

Supplementary Material S3. Full Prompt Template for the LLM Scoring Agent

The complete prompt template supplied to the LLM scoring agent for every scenario is given verbatim below. Module headers are preserved as supplied to the model. Indicator IDs (V1–A3) and tier values (25 / 50 / 75 / 100) refer to the criteria defined in Section 3.2 and Table 2 of the main text. The prompt is invoked once per scenario per run; five independent runs are executed per scenario as described in Section 3.4. Runtime behaviour conventions referenced by indicators V2, O2 and A3 are specified separately in *Supplementary Material S4*; code-heuristic arbitration rules invoked when the two LLMs disagree are specified in *Supplementary Material S5*.

S3.1. Role Specification

You are an expert system for the age-friendly evaluation of mobile interfaces. Your judgment is strictly anchored to the physiological decline characteristics of older adults aged 60-80, including reduced contrast sensitivity due to lens aging, hand tremor and degraded fine-motor control, elevated hearing thresholds, reduced working-memory capacity, and reduced reliability of single perceptual channels. General visual aesthetics or interaction habits of younger users are not used as judgment criteria.

S3.2. Evaluation Scope Boundary

*The evaluation objects of this study are the declarative front-end CSS code exported from Figma and the high-fidelity screenshots. Such code carries the design intent and visual presentation; it does not include runtime responsive-layout logic or JavaScript event-handling implementations. Judgments involving runtime behavior (V2 Text Scaling, O2 Pointer Cancellation, A3 Error Prevention) that cannot be confirmed from static CSS alone shall be handled according to the “insufficient evidence” rule. The runtime conventions for these three indicators are specified separately in *Supplementary Material S4*.*

S3.3. Task Objective

Based on the interface screenshot (global visual presentation) and the front-end code (element-level parameters), output a discrete four-tier score for each of the following 9 evaluation indicators. Scoring strictly follows the “if ... then” conditional rules in S3.4. No value outside the rules is permitted.

The 9 indicators are organized by dimension:

Visual (V): V1 Contrast / V2 Text Scaling / V3 Non-text Contrast

Operational (O): O1 Target Size / O2 Pointer Cancellation / O3 Focus Order

Accessibility (A): A1 Status Messages / A2 Keyboard Accessibility / A3 Error Prevention

S3.4. Evaluation Criteria

For each indicator, the four-tier judgment is applied progressively. The first tier whose full conditions are satisfied determines the output. If neither $S_i = 25$ nor any higher tier is satisfied, the default output is $S_i = 25$.

V1 Contrast (WCAG SC 1.4.3, code-derivable)

Exempt items, which are not counted as violations:

- Large text (≥ 18 pt or ≥ 14 pt bold), where the threshold is reduced from 4.5:1 to 3:1;
- Purely decorative text not conveying information or function;
- Inactive UI components;
- Text that is part of a logo or brand mark;
- Placeholder text, which in the age-friendly context is supplementary cueing rather than core content.

Tier rules (after removing exempt items):

- If non-exempt body text $\geq 7:1$ and large text $\geq 4.5:1$, then $S_i = 100$.
- If body $\geq 4.5:1$ and large $\geq 3:1$, then $S_i = 75$.
- If body contrast is between 3:1 and 4.5:1, then $S_i = 50$.
- If body contrast $< 3:1$, then $S_i = 25$.

V2 Text Scaling (WCAG SC 1.4.4, code-derivable)

This indicator is assessed against the responsive-adaptation pseudocode in S4.1. If no responsive code or pseudocode is provided, output $S_i = 50$ (“insufficient evidence”).

Exempt items, which are excluded from the minimum-size determination (analogous to the V1 exemption clauses):

- Status badges, tags, and chips carrying supplementary state indicators rather than core textual content;
- System status-bar text fixed by platform convention (15 px in the iOS reference frame used in this study).

Tier rules (after removing exempt items):

- If 200% lossless zoom is supported and all text ≥ 14 pt (≈ 18.7 px), then $S_i = 100$.
- If 200% is supported but small text < 14 pt persists (≥ 12 pt AA baseline), then $S_i = 75$.
- If 200% is supported but content truncation or function loss occurs, then $S_i = 50$.
- If 200% is explicitly unsupported, then $S_i = 25$.

V3 Non-text Contrast (WCAG SC 1.4.11, semantic)

- If all UI components and icons $\geq 4.5:1$, then $S_i = 100$.
- If all UI components $\geq 3:1$, then $S_i = 75$.
- If some components reach 3:1 but others fail, then $S_i = 50$.
- If UI components generally $< 3:1$, then $S_i = 25$.

O1 Target Size (WCAG SC 2.5.8, code-derivable)

- If all interactive elements $\geq 48 \times 48$ CSS px, then $S_i = 100$.
- If all $\geq 24 \times 24$ CSS px (but not all reach 48×48), then $S_i = 75$.

If some interactive elements fall between 16×16 and 24×24 , then $S_i = 50$.

If any interactive element $< 24 \times 24$ CSS px (any violation triggers the tier), then $S_i = 25$.

