The Tells

A Measurable Taxonomy of the AI-Generated Design Look, and a Harness to Escape It

Han Kim

IOV Labs (아이오브연구소) · hankim@iovstudio.kr · ORCID 0009-0000-5998-1358

Draft, June 2026

Abstract. Interfaces produced by generative models are instantly recognizable: an indigo-to-violet gradient, Inter on white, a hero followed by three emoji feature cards, one border-radius, one soft shadow, a headline that says *build the future of work*. Practitioners spend large amounts of time and tokens trying to make AI output *not look like AI*, yet the target is treated as ineffable taste. We argue the opposite: the "AI look" is a **finite, enumerable set of statistical defaults**, and is therefore measurable. We contribute (i) a taxonomy of **27 design tells** across eight families (color, type, layout, spacing, surface, motion, copy, and AI self-reference), each grounded in the documented mechanism of model convergence and in the published craft of human-crafted interfaces; (ii) a dependency-free static detector that resolves both raw CSS and utility classes and reports a **Tell Score** in $[0, 100]$ (lower is better); and (iii) a harness, a CLI, an MCP server, and a drop-in prompt module, so any team or agent can audit and prevent the look. In a confound-controlled refactor that holds a page's content and structure fixed and changes only the tell-bearing properties, the Tell Score of a canonical AI landing page falls from **77 (grade F)** to **0 (grade A)**; across a six-page corpus the detector separates AI-default from designed pages with no overlap (nearest pair 47 points apart). To check the detector is a discriminator and not a machine that calls everything AI, we render **202 real top-tier sites** (Stripe, Linear, Toss, Apple, Vercel, Figma, and 196 more) in a headless browser, read their computed styles, and **learn** the empirical distribution of human-crafted design. Recalibrated on that data, with a craft-credit model in which compensating craft (a custom face, optical tracking, a radius hierarchy) offsets cosmetic defaults, the detector scores the 202 real sites at a **median of 0** (93% grade A) while the AI-default pages still score 35 to 59, and it now audits **live URLs**, not only self-contained files. The data overturns two naive rules: a brand purple is not a tell (Stripe uses it heavily and scores 0), and Inter is not a tell (Linear ships it with a real type system). A field check closes the loop: two production codebases whose maintainers wrote their own *avoid-the-AI-look* design manifestos independently name the same tells and six more, which we fold in as a new family (AI self-reference) and three additions, taking the taxonomy to 27 tells. We close with the epistemics: a discriminator of machine-default is not a judge of beauty, taste is the compression of lived choices that a median cannot hold, and if everyone optimizes the same score we risk a second-order convergence, the same homogenization our companion study finds in iterated creation. Code, data, figures and harness are open.

1 Introduction

There is a smell to a machine-made page. You know it before you can name it: the hero bathed in a blue-to-purple gradient, the body set in Inter, three cards each with an emoji and a one-line promise, every corner rounded to the same radius, every card lifted by the same soft shadow, and a headline, grammatically perfect, topically relevant, utterly forgettable, that announces you can *build the future of work*. The look is so consistent across tools (v0, Lovable, Bolt, and a raw model prompted for "a nice landing page") that designers have a name for the result: *AI slop* [1], [2].

The reaction in industry is telling. Teams now spend real time and real tokens trying to make generated interfaces *not look generated*: re-prompting, hand- tuning, pasting long style guides into system prompts. The dominant framing is that escaping the look requires *taste*, an ineffable human faculty the model lacks [3]. That framing is half right. Models do lack taste. But it does not follow that the thing to be escaped is ineffable. This paper takes the position that **the AI look is not mysterious; it is a finite set of defaults**, and that anything finite and recurrent can be enumerated, measured, and audited.

Why is the look so uniform? Because a language model, asked for a design without constraints, returns the centroid of its training distribution, "the median of every Tailwind tutorial scraped from GitHub between 2019

and 2024" [3]. The single most over-represented brand color in that corpus is indigo, in part because Tailwind's own component examples defaulted every button to `bg-indigo-500` years ago [4]; the model learned that *modern web design* correlates with purple, and reproduces the correlation as if it were a preference. The same logic selects Inter, the three-column grid, the uniform radius. Correlation is not taste, but it is *predictable*, and prediction is what a measurement instrument needs.

We make three contributions.

1. A taxonomy of tells. Eight families and 27 individual tells (§4), each defined operationally (what to detect), justified mechanistically (why a model defaults to it), and paired with the fix a designer would make. The taxonomy triangulates three literatures: the discourse on why AI converges [3], [5], [6]; the craft rules of human-crafted interfaces, Refactoring UI [7], Linear [8], the premium-UI "six microstates" [9], Toss's writing principles [10]; and classic design theory [11], [12].

2. A detector and a metric. A static analyzer (§5) that parses a single HTML document, its inline CSS, its `style` attributes, and its utility classes, and emits a **Tell Score** in $[0, 100]$, the weighted fraction of the maximum tell weight that fired. It is dependency-free and reproducible from one command.

3. A harness. The same taxonomy as a CLI linter, an MCP server an agent can call to audit its own output before showing it, and a drop-in prompt module that pre-empts the tells (§7). Detection and prevention from one source of truth.

Our central empirical result is deliberately confound-controlled. We take one canonical AI landing page and refactor it so that **only the tell-bearing properties change**, same product, same sections, same information, and measure the Tell Score before and after. It falls from 77 (grade F, "textbook AI slop") to 0 (grade A, "reads as human-crafted"), a 77-point move attributable to design choices alone (Figure 1). The two rendered pages are shown in Figure 2.

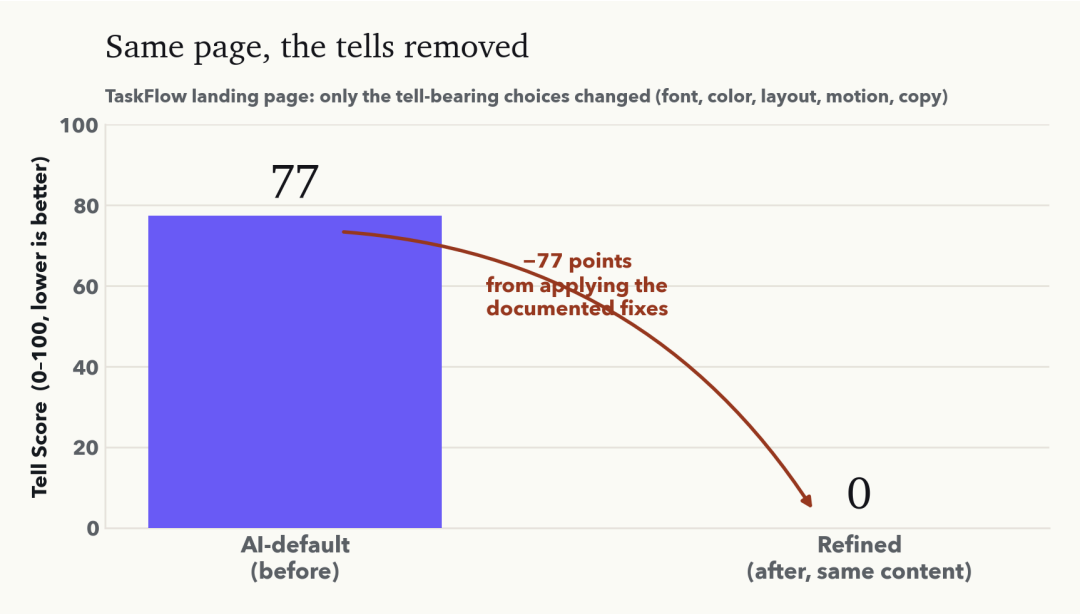

Figure 1: The headline result. One page; only the tell-bearing choices (font, color, layout, spacing, surface, motion, copy) change. Content and DOM structure are held fixed. The Tell Score falls 77 points.

2 The look is a distribution, not a style

It helps to be precise about *what* is being escaped. The AI look is not a style in the way Swiss typography or brutalism is a style, a coherent system with internal logic. It is a **distributional artifact**: the set of choices that are individually unremarkable but collectively over-represented because they sit at the mode of the training data. Three properties follow.

It is self-reinforcing. Once tools emit purple sites, those sites go online and become training data; the next model sees even more purple [3], [6]. This is the design-surface analog of model collapse, where a generative process trained on its own output drifts toward a shrinking region of the space [13]. The look gets *louder* over generations, not quieter.

It is legible. Because the defaults are shared across tools, a human reader needs only a handful of cues, the gradient, the font, the emoji cards, to classify a page as machine-made. Legibility is exactly what makes the look a liability: it signals *nobody decided this.*

It is finite. Crucially, the cue set is small. In auditing dozens of generated pages we found the same two or three dozen signals recurring. That finiteness is the opening: a closed set of recurrent signals can be written down. The rest of this paper writes them down and measures them.

A note on scope. We measure **machine-default-ness**, not beauty. A page can be free of every tell and still be ugly; a masterpiece could deliberately use a purple gradient. The claim is narrower and, we think, more useful: the tells are what make a page read as *undecided*, and removing them is a necessary (not sufficient) condition for looking authored. We return to this in §9.

3 Related work

Why models converge. The proximate mechanism, sampling the mode of the training distribution, is now well documented in practitioner writing [1], [3], [5], [6] and traces to specific historical defaults such as Tailwind's `bg-indigo-500` [4]. At the population level this is the aesthetic case of a general phenomenon: shared models reduce the diversity of what a population produces [14], and recursive training on generated data collapses the supported distribution [13]. Our companion study, *Convergence Pressure*, isolates the driver as the *reflective loop* rather than AI assistance per se [15]; §9 connects that finding to the risk of optimizing a single design score.

