ship bar: the white paper

AI agents build websites fast, and they mostly look the same: glows, gradients, glass cards, a floating phone. This is the method I use to get past that. A separate reviewer agent scores every build out of 10, hands back ranked fixes, and nothing ships below 9.7.

Also available as a pdf. Back to ship-bar.neural-forge.io.

what it is

Two parts. First, art direction rules the building agent follows, ending in a deslop pass (a final sweep that removes the generic AI-site tells). Second, a reviewer agent that never touches the code. It looks at screenshots at 1440px and 390px, runs Lighthouse, checks the console, checks for horizontal overflow, and scans computed styles for banned patterns. It returns a score and a list of fixes, most important first. The builder fixes the top items, ships a new build, and the reviewer scores it again from scratch.

The banned list: `glows, gradients, glassmorphism, floating phones, bento grids, confetti, serif headlines, filler copy`.

the method

01 kill the category default

No near-black with a glowing accent, gradient orbs, glass cards, floating tilted devices or sparkle-emoji ai.

02 pick references

Choose 2-3 real best-in-class sites that fit the product's mood.

03 get a second opinion

Ask another design-minded agent for a direction. Where you converge, that's the signal.

04 write design.md

Warm canvas, ink, hairlines, one accent for the primary action. A display face, a ui grotesk, mono for data. Spacing, radius, motion, do/don't lists.

05 show the transformation

The hero shows input becoming output within one second.

06 **use real imagery**

Real product ui as cropped cards with true data. Plain, specific copy.

07 **one signature motion**

Plus subtle fade-ups, respecting prefers-reduced-motion.

08 **deslop**

Gradient text, pill badge above the h1, identical icon grids, same fade everywhere, rounded-2xl plus shadow-lg, stock shadcn slate, centered-everything, hype words, fake testimonials.

09 **verify, then review**

Screenshots at mobile and desktop, Lighthouse, then a separate reviewer agent scores it before the owner sees it.

the benchmarks

What the reviewer checks on every build.

CHECK	THRESHOLD	CATCHES
Desktop screenshot	`1440px wide`	Layout, hierarchy, polish
Mobile screenshot	`390px wide`	Same, at phone width
Lighthouse performance	`≥ 95`	Speed on mobile
Lighthouse accessibility	`≥ 95`	Labels, contrast, alt text
Lighthouse best practices	`≥ 95`	Console, https, images
Lighthouse SEO	`≥ 95`	Meta, crawlability
Console errors	`0`	Broken scripts
Horizontal overflow	`none at 390px`	Content wider than the phone

CHECK	THRESHOLD	CATCHES
Banned-pattern scan	`0 found`	Computed styles: glows, gradients, glass, floating phones, bento, confetti, serif headlines
Content scan	`0 found`	Debug strings, joke placeholder data, off-palette chips, duplicate rows, raw UTC timestamps, filler copy
Contrast	`WCAG AA`	Text and chips against background
Overall score	`median ≥ 9.7, 0 blockers`	20 reviewers, each with ranked fixes, most important first

The content scan matters more than I expected. Real things it caught: debug strings shown to users, joke placeholder data, off-palette chips, duplicate rows, UTC timestamps where local time belonged, and contrast failures.

The output format, so a builder agent can act on it:

```
score: 9.3 / 10
blockers:
  - sticky cta covers the email field at 390px
fixes (most important first):
  1. remove the gradient on .pricing-card
  2. raise footer link contrast to 4.5:1
  3. ...
```

Format example. The two issues listed are the real v13 regressions.

cohorts: 20 reviewers

One reviewer averages things out. A cohort doesn't. Run 20 reviewer agents, each with one narrow lens, each scoring independently out of 10 with its own ranked fixes. Any reviewer can flag a **blocker**: something broken, untrue or unusable. The final score is the **median** of the 20, and nothing ships under 9.7 or with any open blocker. Fixes are merged into one list ranked by impact.

01 · first-time visitor
Do I understand what this is and what to do within five seconds?

02 · skeptical buyer
Is every claim backed by something real, or is it hand-waving?

03 · art director
Does it have a point of view, or does it look like every other AI site?

