# A2UI Multi-Model Benchmark Report

Local Inference on Apple M1 Max

---

## Comparing 7 Models for A2UI v0.9 Protocol Generation

Structured UI Generation from Natural Language Prompts

| | |
|---|---|
| **Date:** | May 12, 2026 |
| **Hardware:** | Apple M1 Max, 64 GiB RAM |
| **Framework:** | Inspect AI + llama.cpp |
| **Judge:** | Claude Code (interactive, human-supervised) |
| **Protocol:** | A2UI v0.9 Specification |
| **Samples:** | 73 evaluation prompts |

# Contents

# Executive Summary

This report compares **7 local language models** for generating structured user interfaces conforming to the A2UI v0.9 protocol. All evaluations were conducted on a **Apple M1 Max** system running macOS/arm64 using **llama.cpp** as the inference engine.

Each model was evaluated on two axes:

1. **Schema Validation** – programmatic check that output is valid A2UI JSON
2. **Semantic Grading** – LLM-as-a-Judge assessment of UI correctness (C/P/I)

## Models Evaluated

Table 1: Models included in this benchmark.

| Model | ID | Engine | Type | Params | Quant |
|---|---|---|---|---|---|
| Llama 3.1 8B | `NousResearch/Meta-Llama-3.1-8B-Instruct` | llama.cpp | Dense | 8B | none |
| Qwen 3 1.7B | `Qwen/Qwen3-1.7B` | llama.cpp | Dense | 1.7B | none |
| Qwen 3 14B | `Qwen/Qwen3-14B` | llama.cpp | Dense | 14B | none |
| Qwen 3 8B | `Qwen/Qwen3-8B` | llama.cpp | Dense | 8B | none |
| Qwen3 30B MoE NVFP4 | `nvidia/Qwen3-30B-A3B-NVFP4` | llama.cpp | MoE | 30B (3B active) | none |
| Llama 3.2 1B | `unsloth/Llama-3.2-1B-Instruct` | llama.cpp | Dense | 1B | none |
| Gemma 4 26B MoE | `unsloth/gemma-4-26B-A4B-it-GGUF` | llama.cpp | MoE | 26B (4B active) | none |

# Accuracy Comparison

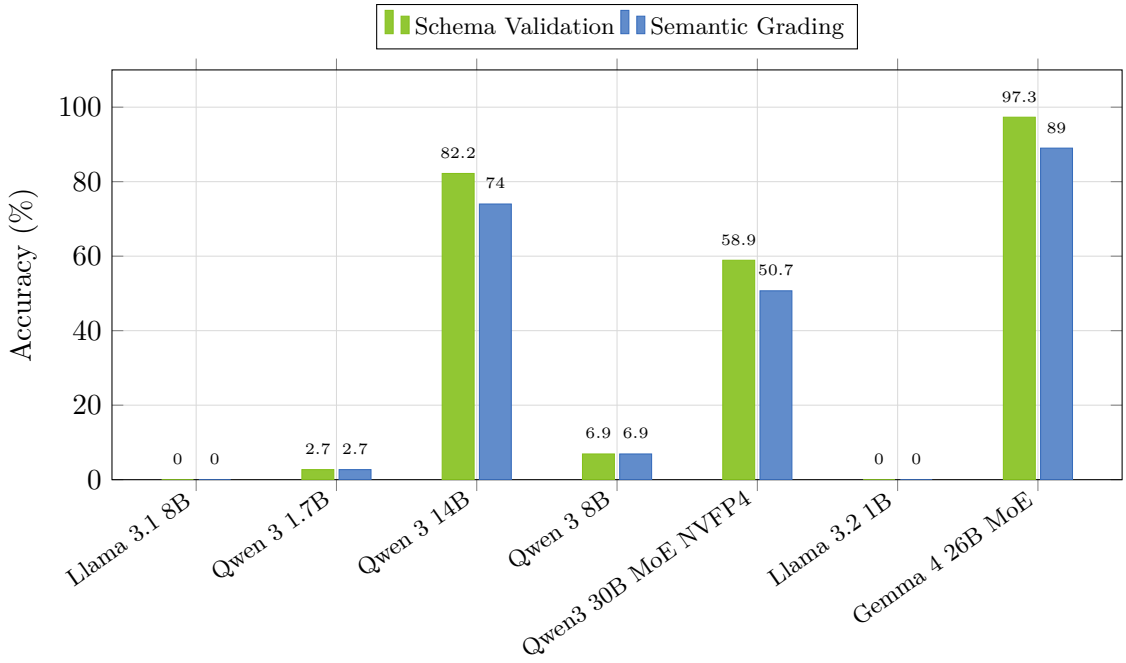


Figure 1: Accuracy comparison across models. Schema validation checks structural JSON correctness; semantic grading evaluates whether the generated UI fulfills the user's intent.

# Inference Performance

## Average Inference Time

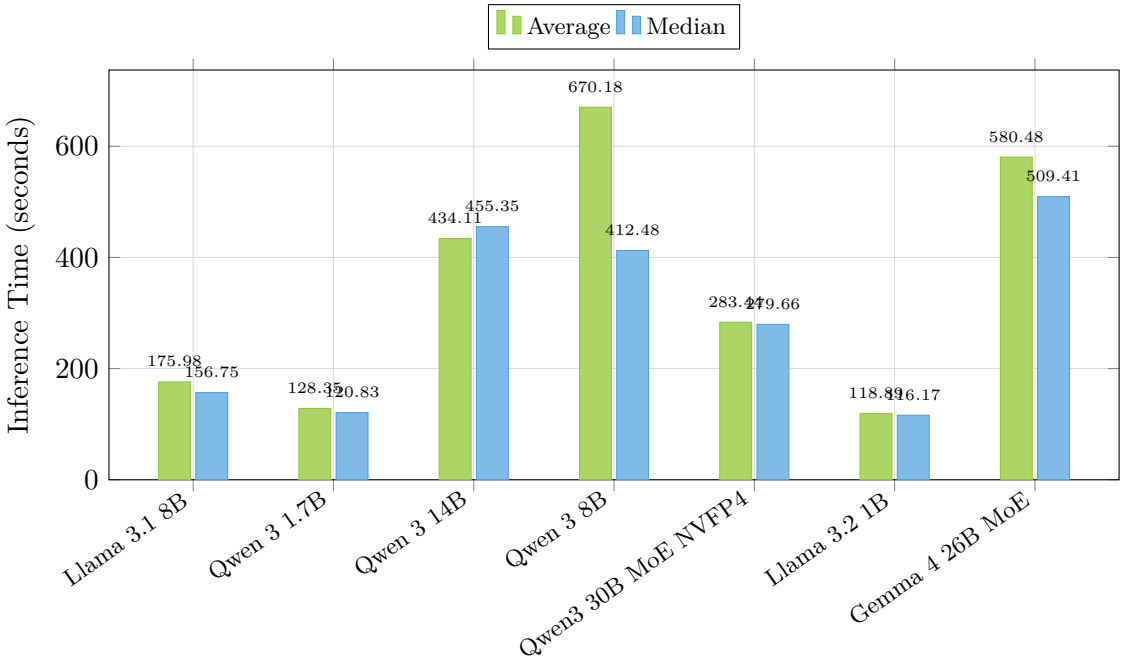


Figure 2: Average and median inference time per sample. Includes full prompt processing (system prompt with A2UI schema ∼8k tokens) and output generation.

