RIVAL RESEARCH · 2026
VOLUME 04 — PREVIEW

The Persona Impact Study

One model. One task. 52 system prompts.
We held everything constant except the persona
and measured what actually moves the needle.

GEMMA 4 31B · 156 GENERATIONS · 3× BLINDED OPUS 4.7 JUDGES

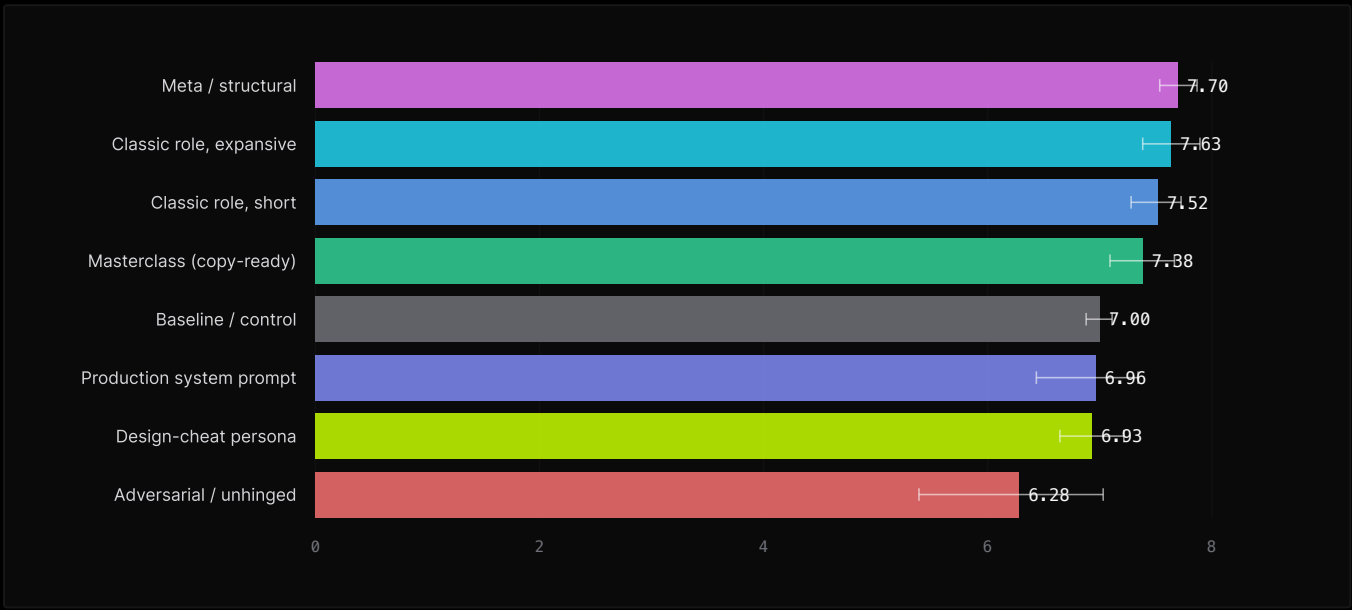
$F = 4.14 \cdot H^2 = 0.164 \cdot A = 0.803$

FINDINGS

What we found

7.70	7.00	4.14	0.164
WINNING BUCKET	BASELINE	ANOVA F	EFFECT SIZE H^2

Bucket leaderboard · composite score · 95% bootstrap CIs



The headline

Across 156 generations from Gemma 4 31B on a fixed luxury-real-estate landing-page task, the **Meta / structural** bucket scored highest (7.70) and the **Adversarial / unhinged** bucket scored lowest (6.28). Between-bucket ANOVA returned F = 4.14 on (7, 148) with η^2 = 0.164. Mean Krippendorff's α across the three judge waves was 0.803, indicating acceptable inter-rater reliability.

