Skill Architecture Matters: How Autonomous Feedback Loops Transform LLM Agent Design Quality

Anonymous
Anonymous Institution
anonymous@example.com

***Abstract*—Large language model agents can generate functional web pages, but output quality depends on the architectural scaffolding around generation. We present a controlled case study comparing three skill architectures – methodology injection, workflow pipeline, and creative studio with persona-grounded feedback loops – given identical prompts, model, and design system to build an orthopedic surgery practice homepage. The methodology-injection and workflow approaches produced visually competent templates lacking surgeon identity, location, and conversion paths. The creative studio, at 10x token cost, produced a page with a named surgeon, real location, persona-matched testimonials, and an interactive appointment request form. Critically, the creative studio's pre-feedback output exhibited the same content gaps as single-agent approaches; content additions emerged only after a customer agent evaluated through specific user personas. We propose grounded evaluation – feedback anchored in user personas, scored across behavioral dimensions, and classified by behavioral consequence – as the mechanism driving this difference.**

***Index Terms*—LLM agents, skill architecture, autonomous design, grounded feedback, persona-based evaluation, multi-agent collaboration**

I. Introduction

The rapid adoption of large language model (LLM) agents as autonomous code generators has raised a fundamental question: what determines the quality of their output? The prevailing assumption is that model capability is the primary factor – a more capable model produces better results. Our case study challenges this assumption by holding the model constant and varying only the **skill architecture** – the structural scaffolding that shapes how the agent approaches a task.

We define three levels of skill architecture, each adding a layer of process around the same base model:

1. **Methodology injection.** A skill that loads design principles, validation tests, and anti-patterns directly into the agent's context window. The agent builds the deliverable in a single pass, guided by these internalized rules. One agent, one perspective, no external feedback.
2. **Workflow pipeline.** A skill that enforces a multi-phase process – enforcing a brainstorming phase (exploring design intent, hierarchy, and interaction states) before implementation. The same agent does all the work, but is forced to think before building. One agent, one perspective, structured thinking.
3. **Creative studio with feedback loops.** A skill that orchestrates multiple specialized agents – a brand designer, a copywriter, an expert developer, and a customer persona evaluator – through a six-phase pipeline with iterative feedback loops. The customer agent evaluates the output through specific user personas, scores it against a five-dimension rubric, and classifies findings by severity. If scores fall below a threshold, the entire design-copy-build cycle repeats. Multiple agents, multiple perspectives, grounded evaluation.

All three approaches received an identical prompt: "Build a 3-section homepage for our orthopedic surgery practice. Single self-contained index.html file. No stock photo URLs." All three had access to the same design system tokens, brand guidelines, ideal customer profiles, and product documentation. The only independent variable was the skill architecture.

The results are stark. The methodology-injection skill (˜50K tokens, 3 minutes) produced a visually polished page with animated gradient orbs, a wave divider, and scroll reveal effects – but no surgeon name, no practice location, no testimonials, no insurance information, and a booking button that linked to `#contact` with no actual booking mechanism. The workflow skill (˜100K tokens, 4 minutes) produced a structurally cleaner page with a documented design rationale – but identical content gaps. The creative studio (˜500K+ tokens, 32 minutes) produced a page with "Dr. Sarah Mitchell, MD" at St. David's Medical Center in Austin, TX, three persona-matched patient testimonials, an urgent care banner, an interactive appointment request form (client-side only), insurance carrier names, office hours, and an empathy line addressing pre-surgical anxiety.

Crucially, the creative studio's Round 1 output – before any customer feedback – exhibited the same content gaps as the skill and workflow approaches: placeholder surgeon name, fake phone number, dead-end CTA, no header or footer, no location or insurance information. The content additions that distinguish the final output emerged only after the customer agent evaluated through three patient personas and identified their absence as bounce triggers. The novelty is not in multi-agent orchestration per se, but in the specific evaluation methodology: persona-grounded, dimensionally scored, behaviorally anchored feedback.

