# TRUSTME: A BENCHMARK FOR LLM-AIDED HARDWARE DESIGN CODE AUDITING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The advancement of Large Language Models (LLMs) has accelerated the adoption of vibe coding in chip design, catalyzing the emergence of LLM-Aided Design (LAD) as an agile development paradigm now widely deployed in practical chip development workflows. This work focuses on a critical task in the chip supply chain: design code security auditing, with a primary emphasis on hardware Trojan detection. Identifying security threats at the early stages of the chip lifecycle is of paramount importance. Existing LAD research has predominantly centered on code generation, verification script synthesis, and debug completion, with limited attention paid to code security alignment and hardware design security awareness. This gap may lead LLMs to generate or overlook hazardous design code, thereby contaminating downstream segments of the supply chain. To bridge this gap, we present TrustMe, a comprehensive benchmark for hardware code Trojan security auditing. TrustMe comprises 816 tasks sampled from real-world projects, with over 1560 complete and stealthy Trojans implanted. It spans evaluation scenarios from module-level to System-on-Chip (SoC) level, assessing full-stack hardware code security monitoring capabilities, including knowledge-based question answering, vulnerability localization, and attack type analysis. The benchmark supports simulation testing across diverse vibe coding environments, including both agentic and pure-LLM setups. We evaluated 10 mainstream LLMs and 6 agentic tools on TrustMe. Results show that even the best-performing combination achieves a hardware Trojan identification success rate below 50%, underscoring the urgent need to prioritize security considerations in LAD. Furthermore, based on our evaluation findings, we propose a Skill-Powered Trojan detection and identification framework. Operating in a self-iterative improvement mode, this framework achieves an average 12.6% performance boost without modifying existing models or agentic tools. We hope this work will encourage LLM-aided chip design from productive into genuinely trustworthy.

## 1 INTRODUCTION

The rapid advancement of Large Language Models (LLMs) (Naveed et al., 2025) is reshaping various domains, including chip design. Pioneering platforms such as NVIDIA ChipNeMo (Liu et al., 2023a) and Synopsys.ai Copilot (Synopsys, Inc., 2026) have demonstrated that LLMs can generate Register-Transfer-Level (RTL) code at scales ranging from individual modules to complete System-on-Chip (SoC) architectures. This progress has catalyzed an emerging AI-native design paradigm known as LLM-Aided Design (LAD) (Huang et al., 2024), accelerating the integration of LLMs into real-world chip-design supply chains and contributing to an industry-wide transition toward increasingly automated design workflows (OpenAI, 2026; Zhu et al., 2026; ChipAgents, Inc., 2026).

However, introducing AI components into established hardware-design supply chains also creates critical security concerns (Ning et al., 2023). Because hardware designs are progressively refined across multiple levels of abstraction, vulnerabilities introduced during early development stages can propagate into downstream implementation and manufacturing. Such vulnerabilities may remain undetected until post-silicon validation, thereby increasing both defect latency and remediation costs (Chakraborty et al., 2009; Hu et al., 2020; Bhunia et al., 2014). Within the LAD paradigm, assessing an LLM's security awareness during code generation is therefore particularly important, especially with respect to hardware Trojans, which represent one of the most challenging threats in



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

silicon security (Yasaei et al., 2022). As LLMs become increasingly common tools for RTL generation and verification, two complementary security capabilities are required. First, LLMs should consistently generate benign and secure RTL code, thereby reducing the number of verification iterations. Second, they should reliably identify malicious design constructs and provide actionable explanations, thereby mitigating downstream security risks (Meng et al., 2021; Witharana et al., 2023; Mankali et al., 2025). Together, these capabilities directly affect the convergence efficiency, reliability, and security of modern chip-design workflows.

Despite the rapid development of LAD, existing benchmarks focus predominantly on design productivity and performance acceleration. Representative efforts evaluate RTL generation (Lu et al., 2024; Garcia-Gasulla et al., 2025; Pinckney et al., 2025; Alberti-Binimelis et al., 2026) and testbench synthesis (Qiu et al., 2024; Wan et al., 2026a;b), while largely overlooking systematic security evaluation. Although prior studies have shown that LLMs can be used to generate hardware Trojans, substantially less attention has been devoted to their ability to audit hardware designs and detect malicious logic. Moreover, initial investigations of LLM-based SoC security verification have not yet provided rigorous evaluations on production-grade hardware codebases (Faruque et al., 2026).

To address this gap, we introduce TrustMe, the first agentic, LLM-driven benchmark designed specifically to evaluate code auditing and hardware Trojan detection in realistic design contexts. TrustMe systematically assesses an LLM's end-to-end security comprehension across multiple dimensions: domain knowledge, Trojan identification, attribution, threat localization, Common Weakness Enumeration (CWE) (The MITRE Corporation, 2026) classification, characterization, and remediation analysis. The benchmark comprises 816 evaluation tasks, including 335 graduate-level conceptual questions and 481 real-world hardware repositories spanning module- and SoC-level designs. Within these repositories, we embedded 1560 functionally complete, triggerable RTL Trojans covering major threat taxonomies and severity levels, complete with multimodal context support. To evaluate these tasks, we implemented an automated scoring system compatible with various LLM backends and agent architectures. Using semantic embeddings alongside ground-truth references, the scoring engine provides fine-grained failure attribution and capability profiling.

We evaluate TrustMe using six leading LLMs, including GPT-5.6, and five widely used agentic frameworks, including Claude Code and CodeX. Across 15 experimental configurations and more than 8,000 compute hours, our experiments establish a comprehensive baseline for LLM-based security auditing in hardware design. The best-performing configuration, a Bash Agent powered by DeepSeek V4 Pro and CodeX, completed 37.5% of the evaluation tasks. Although this result demonstrates substantial potential, it also reveals a significant gap between current systems and fully autonomous, trustworthy hardware security auditing. Motivated by these findings, we further develop FindHT, a lightweight hardware-Trojan detection framework that combines feedback-driven iterative refinement with skill-augmented reasoning. FindHT improves detection accuracy by up to 19.83% over the evaluated baseline systems. Both TrustMe and FindHT are publicly available at `https://anonymous.4open.science/r/ICLR-submit-TrustMe`.

## 2 Background

### 2.1 Hardware Security

Chips form the foundation of contemporary complex cyber-physical systems, and their security is paramount (Ray et al., 2017; Charles & Mishra, 2021; Russell & McGravey, 2025). However, as chip complexity and design cycles have increased, the semiconductor supply chain has become highly fragmented. While this fragmentation reduces development costs, it significantly broadens the attack surface (Raj et al., 2026). In response, hardware security research has evolved from focusing solely on side-channel defenses to encompassing the entire chip lifecycle: design, manufacturing, packaging, and operation (Sami, 2024; Kim & Rho, 2024). Hardware Trojans stand out as one of the most challenging threat vectors (Gubbi et al., 2023), as they can be naturally embedded in design and triggered under obfuscated conditions, and they are nearly impossible to reverse after fabrication. The research community has long sought to push Trojan detection earlier in the design flow: prior studies (Yasaei et al., 2022) and experience have shown that the later a risk is discovered, the higher the costs of repair and remediation become, rising nearly exponentially, as illustrated in Figure 1.

Extensive evidence (Fan et al., 2024; Mishra & Sahay, 2025; Gerlach et al., 2023) demonstrates that identifying and fixing security issues in the pre-silicon phase costs far less than addressing

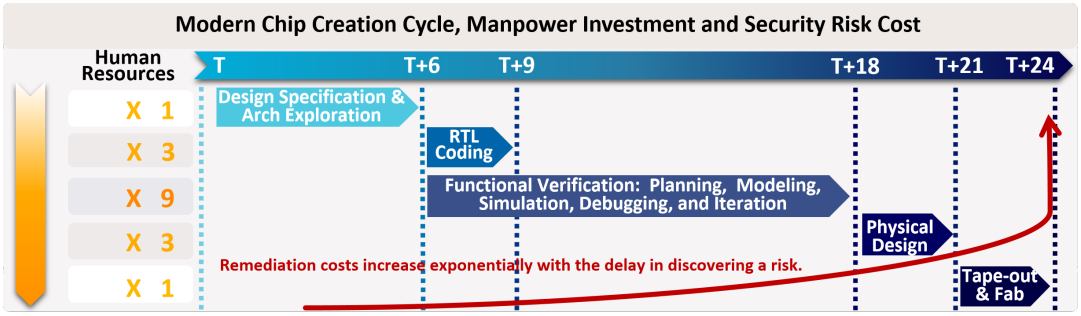



Figure 1: Contemporary Chip Design Flow, Engineering Effort, and Risk Cost

them post-fabrication. RTL design code, as the highest abstraction level in the design cycle and the stage closest to designer intent, directly influences the trustworthiness of all downstream synthesis, verification, and silicon. This motivates our focus on LLM auditing capabilities at the code design stage: in real chip creation, LLMs naturally assume dual roles as designer and verifier, which places higher demands on their security awareness.

### 2.2 LLM-Aided Design and Benchmarks

LLMs have revolutionized code understanding and generation, permeating every step of the hardware design process: RTL generation (Pei et al., 2024; Zhao et al., 2025; Akyash et al., 2025), debugging (Xu et al., 2024; Tsai et al., 2024; Zhang et al., 2026), verification (Hu et al., 2025; Wan et al., 2026a), and system-level design acceleration (Wang et al., 2024; Ghose et al., 2025). As model capabilities have advanced, benchmarks have become essential for assessing LLM performance. Works such as VerilogEval (Liu et al., 2023b) and LiveVerilogEval (Young et al., 2026) focus primarily on RTL generation, while FIXME (Wan et al., 2026b) began evaluating design verification. Yet these benchmarks uniformly prioritize functionality while overlooking security.

Some works have begun to address this gap, though with perspectives fundamentally different from ours. HarmChip (Wang et al., 2026) established a jailbreak benchmark for hardware security. HardSecBench (Chen et al., 2026) examines the safety of generated code. TrojanWhisper (Faruque et al., 2026) preliminarily assessed general-purpose LLMs' ability to detect and localize Trojans in basic RTL, but did not account for realistic deployment scenarios.

That leaves a critical question unanswered: when an LLM is positioned as an auditor facing a real design repository or dataset from untrusted sources, of unknown scale and potentially containing stealthy Trojans, can it perform end-to-end knowledge retrieval, Trojan identification, threat localization, CWE classification, or vulnerability analysis? This gap is what TrustMe aims to fill.

### 2.3 Challenges in LLM-Aided Code Auditing Evaluation

**C1.** Data scarcity. Existing public hardware Trojan corpora rely heavily on established benchmarks such as TrustHub (Krieg, 2023), which focus on digital circuits with specific functions, such as AES. Besides, these datasets are widely cited, and they risk contaminating model pretraining data and inflating evaluation results. Meanwhile, realistic vulnerability resources like Hack@DAC (Farahmandi et al., 2025) are sparse and cannot scale into large evaluation sets.

**C2.** Task complexity. Hardware code auditing is not a simple binary classification task; it spans a long pipeline encompassing security knowledge retrieval, static anomaly detection, bidirectional trigger-and-payload localization, Trojan intent identification, attack type and CWE classification, and exploitability and impact analysis (Ahmad et al., 2022; Shuai et al., 2025). Each phase demands different capabilities and poses distinct benchmarking requirements. Balancing complete capability coverage, difficulty gradation, and annotator verifiability remains the core challenge.

**C3.** Quality assurance constraints. Generative hardware benchmarks can manage quality straightforwardly through automated correctness validation via simulation, keeping annotation costs low and human error minimal. Security auditing benchmarks are fundamentally different. Valid Trojan

samples must satisfy multiple constraints (Saha et al., 2024). Any oversight at any stage corrupts the evaluation. This makes construction far more costly and labor-intensive than generative benchmarks.

These challenges form a compounding bottleneck: data scarcity restricts task design, task complexity inflates quality assurance costs, and quality defects undermine evaluation credibility. To address these issues, we construct TrustMe, a multi-granularity hardware auditing benchmark built on authentic data sources, end-to-end validation, and fine-grained attribution scoring. Furthermore, to address the model deficiencies identified during benchmark evaluation, we propose FindHT, a lightweight enhancement framework designed for hardware Trojan detection.

# 3 TrustMe Benchmark

## 3.1 Overview

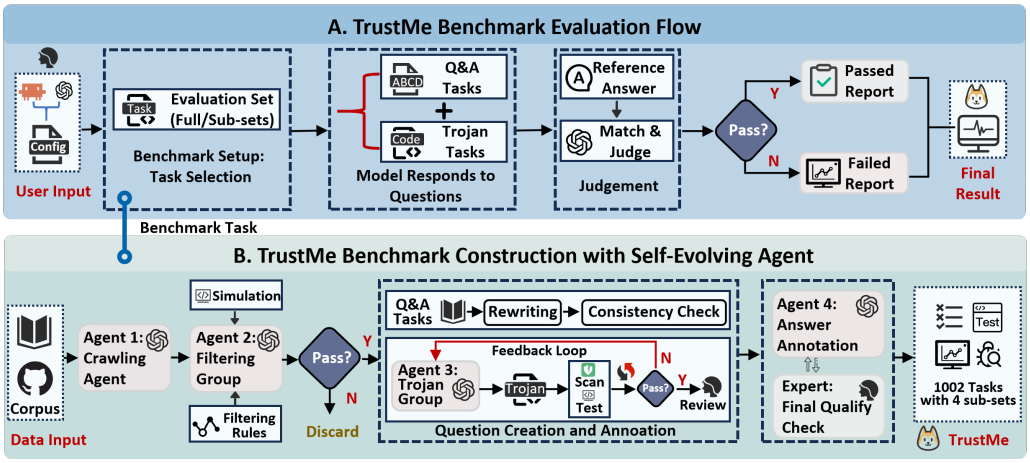



Figure 2: TrustMe: Workflow and Construction. Part A illustrates the general workflow of TrustMe, in which a configuration file specifies the LLM API, agentic tools, evaluation set, and other settings. Part B details the self-evolving, agent-human collaborative construction of TrustMe.

Addressing the need for comprehensive evaluation of LLMs' security awareness in hardware design code, we propose TrustMe, a large-scale benchmark comprising 816 evaluation tasks. Task instances are sourced from high-quality graduate-level textbooks, established hardware Trojan datasets, and 500 open-source designs spanning diverse design paradigms. On this foundation, we designed and embedded 1,560 novel, diverse, and stealthy RTL Trojans to meet the practical requirements of end-to-end evaluation. An overview of TrustMe is presented in Figure 2.

TrustMe further stratifies tasks into scalable subsets, named **HAMS**, organized around atomic capabilities, which serve as evaluation units that balance assessment cost and quality. The atomic capabilities we define are understanding, attribution, classification, characterization, detection, localization, and remediation. The composition of each subset is as follows.

**Hardware Security Understanding *(H)*.** This subset incorporates the foundational knowledge required for code auditing through 335 hardware security Q&A problems: 130 true/false questions, 12 computational problems, and 193 open-ended questions. All problems are derived from graduate-level curricula at leading institutions and are available under open-source licenses. They assess the model's understanding of core concepts and facts in hardware security, forming the epistemic foundation for subsequent auditing capabilities.

**Atomic Capability Verification *(A)*.** This subset contains 246 problems. Each problem presents a complete module-level RTL design with a single, structurally complete Trojan embedded. Each problem focuses on a single atomic capability, enabling fine-grained assessment of the model's abilities in Trojan activation detection, behavioral detection, vulnerability detection, attribution analysis, CWE identification, defect localization, Trojan characterization, and repair recommendation.

**Module-Level Auditing *(M)*.** This subset comprises 225 problems, each corresponding to a complete module-level RTL design with 1 to 20 embedded Trojans that couple multiple atomic capabilities. For instance, tasks may require LLMs to simultaneously identify Trojan locations, classify weakness types, and suggest repairs.

**System-Level Auditing *(S)*.** This subset consists of 10 large-scale SoC designs sampled from tapeout-verified real projects, such as OpenTitan and OpenCores, with 20 to 200 Trojans embedded in each design. The objective is to assess the model's ability to locate security-critical defects across modules and abstraction levels while using all atomic capabilities in large-scale design scenarios.

## 3.2 Construction of TrustMe

### 3.2.1 Design Philosophy

**End-to-End Atomic Capability Evaluation.** Hardware code security auditing is a multi-stage workflow encompassing knowledge retrieval, vulnerability discovery, precise localization, categorization, and exploitability analysis. Because these stages are tightly coupled, evaluation must be conducted end-to-end, with atomic capabilities serving as the fundamental, verifiable units of assessment. This approach balances overall performance tracking with fine-grained attribution, ensuring that all conclusions are based on rigorously verified results.

**Modular and Extensible Tiered Subsets.** To accommodate variations in design scale, modalities, and threat severity, tasks are organized into decoupled, hierarchical subsets. Each subset maintains independent data, task definitions, and scoring criteria. This modular architecture allows seamless, on-demand expansion without sacrificing global comparability, providing a clean interface for community co-development and systematically mitigating data scarcity.

**Self-Evolving Agent-Human Collaborative Paradigm.** TrustMe introduces a self-evolving, agent-driven workflow operating within a feedback-in-the-loop framework to construct novel hardware Trojan samples with human review. This paradigm enables precise control over sample scale, diversity, and covertness. Furthermore, the methodology provides a generalizable blueprint for other security evaluation domains characterized by data scarcity and high annotation overhead.

### 3.2.2 Data Collection and Filtering

**Data Collection.** TrustMe sources data from two categories: a hardware security knowledge corpus and open-source hardware designs. We systematically collect hardware security textbooks, course assignments, and open-source hardware repositories from publicly accessible sources, with the dedicated Crawling Agent coordinating collection and deduplication. This phase yielded more than 100,000 candidate questions and approximately 1 million RTL design modules.

**Data Filtering.** Distinct filtering criteria are applied to each data category, and both categories are processed automatically by a Filtering Agent Group followed by final checks from human experts.

For knowledge questions, we impose three constraints. First, answers must not be directly retrievable through simple searches, preventing the task from degenerating into a retrieval problem. Second, questions and answers are stored separately without being paired, reducing the risk of contamination at the source (Lai et al., 2023). Third, given the knowledge prerequisites, we retain only graduate-level and upper-level textbooks and course materials.

For hardware designs, we prioritize projects with complete specifications and simulation environments, requiring all designs to pass syntax checks and functional simulations. After this filtering round, the candidate pool was reduced to approximately 1,000 questions and 700 RTL designs. A subsequent fine-grained screening removed knowledge questions whose difficulty distributions deviated substantially from the target distribution or that were overly subjective. It also removed RTL modules with highly redundant design types, ultimately forming the base task sets of TrustMe.

**Question Annotation.** This step was completed jointly by an Annotation Agent Group and human experts. For knowledge questions, the agents semantically rewrote the question stems while preserving their original meaning, further mitigating the possibility of pretraining-data contamination.

The primary focus of the annotation effort was on RTL Trojan tasks. Approximately 10% of the tasks were derived from existing datasets, such as TrustHub, although all of them underwent variable

renaming and question-stem rewriting; the remaining 90% were original samples. To accomplish this, we constructed a self-evolving agentic workflow. The process first distills knowledge from software vulnerabilities, CWE entries, hardware Trojan literature, and hardware design principles collected from the web, transferring reasoning capabilities from software vulnerability detection and exploitation to the hardware context. The model is then provided with clean RTL and tasked with designing novel, stealthy, and triggerable RTL Trojans. Candidate Trojans undergo scanning by three mainstream hardware Trojan detection tools at different stages (Xiao et al., 2025; Thorat et al., 2025; Chou et al., 2025) and trigger validation using Synopsys VCS (Synopsys, Inc., 2025). Detection-tool verdicts and simulation-verification results are fed back as automated signals, driving iterative refinement of Trojan structures over multiple rounds. A candidate Trojan is considered to possess preliminary stealth if at least one detection tool fails to flag it while the Trojan remains stably triggerable. Such Trojans proceed to human quality review. Trojans that pass the review enter the formal task candidate pool, after which the agents classify them according to the subset-partition principles and atomic-capability definitions, ensuring that the task distribution aligns with the overall design constraints. The detailed algorithm is presented in Appendix C.

**Answer Annotation.** For each question type, reference answers were written and verified separately. For knowledge questions, human experts determined the unique correct answers or the necessary key points. Reference answers for RTL tasks were generated in conjunction with the Trojan implantation process, explicitly recording trigger conditions, payload behavior, location, and weakness classification, and were verified against simulation evidence.

In total, TrustMe comprises 335 knowledge Q&A items and 481 RTL tasks spanning 200,000 lines of audited code, with 1,560 embedded Trojans and 3,075 atomic assessment points.

### 3.3 I/O Specification and Metrics

The input and output definitions for each subset are as follows.

***(H)*:** The input consists of hardware security knowledge questions, and the model outputs answers partitioned into objective and subjective components. Objective questions have standard reference answers, whereas subjective questions have a set of keyword sequences that must be matched.

***(A)*:** The input is a complete RTL design containing a Trojan and a single atomic-capability evaluation point, such as determining the Trojan type. The model outputs the corresponding result. Reference answers are defined using unambiguous, mutually exclusive standard keywords to minimize scoring noise caused by subjective phrasing.