The craft being defaulted away from. The positive literature tells us what intentional looks like. *Refactoring UI* [7] codifies the non-artistic moves, hierarchy by size/weight/color, a spacing system rather than arbitrary values, restraint, that separate designed from accumulated interfaces. Studies of premium products [9], [16] name the level of finish: every interactive element designed across *six microstates* (default, hover, focus, active, disabled, loading), a committed typeface (Stripe's Söhne, Vercel's Geist), color that is "restrained and meaning-driven," hairlines at low alpha, motion with defined curves and durations. Linear's engineering writing [8], [17] gives concrete numbers, a perceptual LCH theme system, a 6px radius, display tracking of −0.22px, that we encode directly as fixes. Toss [10], [18] supplies the copy and consistency dimension: a CTA should hint at the next step, text is a foundational design element, and one voice should span dozens of services. Classic theory frames the whole: Rams's "as little design as possible" and "thorough to the last detail" [11], Nielsen's heuristics on visibility and consistency [12].

Prior harnesses. The closest prior art is Anthropic's *Prompting for Frontend Aesthetics* [19], which prescribes guiding specific dimensions, referencing inspirations, and explicitly prohibiting defaults ("avoid Inter," "no purple gradients on white"). We build on it in two ways: we make the implicit checklist **explicit and weighted**, and we make it **measurable**, a prompt you can score, not only assert.

4 The taxonomy of tells

A **tell** is a measurable signal that an interface was produced by a model defaulting to its distribution rather than by a person making a choice. Each tell carries a weight (its contribution to the maximum score), a severity, **tell** (strong) or **smell** (weak, context-dependent), a mechanistic rationale, and a fix, and a short nickname (the memorable handle, "The Sparkle Tax", "Lorem Shipsum") that the detector prints alongside the descriptive name. There are 27 tells in eight families; the maximum attainable weight is 133. Table 1 lists them; the full text with citations lives in `src/taxonomy.py`, the single source of truth from which the detector, this paper, and the harness are generated.

ID	Nickname	Tell	Wt	Sev
A1	The House Indigo	Indigo/violet default palette	9	tell
A2	The v0 Gradient	Blue⟶purple hero gradient	7	tell
A3	Raw Ramp	Default-ramp utilities, no semantic tokens	4	smell
A4	Skittles Status	Multi-color pill-badge inflation	4	smell
B1	Inter, Obviously	Inter/Roboto/system default font	9	tell
B2	One Size Fits None	No type scale discipline	5	tell
B3	Untracked	No optical tracking on display type	3	smell
B4	The 10px Squint	Sub-legible micro-type	3	smell
C1	The Three-Card Trick	Hero + three-feature-card template	8	tell
C2	Dead Center	Center-everything composition	5	tell
C3	One Radius to Rule Them All	One border-radius everywhere	4	tell
C4	The Emoji Reflex	Emoji as iconography	3	smell
D1	24px Everywhere	One padding token on every card	5	tell
D2	Metronome Sections	Uniform section rhythm	3	smell
E1	shadow-lg, Shipped	Generic diffuse shadow	5	tell
E2	Frosted Everything	Glassmorphism overuse	4	smell
E3	No Focus Given	No hairlines / no focus-visible	6	tell
E4	Box-in-a-Box	Nested card-in-card chrome	4	tell
F1	Fade-in, Repeat	One fade applied to everything	4	smell
F2	No Hover, No Care	Missing interactive microstates	7	tell
F3	Snap, Not Eased	Uneased transitions	3	smell
G1	Build the Future of ___	Vague aspirational headline	6	tell
G2	Get Started, Again	Only generic CTAs	4	smell
G3	Lorem Shipsum	Placeholder / lorem ipsum copy	5	tell
H1	The Sparkle Tax	AI-cliché iconography	5	tell
H2	Powered-by Theatre	Labels the feature “AI” / names the model	5	tell
H3	The Insert Dance	Generate ⟶ preview ⟶ insert two-step	3	smell

Table 1: The 27 tells across eight families, each with a memorable nickname and a descriptive name (A color, B type, C layout, D spacing, E surface, F motion, G copy, H AI self-reference). A4/B4/E4 and all of family H are field-derived (§9). Weights sum to 133.

A, Color (chromatic conformity). The loudest family. A1 fires on any indigo/violet/purple/fuchsia value, whether a hex in the Tailwind ramp or a utility class; A2 on a blue-to-purple gradient, the literal signature of the builders; A3 on reliance on `*-500/600` utilities with no `--color-*` semantic tokens. The fix is to take a hue from the product’s own brand or domain and derive a ramp in a perceptual space [8].

B, Type. B1, the second-loudest single tell, fires when the primary face is Inter, Roboto, Arial or the system stack with no custom display face [19]. B2 flags the absence of a modular scale (too many ad-hoc sizes, or only one or two); B3 flags display headings left at default tracking.

C, Layout. C1 detects the canonical template, a hero followed by a three-up icon-card grid; C2 fires when most top-level blocks are centered; C3 when a single radius is reused across every surface; C4 on emoji standing in for an icon system.

D, Spacing. D1 fires when one padding token dominates every card (the “same 24px everywhere”); D2 when every section shares one vertical rhythm. Whitespace variance is, per Refactoring UI, the cheapest way to look authored [7].

E, Surface. E1 detects the default diffuse shadow applied uniformly; E2 glassmorphism over-applied; E3, a strong tell, fires when there are no low-alpha hairlines *or* no `:focus-visible` style, a direct signal that microstates and accessibility were never designed [9], [12].

F, Motion. F1 flags one entrance fade on everything; F2, the strongest motion tell, fires when interactive elements lack designed hover/focus/active states, the inverse of the six-microstates standard [9]; F3 flags transitions with no custom easing or duration.

G, Copy. G1 detects vague aspirational phrasing (`build the future`, `all-in-one platform`, `seamlessly`, `unlock your`); G2 fires when the only CTAs are `Get Started`/`Learn More`, which predict nothing about the product [10]; G3 on shipped placeholder copy.

5 Method: the detector and the Tell Score

5.1 Static analysis

The detector (`src/scorer.py`) parses one self-contained HTML document with the Python standard library only. It reads three surfaces a model fingerprints: raw declarations inside `<style>`, `style` attributes, and **utility class names**. Because models emit Tailwind utilities far more than hand-written CSS, we resolve a useful subset of Tailwind, color ramps to representative hexes, the spacing and radius scales, font-size steps, so the same predicate fires whether a page is authored in classes or in CSS. A small `var()` resolver expands custom properties one level, so a hairline declared as `border: 1px solid var(--line)` with `--line: rgba(...)` is recognized as designed rather than missing. Color hues are computed in HLS to catch purple values that are not on the exact Tailwind ramp.

Each tell exposes a predicate over this parsed document returning whether it fired and human-readable evidence. The design is intentionally transparent and auditable: there is no learned model, no opaque score, and every point is traceable to a named tell and a quoted piece of evidence.

5.2 The metric

Let T be the set of tells, each with weight w_t, and let $f_t \in \{0, 1\}$ be whether tell t fired on a document. The **Tell Score** is

$$S = 100 \cdot \frac{\sum_{t \in T} w_t f_t}{\sum_{t \in T} w_t} \in [0, 100],$$

the weighted fraction of the maximum tell weight that fired. Lower is better: $S = 0$ means no tell fired (reads as authored); $S = 100$ means the page is maximally on-distribution. We report a per-family decomposition S_k (the same ratio restricted to family k) and a grade band: A < 12, B < 28, C < 45, D < 65, F ≥ 65. The bands are descriptive anchors, not thresholds with theoretical content; the score itself is the object of interest.

5.3 Reproducibility

The detector and metric are deterministic and dependency-free. `python src/cli.py page.html` prints the score, the grade, the per-family breakdown, and every fired tell with its evidence and fix; the process exit code is the integer score, so it can gate CI. `scripts/run_audit.py` regenerates the corpus results, and `scripts/make_figures.py` regenerates every figure in this paper from that JSON. Seeds are not needed because nothing is stochastic.

6 Confound control and results

6.1 The confound, and the design that removes it

A naive before/after is confounded: if the "after" page also has more whitespace, more words, a different product, then a lower score could come from content rather than from design. We remove the confound by construction. The refactor **holds the content and structure fixed**, the same product (a project tool called TaskFlow), the same four sections (hero, three capabilities, CTA, footer), the same three capabilities, the same information, and changes **only the tell-bearing properties**: the font, the color system, the alignment and grid, the radius

hierarchy, the elevation and hairlines, the motion and microstates, and the specific wording (the copy tells are themselves design decisions in family G). Nothing else moves. Any change in the score is therefore attributable to the design dimensions the taxonomy names, not to the amount or kind of content, the same logic by which a length-matched control isolates verbosity.