The single most counter-intuitive result: the **design-cheat** bucket (6.93), engineered to dominate by loading the prompt with Refactoring UI rules, Tufte principles, WCAG AA, Tailwind scale, 8-pt grids, and modular type ratios, scored **below the blank baseline** (7.00). What won instead: reasoning scaffolds (7.70) and terse role assignments like "Figma designer" or "Apple CPO".

Inter-rater reliability · Krippendorff's α per axis

AXIS	A	INTERPRETATION
design quality	**0.889**	very good agreement
copy quality	**0.878**	very good agreement
information architecture	**0.796**	acceptable agreement
brief adherence	**0.620**	moderate agreement
originality	**0.824**	very good agreement
polish	**0.808**	very good agreement

METHOD

How the study was run

Why a small model on purpose

Gemma 4 31B sits below the frontier. If we had run this on Claude Opus 4.7 or GPT-5, baseline quality would have been high in every condition and the persona signal would have drowned in the model's default competence. Picking a model below the frontier leaves visible headroom, and headroom is what makes the question measurable. The question is not whether Gemma is good. It is how much of the headroom can a well-designed persona actually recover.

Fixed task

Every persona received the same user message: build a single-file HTML landing page for a fictional AI-native CRM for luxury real estate brokers called Keystone. Eight required sections (hero, three feature sections, three-tier pricing, social proof, FAQ, footer). Inline styles only. No external assets. Single self-contained HTML.

Independent variable

Persona / system prompt only. 52 personas across 8 buckets:

- **Masterclass (copy-ready)**, n=7
- **Meta / structural**, n=4
- **Classic role, expansive**, n=6
- **Classic role, short**, n=8
- **Design-cheat persona**, n=8
- **Production system prompt**, n=10
- **Baseline / control**, n=3
- **Adversarial / unhinged**, n=6

Controls and confound mitigation

- **Blinded judging.** Each response was saved under an opaque SHA256 hash. The three judge waves received hashes and files, never persona labels. Prompt-injection defense: <script>, <meta>, and HTML comments were stripped before the judge read the file, so a response could not tell the judge what score to give.
- **Randomized batch order per wave.** Each of the three passes used a different seeded shuffle of the 156 response hashes, so no judge saw the dataset in the same order.
- **Font rendering.** Puppeteer screenshots were taken with a pinned font stack and --font-render-hinting=none to avoid OS drift.
- **Rate-of-call drift.** All 156 generations completed within a single 24-hour window to limit OpenRouter / provider-side model version changes.

Statistical plan (pre-registered)

- One-way ANOVA across buckets on composite score.
- Bootstrap 95% CIs (10,000 resamples) on every reported mean.
- Cohen's d for every pairwise bucket comparison.
- Krippendorff's α (interval) across the three judge waves, per axis.

Model under test

MODEL	google/gemma-4-31b-it via OpenRouter
FALLBACK	google/gemma-4-31b-it:free
TEMPERATURE	0.7
MAX TOKENS	8192
SAMPLES PER PERSONA	3
TOTAL GENERATIONS	156

Judges

JUDGE MODEL	Claude Opus 4.7
DISPATCH	Parallel Claude Code subagents, 33 agents per wave
WAVES	3 independent passes (J1, J2, J3)
RUBRIC	6-axis anchored Likert, 1-10 per axis
PER-WAVE BLINDING	Filename hash only; manifest access explicitly forbidden

RUBRIC

The complete scoring rubric

Each response was scored by three independent agents on six axes. Scores are weighted into a composite using the weights below.

design quality · weight 0.25

Visual coherence, hierarchy, typographic rhythm, spatial taste, use of color. How good does the rendered page actually look?

ANCHOR 1	Visually broken or accidentally ugly. Clashing colors, inconsistent spacing, default browser styles, no hierarchy.
ANCHOR 5	Competent but unremarkable. Consistent spacing and color, readable hierarchy, nothing actively wrong. Looks like a Bootstrap tutorial from 2020.
ANCHOR 10	Design you would pay for. Intentional type scale, restrained palette used confidently, clear rhythm, deliberate whitespace, evidence of taste. Could headline Dribbble.

copy quality · weight 0.15

CTA clarity, voice, concision, benefit-orientation. Is the writing sharp or filler?