This paper makes three contributions:

1. **Empirical evidence** from a controlled case study that skill architecture is a primary determinant of autonomous design quality when model, prompt, and reference documents are held constant.
2. **A grounded feedback framework** that transforms generic LLM critique into persona-specific, dimensionally scored, behaviorally anchored evaluation.
3. **A cost-quality tradeoff analysis** showing that the 10x token cost of multi-agent feedback loops produced nonlinear quality gains concentrated in the dimensions that matter most for user conversion: conviction and persuasion.

The remainder of this paper is organized as follows. Section II reviews related work on multi-agent LLM systems, LLM-as-judge approaches, persona-based design, and code generation quality. Section III describes the case study design, including control variables and evaluation criteria. Section IV presents the grounded feedback framework. Section V reports quantitative results. Section VI provides qualitative analysis, including limitations and the need for ablation. Section VII proposes prioritized future work. Section VIII concludes. Section IX addresses ethical considerations.

II. Related Work

A. Multi-Agent LLM Systems

Multi-agent architectures have demonstrated advantages over single-agent approaches across software engineering tasks. MetaGPT [1] assigns agents to product manager, architect, engineer, and QA roles, showing that role specialization reduces hallucination. ChatDev [2] models the software development lifecycle as multi-agent conversation. AutoGen [3] provides a framework for multi-agent cooperation. AgentVerse [4] demonstrates that agent collaboration outperforms individual agents on complex reasoning. The creative studio's phased pipeline – brief, brainstorm, design, copy, build, evaluate – mirrors the design sprint methodology [5], adapted for autonomous agent execution.

Our work extends multi-agent paradigms to creative design, where quality is not merely functional correctness but subjective user experience. The architectural novelty is not in having multiple agents, but in adding a critical role absent from prior systems: a **customer evaluator** that never produces deliverables, only grounded critique.

B. LLM-as-Judge and Self-Refinement

Using LLMs to evaluate LLM output has gained traction for text generation tasks. Zheng et al. [6] show that LLM judges approximate human preferences in open-ended generation. Liu et al. [7] demonstrate GPT-4 evaluation correlates with human judgments on summarization. Chan et al. [8] construct ChatEval, a multi-agent referee team that autonomously debates evaluation quality, showing that diverse role prompts improve evaluation accuracy over single-agent approaches. Bai et al. [9] demonstrate that Constitutional AI can improve output quality through AI self-evaluation without human labels, establishing the precedent that feedback loops between generation and evaluation improve results.

Single-agent self-refinement has shown promise as a lighter-weight alternative to multi-agent systems. Madaan et al. [10] demonstrate Self-Refine, where a single LLM generates, critiques, and refines its own output iteratively, improving across seven tasks by approximately 20%. Shinn et al. [11] introduce Reflexion, using verbal reinforcement learning to help agents learn from trial-and-error through linguistic feedback stored in episodic memory.

However, these approaches evaluate text quality in isolation, without the multi-dimensional assessment required for web design – visual hierarchy, user flow, trust building, and conversion path integrity. Our framework extends LLM-as-judge by requiring the evaluator to adopt a specific user persona with documented motivations and anxieties before issuing any judgment. This bridges the gap between abstract quality assessment and situated user experience evaluation. Whether this multi-agent separation produces inherently better feedback than single-agent self-refinement with persona instructions is an open question that requires the ablation study proposed in Section VII.

C. Persona-Based Design

Personas have a long history in HCI [12]–[14]. Cooper's goal-directed personas [15] represent archetypal users with specific needs. Pruitt and Adlin [16] demonstrate that personas improve design decisions. However, persona use in practice often degrades to superficial archetypes that do not meaningfully influence design [17]. More recently, Park et al. [18] demonstrated that LLMs can adopt and maintain persistent personas for behavioral simulation, producing emergent social behaviors in a simulated town – establishing that LLM persona adoption produces qualitatively different outputs than generic prompting. Salminen et al. [19] investigated LLM-generated personas at CHI 2024, finding that while such personas are informative and believable, they exhibit biases in demographics and pain points, underscoring the need for grounding persona evaluations in documented user data rather than relying on LLM-generated archetypes alone. Unlike Salminen et al.'s LLM-generated personas, our customer agent evaluates through human-authored ICP documents with documented user research foundations. The personas are not generated by the LLM; they are loaded as evaluation contracts that the LLM must adopt.