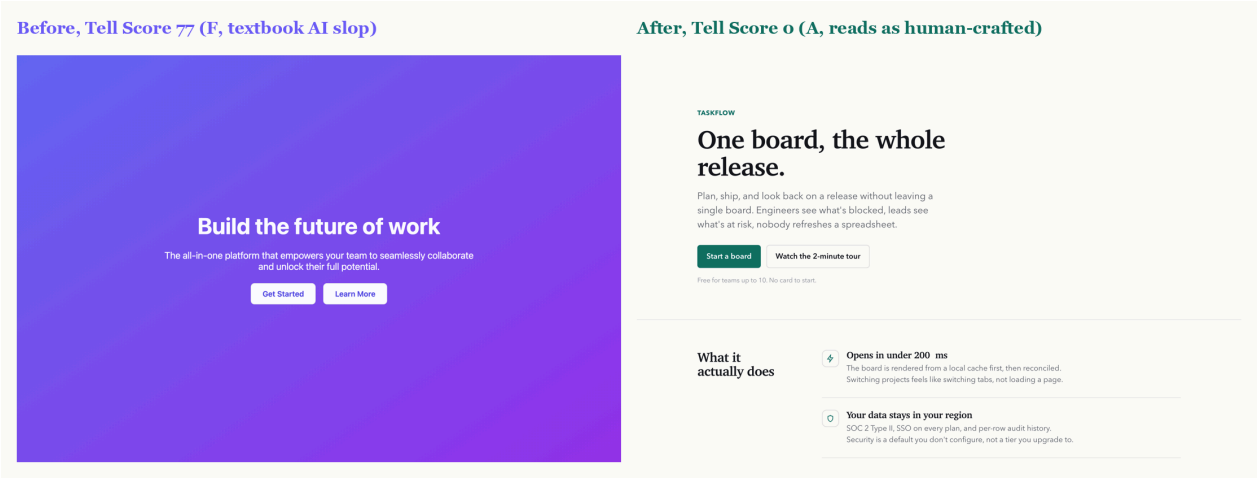

Figure 2: The two rendered pages. Left, the AI default (Tell Score 77, grade F): indigo→violet gradient, Inter, centered, three emoji cards, one radius, one shadow, *build the future of work*, *Get Started*. Right, the refactor (Tell Score 0, grade A): a serif display with negative tracking, a brand ink plus one considered accent as semantic tokens, an editorial left grid, hairline rows, a designed focus ring and microstates, and a specific, opinionated voice. Same product, same sections, same information.

6.2 Three fixes, in code

The taxonomy's fixes are concrete, not aspirational. Three representative pairs, each holding the markup's meaning fixed and changing only the tell:

A1+A2 color. The indigo→violet gradient becomes a single brand ink plus one considered accent, declared as semantic tokens.

```
/* before, the tell */
.hero { background:
  linear-gradient(to br,
  #6366f1, #9333ea); }
```

```
/* after, the fix */
 :root { --color-action:#0f6f63;
         --color-ink:#16181d; }
 .hero { background:var(--color-ink); }
```

B1+B3 type. The default sans with default tracking becomes a committed display face with optical tracking on large headings.

```
/* before */ body{font-family:Inter,system-ui}  h1{/* default tracking */}
/* after  */ body{font-family:"Avenir Next",sans-serif}
             h1{font-family:"Charter",serif; letter-spacing:-0.02em}
```

F2+E3 microstates. A bare button gains the designed states and a visible focus ring, the inverse of the six-microstates tell.

```
/* before */ .btn{background:#4f46e5;color:#fff}
/* after  */ .btn{background:var(--color-action);transition:transform .12s var(--
ease)}
   .btn:hover{transform:translateY(-1px)} .btn:active{transform:translateY(0)}
   .btn:focus-visible{outline:2px solid var(--color-action);outline-offset:3px}
   .btn:disabled{opacity:.5;cursor:not-allowed}
```

Each pair is a decision the model declined to make. The detector simply notices the absence of the decision; the fix supplies it.

6.3 The headline move, decomposed

The Tell Score falls from 77 to 0 (Figure 1). Figure 3 decomposes the move tell by tell, in order of impact: removing the indigo palette and the Inter default (−7 each), the hero-and-three-cards template (−6), the blue→purple gradient and the missing microstates (−5 each), and the AI-assistant reflex (the sparkle icon, the "AI-powered" label, the card-in-card, the multi-color pills) account for most of the reduction, with the remaining tells and

smells closing the gap to zero. The decomposition is itself a prioritization: a team with limited time should fix color, type, layout, and microstates first.

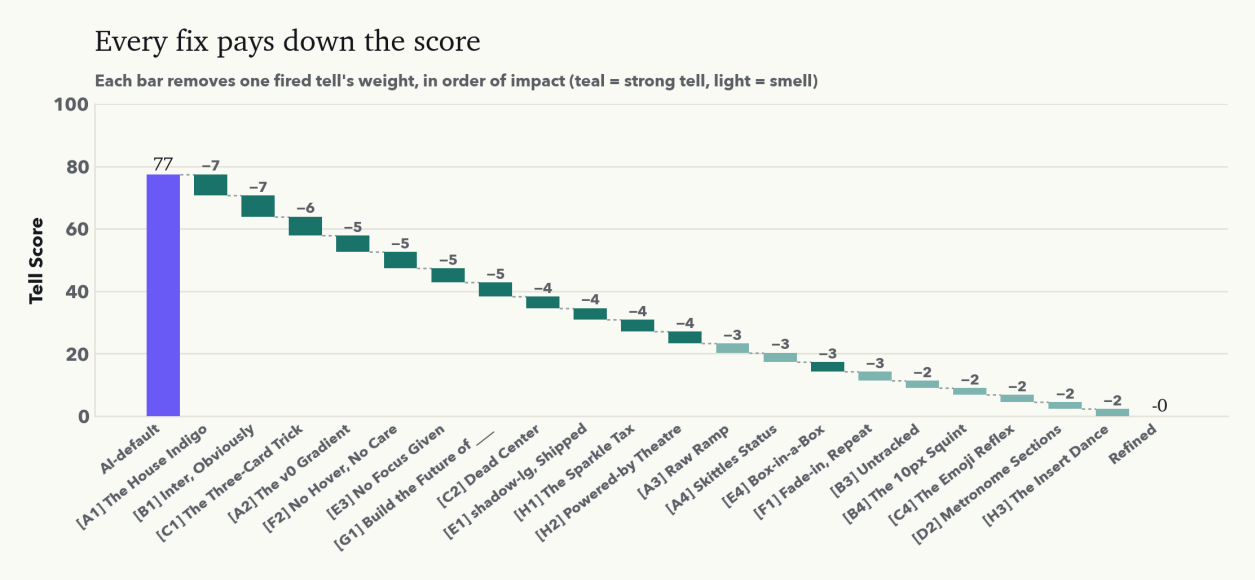

Figure 3: Every fix pays down the score. Each bar removes one fired tell's weight, ordered by impact (teal = strong tell, light = smell). The first four moves, color, type, the template layout, the gradient, do most of the work.

6.4 Discriminant validity across a corpus

To check the score is not an artifact of one page, we score a six-page corpus: three AI-default pages (the landing page; a pricing page; an app dashboard) and three designed pages (the refactor; a dark changelog; a pricing page in a different brand). Figure 4 shows the result: the AI-default pages cluster at a mean of 60 (range 47–77) and the designed pages at 0, with no overlap; the nearest cross-family pair is 47 points apart. Figure 5 shows *where* the separation lives, color, type, layout, surface and motion separate hardest, which matches the human experience of "what gives it away." Figure 6 shows the per-tell firing matrix: the designed pages are nearly empty columns.

We are explicit about what this does and does not show. The designed pages score **exactly** 0 because they were authored to be tell-free; that is a demonstration that the fixes are sufficient to zero the score, **not** a discovery that designed pages in the wild score 0 (they will not, real designed pages make a few defensible default choices). The honest, load-bearing result is the **confound-controlled refactor** and the **separation**: applying the documented fixes moves the score from F to A, and the detector cleanly distinguishes the two families. Scoring live sites in the wild requires resolving external stylesheets, which the single-document detector does not yet do (§9).

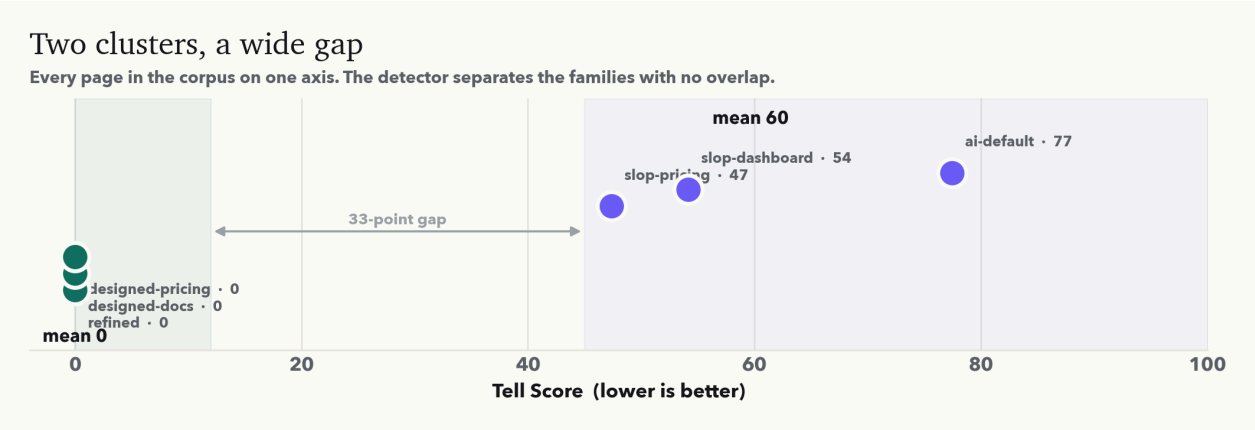

Figure 4: Every page in the corpus on one axis. The detector separates AI-default (mean 60) from designed (mean 0) with no overlap.

