifacts are produced and revised.

Industry schedule, talent, and infrastructure concerns motivate the study [1], but no parallel conventional team or matched project was run. The baseline is therefore methodological rather than quantitative. Any future comparison must normalize team composition, active person-hours, review burden, AI and EDA cost, rework, verification maturity, PPA, and equivalent endpoint maturity.

4. Architecture-Driven Workflow and Implementation Method

4.1 Control Loop

Architects defined workload intent, cycle targets, numerical formats, interface boundaries, reuse expectations, and acceptance criteria. Agents generated candidate code and documentation under those constraints. Deterministic scripts extracted model structure and emitted repeatable artifacts. Functional references and EDA tools then supplied evidence that either accepted an artifact at a bounded level or generated a new constraint for the next iteration. Engineers retained authority to revise, replace, or reject every artifact.

The loop was informal rather than a production orchestration system: prompts, run boundaries, model versions, and human edits are not completely logged. Its observable units are therefore committed artifacts, named tests and reports, and recorded corrections.

4.2 Model and Weight Transformation

The local input is `integ/yolo26n/model_int8.onnx`. The project lacks the exact upstream repository, export command, calibration corpus, and adopted calibration-scale file, so the study begins at this checked-in artifact rather than claiming reproducibility from an original training checkpoint.

`tools/gen_core.py` loads the ONNX graph and runs ONNX shape inference. A balanced scale report supplies the canonical order of 102 convolution nodes, while `scale_pkg_dv.sv` supplies each layer's channel counts, kernel, stride, padding, feature-map dimensions, and selected input/output parallelism. The generator resolves producer tensors back to convolution indices and emits `yolo26n_layers_pkg.sv` with per-layer dimensions and producer indices.

The structural resolver converts graph operations into hardware connectivity. Convolution outputs become typed activation buses. ONNX Slice operations become channel subranges, Concats become bus joins, and residual Adds instantiate banks of `add_rq` units rather than disappearing as wiring. Named graph clusters are replaced with six separately validated integration blocks: SPPF, two upsample/concat paths, two attention paths, and the detect head. Four `skip_buf` banks retain verified skip tensors. The generated `yolo26n_core_impl.sv` therefore captures convolution-to-convolution edges, graph glue, integration blocks, and detection outputs in one structural body.

Weight generation is a separate deterministic path. `tools/gen_all_weight_roms.py` reads convolution weights and parameters from the ONNX graph, expands depthwise kernels to the dense organization used by the hardware, packs lanes in the `conv_stage` consumption order, and derives per-output scale and bias values from calibrated activation scales. It emits content-addressed `rom_<md5>` modules and a layer-to-ROM manifest. The current checkout contains 102 wrappers, 102 ROM modules, and 102 manifest entries.

This transformation has an explicit structural limitation. Two attention positional-encoding convolutions lie inside reshaped attention regions whose ONNX dimension 1 is not the hardware channel axis. The generator feeds those points width-correct V stand-ins to preserve connectivity and floorplanning structure. Consequently, generated-core lint establishes elaboration and width consistency, not exact attention dataflow or whole-model function.

4.3 Hardware Organization and Scaling

Each generated convolution wrapper binds a `conv_stage_core` instance to its weight ROM. The stage combines a line buffer, tiled convolution engine, quantization and activation behavior, and a sequencer over input- and output-channel tiles. Parallelism is selected through `P_CIN` and `P_COUT`. The scale optimizer minimizes their product subject to a per-stage cycle constraint rather than maximizing parallelism without an area bound.

The architecture uses frame-level dataflow. Dedicated stages may process successive frames concurrently, with per-stage double buffering assumed by the throughput model. The internal target is `T_FRAME = 100,000` cycles per frame, equivalent to 10,000 frames/s at a 1 GHz design point. This is a study target introduced during refinement, not a customer requirement or measured clock for the full core.

Memory structures remain abstracted. Generated RTL ROMs identify content and consumers, and behavioral SRAM wrappers expose banking and replacement boundaries. They are not compiled SRAM or ROM macros and therefore do not establish final area, timing, power, initialization, or DFT behavior.

4.4 Functional Evaluation Procedure

Validation proceeds from narrow to broad evidence.

1. Leaf and composite IP testbenches compare synthesizable SystemVerilog against behavioral or independent software references under directed and randomized inputs.
2. Generated convolution layers are compared with ONNX Runtime subgraph outputs using defined numerical tolerances and cosine similarity.
3. Integration tests exercise SPPF, upsample, attention, detect, and related blocks against local references or extracted ONNX samples.
4. Formal sequential equivalence compares selected synthesizable leaf IP with recoded golden models for all inputs and cycles under stated reset assumptions.
5. Image-oriented evaluation carries reconstructed quantized activations through the graph and compares final detections with ONNX Runtime.