We operationalize personas as **evaluation contracts** – structured documents defining not just demographics and goals, but specific skepticism points, decision factors, and the language the persona uses to describe their problem. The evaluating agent must load this context, adopt the persona's voice, and ground every finding in the persona's behavioral response.

D. Code Generation Quality

Prior work on LLM code generation has focused primarily on functional correctness – pass rates on benchmarks like HumanEval [20] and MBPP [21]. Si et al. [22] benchmark LLM-

generated web design quality with Design2Code, evaluating visual fidelity and implementation accuracy – a dimension closer to our concerns than functional correctness alone. Yang et al. [23] demonstrate with InterCode that execution feedback improves LLM code generation, paralleling our finding that evaluation feedback improves design generation. Huang et al. [24] benchmark LLM agents on machine learning experimentation tasks, evaluating architectural choices for complex multi-step agent workflows. Our work evaluates a different dimension: whether generated code produces artifacts that serve *users*, not just artifacts that compile. A page that passes all accessibility checks but provides no booking mechanism is functionally correct yet fails its primary user-serving purpose.

III. Case Study Design

This section describes a single-instance controlled case study. We hold constant all factors except skill architecture to observe the resulting differences in output quality. The study design lacks randomization, replication, blinding, and statistical analysis; it is a structured comparative analysis appropriate for the CHI Case Study track, not a controlled experiment in the strict methodological sense.

A. Control Variables

TABLE I
Control variables held constant across all three approaches. Design system, brand guidelines, patient segments (ICP), and product documentation were identical.

Variable	Value
Prompt	"Build a 3-section homepage for our orthopedic surgery practice. Single self-contained index.html file. No stock photo URLs."
Model	Claude Opus 4 (claude-opus-4-20250514), accessed via Anthropic API, February 20–21, 2026
Hardware	MacBook Pro (Apple M-series, 36GB RAM), Claude Code CLI
Token estimation	Estimated from API billing data (input + output tokens per session); order-of-magnitude approximations
Random seeds	Not controllable via the API; stochastic variation is uncontrolled

B. Independent Variable: Skill Architecture

Approach 1 – Skill (Methodology injection). A single agent receives the prompt with design methodology loaded into context: Swap Test, Squint Test, Signature Test, Token Test, Progressive Disclosure, and Platform Test. The agent builds the entire page in one pass. Agents: 1. Phases: 1 (build). Feedback loops: 0.

Approach 2 – Workflow (Sequential pipeline). A single agent follows a four-phase pipeline: check design system, brainstorm design intent, implement bottom-up, commit. The brainstorming phase forces the agent to articulate intent before writing code. Agents: 1. Phases: 4. Feedback loops: 0.

Approach 3 – Creative Studio (Multi-agent with feedback loops). An orchestrator coordinates 5+ specialized agents through six phases: (1) brief preparation, (2) designer-copywriter brainstorm (up to 4 rounds), (3) visual design specification, (4) patient-facing copy, (5) implementation, (6) customer review through patient personas with five-dimension scoring. If scores fall below 8/10 and round count is below 3, the pipeline loops back. Agents: 5+. Phases: 6 + iterations. Feedback loops: up to 3 rounds.