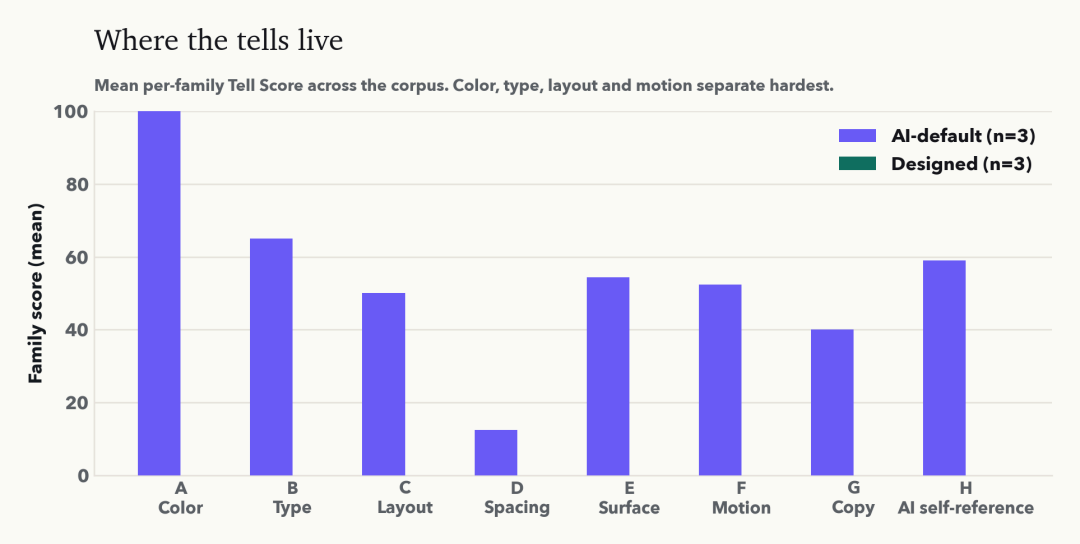

Figure 5: Mean per-family Tell Score across the corpus. Color, type, layout and motion separate the families hardest, matching the human experience of what gives a page away.

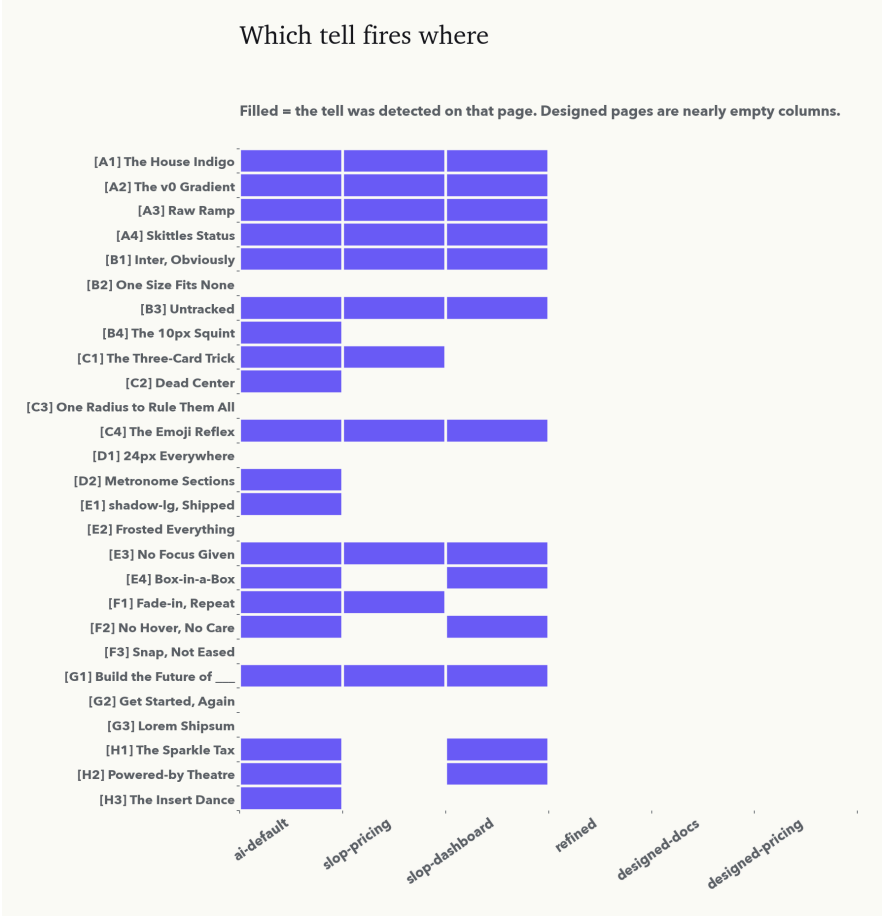

Figure 6: Which tell fires where. Filled cells are detected tells; the three designed pages are near-empty columns.

7 Recalibration and validation on 202 real sites

The results so far rest on authored fixtures. A fair reader objects: maybe the detector is not a discriminator but a machine that flags **any** page, calling the whole web AI. The only way to answer is to point it at real, human-crafted design at scale and see whether the best work in the world scores low. It must, or the instrument is worthless.

7.1 A live-DOM corpus

We curated **202 design-led, human-crafted sites** across categories: developer tools and infrastructure (Stripe, Vercel, Linear, Supabase, Railway, Cloudflare), AI labs (Anthropic, OpenAI, Perplexity), fintech (Toss, Mercury, Ramp, Brex), design and productivity (Figma, Notion, Framer, Canva, Raycast), and consumer brands (Apple, Airbnb, Nike, Spotify, Duolingo). For each we drive headless Chrome (Playwright), let the CSS and webfonts fully apply, and read the **computed styles** off the live DOM, the ground truth a static parse of one file cannot see. This is also the upgrade that lets the detector audit a deployed URL, not only a self-contained document, removing v1′s central limitation.

7.2 What the data overturns

Three of the v1 tells, taken from the literature, are **wrong as stated** once measured against real design (Table 2). The corpus median is **10 distinct font sizes** (p90 = 15), so "more than nine sizes means no scale" would flag nearly every top site; the real signal is only the degenerate one-or-two-size case. **A third of top sites use a purple accent** and Stripe paints 123 of them, so purple hue cannot be a tell; only the exact AI-default indigo ramp, or purple with no compensating craft, discriminates. **A quarter of top sites set Inter or the system stack as their primary face** (Linear among them), so the font alone is not the tell, it is the font with no optical tracking and no type system.

Signal	Human corpus (n=202)	Recalibration
Distinct font sizes	median 10, p90 15	drop the high-count rule; keep only 1-2 sizes
Purple accent present	33% of sites; Stripe 123	exact default-indigo only, else craft-gated
Generic primary font	24% of sites (incl. Linear)	tell only with no tracking and no scale
Distinct border-radii	median 6, p90 12	one-radius tell fires at <= 2
Centered block fraction	median 0.02, p90 0.16	center-everything tell fires at >= 0.5
Optical tracking on display	78% of big headings	confirmed as a craft signal
`:focus-visible` present	96% (where CSS readable)	fire only when CSS is readable

Table 2: Where the literature's tells meet 202 real sites. Thresholds are re-anchored to the human distribution, so top sites sit near zero by data, not by assumption.

7.3 The craft-credit model

The deeper lesson is that **no single signal is the tell**. Stripe has purple and blue-to-purple gradients; Linear has Inter; both are paragons. What separates them from slop is that their cosmetic choices sit on top of real craft. We encode this directly: a page earns a **craft credit** for each of a custom display face, optical tracking on display type, a radius hierarchy, and a designed focus state. Cosmetic tells (purple, generic font, diffuse shadow) fire only when craft credits are absent; structural tells (one radius, center-everything, emoji iconography, missing focus where measurable, vague copy) stand on their own. The AI look is the co-occurrence of defaults **with nothing compensating**, which is exactly what the fixtures have and what no top site has (Figure 7).

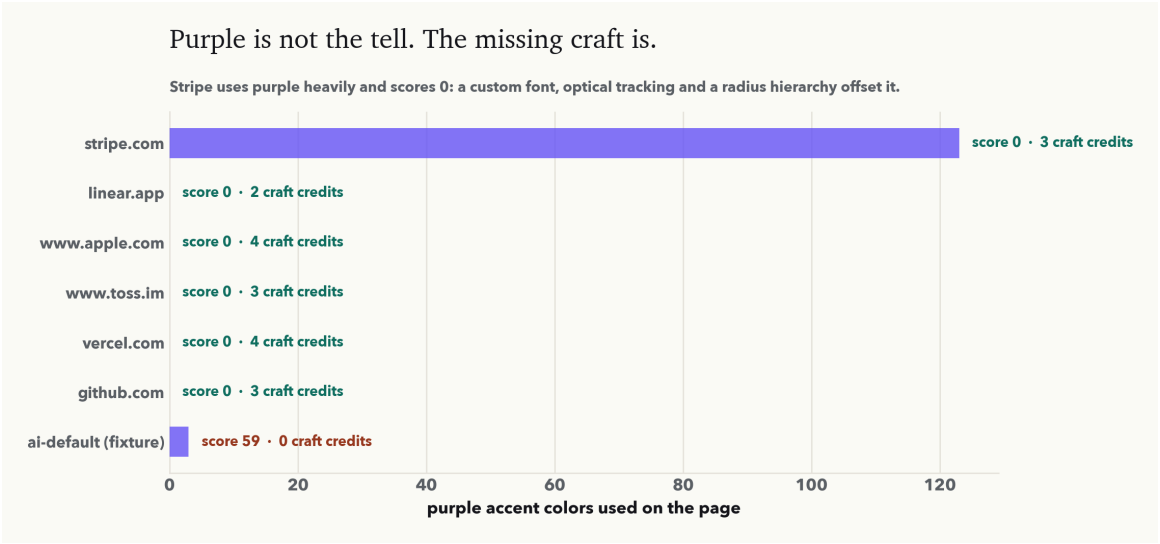

Figure 7: Purple is not the tell. Stripe uses 123 purple accents and scores 0; three craft credits (a custom face, optical tracking, a radius hierarchy) offset it. The AI-default fixture has three purple accents and scores 59, because it has no compensating craft.