The fifth level needs careful interpretation. `tools/e2e/chain.py` interprets the ONNX graph node by node and supports either a NumPy model of the convolution arithmetic or sequential calls to the per-layer `conv_stage` Verilator backend. An early bus-image run used the NumPy path for most convolutions and spot-checked layers 0 through 4 in RTL. The later `sweep_images.sh` path explicitly selects `--conv rtl --rtl-blocks all --scales fixed`, invoking every convolution stage and all six integration-block types through their RTL backends for each image. Graph traversal and data exchange remain a software-orchestrated layer-by-layer chain; this is not simulation of one monolithic generated core.

4.5 Physical Evaluation Procedure

Physical feedback used multiple evidence levels. Yosys supplied generic synthesis checks and flattened ASAP7 logic synthesis. `pnr/route_ips.sh` converts selected SystemVerilog with sv2v, synthesizes with Yosys, and routes with OpenROAD against ASAP7 RVT/TT collateral at a 1000 ps target. Post-fix reroutes use timing repair and record WNS, inferred Fmax, and area for selected blocks. Large `conv_stage_core` and `flash_attn` configurations exhausted local conversion or routing resources and were not routed whole.

The cycle model computes pipeline initiation interval as the maximum stage cycle count and reports frames/s as clock frequency divided by that interval. Latency is the sum along the represented dependency path. These are architecture calculations; only selected block frequencies have routed support. Project accounts also describe commercial full-chip implementation attempts, but raw reports, versions, constraints, and comparable quantitative results are not available for publication.

4.6 Decision Ownership

Agents contributed candidate RTL, test scaffolding, scripts, implementation changes, and documentation. Architects and domain engineers supplied constraints, identified invalid provenance, interpreted functional and physical reports, and decided which outputs remained in the workstream. Because review and rework effort were not logged, the study evaluates technical convergence mechanisms rather than labor displacement or productivity.

5. Evaluation Criteria and Comparison Limits

A structural transformation claim requires a generated artifact, a traceable relation to the model graph, and successful elaboration or lint. A functional claim requires a named reference, inputs, comparison metric, and result at the block or path tested. A physical claim requires a specified PDK or library context, constraints, and synthesis or route result. A production claim would additionally require integrated memories, full-core closure, power and reliability analysis, physical verification, DFT, software, release control, and manufacturing evidence.

The paper applies these criteria asymmetrically: positive results are kept narrow, while missing higher-level evidence remains visible. For example, clean generated-core lint supports structural composition but cannot establish workload correctness; routed leaf IP supports implementation feedback but cannot establish a full-core clock; and a correct analytic throughput calculation cannot establish measured system performance.

The conventional flow is a contextual baseline only. Commercial programs begin with customer requirements and DSE and proceed through controlled implementation, verification, signoff, manufacturing, and post-silicon validation. This study began with a fixed workload and stopped before signoff. It did not run a parallel conventional team or a matched historical implementation. Calendar duration and artifact count are therefore descriptive variables, not evidence of acceleration.

Negative examples were selected because they changed an engineering decision or exposed a failure mode relevant to the workflow. They are causal case records, not a random sample. No defect rate, acceptance rate, or statistical improvement is inferred from them. Similarly, commercial-tool observations are treated as team-reported qualitative evidence unless a publishable report and constraint set is available.

6. Results

6.1 RQ1: Model-to-RTL Structural Transformation

The transformation produced a broad hardware representation rather than isolated code samples. The current checkout contains 102 `conv_layerNN` wrappers and 102 content-addressed ROM modules; `rom_manifest.json` contains 102 corresponding entries. The wrapper range is contiguous from layer 0 through layer 101.

The generator obtained topology from the ONNX graph instead of reconstructing connectivity from prose. Project records identify 55 direct convolution-to-convolution edges. Shape inference supplied tensor widths used for Slice, Concat, and residual glue. Six

integration blocks and four skip-buffer banks were inserted at named graph locations, and detect-head outputs were wired to the generated core boundary.

Table 6-1. Structural Transformation Evidence

Artifact	Observed Result	Evidence Strength and Boundary
Convolution structure	102 wrappers bound to 102 ROM modules	Directly countable; does not establish per-layer correctness
Graph connectivity	Producer index table, 55 recorded convolution edges, structural Slice/Concat/Add glue	Generator and generated RTL are inspectable; two attention PE inputs use documented stand-ins
Integration structure	SPPF, two upsample, two attention, detect, and four skip-buffer banks	Instances are inspectable; functional evidence is separate
Weight binding	Layer, dimensions, tiling, weight-byte count, and content identity in ROM manifest	Strong identity evidence; original model and adopted calibration lineage are incomplete
Generated-core lint	0 warnings and 0 errors reported, approximately 340 s and 14 GB	Supports elaboration and connectivity; raw log and exact version are absent
Product top level	Generated implementation body exists behind a frozen boundary	Physical handoff notes state the AXI/FSM shell does not yet instantiate the datapath; not a complete SoC

These observations answer RQ1 positively at the structural level: the workflow transformed a fixed model artifact into broad, connected RTL. The result is not exact end-to-end function. In addition to the attention stand-ins, the external product shell and memory macros remain incomplete. Those limitations are part of the generated artifact, not merely future publication work.