## Per-Sample Inference Time Distribution

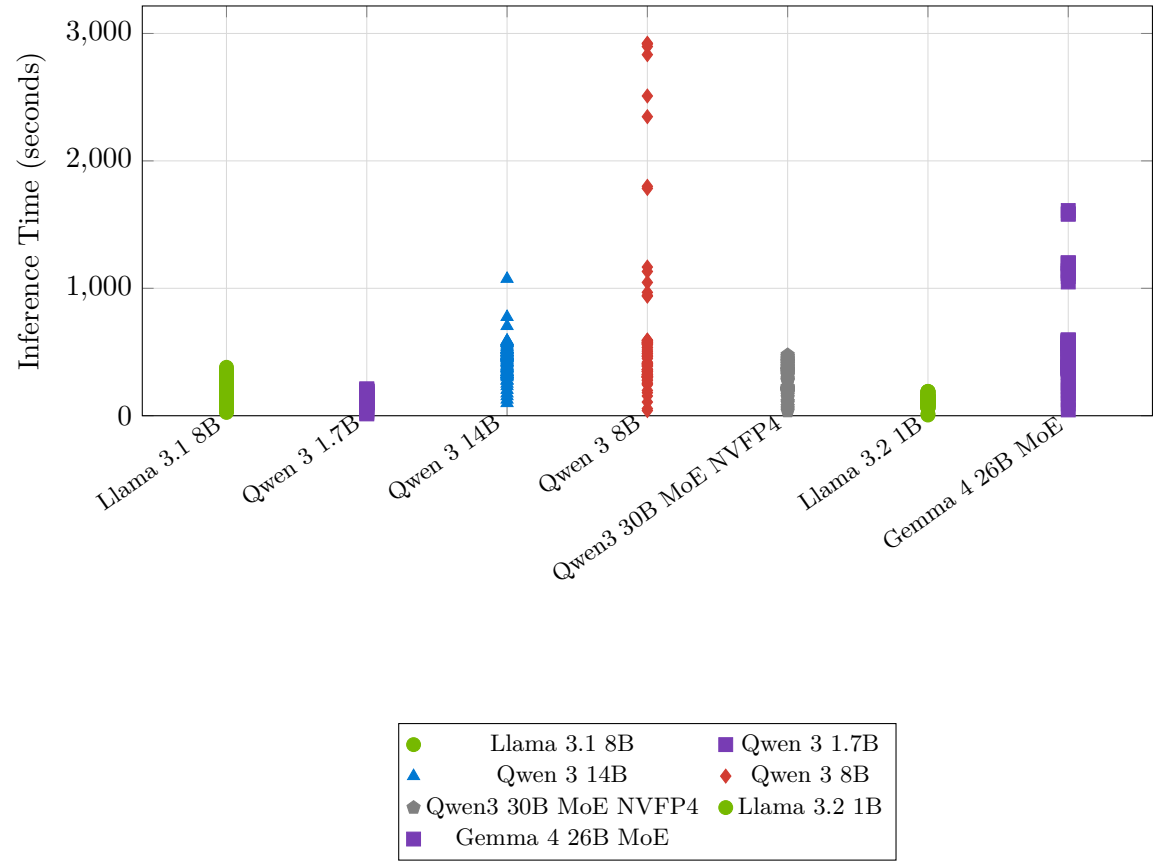


Figure 3: Per-sample inference time scatter plot. Each point represents one evaluation sample. Variation reflects differences in output length and complexity.

## Output Throughput

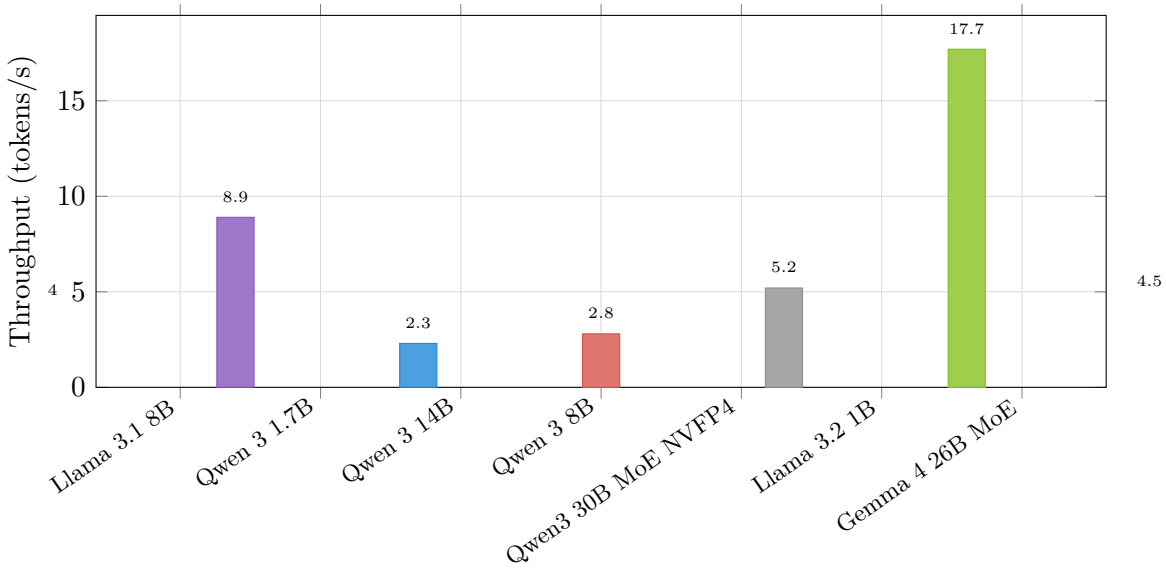


Figure 4: Effective output generation throughput (total output tokens / total inference time). Higher is better. MoE models typically achieve higher throughput due to fewer active parameters per token.

# Memory and Resource Utilization

## Model Memory Footprint vs. Available KV Cache

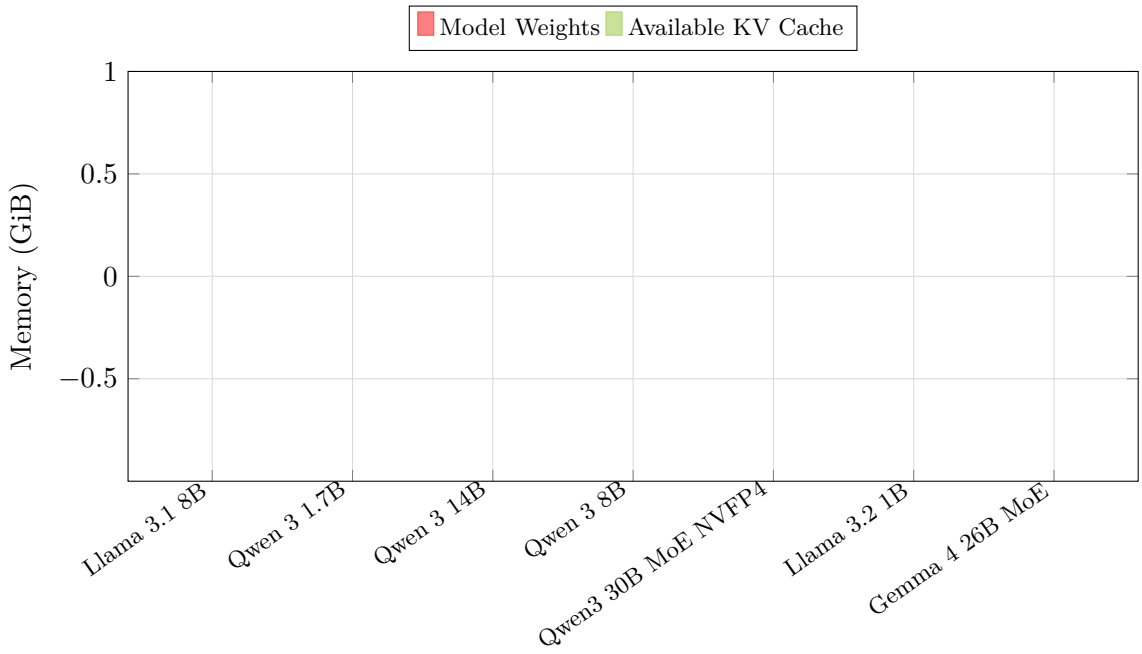


Figure 5: GPU memory allocation per model. Smaller model footprints leave more memory for KV cache, enabling higher concurrency and longer context windows. Total RAM: 64 GiB.

## In-Context Learning (ICL) Prompt Size

The A2UI evaluation injects the full v0.9 JSON Schema and component catalog as a system prompt. This constitutes the in-context learning (ICL) payload that every model must process before generating output.

Table 2: ICL context and memory utilization per model.

| Metric | Llama 3.1 8B | Qwen 3 1.7B | Qwen 3 14B | Qwen 3 8B | Qwen3 30B MoE NVFP4 | Llama 3.2 1B | Gemma 4 26B MoE |
|---|---|---|---|---|---|---|---|
| ICL Context (tokens) | 8,739 | 8,770 | 8,770 | 8,770 | 8,770 | 8,739 | 9,181 |
| Model Memory (GiB) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| KV Cache Available (GiB) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Prefix Cache Hit Rate | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |

## Prefill vs. Decode Throughput

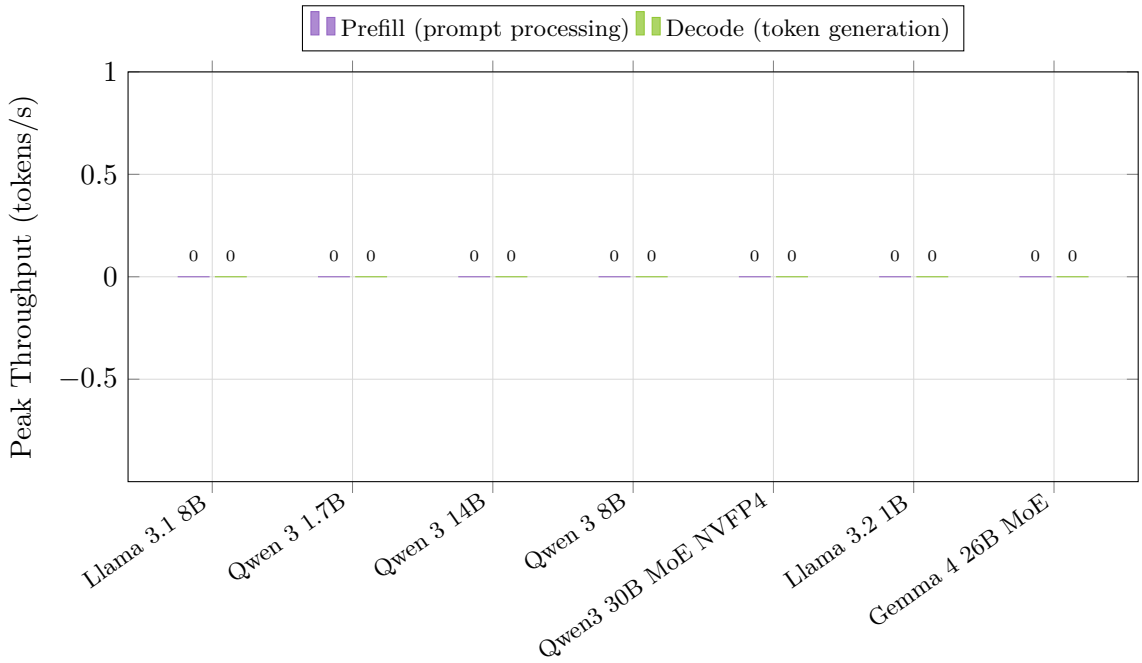


Figure 6: Peak prefill vs. decode throughput. Prefill processes the input prompt (including the ∼8k token ICL context) in parallel; decode generates output tokens autoregressively. The prefill/decode ratio indicates compute vs. memory-bandwidth boundedness.