7.4 Validation

Recalibrated, the detector scores the 202 real sites at a **median of 0** (mean 2.5, p90 10.2); **93% earn grade A** and none reads as slop (Figure 8). The same detector scores the AI-default fixtures at 35 to 59. The few real sites that score above 20 are the genuinely plain ones (a text-first blog, a spec page), which is the correct call, not a false positive. The empirical distributions behind the thresholds are shown in Figure 9; the diversity the score rewards is shown in Figure 10, a montage of 28 of the sites, each having made different choices.

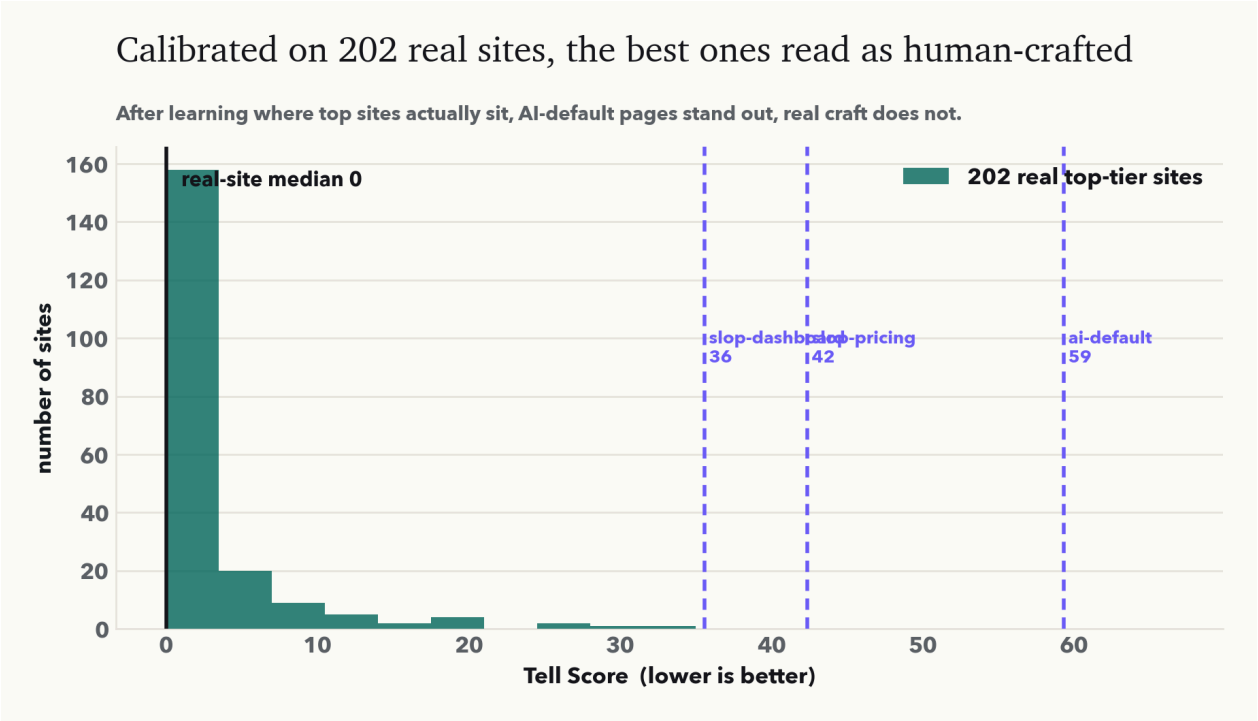

Figure 8: The validation. 202 real top-tier sites cluster at a Tell Score near 0 (median 0); the AI-default pages stand far out at 35 to 59. The instrument distinguishes machine-default from human-crafted on real data, not just fixtures.

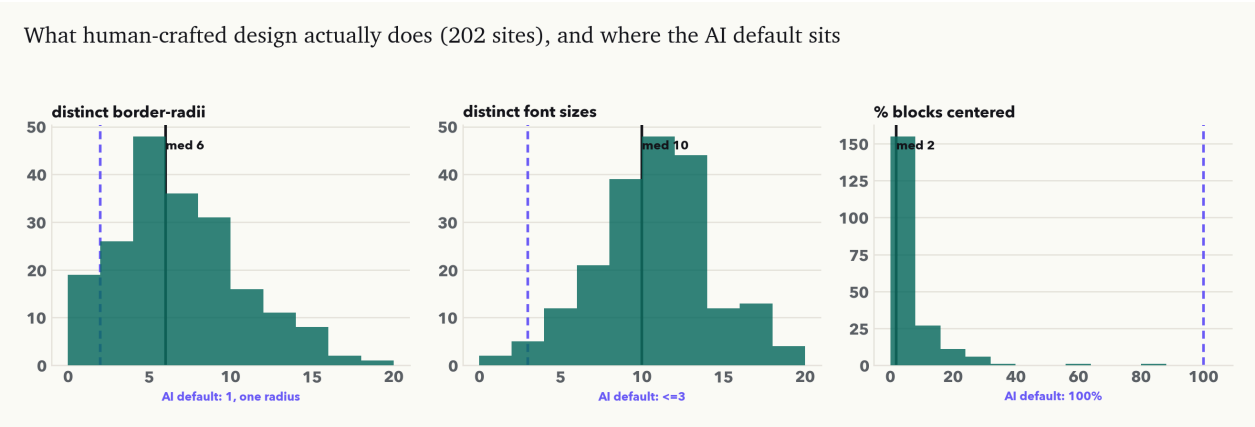

Figure 9: The learned distributions. Real sites use many radii and many type sizes and almost never center everything; the AI default (dashed) sits at the tail of each. Thresholds are anchored here.

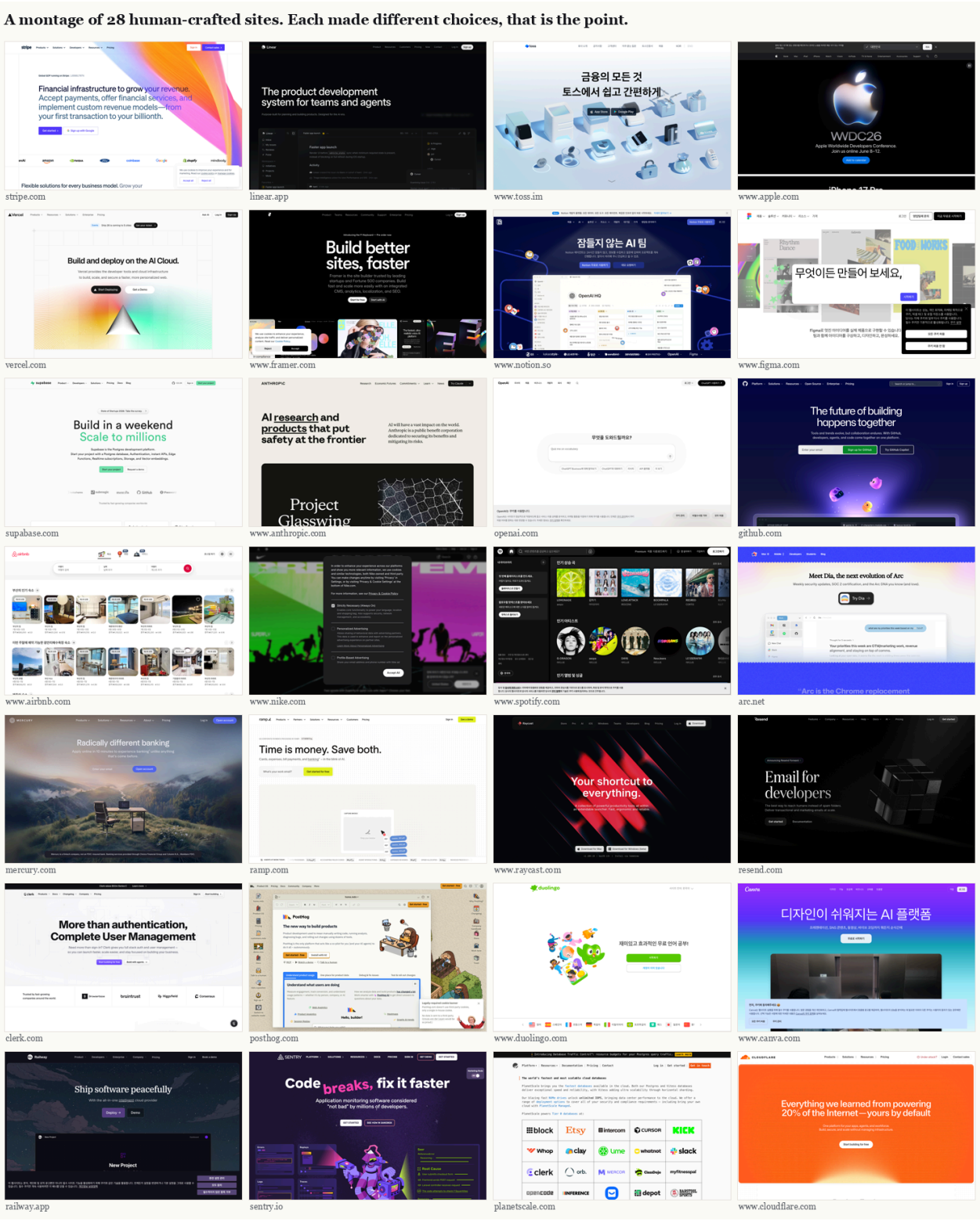

Figure 10: 28 of the 202 human-crafted sites. Dark and light, serif and grotesque, dense and airy: each made a different set of choices. That variance is precisely what the AI default erases and what a low Tell Score is meant to protect.