6.1.1 Worked Trace: First Convolution

Layer 0 provides a concrete transformation trace. The source node is `/model.0/conv/Conv_quant`. Its generated layer record identifies a 3-channel input, 16-channel output, 3x3 kernel, stride 2, padding 1, and a 320x320 output map. The selected parallelism is `P_CIN=3` and `P_COUT=16`, yielding an estimated 51,200 cycles for the stage.

`conv_layer000.sv` carries those values directly into `conv_stage_core`: `CIN=3`, `COUT=16`, `K=3`, `STRIDE=2`, `PAD=1`, `H_IN=W_IN=640`, and the selected parallel factors. The wrapper binds the stage to `rom_ef517ee325e293ceaa13299945a16f4b`. The ROM manifest independently identifies that module as the consumer record for layer 0, with the same channel and kernel dimensions, one tile, and 432 weight bytes. This trace shows model identity, inferred shape, architecture scaling, generated interface, and weight binding agreeing for one ordinary Conv+SiLU stage. It demonstrates the transformation mechanism; validation of its numerical behavior remains a separate result.

6.2 RQ2: Functional Evidence

Leaf DV provided the widest set of explicit test counts. Recorded examples include 71/71 checks for `mac8`, 177/177 for `i32_to_fp16`, 1887/1887 and 1894/1894 for two `dotN` sizes, an exhaustive 65,536-input sweep for `fp16_to_i8_sat`, and directed tests for requantization, residual addition, softmax, and box decode. These results support the tested configurations but lack a consolidated archive of commands, versions, seeds, and raw logs.

Layer validation compared generated convolution behavior with ONNX Runtime subgraphs. `TASKS.md` records all 102 convolution layers as validated and a later 82/82 passing generated regression subset. The distinction matters: 102 is a project-level layer status, whereas 82 is the retained generated-directory count. The checked-in `BATCH_REPORT.md` covers an earlier L27-L101 run and still records generation, build, and cosine failures; no final raw report corresponding to the later 82/82 status is present. The paper therefore treats both counts as project status records rather than independently reproducible closure.

Integration records provide separate evidence for non-convolution behavior. SPPF is recorded at 10/10 checks with 5/5 ONNX Runtime samples. Each upsample path is recorded at 10/10 checks with 5/5 samples. Two attention shims are recorded at 10/10 checks and five ONNX Runtime samples each; the `flash_attn` leaf reports five tested configurations. The detect head reports 7/7 checks across six inputs.

Table 6-2. Functional Evidence by Level

Level	Result	Claim Supported	Principal Gap
Leaf/composite DV	Directed, randomized, and one exhaustive input sweep across named IP	Selected block behavior	Consolidated raw regression and coverage archive
Convolution layers	102 layer statuses and a later 82/82 project summary; retained batch report contains earlier failures	Layer-level convergence across iterations	Final raw report, inputs, per-layer metrics, thresholds, logs, source model

Level	Result	Claim Supported	Principal Gap
Integration	Recorded SPPF, upsample, attention, FlashAttention, and detect checks	Selected non-convolution blocks and interfaces	Exact seeds, raw outputs, retained disposition
Formal	Unbounded equivalence for three leaf IPs; `mac8` safety to K=131071	All-input/all-cycle equivalence under stated assumptions for those leaves	Proof logs, versions, assumptions, broader property coverage
Structural lint	Clean generated-core elaboration	Width and connectivity consistency	No functional behavior
Workload-facing path	Bus image: 91/92 stages at cosine >= 0.99 and five matched detections; later 24-image RTL-chain status: 92/96 detections, 17/24 fully matched, no collapse	Accumulated quantization and partial workload agreement	Missing per-image archive, adopted scales, exact run environment, and monolithic-core execution

Formal evidence is narrow but stronger than simulation within its scope. The recoded golden models were checked by bit-level SAT/induction or word-level SMT. `i32_to_fp16`, `fp16_to_i8_sat`, and `mac8` are recorded as unbounded sequentially equivalent. A separate `mac8` property carries an exact shadow sum and proves no accumulator wrap through 131071 multiply-accumulates; the bound fails at 131072, establishing tightness for that property model. Larger floating-point, protocol-heavy, real-valued-reference, and stateful blocks remain outside this proof level.

The initial image-oriented path tested accumulated quantization behavior. One 640x640 bus image produced 91 of 92 recorded convolution-stage cosine values at or above 0.99, with one depthwise class-logit tail at 0.973. The chain produced the same five objects as ONNX Runtime at IoU at least 0.95. Most convolution execution in that snapshot used the NumPy hardware-arithmetic model, while convolutions 0 through 4 were spot-checked through `conv_stage` RTL.