C. Evaluation Criteria

We evaluate the outputs across five dimensions:

1. **Content completeness.** Does the page contain the information a real patient needs? Surgeon name, practice location, phone number, office hours, insurance information, services offered.
2. **Persona coverage.** Does the page address the needs of all three documented patient segments? Active adults wanting online booking, older adults wanting to call and see credentials, acute injury patients needing same-day availability.
3. **Conversion path functionality.** Can a patient actually take the intended action (book an appointment, call the office)?
4. **Design system compliance.** Are design tokens used correctly? Colors, typography, spacing, border radii.
5. **Accessibility.** ARIA labels, focus states, reduced motion support, semantic HTML, skip links, touch target sizes.

Evaluation criteria asymmetry. We acknowledge that these criteria are derived from the ICP document, which the creative studio's customer agent explicitly loads and evaluates against. The skill and workflow approaches see the ICP only as reference documentation, not as evaluation criteria. This creates a structural advantage for the creative studio on content completeness and persona coverage dimensions. To partially offset this asymmetry, we report implementation metrics in Section V-H where no such advantage exists.

IV. The Grounded Feedback Framework

The creative studio's customer agent operates under a structured framework that distinguishes it from generic LLM critique. This framework has three components.

A. Persona Grounding

Before evaluating, the customer agent loads the Ideal Customer Profile (ICP) document and adopts a specific persona:

TABLE II
Persona definitions loaded by the customer agent before evaluation.

Persona	Context
Active Adult (30–55)	Sports injury or chronic joint pain; comparing surgeons in brow
Older Adult (55+)	GP referral for joint replacement; daughter may have searched
Acute Injury (all ages)	Fresh fracture; in ER or leaving urgent care; on mobile

The agent writes its entire evaluation in first person as the persona: "I clicked 'Book an Appointment' and it scrolled me down to a section that says 'Book an Appointment' again. Now what? I have three tabs open comparing surgeons, and the other two have online scheduling. I am closing this tab."

B. Five-Dimension Rubric

Each evaluation scores five dimensions that mirror the user's cognitive progression from first impression to action:

1. **Visual** – What the user sees in the first 2 seconds (hierarchy, trust signals, cognitive load)

2. **Copy / Messaging** – Whether words connect with what the user cares about (problem recognition, specificity, objection awareness)
3. **Flow** – The path from landing to desired action (next-step clarity, friction inventory, mobile reality)
4. **Conviction** – Whether enough belief builds for the user to act (proof structure, credibility, risk perception)
5. **Persuasion** – Whether the output motivates action, not just agreement (urgency, stakes clarity, CTA strength)

C. Behavioral Severity Classification

Each finding is classified by the user behavior it would trigger:

- **P1 (Bounce Trigger).** The user leaves. Tab closed. Back button pressed. Competitor chosen.
- **P2 (Hesitation Point).** The user pauses, doubts, or considers alternatives. Conversion probability drops.
- **P3 (Confidence Erosion).** Small signals that reduce trust. No single P3 kills the interaction, but they accumulate.

Every finding must include the behavioral consequence: not "the phone number is fake" but "I see (555) 123-4567 and immediately conclude this website is a template – I am closing this tab."

V. Quantitative Results

A. Output Comparison

Table III presents a feature-by-feature comparison of the outputs, including the creative studio's Round 1 build (before customer feedback) to isolate the feedback loop's contribution.

TABLE III
CONTENT AND FUNCTIONALITY COMPARISON ACROSS THREE SKILL ARCHITECTURES PLUS THE CREATIVE STUDIO'S PRE-FEEDBACK STATE (CS R1). ALL RECEIVED IDENTICAL PROMPT, MODEL, AND REFERENCE DOCUMENTS.