7.5 Auditing a live site

The recalibrated model runs over computed styles, so it audits a deployed URL. `python scripts/audit_url.py https://your-site.com` renders the page, scores it, and lists the craft credits offsetting its cosmetic defaults; the same is exposed as the MCP tool `audit_url`. On the corpus this is the difference between

a tool that lectures from a style guide and one that can look at the running product and say, with evidence, whether anyone decided anything.

8 From negative to positive: a component-spec catalog

A detector is a negative instrument: it says what *not* to do. The recurring question from anyone trying to use it to actually build is the positive one, *then what numbers should I use?* To answer it without inventing a house style, we read the answer off the same kind of evidence. We re-rendered **199** design-led sites a second time, but instead of aggregate signals we recorded the *concrete CSS values* they ship per component, and we did it twice per site, once under `prefers-color-scheme: light` and once under `dark`, to capture both palettes (`src/scrape_detail.py`, aggregated by `scripts/build_spec_catalog.py`).

The result (Figure 11, full tables in `reference/COMPONENT-SPECS.md`) is a set of empirical targets. Primary-button corner radius does not converge on one value; it splits between a soft-rounded 8 to 12`px` cluster and a full pill, with sharp 0`px` corners a deliberate minority and only **13%** of buttons carrying any shadow. The type scale lands near 64/48/32`px` for `h1`/`h2`/`h3` over a 16`px` body, headlines on a tight ~ 1.1 line-height with frequent negative tracking, body around 1.5. Content containers center on a **1200px** median; vertical section rhythm on ~ 64`px`. Spacing follows a 4/8`px` grid, but only loosely: 70% of padding values are multiples of four, not the religious adherence a generator assumes.

Two findings matter for the thesis. First, accent color is genuinely unconstrained: the most common brand hues each appear once across the corpus, scattered around the wheel, which is the positive form of "the hue is never the tell." Second, dark mode has a grammar. Of the sites that repaint automatically for the OS, the page background is almost never pure `#000000`; it is a *tinted* near-black (`#0b0f19`, `#111111`, `#18181b`, and similar), raised surfaces sit a step lighter rather than a hairline border away, and text is an off-white rather than pure `#ffffff`. Pure black on pure white, in either mode, is itself a tell.

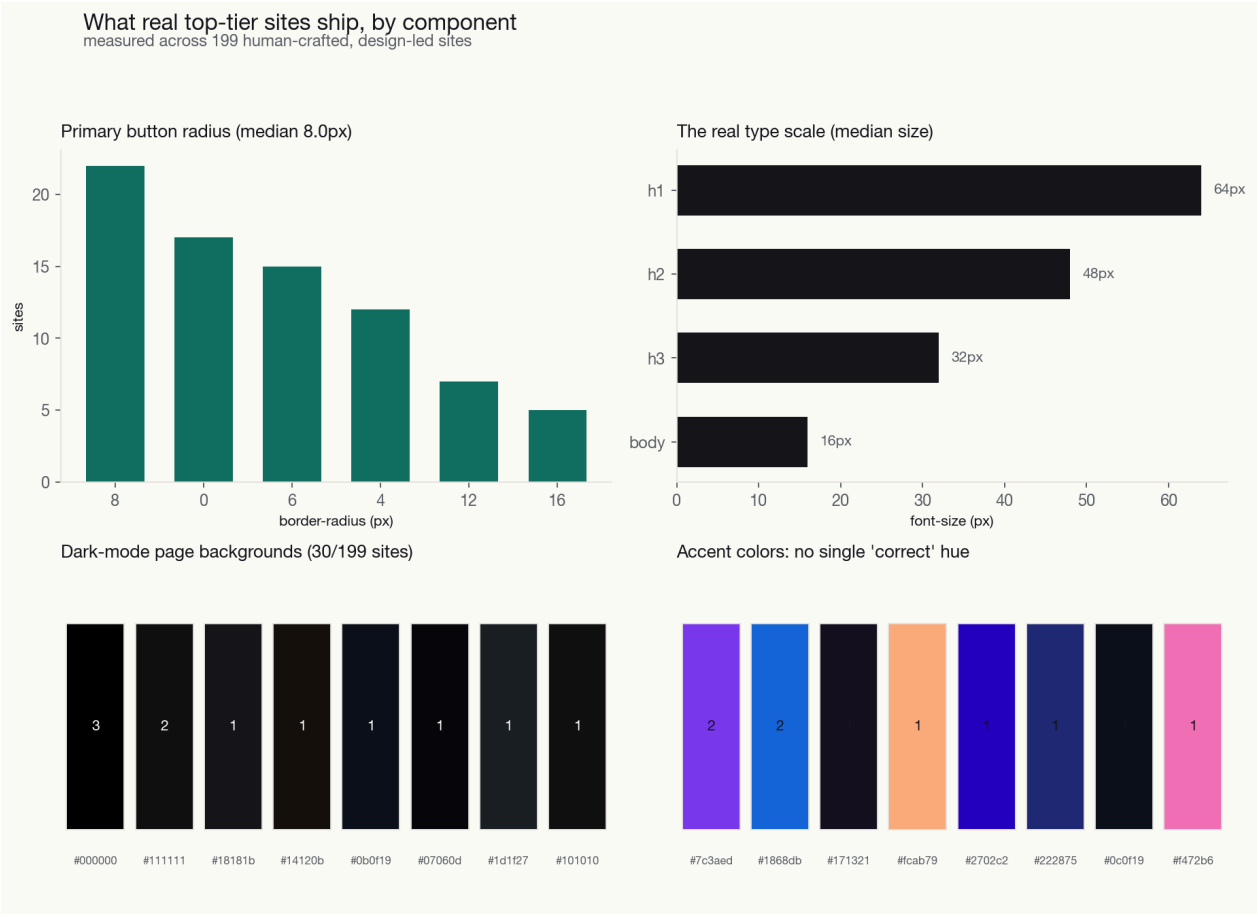

Figure 11: Measured component specs across 199 design-led sites: primary-button radius distribution, the real type scale, the actual dark-mode page backgrounds (tinted near-blacks, not pure black), and the spread of accent hues.

To show the targets are not just descriptive but buildable, we assembled one landing page directly from the catalog medians, `fixtures/catalog-sample.html`: a 1200px container, a 64/48/32px type scale on a serif display, 8px buttons with a full set of microstates, an owned amber accent rather than the default indigo, and a dark mode that follows the measured grammar (a tinted `#0e1014` near-black, surfaces a step lighter, off-white text). The page scores a Tell Score of 0 in both palettes (Figure 12). The same file is the v3 entry in the template gallery.

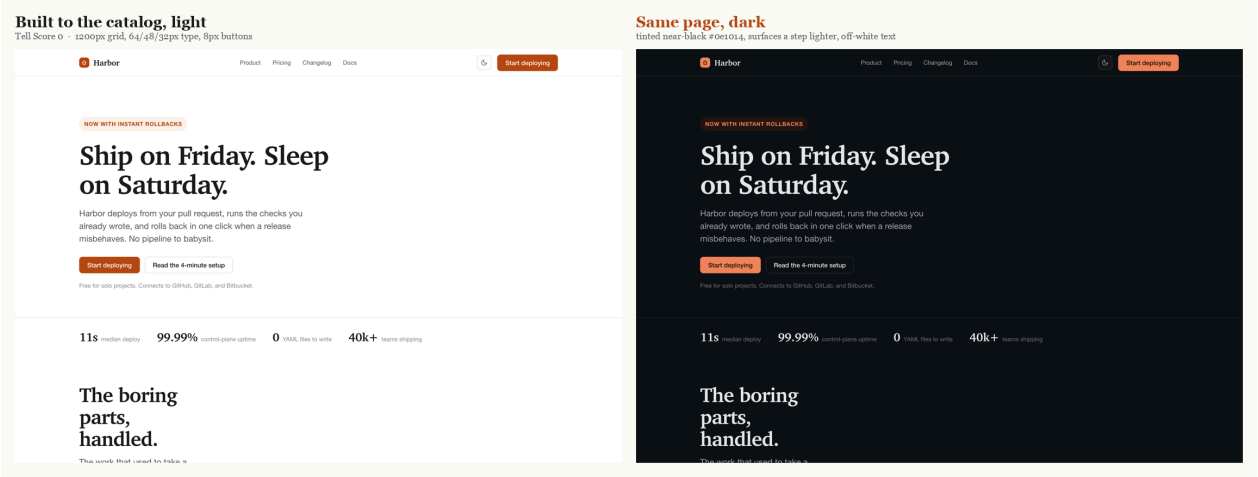

Figure 12: A landing page built straight from the catalog medians, shown in light and dark from a single self-contained file. It scores 0. The dark palette is the measured grammar, a tinted near-black with surfaces a step lighter, not an invert to pure black.