The validation process then falsified the apparent single-image success. Broader images exposed activation-scale sensitivity and a stale fixed-scale build cache. After percentile recalibration and cache correction, the dated project dashboard records a 24-image run using all convolution and integration RTL backends: 92 of 96 ONNX Runtime detections matched, 17 of 24 images matched all reference detections, and no image collapsed. The sweep driver and 24-image path list remain, but the generated gallery, per-image JSON and logs, and adopted calibrated-scale file do not. The result is therefore a quantified project status supported by an inspectable execution path, not an independently rerunnable regression package or a claim of exact workload equivalence.

6.3 RQ3: Performance and Physical Evidence

The throughput calculation models a frame-level pipeline. For stage cycle counts `C_i`, the initiation interval is `II = max_i(C_i)` and the modeled throughput is `f_clk / II`. The report uses validated block-DV counts where recorded and analytic frame-store estimates for small support stages. Under per-stage double buffering, the slowest represented stage is an attention block at 89,651 cycles/frame. At the assumed 1 GHz design point, this gives 11,154 frames/s and satisfies the internal 100,000-cycle target. The critical-path fill estimate is 5,939,396 cycles, or approximately 5.94 ms at the same assumed clock. Other project summaries retain 84,700-cycle/11,806-frames/s and 82,951-cycle/12,055-frames/s snapshots; this paper selects the current `TIMING.md` values and does not combine measurements across those revisions.

The 1 GHz value is not a full-core timing result. It is a design point informed by selected routed blocks. The architecture result is therefore the cycle count and its conditional conversion to throughput; a final implementation must establish the clock independently.

Flattened ASAP7 synthesis gives the strongest area evidence for the convolution logic. `pnr/LAYER_AREA.md` reports 11.661 mm2 for 102 convolution layers and approximately 11.8 mm2 for currently represented logic with selected support blocks. The report separately identifies approximately 4.7 MB of ROM and SRAM content that has not been compiled or included as final macros. Attention routing, full-core routing, utilization, and power are also absent.

Table 6-3. Physical and Performance Evidence

Quantity	Result	Interpretation Boundary
Stage initiation interval	89,651 cycles/frame	Architecture/DV-derived model
Conditional throughput	11,154 frames/s at 1 GHz	Assumed full-design clock and double buffering
Conditional fill latency	5.94 ms at 1 GHz	Represented dependency path, not measured silicon
Convolution logic area	11.661 mm2 flattened ASAP7 synthesis	Excludes final routing and memories
Selected post-fix routes	`dotN` 1.02 GHz, `requant` 1.05 GHz, `reduce_max_n` 1.17 GHz	Leaf or reduced blocks under recorded ASAP7 flow
Open physical items	Whole `conv_stage_core` and `flash_attn` route, macros, utilization, power, signoff	Prevents full-core PPA and GDS claims

The routed experiments did more than provide favorable examples. First-pass results exposed timing failures in arithmetic, reduction, requantization, and pooling structures. Pipeline changes, timing repair, and a split comparator tree changed the selected reroute set; for example, the committed `reduce_max_n` result is reported at 1.17 GHz after its two-stage split. Other measured blocks, including `box_decode` and FP16 units outside the reroute set, remain below 1 GHz in the first-pass summary. Per-block closure therefore does not compose into a full-core frequency claim.

RQ3 is answered at the feedback level: synthesis and routing produced concrete measurements and caused implementation changes. It is not answered at the final implementation level because memories, attention, interconnect, clocking, congestion, power, and signoff are not closed.

6.4 RQ4: Evidence That Changed Engineering Decisions

Target revision. An early 80K-FPS-style ambition was revised as area, cycle, and workload constraints became clearer. The retained target became 100,000 cycles/frame, approximately 10K frames/s at 1 GHz. This case shows an agent-assisted estimate being replaced by an explicit architecture constraint. It does not quantify the time saved or lost.

Provenance correction. An FP16 implementation was represented as extracted Erbium IP but was found to be newly generated and physically poor. The correction required engineers to inspect origin and implementation quality separately. This case supports machine-enforced source identity and license lineage as acceptance criteria; functional plausibility alone was insufficient.

Memory restructuring. RTL ROM inference and behavioral SRAM were useful for function and structure but produced misleading physical implications. The work introduced explicit macro-swap points and banked wrappers so bandwidth and replacement boundaries were visible. No compiled macro result exists, so this is evidence of problem identification and interface correction rather than memory closure.

6.4.1 Physical-Feedback Trace: Class-Score Reduction

The detect head ranks each anchor by the maximum of 80 signed-int8 class logits. The first implementation of `reduce_max_n` computed that maximum with a balanced seven-level comparator tree in one cycle. The function was simple and its leaf DV passed, but a single-cycle result did not satisfy the physical objective. The authoritative post-fix record identifies the pre-split netlist at -248 ps slack and 0.80 GHz under the 1000 ps ASAP7 RVT/TT route. An earlier first-pass table records a still worse version at -517 ps and 0.66 GHz; these are distinct implementation snapshots, not interchangeable measurements.