***(M) and (S)*:** The input is a complete module or SoC containing one or more Trojans, and the model outputs all relevant information, including Trojan location, type, and classification, etc. Reference answers consist of an unambiguous, mutually exclusive sequence of standard keywords.

Scoring is performed by an independent judge LLM that did not participate in the evaluation. Objective questions are scored by comparison with the reference answers, whereas the other question types are evaluated using a keyword-matching mechanism and assigned linear quantitative scores based on the number of matched keywords. Final results are reported as Pass@5.

## 4 Evaluation

We observe that real-world LAD engineering scenarios are seldom completed in a single dialogue. They instead unfold through an agentic LLM loop that retrieves code, invokes simulation tools, and iteratively corrects its outputs. To emulate this workflow, TrustMe goes beyond standard single-turn inference evaluation by introducing an agent harness, making it the first benchmark to support large-scale evaluation of agentic LLMs.

Specifically, we selected six mainstream models as inference cores: GPT-5.6-sol, Qwen3.8-Flash, DeepSeek V4 Flash, DeepSeek V4 Pro, GLM-5.3, and GLM-X. We further integrated five mainstream agent frameworks as execution backbones: Codex, Claude Code, OpenCode, Trae Agent, and DeepSeek Harness. All models were tested on Codex and Claude Code, and the performance of the agent fraeworks is reported based on GLM-5.3, yielding 15 model–agent combinations in total. All combinations were run under the same task set, prompt template, and hyperparameter settings to ensure comparability. Each configuration was run five times, with pass@5 taken as the result, for a total runtime of 8,125 hours. The judge model is Claude-5-Opus.

Table 1: Overview of the Evaluation Results. The best value in each column is highlighted in **bold**.

| | | Overall (%) | | | Resources | |
|---|---|---|---|---|---|---|
| Model | Base | Avg. score[a] | Full-score rate | Pass rate | Token (M)[b] | Time (s)[b] |
| deepseek-v4-flash | claude_code | 58.6 | 32.0 | 33.6 | 104.46 | **32.0** |
| deepseek-v4-pro | claude_code | 59.0 | 30.9 | 32.5 | 101.65 | 96.7 |
| glm-5.3 | claude_code | 59.9 | 32.9 | 34.6 | 49.27 | 146.7 |
| glm-x-preview-260804 | claude_code | 58.1 | 31.9 | 33.5 | 69.43 | 302.4 |
| gpt-5.6-sol | claude_code | 54.5 | 28.9 | 30.3 | **19.63** | 86.4 |
| qwen3.8-flash | claude_code | 59.9 | 33.2 | 34.8 | 62.09 | 105.1 |
| deepseek-v4-flash | codex | 61.4 | 34.8 | 36.3 | 178.35 | 60.1 |
| deepseek-v4-pro | codex | **62.3** | **36.0** | **37.5** | 225.79 | 149.6 |
| glm-5.3 | codex | 48.2 | 23.8 | 28.7 | 65.54 | 174.5 |
| glm-x-preview-260804 | codex | 51.0 | 29.8 | 31.4 | 67.81 | 148.3 |
| gpt-5.6-sol | codex | 57.0 | 29.0 | 28.7 | 73.05 | 88.5 |
| qwen3.8-flash | codex | 59.5 | 33.6 | 34.8 | 84.16 | 178.5 |
| glm-5.3 | deepseek_harness | 48.3 | 30.6 | 34.7 | 113.50 | 453.7 |
| glm-5.3 | opencode | 55.8 | 35.1 | 32.7 | 100.29 | 266.2 |
| glm-5.3 | trae_agent | 41.8 | 25.3 | 27.1 | 145.93 | 419.7 |

[a] Computed as the mean across all questions that received a non-zero score, with zero-score items excluded.
[b] Token unit M ($= 10^6$); average time is converted from ms/question to s/question.

### 4.1 Overview

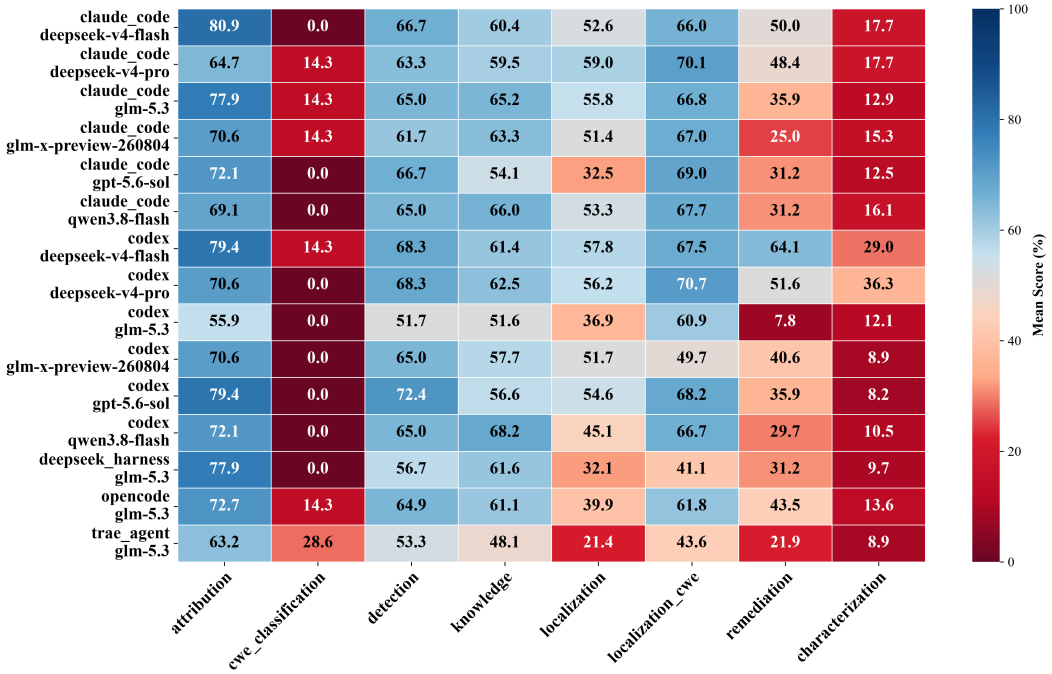



Figure 3: Evaluation Results of Pass Rate on Atomic Capability.

## 5 Result and Discussion

### 5.1 Result Analysis

Table 1 shows the final results. The overall pass rate is in the 5th column. After excluding zero-score and execution-failure cases, the average score and full-score rate of valid responses are reported in

the 3rd and 4th columns, respectively. The average scores for atomic capabilities are in Figure. 3. More raw experimental data are provided in Appendix A.

Among all 15 LLM–Agent configurations, **codex + deepseek-v4-pro** achieves the highest pass rate of 37.5%, along with the highest average score (62.3%) and full-score accuracy (36.0%) on valid responses. These results indicate that this configuration consistently leads in both complete task correctness and average response quality. However, this performance comes at a substantial resource cost. Its Agent Token consumption reaches 225.79M, approximately 11.5× that of the most resource-efficient configuration, **claude_code + gpt-5.6-sol** (19.63M). Its runtime of 149.6s is also considerably longer than that of the fastest configuration, **claude_code + deepseek-v4-flash** (32.0s). Considering pass rate, token consumption, and runtime jointly, **claude_code + glm-5.3** offers the most favorable efficiency–performance trade-off. With the second-lowest token consumption (49.27M), it achieves a pass rate of 34.6%, just 2.9 percentage points below the best-performing configuration, while consuming only approximately 22% of its tokens. This configuration therefore lies on the Pareto frontier of efficiency and performance.

When the LLM is held fixed, the five Agent backends exhibit substantially different resource-management and consumption patterns. Taking glm-5.3 as an example, its pass rate varies from 27.1% to 34.7% when deployed with claude_code, codex, deepseek_harness, opencode, and trae_agent. Token consumption varies by nearly 3× (49.27M–145.93M), while average runtime also differs by approximately 3× (146.7s–453.7s). These results suggest that, even with the task and LLM held constant, the choice of Agent backend is a major determinant of resource efficiency. Across all six LLMs, Codex and Claude Code exhibit the most consistent performance. Claude Code maintains average scores within a relatively narrow range of 54.5%–59.9%, exhibiting the lowest variance, whereas Codex achieves a higher performance ceiling, with its average score reaching 62.3%. In contrast, backends such as deepseek_harness and trae_agent not only achieve lower average scores (as low as 48.3% and 41.8%, respectively) but also incur the longest runtimes (453.7s and 419.7s), indicating substantially lower efficiency.

At the atomic-capability level, attribution, detection, and localization_cwe are tasks for which nearly all configurations demonstrate some degree of proficiency, with average scores across all models of 71.8%, 63.6%, and 62.5%, respectively. In contrast, cwe_classification and characterization (Trojan behavior characterization) remain broadly challenging, with average scores of only 7.6% and 15.3%, respectively. Notably, cwe_classification receives a score of zero in 9 of the 15 configurations, indicating that this capability is largely beyond the reach of the evaluated configurations.

### 5.2 Discussion

**Overall performance on hardware code auditing remains limited.** None of the evaluated configurations achieves a pass rate above 50%, with the highest reaching only 37.5%. This indicates that current LLM-based agents remain limited in their ability to detect and analyze hardware Trojans, with a substantial gap between current capabilities and practical deployment.

**Performance degrades as design complexity and analysis depth increase.** Knowledge-based QA tasks in book_qa achieve a substantially higher average score (approximately 59.8%) than IP-/module-level Trojan analysis tasks in htbench_core (approximately 44.0%). A similar trend emerges across atomic capabilities, with a clear decline in performance as analysis becomes more demanding: scores decrease from 71.8% for the relatively shallow attribution task to 46.7% for localization, 36.5% for remediation, and 15.3% for characterization.

**Efficiency–performance trade-offs are pronounced, with diminishing returns from additional resource consumption.** Compared with the most token-efficient configuration, claude_code + gpt-5.6-sol, the highest-performing configuration, codex + deepseek-v4-pro, consumes approximately 11.5× as many tokens and incurs a longer runtime, yet improves the pass rate by only 7.2 percentage points (from 30.3% to 37.5%). This suggests that the performance gains from adopting a more resource-intensive configuration may diminish substantially as resource consumption increases.

### 5.3 Proposed Mitigation: FindHT

Motivated by the capability gaps, we propose FindHT, a lightweight enhancement framework for the current LAD paradigm. FindHT introduces an external Controller, turning code auditing into a self-iterative reasoning process. The central design is a structured detection loop that proceeds through

source inventory, global lightweight scanning, focused code review, bidirectional hypothesis generation, structural checking, decision-making, adversarial challenge, and self-iterative conformance. It treats RTL Trojan auditing as a sequence of atomic capabilities—understanding, detection, attribution, classification, characterization, localization, and remediation—and routes different task families into either an EvidenceLoop or a TaskReasoningLoop. Detection, attribution, and localization tasks share hypothesis and decision representations, which are later projected into task-specific answer formats. Knowledge, mathematical, repair, and classification tasks follow requirement-driven reasoning without forcing Trojan assumptions onto unrelated questions. This separation preserves the generality of the underlying agent while imposing domain-specific constraints only where they matter. The structured loop further enforces source-version binding, issue reopening after hypothesis changes, scope sufficiency review, and independent adversarial challenge, thereby reducing the risk that an answer is formally valid but evidentially unsupported. The detailed algorithm of FindHT is provided in Appendix B.

The experimental results of FindHT are shown in Figure 4. Compared with the baseline, FindHT improves the average pass rate by 12.6%, with a maximum improvement of 19.83%.

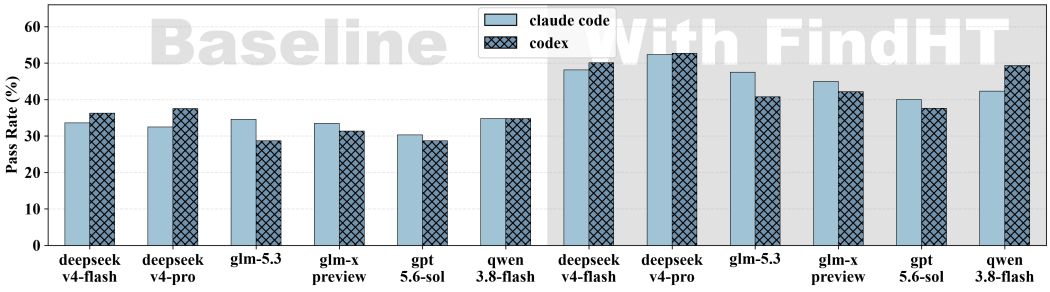



Figure 4: Evaluation Results of FindHT on 6 Models with 2 Agent Frameworks.

## 6 Conclusion

This work introduced TrustMe, the first large-scale benchmark for evaluating the security awareness of LLMs and agentic systems when auditing hardware design code. The benchmark combines 816 tasks, including 335 hardware-security knowledge questions and 481 RTL auditing tasks, covering approximately 200,000 lines of code, 1,560 embedded Trojans, and 3,075 atomic assessment points. It has 4 subsets: Hardware Security Understanding, Atomic Capability Verification, Module-Level Auditing, and System-Level Auditing, enabling evaluation from isolated skills to realistic, multi-Trojan SoC scenarios with fine-grained assessment.

Our evaluation of six LLMs across five agent frameworks and 15 model–agent configurations shows that current systems remain far from reliable autonomous hardware security auditors. The strongest configuration achieved a pass rate of only 37.5%. Performance declined as tasks required deeper semantic and causal reasoning: attribution and basic detection were comparatively accessible, whereas classification and characterization remained particularly difficult, with average scores of 7.6% and 15.3%, respectively. The results further expose a substantial efficiency–performance trade-off, indicating that additional tokens and runtime alone do not close the security gap.

Besides, to address these limitations, we proposed FindHT, a lightweight framework that introduces structured evidence gathering, hypothesis-driven analysis, adversarial challenge, and self-iterative refinement. Without modifying underlying LLMs or agent frameworks, FindHT improved the average pass rate by 12.6% and achieved a maximum gain of 19.83%. Beyond its empirical value, TrustMe offers a reusable methodology for constructing verifiable security benchmarks under data scarcity and annotation constraints. We hope this work could help transform LLM-aided design from productive into genuinely trustworthy.

### USE OF GENERATIVE AI

Generative AI tools were used in this work in two ways. First, the benchmark studied in this paper is constructed through a self-evolving agentic workflow that orchestrates large language models to generate, refine, and annotate RTL Trojan samples, and the evaluation itself targets large language models as auditing agents. Therefore, the methodological description necessarily concerns model behavior, and the experiments were run with the models and agent frameworks reported in Figure 2. Second, generative AI tools were used as writing assistance for language editing of the manuscript. They were not used to generate research ideas, to design the method, to analyze results, or to produce the claims made in this paper.

All AI-assisted output was reviewed by the authors. The framework-generated RTL is validated by the verification pipeline described in Appendix **??** rather than accepted on the model's assertion, so every reported instance has passed an executable check. We take responsibility for the final content of this work, including all text, claims, and artifacts.

### ETHICS STATEMENT

This work studies hardware Trojans from the defender's perspective, with the goal of stress-testing RTL security evaluation before deployment. All experiments are conducted exclusively on the public datasets and open-source designs, and no real product, vendor design, or deployed system is targeted or tested. Generated Trojan cases are research artifacts intended for analysis, and the framework operates on designs already available to any researcher. We consider the work to raise no ethical concerns beyond those inherent to security research.

## REFERENCES

Baleegh Ahmad, Wei-Kai Liu, Luca Collini, Hammond Pearce, Jason M Fung, Jonathan Valamehr, Mohammad Bidmeshki, Pawel Sapiecha, Steve Brown, Krishnendu Chakrabarty, Ramesh Karri, and Benjamin Tan. Don't CWEAT it: Toward cwe analysis techniques in early stages of hardware design. *arXiv preprint arXiv:2209.01291*, 2022. doi: 10.1145/3508352.3549506.

Mohammad Akyash, Kimia Azar, and Hadi Kamali. Rtl++: Graph-enhanced llm for rtl code generation. In *2025 IEEE International Conference on LLM-Aided Design (ICLAD)*, pp. 44–50. IEEE, 2025.

Miquel Alberti-Binimelis, Cristian Gutierrez-Gomez, Dario Garcia-Gasulla, Emanuele Parisi, Razine Moundir Ghorab, Orlando Montenegro, Bernat Homs, Miquel Moreto, and Gokcen Kestor. Revisiting turtle: A comprehensive evaluation of llms for rtl generation. *ACM Transactions on Design Automation of Electronic Systems*, 2026.

Swarup Bhunia, Michael S Hsiao, Mainak Banga, and Seetharam Narasimhan. Hardware trojan attacks: Threat analysis and countermeasures. *Proceedings of the IEEE*, 102(8):1229–1247, 2014.

Rajat Subhra Chakraborty, Seetharam Narasimhan, and Swarup Bhunia. Hardware trojan: Threats and emerging solutions. In *2009 IEEE International high level design validation and test workshop*, pp. 166–171. IEEE, 2009.

Subodha Charles and Prabhat Mishra. A survey of network-on-chip security attacks and countermeasures. *ACM Computing Surveys (CSUR)*, 54(5):1–36, 2021.

Qirui Chen, Jingxian Shuai, Shuangwu Chen, Shenghao Ye, Zijian Wen, Xufei Su, Jie Jin, Jiangming Li, Jun Chen, Xiaobin Tan, et al. Hardsecbench: Benchmarking the security awareness of llms for hardware code generation. *arXiv preprint arXiv:2601.13864*, 2026.

ChipAgents, Inc. ChipAgents: The agentic AI chip design environment. https://chipagents.ai/, 2026. Agentic AI platform for semiconductor design and verification; founded by William Wang (UC Santa Barbara). Accessed: 2026-09-26.

Chung Han Chou, Chih Jen Hsu, Hung Chun Chiu, Chung-Kuan Cheng, Shao-Yun Fang, Yi-Yu Liu, and Zhuo Li. Invited paper: 2025 ICCAD CAD contest problem a: Hardware trojan detection on gate-level netlist. In *2025 IEEE/ACM International Conference on Computer-Aided Design*, pp. 1–8. IEEE, 2025. doi: 10.1109/ICCAD66269.2025.11240658.

Ruchao Fan, Yongming Tang, Hao Sun, Jiyuan Liu, and He Li. An efficient ml-based hardware trojan localization framework for rtl security analysis. In *Proceedings of the 2024 ACM/IEEE International Symposium on Machine Learning for CAD*, pp. 1–7, 2024.

Farimah Farahmandi et al. Lessons from the hardware hacking competitions: Verification techniques, findings, and insights. In *arXiv preprint arXiv:2501.xxxxx*, 2025.

Md Omar Faruque, Peter Jamieson, Ahmad Patooghy, and Abdel-Hameed A Badawy. Trojanwhisper: Evaluating pre-trained llms to detect and localize hardware trojans. *IEEE Access*, 2026.

Dario Garcia-Gasulla, Gokcen Kestor, Emanuele Parisi, Miquel Albertí-Binimelis, Cristian Gutierrez, Razine Moundir Ghorab, Orlando Montenegro, Bernat Homs, and Miquel Moreto. Turtle: A unified evaluation of llms for rtl generation. In *2025 ACM/IEEE 7th Symposium on Machine Learning for CAD (MLCAD)*, pp. 1–12. IEEE, 2025.

Lukas Gerlach, Daniel Weber, Ruiyi Zhang, and Michael Schwarz. A security risc: microarchitectural attacks on hardware risc-v cpus. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 2321–2338. IEEE, 2023.

Amur Ghose, Andrew B Kahng, Sayak Kundu, and Zhiang Wang. Orfs-agent: Tool-using agents for chip design optimization. In *2025 ACM/IEEE 7th Symposium on Machine Learning for CAD (MLCAD)*, pp. 1–13. IEEE, 2025.

Kevin Immanuel Gubbi, Banafsheh Saber Latibari, Anirudh Srikanth, Tyler Sheaves, Sayed Arash Beheshti-Shirazi, Sai Manoj PD, Satareh Rafatirad, Avesta Sasan, Houman Homayoun, and Soheil Salehi. Hardware trojan detection using machine learning: A tutorial. *ACM Transactions on Embedded Computing Systems*, 22(3):1–26, 2023.

Wei Hu, Chip-Hong Chang, Anirban Sengupta, Swarup Bhunia, Ryan Kastner, and Hai Li. An overview of hardware security and trust: Threats, countermeasures, and design tools. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 40(6):1010–1038, 2020.

Yuchen Hu, Junhao Ye, Ke Xu, Jialin Sun, Shiyue Zhang, Xinyao Jiao, Dingrong Pan, Jie Zhou, Ning Wang, Weiwei Shan, et al. Uvllm: An automated universal rtl verification framework using llms. In *2025 62nd ACM/IEEE Design Automation Conference (DAC)*, pp. 1–7. IEEE, 2025.

Yingbing Huang, Lily Jiaxin Wan, Hanchen Ye, Manvi Jha, Jinghua Wang, Yuhong Li, Xiaofan Zhang, and Deming Chen. New solutions on llm acceleration, optimization, and application. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*, pp. 1–4, 2024.

Yongshin Kim and Sungho Rho. The us–china chip war, economy–security nexus, and asia: Kim and rho. *Journal of Chinese Political Science*, 29(3):433–460, 2024.