These numbers feed back into the harness as concrete targets, so the prompt module tells a builder not only what to avoid but what range to aim inside.

8.1 A Korean-web companion, and what actually differs

The catalog above is mostly Western. Hangul is set differently from Latin, so we ran the same extraction over **48 Korean design-led sites** (Toss, Kakao, 당근, 무신사, 29CM, 오늘의집, 배민, 업비트, and more) and compared (`reference/KOREAN-SPECS.md`, Figure 13). The honest result separates a real difference from a folk one. The real differences are two. First, the default face: Pretendard is the Korean web's Inter, on 44% of these sites as the body face and a hangul sans on 69%, so the type tell translates directly, bare Pretendard with no scale reads as machine-default exactly as bare Inter does. Second, body size: Korean body text sets smaller, a 14`px` median against 16`px` globally, because hangul carries more ink per glyph and Korean product culture is denser. The folk difference is line-height: hangul is often said to need much more leading, but the measured median is ~ 1.5 in both corpora, because Pretendard already ships generous leading. One caveat is honest to flag: the Korean h1 median is bimodal, design-led product sites use 56 to 90`px` heroes like the West while portals and commerce are banner-driven with small or absent display headlines, so the low aggregate is a density culture, not a different notion of a headline.

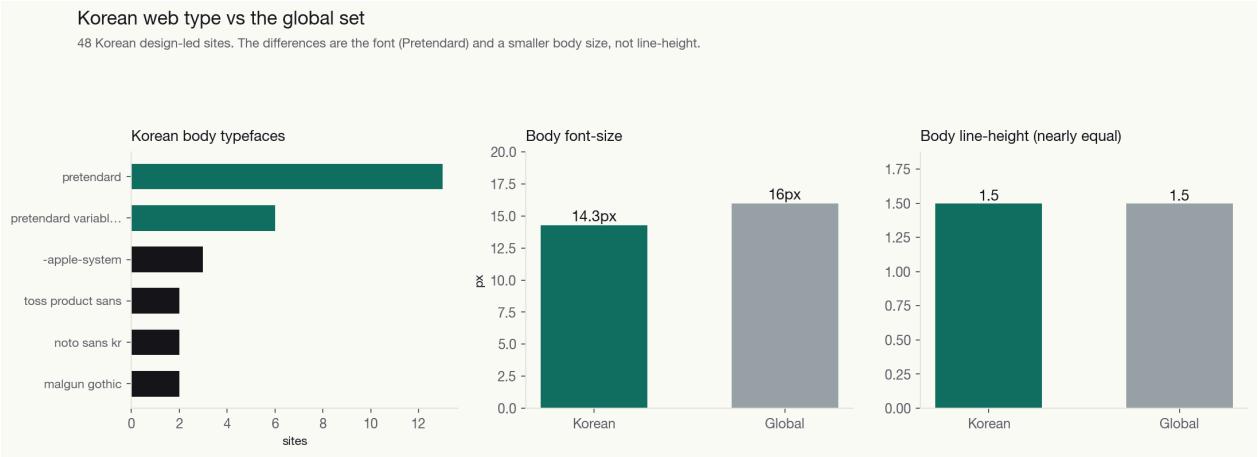

Figure 13: Korean web type vs the global set across 48 Korean sites. The differences are the font (Pretendard dominance) and a smaller body size; line-height is nearly identical.

The taxonomy itself is not language-bound, the same tells apply, but a builder targeting a Korean audience should read the catalog's body-size and font rows through this companion.

9 Field evidence: two production design manifestos

Sections 4 through 8 built the taxonomy from public craft writing and from the statistics of 202 scraped sites. The sharpest test of whether the taxonomy names the *right* things came from the opposite direction: two production codebases whose maintainers, independently of this work, had already written their own internal design manifestos titled, in effect, "avoid the AI look." Both are private commercial products, so we treat them as anonymized field evidence: *Manifesto A*, a dark-mode media tool built on a Toss-style minimalism with a single owned accent, and *Manifesto B*, a productivity assistant on Pretendard with a strictly neutral hierarchy. Each manifesto is a versioned document in its repository, paired with a remediation log that counts instances: one team patched roughly 600 sub-12`px` labels up to a 12`px` floor; the other enumerated every `Sparkles` and `Wand2` icon slated for removal.

Two findings matter. First, *convergent validity*: both manifestos independently name tells the detector already had, the indigo default and the blue-to-purple gradient (banned as "AI 풍 그라데이션"), emoji-as-iconography, the generic font (both commit to a distinctive or commissioned face), and vague system-voice copy. Practitioners who never saw this taxonomy, writing only to stop their own products from looking generated, arrived at the same list. That is the strongest evidence we have that the families are real and not an artifact of our framing.

Second, *the field saw further*. Both manifestos lead with a register the v1 to v3 taxonomy did not cover at all: the interface announcing itself as AI. We add it as a new family, **H, AI self-reference**, with three tells, and three more in existing families (Figure 14):

H1, AI-cliché iconography. The `Sparkles` / `Wand` / `Bot` / `Brain` / `Cpu` icon set bolted onto any "AI" feature. *Manifesto B ranks it the single loudest tell.* The fix is a function-true icon or the product's own brand mark.

H2, labels the feature "AI" or names the model. "AI-powered", "AI 분석", or an exposed `GPT-4` / `Claude` / `OpenAI` in the interface. Users care about the outcome, not the engine; name the function by what it does and reveal the model only in settings.

H3, generate then preview then insert. The assistant-panel ceremony (produce a result, show a preview card, ask the user to "Insert"). Apply the result into the content directly and let the user undo. *(smell; the weakest of the three.)*

A4, multi-color pill-badge inflation. A row of status pills each in a different bright hue. When everything is colored, color stops carrying meaning.

B4, sub-legible micro-type. Scattered 9 to 11`px` labels. Below ~ 12`px` both hangul and dense Latin lose legibility; build hierarchy with weight, not by shrinking.

E4, nested card-in-card chrome. The "double box", a bordered rounded card wrapped directly inside another. Use one outer card with a flat divided list.

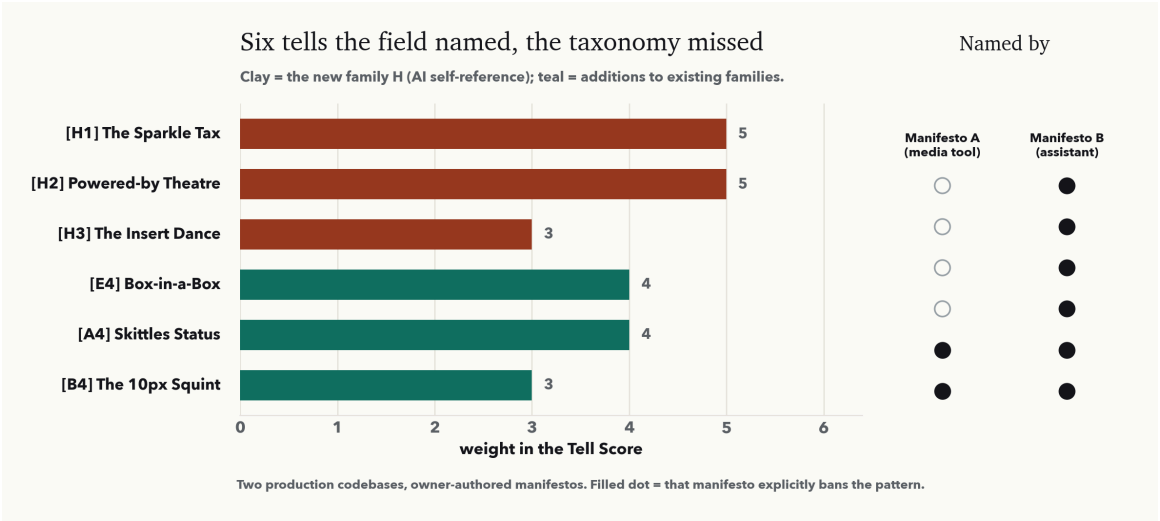

Figure 14: The six field-derived tells, their weights, and which manifesto named each. The new family H (clay) is the AI-self-reference register; A4/B4/E4 (teal) are additions to existing families. Manifesto B, the assistant, enumerates the self-reference family in the most detail; both name the color and type-floor tells.

These six raise the taxonomy to 27 tells and the maximum weight to 133. They are detected by the **file/code detector** (`score_design`, run on the markup an agent just wrote), where icon imports, label text, and nesting are visible; they are not yet wired into the live computed-style audit, which sees rendered CSS but not component names. We state that as a boundary, not a result. Two honest caveats on the evidence itself: the two codebases are a convenience sample (products the authors had access to), not a random draw, so they establish that practitioners converge on these tells, not how common the manifestos are; and both are Korean products, which is why the self-reference tells carry Korean phrasings alongside the English. The mechanism, a model marking its own output as AI, is not language-specific, and the detector matches both.

10 The harness

The taxonomy ships as a usable tool in three forms, all generated from `src/taxonomy.py` so guidance can never drift from measurement.