The corrective change divided the comparator tree at `CUT=floor(ceil(log2(N))/2)`, registered the surviving partial maxima, and completed the remaining comparisons in a second stage. This changed output latency from one to two cycles, so the independent reference and detect-head control path were delayed to preserve the protocol. The committed RTL exposes that boundary as `cut_q` and delays enable with `en_q`; it does not change the signed-max operation or its one-vector-per-cycle steady-state rate. The updated block retained its recorded 82/82 functional checks.

OpenROAD was then rerun with the same named 1 GHz ASAP7 flow. The resulting row reports +148 ps slack, 1.17 GHz derived Fmax, and 219 um2 area. Thus the evidence chain is inspectable: a workload requirement selected an 80-way reducer; DV established the operation; P&R rejected its single-cycle structure; the RTL and latency contract changed; DV was retained; and rerouting moved the same block above the target. This trace proves that physical evidence altered an implementation decision in this case. It does not prove that the detect head or full core closes at 1 GHz, because `box_decode` remains at 0.79 GHz in the retained first-pass record and the full core was never routed.

6.4.2 Validation-Feedback Trace: Calibration Scope

The first end-to-end validation path calibrated quantization from one image and then exercised that same narrow path. Layer agreement and image output initially appeared acceptable, but the result did not survive expansion to the broader test image set. The project narrative records the failure after several hours of multi-image execution and attributes it to activation scales derived from the single calibration image. The response was to expand calibration inputs to COCO and QOI images and regenerate scale-dependent weights and lookup data.

The retained tooling makes the intended correction concrete. `calibrate_full.sh` enumerates COCO val2017, approximately 5,000 in-distribution images, and the approximately 2,800-image QOI benchmark collection, invokes percentile calibration, writes `calib_scales_full.json`, and requires an explicit adoption and RTL-sweep step. The generated ROM identities include scale and bias content, so adopting new scales also requires ROM and wrapper regeneration rather than changing only a runtime parameter. This connection explains why calibration provenance belongs in the hardware release record.

The dashboard records the outcome of the correction on the retained 24-image set: 92 of 96 reference detections matched, 17 images fully matched, and no image classified as a collapse under its detection-agreement criterion. This is evidence that the calibration and cache changes improved generalization beyond the bus image, but it is not exact equivalence: four detections remained unmatched and seven images were not fully matched. The adopted scale file, generated gallery, metrics JSON, logs, tool versions, and exact

rerun environment are absent. Full-corpus calibration over the roughly 8,000 COCO-plus-QOI inputs named by `calibrate_full.sh` also remains an instructed future run rather than a retained result.

Resource-aware decomposition. Whole-core simulation and selected large routes exceeded local memory. The team moved to leaf DV, layer comparison, chained interpretation, formal checks, and reduced physical runs. This decomposition enabled progress but also created a risk of composing narrow passes into a broad claim; the evidence hierarchy in Section 2 was adopted to prevent that inference.

These cases answer RQ4: architect review, executable workload comparisons, provenance inspection, and EDA reports each rejected or changed plausible artifacts. The evidence is causal at the case level because a recorded failure led to a specific workflow or RTL response. Without dated effort logs and a complete issue history, it cannot establish frequency, efficiency, or economic benefit.

6.5 Answers to the Research Questions

RQ1 is supported for broad structural transformation: the checked-in model artifact was converted into a generated, linted core body with 102 convolution/ROM bindings and named integration structures. Exact whole-model function and product integration remain incomplete.

RQ2 is supported at multiple bounded levels: named leaf tests, layer comparisons, integration checks, three unbounded leaf proofs, structural lint, a single-image stage trace, and a recorded 24-image block-chained RTL result. Exact whole-model equivalence, monolithic-core RTL execution, and independently reproducible multi-image closure are not demonstrated.

RQ3 is supported as implementation feedback: cycle, partial area, and global routing results exist and caused changes. Full-core PPA, detailed routing, memory integration, signoff, and GDS are not demonstrated.

RQ4 is supported by five decision-changing cases. They establish necessary controls for this workflow, while productivity and economics remain untested.

7. Discussion and Lessons Learned

The case study supports six operating lessons.

Architecture must remain explicit. Agents made local progress without reliably preserving system-wide intent. Cycle targets, numerical formats, interfaces, memory assumptions, reuse rules, and acceptance criteria must remain visible across iterations.

Generation and acceptance are different functions. Agents were useful for candidate RTL, tests, scripts, and documentation, but experts remained responsible for integration and readiness. Review effort is part of the workflow cost and must be measured rather than treated as free.

Functional evidence must follow the workload. Syntax, leaf tests, and isolated tensor agreement can all pass while workload behavior remains wrong. Executable model references, broader input sets, intermediate comparisons, and predeclared thresholds are necessary to avoid narrow-validation false confidence.

Physical evidence must arrive early. Syntactically valid RTL can be unsuitable silicon structure. Reduced synthesis and routing experiments exposed timing, fanout, hierarchy, ROM, memory, and area problems early enough to revise the design. Full-core signoff remains necessary; reduced experiments are feedback, not substitutes.