Christian Krieg. Reflections on trusting trusthub. In *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pp. 1–9. IEEE, 2023.

Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Wen-tau Yih, Daniel Fried, Sida Wang, and Tao Yu. Ds-1000: A natural and reliable benchmark for data science code generation. In *International Conference on Machine Learning*, pp. 18319–18345. PMLR, 2023.

Mingjie Liu, Teodor-Dumitru Ene, Robert Kirby, Chris Cheng, Nathaniel Pinckney, Rongjian Liang, Jonah Alben, Himyanshu Anand, Sanmitra Banerjee, Ismet Bayraktaroglu, et al. Chipnemo: Domain-adapted llms for chip design. *arXiv preprint arXiv:2311.00176*, 2023a.

Mingjie Liu, Nathaniel Pinckney, Brucek Khailany, and Haoxing Ren. Verilogeval: Evaluating large language models for verilog code generation. In *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pp. 1–8. IEEE, 2023b.

Yao Lu, Shang Liu, Qijun Zhang, and Zhiyao Xie. Rtllm: An open-source benchmark for design rtl generation with large language model. In *2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 722–727. IEEE, 2024.

Lakshmi Likhitha Mankali, Jitendra Bhandari, Manaar Alam, Ramesh Karri, Michail Maniatakos, Ozgur Sinanoglu, and Johann Knechtel. Rtl-breaker: Assessing the security of llms against backdoor attacks on hdl code generation. In *2025 Design, Automation & Test in Europe Conference (DATE)*, pp. 1–7. IEEE, 2025.

Xingyu Meng, Shamik Kundu, Arun K Kanuparthi, and Kanad Basu. Rtl-contest: Concolic testing on rtl for detecting security vulnerabilities. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(3):466–477, 2021.

Jyotiprakash Mishra and Sanjay K Sahay. Modern hardware security: A review of attacks and countermeasures. *arXiv preprint arXiv:2501.04394*, 2025.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72, 2025.

August Ning, Georgios Tziantzioulis, and David Wentzlaff. Supply chain aware computer architecture. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pp. 1–15, 2023.

OpenAI. Jalapeño's first results show industry-leading speed and efficiency in AI inference. https://openai.com/index/jalapeno-first-results/, August 2026. Accessed: 2026-09-22.

Zehua Pei, Huiling Zhen, Mingxuan Yuan, Yu Huang, and Bei Yu. BetterV: Controlled verilog generation with discriminative guidance. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40145–40153. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/pei24e.html.

Nathaniel Pinckney, Christopher Batten, Mingjie Liu, Haoxing Ren, and Brucek Khailany. Revisiting verilogeval: A year of improvements in large-language models for hardware code generation. *ACM Transactions on Design Automation of Electronic Systems*, 30(6):1–20, 2025.

Ruidi Qiu, Grace Li Zhang, Rolf Drechsler, Ulf Schlichtmann, and Bing Li. Autobench: Automatic testbench generation and evaluation using llms for hdl design. In *Proceedings of the 2024 ACM/IEEE International Symposium on Machine Learning for CAD*, pp. 1–10, 2024.

Kshitij Raj, Atri Chatterjee, Patanjali Slpsk, Swarup Bhunia, and Sandip Ray. Sentry: Protecting system-on-chip designs against supply-chain attacks. *ACM Transactions on Embedded Computing Systems*, 25(3):1–25, 2026.

Sandip Ray, Eric Peeters, Mark M Tehranipoor, and Swarup Bhunia. System-on-chip platform security assurance: Architecture and validation. *Proceedings of the IEEE*, 106(1):21–37, 2017.

Alison Lawlor Russell and Kevin McGravey. Digital oil: chips, artificial intelligence and us national security. *International Affairs*, 101(3):1087–1101, 2025.

Debjit Saha, Suryadipta Saha, et al. Hardware trojan detection using pre-trained language models. In *2024 IEEE International Symposium on Hardware Oriented Security and Trust*. IEEE, 2024.

Md Sami Ul Islam Sami. *End-to-End Secure IC Lifecycle Management*. University of Florida, 2024.

Jingxian Shuai et al. Unleashing GHOST: An llm-powered framework for automated hardware trojan design. In *2025 IEEE International Symposium on Hardware Oriented Security and Trust*. IEEE, 2025.

Synopsys, Inc. VCS: Functional verification solution. https://www.synopsys.com/verification/simulation/vcs.html, 2025. Accessed: 2026-09-26.

Synopsys, Inc. Generative AI for chip design. https://www.synopsys.com/ai/generative-ai.html, 2026. Accessed: 2026-09-22.

The MITRE Corporation. Common weakness enumeration (cwe). https://cwe.mitre.org/, 2026. Accessed: 2026-09-26.

Kiran Thorat, Amit Hasan, Caiwen Ding, and Zhijie Shi. Trojan-guard: Hardware trojans detection using gnn in rtl designs. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2025.

YunDa Tsai, Mingjie Liu, and Haoxing Ren. Rtlfixer: Automatically fixing rtl syntax errors with large language model. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*, pp. 1–6, 2024.

Gwok-Waa Wan, Shengchu Su, Jingyi Zhang, Sam Zaak Wong, Mengnv Xing, Lei Ji, Zhe Jiang, Xi Wang, and Jun Yang. Chattest: Coverage-enhanced testbench generation for agile hardware verification with llms. In *2026 Design, Automation & Test in Europe Conference (DATE)*, pp. 1–7. IEEE, 2026a.

Gwok-Waa Wan, SamZaak Wong, Shengchu Su, Chenxu Niu, Ning Wang, Xinlai Wan, Qixiang Chen, Mengnv Xing, Jingyi Zhang, Jianmin Ye, et al. Fixme: Towards end-to-end benchmarking of llm-aided design verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1087–1095, 2026b.

Xi Wang, Gwok-Waa Wan, Sam-Zaak Wong, Layton Zhang, Tianyang Liu, Qi Tian, and Jianmin Ye. Chatcpu: An agile cpu design and verification platform with llm. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*, pp. 1–6, 2024.

Zeng Wang, Minghao Shao, Weimin Fu, Prithwish Basu Roy, Xiaolong Guo, Ramesh Karri, Muhammad Shafique, Johann Knechtel, and Ozgur Sinanoglu. Harmchip: Evaluating hardware security centric llm safety via jailbreak benchmarking. *arXiv preprint arXiv:2604.17093*, 2026.

Hasini Witharana, Aruna Jayasena, Andrew Whigham, and Prabhat Mishra. Automated generation of security assertions for rtl models. *ACM Journal on Emerging Technologies in Computing Systems*, 19(1):1–27, 2023.

Weihua Xiao, Zeng Wang, Minghao Shao, Raghu Vamshi Hemadri, Ozgur Sinanoglu, Muhammad Shafique, Johann Knechtel, Siddharth Garg, and Ramesh Karri. Trojanloc: Llm-based framework for rtl trojan localization. *arXiv preprint arXiv:2512.00591*, 2025.

Ke Xu, Jialin Sun, Yuchen Hu, Xinwei Fang, Weiwei Shan, Xi Wang, and Zhe Jiang. Meic: Rethinking rtl debug automation using llms. In *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, pp. 1–9, 2024.

Rozhin Yasaei, Luke Chen, Shih-Yuan Yu, and Mohammad Abdullah Al Faruque. Hardware trojan detection using graph neural networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 44(1):25–38, 2022.

Charles Young, Hao Yu, Dezhi Ran, Qingchen Zhai, Tianqi Qiu, Frank Qu, Bangyan Wang, Yuan Xie, and Tao Xie. Liveverilogeval: Contamination free and automatically scalable benchmark for verilog code generation. In *2026 Design, Automation & Test in Europe Conference (DATE)*, pp. 1–7. IEEE, 2026.

Jiazheng Zhang, Cheng Liu, Long Cheng, Xiaowei Li, and Huawei Li. Understanding and mitigating errors of llm-generated rtl code. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2026.

Yujie Zhao, Hejia Zhang, Hanxian Huang, Zhongming Yu, and Jishen Zhao. Mage: A multi-agent engine for automated rtl code generation. In *2025 62nd ACM/IEEE Design Automation Conference (DAC)*, pp. 1–7. IEEE, 2025.

Zhantong Zhu, Kangbo Bai, and Tianyu Jia. Chat-a 2: An llm-aided design space exploration framework for high-performance cpu design. In *2026 31st Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 540–546. IEEE, 2026.

# 7 APPENDIX

This appendix supplies the material referenced from the main text. Every claim in the main text is supported within the main text itself; the sections below provide the underlying detail.

- **Appendix A** — Per-combination detailed results: one longtable per LLM×agent-base combination, with the average-score column heat-colored on the paper's color scale.
- **Appendix B** — The FindHT mitigation as a full algorithm listing and its control mechanisms.
- **Appendix C** — The TrustMe self-iterative Trojan-design loop as an algorithm listing.
- **Appendix D** — Benchmark reproduction: the complete run workflow.
- **Appendix E** — Evaluation protocol: benchmark composition, scoring rubric, metric definitions, judge prompt structure, and the failure taxonomy.
- **Appendix F** — Benchmark artifacts and the directory structure of a run.
- **Appendix G** — A worked Trojan example traced through the characterization task.
- **Appendix H** — Responsible-disclosure and dual-use considerations.

# A PER-COMBINATION DETAILED RESULTS

Table 1 in the main text condenses each of the 15 combinations to seven summary numbers. This section reports the full per-dimension breakdown for every combination. For each combination we extract the "group summary" sheet of its final workbook and reproduce columns A–J (dimension, group, question count, executed, completed, scored, average score in percent, full-score accuracy in percent, and failure rate in percent). The raw fraction column for the average score is omitted because it is identical to the percentage column rescaled by $100\times$.

The *average-score* column is heat-colored with the same ten-color scale used for the category heatmap of the main text, so the reader can read each combination against the same visual grammar. The scale runs from 0–10% in dark red `#6A0624` through the light middle to 90–100% in dark blue `#053061`; the cell text is white on dark cells and black on light cells, chosen by relative luminance. Figure 5 shows the legend.

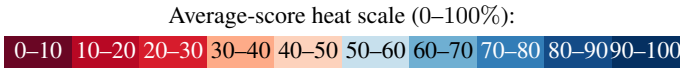



Figure 5: Heat-scale legend for the average-score column in the fifteen per-combination tables.

Figure **??** reproduces the eight-category average-score heatmap at a readable size, placing the fifteen per-combination tables in context. The four dataset codes used throughout are **H** = book_qa (hardware-security knowledge QA), **A** = htbench_core (core Trojan-detection suite), **M** = module_loc_cwe (module-level CWE localization), and **S** = soc_loc_extension (SoC-level extension). The remaining groups (L-level, suite, design level, task category, question type, execution mode, backend, agent framework, and their cross product) use their native codes unchanged.

## A.1 CROSS-COMBINATION DATASET PROFILE

Before reading the fifteen per-combination tables, it helps to see the four datasets side by side. Table 2 places the per-dataset average score of every combination in one view, together with the per-dataset Agent-Token cost and per-question time, using the same heat scale as the per-combination tables.

Three patterns stand out. First, **A** (htbench_core) is the hardest dataset by a clear margin: its cross-combination mean average score is 44.0% and no combination exceeds 55.7%, whereas **H** (book_qa, mean 59.8%), **M** (module_loc_cwe, mean 62.3%), and **S** (soc_loc_extension, mean 67.7%) all sit well above it. This is by construction: A concentrates the agentic L2/L3 attack-detection tasks, while M and S award substantial partial credit for evidence-backed localization. Second, a high average score on M and S does *not* imply solved tasks — the per-combination tables show that the full-score accuracy on M is at or near 0% for most combinations, meaning agents routinely produce plausible but not exactly-CWE-correct localizations. Third, the cost is concentrated in A: for the

strongest combination (codex + deepseek-v4-pro), the 246 questions of A consume $174.7$M of the $225.8$M total tokens ($77\%$) and $337$ s per question on average, while the $335$ knowledge questions of H consume only $13.5$M. The cheapest combination (claude_code + gpt-5.6-sol) spends $19.6$M in total, an $11.5\times$ gap that Appendix A traces combination by combination.

Table 2: Per-dataset average score (heat-colored, %) across the 15 combinations, with the per-dataset Agent-Token cost (M) and average time (s/question). Dataset codes: H = book_qa, A = htbench_core, M = module_loc_cwe, S = soc_loc_extension.

| | | Avg. score (%) | | | | Token (M) / Time (s) | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Base | H | A | M | S | H | A | M | S |
| deepseek-v4-flash | claude_code | 60.4 | 49.3 | 65.7 | 70.9 | 3.60 / 9.6 | 88.86 / 67.4 | 6.75 / 24.5 | 5.25 / 46.0 |
| deepseek-v4-pro | claude_code | 59.5 | 47.8 | 69.2 | 91.4 | 4.70 / 25.9 | 79.89 / 178.4 | 8.87 / 106.6 | 8.19 / 137.4 |
| glm-5.3 | claude_code | 65.2 | 46.5 | 66.7 | 69.0 | 5.65 / 96.9 | 31.00 / 125.9 | 11.05 / 242.2 | 1.57 / 106.7 |
| glm-x-preview-260804 | claude_code | 63.3 | 43.0 | 66.8 | 71.3 | 7.21 / 90.8 | 36.25 / 514.4 | 19.67 / 375.4 | 6.30 / 234.0 |
| gpt-5.6-sol | claude_code | 54.1 | 40.9 | 69.0 | 67.3 | 3.01 / 65.7 | 11.82 / 124.3 | 3.93 / 74.4 | 0.86 / 92.6 |
| qwen3.8-flash | claude_code | 66.0 | 44.6 | 67.0 | 82.6 | 5.40 / 43.8 | 43.52 / 192.4 | 8.89 / 95.7 | 4.28 / 133.2 |
| deepseek-v4-flash | codex | 61.4 | 55.7 | 67.0 | 79.6 | 7.25 / 30.6 | 152.25 / 119.7 | 4.39 / 33.4 | 14.46 / 143.3 |
| deepseek-v4-pro | codex | 62.5 | 53.9 | 70.6 | 73.8 | 13.54 / 34.4 | 174.68 / 337.4 | 28.02 / 109.3 | 9.55 / 134.7 |
| glm-5.3 | codex | 51.6 | 32.0 | 61.1 | 56.8 | 10.32 / 189.3 | 39.26 / 233.0 | 12.31 / 92.3 | 3.65 / 108.6 |
| glm-x-preview-260804 | codex | 57.7 | 43.9 | 48.9 | 67.9 | 5.73 / 58.2 | 53.36 / 230.5 | 6.21 / 187.7 | 2.51 / 128.8 |
| gpt-5.6-sol | codex | 56.6 | 46.6 | 68.0 | 74.4 | 9.63 / 68.2 | 49.31 / 92.8 | 9.18 / 112.2 | 4.93 / 103.2 |
| qwen3.8-flash | codex | 68.2 | 41.7 | 67.7 | 43.1 | 10.56 / 56.4 | 58.72 / 328.1 | 9.94 / 175.4 | 4.94 / 486.0 |
| glm-5.3 | deepseek_harness | 61.6 | 37.8 | 40.6 | 50.8 | 10.91 / 178.6 | 74.40 / 552.2 | 21.92 / 744.8 | 6.27 / 317.0 |
| glm-5.3 | opencode | 61.1 | 43.9 | 61.7 | 81.0 | 7.80 / 87.1 | 73.72 / 383.7 | 17.10 / 400.8 | 1.67 / 96.9 |
| glm-5.3 | trae_agent | 48.1 | 32.1 | 43.9 | 35.5 | 13.29 / 161.4 | 81.27 / 584.6 | 40.15 / 606.7 | 11.22 / 512.3 |

Table 3: Detailed per-dimension breakdown for deepseek-v4-flash + claude_code.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 321 | 316 | 60.4 | 53.2 | 46.1 |
| Dataset | A | 246 | 246 | 246 | 246 | 49.3 | 35.0 | 65.0 |
| Dataset | M | 225 | 225 | 225 | 225 | 65.7 | 0.0 | 100.0 |
| Dataset | S | 10 | 10 | 10 | 10 | 70.9 | 10.0 | 90.0 |
| L-Level | L2 | 139 | 139 | 139 | 139 | 49.2 | 34.5 | 65.5 |
| L-Level | L3 | 107 | 107 | 107 | 107 | 49.4 | 35.5 | 64.5 |
| L-Level | N/A | 570 | 556 | 556 | 551 | 62.8 | 30.7 | 68.7 |
| Suite | BOOK | 335 | 321 | 321 | 316 | 60.4 | 53.2 | 46.1 |
| Suite | L2 | 139 | 139 | 139 | 139 | 49.2 | 34.5 | 65.5 |
| Suite | L3 | 107 | 107 | 107 | 107 | 49.4 | 35.5 | 64.5 |
| Suite | MODULE-EXT | 225 | 225 | 225 | 225 | 65.7 | 0.0 | 100.0 |
| Suite | SOC-EXT | 10 | 10 | 10 | 10 | 70.9 | 10.0 | 90.0 |
| Design Level | ip | 126 | 126 | 126 | 126 | 44.7 | 30.2 | 69.8 |
| Design Level | knowledge | 335 | 321 | 321 | 316 | 60.4 | 53.2 | 46.1 |
| Design Level | module | 289 | 289 | 289 | 289 | 64.0 | 9.0 | 91.0 |
| Design Level | soc | 66 | 66 | 66 | 66 | 53.0 | 34.8 | 65.2 |
| Task Category | attribution | 34 | 34 | 34 | 34 | 80.9 | 73.5 | 26.5 |
| Task Category | cwe_classification | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Task Category | detection | 60 | 60 | 60 | 60 | 66.7 | 66.7 | 33.3 |
| Task Category | knowledge | 335 | 321 | 321 | 316 | 60.4 | 53.2 | 46.1 |
| Task Category | localization | 51 | 51 | 51 | 51 | 52.6 | 3.9 | 96.1 |
| Task Category | localization_cwe | 235 | 235 | 235 | 235 | 66.0 | 0.4 | 99.6 |
| Task Category | remediation | 32 | 32 | 32 | 32 | 50.0 | 46.9 | 53.1 |
| Task Category | characterization | 62 | 62 | 62 | 62 | 17.7 | 6.5 | 93.5 |
| Question Type | ATTR | 34 | 34 | 34 | 34 | 80.9 | 73.5 | 26.5 |
| Question Type | DET_ACT | 12 | 12 | 12 | 12 | 91.7 | 91.7 | 8.3 |
| Question Type | DET_BEH | 40 | 40 | 40 | 40 | 55.0 | 55.0 | 45.0 |
| Question Type | DET_VUL | 8 | 8 | 8 | 8 | 87.5 | 87.5 | 12.5 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Question Type | FIX | 16 | 16 | 16 | 16 | 50.0 | 50.0 | 50.0 |
| Question Type | FIX_CWE | 16 | 16 | 16 | 16 | 50.0 | 43.8 | 56.2 |
| Question Type | LOC | 61 | 61 | 61 | 61 | 55.6 | 4.9 | 95.1 |
| Question Type | TRUSTHUB | 62 | 62 | 62 | 62 | 17.7 | 6.5 | 93.5 |
| Question Type | book_long_answer | 105 | 95 | 95 | 93 | 32.3 | 17.2 | 81.1 |
| Question Type | book_mathematical | 12 | 10 | 10 | 10 | 30.0 | 30.0 | 70.0 |
| Question Type | book_short_answer | 88 | 86 | 86 | 84 | 58.3 | 47.6 | 51.2 |
| Question Type | book_true_false | 130 | 130 | 130 | 129 | 84.5 | 84.5 | 15.4 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 225 | 225 | 65.7 | 0.0 | 100.0 |
| Execution Mode | agentic | 802 | 802 | 802 | 797 | 58.6 | 32.0 | 67.6 |
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | universal_agent_adapter | 802 | 802 | 802 | 797 | 58.6 | 32.0 | 67.6 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | claude_code | 802 | 802 | 802 | 797 | 58.6 | 32.0 | 67.6 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 24 | 24 | 83.3 | 79.2 | 20.8 |
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 11 | 11 | 90.9 | 90.9 | 9.1 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 27 | 27 | 48.1 | 48.1 | 51.9 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 5 | 5 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 22 | 22 | 72.4 | 9.1 | 90.9 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 50 | 50 | 19.0 | 8.0 | 92.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 10 | 10 | 75.0 | 60.0 | 40.0 |
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 1 | 1 | 100.0 | 100.0 | 0.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 13 | 13 | 69.2 | 69.2 | 30.8 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 8 | 8 | 87.5 | 87.5 | 12.5 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 2 | 2 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 16 | 16 | 50.0 | 50.0 | 50.0 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 16 | 16 | 50.0 | 43.8 | 56.2 |
| L-Level × Type | L3 × LOC | 29 | 29 | 29 | 29 | 37.5 | 0.0 | 100.0 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 12 | 12 | 12.5 | 0.0 | 100.0 |
| L-Level × Type | N/A × LOC | 10 | 10 | 10 | 10 | 70.9 | 10.0 | 90.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 95 | 93 | 32.3 | 17.2 | 81.1 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 10 | 10 | 30.0 | 30.0 | 70.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 86 | 84 | 58.3 | 47.6 | 51.2 |
| L-Level × Type | N/A × book_true_false | 130 | 130 | 130 | 129 | 84.5 | 84.5 | 15.4 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 225 | 225 | 65.7 | 0.0 | 100.0 |