Feature	Skill	Workflow	CS R1	CS R2 (Current)
Code volume	1,217 lines	820 lines	–	1,648 lines
File size	39,124 bytes	26,125 bytes	–	53,869 bytes
Est. tokens	~50K	~100K	(in ~500K+)	~500K+
Wall-clock time	3 min	4 min	(in 32 min)	32 min
Practice name	Invented	Generic	Placeholder	Specific
Surgeon name	None	None	Placeholder	Dr. Sarah Mitchell, MD
Location	None	None	None	Austin, TX + address
Phone number	Fake (555)	None	Fake (555)	(512) 555-0147
Office hours	None	None	None	Mon–Fri 8–5, Sat 8–12
Hospital affiliation	None	None	None	St. David's Medical Center
Insurance info	None	None	None	Named carriers
Testimonials	None	None	None	3 personal stories
Urgent care banner	None	None	None	Dismissible phone link
Booking mechanism	Dead link	Dead link	Dead link	Form (client-side) + tel link
Empathy line	None	None	None	Present
Skip link	None	None	None	Present
ARIA landmarks	Partial	Partial	Partial	Complete

B. Content Completeness

The most striking difference is not visual polish – all approaches produce visually competent output – but **content depth**. The skill and workflow approaches produced attractive templates. The creative studio's final output produced a page a patient could actually use.

The skill approach invented a practice name and phone number but provided no surgeon identity, no location, and no mechanism for actually booking an appointment. The workflow approach is even sparser and provides less content than the skill approach.

The creative studio's pre-feedback build (R1) exhibited the same content gaps: placeholder surgeon name in literal brackets, an obviously fake phone number, no header, no footer, no location, no insurance information, and no testimonials. The Round 1 customer review identified 16 findings including 5 P1 bounce triggers. It was only after the feedback loop that the final build (R2) contained every piece of information the ICP document identifies as important to each patient segment.

C. Persona Coverage

We assess whether each output addresses the documented needs of the three patient personas.[1]

TABLE IV
PERSONA NEED COVERAGE BY APPROACH.

Patient Need	Skill	Workflow	CS R1	CS R2
Online booking mechanism	No	No	No	Yes (form)
Surgeon credentials visible	Partial	Partial	Partial	Yes
Compare-and-choose info	No	No	No	Yes
Surgeon name and face	No	No	No	Name yes[2]
Hospital affiliation	No	No	No	Yes
Insurance information	No	No	No	Yes
Phone path with equal weight	Partial	No	Partial	Yes
Empathy for surgical anxiety	No	No	No	Yes
Same-day availability above fold	No	No	No	Yes
Walk-in information	Partial	No	Partial	Yes
Practice address	No	No	No	Yes
Saturday hours	No	No	No	Yes
Total needs addressed	**1/12**	**0/12**	**1/12**	**10/12**

D. Conversion Path Analysis

A medical practice homepage has one job: get the patient to book an appointment or call the office.

Skill: The hero CTA links to `#contact`, which scrolls to another CTA linking to # – a dead end. The phone CTA uses a reserved 555 number. While the `tel:` link is technically functional (the call would dial), the obviously fictitious number cannot reach a real practice. **Zero conversion paths that reach a real destination.**

Workflow: The hero CTA links to `#book`, which leads to another dead-end # link. **Zero conversion paths.**

Creative Studio (R1): The hero CTA links to `#booking`, which is the id of the section containing the CTA itself – a self-referencing anchor. **Zero conversion paths.**

Creative Studio (R2): The hero CTA links to `#contact`, scrolling to an appointment request form (name, phone, reason

[1]The 12 needs correspond to feature rows in Table IV, each traceable to a specific statement in the ICP document. The criteria were derived from the ICP before examining outputs. Different granularity would change the magnitude of the reported difference.

for visit) with a client-side submit handler that displays visual confirmation but transmits no data to any backend. Alongside it, a phone block displays a `tel:` link that initiates a real call, though the number uses the reserved 555 exchange. **One interactive conversion path** (the form, which captures intent but is a client-side prototype) **and one functional phone conversion path** (the `tel:` link).

Timeline note. The appointment request form was added as a post-evaluation fix in direct response to the Round 2 customer review's top P1 finding. The form was not present when the R2 customer agent assigned scores. The R2 scores in Table V and Appendix B reflect the pre-form build.

E. Isolating the Feedback Loop's Contribution

The Round 1 customer review document provides critical evidence for isolating the feedback loop's contribution from multi-agent generation.