Provenance is a correctness property. The invented-IP substitution showed that functional plausibility does not establish origin or licensing. A production workflow needs machine-enforced source attribution, generated-versus-reused boundaries, human-edit records, and release decisions.

Operational scope must be controlled. Agents can spend substantial time on oversized simulations, repeated regressions, or reopened decisions. Task boundaries, resource budgets, authoritative artifacts, and termination criteria are engineering controls, not merely prompt-writing practices.

Together, these lessons suggest that the scalable asset is not a collection of generated files. It is a governed system that retains architectural intent, artifact lineage, acceptance criteria, and tool evidence. The project demonstrates elements of that system, but not yet its repeatability across teams or designs.

8. Threats to Validity

Single case and expert dependence. Evidence comes from one compressed YOLO-oriented project directed by experienced semiconductor engineers. Results may not generalize to other workloads, nodes, mixed-signal or safety-critical designs, larger systems, or less experienced teams.

No experimental baseline. No parallel conventional team or matched historical project produced an equivalent deliverable. Calendar duration, artifact breadth, and observed feedback loops cannot establish productivity, schedule, cost, or quality improvement.

Incomplete measurement and verification. Human effort, review time, agent runs, model and tool cost, retained artifacts, rework, coverage, reproducible image results, full-core PPA, and workload correctness are incomplete. The 24-image dashboard result lacks its raw package, and reported throughput is analytic at an assumed 1 GHz design point supported by selected routed blocks, not full-core signoff.

Artifact and environment reproducibility. Exact upstream model provenance, calibration and validation data, agent identities and prompts, original tool versions, host environment, raw logs, and some rerun inputs are missing. Repository checksums improve identity but do not replace a reproducible release.

Provenance and licensing. Generated, reused, derived, dataset, and commercial-tool artifacts are not yet covered by a complete file-level provenance and redistribution audit. The repository-level license does not settle every third-party or generated-file obligation.

Physical-evidence boundary. OpenROAD/ASAP7 routed leaf or reduced-block results, flattened logic synthesis, analytic models, and qualitative commercial-tool observations represent different evidence levels. The study has no full-core route, integrated memory macros, final power or utilization, DRC/LVS-equivalent closure, GDS, or silicon.

Disclosure limits. Commercial EDA reports and some agent context may be restricted. Where raw evidence cannot be published, those observations remain team-reported and should not be treated as independently inspectable measurements.

These threats bound the result to an artifact-informed feasibility study. They do not negate the observed convergence loop, but they define the evidence required before broader technical or economic claims can be made.

9. Conclusions

9.1 What the Case Study Established

This study evaluated whether experienced semiconductor architects could use AI agents, executable workload references, and implementation tools as a connected engineering workflow rather than as isolated aids. Starting from a fixed YOLO26-based model artifact and weights, the project produced a Python ONNX-to-SystemVerilog path, a structural implementation with 102 convolution wrappers and corresponding weight ROMs, integration RTL, verification collateral, selected formal proofs, generated-core structural lint, a recorded 24-image block-chained RTL evaluation, analytic performance and area models, and routed experiments for selected physical blocks.

Those artifacts support a bounded but meaningful conclusion. Under expert direction, the workstream progressed from model intent into substantial hardware structure and repeated functional and physical feedback. Generated artifacts became useful when architects constrained the solution space, executable references exposed behavioral divergence, and EDA reports forced changes to timing, hierarchy, memory treatment, and implementation structure. The project therefore provides evidence for architecture-driven hardware-workflow convergence, not merely for language-model code generation.

The case study did not produce a customer-derived architecture, complete SoC, production software stack, full-core functional closure, integrated memory-macro implementation, full-core route, signoff database, GDS handoff, or manufactured silicon. It also did not measure a productivity or cost advantage against a conventional baseline. These are not semantic qualifications: they define the difference between the present feasibility result and a commercially deployable design flow.

9.2 Strategic Implication

The central implication is that architecture can serve as the control layer for AI-assisted semiconductor development. Architects supply the durable content that generated artifacts alone do not provide: system boundaries, interfaces, numerical intent, performance constraints, implementation assumptions, provenance requirements, and acceptance criteria. Agents can then expand that intent into candidate specifications, compiler transformations, RTL, tests, implementation scripts, and documentation, while executable references and EDA tools provide increasingly authoritative evidence.

This structure suggests a path to leverage that is more durable than faster first-draft RTL. If architectural decisions and acceptance gates can be represented explicitly, retained across iterations, and linked to generated artifacts and tool results, then engineering knowledge can become part of a repeatable workflow rather than remaining only in manual handoffs and individual review. Whether that structure improves schedule, cost, or team utilization remains to be measured, but it is a concrete technical hypothesis with auditable outputs.

The engineering direction suggested by this case is therefore an architecture-driven weights-to-GDS workflow. In such a workflow, a versioned model and weights are transformed under architect-approved constraints through compiler representation, RTL, verification, physical implementation, signoff, and release-controlled GDS. The significance is not that signoff disappears. It is that functional and physical signoff evidence becomes part of the same machine-assisted convergence loop that produced the earlier artifacts.