Table 4: Detailed per-dimension breakdown for deepseek-v4-pro + claude_code.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 321 | 316 | 59.5 | 52.2 | 47.0 |
| Dataset | A | 246 | 246 | 244 | 246 | 47.8 | 32.1 | 67.9 |
| Dataset | M | 225 | 225 | 225 | 225 | 69.2 | 0.0 | 100.0 |
| Dataset | S | 10 | 10 | 10 | 10 | 91.4 | 20.0 | 80.0 |
| L-Level | L2 | 139 | 139 | 138 | 139 | 46.0 | 30.9 | 69.1 |
| L-Level | L3 | 107 | 107 | 106 | 107 | 50.2 | 33.6 | 66.4 |
| L-Level | N/A | 570 | 556 | 556 | 551 | 64.0 | 30.3 | 69.1 |
| Suite | BOOK | 335 | 321 | 321 | 316 | 59.5 | 52.2 | 47.0 |
| Suite | L2 | 139 | 139 | 138 | 139 | 46.0 | 30.9 | 69.1 |
| Suite | L3 | 107 | 107 | 106 | 107 | 50.2 | 33.6 | 66.4 |
| Suite | MODULE-EXT | 225 | 225 | 225 | 225 | 69.2 | 0.0 | 100.0 |
| Suite | SOC-EXT | 10 | 10 | 10 | 10 | 91.4 | 20.0 | 80.0 |
| Design Level | ip | 126 | 126 | 125 | 126 | 41.1 | 25.4 | 74.6 |
| Design Level | knowledge | 335 | 321 | 321 | 316 | 59.5 | 52.2 | 47.0 |
| Design Level | module | 289 | 289 | 289 | 289 | 66.4 | 9.0 | 91.0 |
| Design Level | soc | 66 | 66 | 65 | 66 | 58.6 | 34.8 | 65.2 |
| Task Category | attribution | 34 | 34 | 34 | 34 | 64.7 | 61.8 | 38.2 |
| Task Category | cwe_classification | 7 | 7 | 7 | 7 | 14.3 | 14.3 | 85.7 |
| Task Category | detection | 60 | 60 | 59 | 60 | 63.3 | 63.3 | 36.7 |
| Task Category | knowledge | 335 | 321 | 321 | 316 | 59.5 | 52.2 | 47.0 |
| Task Category | localization | 51 | 51 | 50 | 51 | 59.0 | 5.9 | 94.1 |
| Task Category | localization_cwe | 235 | 235 | 235 | 235 | 70.1 | 0.9 | 99.1 |
| Task Category | remediation | 32 | 32 | 32 | 32 | 48.4 | 43.8 | 56.2 |
| Task Category | characterization | 62 | 62 | 62 | 62 | 17.7 | 3.2 | 96.8 |
| Question Type | ATTR | 34 | 34 | 34 | 34 | 64.7 | 61.8 | 38.2 |
| Question Type | DET_ACT | 12 | 12 | 12 | 12 | 91.7 | 91.7 | 8.3 |
| Question Type | DET_BEH | 40 | 40 | 39 | 40 | 50.0 | 50.0 | 50.0 |
| Question Type | DET_VUL | 8 | 8 | 8 | 8 | 87.5 | 87.5 | 12.5 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 7 | 7 | 14.3 | 14.3 | 85.7 |
| Question Type | FIX | 16 | 16 | 16 | 16 | 53.1 | 50.0 | 50.0 |
| Question Type | FIX_CWE | 16 | 16 | 16 | 16 | 43.8 | 37.5 | 62.5 |
| Question Type | LOC | 61 | 61 | 60 | 61 | 64.3 | 8.2 | 91.8 |
| Question Type | TRUSTHUB | 62 | 62 | 62 | 62 | 17.7 | 3.2 | 96.8 |
| Question Type | book_long_answer | 105 | 95 | 95 | 93 | 24.7 | 9.7 | 88.4 |
| Question Type | book_mathematical | 12 | 10 | 10 | 10 | 30.0 | 30.0 | 70.0 |
| Question Type | book_short_answer | 88 | 86 | 86 | 84 | 60.7 | 50.0 | 48.8 |
| Question Type | book_true_false | 130 | 130 | 130 | 129 | 86.0 | 86.0 | 13.8 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 225 | 225 | 69.2 | 0.0 | 100.0 |
| Execution Mode | agentic | 802 | 802 | 800 | 797 | 59.0 | 30.9 | 68.7 |
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | universal_agent_adapter | 802 | 802 | 800 | 797 | 59.0 | 30.9 | 68.7 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | claude_code | 802 | 802 | 800 | 797 | 59.0 | 30.9 | 68.7 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 24 | 24 | 66.7 | 66.7 | 33.3 |
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 11 | 11 | 90.9 | 90.9 | 9.1 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 26 | 27 | 40.7 | 40.7 | 59.3 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 5 | 5 | 20.0 | 20.0 | 80.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 22 | 22 | 74.5 | 13.6 | 86.4 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 50 | 50 | 19.0 | 4.0 | 96.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 10 | 10 | 60.0 | 50.0 | 50.0 |
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 1 | 1 | 100.0 | 100.0 | 0.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 13 | 13 | 69.2 | 69.2 | 30.8 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 8 | 8 | 87.5 | 87.5 | 12.5 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 2 | 2 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 16 | 16 | 53.1 | 50.0 | 50.0 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 16 | 16 | 43.8 | 37.5 | 62.5 |
| L-Level × Type | L3 × LOC | 29 | 29 | 28 | 29 | 47.3 | 0.0 | 100.0 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 12 | 12 | 12.5 | 0.0 | 100.0 |
| L-Level × Type | N/A × LOC | 10 | 10 | 10 | 10 | 91.4 | 20.0 | 80.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 95 | 93 | 24.7 | 9.7 | 88.4 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 10 | 10 | 30.0 | 30.0 | 70.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 86 | 84 | 60.7 | 50.0 | 48.8 |

*Table 4 (continued).*

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| L-Level × Type | N/A × book_true_false | 130 | 130 | 130 | 129 | 86.0 | 86.0 | 13.8 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 225 | 225 | 69.2 | 0.0 | 100.0 |

Table 5: Detailed per-dimension breakdown for glm-5.3 + claude_code.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 317 | 316 | 65.2 | 57.9 | 41.4 |
| Dataset | A | 246 | 246 | 239 | 245 | 46.5 | 31.8 | 67.9 |
| Dataset | M | 225 | 225 | 223 | 225 | 66.7 | 0.0 | 100.0 |
| Dataset | S | 10 | 10 | 10 | 10 | 69.0 | 10.0 | 90.0 |
| L-Level | L2 | 139 | 139 | 139 | 139 | 46.1 | 30.9 | 69.1 |
| L-Level | L3 | 107 | 107 | 100 | 106 | 47.0 | 33.0 | 66.4 |
| L-Level | N/A | 570 | 556 | 550 | 551 | 65.9 | 33.4 | 66.0 |
| Suite | BOOK | 335 | 321 | 317 | 316 | 65.2 | 57.9 | 41.4 |
| Suite | L2 | 139 | 139 | 139 | 139 | 46.1 | 30.9 | 69.1 |
| Suite | L3 | 107 | 107 | 100 | 106 | 47.0 | 33.0 | 66.4 |
| Suite | MODULE-EXT | 225 | 225 | 223 | 225 | 66.7 | 0.0 | 100.0 |
| Suite | SOC-EXT | 10 | 10 | 10 | 10 | 69.0 | 10.0 | 90.0 |
| Design Level | ip | 126 | 126 | 126 | 126 | 41.9 | 27.0 | 73.0 |
| Design Level | knowledge | 335 | 321 | 317 | 316 | 65.2 | 57.9 | 41.4 |
| Design Level | module | 289 | 289 | 287 | 289 | 65.6 | 10.0 | 90.0 |
| Design Level | soc | 66 | 66 | 59 | 65 | 43.7 | 24.6 | 74.2 |
| Task Category | attribution | 34 | 34 | 34 | 34 | 77.9 | 73.5 | 26.5 |
| Task Category | cwe_classification | 7 | 7 | 7 | 7 | 14.3 | 14.3 | 85.7 |
| Task Category | detection | 60 | 60 | 59 | 60 | 65.0 | 65.0 | 35.0 |
| Task Category | knowledge | 335 | 321 | 317 | 316 | 65.2 | 57.9 | 41.4 |
| Task Category | localization | 51 | 51 | 47 | 50 | 55.8 | 4.0 | 94.1 |
| Task Category | localization_cwe | 235 | 235 | 233 | 235 | 66.8 | 0.4 | 99.6 |
| Task Category | remediation | 32 | 32 | 30 | 32 | 35.9 | 31.2 | 68.8 |
| Task Category | characterization | 62 | 62 | 62 | 62 | 12.9 | 1.6 | 98.4 |
| Question Type | ATTR | 34 | 34 | 34 | 34 | 77.9 | 73.5 | 26.5 |
| Question Type | DET_ACT | 12 | 12 | 12 | 12 | 91.7 | 91.7 | 8.3 |
| Question Type | DET_BEH | 40 | 40 | 40 | 40 | 57.5 | 57.5 | 42.5 |
| Question Type | DET_VUL | 8 | 8 | 7 | 8 | 62.5 | 62.5 | 37.5 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 7 | 7 | 14.3 | 14.3 | 85.7 |
| Question Type | FIX | 16 | 16 | 15 | 16 | 37.5 | 37.5 | 62.5 |
| Question Type | FIX_CWE | 16 | 16 | 15 | 16 | 34.4 | 25.0 | 75.0 |
| Question Type | LOC | 61 | 61 | 57 | 60 | 58.0 | 5.0 | 93.4 |
| Question Type | TRUSTHUB | 62 | 62 | 62 | 62 | 12.9 | 1.6 | 98.4 |
| Question Type | book_long_answer | 105 | 95 | 95 | 93 | 46.8 | 31.2 | 67.4 |
| Question Type | book_mathematical | 12 | 10 | 9 | 10 | 30.0 | 30.0 | 70.0 |
| Question Type | book_short_answer | 88 | 86 | 85 | 84 | 58.9 | 48.8 | 50.0 |
| Question Type | book_true_false | 130 | 130 | 128 | 129 | 85.3 | 85.3 | 14.6 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 223 | 225 | 66.7 | 0.0 | 100.0 |
| Execution Mode | agentic | 802 | 802 | 789 | 796 | 59.9 | 32.9 | 66.6 |
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | universal_agent_adapter | 802 | 802 | 789 | 796 | 59.9 | 32.9 | 66.6 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | claude_code | 802 | 802 | 789 | 796 | 59.9 | 32.9 | 66.6 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 24 | 24 | 77.1 | 75.0 | 25.0 |
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 11 | 11 | 90.9 | 90.9 | 9.1 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 27 | 27 | 44.4 | 44.4 | 55.6 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 5 | 5 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 22 | 22 | 75.4 | 9.1 | 90.9 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 50 | 50 | 14.0 | 2.0 | 98.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 10 | 10 | 80.0 | 70.0 | 30.0 |
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 1 | 1 | 100.0 | 100.0 | 0.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 13 | 13 | 84.6 | 84.6 | 15.4 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 7 | 8 | 62.5 | 62.5 | 37.5 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 2 | 2 | 50.0 | 50.0 | 50.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 15 | 16 | 37.5 | 37.5 | 62.5 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 15 | 16 | 34.4 | 25.0 | 75.0 |

*Table 5 (continued).*

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| L-Level × Type | L3 × LOC | 29 | 29 | 25 | 28 | 40.3 | 0.0 | 96.6 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 12 | 12 | 8.3 | 0.0 | 100.0 |
| L-Level × Type | N/A × LOC | 10 | 10 | 10 | 10 | 69.0 | 10.0 | 90.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 95 | 93 | 46.8 | 31.2 | 67.4 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 9 | 10 | 30.0 | 30.0 | 70.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 85 | 84 | 58.9 | 48.8 | 50.0 |
| L-Level × Type | N/A × book_true_false | 130 | 130 | 128 | 129 | 85.3 | 85.3 | 14.6 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 223 | 225 | 66.7 | 0.0 | 100.0 |

Table 6: Detailed per-dimension breakdown for glm-x-preview-260804 + claude_code.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 319 | 316 | 63.3 | 57.0 | 42.4 |
| Dataset | A | 246 | 246 | 212 | 246 | 43.0 | 29.7 | 70.3 |
| Dataset | M | 225 | 225 | 223 | 225 | 66.8 | 0.0 | 100.0 |
| Dataset | S | 10 | 10 | 10 | 10 | 71.3 | 10.0 | 90.0 |
| L-Level | L2 | 139 | 139 | 125 | 139 | 45.2 | 31.7 | 68.3 |
| L-Level | L3 | 107 | 107 | 87 | 107 | 40.2 | 27.1 | 72.9 |
| L-Level | N/A | 570 | 556 | 552 | 551 | 64.9 | 32.8 | 66.5 |
| Suite | BOOK | 335 | 321 | 319 | 316 | 63.3 | 57.0 | 42.4 |
| Suite | L2 | 139 | 139 | 125 | 139 | 45.2 | 31.7 | 68.3 |
| Suite | L3 | 107 | 107 | 87 | 107 | 40.2 | 27.1 | 72.9 |
| Suite | MODULE-EXT | 225 | 225 | 223 | 225 | 66.8 | 0.0 | 100.0 |
| Suite | SOC-EXT | 10 | 10 | 10 | 10 | 71.3 | 10.0 | 90.0 |
| Design Level | ip | 126 | 126 | 110 | 126 | 39.7 | 26.2 | 73.8 |
| Design Level | knowledge | 335 | 321 | 319 | 316 | 63.3 | 57.0 | 42.4 |
| Design Level | module | 289 | 289 | 284 | 289 | 65.5 | 10.0 | 90.0 |
| Design Level | soc | 66 | 66 | 51 | 66 | 36.4 | 18.2 | 81.8 |
| Task Category | attribution | 34 | 34 | 34 | 34 | 70.6 | 67.6 | 32.4 |
| Task Category | cwe_classification | 7 | 7 | 4 | 7 | 14.3 | 14.3 | 85.7 |
| Task Category | detection | 60 | 60 | 52 | 60 | 61.7 | 61.7 | 38.3 |
| Task Category | knowledge | 335 | 321 | 319 | 316 | 63.3 | 57.0 | 42.4 |
| Task Category | localization | 51 | 51 | 44 | 51 | 51.4 | 5.9 | 94.1 |
| Task Category | localization_cwe | 235 | 235 | 233 | 235 | 67.0 | 0.4 | 99.6 |
| Task Category | remediation | 32 | 32 | 25 | 32 | 25.0 | 18.8 | 81.2 |
| Task Category | characterization | 62 | 62 | 53 | 62 | 15.3 | 4.8 | 95.2 |
| Question Type | ATTR | 34 | 34 | 34 | 34 | 70.6 | 67.6 | 32.4 |
| Question Type | DET_ACT | 12 | 12 | 9 | 12 | 75.0 | 75.0 | 25.0 |
| Question Type | DET_BEH | 40 | 40 | 36 | 40 | 57.5 | 57.5 | 42.5 |
| Question Type | DET_VUL | 8 | 8 | 7 | 8 | 62.5 | 62.5 | 37.5 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 4 | 7 | 14.3 | 14.3 | 85.7 |
| Question Type | FIX | 16 | 16 | 11 | 16 | 28.1 | 25.0 | 75.0 |
| Question Type | FIX_CWE | 16 | 16 | 14 | 16 | 21.9 | 12.5 | 87.5 |
| Question Type | LOC | 61 | 61 | 54 | 61 | 54.7 | 6.6 | 93.4 |
| Question Type | TRUSTHUB | 62 | 62 | 53 | 62 | 15.3 | 4.8 | 95.2 |
| Question Type | book_long_answer | 105 | 95 | 95 | 93 | 37.6 | 24.7 | 73.7 |
| Question Type | book_mathematical | 12 | 10 | 9 | 10 | 30.0 | 30.0 | 70.0 |
| Question Type | book_short_answer | 88 | 86 | 86 | 84 | 61.9 | 52.4 | 46.5 |
| Question Type | book_true_false | 130 | 130 | 129 | 129 | 85.3 | 85.3 | 14.6 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 223 | 225 | 66.8 | 0.0 | 100.0 |
| Execution Mode | agentic | 802 | 802 | 764 | 797 | 58.1 | 31.9 | 67.7 |
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | universal_agent_adapter | 802 | 802 | 764 | 797 | 58.1 | 31.9 | 67.7 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | claude_code | 802 | 802 | 764 | 797 | 58.1 | 31.9 | 67.7 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 24 | 24 | 70.8 | 70.8 | 29.2 |
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 9 | 11 | 81.8 | 81.8 | 18.2 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 25 | 27 | 44.4 | 44.4 | 55.6 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 2 | 5 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 22 | 22 | 73.9 | 13.6 | 86.4 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 43 | 50 | 17.0 | 6.0 | 94.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 10 | 10 | 70.0 | 60.0 | 40.0 |

*Table 6 (continued).*

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 0 | 1 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 11 | 13 | 84.6 | 84.6 | 15.4 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 7 | 8 | 62.5 | 62.5 | 37.5 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 2 | 2 | 50.0 | 50.0 | 50.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 11 | 16 | 28.1 | 25.0 | 75.0 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 14 | 16 | 21.9 | 12.5 | 87.5 |
| L-Level × Type | L3 × LOC | 29 | 29 | 22 | 29 | 34.4 | 0.0 | 100.0 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 10 | 12 | 8.3 | 0.0 | 100.0 |
| L-Level × Type | N/A × LOC | 10 | 10 | 10 | 10 | 71.3 | 10.0 | 90.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 95 | 93 | 37.6 | 24.7 | 73.7 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 9 | 10 | 30.0 | 30.0 | 70.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 86 | 84 | 61.9 | 52.4 | 46.5 |
| L-Level × Type | N/A × book_true_false | 130 | 130 | 129 | 129 | 85.3 | 85.3 | 14.6 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 223 | 225 | 66.8 | 0.0 | 100.0 |

Table 7: Detailed per-dimension breakdown for gpt-5.6-sol + claude_code.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 319 | 314 | 54.1 | 48.7 | 50.8 |
| Dataset | A | 246 | 246 | 246 | 240 | 40.9 | 31.2 | 69.5 |
| Dataset | M | 225 | 225 | 225 | 225 | 69.0 | 0.0 | 100.0 |
| Dataset | S | 10 | 10 | 10 | 10 | 67.3 | 0.0 | 100.0 |
| L-Level | L2 | 139 | 139 | 139 | 133 | 39.3 | 30.8 | 70.5 |
| L-Level | L3 | 107 | 107 | 107 | 107 | 42.8 | 31.8 | 68.2 |
| L-Level | N/A | 570 | 556 | 554 | 549 | 60.5 | 27.9 | 71.6 |
| Suite | BOOK | 335 | 321 | 319 | 314 | 54.1 | 48.7 | 50.8 |
| Suite | L2 | 139 | 139 | 139 | 133 | 39.3 | 30.8 | 70.5 |
| Suite | L3 | 107 | 107 | 107 | 107 | 42.8 | 31.8 | 68.2 |
| Suite | MODULE-EXT | 225 | 225 | 225 | 225 | 69.0 | 0.0 | 100.0 |
| Suite | SOC-EXT | 10 | 10 | 10 | 10 | 67.3 | 0.0 | 100.0 |
| Design Level | ip | 126 | 126 | 126 | 120 | 34.0 | 26.7 | 74.6 |
| Design Level | knowledge | 335 | 321 | 319 | 314 | 54.1 | 48.7 | 50.8 |
| Design Level | module | 289 | 289 | 289 | 289 | 66.2 | 9.3 | 90.7 |
| Design Level | soc | 66 | 66 | 66 | 66 | 42.4 | 24.2 | 75.8 |
| Task Category | attribution | 34 | 34 | 34 | 34 | 72.1 | 70.6 | 29.4 |
| Task Category | cwe_classification | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Task Category | detection | 60 | 60 | 60 | 60 | 66.7 | 66.7 | 33.3 |
| Task Category | knowledge | 335 | 321 | 319 | 314 | 54.1 | 48.7 | 50.8 |
| Task Category | localization | 51 | 51 | 51 | 51 | 32.5 | 0.0 | 100.0 |
| Task Category | localization_cwe | 235 | 235 | 235 | 235 | 69.0 | 0.0 | 100.0 |
| Task Category | remediation | 32 | 32 | 32 | 32 | 31.2 | 28.1 | 71.9 |
| Task Category | characterization | 62 | 62 | 62 | 56 | 12.5 | 3.6 | 96.8 |
| Question Type | ATTR | 34 | 34 | 34 | 34 | 72.1 | 70.6 | 29.4 |
| Question Type | DET_ACT | 12 | 12 | 12 | 12 | 83.3 | 83.3 | 16.7 |
| Question Type | DET_BEH | 40 | 40 | 40 | 40 | 57.5 | 57.5 | 42.5 |
| Question Type | DET_VUL | 8 | 8 | 8 | 8 | 87.5 | 87.5 | 12.5 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Question Type | FIX | 16 | 16 | 16 | 16 | 37.5 | 37.5 | 62.5 |
| Question Type | FIX_CWE | 16 | 16 | 16 | 16 | 25.0 | 18.8 | 81.2 |
| Question Type | LOC | 61 | 61 | 61 | 61 | 38.2 | 0.0 | 100.0 |
| Question Type | TRUSTHUB | 62 | 62 | 62 | 56 | 12.5 | 3.6 | 96.8 |
| Question Type | book_long_answer | 105 | 95 | 94 | 92 | 21.7 | 12.0 | 86.3 |
| Question Type | book_mathematical | 12 | 10 | 10 | 10 | 30.0 | 30.0 | 70.0 |
| Question Type | book_short_answer | 88 | 86 | 85 | 83 | 49.4 | 39.8 | 59.3 |
| Question Type | book_true_false | 130 | 130 | 130 | 129 | 82.2 | 82.2 | 17.7 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 225 | 225 | 69.0 | 0.0 | 100.0 |
| Execution Mode | agentic | 802 | 802 | 800 | 789 | 54.5 | 28.9 | 70.9 |
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | universal_agent_adapter | 802 | 802 | 800 | 789 | 54.5 | 28.9 | 70.9 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | claude_code | 802 | 802 | 800 | 789 | 54.5 | 28.9 | 70.9 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 24 | 24 | 72.9 | 70.8 | 29.2 |