04 · typographer
Type scale, line length, wraps, orphans, the mono/grotesk split.

05 · copy editor
Plain, specific, correct sentences. No filler, no hype.

06 · accessibility auditor
WCAG AA contrast, landmarks, focus states, alt text, headings.

07 · mobile thumb-user
Tap targets of 24px or more, reach, nothing hidden behind sticky bars at 390px.

08 · performance engineer
Lighthouse performance, payload, fonts, layout shift.

09 · seo
Title, meta, canonical, headings, sitemap, crawlability.

10 · aeo / llm answer engine
Can an answer engine quote a clean, dated answer? JSON-LD, FAQ, llms.txt.

11 · brand guardian
The banned list, the palette, and one accent reserved for the primary action.

12 · conversion / cta
Is there one obvious next step, and does it work?

13 · information architect
Section order, labels, and whether you can find things again.

14 · developer copying the skill
Can I paste this and run it today? Is the output format defined?

15 · non-native english reader
Jargon, idioms, and sentences that need a second read.

16 · screen-reader user
Reading order, button names, live feedback, skip link.

17 · trust / legal
Licence, attribution, no fake data, no personal info, honest claims.

18 · competitor's designer
What would I steal, and what would I mock?

19 · qa breaker
Edge cases: no clipboard, no js, slow images, odd widths, broken links.

20 · would i share this
Would I post this link, and does the preview card make sense?

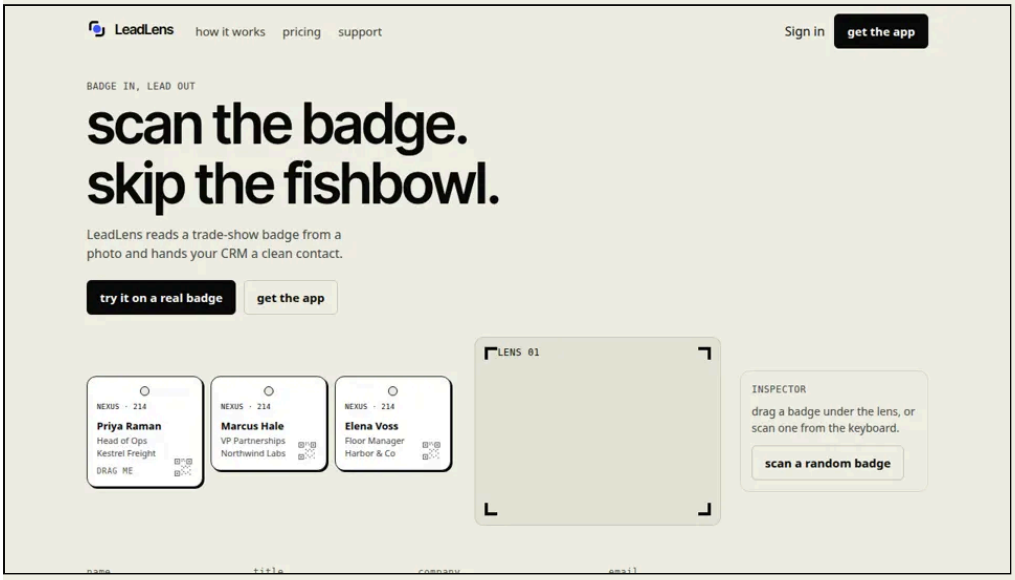
On real projects I also run smaller themed cohorts, for example fun, art, security and pitch, each with its own bar.

an example score trajectory

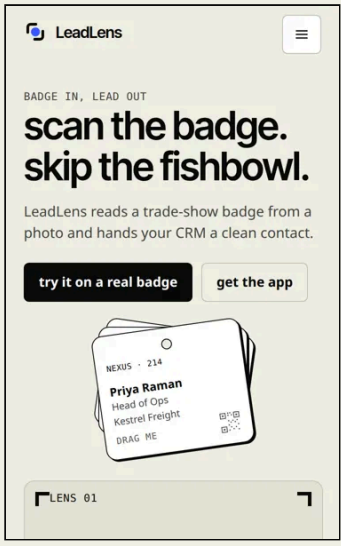
A real product landing page, reviewed build by build.