# Detailed Per-Sample Results

Table 3: Per-sample results across all models. Schema: structural JSON validation (PASS/FAIL). Grade: semantic correctness (C=Correct, P=Partial, I=Incorrect). Time: inference seconds.

| Sample | Llama 3.1 8B | | | Qwen 3 1.7B | | | Qwen 3 14B | | | Qwen 3 8B | | | Qwen3 30B MoE NVFP4 | | | Llama 3.2 1B | | | Gemma 4 26B MoE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Schema | Grade | Time | Schema | Grade | Time | Schema | Grade | Time | Schema | Grade | Time | Schema | Grade | Time | Schema | Grade | Time | Schema | Grade | Time |
| text_heading | FAIL | I | 51.6 | PASS | C | 81.9 | PASS | C | 251.8 | FAIL | I | 198.6 | PASS | C | 177.7 | FAIL | I | 114.4 | PASS | C | 255.8 |
| text_caption | FAIL | I | 90.0 | FAIL | I | 29.2 | PASS | C | 203.3 | FAIL | I | 40.5 | PASS | C | 111.0 | FAIL | I | 123.9 | PASS | C | 230.7 |
| text_markdown | FAIL | I | 56.0 | FAIL | I | 71.5 | PASS | C | 231.9 | FAIL | I | 107.5 | PASS | C | 53.2 | FAIL | I | 63.3 | PASS | C | 174.3 |
| image_avatar | FAIL | I | 26.1 | FAIL | I | 90.5 | PASS | C | 175.3 | PASS | C | 154.8 | PASS | C | 71.9 | FAIL | I | 61.3 | PASS | C | 202.0 |
| image_header_banner | FAIL | I | 85.7 | FAIL | I | 14.2 | PASS | C | 99.0 | FAIL | I | 275.3 | PASS | C | 35.9 | FAIL | I | 12.8 | PASS | C | 290.1 |
| icon_inline | FAIL | I | 35.5 | FAIL | I | 41.8 | PASS | C | 278.2 | PASS | C | 299.7 | PASS | C | 160.5 | FAIL | I | 5.5 | PASS | C | 45.4 |
| video_player | FAIL | I | 44.0 | PASS | C | 51.1 | PASS | C | 126.6 | FAIL | I | 56.3 | PASS | C | 98.0 | FAIL | I | 59.6 | PASS | C | 384.1 |
| audio_player | FAIL | I | 78.6 | FAIL | I | 111.7 | PASS | C | 341.3 | FAIL | I | 326.9 | PASS | C | 140.5 | FAIL | I | 140.3 | PASS | C | 92.4 |
| divider_horizontal | FAIL | I | 62.8 | FAIL | I | 20.6 | PASS | C | 153.3 | FAIL | I | 180.7 | PASS | C | 189.8 | FAIL | I | 65.9 | PASS | C | 339.2 |
| button_primary | FAIL | I | 95.3 | FAIL | I | 102.0 | PASS | C | 314.1 | FAIL | I | 242.8 | FAIL | I | 207.7 | FAIL | I | 63.1 | PASS | C | 124.6 |
| button_borderless | FAIL | I | 74.5 | FAIL | I | 111.5 | PASS | C | 274.6 | FAIL | I | 305.5 | PASS | C | 196.7 | FAIL | I | 141.9 | PASS | C | 392.3 |
| textfield_password | FAIL | I | 70.2 | FAIL | I | 122.0 | FAIL | I | 310.8 | FAIL | I | 310.1 | PASS | C | 209.7 | FAIL | I | 139.6 | PASS | C | 376.8 |
| text_then_button | FAIL | I | 68.1 | FAIL | I | 120.8 | PASS | C | 322.1 | FAIL | I | 280.7 | PASS | C | 201.8 | FAIL | I | 141.2 | PASS | C | 375.4 |
| image_caption | FAIL | I | 73.4 | FAIL | I | 115.5 | PASS | C | 326.9 | FAIL | I | 250.6 | PASS | P | 188.5 | FAIL | I | 141.4 | PASS | C | 358.9 |
| icon_label | FAIL | I | 75.9 | FAIL | I | 117.0 | PASS | C | 324.7 | FAIL | I | 256.6 | PASS | C | 194.4 | FAIL | I | 188.0 | PASS | C | 377.4 |
| two_buttons_row | FAIL | I | 106.1 | FAIL | I | 103.7 | PASS | C | 325.3 | FAIL | I | 322.3 | PASS | C | 204.6 | FAIL | I | 191.5 | PASS | C | 413.7 |
| heading_paragraph | FAIL | I | 100.5 | FAIL | I | 108.7 | PASS | C | 346.6 | FAIL | I | 300.7 | PASS | C | 201.4 | FAIL | I | 189.3 | PASS | C | 427.5 |
| divider_between_texts | FAIL | I | 187.1 | FAIL | I | 127.7 | PASS | C | 347.6 | PASS | C | 305.7 | PASS | P | 198.2 | FAIL | I | 157.0 | PASS | C | 421.4 |
| checkbox_simple | FAIL | I | 185.8 | FAIL | I | 121.2 | PASS | C | 340.0 | FAIL | I | 315.6 | PASS | C | 207.3 | FAIL | I | 152.9 | PASS | C | 394.1 |
| slider_volume | FAIL | I | 182.7 | FAIL | I | 118.2 | PASS | C | 344.3 | FAIL | I | 305.3 | PASS | C | 214.1 | FAIL | I | 145.9 | PASS | C | 369.1 |
| date_picker_only | FAIL | I | 188.2 | FAIL | I | 110.5 | FAIL | I | 330.8 | PASS | C | 344.8 | PASS | C | 200.5 | FAIL | I | 187.1 | PASS | C | 367.3 |
| choice_picker_chips | FAIL | I | 189.5 | FAIL | I | 103.0 | PASS | C | 301.8 | FAIL | I | 346.7 | FAIL | I | 183.6 | FAIL | I | 187.2 | PASS | C | 368.9 |
| textfield_search | FAIL | I | 191.0 | FAIL | I | 99.3 | PASS | C | 362.6 | FAIL | I | 362.9 | PASS | P | 185.3 | FAIL | I | 139.7 | PASS | C | 365.4 |
| textfield_multiline | FAIL | I | 185.7 | FAIL | I | 101.5 | PASS | C | 372.7 | FAIL | I | 361.5 | FAIL | I | 189.7 | FAIL | I | 163.2 | PASS | C | 381.4 |
| card_with_text | FAIL | I | 193.2 | FAIL | I | 104.7 | PASS | C | 371.7 | FAIL | I | 357.3 | PASS | C | 210.5 | FAIL | I | 116.2 | PASS | C | 371.6 |
| card_with_image_text_button | FAIL | I | 172.8 | FAIL | I | 127.6 | PASS | C | 378.2 | FAIL | I | 392.9 | PASS | P | 193.1 | FAIL | I | 115.3 | PASS | C | 370.9 |
| card_profile_summary | FAIL | I | 177.6 | FAIL | I | 118.6 | PASS | C | 379.6 | FAIL | I | 306.5 | PASS | C | 201.3 | FAIL | I | 117.6 | PASS | C | 382.2 |
| card_stats | FAIL | I | 95.1 | FAIL | I | 111.3 | PASS | C | 390.5 | FAIL | I | 324.4 | FAIL | I | 202.5 | FAIL | I | 104.1 | PASS | C | 392.0 |
| card_with_divider | FAIL | I | 106.6 | FAIL | I | 136.9 | PASS | C | 387.0 | FAIL | I | 402.5 | PASS | C | 206.8 | FAIL | I | 102.4 | PASS | C | 412.4 |
| card_with_icon_title | FAIL | I | 118.0 | FAIL | I | 163.3 | FAIL | I | 440.3 | FAIL | I | 407.0 | PASS | C | 225.0 | FAIL | I | 94.7 | PASS | C | 445.0 |
| notification_card | FAIL | I | 121.8 | FAIL | I | 165.9 | PASS | C | 455.5 | FAIL | I | 390.9 | FAIL | I | 270.4 | FAIL | I | 93.7 | PASS | C | 471.1 |
| card_quote | FAIL | I | 120.9 | FAIL | I | 188.8 | PASS | C | 471.0 | FAIL | I | 420.6 | PASS | C | 292.9 | FAIL | I | 94.2 | PASS | C | 468.8 |
| list_simple_text | FAIL | I | 129.5 | FAIL | I | 192.0 | PASS | C | 401.7 | FAIL | I | 412.5 | PASS | C | 312.8 | FAIL | I | 97.0 | PASS | P | 464.8 |
| list_horizontal_cards | FAIL | I | 138.4 | FAIL | I | 197.8 | PASS | C | 449.9 | FAIL | I | 486.6 | PASS | C | 352.6 | FAIL | I | 77.8 | PASS | C | 488.2 |
| list_checkable_todo | FAIL | I | 140.8 | FAIL | I | 192.6 | PASS | P | 455.4 | FAIL | I | 516.2 | PASS | P | 354.7 | FAIL | I | 80.6 | PASS | P | 486.0 |