The current project reached the early and middle portions of that chain and identified the controls required to continue it. Completing the chain is the next logical system-level test of the methodology. It would convert a promising implementation workstream into a falsifiable demonstration with a precise technical endpoint.

9.3 Next Demonstration and Diligence Gates

A credible follow-on study should predeclare the claim and evidence standard before implementation begins. At minimum, it should require:

1. a versioned source model, weights, dataset provenance, preprocessing, quantization, and acceptance metrics;
2. an architect-approved system contract linking customer or workload requirements to interfaces, numerical behavior, PPA constraints, memories, clocks, and implementation assumptions;
3. a deterministic artifact manifest distinguishing generated, edited, reused, replaced, and retained files, together with model, agent, tool, and human-intervention records;
4. a complete integrated top level and reproducible functional-verification package, including workload accuracy, coverage, formal scope, protocol checks, and CDC/RDC status;
5. compiled memory macros and a full-core physical implementation with floorplan, utilization, congestion, clocking, multi-corner timing, power, signal-integrity, and power-integrity evidence;
6. a release-controlled GDS or OASIS artifact with checksum, DRC/LVS-equivalent results, waiver and ECO status, tool and PDK versions, and an archival build record; and
7. normalized human effort, agent execution, review, rework, infrastructure cost, and a conventional or matched historical baseline at equivalent deliverable maturity.

These gates would answer the two questions the present case study cannot: whether the technical loop closes at full-system scale, and whether doing so creates measurable engineering or economic advantage. Repeating the experiment on a second workload and obtaining independent reproduction would then test whether the capability generalizes beyond the founding team and the YOLO26-oriented design.

9.4 Closing Perspective

The evidence in this paper is sufficient to justify the next experiment, not to claim its outcome in advance. It shows that a model-and-weights starting point can be expanded into a broad hardware workstream when experienced architects retain control and functional and physical evidence continuously constrain generated artifacts. It also shows where that process fails when provenance, validation scope, or implementation structure is weak.

The next step for this line of semiconductor engineering is to carry the same discipline through the stages that remain open: complete integration, verification closure, memory implementation, full-core physical closure, signoff, and GDS release. A successful demonstration would not be autonomous chip design. It would be something more operationally useful: a reproducible architecture-driven path from weights to GDS in which expert intent remains visible, tool evidence remains authoritative, and every major transformation can be inspected and challenged.

10. Appendix: Artifact and Reproducibility Handoff

The paper keeps only the evidence needed to evaluate its claims. Detailed repository paths, checksums, command summaries, provenance status, and release gaps are maintained in `ARTIFACT_MANIFEST.md`, with implementation-specific rerun notes in the referenced module, formal, timing, area, and physical-design documentation. This separation keeps the paper compact without treating the supplementary records as stronger evidence than they contain.

Table 10-1. Supplementary Evidence Status

Area	Available Supplement	Principal Missing Item
Workload and data	Local quantized ONNX artifact, preprocessing and calibration scripts, 24-image sweep set, and dashboard result summary	Upstream model/export lineage, calibration-corpus manifest, adopted scale file, licenses, and raw per-image result archive
Generated hardware	Compiler and generator sources, 102 wrappers, 102 ROMs, structural core, integration RTL, and partial checksum manifest	Complete generated/edited/reused/replaced/retained lineage and exact generation record
Verification	Leaf and integration DV sources, layer summaries, selected formal sources, lint instructions, and image-oriented scripts	Raw logs, exact versions, seeds, stimuli, thresholds, coverage rationale, and reproducible consolidated results
Physical implementation	ASAP7 synthesis and selected routed-block summaries, scripts, timing and area reports, and commercial-flow scaffolding	Full-core route, memory macros, power/utilization, signoff, raw tool records, and GDS/OASIS
AI and human workflow	Qualitative project accounts and representative negative cases	Model/agent identities, prompts or task records, run boundaries, human edits, review effort, rework, and execution cost
Provenance and release	Repository license, partial SPDX inventory, and category-level provenance notes	File-level third-party licenses, dataset rights, redistribution decisions, and commercial-report disclosure status
Environment	Python version request and locked Python dependencies	Original host/container, EDA/formal tool versions, PDK revision, resource limits, and complete build environment

The current artifact package supports repository inspection but not independent reproduction of every reported result. A publication release should freeze the source commit, populate the missing fields above, archive raw or license-permitted evidence, and identify restricted artifacts explicitly. A future comparative study should additionally record active person-hours, role allocation, agent runs, review and rework time, infrastructure cost, and an equivalent conventional baseline.

References

[1] Semiconductor Industry Association and Boston Consulting Group, *The Growing Challenge of Semiconductor Design Leadership*, 2022. Available: https://www.semiconductors.org/wp-content/uploads/2022/11/2022_The-Growing-Challenge-of-Semiconductor-Design-Leadership_FINAL.pdf. Accessed: July 8, 2026.