BUILD	SCORE	NOTE
v3–v5	below bar	early builds; fixes ranked and applied each round
v6	below bar	“still felt a tad vibe coded”

BUILD	SCORE	NOTE
v7–v11	below bar	fixes merged by rank, re-scored each build
v12	**9.7 · pass**	accessibility 100
v13	9.3 · fail	sticky cta covered a form field; a gradient slipped in



v12, desktop



v12, mobile

The v13 drop is the point. A score isn't kept; it's re-earned every build.

we ran it on this page

A page about a benchmark should pass it. These are the real scores from running the 20-reviewer cohort on this site, round by round. The measurements came from headless Chrome and Lighthouse on the live URL. The reviewer judgments in these rounds were done by one agent working through each lens in turn, not by 20 separate models, and the table says exactly what they said.

Note: rounds r1 to r3 scored the earlier one-page version of this site. After feedback that it was too much information, the detail moved into this white paper and the homepage became a short slideshow. The new homepage was re-checked with the automated measurements only (0 console errors, no horizontal overflow at 1440 or 390, 0 banned patterns, Lighthouse 100 in all four categories on mobile and desktop); the 20 reviewers have not re-scored it yet.

1. **r1 · median 9.20 · 2 blockers · fail**

   Measured: 0 console errors, no horizontal overflow at 1440 or 390, 0 banned patterns in the computed-style scan, Lighthouse 100/100/100/100 mobile and desktop.

   Blockers: qa breaker: the two v12 screenshots rendered blank in a full-page capture (lazy loading); brand guardian: the accent was used on three buttons, not one primary action.

   Top fixes, ranked:

   - fix blank screenshots: load them eagerly
   - make download the only accent button; copy becomes secondary
   - tap targets under 24px: 'copy it below' and the footer link
   - define 'deslop' for non-native readers
   - replace 'report score' with real Lighthouse thresholds
   - show the reviewer's output format
   - add a dated byline for answer engines
   - copy button: live status and a fallback when the clipboard is blocked
   - add a skip link and visible focus states
   - expand to the 20-reviewer cohort

2. **r2 · median 9.70 · 0 blockers · pass**

   Measured: 0 console errors, no horizontal overflow at 1440 or 390, 0 banned patterns, screenshots now load, Lighthouse 100/100/100/100 mobile and desktop.

   Top fixes, ranked:

   - add a short section nav: the page is long, especially at 390px
   - stack the two screenshots on mobile; side by side they are too small to read
   - turn the round fixes into a list instead of one long paragraph
   - the accessibility row's 'catches' column said 'v12 hit 100'; describe what it catches

3. **r3 · median 9.70 · 0 blockers · pass**

   Measured: 0 console errors, no horizontal overflow at 1440 or 390, 0 banned patterns, Lighthouse 100/100/100/100 mobile and desktop.

   Top fixes, ranked:

   - open notes, not fixed: the hero paragraph still describes one reviewer; it could name the 20-reviewer median up front
   - competitor's designer: the self-test tables are dense on a phone

REVIEWER	R1	R2	R3
01 first-time visitor	9.0	9.6	9.7
02 skeptical buyer	8.8	9.6	9.7
03 art director	9.2	9.6	9.7
04 typographer	9.3	9.5	9.7
05 copy editor	9.4	9.6	9.7
06 accessibility auditor	9.3	9.8	9.8
07 mobile thumb-user	9.0	9.5	9.7
08 performance engineer	9.9	10.0	10.0
09 seo	9.6	9.8	9.8
10 aeo / llm answer engine	9.2	9.8	9.8
11 brand guardian	9.1	9.8	9.8
12 conversion / cta	9.3	9.7	9.7
13 information architect	9.4	9.4	9.7
14 developer copying the skill	9.0	9.8	9.8
15 non-native english reader	9.2	9.7	9.7
16 screen-reader user	9.2	9.8	9.8
17 trust / legal	9.5	9.8	9.8
18 competitor's designer	9.0	9.5	9.6
19 qa breaker	8.7	9.7	9.8
20 would i share this	9.1	9.6	9.7
median	**9.20**	**9.70**	**9.70**
blockers	2	0	0

skill.md

Generic version, free to use and adapt. Paste it into your agent's skills or instructions.