| list_settings_menu | FAIL | I | 140.8 | FAIL | I | 187.4 | PASS | P | 522.2 | FAIL | I | 574.1 | FAIL | I | 352.6 | FAIL | I | 83.6 | PASS | P | 526.4 |
| list_with_dividers | FAIL | I | 146.8 | FAIL | I | 210.9 | PASS | C | 538.9 | FAIL | I | 1046.2 | FAIL | I | 352.1 | FAIL | I | 83.9 | PASS | C | 509.4 |
| list_users_avatars | FAIL | I | 154.1 | FAIL | I | 208.8 | PASS | P | 572.7 | FAIL | I | 566.8 | PASS | C | 405.1 | FAIL | I | 85.4 | PASS | C | 546.9 |
| list_notifications | FAIL | I | 166.3 | FAIL | I | 190.0 | FAIL | I | 1074.2 | FAIL | I | 573.9 | PASS | P | 477.4 | FAIL | I | 83.8 | PASS | P | 594.8 |
| list_pricing_options | FAIL | I | 179.6 | FAIL | I | 192.4 | PASS | P | 572.9 | FAIL | I | 592.9 | FAIL | I | 461.1 | FAIL | I | 84.9 | PASS | P | 1116.4 |
| form_login | FAIL | I | 207.1 | FAIL | I | 191.1 | PASS | C | 577.3 | FAIL | I | 968.3 | PASS | C | 421.8 | FAIL | I | 86.1 | PASS | C | 583.6 |
| form_signup | FAIL | I | 237.1 | FAIL | I | 173.7 | PASS | C | 575.6 | FAIL | I | 944.4 | FAIL | I | 404.8 | FAIL | I | 92.6 | PASS | C | 1117.3 |
| form_contact | FAIL | I | 264.6 | FAIL | I | 176.4 | FAIL | I | 584.8 | FAIL | I | 938.1 | FAIL | I | 391.2 | FAIL | I | 102.0 | PASS | C | 1134.8 |
| form_reservation | FAIL | I | 279.9 | FAIL | I | 171.1 | PASS | C | 544.8 | FAIL | I | 537.7 | FAIL | I | 361.0 | FAIL | I | 101.9 | PASS | C | 575.6 |
| form_search_with_filter | FAIL | I | 275.3 | FAIL | I | 187.2 | PASS | C | 585.8 | FAIL | I | 532.8 | FAIL | I | 338.8 | FAIL | I | 103.7 | PASS | C | 591.1 |
| form_subscribe | FAIL | I | 279.4 | FAIL | I | 170.6 | PASS | C | 523.5 | FAIL | I | 465.0 | FAIL | I | 335.0 | FAIL | I | 101.2 | PASS | C | 523.3 |
| form_feedback_rating | FAIL | I | 265.7 | FAIL | I | 150.1 | PASS | C | 500.6 | FAIL | I | 448.0 | FAIL | I | 330.2 | FAIL | I | 128.5 | PASS | C | 529.4 |
| form_settings_toggles | FAIL | I | 262.6 | FAIL | I | 141.7 | PASS | C | 463.8 | FAIL | I | 365.1 | FAIL | I | 279.7 | FAIL | I | 148.2 | PASS | C | 513.2 |
| form_address | FAIL | I | 252.1 | FAIL | I | 142.9 | FAIL | I | 471.3 | FAIL | I | 345.1 | FAIL | I | 203.8 | FAIL | I | 148.8 | PASS | C | 474.8 |
| form_event | FAIL | I | 246.7 | FAIL | I | 117.2 | FAIL | I | 483.4 | FAIL | I | 391.6 | FAIL | I | 200.3 | FAIL | I | 149.8 | PASS | C | 511.7 |
| tabs_two_simple | FAIL | I | 221.4 | FAIL | I | 118.2 | PASS | C | 463.4 | FAIL | I | 387.4 | PASS | C | 197.1 | FAIL | I | 104.1 | PASS | C | 529.7 |
| tabs_three_panels | FAIL | I | 197.8 | FAIL | I | 118.8 | PASS | C | 495.9 | FAIL | I | 474.1 | PASS | C | 224.2 | FAIL | I | 101.9 | PASS | C | 530.9 |
| tabs_dashboard | FAIL | I | 169.4 | FAIL | I | 117.7 | PASS | C | 493.9 | FAIL | I | 462.9 | PASS | C | 278.8 | FAIL | I | 96.2 | PASS | C | 574.9 |
| tabs_settings_sections | FAIL | I | 153.3 | FAIL | I | 120.1 | PASS | C | 473.4 | FAIL | I | 593.1 | PASS | C | 319.4 | FAIL | I | 138.6 | PASS | C | 1198.6 |
| modal_confirm | FAIL | I | 153.3 | FAIL | I | 109.1 | FAIL | I | 503.9 | FAIL | I | 1806.5 | FAIL | I | 321.0 | FAIL | I | 137.6 | PASS | C | 1199.3 |
| modal_info | FAIL | I | 145.7 | FAIL | I | 109.3 | FAIL | I | 503.3 | FAIL | I | 2833.6 | FAIL | I | 347.9 | FAIL | I | 137.9 | PASS | C | 1199.1 |
| modal_signup_form | FAIL | I | 154.7 | FAIL | I | 109.8 | FAIL | I | 527.7 | FAIL | I | 2897.2 | FAIL | I | 417.6 | FAIL | I | 156.0 | PASS | C | 568.1 |
| modal_image_zoom | FAIL | I | 156.8 | FAIL | I | 108.1 | FAIL | I | 517.9 | FAIL | I | 2923.4 | FAIL | I | 449.5 | FAIL | I | 140.1 | PASS | C | 568.4 |
| nested_layout | FAIL | I | 153.3 | FAIL | I | 105.6 | PASS | C | 498.1 | FAIL | I | 2917.6 | PASS | C | 439.6 | FAIL | I | 140.5 | PASS | C | 568.8 |
| simple_column_layout | FAIL | I | 136.9 | FAIL | I | 105.2 | PASS | C | 475.9 | FAIL | I | 563.9 | PASS | C | 426.6 | FAIL | I | 158.7 | PASS | C | 570.8 |
| row_with_buttons | FAIL | I | 128.3 | FAIL | I | 138.3 | PASS | C | 469.3 | PASS | C | 585.0 | PASS | C | 464.2 | FAIL | I | 158.4 | PASS | C | 593.4 |
| simple_text | FAIL | I | 123.9 | FAIL | I | 128.8 | PASS | P | 413.8 | FAIL | I | 491.3 | PASS | C | 432.0 | FAIL | I | 151.2 | PASS | C | 558.1 |
| dashboard_overview | FAIL | I | 315.5 | FAIL | I | 133.9 | FAIL | I | 420.2 | FAIL | I | 1166.9 | FAIL | I | 423.8 | FAIL | I | 149.7 | PASS | C | 581.9 |
| product_detail | FAIL | I | 313.2 | FAIL | I | 134.6 | PASS | C | 421.5 | FAIL | I | 556.0 | FAIL | I | 377.8 | FAIL | I | 107.2 | PASS | C | 592.8 |
| chat_message_thread | FAIL | I | 321.8 | FAIL | I | 131.9 | PASS | C | 431.8 | FAIL | I | 2347.0 | FAIL | I | 382.2 | FAIL | I | 106.9 | PASS | C | 1139.8 |
| profile_screen | FAIL | I | 322.6 | FAIL | I | 133.9 | PASS | C | 459.7 | FAIL | I | 1783.5 | FAIL | I | 368.3 | FAIL | I | 110.7 | PASS | C | 1191.1 |
| shopping_cart_summary | FAIL | I | 334.7 | FAIL | I | 143.0 | PASS | P | 479.4 | FAIL | I | 2509.3 | FAIL | I | 358.3 | FAIL | I | 110.6 | PASS | P | 1585.0 |
| settings_account_screen | FAIL | I | 334.4 | FAIL | I | 147.2 | PASS | C | 516.7 | FAIL | I | 1131.9 | FAIL | I | 340.3 | FAIL | I | 107.4 | FAIL | P | 1581.2 |
| media_gallery | FAIL | I | 341.6 | FAIL | I | 146.8 | PASS | C | 519.9 | FAIL | I | 571.6 | PASS | C | 362.0 | FAIL | I | 136.4 | PASS | C | 1610.3 |
| media_player_with_controls | FAIL | I | 352.4 | FAIL | I | 146.3 | PASS | C | 558.9 | FAIL | I | 586.5 | FAIL | I | 386.8 | FAIL | I | 118.8 | FAIL | P | 581.0 |
| onboarding_screen | FAIL | I | 360.0 | FAIL | I | 113.2 | PASS | C | 703.7 | FAIL | I | 559.4 | FAIL | I | 365.2 | FAIL | I | 120.4 | PASS | C | 559.9 |
| pricing_table | FAIL | I | 380.1 | FAIL | I | 123.7 | PASS | C | 774.8 | FAIL | I | 505.0 | PASS | C | 445.0 | FAIL | I | 166.1 | PASS | C | 1049.4 |
| appointment_booking | FAIL | I | 198.2 | FAIL | I | 119.8 | FAIL | I | 588.5 | FAIL | I | 406.9 | FAIL | I | 403.8 | FAIL | I | 169.5 | PASS | C | 1092.1 |