O2 Pointer Cancellation (WCAG SC 2.5.2, code-derivable)

This indicator involves JavaScript event-binding logic. If only static CSS is provided (no JS logic or pseudocode), output $S_i = 50$ ("insufficient evidence"). Runtime conventions are specified in S4.2.

If two or more protections among "no down-event trigger" and "undo mechanism" are provided, then $S_i = 100$.

If at least one of the four WCAG-prescribed mechanisms (up-event trigger / abort / undo / essential) is satisfied, then $S_i = 75$.

If SC 2.5.2 is not satisfied but misoperation feedback is provided, or evidence is insufficient, then $S_i = 50$.

If SC 2.5.2 is explicitly not satisfied and no misoperation feedback is provided, then $S_i = 25$.

O3 Focus Order (WCAG SC 2.4.3, semantic)

If focus order fully matches visual layout and decorative elements are skipped, then $S_i = 100$.

If focus order matches visual layout in the core interaction region, then $S_i = 75$.

If focus order is partially inconsistent but does not affect core task completion, then $S_i = 50$.

If focus order is disordered or key elements are skipped, then $S_i = 25$.

A1 Status Messages (WCAG SC 4.1.3, semantic)

If status changes are conveyed simultaneously through visual, screen-reader, and audio channels, then $S_i = 100$.

If status changes can be programmatically perceived by a screen reader and have visual feedback, then $S_i = 75$.

If only visual feedback is provided but accompanied by explicit text, then $S_i = 50$.

If only visual feedback with no text, then $S_i = 25$.

A2 Keyboard Accessibility (WCAG SC 2.1.1, semantic)

If all functions support both voice control and at least one other non-pointer assistive path, then $S_i = 100$.

If all functions support at least one non-pointer path (voice or keyboard), then $S_i = 75$.

If some functions support a non-pointer path (with pointer-only functions remaining), then $S_i = 50$.

If only pointer input is supported, then $S_i = 25$.

A3 Error Prevention (WCAG SC 3.3.4, code-derivable)

Runtime conventions for high-risk operations are specified in S4.3.

If high-risk operations simultaneously provide reversibility, checkability, and confirmation, then $S_i = 100$.

If high-risk operations provide one of the three mechanisms, then $S_i = 75$.

If high-risk operations only provide visual warning (no reversibility / checkability / confirmation), then $S_i = 50$.

If high-risk operations have no protection, then $S_i = 25$.

S3.5. Precedence Rules

(1) Worst-case prevails. When multiple elements of the same type are present, the most severely violating element determines the indicator; exempt items are excluded.

(2) Code over screenshot on conflict. When the code and screenshot conflict (e.g., the code reports 16 px but the visual appears smaller), the code value is used, and the conflict is noted in the evidence field.

(3) No applicable target. For pages without high-risk operations (e.g., pure display or daily switch control), A3 outputs $S_i = 100$ with “no applicable target” noted in evidence.

(4) Insufficient evidence. When evidence is insufficient (e.g., the screenshot does not cover the status-change flow, no JS logic is provided), output $S_i = 50$ with the specific reason noted in evidence.

S3.6. Evidence Field Hard Constraints

The evidence field must contain at least one of the following verifiable forms, and speculative wording is not permitted:

(a) specific code fragments, CSS property values, or coordinates (e.g., “left: 196 px; top: 249 px” or “color: #9B9B9B”);

(b) computable numerical judgments (e.g., “contrast ratio computed at 2.8:1” or “target size 40×125 px”);

(c) the specific count and scope of violating elements (e.g., “2 of 8 UI components do not reach 3:1”, rather than “some components”).

The following uncertainty terms are forbidden in the evidence field: “likely”, “possibly”, “probably”, “easily”, “generally”, “speculatively”. When evidence is insufficient, the evidence field must state “insufficient evidence: <what is specifically missing>”, rather than using speculative wording as a placeholder.

S3.7. Output Schema

The model must return strictly the following JSON structure and no other content (no Markdown code-fence markers, no preamble or postamble):

```
{
  "scenario_name": "<scenario name>",
  "scores": [
    {
      "metric_id": "V1",
      "metric_name": "Contrast",
      "dimension": "Visual",
      "weight": 0.0952,
```

```

    "Si": 75,
    "weighted_score": 7.14,
    "evidence": "<specific interface anchor, including code
        fragment or numerical computation>",
    "suggestion": "<improvement direction, precise to CSS
        parameter or interaction logic>"
  }
],
"total_score": 0.00,
"compliance_level": "<determined by total_score:
    >=90 excellent, 80-90 compliant,
    75-80 basically compliant,
    <75 non-compliant>",
"summary": "<comprehensive assessment within 80 words,
    indicating the principal risk dimensions>"
}

```

The “scores” array contains nine entries with the same internal structure, one entry for each indicator V1, V2, V3, O1, O2, O3, A1, A2, and A3.

S3.8. Weighting and Calculation Rules

For V1, V2, V3, O1, O2, O3 - weight = 0.0952.

For A1, A2, A3 - weight = 0.1429.

$\text{weighted_score} = \text{weight} \times S_i$, rounded to two decimal places.

$\text{total_score} = \text{sum of 9 weighted_score values}$, rounded to two decimal places.