*Table 7 (continued).*

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 11 | 11 | 81.8 | 81.8 | 18.2 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 27 | 27 | 48.1 | 48.1 | 51.9 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 5 | 5 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 22 | 22 | 33.0 | 0.0 | 100.0 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 50 | 44 | 12.5 | 4.5 | 96.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 10 | 10 | 70.0 | 70.0 | 30.0 |
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 1 | 1 | 100.0 | 100.0 | 0.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 13 | 13 | 76.9 | 76.9 | 23.1 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 8 | 8 | 87.5 | 87.5 | 12.5 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 2 | 2 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 16 | 16 | 37.5 | 37.5 | 62.5 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 16 | 16 | 25.0 | 18.8 | 81.2 |
| L-Level × Type | L3 × LOC | 29 | 29 | 29 | 29 | 32.1 | 0.0 | 100.0 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 12 | 12 | 12.5 | 0.0 | 100.0 |
| L-Level × Type | N/A × LOC | 10 | 10 | 10 | 10 | 67.3 | 0.0 | 100.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 94 | 92 | 21.7 | 12.0 | 86.3 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 10 | 10 | 30.0 | 30.0 | 70.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 85 | 83 | 49.4 | 39.8 | 59.3 |
| L-Level × Type | N/A × book_true_false | 130 | 130 | 130 | 129 | 82.2 | 82.2 | 17.7 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 225 | 225 | 69.0 | 0.0 | 100.0 |

Table 8: Detailed per-dimension breakdown for qwen3.8-flash + claude_code.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 321 | 316 | 66.0 | 58.5 | 40.8 |
| Dataset | A | 246 | 246 | 241 | 246 | 44.6 | 31.3 | 68.7 |
| Dataset | M | 225 | 225 | 225 | 225 | 67.0 | 0.0 | 100.0 |
| Dataset | S | 10 | 10 | 10 | 10 | 82.6 | 30.0 | 70.0 |
| L-Level | L2 | 139 | 139 | 139 | 139 | 45.8 | 32.4 | 67.6 |
| L-Level | L3 | 107 | 107 | 102 | 107 | 42.9 | 29.9 | 70.1 |
| L-Level | N/A | 570 | 556 | 556 | 551 | 66.7 | 34.1 | 65.3 |
| Suite | BOOK | 335 | 321 | 321 | 316 | 66.0 | 58.5 | 40.8 |
| Suite | L2 | 139 | 139 | 139 | 139 | 45.8 | 32.4 | 67.6 |
| Suite | L3 | 107 | 107 | 102 | 107 | 42.9 | 29.9 | 70.1 |
| Suite | MODULE-EXT | 225 | 225 | 225 | 225 | 67.0 | 0.0 | 100.0 |
| Suite | SOC-EXT | 10 | 10 | 10 | 10 | 82.6 | 30.0 | 70.0 |
| Design Level | ip | 126 | 126 | 126 | 126 | 41.9 | 27.8 | 72.2 |
| Design Level | knowledge | 335 | 321 | 321 | 316 | 66.0 | 58.5 | 40.8 |
| Design Level | module | 289 | 289 | 289 | 289 | 64.7 | 9.3 | 90.7 |
| Design Level | soc | 66 | 66 | 61 | 66 | 44.0 | 27.3 | 72.7 |
| Task Category | attribution | 34 | 34 | 34 | 34 | 69.1 | 67.6 | 32.4 |
| Task Category | cwe_classification | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Task Category | detection | 60 | 60 | 60 | 60 | 65.0 | 65.0 | 35.0 |
| Task Category | knowledge | 335 | 321 | 321 | 316 | 66.0 | 58.5 | 40.8 |
| Task Category | localization | 51 | 51 | 46 | 51 | 53.3 | 2.0 | 98.0 |
| Task Category | localization_cwe | 235 | 235 | 235 | 235 | 67.7 | 1.3 | 98.7 |
| Task Category | remediation | 32 | 32 | 32 | 32 | 31.2 | 28.1 | 71.9 |
| Task Category | characterization | 62 | 62 | 62 | 62 | 16.1 | 8.1 | 91.9 |
| Question Type | ATTR | 34 | 34 | 34 | 34 | 69.1 | 67.6 | 32.4 |
| Question Type | DET_ACT | 12 | 12 | 12 | 12 | 91.7 | 91.7 | 8.3 |
| Question Type | DET_BEH | 40 | 40 | 40 | 40 | 55.0 | 55.0 | 45.0 |
| Question Type | DET_VUL | 8 | 8 | 8 | 8 | 75.0 | 75.0 | 25.0 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Question Type | FIX | 16 | 16 | 16 | 16 | 40.6 | 37.5 | 62.5 |
| Question Type | FIX_CWE | 16 | 16 | 16 | 16 | 21.9 | 18.8 | 81.2 |
| Question Type | LOC | 61 | 61 | 56 | 61 | 58.1 | 6.6 | 93.4 |
| Question Type | TRUSTHUB | 62 | 62 | 62 | 62 | 16.1 | 8.1 | 91.9 |
| Question Type | book_long_answer | 105 | 95 | 95 | 93 | 43.5 | 26.9 | 71.6 |
| Question Type | book_mathematical | 12 | 10 | 10 | 10 | 30.0 | 30.0 | 70.0 |
| Question Type | book_short_answer | 88 | 86 | 86 | 84 | 63.1 | 53.6 | 45.3 |
| Question Type | book_true_false | 130 | 130 | 130 | 129 | 86.8 | 86.8 | 13.1 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 225 | 225 | 67.0 | 0.0 | 100.0 |
| Execution Mode | agentic | 802 | 802 | 797 | 797 | 59.9 | 33.2 | 66.3 |

*Table 8 (continued).*

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | universal_agent_adapter | 802 | 802 | 797 | 797 | 59.9 | 33.2 | 66.3 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | claude_code | 802 | 802 | 797 | 797 | 59.9 | 33.2 | 66.3 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 24 | 24 | 70.8 | 70.8 | 29.2 |
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 11 | 11 | 90.9 | 90.9 | 9.1 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 27 | 27 | 44.4 | 44.4 | 55.6 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 5 | 5 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 22 | 22 | 73.8 | 4.5 | 95.5 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 50 | 50 | 17.0 | 10.0 | 90.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 10 | 10 | 65.0 | 60.0 | 40.0 |
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 1 | 1 | 100.0 | 100.0 | 0.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 13 | 13 | 76.9 | 76.9 | 23.1 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 8 | 8 | 75.0 | 75.0 | 25.0 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 2 | 2 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 16 | 16 | 40.6 | 37.5 | 62.5 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 16 | 16 | 21.9 | 18.8 | 81.2 |
| L-Level × Type | L3 × LOC | 29 | 29 | 24 | 29 | 37.7 | 0.0 | 100.0 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 12 | 12 | 12.5 | 0.0 | 100.0 |
| L-Level × Type | N/A × LOC | 10 | 10 | 10 | 10 | 82.6 | 30.0 | 70.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 95 | 93 | 43.5 | 26.9 | 71.6 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 10 | 10 | 30.0 | 30.0 | 70.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 86 | 84 | 63.1 | 53.6 | 45.3 |
| L-Level × Type | N/A × book_true_false | 130 | 130 | 130 | 129 | 86.8 | 86.8 | 13.1 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 225 | 225 | 67.0 | 0.0 | 100.0 |

Table 9: Detailed per-dimension breakdown for deepseek-v4-flash + codex.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 320 | 316 | 61.4 | 54.7 | 44.5 |
| Dataset | A | 246 | 246 | 244 | 246 | 55.7 | 41.9 | 58.1 |
| Dataset | M | 225 | 225 | 225 | 225 | 67.0 | 0.0 | 100.0 |
| Dataset | S | 10 | 10 | 10 | 10 | 79.6 | 10.0 | 90.0 |
| L-Level | L2 | 139 | 139 | 138 | 139 | 54.3 | 39.6 | 60.4 |
| L-Level | L3 | 107 | 107 | 106 | 107 | 57.4 | 44.9 | 55.1 |
| L-Level | N/A | 570 | 556 | 555 | 551 | 64.0 | 31.6 | 67.8 |
| Suite | BOOK | 335 | 321 | 320 | 316 | 61.4 | 54.7 | 44.5 |
| Suite | L2 | 139 | 139 | 138 | 139 | 54.3 | 39.6 | 60.4 |
| Suite | L3 | 107 | 107 | 106 | 107 | 57.4 | 44.9 | 55.1 |
| Suite | MODULE-EXT | 225 | 225 | 225 | 225 | 67.0 | 0.0 | 100.0 |
| Suite | SOC-EXT | 10 | 10 | 10 | 10 | 79.6 | 10.0 | 90.0 |
| Design Level | ip | 126 | 126 | 125 | 126 | 49.9 | 35.7 | 64.3 |
| Design Level | knowledge | 335 | 321 | 320 | 316 | 61.4 | 54.7 | 44.5 |
| Design Level | module | 289 | 289 | 289 | 289 | 65.7 | 9.7 | 90.3 |
| Design Level | soc | 66 | 66 | 65 | 66 | 64.7 | 47.0 | 53.0 |
| Task Category | attribution | 34 | 34 | 34 | 34 | 79.4 | 73.5 | 26.5 |
| Task Category | cwe_classification | 7 | 7 | 7 | 7 | 14.3 | 14.3 | 85.7 |
| Task Category | detection | 60 | 60 | 60 | 60 | 68.3 | 68.3 | 31.7 |
| Task Category | knowledge | 335 | 321 | 320 | 316 | 61.4 | 54.7 | 44.5 |
| Task Category | localization | 51 | 51 | 50 | 51 | 57.8 | 7.8 | 92.2 |
| Task Category | localization_cwe | 235 | 235 | 235 | 235 | 67.5 | 0.4 | 99.6 |
| Task Category | remediation | 32 | 32 | 32 | 32 | 64.1 | 62.5 | 37.5 |
| Task Category | characterization | 62 | 62 | 61 | 62 | 29.0 | 19.4 | 80.6 |
| Question Type | ATTR | 34 | 34 | 34 | 34 | 79.4 | 73.5 | 26.5 |
| Question Type | DET_ACT | 12 | 12 | 12 | 12 | 91.7 | 91.7 | 8.3 |
| Question Type | DET_BEH | 40 | 40 | 40 | 40 | 57.5 | 57.5 | 42.5 |
| Question Type | DET_VUL | 8 | 8 | 8 | 8 | 87.5 | 87.5 | 12.5 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 7 | 7 | 14.3 | 14.3 | 85.7 |
| Question Type | FIX | 16 | 16 | 16 | 16 | 84.4 | 81.2 | 18.8 |
| Question Type | FIX_CWE | 16 | 16 | 16 | 16 | 43.8 | 43.8 | 56.2 |
| Question Type | LOC | 61 | 61 | 60 | 61 | 61.4 | 8.2 | 91.8 |
| Question Type | TRUSTHUB | 62 | 62 | 61 | 62 | 29.0 | 19.4 | 80.6 |

*Table 9 (continued).*

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Question Type | book_long_answer | 105 | 95 | 95 | 93 | 36.0 | 23.7 | 74.7 |
| Question Type | book_mathematical | 12 | 10 | 10 | 10 | 40.0 | 40.0 | 60.0 |
| Question Type | book_short_answer | 88 | 86 | 86 | 84 | 58.9 | 47.6 | 51.2 |
| Question Type | book_true_false | 130 | 130 | 129 | 129 | 82.9 | 82.9 | 16.9 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 225 | 225 | 67.0 | 0.0 | 100.0 |
| Execution Mode | agentic | 802 | 802 | 799 | 797 | 61.4 | 34.8 | 64.8 |
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | universal_agent_adapter | 802 | 802 | 799 | 797 | 61.4 | 34.8 | 64.8 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | codex | 802 | 802 | 799 | 797 | 61.4 | 34.8 | 64.8 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 24 | 24 | 79.2 | 75.0 | 25.0 |
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 11 | 11 | 90.9 | 90.9 | 9.1 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 27 | 27 | 51.9 | 51.9 | 48.1 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 5 | 5 | 20.0 | 20.0 | 80.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 22 | 22 | 72.8 | 4.5 | 95.5 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 49 | 50 | 31.0 | 22.0 | 78.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 10 | 10 | 80.0 | 70.0 | 30.0 |
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 1 | 1 | 100.0 | 100.0 | 0.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 13 | 13 | 69.2 | 69.2 | 30.8 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 8 | 8 | 87.5 | 87.5 | 12.5 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 2 | 2 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 16 | 16 | 84.4 | 81.2 | 18.8 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 16 | 16 | 43.8 | 43.8 | 56.2 |
| L-Level × Type | L3 × LOC | 29 | 29 | 28 | 29 | 46.4 | 10.3 | 89.7 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 12 | 12 | 20.8 | 8.3 | 91.7 |
| L-Level × Type | N/A × LOC | 10 | 10 | 10 | 10 | 79.6 | 10.0 | 90.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 95 | 93 | 36.0 | 23.7 | 74.7 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 10 | 10 | 40.0 | 40.0 | 60.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 86 | 84 | 58.9 | 47.6 | 51.2 |
| L-Level × Type | N/A × book_true_false | 130 | 130 | 129 | 129 | 82.9 | 82.9 | 16.9 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 225 | 225 | 67.0 | 0.0 | 100.0 |

Table 10: Detailed per-dimension breakdown for deepseek-v4-pro + codex.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 321 | 316 | 62.5 | 56.0 | 43.3 |
| Dataset | A | 246 | 246 | 241 | 246 | 53.9 | 39.0 | 61.0 |
| Dataset | M | 225 | 225 | 225 | 225 | 70.6 | 5.8 | 94.2 |
| Dataset | S | 10 | 10 | 10 | 10 | 73.8 | 10.0 | 90.0 |
| L-Level | L2 | 139 | 139 | 138 | 139 | 53.9 | 40.3 | 59.7 |
| L-Level | L3 | 107 | 107 | 103 | 107 | 53.9 | 37.4 | 62.6 |
| L-Level | N/A | 570 | 556 | 556 | 551 | 66.0 | 34.7 | 64.7 |
| Suite | BOOK | 335 | 321 | 321 | 316 | 62.5 | 56.0 | 43.3 |
| Suite | L2 | 139 | 139 | 138 | 139 | 53.9 | 40.3 | 59.7 |
| Suite | L3 | 107 | 107 | 103 | 107 | 53.9 | 37.4 | 62.6 |
| Suite | MODULE-EXT | 225 | 225 | 225 | 225 | 70.6 | 5.8 | 94.2 |
| Suite | SOC-EXT | 10 | 10 | 10 | 10 | 73.8 | 10.0 | 90.0 |
| Design Level | ip | 126 | 126 | 126 | 126 | 52.0 | 35.7 | 64.3 |
| Design Level | knowledge | 335 | 321 | 321 | 316 | 62.5 | 56.0 | 43.3 |
| Design Level | module | 289 | 289 | 288 | 289 | 67.5 | 13.8 | 86.2 |
| Design Level | soc | 66 | 66 | 62 | 66 | 58.2 | 37.9 | 62.1 |
| Task Category | attribution | 34 | 34 | 34 | 34 | 70.6 | 67.6 | 32.4 |
| Task Category | cwe_classification | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Task Category | detection | 60 | 60 | 60 | 60 | 68.3 | 68.3 | 31.7 |
| Task Category | knowledge | 335 | 321 | 321 | 316 | 62.5 | 56.0 | 43.3 |
| Task Category | localization | 51 | 51 | 49 | 51 | 56.2 | 3.9 | 96.1 |
| Task Category | localization_cwe | 235 | 235 | 235 | 235 | 70.7 | 6.0 | 94.0 |
| Task Category | remediation | 32 | 32 | 29 | 32 | 51.6 | 46.9 | 53.1 |
| Task Category | characterization | 62 | 62 | 62 | 62 | 36.3 | 24.2 | 75.8 |
| Question Type | ATTR | 34 | 34 | 34 | 34 | 70.6 | 67.6 | 32.4 |
| Question Type | DET_ACT | 12 | 12 | 12 | 12 | 91.7 | 91.7 | 8.3 |
| Question Type | DET_BEH | 40 | 40 | 40 | 40 | 57.5 | 57.5 | 42.5 |

*Table 10 (continued).*

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Question Type | DET_VUL | 8 | 8 | 8 | 8 | 87.5 | 87.5 | 12.5 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Question Type | FIX | 16 | 16 | 14 | 16 | 46.9 | 43.8 | 56.2 |
| Question Type | FIX_CWE | 16 | 16 | 15 | 16 | 56.2 | 50.0 | 50.0 |
| Question Type | LOC | 61 | 61 | 59 | 61 | 59.1 | 4.9 | 95.1 |
| Question Type | TRUSTHUB | 62 | 62 | 62 | 62 | 36.3 | 24.2 | 75.8 |
| Question Type | book_long_answer | 105 | 95 | 95 | 93 | 31.2 | 19.4 | 78.9 |
| Question Type | book_mathematical | 12 | 10 | 10 | 10 | 50.0 | 50.0 | 50.0 |
| Question Type | book_short_answer | 88 | 86 | 86 | 84 | 57.7 | 46.4 | 52.3 |
| Question Type | book_true_false | 130 | 130 | 130 | 129 | 89.1 | 89.1 | 10.8 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 225 | 225 | 70.6 | 5.8 | 94.2 |
| Execution Mode | agentic | 802 | 802 | 797 | 797 | 62.3 | 36.0 | 63.6 |
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | universal_agent_adapter | 802 | 802 | 797 | 797 | 62.3 | 36.0 | 63.6 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | codex | 802 | 802 | 797 | 797 | 62.3 | 36.0 | 63.6 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 24 | 24 | 70.8 | 70.8 | 29.2 |
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 11 | 11 | 90.9 | 90.9 | 9.1 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 27 | 27 | 55.6 | 55.6 | 44.4 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 5 | 5 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 21 | 22 | 68.0 | 0.0 | 100.0 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 50 | 50 | 36.0 | 28.0 | 72.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 10 | 10 | 70.0 | 60.0 | 40.0 |
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 1 | 1 | 100.0 | 100.0 | 0.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 13 | 13 | 61.5 | 61.5 | 38.5 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 8 | 8 | 87.5 | 87.5 | 12.5 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 2 | 2 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 14 | 16 | 46.9 | 43.8 | 56.2 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 15 | 16 | 56.2 | 50.0 | 50.0 |
| L-Level × Type | L3 × LOC | 29 | 29 | 28 | 29 | 47.3 | 6.9 | 93.1 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 12 | 12 | 37.5 | 8.3 | 91.7 |
| L-Level × Type | N/A × LOC | 10 | 10 | 10 | 10 | 73.8 | 10.0 | 90.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 95 | 93 | 31.2 | 19.4 | 78.9 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 10 | 10 | 50.0 | 50.0 | 50.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 86 | 84 | 57.7 | 46.4 | 52.3 |
| L-Level × Type | N/A × book_true_false | 130 | 130 | 130 | 129 | 89.1 | 89.1 | 10.8 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 225 | 225 | 70.6 | 5.8 | 94.2 |