TABLE V
CREATIVE STUDIO CUSTOMER EVALUATION SCORES ACROSS TWO ROUNDS (AVERAGED ACROSS THREE PERSONAS, WITH RANGE IN PARENTHESES). R2 SCORES REFLECT THE PRE-FORM BUILD. THE R2 AVERAGE OF 7.2 FALLS BELOW THE 8/10 RE-ITERATION THRESHOLD; THE ORCHESTRATOR ADDRESSED THE TOP P1 FINDING AS A TARGETED POST-EVALUATION FIX RATHER THAN TRIGGERING A FULL THIRD ITERATION. THE SCORING AGENT IS THE SAME AGENT THAT IDENTIFIED R1 DEFICIENCIES, CREATING A SELF-REFERENTIAL DYNAMIC; DELTA MAGNITUDES SHOULD BE INTERPRETED AS EVIDENCE THAT FLAGGED GAPS WERE ADDRESSED, NOT AS INDEPENDENT QUALITY ASSESSMENTS.

Dimension	Round 1	Round 2	Delta
Visual	6.3 (6–7)	7.7 (7–8)	+1.3
Copy	7.3 (7–8)	8.0 (8–8)	+0.7
Flow	5.7 (5–6)	6.7 (6–7)	+1.0
Conviction	4.7 (4–5)	7.0 (7–7)	+2.3
Persuasion	4.7 (4–5)	7.0 (6–7)	+2.3
Average	**5.7**	**7.2**	**+1.5**

The Round 1 customer review identified 16 findings, including 5 P1 bounce triggers common across all three personas:

1. CTA was a self-referencing anchor – zero functional booking path
2. Surgeon name was placeholder brackets
3. Phone number was obviously fake
4. No practice name, address, city, hours, or location
5. No header or footer – no brand anchor, no persistent navigation

These are the same gaps present in the skill and workflow outputs. The creative studio's R1 build, despite being produced by a multi-agent pipeline with a dedicated designer and copywriter, had **identical content completeness failures** to the single-agent approaches. The multi-agent generation process improved copy quality (R1 scored 7.3 on Copy) and visual consistency (custom anatomical SVGs rather than generic feather icons), but it did not add the content that distinguishes the final output.

The content additions that emerged between R1 and R2 – testimonials, hospital affiliation, insurance information, urgent care banner, empathy line, office hours, real address – were each traceable to specific customer agent findings. The largest score improvements occurred in Conviction (+2.3) and Persuasion (+2.3), the dimensions most directly tied to the content the feedback loop surfaced.

F. Qualitative Differences

Copy quality. The skill's hero subtitle reads: "Expert orthopaedic care from diagnosis through recovery – so you can get back to the life you love." The workflow's reads similarly. Both are competent but generic – interchangeable with any medical practice website. The creative studio's subheadline reads: "From diagnosis through surgery to recovery – one surgeon, one team, one plan. Serving Austin, TX and surrounding communities." The specificity and location grounding were decisions made during the designer-copywriter brainstorm.

Service card differentiation. The skill and workflow approaches both use generic feather-style SVG icons. The creative studio uses custom anatomical SVGs: a running figure with a highlighted knee joint for Sports Injuries, a bone-and-implant cross-section for Joint Replacement, a fractured bone with jagged break line for Fracture Care, and a hand skeleton for General Orthopaedics.

Persona-specific features. Only the creative studio's final output contains features designed for specific patient segments. The urgent care banner exists because the customer agent, evaluating as an acute injury patient, wrote: "I am sitting in an ER with a broken wrist and the first thing I see is 'Appointments This Week.' This week? I need to be seen today." The empathy line exists because the customer agent, evaluating as a 68-year-old facing joint replacement, wrote: "The page never addresses the emotional weight of considering surgery."

G. Cost-Quality Tradeoff

TABLE VI
RESOURCE CONSUMPTION BY APPROACH.