[2] H. Pearce, B. Tan, and R. Karri, "DAVE: Deriving Automatically Verilog from English," arXiv:2009.01026, 2020. Available: https://arxiv.org/abs/2009.01026. Accessed: July 8, 2026.

[3] S. Thakur, B. Ahmad, H. Pearce, B. Tan, B. Dolan-Gavitt, R. Karri, and S. Garg, "VeriGen: A Large Language Model for Verilog Code Generation," arXiv:2308.00708, 2023. Available: https://arxiv.org/abs/2308.00708. Accessed: July 8, 2026.

[4] M. Liu, N. Pinckney, B. Khailany, and H. Ren, "VerilogEval: Evaluating Large Language Models for Verilog Code Generation," arXiv:2309.07544, 2023. Available: https://arxiv.org/abs/2309.07544. Accessed: July 8, 2026.

[5] M. Liu et al., "ChipNeMo: Domain-Adapted LLMs for Chip Design," arXiv:2311.00176, 2023. Available: https://arxiv.org/abs/2311.00176. Accessed: July 8, 2026.

[6] S. Liu, W. Fang, Y. Lu, J. Wang, Q. Zhang, H. Zhang, and Z. Xie, "RTLCoder: Fully Open-Source and Efficient LLM-Assisted RTL Code Generation Technique," arXiv:2312.08617, 2023. Available: https://arxiv.org/abs/2312.08617. Accessed: July 8, 2026.

[7] A. Canis et al., "LegUp: High-Level Synthesis for FPGA-Based Processor/Accelerator Systems," FPGA '11, 2011. doi:10.1145/1950413.1950423.

[8] H. Ye, C. Hao, J. Cheng, H. Jeong, J. Huang, S. Neuendorffer, and D. Chen, "ScaleHLS: A New Scalable High-Level Synthesis Framework on Multi-Level Intermediate Representation," DAC '22, 2022. doi:10.1145/3489517.3530631.

[9] C. Wolf, J. Glaser, and F. Schupfer, "Yosys: A Free Verilog Synthesis Suite," Austrochip, 2013. Available: https://yosyshq.net/yosys/files/yosys-austrochip2013.pdf. Accessed: July 8, 2026.

[10] T. Ajayi et al., "OpenROAD: Toward a Self-Driving, Open-Source Digital Layout Implementation Tool Chain," GOMACTech, 2019. Available: https://theopenroadproject.org/publications/. Accessed: July 8, 2026.

[11] YosysHQ, "SymbiYosys Documentation," Read the Docs. Available: https://symbiyosys.readthedocs.io/. Accessed: July 8, 2026.

[12] RISC-V International, "Ratified Specifications." Available: https://riscv.org/specifications/ratified/. Accessed: July 8, 2026.

[13] L. T. Clark, V. Vashishtha, L. Shifren, A. Gujja, S. Sinha, B. Cline, C. Ramamurthy, and G. Yeric, "ASAP7: A 7-nm FinFET Predictive Process Design Kit," *Microelectronics Journal*, vol. 53, pp. 105-115, 2016. doi:10.1016/j.mejo.2016.04.006.

[14] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," arXiv:1405.0312, 2014. Available: https://arxiv.org/abs/1405.0312. Accessed: July 8, 2026.

[15] D. Szablewski, "The Quite OK Image Format: Specification Version 1.0," 2022. Available: https://qoiformat.org/qoi-specification.pdf. Accessed: July 8, 2026.

[16] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Re, "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness," arXiv:2205.14135, 2022. Available: https://arxiv.org/abs/2205.14135. Accessed: July 8, 2026.

[17] Z. Zang, Y. Song, B. W.-K. Ling, A. Wang, and F. Yang, "The Dawn of Agentic EDA: A Survey of Autonomous Digital Chip Design," arXiv:2512.23189, 2025. Available: https://arxiv.org/abs/2512.23189. Accessed: July 13, 2026.

[18] D. N. Gadde, K. K. Radhakrishna, V. N. Viswambharan, A. Kumar, D. Lettnin, W. Kunz, and S. Simon, "Hey AI, Generate Me a Hardware Code! Agentic AI-Based Hardware Design and Verification," arXiv:2507.02660, 2025. Available: https://arxiv.org/abs/2507.02660. Accessed: July 13, 2026.

[19] A. Ghose, A. B. Kahng, S. Kundu, and B. Pramanik, "Invited: Agentic AI for Physical Design R&D: Status and Prospects," in *Proc. ACM Int. Symp. Physical Design (ISPD)*, 2026. doi:10.1145/3764386.3779612.

[20] Z. Yu, M. Liu, M. Zimmer, Y. C. Lin, Y. Liu, and H. Ren, "Spec2RTL-Agent: Automated Hardware Code Generation from Complex Specifications Using LLM Agent Systems," in *Proc. IEEE Int. Conf. LLM-Aided Design*, 2025. Available: https://research.nvidia.com/publication/2025-06_spec2rtl-agent-automated-hardware-code-generation-complex-specifications-using. Accessed: July 13, 2026.