Table 11: Detailed per-dimension breakdown for glm-5.3 + codex.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 306 | 283 | 51.6 | 45.6 | 48.0 |
| Dataset | A | 246 | 246 | 240 | 246 | 32.0 | 20.7 | 79.3 |
| Dataset | M | 225 | 225 | 225 | 225 | 61.1 | 0.9 | 99.1 |
| Dataset | S | 10 | 10 | 10 | 10 | 56.8 | 0.0 | 100.0 |
| L-Level | L2 | 139 | 139 | 139 | 139 | 39.2 | 26.6 | 73.4 |
| L-Level | L3 | 107 | 107 | 101 | 107 | 22.8 | 13.1 | 86.9 |
| L-Level | N/A | 570 | 556 | 541 | 518 | 55.8 | 25.3 | 69.6 |
| Suite | BOOK | 335 | 321 | 306 | 283 | 51.6 | 45.6 | 48.0 |
| Suite | L2 | 139 | 139 | 139 | 139 | 39.2 | 26.6 | 73.4 |
| Suite | L3 | 107 | 107 | 101 | 107 | 22.8 | 13.1 | 86.9 |
| Suite | MODULE-EXT | 225 | 225 | 225 | 225 | 61.1 | 0.9 | 99.1 |
| Suite | SOC-EXT | 10 | 10 | 10 | 10 | 56.8 | 0.0 | 100.0 |
| Design Level | ip | 126 | 126 | 124 | 126 | 31.9 | 19.8 | 80.2 |
| Design Level | knowledge | 335 | 321 | 306 | 283 | 51.6 | 45.6 | 48.0 |
| Design Level | module | 289 | 289 | 289 | 289 | 57.6 | 7.3 | 92.7 |
| Design Level | soc | 66 | 66 | 62 | 66 | 23.3 | 10.6 | 89.4 |
| Task Category | attribution | 34 | 34 | 34 | 34 | 55.9 | 50.0 | 50.0 |
| Task Category | cwe_classification | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Task Category | detection | 60 | 60 | 58 | 60 | 51.7 | 51.7 | 48.3 |
| Task Category | knowledge | 335 | 321 | 306 | 283 | 51.6 | 45.6 | 48.0 |
| Task Category | localization | 51 | 51 | 48 | 51 | 36.9 | 0.0 | 100.0 |

*Table 11 (continued).*

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Task Category | localization_cwe | 235 | 235 | 235 | 235 | 60.9 | 0.9 | 99.1 |
| Task Category | remediation | 32 | 32 | 31 | 32 | 7.8 | 6.2 | 93.8 |
| Task Category | characterization | 62 | 62 | 62 | 62 | 12.1 | 1.6 | 98.4 |
| Question Type | ATTR | 34 | 34 | 34 | 34 | 55.9 | 50.0 | 50.0 |
| Question Type | DET_ACT | 12 | 12 | 11 | 12 | 75.0 | 75.0 | 25.0 |
| Question Type | DET_BEH | 40 | 40 | 39 | 40 | 42.5 | 42.5 | 57.5 |
| Question Type | DET_VUL | 8 | 8 | 8 | 8 | 62.5 | 62.5 | 37.5 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Question Type | FIX | 16 | 16 | 16 | 16 | 15.6 | 12.5 | 87.5 |
| Question Type | FIX_CWE | 16 | 16 | 15 | 16 | 0.0 | 0.0 | 100.0 |
| Question Type | LOC | 61 | 61 | 58 | 61 | 40.2 | 0.0 | 100.0 |
| Question Type | TRUSTHUB | 62 | 62 | 62 | 62 | 12.1 | 1.6 | 98.4 |
| Question Type | book_long_answer | 105 | 95 | 89 | 81 | 25.9 | 13.6 | 73.7 |
| Question Type | book_mathematical | 12 | 10 | 9 | 6 | 50.0 | 50.0 | 30.0 |
| Question Type | book_short_answer | 88 | 86 | 82 | 69 | 49.3 | 39.1 | 48.8 |
| Question Type | book_true_false | 130 | 130 | 126 | 127 | 69.3 | 69.3 | 30.0 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 225 | 225 | 61.1 | 0.9 | 99.1 |
| Execution Mode | agentic | 802 | 802 | 781 | 764 | 48.2 | 23.8 | 72.6 |
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | universal_agent_adapter | 802 | 802 | 781 | 764 | 48.2 | 23.8 | 72.6 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | codex | 802 | 802 | 781 | 764 | 48.2 | 23.8 | 72.6 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 24 | 24 | 64.6 | 62.5 | 37.5 |
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 11 | 11 | 81.8 | 81.8 | 18.2 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 27 | 27 | 44.4 | 44.4 | 55.6 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 5 | 5 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 22 | 22 | 52.2 | 0.0 | 100.0 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 50 | 50 | 13.0 | 2.0 | 98.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 10 | 10 | 35.0 | 20.0 | 80.0 |
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 0 | 1 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 12 | 13 | 38.5 | 38.5 | 61.5 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 8 | 8 | 62.5 | 62.5 | 37.5 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 2 | 2 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 16 | 16 | 15.6 | 12.5 | 87.5 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 15 | 16 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × LOC | 29 | 29 | 26 | 29 | 25.4 | 0.0 | 100.0 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 12 | 12 | 8.3 | 0.0 | 100.0 |
| L-Level × Type | N/A × LOC | 10 | 10 | 10 | 10 | 56.8 | 0.0 | 100.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 89 | 81 | 25.9 | 13.6 | 73.7 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 9 | 6 | 50.0 | 50.0 | 30.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 82 | 69 | 49.3 | 39.1 | 48.8 |
| L-Level × Type | N/A × book_true_false | 130 | 130 | 126 | 127 | 69.3 | 69.3 | 30.0 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 225 | 225 | 61.1 | 0.9 | 99.1 |

Table 12: Detailed per-dimension breakdown for glm-x-preview-260804 + codex.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 317 | 313 | 57.7 | 50.2 | 49.5 |
| Dataset | A | 246 | 246 | 242 | 246 | 43.9 | 31.7 | 68.3 |
| Dataset | M | 225 | 225 | 223 | 225 | 48.9 | 0.0 | 100.0 |
| Dataset | S | 10 | 10 | 10 | 10 | 67.9 | 20.0 | 80.0 |
| L-Level | L2 | 139 | 139 | 138 | 139 | 43.9 | 31.7 | 68.3 |
| L-Level | L3 | 107 | 107 | 104 | 107 | 43.8 | 31.8 | 68.2 |
| L-Level | N/A | 570 | 556 | 550 | 548 | 54.3 | 29.0 | 70.5 |
| Suite | BOOK | 335 | 321 | 317 | 313 | 57.7 | 50.2 | 49.5 |
| Suite | L2 | 139 | 139 | 138 | 139 | 43.9 | 31.7 | 68.3 |
| Suite | L3 | 107 | 107 | 104 | 107 | 43.8 | 31.8 | 68.2 |
| Suite | MODULE-EXT | 225 | 225 | 223 | 225 | 48.9 | 0.0 | 100.0 |
| Suite | SOC-EXT | 10 | 10 | 10 | 10 | 67.9 | 20.0 | 80.0 |
| Design Level | ip | 126 | 126 | 125 | 126 | 38.1 | 26.2 | 73.8 |
| Design Level | knowledge | 335 | 321 | 317 | 313 | 57.7 | 50.2 | 49.5 |
| Design Level | module | 289 | 289 | 287 | 289 | 50.7 | 9.3 | 90.7 |

*Table 12 (continued).*

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Design Level | soc | 66 | 66 | 63 | 66 | 45.8 | 30.3 | 69.7 |
| Task Category | attribution | 34 | 34 | 34 | 34 | 70.6 | 67.6 | 32.4 |
| Task Category | cwe_classification | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Task Category | detection | 60 | 60 | 60 | 60 | 65.0 | 65.0 | 35.0 |
| Task Category | knowledge | 335 | 321 | 317 | 313 | 57.7 | 50.2 | 49.5 |
| Task Category | localization | 51 | 51 | 51 | 51 | 51.7 | 5.9 | 94.1 |
| Task Category | localization_cwe | 235 | 235 | 233 | 235 | 49.7 | 0.9 | 99.1 |
| Task Category | remediation | 32 | 32 | 29 | 32 | 40.6 | 37.5 | 62.5 |
| Task Category | characterization | 62 | 62 | 61 | 62 | 8.9 | 1.6 | 98.4 |
| Question Type | ATTR | 34 | 34 | 34 | 34 | 70.6 | 67.6 | 32.4 |
| Question Type | DET_ACT | 12 | 12 | 12 | 12 | 91.7 | 91.7 | 8.3 |
| Question Type | DET_BEH | 40 | 40 | 40 | 40 | 57.5 | 57.5 | 42.5 |
| Question Type | DET_VUL | 8 | 8 | 8 | 8 | 62.5 | 62.5 | 37.5 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Question Type | FIX | 16 | 16 | 13 | 16 | 62.5 | 62.5 | 37.5 |
| Question Type | FIX_CWE | 16 | 16 | 16 | 16 | 18.8 | 12.5 | 87.5 |
| Question Type | LOC | 61 | 61 | 61 | 61 | 54.4 | 8.2 | 91.8 |
| Question Type | TRUSTHUB | 62 | 62 | 61 | 62 | 8.9 | 1.6 | 98.4 |
| Question Type | book_long_answer | 105 | 95 | 94 | 92 | 23.9 | 8.7 | 89.5 |
| Question Type | book_mathematical | 12 | 10 | 7 | 8 | 31.2 | 25.0 | 80.0 |
| Question Type | book_short_answer | 88 | 86 | 86 | 84 | 56.0 | 45.2 | 53.5 |
| Question Type | book_true_false | 130 | 130 | 130 | 129 | 84.5 | 84.5 | 15.4 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 223 | 225 | 48.9 | 0.0 | 100.0 |
| Execution Mode | agentic | 802 | 802 | 792 | 794 | 51.0 | 29.8 | 69.8 |
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | universal_agent_adapter | 802 | 802 | 792 | 794 | 51.0 | 29.8 | 69.8 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | codex | 802 | 802 | 792 | 794 | 51.0 | 29.8 | 69.8 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 24 | 24 | 72.9 | 70.8 | 29.2 |
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 11 | 11 | 90.9 | 90.9 | 9.1 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 27 | 27 | 51.9 | 51.9 | 48.1 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 5 | 5 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 22 | 22 | 70.4 | 9.1 | 90.9 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 49 | 50 | 8.0 | 2.0 | 98.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 10 | 10 | 65.0 | 60.0 | 40.0 |
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 1 | 1 | 100.0 | 100.0 | 0.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 13 | 13 | 69.2 | 69.2 | 30.8 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 8 | 8 | 62.5 | 62.5 | 37.5 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 2 | 2 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 13 | 16 | 62.5 | 62.5 | 37.5 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 16 | 16 | 18.8 | 12.5 | 87.5 |
| L-Level × Type | L3 × LOC | 29 | 29 | 29 | 29 | 37.5 | 3.4 | 96.6 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 12 | 12 | 12.5 | 0.0 | 100.0 |
| L-Level × Type | N/A × LOC | 10 | 10 | 10 | 10 | 67.9 | 20.0 | 80.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 94 | 92 | 23.9 | 8.7 | 89.5 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 7 | 8 | 31.2 | 25.0 | 80.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 86 | 84 | 56.0 | 45.2 | 53.5 |
| L-Level × Type | N/A × book_true_false | 130 | 130 | 130 | 129 | 84.5 | 84.5 | 15.4 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 223 | 225 | 48.9 | 0.0 | 100.0 |

Table 13: Detailed per-dimension breakdown for gpt-5.6-sol + codex.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 272 | 267 | 56.6 | 49.8 | 57.0 |
| Dataset | A | 246 | 246 | 243 | 243 | 46.6 | 32.9 | 67.5 |
| Dataset | M | 225 | 225 | 222 | 222 | 68.0 | 0.0 | 100.0 |
| Dataset | S | 10 | 10 | 10 | 10 | 74.4 | 20.0 | 80.0 |
| L-Level | L2 | 139 | 139 | 138 | 138 | 46.7 | 33.3 | 66.9 |
| L-Level | L3 | 107 | 107 | 105 | 105 | 46.6 | 32.4 | 68.2 |
| L-Level | N/A | 570 | 556 | 504 | 499 | 62.0 | 27.1 | 74.8 |
| Suite | BOOK | 335 | 321 | 272 | 267 | 56.6 | 49.8 | 57.0 |
| Suite | L2 | 139 | 139 | 138 | 138 | 46.7 | 33.3 | 66.9 |

*Table 13 (continued).*

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Suite | L3 | 107 | 107 | 105 | 105 | 46.6 | 32.4 | 68.2 |
| Suite | MODULE-EXT | 225 | 225 | 222 | 222 | 68.0 | 0.0 | 100.0 |
| Suite | SOC-EXT | 10 | 10 | 10 | 10 | 74.4 | 20.0 | 80.0 |
| Design Level | ip | 126 | 126 | 123 | 123 | 41.2 | 28.5 | 72.2 |
| Design Level | knowledge | 335 | 321 | 272 | 267 | 56.6 | 49.8 | 57.0 |
| Design Level | module | 289 | 289 | 286 | 286 | 66.8 | 10.5 | 89.6 |
| Design Level | soc | 66 | 66 | 66 | 66 | 45.2 | 25.8 | 74.2 |
| Task Category | attribution | 34 | 34 | 34 | 34 | 79.4 | 76.5 | 23.5 |
| Task Category | cwe_classification | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Task Category | detection | 60 | 60 | 58 | 58 | 72.4 | 72.4 | 30.0 |
| Task Category | knowledge | 335 | 321 | 272 | 267 | 56.6 | 49.8 | 57.0 |
| Task Category | localization | 51 | 51 | 51 | 51 | 54.6 | 2.0 | 98.0 |
| Task Category | localization_cwe | 235 | 235 | 232 | 232 | 68.2 | 0.9 | 99.1 |
| Task Category | remediation | 32 | 32 | 32 | 32 | 35.9 | 31.2 | 68.8 |
| Task Category | characterization | 62 | 62 | 61 | 61 | 8.2 | 1.6 | 98.4 |
| Question Type | ATTR | 34 | 34 | 34 | 34 | 79.4 | 76.5 | 23.5 |
| Question Type | DET_ACT | 12 | 12 | 11 | 11 | 100.0 | 100.0 | 8.3 |
| Question Type | DET_BEH | 40 | 40 | 39 | 39 | 66.7 | 66.7 | 35.0 |
| Question Type | DET_VUL | 8 | 8 | 8 | 8 | 62.5 | 62.5 | 37.5 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Question Type | FIX | 16 | 16 | 16 | 16 | 46.9 | 43.8 | 56.2 |
| Question Type | FIX_CWE | 16 | 16 | 16 | 16 | 25.0 | 18.8 | 81.2 |
| Question Type | LOC | 61 | 61 | 61 | 61 | 57.8 | 4.9 | 95.1 |
| Question Type | TRUSTHUB | 62 | 62 | 61 | 61 | 8.2 | 1.6 | 98.4 |
| Question Type | book_long_answer | 105 | 95 | 80 | 78 | 26.3 | 12.8 | 87.4 |
| Question Type | book_mathematical | 12 | 10 | 10 | 10 | 30.0 | 30.0 | 70.0 |
| Question Type | book_short_answer | 88 | 86 | 72 | 70 | 50.7 | 40.0 | 65.1 |
| Question Type | book_true_false | 130 | 130 | 110 | 109 | 84.4 | 84.4 | 28.5 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 222 | 222 | 68.0 | 0.0 | 100.0 |
| Execution Mode | agentic | 802 | 802 | 747 | 742 | 57.0 | 29.0 | 72.6 |
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | universal_agent_adapter | 802 | 802 | 747 | 742 | 57.0 | 29.0 | 72.6 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | codex | 802 | 802 | 747 | 742 | 57.0 | 29.0 | 72.6 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 24 | 24 | 77.1 | 75.0 | 25.0 |
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 11 | 11 | 100.0 | 100.0 | 0.0 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 27 | 27 | 55.6 | 55.6 | 44.4 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 5 | 5 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 22 | 22 | 70.1 | 4.5 | 95.5 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 49 | 49 | 9.2 | 2.0 | 98.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 10 | 10 | 85.0 | 80.0 | 20.0 |
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 0 | 0 | — | — | 100.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 12 | 12 | 91.7 | 91.7 | 15.4 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 8 | 8 | 62.5 | 62.5 | 37.5 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 2 | 2 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 16 | 16 | 46.9 | 43.8 | 56.2 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 16 | 16 | 25.0 | 18.8 | 81.2 |
| L-Level × Type | L3 × LOC | 29 | 29 | 29 | 29 | 42.8 | 0.0 | 100.0 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 12 | 12 | 4.2 | 0.0 | 100.0 |
| L-Level × Type | N/A × LOC | 10 | 10 | 10 | 10 | 74.4 | 20.0 | 80.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 80 | 78 | 26.3 | 12.8 | 87.4 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 10 | 10 | 30.0 | 30.0 | 70.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 72 | 70 | 50.7 | 40.0 | 65.1 |
| L-Level × Type | N/A × book_true_false | 130 | 130 | 110 | 109 | 84.4 | 84.4 | 28.5 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 222 | 222 | 68.0 | 0.0 | 100.0 |

Table 14: Detailed per-dimension breakdown for qwen3.8-flash + codex.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 313 | 308 | 68.2 | 62.3 | 38.6 |
| Dataset | A | 246 | 246 | 233 | 246 | 41.7 | 29.7 | 70.3 |
| Dataset | M | 225 | 225 | 225 | 225 | 67.7 | 0.0 | 100.0 |

*Table 14 (continued).*

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | S | 10 | 10 | 9 | 10 | 43.1 | 0.0 | 100.0 |
| L-Level | L2 | 139 | 139 | 136 | 139 | 41.1 | 28.8 | 71.2 |
| L-Level | L3 | 107 | 107 | 97 | 107 | 42.4 | 30.8 | 69.2 |
| L-Level | N/A | 570 | 556 | 547 | 543 | 67.5 | 35.4 | 64.6 |
| Suite | BOOK | 335 | 321 | 313 | 308 | 68.2 | 62.3 | 38.6 |
| Suite | L2 | 139 | 139 | 136 | 139 | 41.1 | 28.8 | 71.2 |
| Suite | L3 | 107 | 107 | 97 | 107 | 42.4 | 30.8 | 69.2 |
| Suite | MODULE-EXT | 225 | 225 | 225 | 225 | 67.7 | 0.0 | 100.0 |
| Suite | SOC-EXT | 10 | 10 | 9 | 10 | 43.1 | 0.0 | 100.0 |
| Design Level | ip | 126 | 126 | 122 | 126 | 34.3 | 23.0 | 77.0 |
| Design Level | knowledge | 335 | 321 | 313 | 308 | 68.2 | 62.3 | 38.6 |
| Design Level | module | 289 | 289 | 289 | 289 | 66.5 | 10.0 | 90.0 |
| Design Level | soc | 66 | 66 | 56 | 66 | 35.9 | 22.7 | 77.3 |
| Task Category | attribution | 34 | 34 | 34 | 34 | 72.1 | 67.6 | 32.4 |
| Task Category | cwe_classification | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Task Category | detection | 60 | 60 | 58 | 60 | 65.0 | 65.0 | 35.0 |
| Task Category | knowledge | 335 | 321 | 313 | 308 | 68.2 | 62.3 | 38.6 |
| Task Category | localization | 51 | 51 | 46 | 51 | 45.1 | 0.0 | 100.0 |
| Task Category | localization_cwe | 235 | 235 | 234 | 235 | 66.7 | 0.0 | 100.0 |
| Task Category | remediation | 32 | 32 | 28 | 32 | 29.7 | 28.1 | 71.9 |
| Task Category | characterization | 62 | 62 | 60 | 62 | 10.5 | 3.2 | 96.8 |
| Question Type | ATTR | 34 | 34 | 34 | 34 | 72.1 | 67.6 | 32.4 |
| Question Type | DET_ACT | 12 | 12 | 10 | 12 | 58.3 | 58.3 | 41.7 |
| Question Type | DET_BEH | 40 | 40 | 40 | 40 | 65.0 | 65.0 | 35.0 |
| Question Type | DET_VUL | 8 | 8 | 8 | 8 | 75.0 | 75.0 | 25.0 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 7 | 7 | 0.0 | 0.0 | 100.0 |
| Question Type | FIX | 16 | 16 | 13 | 16 | 25.0 | 25.0 | 75.0 |
| Question Type | FIX_CWE | 16 | 16 | 15 | 16 | 34.4 | 31.2 | 68.8 |
| Question Type | LOC | 61 | 61 | 55 | 61 | 44.8 | 0.0 | 100.0 |
| Question Type | TRUSTHUB | 62 | 62 | 60 | 62 | 10.5 | 3.2 | 96.8 |
| Question Type | book_long_answer | 105 | 95 | 92 | 90 | 44.4 | 31.1 | 68.4 |
| Question Type | book_mathematical | 12 | 10 | 9 | 9 | 44.4 | 44.4 | 60.0 |
| Question Type | book_short_answer | 88 | 86 | 82 | 80 | 63.8 | 56.2 | 45.3 |
| Question Type | book_true_false | 130 | 130 | 130 | 129 | 89.1 | 89.1 | 10.8 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 225 | 225 | 67.7 | 0.0 | 100.0 |
| Execution Mode | agentic | 802 | 802 | 780 | 789 | 59.5 | 33.6 | 66.3 |
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | universal_agent_adapter | 802 | 802 | 780 | 789 | 59.5 | 33.6 | 66.3 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | codex | 802 | 802 | 780 | 789 | 59.5 | 33.6 | 66.3 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 24 | 24 | 72.9 | 70.8 | 29.2 |
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 10 | 11 | 63.6 | 63.6 | 36.4 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 27 | 27 | 55.6 | 55.6 | 44.4 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 5 | 5 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 22 | 22 | 57.6 | 0.0 | 100.0 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 48 | 50 | 10.0 | 2.0 | 98.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 10 | 10 | 70.0 | 60.0 | 40.0 |
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 0 | 1 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 13 | 13 | 84.6 | 84.6 | 15.4 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 8 | 8 | 75.0 | 75.0 | 25.0 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 2 | 2 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 13 | 16 | 25.0 | 25.0 | 75.0 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 15 | 16 | 34.4 | 31.2 | 68.8 |
| L-Level × Type | L3 × LOC | 29 | 29 | 24 | 29 | 35.7 | 0.0 | 100.0 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 12 | 12 | 12.5 | 8.3 | 91.7 |
| L-Level × Type | N/A × LOC | 10 | 10 | 9 | 10 | 43.1 | 0.0 | 100.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 92 | 90 | 44.4 | 31.1 | 68.4 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 9 | 9 | 44.4 | 44.4 | 60.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 82 | 80 | 63.8 | 56.2 | 45.3 |
| L-Level × Type | N/A × book_true_false | 130 | 130 | 130 | 129 | 89.1 | 89.1 | 10.8 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 225 | 225 | 67.7 | 0.0 | 100.0 |