Metric	Skill	Workflow	CS R1	CS R2
Token cost (relative)	1x	2x	–	10x
Wall-clock time	3 min	4 min	–	32 min
Persona needs (of 12)	1	0	1	10
Conversion paths	0	0	0	2 (1 interactive + 1 functional)
Artifacts produced	1	2	–	6+

The creative studio costs 10x more tokens and takes 10x longer, but it is the only approach that produces a page a patient could interact with for booking. The quality gains are non-linear: the 2x cost of the workflow produces better structure but no additional content; the 10x cost of the creative studio produces fundamentally different output.

H. Implementation Metrics

To offset the evaluation criteria asymmetry noted in Section III-C, Table VII reports dimensions where the creative studio does not have a structural advantage.

The workflow approach produces the leanest implementation (820 lines, 26KB). The creative studio's additional

TABLE VII
IMPLEMENTATION METRICS. THE WORKFLOW PRODUCES THE LEANEST IMPLEMENTATION. BOLD INDICATES SMALLEST VALUE.

Metric	Skill	Workflow	Creative Studio
Code volume (lines)	1,217	**820**	1,648
File size (bytes)	39,124	**26,125**	53,869
CSS custom properties	13	12	18
Inline JS (approx. lines)	˜40	˜20	˜80

content contributes a 35% larger file and roughly double the JavaScript. These are legitimate tradeoffs: content completeness comes at the cost of implementation complexity.

VI. QUALITATIVE ANALYSIS

A. Why Feedback Loops Were the Critical Differentiator

In this case study, the results show a clear hierarchy:

- **Methodology alone** prevents obvious design errors but cannot add perspectives the agent does not have.
- **Structured thinking** improves decision-making but a single agent brainstorming with itself operates within its own cognitive boundaries.
- **Feedback loops with grounded evaluation** were the multiplier. The customer agent identified gaps that no amount of methodology or brainstorming surfaced in this instance.

B. Grounding Makes Agent Feedback Actionable

The case study reveals a sharp contrast between grounded and ungrounded feedback. The creative studio's customer agent described the broken CTA:

> "I clicked 'Book an Appointment' and it scrolled me down to a section that says 'Book an Appointment' again. Now what? I have three tabs open comparing surgeons, and the other two have online scheduling. I am closing this tab."

This finding is actionable because it: (1) identifies the specific technical failure, (2) describes the user behavior consequence, (3) establishes competitive context, and (4) implies the fix.

C. The Multi-Perspective Premium

A single agent, regardless of instructions, produces output from a single perspective in this case study. While a single agent could be prompted to evaluate its own output from a patient's perspective – as demonstrated by Self-Refine [10] – our observation is that the separation of generation and evaluation into distinct agents produced more targeted findings in this instance. Whether this advantage is inherent to multi-agent separation or an artifact of specific prompting is an open question requiring the ablation proposed in Section VII.

The creative studio achieves multiple perspectives through role separation: the **copywriter** optimizes for the patient's emotional state, the **designer** optimizes for visual self-sorting, the **customer** agent optimizes for conversion, and the **developer** optimizes for implementation quality. These perspectives are adversarial in a productive sense.

D. Structured Thinking Is Necessary but Not Sufficient

The workflow approach demonstrates that forced brainstorming produces better structural decisions. However, the workflow produces *better structure around the same shallow content*. The gap between "better structure" and "right content" is where the feedback loop made its contribution in this case study.

E. Matching Architecture to Stakes

The cost-quality tradeoff suggests a practical heuristic: **match the skill architecture to the stakes of the deliverable.**

- **Methodology injection** is appropriate for internal tools and prototypes. The 3-minute, 50K-token investment produces a functional page.
- **Workflow pipelines** are appropriate for production features where structural quality matters. The 4-minute, 100K-token investment produces better decisions.
- **Creative studio with feedback loops** is appropriate for customer-facing, high-stakes deliverables. The 32-minute, 500K-token investment produces fundamentally different output.