# Summary Statistics

Table 4: Aggregate benchmark statistics per model.

| Metric | Llama 3.1 8B | Qwen 3 1.7B | Qwen 3 14B | Qwen 3 8B | Qwen3 30B MoE NVFP4 | Llama 3.2 1B | Gemma 4 26B MoE |
|---|---|---|---|---|---|---|---|
| Schema Accuracy | 0% | 3% | 82% | 7% | 59% | 0% | 97% |
| Semantic Accuracy | 0% | 3% | 74% | 7% | 51% | 0% | 89% |
| Avg Inference Time | 176.0 s | 128.3 s | 434.1 s | 670.2 s | 283.4 s | 118.9 s | 580.5 s |
| Median Inference Time | 156.8 s | 120.8 s | 455.4 s | 412.5 s | 279.7 s | 116.2 s | 509.4 s |
| Output Throughput | 4.0 tok/s | 8.9 tok/s | 2.3 tok/s | 2.8 tok/s | 5.2 tok/s | 17.7 tok/s | 4.5 tok/s |
| ICL Context Size | 8,739 tok | 8,770 tok | 8,770 tok | 8,770 tok | 8,770 tok | 8,739 tok | 9,181 tok |
| Prefill Throughput (peak) | 0 tok/s | 0 tok/s | 0 tok/s | 0 tok/s | 0 tok/s | 0 tok/s | 0 tok/s |
| Decode Throughput (peak) | 0 tok/s | 0 tok/s | 0 tok/s | 0 tok/s | 0 tok/s | 0 tok/s | 0 tok/s |
| Prefix Cache Hit Rate | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| Model Memory | 0.0 GiB | 0.0 GiB | 0.0 GiB | 0.0 GiB | 0.0 GiB | 0.0 GiB | 0.0 GiB |
| KV Cache Available | 0.0 GiB | 0.0 GiB | 0.0 GiB | 0.0 GiB | 0.0 GiB | 0.0 GiB | 0.0 GiB |
| Total Input Tokens | 638,958 | 641,239 | 641,239 | 641,239 | 641,239 | 638,958 | 671,278 |
| Total Output Tokens | 52,013 | 83,841 | 74,108 | 135,616 | 107,137 | 153,241 | 191,860 |
| Samples Evaluated | 73 | 73 | 73 | 73 | 73 | 73 | 73 |

# Hardware and Software Configuration

Table 5: System configuration.

| Component | Specification |
|---|---|
| GPU | Apple M1 Max |
| Memory | 64 GiB RAM |
| Architecture | arm64 |
| Engine | llama.cpp |
| Inspect AI | $\geq$0.3.217 |
| OS | macOS (kernel 25.4.0) |

# Methodology

## Evaluation Pipeline

Each model was evaluated through the following pipeline:

1. **Model serving:** llama.cpp starts an OpenAI-compatible API server with the model
2. **System prompt injection:** The full A2UI v0.9 JSON Schema and component catalog are injected as a system prompt (~8,000 tokens)
3. **Generation:** The model generates structured JSON output for each evaluation prompt in a *single-turn* setting — one user prompt yields one assistant response, with no follow-up turns, no self-correction loop, and no schema-error retries
4. **Schema validation:** Programmatic validation against A2UI schema and integrity rules
5. **Semantic grading:** An external judge (Claude Code (interactive, human-supervised)) assigns C/P/I per sample using the rubric in Appendix **??**

## Limitations

- **LLM-based grading:** Grades are assigned by an LLM judge (Claude Code (interactive, human-supervised)), not a domain expert. A human evaluator or multi-judge consensus would reduce single-grader bias.
- **Test dataset:** Results use a test dataset. The full encrypted production dataset (~100 samples) provides more comprehensive coverage.

- **Single-run:** Each model was benchmarked once. Multiple runs would provide statistical significance for timing measurements.
- **Single-turn generation:** Every sample is a one-shot generation. The model receives the A2UI system prompt plus the user's task, and its first response is what we score. A real agent could re-emit JSON after seeing a schema error, or split a complex UI across several turns; we do not exercise either capability here. Schema-invalid outputs therefore count as incorrect even when a follow-up turn would have repaired them.

### Output-Format Strictness and Small-Model Failures

The schema validator is the upstream A2UI Python SDK (`a2ui.parser.parser.parse_response` + `a2ui.schema.manager.A2uiSchemaManager`, Apache 2.0, Copyright Google LLC, vendored in `A2UI/eval/a2ui_eval/scorers.py`). It only recognises payloads wrapped in literal `<a2ui-json>...</a2ui-jso` tags and then runs the full A2UI v0.9 JSON schema against the parsed message graph. There is no markdown fence fallback, no automatic unwrap of "thinking" preambles, and no acceptance of structurally valid JSON in any other envelope.

The validator's strictness is a faithful gate for a renderer that would consume the output, but it explains most of the near-zero scores in the small-model rows. Spot-checking sample 0 (`text_heading`) for each failing model shows the issue is *output format compliance*, not context size — inputs were 8.7–9.7k tokens against a 16,384-token serving limit, and no completion approached the output-token cap. The recurring patterns:

- **Llama 3.2 1B** (both engines): emitted a JSON Schema *definition* (`$schema: https://json-schema.org` describing what a Text component should look like, instead of an A2UI *instance*. No `<a2ui-json>` tags appear in the response at all.
- **Qwen 3 1.7B** (vLLM): used the tags but produced a malformed structure — `version` placed inside `updateComponents`, plus a duplicated nested `updateComponents` block.
- **Qwen 3 8B** (vLLM): wrapped otherwise-correct A2UI JSON in `"'json` markdown fences instead of `<a2ui-json>` tags. The validator therefore reports "tags not found" and the run scores near zero despite the content largely being valid.
- **Nemotron 3 Nano 30B** (vLLM): a reasoning model that emits a `<think>...</think>` preamble followed by the final answer, but the final answer double-wraps `createSurface` inside another `createSurface` (`{createSurface:{version:v0.9,createSurface:{...}}}`).

These are not capability ceilings so much as instruction-following and format-compliance gaps that a single follow-up turn or a more lenient extractor (e.g. accepting any fenced JSON block whose root matches the A2UI message schema) would close in many cases. We deliberately do not apply such fallbacks here so that scores reflect what a renderer running the production parser would actually accept.

## Conclusion

This benchmark demonstrates that local inference on Apple M1 Max is viable for A2UI protocol evaluation across multiple model architectures. The results provide a baseline for comparing model capabilities in structured UI generation tasks.

**Key findings:**

- MoE models (Gemma 4) offer the best throughput-to-parameter ratio on memory-bandwidth-limited hardware
- Dense models with fewer parameters achieve lower latency per sample
- All tested models benefit from the engine's prefix caching for the shared A2UI system prompt

## Grading Methodology

### Judge

The semantic grade column (C/P/I) in this report was assigned by:

> **Claude Code (interactive, human-supervised)**

### Grading sessions

- Graded in Claude Code session 2026-05-12T05:50:24Z
- Graded in Claude Code session 2026-05-12T07:24:34Z; added Gemma 4 26B

### Rubric

Each generated A2UI message was judged against the prompt's target description and assigned one of three grades:

- **C (Correct)** – the response satisfies the target. Optional/cosmetic additions and ID/label variations are accepted.
- **P (Partial)** – the response is substantively right but has minor variations (e.g. wrong wrapper format around correct components, slight punctuation differences).
- **I (Incorrect)** – the response is missing required components or has substantive errors (wrong variant, missing fields, broken structure, fabricated component names).