CLI. `python src/cli.py page.html` scores a file, prints the fired tells with fixes, and returns the score as its exit code for CI gating. A `--quiet` mode prints a leaderboard across many files.

MCP server. `mcp/server.py` exposes `score_design(html)`, `score_file(path)`, `audit_url(url)`, `list_tells()`, and `harness_prompt()` over the Model Context Protocol. An agent can therefore **audit the UI it just wrote before showing it to the user** (or audit a deployed URL with `audit_url`), receive the specific fixes, and iterate, closing the loop without a human in the middle for the mechanical part.

Drop-in prompt module. `harness/AI-DESIGN-TELLS.md` turns each *detected* tell into a *preventive* instruction ("Don't: Inter default. Do instead: commit to a display face …") plus a pre-ship checklist, and now carries the measured component-spec targets from the 199-site catalog so the guidance is two-sided: what to avoid, and the concrete range to aim inside. Pasted into a system prompt, CLAUDE.md or a builder's custom-instructions field, it pre-empts the look at generation time; the detector then verifies the output. Prevention and detection share one source of truth.

The intended workflow is *generate* → *score* → *fix* → *re-score*, with the prompt module shifting the starting score down and the detector certifying the result.

11 Threats to validity, and the epistemics of measuring a look

11.1 Threats to validity

A discriminator, not a beauty judge. The Tell Score measures machine-default- ness, not quality. Low is necessary, not sufficient, for good design; the score cannot see whether the authored choices are *good*, only that choices were made. We name the metric "Tell Score," not "Design Score," for this reason.

Authored fixtures (now backed by real data). The headline refactor still uses controlled fixtures, where the designed page scores 0 by construction. That alone would be weak, which is why §7 validates against 202 real human-crafted sites that were **not** authored by us and yet score at a median of 0: the low scores are a property of real craft, not an artifact of our fixtures.

Live auditing, with one residual blind spot. v2 removes the single-file limit: the recalibrated detector renders the page and reads computed styles, so it audits deployed URLs (§7). One blind spot remains, and we measure it rather than hide it: `:focus-visible` is read from `cssRules`, which the browser blocks for cross-origin stylesheets, so on a site whose CSS is served from another origin (e.g. Stripe, 0 readable sheets) the focus tell is **not measurable**. We confidence-gate it, firing only when the CSS is readable, so the failure mode is a missed tell, never a false accusation.

The look is time-bound. The tells are the defaults of mid-2020s models. As training distributions shift, as today's "designed" choices themselves become over-represented, the taxonomy will need revision. The method (enumerate the current mode, weight it, measure it) outlives any particular list.

Goodhart. If the score becomes a target, it can be gamed: swap indigo for a brand teal, Inter for any non-default face, and pass while remaining lazy in every way the detector cannot see. We mitigate by weighting strong structural and microstate tells above easy cosmetic ones, but no static metric is Goodhart-proof [20]. The score is a floor on intentionality, not a ceiling on craft.

11.2 What "looking human" actually is

It is worth asking what we are really detecting. The honest answer is *not* humanity, a human using the same defaults produces the same tells, and a model given constraints produces none. What the score detects is **intentionality made legible in the artifact**: the visible residue of decisions. Taste, on this view, is not a faculty the model lacks so much as a *compression of a person's accumulated choices*, every interface they have loved, every alignment they have nudged, into a prior the median of a corpus cannot represent. The model returns the mean; a person returns a sample from their own much narrower, much weirder distribution. The tells are precisely the places where the mean shows through.

This reframes the practitioner's complaint. "It looks like AI" does not mean "a machine made it"; it means "no one in particular made it." The gradient is not ugly; it is *unowned*. Removing the tells is the act of putting an owner back into the page.

11.3 The map is not the territory

The Tell Score is a map. It is useful exactly to the degree that it stays a map, a low-dimensional, legible projection of a high-dimensional thing (whether a page reads as authored). The danger of any such map is that optimizing it can pull the territory toward the map's blind spots: a page engineered to score 0 while being, in every unmeasured respect, thoughtless. We have tried to keep the map honest by making every point traceable to a quoted piece of evidence and a real design principle, so that "improving the score" and "making a real decision" coincide as often as possible. But the reader should hold the number lightly. The purpose of the instrument is not the number; it is to make a previously ineffable judgment *discussable*, to let a team point at a specific tell and argue about it, which is something taste alone never allowed.

11.4 The second-order convergence

There is a final irony the companion study sharpens. *Convergence Pressure* finds that what homogenizes a population is not AI assistance but the **reflective loop** , everyone consulting the same oracle and converging on its center [14], [15]. A single, widely adopted design score is such an oracle. If every team optimizes the same Tell Score with the same fixes, the escape from the indigo mean could become a new mean, a monoculture of "de-slopped" pages as recognizable, in time, as the slop they replaced. The taxonomy cannot prevent this; only the diversity of human intent can. The right use of the harness is therefore the one we have argued for throughout: as a detector of the *absence* of decision, a prompt to make a choice, not as a prescription of which choice to make. The score should send you toward your own distribution, not toward a new shared one.

12 Conclusion

The AI look is not taste and not mystery; it is a finite set of statistical defaults, and finiteness makes it measurable. We enumerated 27 tells, weighted them, and built a transparent detector that scores any page in $[0, 100]$. Holding content fixed and changing only the tells, a canonical AI page moves from 77 to 0; across a corpus the detector separates machine-default from designed with no overlap. The same taxonomy ships as a CLI, an MCP server, and a drop-in prompt, so the judgment "this looks like AI" becomes an auditable, fixable, preventable property rather than a sigh. The instrument's value is not its number but its effect: it turns an ineffable complaint into a list of decisions waiting to be made, and asks that you, not the median, make them.

Data and code. Taxonomy, detector, fixtures, figures, paper, and harness (CLI, MCP server with live `audit_url`, drop-in prompt), and the 202-site corpus (signals + scores) are open at github.com/hankimis/ai-design-tells. Reproduce with `python scripts/run_audit.py` (fixtures), `python src/scrape.py` (re-scrape), `python scripts/analyze_corpus.py` (learn the calibration), and `python scripts/make_figures.py && python scripts/make_figures_v2.py` (figures). Companion study: *Convergence Pressure* [15].

References

[1] 925 Studios, "AI Slop Web Design: A Guide to Spotting and Fixing Generic Websites." [Online]. Available: https://www.925studios.co/blog/ai-slop-web-design-guide

[2] AXE-WEB, "Why AI Websites All Look the Same (And When It Matters)." [Online]. Available: https://axe-web.com/insights/ai-website-design-sameness/

[3] R. Prashant, "Why Your AI Keeps Building the Same Purple Gradient Website." [Online]. Available: https://prg.sh/ramblings/Why-Your-AI-Keeps-Building-the-Same-Purple-Gradient-Website

[4] A. Wathan, "On Tailwind UI defaulting every button to bg-indigo-500." 2024.

[5] K. Ni, "Design Observation: Why Do AI-Generated Websites Always Favour Blue-Purple Gradients?." [Online]. Available: https://medium.com/@kai.ni

[6] J. Pearce, "Where does that purple gradient come from?." [Online]. Available: https://www.jackpearce.co.uk/notes/purple-gradient-ai-aesthetics/

[7] A. Wathan and S. Schoger, *Refactoring UI*. Self-published, 2018. [Online]. Available: https://refactoringui.com/

[8] K. Saarinen and Linear, "How we redesigned the Linear UI; behind the latest design refresh." [Online]. Available: https://linear.app/now/how-we-redesigned-the-linear-ui

[9] Mantlr, "How Stripe, Linear, and Vercel Ship Premium UI." [Online]. Available: https://mantlr.com/blog/stripe-linear-vercel-premium-ui

[10] Toss, "The 8 Writing Principles of Toss." [Online]. Available: https://toss.tech/article/8-writing-principles-of-toss

[11] D. Rams, "Ten Principles for Good Design." [Online]. Available: https://www.vitsoe.com/us/about/good-design

[12] J. Nielsen, "10 Usability Heuristics for User Interface Design." [Online]. Available: https://www.nngroup.com/articles/ten-usability-heuristics/

[13] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, "AI models collapse when trained on recursively generated data," *Nature*, vol. 631, pp. 755–759, 2024, doi: 10.1038/s41586-024-07566-y.

[14] A. R. Doshi and O. P. Hauser, "Generative AI enhances individual creativity but reduces the collective diversity of novel content," *Science Advances*, vol. 10, no. 28, p. eadn5290, 2024, doi: 10.1126/sciadv.adn5290.

[15] H. Kim, "Convergence Pressure: Measuring AI-Mediated Cultural Homogenization in Iterated Creation." [Online]. Available: https://github.com/hankimis/convergence-pressure

[16] Pixeldarts, "Four design principles behind Stripe, Linear, and Vercel." [Online]. Available: https://www.pixeldarts.com/en/post/four-design-principles-behind-stripe-linear-and-vercel

[17] A. Xu and LogRocket, "Linear design: the SaaS design trend that is boring and bettering UI." [Online]. Available: https://blog.logrocket.com/ux-design/linear-design/

[18] Toss, "The Toss Design System and the error-message system." [Online]. Available: https://toss.tech/article/introducing-toss-error-message-system

[19] Anthropic, "Prompting for Frontend Aesthetics." [Online]. Available: https://platform.claude.com/cookbook/coding-prompting-for-frontend-aesthetics

[20] M. Strathern, "Improving ratings: audit in the British university system." 1997.