# **Accessibility Benchmark Comparison Across IDEs & Phases**

## **1. Error Volumes by IDE & Phase**

| **IDE / Phase** | **Phase 1** | **Phase 2** | **Phase 3** | **Trend** |
| --- | --- | --- | --- | --- |
| **Builder Fusion** | 10 | 2 | 2 | Sharp drop after Phase 1, but persistent structural issues |
| **Copilot (GPT-5)** | 5 | 6 | 3 | Regression in Phase 2, improvement in Phase 3 |
| **Copilot (Claude Sonnet 4)** | 10 | 6 | 3 | Similar to GPT-5: starts with many errors, improves by Phase 3 |
| **Vercel V0** | 5 | 6 | 3 | Relatively flat, persistent link/contrast issue’ |

🔎 **Insight:** All IDEs show a **converging pattern toward fewer errors by Phase 3**, but types of errors differ (structural vs. semantic vs. visual contrast).

## **2. Error Severity Distribution**

### **Builder Fusion**

* **Phase 1**: 1 critical, 7 high, 1 low, 1 info → heavily skewed to serious foundational problems.  
  **Phase 2–3**: Only 2 errors each, but still **high severity structural issues** (Info & Relationships).

### **Copilot (GPT-5)**

* **Phase 1**: Mix of low + high, including **contrast failures**.
* **Phase 2**: 6 errors (mostly ARIA/role misuse), high severity.
* **Phase 3**: 3 errors (1 critical on combobox name, 2 high).

### **Copilot (Claude Sonnet 4)**

* **Phase 1**: 10 total (1 critical, 6 high, 3 low) → diverse issues including **captions, audio control, contrast**.
* **Phase 2**: 6 total (mostly ARIA/roles, some contrast).
* **Phase 3**: 3 total (1 critical, 2 high) → same issues as GPT-5: **contrast + role/name problems**.

### **Vercel V0**

* **Phase 1**: 5 errors (mostly high, link purpose & placeholder issues).
* **Phase 2**: 6 errors (contrast failures dominate).
* **Phase 3**: 3 errors (all high: **contrast + non-unique link names**).

## **3. Guideline Patterns (WCAG 2.1)**

### **🔧 Structural / Semantic**

* **Builder Fusion**: Mostly **1.3.1 Info & Relationships** → missing headings, list structures.
* **Copilot GPT-5 & Claude Sonnet 4**: Persistent **4.1.2 Name, Role, Value** errors → misuse of ARIA, redundant roles, combobox issues.

### **🎨 Visual Contrast**

* **All IDEs** fail **1.4.3 Contrast (Minimum)** in multiple phases.
* **Vercel V0** → worst offender, especially in Phase 2 & 3 (7 contrast errors Phase 3 alone).
* **Claude Sonnet 4** also flagged **1.4.2 Audio Control** and **captions** (multimedia issues not seen in others).

### **🔗 Link Purpose & Naming**

* **Vercel V0** repeatedly fails **2.4.4 Link Purpose** (non-unique or suspicious names).
* **Builder Fusion & Copilot** less affected here.

### **🎥 Multimedia Accessibility**

* Only **Claude Sonnet 4 (Phase 1)** flagged multimedia issues (captions & audio control), showing broader coverage.

## **4. Comparative Trends**

* **Builder Fusion**:  
   Starts with many fundamental issues (10, including 1 critical) but almost flatlines at 2–2 later. However, the remaining issues are **basic structural failures** → indicating it doesn’t evolve to cover advanced accessibility at all.
* **Copilot GPT-5**:  
   Mid-range overall. Phase 2 worsened (more ARIA issues), but Phase 3 shows clear **improvement** (down to 3 issues, 1 critical). Stronger at **semantic scaffolding**, weaker at **ARIA correctness & contrast**.
* **Copilot Claude Sonnet 4**:  
   Broadest error coverage (even beyond contrast/roles, into multimedia). Phase 3 converges with GPT-5 (3 issues, critical + high). Suggests Sonnet is more **comprehensive early on**, but still struggles with **role/contrast issues later**.
* **Vercel V0**:  
   Most consistent but **consistently poor at contrast & link naming**. Doesn’t regress, but also doesn’t resolve persistent weaknesses. Likely produces more “visual-first” output with little accessibility focus.

## **5. 🏆 Key Insights & Rankings**

### **By Foundational Accessibility (Headings, Structure)**

* **Winner**: Copilot GPT-5 / Claude Sonnet 4 (better semantic coverage)
* **Weakest**: Builder Fusion (basic missing scaffolding)

### **By Visual Accessibility (Contrast Compliance)**

* **Winner**: Builder Fusion (fewer color contrast errors, though maybe due to simplistic design)
* **Weakest**: Vercel V0 (chronic contrast failures, especially Phase 2–3)

### **By Comprehensiveness**

* **Winner**: Claude Sonnet 4 (caught captions, audio control, and broader WCAG categories)
* **Weakest**: Builder Fusion (very narrow scope: mostly 1.3.1)

### **By Improvement Over Phases**

* **Best Improvement**: Copilot GPT-5 and Claude Sonnet 4 (both drop to 3 errors by Phase 3).
* **Least Change**: Vercel V0 (remains flat at 5–6–3, with the same contrast/link problems).

## **📌 Overall Conclusion**

* **Builder Fusion**: Lightweight, fewer issues but fails accessibility fundamentals → *not reliable for production*.
* **Copilot GPT-5**: More balanced, improves with phases, but still **ARIA & contrast weak spots**.
* **Copilot Claude Sonnet 4**: Most comprehensive, surfaces a wider spectrum of issues, converges toward GPT-5 in Phase 3.
* **Vercel V0**: Strong visual/UI generator, but **consistently fails contrast and naming**, making it weakest for compliance.