Table 15: Detailed per-dimension breakdown for glm-5.3 + deepseek_harness.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 302 | 310 | 61.6 | 52.9 | 45.5 |
| Dataset | A | 246 | 246 | 175 | 246 | 37.8 | 28.9 | 71.1 |
| Dataset | M | 225 | 225 | 144 | 202 | 40.6 | 0.0 | 89.8 |
| Dataset | S | 10 | 10 | 6 | 10 | 50.8 | 0.0 | 100.0 |
| L-Level | L2 | 139 | 139 | 103 | 139 | 39.0 | 30.9 | 69.1 |
| L-Level | L3 | 107 | 107 | 72 | 107 | 36.1 | 26.2 | 73.8 |
| L-Level | N/A | 570 | 556 | 452 | 522 | 53.3 | 31.4 | 64.4 |
| Suite | BOOK | 335 | 321 | 302 | 310 | 61.6 | 52.9 | 45.5 |
| Suite | L2 | 139 | 139 | 103 | 139 | 39.0 | 30.9 | 69.1 |
| Suite | L3 | 107 | 107 | 72 | 107 | 36.1 | 26.2 | 73.8 |
| Suite | MODULE-EXT | 225 | 225 | 144 | 202 | 40.6 | 0.0 | 89.8 |
| Suite | SOC-EXT | 10 | 10 | 6 | 10 | 50.8 | 0.0 | 100.0 |
| Design Level | ip | 126 | 126 | 89 | 126 | 32.8 | 26.2 | 73.8 |
| Design Level | knowledge | 335 | 321 | 302 | 310 | 61.6 | 52.9 | 45.5 |
| Design Level | module | 289 | 289 | 198 | 266 | 44.2 | 9.8 | 83.0 |
| Design Level | soc | 66 | 66 | 38 | 66 | 32.2 | 18.2 | 81.8 |
| Task Category | attribution | 34 | 34 | 32 | 34 | 77.9 | 73.5 | 26.5 |
| Task Category | cwe_classification | 7 | 7 | 4 | 7 | 0.0 | 0.0 | 100.0 |
| Task Category | detection | 60 | 60 | 47 | 60 | 56.7 | 56.7 | 43.3 |
| Task Category | knowledge | 335 | 321 | 302 | 310 | 61.6 | 52.9 | 45.5 |
| Task Category | localization | 51 | 51 | 30 | 51 | 32.1 | 2.0 | 98.0 |
| Task Category | localization_cwe | 235 | 235 | 150 | 212 | 41.1 | 0.0 | 90.2 |
| Task Category | remediation | 32 | 32 | 19 | 32 | 31.2 | 28.1 | 71.9 |
| Task Category | characterization | 62 | 62 | 43 | 62 | 9.7 | 3.2 | 96.8 |
| Question Type | ATTR | 34 | 34 | 32 | 34 | 77.9 | 73.5 | 26.5 |
| Question Type | DET_ACT | 12 | 12 | 10 | 12 | 75.0 | 75.0 | 25.0 |
| Question Type | DET_BEH | 40 | 40 | 32 | 40 | 55.0 | 55.0 | 45.0 |
| Question Type | DET_VUL | 8 | 8 | 5 | 8 | 37.5 | 37.5 | 62.5 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 4 | 7 | 0.0 | 0.0 | 100.0 |
| Question Type | FIX | 16 | 16 | 12 | 16 | 40.6 | 37.5 | 62.5 |
| Question Type | FIX_CWE | 16 | 16 | 7 | 16 | 21.9 | 18.8 | 81.2 |
| Question Type | LOC | 61 | 61 | 36 | 61 | 35.2 | 1.6 | 98.4 |
| Question Type | TRUSTHUB | 62 | 62 | 43 | 62 | 9.7 | 3.2 | 96.8 |
| Question Type | book_long_answer | 105 | 95 | 93 | 93 | 45.7 | 29.0 | 69.5 |
| Question Type | book_mathematical | 12 | 10 | 6 | 10 | 30.0 | 30.0 | 70.0 |
| Question Type | book_short_answer | 88 | 86 | 82 | 84 | 56.5 | 42.9 | 55.8 |
| Question Type | book_true_false | 130 | 130 | 121 | 123 | 79.7 | 79.7 | 19.2 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 144 | 202 | 40.6 | 0.0 | 89.8 |
| Execution Mode | agentic | 802 | 802 | 627 | 768 | 48.3 | 30.6 | 66.5 |
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | deepseek_harness_docker_agentic_adapter | 802 | 802 | 627 | 768 | 48.3 | 30.6 | 66.5 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | deepseek_harness | 802 | 802 | 627 | 768 | 48.3 | 30.6 | 66.5 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 22 | 24 | 77.1 | 75.0 | 25.0 |
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 9 | 11 | 72.7 | 72.7 | 27.3 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 21 | 27 | 51.9 | 51.9 | 48.1 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 3 | 5 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 14 | 22 | 42.1 | 4.5 | 95.5 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 34 | 50 | 9.0 | 4.0 | 96.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 10 | 10 | 80.0 | 70.0 | 30.0 |
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 1 | 1 | 100.0 | 100.0 | 0.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 11 | 13 | 61.5 | 61.5 | 38.5 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 5 | 8 | 37.5 | 37.5 | 62.5 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 1 | 2 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 12 | 16 | 40.6 | 37.5 | 62.5 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 7 | 16 | 21.9 | 18.8 | 81.2 |
| L-Level × Type | L3 × LOC | 29 | 29 | 16 | 29 | 24.6 | 0.0 | 100.0 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 9 | 12 | 12.5 | 0.0 | 100.0 |
| L-Level × Type | N/A × LOC | 10 | 10 | 6 | 10 | 50.8 | 0.0 | 100.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 93 | 93 | 45.7 | 29.0 | 69.5 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 6 | 10 | 30.0 | 30.0 | 70.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 82 | 84 | 56.5 | 42.9 | 55.8 |
| L-Level × Type | N/A × book_true_false | 130 | 130 | 121 | 123 | 79.7 | 79.7 | 19.2 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 144 | 202 | 40.6 | 0.0 | 89.8 |

Table 16: Detailed per-dimension breakdown for glm-5.3 + opencode.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 315 | 316 | 61.1 | 54.7 | 44.5 |
| Dataset | A | 246 | 246 | 188 | 224 | 43.9 | 33.5 | 69.5 |
| Dataset | M | 225 | 225 | 151 | 165 | 61.7 | 0.0 | 100.0 |
| Dataset | S | 10 | 10 | 1 | 1 | 81.0 | 0.0 | 100.0 |
| L-Level | L2 | 139 | 139 | 112 | 127 | 42.9 | 32.3 | 70.5 |
| L-Level | L3 | 107 | 107 | 76 | 97 | 45.2 | 35.1 | 68.2 |
| L-Level | N/A | 570 | 556 | 467 | 482 | 61.3 | 35.9 | 68.0 |
| Suite | BOOK | 335 | 321 | 315 | 316 | 61.1 | 54.7 | 44.5 |
| Suite | L2 | 139 | 139 | 112 | 127 | 42.9 | 32.3 | 70.5 |
| Suite | L3 | 107 | 107 | 76 | 97 | 45.2 | 35.1 | 68.2 |
| Suite | MODULE-EXT | 225 | 225 | 151 | 165 | 61.7 | 0.0 | 100.0 |
| Suite | SOC-EXT | 10 | 10 | 1 | 1 | 81.0 | 0.0 | 100.0 |
| Design Level | ip | 126 | 126 | 97 | 111 | 37.6 | 28.8 | 74.6 |
| Design Level | knowledge | 335 | 321 | 315 | 316 | 61.1 | 54.7 | 44.5 |
| Design Level | module | 289 | 289 | 210 | 227 | 60.6 | 11.0 | 91.3 |
| Design Level | soc | 66 | 66 | 33 | 52 | 41.5 | 34.6 | 72.7 |
| Task Category | attribution | 34 | 34 | 32 | 33 | 72.7 | 66.7 | 35.3 |
| Task Category | cwe_classification | 7 | 7 | 5 | 7 | 14.3 | 14.3 | 85.7 |
| Task Category | detection | 60 | 60 | 52 | 57 | 64.9 | 64.9 | 38.3 |
| Task Category | knowledge | 335 | 321 | 315 | 316 | 61.1 | 54.7 | 44.5 |
| Task Category | localization | 51 | 51 | 28 | 37 | 39.9 | 2.7 | 98.0 |
| Task Category | localization_cwe | 235 | 235 | 152 | 166 | 61.8 | 0.0 | 100.0 |
| Task Category | remediation | 32 | 32 | 21 | 31 | 43.5 | 41.9 | 59.4 |
| Task Category | characterization | 62 | 62 | 50 | 59 | 13.6 | 1.7 | 98.4 |
| Question Type | ATTR | 34 | 34 | 32 | 33 | 72.7 | 66.7 | 35.3 |
| Question Type | DET_ACT | 12 | 12 | 10 | 11 | 81.8 | 81.8 | 25.0 |
| Question Type | DET_BEH | 40 | 40 | 36 | 38 | 60.5 | 60.5 | 42.5 |
| Question Type | DET_VUL | 8 | 8 | 6 | 8 | 62.5 | 62.5 | 37.5 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 5 | 7 | 14.3 | 14.3 | 85.7 |
| Question Type | FIX | 16 | 16 | 10 | 15 | 46.7 | 46.7 | 56.2 |
| Question Type | FIX_CWE | 16 | 16 | 11 | 16 | 40.6 | 37.5 | 62.5 |
| Question Type | LOC | 61 | 61 | 29 | 38 | 41.0 | 2.6 | 98.4 |
| Question Type | TRUSTHUB | 62 | 62 | 50 | 59 | 13.6 | 1.7 | 98.4 |
| Question Type | book_long_answer | 105 | 95 | 95 | 93 | 32.8 | 20.4 | 77.9 |
| Question Type | book_mathematical | 12 | 10 | 4 | 10 | 30.0 | 30.0 | 70.0 |
| Question Type | book_short_answer | 88 | 86 | 86 | 84 | 58.9 | 48.8 | 50.0 |
| Question Type | book_true_false | 130 | 130 | 130 | 129 | 85.3 | 85.3 | 14.6 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 151 | 165 | 61.7 | 0.0 | 100.0 |
| Execution Mode | agentic | 802 | 802 | 655 | 706 | 55.8 | 35.1 | 68.5 |
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | opencode_docker_agentic_adapter | 802 | 802 | 655 | 706 | 55.8 | 35.1 | 68.5 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | opencode | 802 | 802 | 655 | 706 | 55.8 | 35.1 | 68.5 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 23 | 24 | 75.0 | 70.8 | 29.2 |
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 9 | 10 | 80.0 | 80.0 | 27.3 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 24 | 26 | 53.8 | 53.8 | 48.1 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 3 | 5 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 13 | 15 | 52.9 | 6.7 | 95.5 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 40 | 47 | 13.8 | 2.1 | 98.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 9 | 9 | 66.7 | 55.6 | 50.0 |
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 1 | 1 | 100.0 | 100.0 | 0.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 12 | 12 | 75.0 | 75.0 | 30.8 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 6 | 8 | 62.5 | 62.5 | 37.5 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 2 | 2 | 50.0 | 50.0 | 50.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 10 | 15 | 46.7 | 46.7 | 56.2 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 11 | 16 | 40.6 | 37.5 | 62.5 |
| L-Level × Type | L3 × LOC | 29 | 29 | 15 | 22 | 31.1 | 0.0 | 100.0 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 10 | 12 | 12.5 | 0.0 | 100.0 |
| L-Level × Type | N/A × LOC | 10 | 10 | 1 | 1 | 81.0 | 0.0 | 100.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 95 | 93 | 32.8 | 20.4 | 77.9 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 4 | 10 | 30.0 | 30.0 | 70.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 86 | 84 | 58.9 | 48.8 | 50.0 |

*Table 16 (continued).*

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| L-Level × Type | N/A × book_true_false | 130 | 130 | 130 | 129 | 85.3 | 85.3 | 14.6 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 151 | 165 | 61.7 | 0.0 | 100.0 |

Table 17: Detailed per-dimension breakdown for glm-5.3 + trae_agent.

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| Dataset | H | 335 | 321 | 313 | 316 | 48.1 | 43.0 | 56.1 |
| Dataset | A | 246 | 246 | 187 | 246 | 32.1 | 26.0 | 74.0 |
| Dataset | M | 225 | 225 | 190 | 225 | 43.9 | 0.4 | 99.6 |
| Dataset | S | 10 | 10 | 8 | 10 | 35.5 | 10.0 | 90.0 |
| L-Level | L2 | 139 | 139 | 109 | 139 | 30.9 | 25.9 | 74.1 |
| L-Level | L3 | 107 | 107 | 78 | 107 | 33.6 | 26.2 | 73.8 |
| L-Level | N/A | 570 | 556 | 511 | 551 | 46.2 | 25.0 | 74.3 |
| Suite | BOOK | 335 | 321 | 313 | 316 | 48.1 | 43.0 | 56.1 |
| Suite | L2 | 139 | 139 | 109 | 139 | 30.9 | 25.9 | 74.1 |
| Suite | L3 | 107 | 107 | 78 | 107 | 33.6 | 26.2 | 73.8 |
| Suite | MODULE-EXT | 225 | 225 | 190 | 225 | 43.9 | 0.4 | 99.6 |
| Suite | SOC-EXT | 10 | 10 | 8 | 10 | 35.5 | 10.0 | 90.0 |
| Design Level | ip | 126 | 126 | 96 | 126 | 21.8 | 19.8 | 80.2 |
| Design Level | knowledge | 335 | 321 | 313 | 316 | 48.1 | 43.0 | 56.1 |
| Design Level | module | 289 | 289 | 250 | 289 | 46.3 | 9.0 | 91.0 |
| Design Level | soc | 66 | 66 | 39 | 66 | 30.5 | 22.7 | 77.3 |
| Task Category | attribution | 34 | 34 | 29 | 34 | 63.2 | 58.8 | 41.2 |
| Task Category | cwe_classification | 7 | 7 | 5 | 7 | 28.6 | 28.6 | 71.4 |
| Task Category | detection | 60 | 60 | 51 | 60 | 53.3 | 53.3 | 46.7 |
| Task Category | knowledge | 335 | 321 | 313 | 316 | 48.1 | 43.0 | 56.1 |
| Task Category | localization | 51 | 51 | 36 | 51 | 21.4 | 0.0 | 100.0 |
| Task Category | localization_cwe | 235 | 235 | 198 | 235 | 43.6 | 0.9 | 99.1 |
| Task Category | remediation | 32 | 32 | 16 | 32 | 21.9 | 21.9 | 78.1 |
| Task Category | characterization | 62 | 62 | 50 | 62 | 8.9 | 4.8 | 95.2 |
| Question Type | ATTR | 34 | 34 | 29 | 34 | 63.2 | 58.8 | 41.2 |
| Question Type | DET_ACT | 12 | 12 | 9 | 12 | 66.7 | 66.7 | 33.3 |
| Question Type | DET_BEH | 40 | 40 | 34 | 40 | 42.5 | 42.5 | 57.5 |
| Question Type | DET_VUL | 8 | 8 | 8 | 8 | 87.5 | 87.5 | 12.5 |
| Question Type | FILL_IN_THE_BLANK | 7 | 7 | 5 | 7 | 28.6 | 28.6 | 71.4 |
| Question Type | FIX | 16 | 16 | 8 | 16 | 25.0 | 25.0 | 75.0 |
| Question Type | FIX_CWE | 16 | 16 | 8 | 16 | 18.8 | 18.8 | 81.2 |
| Question Type | LOC | 61 | 61 | 44 | 61 | 23.8 | 1.6 | 98.4 |
| Question Type | TRUSTHUB | 62 | 62 | 50 | 62 | 8.9 | 4.8 | 95.2 |
| Question Type | book_long_answer | 105 | 95 | 95 | 93 | 33.9 | 22.6 | 75.8 |
| Question Type | book_mathematical | 12 | 10 | 8 | 10 | 10.0 | 10.0 | 90.0 |
| Question Type | book_short_answer | 88 | 86 | 85 | 84 | 41.1 | 34.5 | 64.0 |
| Question Type | book_true_false | 130 | 130 | 125 | 129 | 65.9 | 65.9 | 33.8 |
| Question Type | enhanced_loc_cwe | 225 | 225 | 190 | 225 | 43.9 | 0.4 | 99.6 |
| Execution Mode | agentic | 802 | 802 | 698 | 797 | 41.8 | 25.3 | 74.2 |
| Execution Mode | unsupported_vision | 14 | 0 | 0 | 0 | — | — | — |
| Backend | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Backend | universal_agent_adapter | 802 | 802 | 698 | 797 | 41.8 | 25.3 | 74.2 |
| Agent Framework | N/A | 14 | 0 | 0 | 0 | — | — | — |
| Agent Framework | trae_agent | 802 | 802 | 698 | 797 | 41.8 | 25.3 | 74.2 |
| L-Level × Type | L2 × ATTR | 24 | 24 | 19 | 24 | 60.4 | 58.3 | 41.7 |
| L-Level × Type | L2 × DET_ACT | 11 | 11 | 8 | 11 | 63.6 | 63.6 | 36.4 |
| L-Level × Type | L2 × DET_BEH | 27 | 27 | 22 | 27 | 37.0 | 37.0 | 63.0 |
| L-Level × Type | L2 × FILL_IN_THE_BLANK | 5 | 5 | 3 | 5 | 40.0 | 40.0 | 60.0 |
| L-Level × Type | L2 × LOC | 22 | 22 | 17 | 22 | 20.3 | 0.0 | 100.0 |
| L-Level × Type | L2 × TRUSTHUB | 50 | 50 | 40 | 50 | 10.0 | 6.0 | 94.0 |
| L-Level × Type | L3 × ATTR | 10 | 10 | 10 | 10 | 70.0 | 60.0 | 40.0 |
| L-Level × Type | L3 × DET_ACT | 1 | 1 | 1 | 1 | 100.0 | 100.0 | 0.0 |
| L-Level × Type | L3 × DET_BEH | 13 | 13 | 12 | 13 | 53.8 | 53.8 | 46.2 |
| L-Level × Type | L3 × DET_VUL | 8 | 8 | 8 | 8 | 87.5 | 87.5 | 12.5 |
| L-Level × Type | L3 × FILL_IN_THE_BLANK | 2 | 2 | 2 | 2 | 0.0 | 0.0 | 100.0 |
| L-Level × Type | L3 × FIX | 16 | 16 | 8 | 16 | 25.0 | 25.0 | 75.0 |
| L-Level × Type | L3 × FIX_CWE | 16 | 16 | 8 | 16 | 18.8 | 18.8 | 81.2 |

*Table 17 (continued).*

| Dimension | Group | #Q | Run | Done | Scored | Avg. (%) | Full (%) | Fail (%) |
|---|---|---|---|---|---|---|---|---|
| L-Level × Type | L3 × LOC | 29 | 29 | 19 | 29 | 22.3 | 0.0 | 100.0 |
| L-Level × Type | L3 × TRUSTHUB | 12 | 12 | 10 | 12 | 4.2 | 0.0 | 100.0 |
| L-Level × Type | N/A × LOC | 10 | 10 | 8 | 10 | 35.5 | 10.0 | 90.0 |
| L-Level × Type | N/A × book_long_answer | 105 | 95 | 95 | 93 | 33.9 | 22.6 | 75.8 |
| L-Level × Type | N/A × book_mathematical | 12 | 10 | 8 | 10 | 10.0 | 10.0 | 90.0 |
| L-Level × Type | N/A × book_short_answer | 88 | 86 | 85 | 84 | 41.1 | 34.5 | 64.0 |
| L-Level × Type | N/A × book_true_false | 130 | 130 | 125 | 129 | 65.9 | 65.9 | 33.8 |
| L-Level × Type | N/A × enhanced_loc_cwe | 225 | 225 | 190 | 225 | 43.9 | 0.4 | 99.6 |

## B THE FINDHT MITIGATION ALGORITHM

FindHT is a lightweight orchestration layer placed *outside* any general-purpose agent. It does not change the underlying model weights. An external *Controller* maintains the evidence base, the set of open obligations, and the candidate state; the agent is responsible for semantic judgment, while the Controller is responsible for mechanically verifiable consistency and state advancement. This division turns heuristic question-answering into an auditable reasoning orchestration. The Controller relies only on a `pyslang` front end and self-contained `read`/`slice`/`arithmetic`/`syntax` checks; it never depends on external verification tools. Algorithm 1 gives the full loop.