ANCHOR 1	Lorem-ipsum vibes. Placeholder copy, vague claims, passive voice, feature lists dressed as benefits, generic CTAs like 'Click here'.
ANCHOR 5	Serviceable. Real sentences, plausible benefits, standard CTAs like 'Get started' or 'Learn more'. Nothing memorable, nothing wrong.
ANCHOR 10	Tight, specific, earned. CTAs are verbs with objects. Benefits are concrete, numbers are used correctly, voice is consistent. Reads like a top SaaS page.

information architecture · weight 0.15

Hierarchy legibility, scannability, order of sections, clarity of navigation within the page.

ANCHOR 1	Wall of undifferentiated content. No visual hierarchy, sections blur together, required sections missing or in bizarre order.
ANCHOR 5	Standard vertical stack, sections clearly demarcated, all required sections present in a sensible order. No navigation aids beyond section breaks.
ANCHOR 10	Every section earns its place. Clear narrative arc from hero to footer. Sub-hierarchy within sections is readable at a glance. Uses visual weight to guide the eye.

brief adherence · weight 0.2

Did the response do what the brief asked — all 8 required sections present, pricing correct, no external assets, single-file HTML?

ANCHOR 1	Major omissions. Missing multiple required sections, wrong product, external asset URLs, multi-file output, or fictional pricing.
ANCHOR 5	Most elements present. May be missing one sub-section or have a minor external asset. Pricing and product name correct.
ANCHOR 10	Complete compliance. All 8 sections present, correct pricing in all three tiers, no external assets whatsoever, truly self-contained.

originality · weight 0.1

Distinctive vs generic. Does this look like a unique voice, or is it the median AI-generated SaaS page?

ANCHOR 1	Indistinguishable from ten thousand generic AI-generated SaaS pages. Purple-to-blue gradient hero, three icon cards, rocket emoji.
ANCHOR 5	Has one or two distinctive elements but mostly plays it safe. A novel layout choice or an unusual color use breaks up otherwise-generic output.
ANCHOR 10	Has a point of view. Visual and copy choices feel deliberate and specific to this product, not template-fill. Memorable.

polish · weight 0.15

Finish, attention to detail, interaction states, micro-copy, edge alignment, typography nits.

ANCHOR 1	Visibly unfinished. Broken alignment, typos, inconsistent border radii, hover states missing, footer links that lead nowhere stated.
ANCHOR 5	Clean at first glance. Consistent radii, aligned edges, readable type. Lacks micro-details like hover states, legal microcopy, form affordances.
ANCHOR 10	Every small thing is handled. Consistent interaction states, well-designed form fields, trust microcopy near CTAs, clean footer, considered edge cases.

TASK PROMPT

The exact user message every persona received

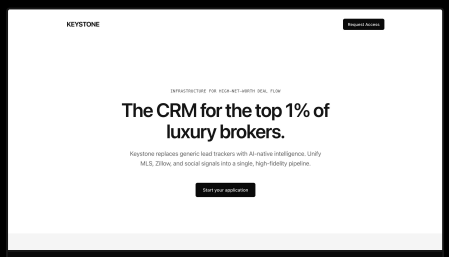
```
Design and build a single-file HTML landing page for **Keystone**, an AI-native CRM and deal-flow platform built for top-producing luxury real estate brokers.

About Keystone:
- Unifies client activity across MLS, Zillow, and social channels
- Surfaces buyer intent signals with AI (predictive lead scoring, listing-match notifications, contract-stage nudges)
- Lives natively on phone and CarPlay — the places luxury agents actually work
- Powers over $4B in annual luxury real estate transactions on the platform
- Premium pricing: $2,400 per seat per year, positioned against $600/seat/yr generic CRMs

The page must include:
1. A hero section with a primary call-to-action
2. Three feature sections (AI Client Intelligence, Deal-Flow Pipeline, Mobile-First Workflow)
3. A pricing block with three tiers (Producer $2,400/yr, Team $7,200/yr, Brokerage custom)
4. A social-proof or testimonial section
5. An FAQ with four items
6. A footer

Constraints:
- Inline styles only (a single <style> block or style attributes)
- No external assets: no external fonts, scripts, stylesheets, or image URLs
- Emoji and inline SVG are allowed
- Output one single self-contained HTML file — nothing else
```