download skill.md

```
---
name: ship-bar-design-review
description: >-
  use this when an AI agent builds or redesigns a website and you want it to
  stop looking vibe-coded: art direction, a deslop pass, and a separate
  reviewer agent that scores each build against a 9.7/10 ship bar
---
# Ship bar: design review for AI-built websites

## Build
1. Kill the category default. No near-black with a glowing teal/cyan/purple accent, gradient orbs, glass
cards, floating tilted devices, or sparkle-emoji "AI".
2. Pick 2-3 real best-in-class reference sites that fit the product's mood.
3. Get a second opinion from another design-minded agent. Where opinions converge, follow them.
4. Write design.md at the repo root: a warm canvas, ink, secondary and hairline colors, ONE accent reserved
for the primary action; a display face, a UI grotesk, and mono for data; spacing, radius, hairlines over
shadows, motion rules, do/don't lists.
5. Hero = the product's transformation shown working (input becomes output) within one second.
6. Imagery: real product UI as cropped cards with true data. Copy is plain, specific, concrete.
7. Motion: one signature animation plus subtle fade-ups; respect prefers-reduced-motion.
8. Deslop before shipping: gradient text, pill badge above the H1, identical icon-card grids, same fade on
everything, uniform rounded-2xl + shadow-lg, unmodified shadcn slate, centered-everything, hype words, fake
testimonials.
9. Verify: screenshots at mobile and desktop, Lighthouse, then a reviewer agent scores it before the owner
sees it.

## Review (separate reviewer agent; it never edits the repo)
- Screenshots at desktop 1440px and mobile 390px.
- Lighthouse: performance, accessibility, best practices, SEO.
- Console errors: zero.
- Horizontal overflow at 390px: none.
- Computed-style scan for banned patterns: glows, gradients, glassmorphism, floating phones, bento grids,
confetti, serif headlines (if your design.md bans them), filler copy.
- Content checks: debug strings shown to users, joke placeholder data, off-palette chips, duplicate rows, raw
UTC timestamps, contrast failures, sticky elements covering inputs.

Output: a score out of 10 plus fixes ranked most important first.
Ship bar: 9.7. Below it, the builder fixes the top items and the reviewer re-scores a new build.

## Cohorts (optional)
Run several narrow-lens reviewer agents (for example fun, art, security, pitch), each with its own bar,
scoring independently. Merge their fixes into one list ranked by impact. Ship when every lens clears its bar.
```

faq

What is the ship bar method?

A way to review AI-built websites before a human sees them. A separate reviewer agent scores each build out of 10 from screenshots, Lighthouse and automated checks, returns ranked fixes, and the build only ships at 9.7 or above.

Why a separate reviewer agent?

The agent that built the page tends to grade its own work kindly. A reviewer that only reads screenshots and reports, and never edits the repo, keeps the score honest and the fixes specific.

What counts as vibe-coded?

The category defaults: glows, gradients, glassmorphism, floating phones, bento grids, confetti, decorative serif headlines and filler copy. The reviewer scans computed styles for these, not just the screenshot.

Why 9.7 and not 10?

9.7 was the point where the remaining notes were taste, not defects. Below it, there was always something a visitor would notice.

What are cohorts?

Several narrow-lens reviewer agents, for example fun, art, security and pitch, each scoring independently against its own bar. Their fixes are merged into one list ranked by impact.

Can a score go down?

Yes. In the example below, v13 scored 9.3 after v12 passed at 9.7, because a sticky call-to-action covered a form field and a gradient slipped in. Every build is re-scored from scratch.

How do I use it?

Copy the SKILL.md block into your agent's skills or instructions, write a design.md for your project, and run a second agent as the reviewer on every build.

How does the 20-reviewer cohort decide?

Each of 20 narrow-lens reviewers scores out of 10 and can flag blockers. The final score is the median. Nothing ships under 9.7 or with any open blocker.

Did this page pass its own benchmark?

See we ran it on this page for every round, score and blocker.

Made by Travis Curnutte. Share it, fork it, make it yours. Licensed CC BY 4.0. llms.txt