The rubric is the same one used by the production Inspect-AI grader (`A2UI/eval/tasks.py:GRADER_INSTRUCT`). Variations in capitalization, punctuation, component IDs, label synonyms, and data-binding paths are explicitly allowed.

## Per-Sample Grading Rationales

Each row references one of the 73 evaluation prompts by its dataset **Sample** slug. The slug is consistent across all per-model tables, so a given row identifies the same prompt for every model. The rationale is the one-line note recorded at grading time; rationales over 240 characters are truncated (the full text is in the result JSON).

For reference, the first prompt `text_heading` reads in full:

> *Render a single Text component with the heading 'Settings' using the h1 variant.*

with grading target:

> A valid A2UI response with a Column root containing a single Text component whose variant is h1 and whose content is 'Settings'.

## Llama 3.1 8B

| Sample | Grade | |
|---|---|---|
| text_heading | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| text_caption | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| text_markdown | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| image_avatar | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| image_header_banner | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| icon_inline | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| video_player | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| audio_player | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| divider_horizontal | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| button_primary | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| button_borderless | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| textfield_password | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| text_then_button | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| image_caption | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| icon_label | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| two_buttons_row | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| heading_paragraph | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| divider_between_texts | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| checkbox_simple | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| slider_volume | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| date_picker_only | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| choice_picker_chips | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| textfield_search | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| textfield_multiline | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| card_with_text | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| card_with_image_text_button | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| card_profile_summary | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| card_stats | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| card_with_divider | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| card_with_icon_title | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| notification_card | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| card_quote | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| list_simple_text | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| list_horizontal_cards | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| list_checkable_todo | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| list_settings_menu | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| list_with_dividers | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| list_users_avatars | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| list_notifications | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| list_pricing_options | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| form_login | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| form_signup | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| form_contact | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| form_reservation | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| form_search_with_filter | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| form_subscribe | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| form_feedback_rating | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| form_settings_toggles | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| form_address | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| form_event | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| tabs_two_simple | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| tabs_three_panels | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| tabs_dashboard | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| tabs_settings_sections | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| modal_confirm | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| modal_info | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| modal_signup_form | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| modal_image_zoom | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| nested_layout | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| simple_column_layout | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| row_with_buttons | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| simple_text | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| dashboard_overview | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| product_detail | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| chat_message_thread | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| profile_screen | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |
| shopping_cart_summary | I | Schema failed; completion either lacked <a2ui-json> tags or used invalid structure |

## Qwen 3 1.7B

| Sample | Grade | |
|---|---|---|
| text_heading | C | Schema passed; produced the requested component |
| text_caption | I | Schema failed with substantive structural errors (missing message envelope, missi |
| text_markdown | I | Schema failed with substantive structural errors (missing message envelope, missi |
| image_avatar | I | Schema failed with substantive structural errors (missing message envelope, missi |
| image_header_banner | I | Schema failed with substantive structural errors (missing message envelope, missi |
| icon_inline | I | Schema failed with substantive structural errors (missing message envelope, missi |
| video_player | C | Schema passed; produced the requested component |
| audio_player | I | Schema failed with substantive structural errors (missing message envelope, missi |
| divider_horizontal | I | Schema failed with substantive structural errors (missing message envelope, missi |
| button_primary | I | Schema failed with substantive structural errors (missing message envelope, missi |
| button_borderless | I | Schema failed with substantive structural errors (missing message envelope, missi |
| textfield_password | I | Schema failed with substantive structural errors (missing message envelope, missi |
| text_then_button | I | Schema failed with substantive structural errors (missing message envelope, missi |
| image_caption | I | Schema failed with substantive structural errors (missing message envelope, missi |
| icon_label | I | Schema failed with substantive structural errors (missing message envelope, missi |
| two_buttons_row | I | Schema failed with substantive structural errors (missing message envelope, missi |
| heading_paragraph | I | Schema failed with substantive structural errors (missing message envelope, missi |
| divider_between_texts | I | Schema failed with substantive structural errors (missing message envelope, missi |
| checkbox_simple | I | Schema failed with substantive structural errors (missing message envelope, missi |
| slider_volume | I | Schema failed with substantive structural errors (missing message envelope, missi |
| date_picker_only | I | Schema failed with substantive structural errors (missing message envelope, missi |
| choice_picker_chips | I | Schema failed with substantive structural errors (missing message envelope, missi |
| textfield_search | I | Schema failed with substantive structural errors (missing message envelope, missi |
| textfield_multiline | I | Schema failed with substantive structural errors (missing message envelope, missi |
| card_with_text | I | Schema failed with substantive structural errors (missing message envelope, missi |
| card_with_image_text_button | I | Schema failed with substantive structural errors (missing message envelope, missi |
| card_profile_summary | I | Schema failed with substantive structural errors (missing message envelope, missi |
| card_stats | I | Schema failed with substantive structural errors (missing message envelope, missi |
| card_with_divider | I | Schema failed with substantive structural errors (missing message envelope, missi |
| card_with_icon_title | I | Schema failed with substantive structural errors (missing message envelope, missi |
| notification_card | I | Schema failed with substantive structural errors (missing message envelope, missi |
| card_quote | I | Schema failed with substantive structural errors (missing message envelope, missi |
| list_simple_text | I | Schema failed with substantive structural errors (missing message envelope, missi |
| list_horizontal_cards | I | Schema failed with substantive structural errors (missing message envelope, missi |
| list_checkable_todo | I | Schema failed with substantive structural errors (missing message envelope, missi |
| list_settings_menu | I | Schema failed with substantive structural errors (missing message envelope, missi |
| list_with_dividers | I | Schema failed with substantive structural errors (missing message envelope, missi |
| list_users_avatars | I | Schema failed with substantive structural errors (missing message envelope, missi |
| list_notifications | I | Schema failed with substantive structural errors (missing message envelope, missi |
| list_pricing_options | I | Schema failed with substantive structural errors (missing message envelope, missi |
| form_login | I | Schema failed with substantive structural errors (missing message envelope, missi |
| form_signup | I | Schema failed with substantive structural errors (missing message envelope, missi |
| form_contact | I | Schema failed with substantive structural errors (missing message envelope, missi |
| form_reservation | I | Schema failed with substantive structural errors (missing message envelope, missi |
| form_search_with_filter | I | Schema failed with substantive structural errors (missing message envelope, missi |
| form_subscribe | I | Schema failed with substantive structural errors (missing message envelope, missi |
| form_feedback_rating | I | Schema failed with substantive structural errors (missing message envelope, missi |
| form_settings_toggles | I | Schema failed with substantive structural errors (missing message envelope, missi |
| form_address | I | Schema failed with substantive structural errors (missing message envelope, missi |
| form_event | I | Schema failed with substantive structural errors (missing message envelope, missi |
| tabs_two_simple | I | Schema failed with substantive structural errors (missing message envelope, missi |
| tabs_three_panels | I | Schema failed with substantive structural errors (missing message envelope, missi |
| tabs_dashboard | I | Schema failed with substantive structural errors (missing message envelope, missi |
| tabs_settings_sections | I | Schema failed with substantive structural errors (missing message envelope, missi |
| modal_confirm | I | Schema failed with substantive structural errors (missing message envelope, missi |
| modal_info | I | Schema failed with substantive structural errors (missing message envelope, missi |
| modal_signup_form | I | Schema failed with substantive structural errors (missing message envelope, missi |
| modal_image_zoom | I | Schema failed with substantive structural errors (missing message envelope, missi |
| nested_layout | I | Schema failed with substantive structural errors (missing message envelope, missi |
| simple_column_layout | I | Schema failed with substantive structural errors (missing message envelope, missi |
| row_with_buttons | I | Schema failed with substantive structural errors (missing message envelope, missi |
| simple_text | I | Schema failed with substantive structural errors (missing message envelope, missi |
| dashboard_overview | I | Schema failed with substantive structural errors (missing message envelope, missi |
| product_detail | I | Schema failed with substantive structural errors (missing message envelope, missi |
| chat_message_thread | I | Schema failed with substantive structural errors (missing message envelope, missi |
| profile_screen | I | Schema failed with substantive structural errors (missing message envelope, missi |
| shopping_cart_summary | I | Schema failed with substantive structural errors (missing message envelope, missi |