**Algorithm 1** FindHT — RTL Trojan-audit orchestrator

**Require:** design source $D$, task contract $T$, `pyslang` front end $\mathcal{F}$
**Ensure:** verdict $V$, typed evidence $\mathcal{E}$, resolved obligations $\mathcal{O}$

$\mathcal{I} \leftarrow \textsc{IndexSources}(D, \mathcal{F})$ ▷ *symbols, modules, ports, always blocks*
$\mathcal{R} \leftarrow \textsc{FocusReading}(\mathcal{I}, T)$ ▷ *candidate regions per task contract*
$\mathcal{H} \leftarrow \textsc{BidirectionalHypotheses}(\mathcal{R})$ ▷ *trigger→payload and payload→trigger*
**for** each hypothesis $h \in \mathcal{H}$ **do**
  $e_h \leftarrow \textsc{InvestigateAndCheck}(h, \mathcal{F})$ ▷ *read/slice/arithmetic/syntax checks*
  **if** $\neg\textsc{TypedCheck}(h, e_h)$ **then**
    discard $h$
  **else**
    $b_h \leftarrow \textsc{SourceBinding}(h, e_h)$ ▷ *bind every claim to file:line and net*
    $\mathcal{O} \leftarrow \mathcal{O} \cup \textsc{OpenIssues}(h, b_h)$
  **end if**
**end for**
**for** each surviving $h \in \mathcal{H}$ **do**
  $j_h \leftarrow \textsc{Judge}(h, e_h, b_h)$ ▷ *agent semantic judgment*
  $q_h \leftarrow \textsc{Challenge}(h, j_h)$ ▷ *independent questioning of the judgment*
  **if** $q_h$ exposes a gap **then**
    $\mathcal{O} \leftarrow \mathcal{O} \cup \{q_h\}$; re-investigate focused region
  **end if**
  $\textsc{ScopeChallenge}(h, T)$ ▷ *confirm the claim stays within the task contract*
**end for**
$\textsc{CloseIssues}(\mathcal{O})$ ▷ *close each obligation with evidence, or mark unresolved*
$V \leftarrow \textsc{FormatProjection}(\mathcal{H}, T)$ ▷ *task-contract projection to the output format*
**if** $V$ violates $T$ **then**
  $\textsc{Fallback}(T)$ ▷ *candidate persistence, time-limited rollup, explicit fallback*
**end if**
**return** $(V, \mathcal{E}, \mathcal{O})$

**Control mechanisms.** FindHT uses three mechanisms to keep the reasoning both auditable and reliable. First, *typed checks* enforce that every intermediate claim carries a concrete type—signal width, clock domain, control-versus-data role—so that a claim that cannot be typed cannot silently survive. Second, *source binding* requires each claim to point at a concrete source location (file, line,

net), which is what makes the evidence base checkable after the fact. Third, *issue closure* forces every open sub-claim to be either resolved against evidence or explicitly recorded as unresolved, so the audit never silently drops a hard question.

**Fault recovery.** FindHT persists partial candidates and rolls up a time-limited verdict when the budget is exhausted, then emits an explicit fallback answer. This path exists specifically to convert "analyzed but produced no valid answer" failures—a dominant reliability loss observed for some agent bases—into a well-formed, low-confidence response that downstream scoring can still consume. FindHT is an auditable reasoning-orchestration framework, not a behavior verifier; its causal gain and cross-combination generalization remain subject to budget-paired and component-ablation study.

## C TrustMe Self-Iterative Trojan Design

The benchmark questions are derived from Trojans produced by TrustMe, a self-iterative design framework. TrustMe mines candidate Trojan seeds from a hardware-weakness corpus and then refines them through a verification-gated, feedback-driven loop until they cover the intended threat space. Algorithm 2 gives the full loop; the paragraph after it explains the three design operators.

**Algorithm 2** TrustMe — self-iterative Trojan design

**Require:** clean design $D$, weakness corpus $\mathcal{K}$, verification backbone $\mathcal{V}$, detector ensemble $\mathcal{E}$, generations $G$, offspring per generation $K$
**Ensure:** elite Trojan archive $\mathcal{A}$ and updated corpus $\mathcal{K}$

$\mathcal{S} \leftarrow \textsc{MineSeeds}(D, \mathcal{K})$ ▷ *seed candidates from CWE and known cases*
$\mathcal{P} \leftarrow \textsc{InitializePopulation}(\mathcal{S})$ ▷ *verified seeds only*
$\mathcal{A} \leftarrow \textsc{InitializeArchive}(\mathcal{P})$
**while** generation budget is not exhausted **do**
  $\mathcal{C} \leftarrow \emptyset$
  **for** $i = 1$ to $K$ **do**
    $p \leftarrow \textsc{BucketAwareSampling}(\mathcal{P})$ ▷ *by harmfulness × evasion × weakness identity*
    $o \leftarrow \textsc{AdaptiveOperatorPolicy}(\mathcal{P}, p)$ ▷ *Mutate / Deepen / Recombine*
    $c \leftarrow \textsc{Apply}(o, p, \mathcal{K})$ ▷ *offspring with corpus-guided context*
    **if** $\mathcal{V}(c) = 1$ **then** ▷ *lint + functional simulation + logic synthesis*
      $(f_R, f_E) \leftarrow \textsc{Evaluate}(c, \mathcal{E})$ ▷ *harmfulness; evasion* $= 1 - P_{\text{detect}}$
      $c.\text{fitness} \leftarrow \omega f_R + (1 - \omega) f_E$
      $\mathcal{C} \leftarrow \mathcal{C} \cup \{c\}$
      $\mathcal{K} \leftarrow \textsc{UpdateCorpus}(\mathcal{K}, c)$ ▷ *feed novel threat tuples back*
    **end if**
  **end for**
  $\mathcal{A} \leftarrow \textsc{UpdateArchive}(\mathcal{A}, \mathcal{C})$
  $\mathcal{P} \leftarrow \textsc{DiversityPreservingPoolUpdate}(\mathcal{P} \cup \mathcal{C})$
  $\textsc{UpdateOperatorPreference}(\mathcal{P}, \mathcal{A}, \mathcal{C})$
**end while**
**return** $\mathcal{A}, \mathcal{K}$

**Verification gate.** A candidate is accepted only if it passes three sequential checks against the original module and its unmodified testbench: (i) Verilator lint with all warnings enabled; (ii) functional simulation requiring a clean exit and no failure markers; and (iii) logic synthesis through the module's own target. A candidate failing any check is repaired with a bounded number of attempts and discarded on budget exhaustion. This gate is what keeps the benchmark's ground-truth answers realizable in hardware rather than merely textually plausible.

**Design operators.** *Mutate* preserves the trigger–state–payload mechanism while rewriting its realization (renaming signals, splitting a conjunction, re-targeting ports), attacking detectors that key on surface structure. *Deepen* keeps the trigger fixed and inserts a temporal stage (a window counter plus a confirmation condition) so that activation requires a sequence of events spread over time rather than a single instant. *Recombine* joins the guard of one family with the payload of another through a

freshly latched signal that neither parent contains, producing a composite whose parts are individually familiar but whose combination is novel. The three operators are complements, not substitutes: each raises a different property of the threat while leaving the verification pipeline intact.

## D BENCHMARK REPRODUCTION

The benchmark is a self-contained Python harness. A single configuration is executed with `harness/run_paper.py` against a JSON configuration under `config/`; a batch of configurations is executed with `batch/run_batch.py` against a list under `batch/`. A batch can be smoke-tested on the first $N$ questions with `-question-limit`.

```
# run a single configuration
python3 harness/run_paper.py config/l1_chat_gpt.json
python3 harness/run_paper.py config/l2_agentic_claude.json
python3 harness/run_paper.py config/l3_agentic_qwen.json

# run a batch of configurations
python3 batch/run_batch.py batch/l1_chat.txt
python3 batch/run_batch.py batch/l2_agentic.txt
python3 batch/run_batch.py batch/l3_agentic.txt

# smoke-test the first N questions of a batch
python3 batch/run_batch.py batch/l2_agentic.txt --question-limit 10
python3 batch/run_batch.py batch/l3_agentic.txt --question-limit 10
```

Each configuration instantiates an *agent base* (e.g. `claude_code`, `codex`, `deepseek_harness`, `opencode`, `trae_agent`) driving a *large language model* through a universal agent adapter. The reported runs used `temperature`=0.1, `timeout`=1200 s per question, an infrastructure-only retry policy (up to 2 retries with exponential backoff from 30 s to 120 s), and an infrastructure circuit breaker (threshold 3, cooldown 300–900 s). Concurrency is adaptive: it steps down one worker on a failure and steps back up after three consecutive successes. Research capture records the observable trace and, where exposed, the chain-of-thought. The authoritative run configuration is reproduced verbatim in Figure 6.

```
{
  "execution_mode": "agentic",
  "agent_framework": "claude_code",
  "adapter": "Agent Adapter/universal_agent_adapter.py",
  "agent_driver": "claude_code_docker",
  "model_spec": {
    "provider": "deepseek_official", "model": "deepseek-v4-flash",
    "api_env": "DEEPSEEK_API_KEY", "base_url": "https://api.deepseek.com/anthropic"
  },
  "runtime": { "temperature": 0.1, "timeout_seconds": 1200,
               "retry_policy": "infrastructure_only", "max_retries": 2,
               "docker_memory": "4g", "docker_cpus": "2", "docker_pids_limit": 256 },
  "infrastructure_circuit_breaker": { "enabled": true, "failure_threshold": 3 },
  "adaptive_concurrency": { "enabled": true, "min_workers": 1,
                            "recovery_successes_to_step_up": 3 },
  "research_capture": { "enabled": true,
                        "capture_observable_trace": true,
                        "capture_exposed_chain_of_thought": true },
  "suite_mode": "agent_only", "paper": "AGENT-ONLY-816",
  "capabilities": { "vision": false }
}
```

Figure 6: Authoritative run configuration (`run_config.snapshot.json`) for one representative combination (claude_code + deepseek-v4-flash). Other combinations differ only in agent framework, model spec, and capability flags.

# E Evaluation Protocol and Metric Definitions

## E.1 Benchmark composition

The benchmark contains 816 questions organized into four datasets and, orthogonally, into eight task categories (atomic capabilities) and fourteen question types. Table 18 reports the full composition. The four datasets are **H** (book_qa, 335 hardware-security knowledge questions), **A** (htbench_core, 246 core Trojan-detection questions split into L2 and L3 difficulty), **M** (module_loc_cwe, 225 module-level CWE localization questions), and **S** (soc_loc_extension, 10 SoC-level extension questions).

Table 18: Benchmark composition. Left: the four datasets. Middle: the eight task categories. Right: the fourteen question types. Each decomposition sums to 816.

| Dataset | # | Task category | # | Question type | # |
|---|---|---|---|---|---|
| H (book_qa) | 335 | attribution | 34 | ATTR | 34 |
| A (htbench_core) | 246 | cwe_classification | 7 | DET_ACT | 12 |
| M (module_loc_cwe) | 225 | detection | 60 | DET_BEH | 40 |
| S (soc_loc_extension) | 10 | knowledge | 335 | DET_VUL | 8 |
| | | localization | 51 | FILL_IN_THE_BLANK | 7 |
| | | localization_cwe | 235 | FIX | 16 |
| | | remediation | 32 | FIX_CWE | 16 |
| | | characterization | 62 | LOC | 61 |
| | | | | TRUSTHUB | 62 |
| | | | | book_long_answer | 105 |
| | | | | book_mathematical | 12 |
| | | | | book_short_answer | 88 |
| | | | | book_true_false | 130 |
| | | | | enhanced_loc_cwe | 225 |
| **Total** | **816** | **Total** | **816** | **Total** | **816** |

The same 816 questions are partitioned three ways; the three decompositions are cross-cutting views, not disjoint subsets.

## E.2 Scoring and metrics

Every question carries a deterministic scoring rubric (e.g. trigger/effect/location/description weights for characterization questions), and a single judge model (glm-5.2) scores each answer against a reference answer. The metrics reported throughout are defined as follows.

- *Average score* — the mean rubric-normalized score over all *scored* questions, as a percentage. Zero-score and execution-failed questions are excluded from this average.
- *Full-score accuracy* — the fraction of questions answered with the maximum possible score.
- *Failure rate* — the fraction of the 816 questions whose execution failed (agent failure, LLM failure, or infrastructure failure). *Pass rate* $= 1-$ failure rate.
- *Agent Token* — total tokens consumed by the agent, including input, output, and cache-read tokens.
- *Average time* — mean wall-clock time per question (converted from milliseconds).

## E.3 Answer-quality note

"Full-score accuracy" and "average score" describe different ability profiles. A combination with a high full-score accuracy but a middling average score (e.g. opencode + glm-5.3, full-score $35.1\%$ vs. average $55.8\%$) tends toward "all-or-nothing" answering, whereas a combination with a low average score despite a passable full-score accuracy is concentrated in low-score partial answers. The two numbers are therefore reported side by side rather than merged.

### E.4 JUDGE PROMPT STRUCTURE

Every judged question is scored against a fixed prompt template that bundles five blocks: the judge-rule fingerprint (a cache identity), the question record, the reference answer, the candidate answer, and the response-format instruction. Figure 7 reproduces the structure of the template for a characterization item (L2-090, a TrustHub case), trimmed to its field layout.

```
Evaluate the following private benchmark item using the system rubric.

Judge rule/config fingerprint (cache identity only): 49dafd26...

{
  "question_id": "L2-090",
  "question_type": "trusthub",
  "question": {
      task description, trigger/effect/location option lists,
      points ("2"), scoring rubric, strict-JSON response format
  },
  "reference_answer": { trigger, effect, location, description },
  "candidate_answer": { trigger, effect, location, description }
}

Return only the required JSON object.
```

Figure 7: Judge prompt template for a characterization item. The scoring rubric inside the question record for L2-090 awards 0.3 points each for trigger, effect, and location, and 1.1 points for the description, evaluated for coverage and semantic correctness.

Three properties of this template matter for interpreting the results. First, scoring is deterministic for the closed-choice fields (trigger, effect, location) and judge-mediated only for the free-text description, so partial credit is anchored to a fixed rubric rather than to judge taste. Second, the fingerprint ties the prompt to a judge-rule version, which prevents silent rubric drift across the 15 combinations. Third, the candidate answer is inserted verbatim, so an agent that fails to produce the strict-JSON contract scores zero on every field regardless of the quality of its reasoning — one of the failure modes that FindHT's format-projection stage targets directly.

### E.5 FAILURE TAXONOMY

Execution failures are attributed to one of three causes, recorded per question in the metrics file and aggregated in the group-summary sheet:

- *Agent failure* — the agent base crashed, timed out, or produced no admissible answer. This is a property of the agent scaffold rather than of the model: the same model on a different base can avoid it entirely. The spread across bases is large (claude_code recorded $0$ agent failures, while deepseek_harness recorded up to $175$ on the same question set).
- *LLM failure* — the model returned a response that could not be scored (missing required fields, refusal, or an output contract violation). This is the dominant cause overall and is counted against the pass rate.
- *Infrastructure failure* — network, provider, or harness faults. These are retried transparently (infrastructure-only policy, at most $2$ retries with exponential backoff) and are therefore rare in the final tallies.

Two further reliability mechanisms shape the observed numbers. An *infrastructure circuit breaker* opens after three consecutive infrastructure failures and cools down for $300$–$900$ s, preventing a provider outage from burning the whole question budget. *Adaptive concurrency* steps the worker count down by one on any failure and back up after three consecutive successes, so flaky regions of the run execute with less parallel pressure. Finally, $14$ questions require vision and are recorded as `unsupported_vision` for every non-vision combination; they appear in the per-combination tables as executed-zero rows rather than as failures, since no admissible answer was possible.

# F BENCHMARK ARTIFACTS AND DIRECTORY STRUCTURE

A finished run is packaged into a self-describing directory. The package index records, for each of the 15 combinations, the agent base, the model, the judge, the run identifier, the authoritative run source, and the SHA-256 of both the final workbook and the question-results file, so that every reported number is traceable to a reproducible artifact. Figure 8 shows the directory structure.

```
HTBench/
|-- config/                          # 24 single-run configurations
|-- batch/                           # batch lists + run_batch.py
|-- harness/run_paper.py             # single-config runner
|-- data/                            # benchmark question catalog
|-- prompts/                         # per-question prompts
|-- meta/                            # schema + metadata
|-- template/                        # report templates
|-- ref/                             # reference material
|-- runtime/                         # runtime helpers
`-- RESULT/
    `-- final/
        |-- INDEX.json               # 15-package index (judge = glm-5.2)
        |-- ORGANIZATION_STATUS.json# complete, 15/15 verified
        |-- agent/                   # 15 agentic combinations
        |   `-- <agent>+<llm>/
        |       |-- PACKAGE_MANIFEST.json
        |       |-- <run_id>_816_judge-glm-5.2.xlsx   # final workbook
        |       `-- run/
        |           |-- run_config.snapshot.json      # authoritative config
        |           |-- metrics.json                  # aggregate metrics
        |           |-- question_results.json         # per-question results
        |           |-- complete.json / source.json / prepared.json
        |           |-- judge/{answers,attempts,errors,prompts,requests,traces}/
        |           `-- report/
        |               |-- REPORT.md                 # human-readable summary
        |               |-- report.json               # machine-readable summary
        |               |-- group_summary.csv / question_details.csv
        |               |-- failure_attributions.json
        |               `-- trace_outcome_analysis.{json,md,xlsx,csv}
        `-- llm/                      # 7 direct-LLM (non-agentic) baselines
            `-- *_llm_186_judge-glm-5.2.xlsx
```

Figure 8: Directory structure of a finished benchmark run. The final workbook of each combination aggregates five sheets — overview, group summary, per-question detail, unscored-and-skipped, and run statistics — from which all reported metrics are derived.

The final workbook of each combination contains five sheets. The *overview* sheet reports the headline numbers (average score, full-score accuracy, failure count, Agent Token, average time). The *group summary* sheet cross-tabulates every metric over nine dimensions (dataset, L-level, suite, design level, task category, question type, execution mode, backend, agent framework, and the L-level×type cross product) and is the source of the fifteen per-combination tables in Appendix A. The *per-question detail* sheet records each question's rubric score, and the *run statistics* sheet records token, time, and failure attribution per question. The *unscored-and-skipped* sheet records the questions that reached no score, including the 14 vision-requiring questions that every combination marked `unsupported_vision`.

# G A WORKED TROJAN EXAMPLE

This section traces one concrete benchmark item to make the characterization task concrete. The item is a TrustHub case known to contain a hardware Trojan; the agent must characterize its trigger, effect, and location, and produce an evidence-based description, without re-evaluating a trivial present/absent flag.

**Ground truth.** The reference answer characterizes the Trojan as *trigger* = Always on, *effect* = Leak Information, *location* = Processor. The mechanism is a power side-channel that leaks a secret key from a cryptographic core running AES: a pseudo-random number generator (PRNG), initialized to a predefined value, produces a CDMA spreading sequence; the secret bits are XOR-modulated by that sequence; and the modulated signal drives a leakage circuit realized by eight identical flip-flops tied to a single XOR output so as to mimic a large capacitance in the power trace.

**Mechanism sketch.** Figure 9 shows an illustrative sketch of the mechanism, not the exact benchmark source. The trigger is always-on (the PRNG and modulator run unconditionally); the state element is the PRNG; and the payload is the leakage circuit. The characterization rubric awards partial credit for the trigger/effect/location triple and the remainder for the evidence-based description, which is why agents that only state "a Trojan is present" score near zero on this item.

```
// illustrative sketch of an always-on CDMA power side-channel
module trojan_sketch(clk, key_bit, out);
  input  clk, key_bit;          // key_bit: one secret bit stream
  output out;
  reg [7:0] lfsr = 8'hA5;       // PRNG, predefined seed  -> state
  reg       mod;                // spread-spectrum modulator
  // trigger: always on (no arming condition)
  always @(posedge clk) begin
    lfsr <= {lfsr[6:0], lfsr[7] ^ lfsr[5] ^ lfsr[4] ^ lfsr[3]};
    mod  <= key_bit ^ lfsr[0];  // XOR-modulate secret by spreading code
  end
  // payload: leakage circuit mimics large capacitance (8 FF fan-out)
  assign out = mod;             // drives heavy capacitive load
endmodule
```

Figure 9: Illustrative mechanism sketch of the always-on CDMA power side-channel leakage Trojan. This is a conceptual rendering for exposition; it is not the exact benchmark source.

## H Responsible Disclosure and Dual-Use Considerations

The benchmark consists of RTL Trojan-detection questions derived from publicly documented cases (TrustHub and CWE-based weaknesses). No proprietary design, commercial IP, or fabricated device is involved, and no result describes the behavior of a specific product. The generated artifacts are benchmark test cases rather than exploits: they are built to satisfy a verification pipeline on public modules and depend on those modules' own ports and signals, so a case does not transfer to a real product without the access an attacker would already need to insert a Trojan there at all. No vulnerability in a third-party product was discovered, so no vendor disclosure applies. We intend to release the benchmark and the evaluation harness for research use so that the same cases serve both the detection and the generation directions of work.