LEADERBOARD

Top 10 personas overall

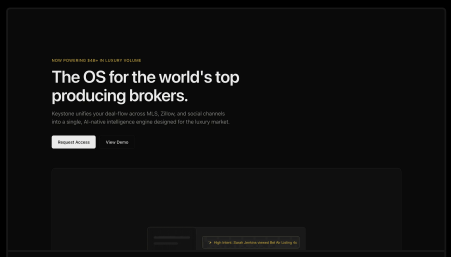
#	PERSONA	BUCKET	Σ	95% CI	MEAN
1	Stripe SVP of Design (expansive)	CLASSIC ROLE, EXPANSIVE	0.15	[8.40, 8.70]	**8.54**
2	Reference-pinned prompt	MASTERCLASS (COPY-READY)	0.41	[8.02, 8.77]	**8.30**
3	Figma principal designer (short)	CLASSIC ROLE, SHORT	0.18	[7.98, 8.33]	**8.14**
4	Vercel-style monochrome	DESIGN-CHEAT PERSONA	0.33	[7.82, 8.48]	**8.13**
5	The self-correcting loop	MASTERCLASS (COPY-READY)	0.05	[7.92, 8.02]	**7.97**
6	Brutalist web designer, 20 years in (expansive)	CLASSIC ROLE, EXPANSIVE	0.26	[7.78, 8.27]	**7.97**
7	Draft, critique, revise	META / STRUCTURAL	0.37	[7.45, 8.13]	**7.87**
8	Apple CPO (short)	CLASSIC ROLE, SHORT	0.23	[7.60, 8.05]	**7.83**
9	v0 by Vercel style prompt	PRODUCTION SYSTEM PROMPT	0.04	[7.80, 7.88]	**7.83**
10	Few-shot exemplar patterns	MASTERCLASS (COPY-READY)	0.35	[7.55, 8.22]	**7.82**

Exemplar screenshots · top performers



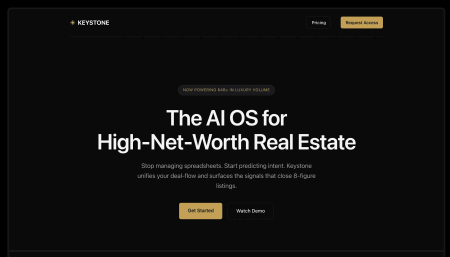
Stripe SVP of Design (expansive) **8.54**

CLASSIC ROLE, EXPANSIVE



Reference-pinned prompt **8.30**

MASTERCLASS (COPY-READY)



Figma principal designer (short) **8.14**

CLASSIC ROLE, SHORT

END OF PREVIEW

The full report has the rest

The full editorial deck is ~60 pages. It includes:

- The agent-authored editorial review of the full corpus, with exemplar comparisons and the unexpected-winners / noble-failures gallery
- One deep-dive page per bucket with the complete ranked persona table inside it
- The full 52-persona catalog with every system prompt in full, the rendered screenshot, and the per-axis score breakdown
- The complete pairwise Cohen's d table between all bucket pairs
- Prompt length vs composite scatter, within-persona variance ranking
- The raw data appendix for re-analysis

Available at rival.tips/research/persona-impact