## Qwen 3 14B

| Sample | Grade | |
|---|---|---|
| text_heading | C | Schema passed and completion produced the required components with correct pr |
| text_caption | C | Schema passed and completion produced the required components with correct pr |
| text_markdown | C | Schema passed and completion produced the required components with correct pr |
| image_avatar | C | Schema passed and completion produced the required components with correct pr |
| image_header_banner | C | Schema passed and completion produced the required components with correct pr |
| icon_inline | C | Schema passed and completion produced the required components with correct pr |
| video_player | C | Schema passed and completion produced the required components with correct pr |
| audio_player | C | Schema passed and completion produced the required components with correct pr |
| divider_horizontal | C | Schema passed and completion produced the required components with correct pr |
| button_primary | C | Schema passed and completion produced the required components with correct pr |
| button_borderless | C | Schema passed and completion produced the required components with correct pr |
| textfield_password | I | Schema failed with substantive structural |
| text_then_button | C | Schema passed and completion produced the required components with correct pr |
| image_caption | C | Schema passed and completion produced the required components with correct pr |
| icon_label | C | Schema passed and completion produced the required components with correct pr |
| two_buttons_row | C | Schema passed and completion produced the required components with correct pr |
| heading_paragraph | C | Schema passed and completion produced the required components with correct pr |
| divider_between_texts | C | Schema passed and completion produced the required components with correct pr |
| checkbox_simple | C | Schema passed and completion produced the required components with correct pr |
| slider_volume | C | Schema passed and completion produced the required components with correct pr |
| date_picker_only | I | Schema failed with substantive structural |
| choice_picker_chips | C | Schema passed and completion produced the required components with correct pr |
| textfield_search | C | Schema passed and completion produced the required components with correct pr |
| textfield_multiline | C | Schema passed and completion produced the required components with correct pr |
| card_with_text | C | Schema passed and completion produced the required components with correct pr |
| card_with_image_text_button | C | Schema passed and completion produced the required components with correct pr |
| card_profile_summary | C | Schema passed and completion produced the required components with correct pr |
| card_stats | C | Schema passed and completion produced the required components with correct pr |
| card_with_divider | C | Schema passed and completion produced the required components with correct pr |
| card_with_icon_title | I | Schema failed with substantive structural |
| notification_card | C | Schema passed and completion produced the required components with correct pr |
| card_quote | C | Schema passed and completion produced the required components with correct pr |
| list_simple_text | C | Schema passed and completion produced the required components with correct pr |
| list_horizontal_cards | C | Schema passed and completion produced the required components with correct pr |
| list_checkable_todo | P | Schema valid, but used Column container instead |
| list_settings_menu | P | Schema valid, but used Column container instead |
| list_with_dividers | C | Schema passed and completion produced the required components with correct pr |
| list_users_avatars | P | Schema valid, but used Column container instead |
| list_notifications | I | Schema failed with substantive structural |
| list_pricing_options | P | Schema valid, but used Column container instead |
| form_login | C | Schema passed and completion produced the required components with correct pr |
| form_signup | C | Schema passed and completion produced the required components with correct pr |
| form_contact | I | Schema failed with substantive structural |
| form_reservation | C | Schema passed and completion produced the required components with correct pr |
| form_search_with_filter | C | Schema passed and completion produced the required components with correct pr |
| form_subscribe | C | Schema passed and completion produced the required components with correct pr |
| form_feedback_rating | C | Schema passed and completion produced the required components with correct pr |
| form_settings_toggles | C | Schema passed and completion produced the required components with correct pr |
| form_address | I | Schema failed with substantive structural |
| form_event | I | Schema failed with substantive structural |
| tabs_two_simple | C | Schema passed and completion produced the required components with correct pr |
| tabs_three_panels | C | Schema passed and completion produced the required components with correct pr |
| tabs_dashboard | C | Schema passed and completion produced the required components with correct pr |
| tabs_settings_sections | C | Schema passed and completion produced the required components with correct pr |
| modal_confirm | I | Schema failed with substantive structural |
| modal_info | I | Schema failed with substantive structural |
| modal_signup_form | I | Schema failed with substantive structural |
| modal_image_zoom | I | Schema failed with substantive structural |
| nested_layout | C | Schema passed and completion produced the required components with correct pr |
| simple_column_layout | C | Schema passed and completion produced the required components with correct pr |
| row_with_buttons | C | Schema passed and completion produced the required components with correct pr |
| simple_text | P | Schema valid, but produced a bare Text without the r |
| dashboard_overview | I | Schema failed with substantive structural |
| product_detail | C | Schema passed and completion produced the required components with correct pr |
| chat_message_thread | C | Schema passed and completion produced the required components with correct pr |
| profile_screen | C | Schema passed and completion produced the required components with correct pr |
| shopping_cart_summary | P | Schema valid, but omitted the inner List container |

## Qwen 3 8B

| Sample | Grade | |
|---|---|---|
| text_heading | I | Schema failed with substantive str |
| text_caption | I | Schema failed with substantive str |
| text_markdown | I | Schema failed with substantive str |
| image_avatar | C | Schema passed; produced the required component(s) with correct properties and s |
| image_header_banner | I | Schema failed with substantive str |
| icon_inline | C | Schema passed; produced the required component(s) with correct properties and s |
| video_player | I | Schema failed with substantive str |
| audio_player | I | Schema failed with substantive str |
| divider_horizontal | I | Schema failed with substantive str |
| button_primary | I | Schema failed with substantive str |
| button_borderless | I | Schema failed with substantive str |
| textfield_password | I | Schema failed with substantive str |
| text_then_button | I | Schema failed with substantive str |
| image_caption | I | Schema failed with substantive str |
| icon_label | I | Schema failed with substantive str |
| two_buttons_row | I | Schema failed with substantive str |
| heading_paragraph | I | Schema failed with substantive str |
| divider_between_texts | C | Schema passed; produced the required component(s) with correct properties and s |
| checkbox_simple | I | Schema failed with substantive str |
| slider_volume | I | Schema failed with substantive str |
| date_picker_only | C | Schema passed; produced the required component(s) with correct properties and s |
| choice_picker_chips | I | Schema failed with substantive str |
| textfield_search | I | Schema failed with substantive str |
| textfield_multiline | I | Schema failed with substantive str |
| card_with_text | I | Schema failed with substantive str |
| card_with_image_text_button | I | Schema failed with substantive str |
| card_profile_summary | I | Schema failed with substantive str |
| card_stats | I | Schema failed with substantive str |
| card_with_divider | I | Schema failed with substantive str |
| card_with_icon_title | I | Schema failed with substantive str |
| notification_card | I | Schema failed with substantive str |
| card_quote | I | Schema failed with substantive str |
| list_simple_text | I | Schema failed with substantive str |
| list_horizontal_cards | I | Schema failed with substantive str |
| list_checkable_todo | I | Schema failed with substantive str |
| list_settings_menu | I | Schema failed with substantive str |
| list_with_dividers | I | Schema failed with substantive str |
| list_users_avatars | I | Schema failed with substantive str |
| list_notifications | I | Schema failed with substantive str |
| list_pricing_options | I | Schema failed with substantive str |
| form_login | I | Schema failed with substantive str |
| form_signup | I | Schema failed with substantive str |
| form_contact | I | Schema failed with substantive str |
| form_reservation | I | Schema failed with substantive str |
| form_search_with_filter | I | Schema failed with substantive str |
| form_subscribe | I | Schema failed with substantive str |
| form_feedback_rating | I | Schema failed with substantive str |
| form_settings_toggles | I | Schema failed with substantive str |
| form_address | I | Schema failed with substantive str |
| form_event | I | Schema failed with substantive str |
| tabs_two_simple | I | Schema failed with substantive str |
| tabs_three_panels | I | Schema failed with substantive str |
| tabs_dashboard | I | Schema failed with substantive str |
| tabs_settings_sections | I | Schema failed with substantive str |
| modal_confirm | I | Schema failed with substantive str |
| modal_info | I | Schema failed with substantive str |
| modal_signup_form | I | Schema failed with substantive str |
| modal_image_zoom | I | Schema failed with substantive str |
| nested_layout | I | Schema failed with substantive str |
| simple_column_layout | I | Schema failed with substantive str |
| row_with_buttons | C | Schema passed; produced the required component(s) with correct properties and s |
| simple_text | I | Schema failed with substantive str |
| dashboard_overview | I | Schema failed with substantive str |
| product_detail | I | Schema failed with substantive str |
| chat_message_thread | I | Schema failed with substantive str |
| profile_screen | I | Schema failed with substantive str |
| shopping_cart_summary | I | Schema failed with substantive str |

## Qwen3 30B MoE NVFP4

| Sample | Grade | |
|---|---|---|
| text_heading | C | Schema passed and completion produced the required components with correct p |
| text_caption | C | Schema passed and completion produced the required components with correct p |
| text_markdown | C | Schema passed and completion produced the required components with correct p |
| image_avatar | C | Schema passed and completion produced the required components with correct p |
| image_header_banner | C | Schema passed and completion produced the required components with correct p |
| icon_inline | C | Schema passed and completion produced the required components with correct p |
| video_player | C | Schema passed and completion produced the required components with correct p |
| audio_player | C | Schema passed and completion produced the required components with correct p |
| divider_horizontal | C | Schema passed and completion produced the required components with correct p |
| button_primary | I | Schema failed with substantive structural |
| button_borderless | C | Schema passed and completion produced the required components with correct p |
| textfield_password | C | Schema passed and completion produced the required components with correct p |
| text_then_button | C | Schema passed and completion produced the required components with correct p |
| image_caption | P | Schema valid, but caption Text used variant="bo |
| icon_label | C | Schema passed and completion produced the required components with correct p |
| two_buttons_row | C | Schema passed and completion produced the required components with correct p |
| heading_paragraph | C | Schema passed and completion produced the required components with correct p |
| divider_between_texts | P | Schema valid, but used Row container instead of the requested Column for sequer |
| checkbox_simple | C | Schema passed and completion produced the required components with correct p |
| slider_volume | C | Schema passed and completion produced the required components with correct p |
| date_picker_only | C | Schema passed and completion produced the required components with correct p |
| choice_picker_chips | I | Schema failed with substantive structural |
| textfield_search | P | Schema valid, but used label="Search..." instead of |
| textfield_multiline | I | Schema failed with substantive structural |
| card_with_text | C | Schema passed and completion produced the required components with correct p |
| card_with_image_text_button | P | Schema valid, but omitted the Card wrapper around |
| card_profile_summary | C | Schema passed and completion produced the required components with correct p |
| card_stats | I | Schema failed with substantive structural |
| card_with_divider | C | Schema passed and completion produced the required components with correct p |
| card_with_icon_title | C | Schema passed and completion produced the required components with correct p |
| notification_card | I | Schema failed with substantive structural |
| card_quote | C | Schema passed and completion produced the required components with correct p |
| list_simple_text | C | Schema passed and completion produced the required components with correct p |
| list_horizontal_cards | C | Schema passed and completion produced the required components with correct p |
| list_checkable_todo | P | Schema valid, but used a Column container instead of the request |
| list_settings_menu | I | Schema failed with substantive structural |
| list_with_dividers | I | Schema failed with substantive structural |
| list_users_avatars | C | Schema passed and completion produced the required components with correct p |
| list_notifications | P | Schema valid, but used a Column container instea |
| list_pricing_options | I | Schema failed with substantive structural |
| form_login | C | Schema passed and completion produced the required components with correct p |
| form_signup | I | Schema failed with substantive structural |
| form_contact | I | Schema failed with substantive structural |
| form_reservation | I | Schema failed with substantive structural |
| form_search_with_filter | I | Schema failed with substantive structural |
| form_subscribe | I | Schema failed with substantive structural |
| form_feedback_rating | I | Schema failed with substantive structural |
| form_settings_toggles | I | Schema failed with substantive structural |
| form_address | I | Schema failed with substantive structural |
| form_event | I | Schema failed with substantive structural |
| tabs_two_simple | C | Schema passed and completion produced the required components with correct p |
| tabs_three_panels | C | Schema passed and completion produced the required components with correct p |
| tabs_dashboard | C | Schema passed and completion produced the required components with correct p |
| tabs_settings_sections | C | Schema passed and completion produced the required components with correct p |
| modal_confirm | I | Schema failed with substantive structural |
| modal_info | I | Schema failed with substantive structural |
| modal_signup_form | I | Schema failed with substantive structural |
| modal_image_zoom | I | Schema failed with substantive structural |
| nested_layout | C | Schema passed and completion produced the required components with correct p |
| simple_column_layout | C | Schema passed and completion produced the required components with correct p |
| row_with_buttons | C | Schema passed and completion produced the required components with correct p |
| simple_text | C | Schema passed and completion produced the required components with correct p |
| dashboard_overview | I | Schema failed with substantive structural |
| product_detail | I | Schema failed with substantive structural |
| chat_message_thread | I | Schema failed with substantive structural |
| profile_screen | I | Schema failed with substantive structural |
| shopping_cart_summary | I | Schema failed with substantive structural |

## Llama 3.2 1B

| Sample | Grade | |
|---|---|---|
| text_heading | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| text_caption | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| text_markdown | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| image_avatar | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| image_header_banner | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| icon_inline | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| video_player | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| audio_player | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| divider_horizontal | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| button_primary | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| button_borderless | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| textfield_password | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| text_then_button | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| image_caption | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| icon_label | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| two_buttons_row | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| heading_paragraph | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| divider_between_texts | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| checkbox_simple | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| slider_volume | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| date_picker_only | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| choice_picker_chips | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| textfield_search | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| textfield_multiline | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| card_with_text | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| card_with_image_text_button | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| card_profile_summary | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| card_stats | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| card_with_divider | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| card_with_icon_title | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| notification_card | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| card_quote | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| list_simple_text | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| list_horizontal_cards | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| list_checkable_todo | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| list_settings_menu | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| list_with_dividers | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| list_users_avatars | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| list_notifications | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| list_pricing_options | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| form_login | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| form_signup | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| form_contact | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| form_reservation | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| form_search_with_filter | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| form_subscribe | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| form_feedback_rating | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| form_settings_toggles | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| form_address | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| form_event | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| tabs_two_simple | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| tabs_three_panels | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| tabs_dashboard | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| tabs_settings_sections | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| modal_confirm | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| modal_info | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| modal_signup_form | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| modal_image_zoom | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| nested_layout | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| simple_column_layout | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| row_with_buttons | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| simple_text | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| dashboard_overview | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| product_detail | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| chat_message_thread | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| profile_screen | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |
| shopping_cart_summary | I | Off-protocol: completion did not produce a valid <a2ui-json> updateSurface mess |

## Gemma 4 26B MoE

| Sample | Grade | |
|---|---|---|
| text_heading | C | Schema passed and completion produced the required c |
| text_caption | C | Schema passed and completion produced the required c |
| text_markdown | C | Schema passed and completion produced the required c |
| image_avatar | C | Schema passed and completion produced the required c |
| image_header_banner | C | Schema passed and completion produced the required c |
| icon_inline | C | Schema passed and completion produced the required c |
| video_player | C | Schema passed and completion produced the required c |
| audio_player | C | Schema passed and completion produced the required c |
| divider_horizontal | C | Schema passed and completion produced the required c |
| button_primary | C | Schema passed and completion produced the required c |
| button_borderless | C | Schema passed and completion produced the required c |
| textfield_password | C | Schema passed and completion produced the required c |
| text_then_button | C | Schema passed and completion produced the required c |
| image_caption | C | Schema passed and completion produced the required c |
| icon_label | C | Schema passed and completion produced the required c |
| two_buttons_row | C | Schema passed and completion produced the required c |
| heading_paragraph | C | Schema passed and completion produced the required c |
| divider_between_texts | C | Schema passed and completion produced the required c |
| checkbox_simple | C | Schema passed and completion produced the required c |
| slider_volume | C | Schema passed and completion produced the required c |
| date_picker_only | C | Schema passed and completion produced the required c |
| choice_picker_chips | C | Schema passed and completion produced the required c |
| textfield_search | C | Schema passed and completion produced the required c |
| textfield_multiline | C | Schema passed and completion produced the required c |
| card_with_text | C | Schema passed and completion produced the required c |
| card_with_image_text_button | C | Schema passed and completion produced the required c |
| card_profile_summary | C | Schema passed and completion produced the required c |
| card_stats | C | Schema passed and completion produced the required c |
| card_with_divider | C | Schema passed and completion produced the required c |
| card_with_icon_title | C | Schema passed and completion produced the required c |
| notification_card | C | Schema passed and completion produced the required c |
| card_quote | C | Schema passed and completion produced the required c |
| list_simple_text | P | Schema valid, but use |
| list_horizontal_cards | C | Schema passed and completion produced the required c |
| list_checkable_todo | P | Schema valid, but use |
| list_settings_menu | P | Schema valid, but use |
| list_with_dividers | C | Schema passed and completion produced the required c |
| list_users_avatars | C | Schema passed and completion produced the required c |
| list_notifications | P | Schema valid, but use |
| list_pricing_options | P | Schema valid, but use |
| form_login | C | Schema passed and completion produced the required c |
| form_signup | C | Schema passed and completion produced the required c |
| form_contact | C | Schema passed and completion produced the required c |
| form_reservation | C | Schema passed and completion produced the required c |
| form_search_with_filter | C | Schema passed and completion produced the required c |
| form_subscribe | C | Schema passed and completion produced the required c |
| form_feedback_rating | C | Schema passed and completion produced the required c |
| form_settings_toggles | C | Schema passed and completion produced the required c |
| form_address | C | Schema passed and completion produced the required c |
| form_event | C | Schema passed and completion produced the required c |
| tabs_two_simple | C | Schema passed and completion produced the required c |
| tabs_three_panels | C | Schema passed and completion produced the required c |
| tabs_dashboard | C | Schema passed and completion produced the required c |
| tabs_settings_sections | C | Schema passed and completion produced the required c |
| modal_confirm | C | Schema passed and completion produced the required c |
| modal_info | C | Schema passed and completion produced the required c |
| modal_signup_form | C | Schema passed and completion produced the required c |
| modal_image_zoom | C | Schema passed and completion produced the required c |
| nested_layout | C | Schema passed and completion produced the required c |
| simple_column_layout | C | Schema passed and completion produced the required c |
| row_with_buttons | C | Schema passed and completion produced the required c |
| simple_text | C | Schema passed and completion produced the required c |
| dashboard_overview | C | Schema passed and completion produced the required c |
| product_detail | C | Schema passed and completion produced the required c |
| chat_message_thread | C | Schema passed and completion produced the required c |
| profile_screen | C | Schema passed and completion produced the required c |
| shopping_cart_summary | P | Schema valid, but omitt |