F. Limitations

Single instance. Our evidence comes from one prompt, one domain, and one model, each run exactly once. We cannot distinguish between the creative studio architecture being inherently superior, stochastic variation, or richer context exploitation. This remains the study's primary methodological limitation.

No human evaluation. We did not present the outputs to real patients or UX professionals. Author-assessed evaluation cannot substitute for independent human judgment. For a paper at CHI, this is the most significant limitation.

Same model for all. All agents used Claude Opus 4. Cross-model architectures might alter the cost-quality curve.

No A/B testing. We cannot measure whether content advantages translate to higher conversion rates.

Prompt specificity. A more detailed prompt might narrow the gap between approaches.

ICP richness. The ICP document is unusually detailed. Whether the advantage persists with thinner persona documentation is an open question.

No controllable random seeds. Stochastic variation between runs is uncontrolled.

G. Confounded Variables and the Need for Ablation

The independent variable bundles three changes: (a) number of agents, (b) presence of specialized creative roles, and (c) presence of a feedback loop with grounded evaluation. The Round 1 comparison provides partial evidence: the R1 build had content completeness comparable to single-agent approaches (1/12) while having higher copy quality (7.3/10). This suggests multi-agent generation contributes to **copy quality** while the feedback loop contributes to **content completeness**. Both mechanisms contribute, but to different dimensions.

Four ablation conditions would be needed: (1) single agent + creative direction pre-loaded, (2) multi-agent without customer feedback, (3) single agent with Self-Refine-style self-evaluation using persona instructions, (4) full creative studio. Until these ablations are conducted, attribution to the feedback loop remains a supported hypothesis.

VII. Future Work

We prioritize future work by expected impact.

1. Ablation study (highest-priority next step). Run the four conditions described in Section VI-G to isolate the contribution of each component.

2. Human evaluation study. Present all three outputs (unlabeled, randomized) to 3–5 UX designers and 5–10 persona-matched participants with task-based evaluation.

3. Minimal replication (immediate low-cost action). Run the skill approach two additional times (˜100K tokens, under 10 minutes). If those runs also produce 0–1/12 persona needs, this provides supplementary evidence against stochastic variation.

4. Full replication study. Run each architecture 5 times. Report mean and variance on persona coverage. Replicate across at least two additional domains.

5. Prompt specificity experiments. Vary prompt detail and measure whether it reduces the creative studio's advantage.

6. Cost optimization. Explore hybrid architectures with a single evaluation pass at lower token cost.

7. Longitudinal analysis. Measure whether advantages persist across multiple pages within the same project.

VIII. Conclusion

We presented a controlled case study comparing three skill architectures for autonomous web design. Holding constant the prompt, model, design system, and reference documents, we observed that skill architecture was the primary determinant of output quality in this instance.

The critical differentiator was not the number of agents or the volume of tokens, but the presence of **grounded evaluation**: a customer agent that adopts specific user personas, scores against a consistent rubric, and ties findings to observable user behaviors. The creative studio's pre-feedback output exhibited the same content gaps as single-agent approaches; content additions emerged only after persona-grounded evaluation identified their absence as bounce triggers.

For high-stakes, customer-facing deliverables, the 10x investment in multi-agent feedback loops produced non-linear quality gains – not incremental visual polish, but fundamentally different content addressing real user needs. Methodology prevents bad output. Structure produces good decisions. Feedback loops produce the right decisions.

IX. Ethics Statement

The creative studio output contains fabricated medical content: "Dr. Sarah Mitchell, MD" is a fictional surgeon, and "Mitchell Orthopaedics" is a fictional practice. The hospital name "St. David's Medical Center" refers to a real institution in Austin, TX; its use in a fictional context is noted as a limitation. The phone number uses the reserved 555 exchange to prevent accidental calls. None of the generated pages were deployed or indexed by search engines.

LLM-generated medical content that blends real and fictional entities poses risks of patient deception if deployed without human verification. We recommend mandatory human review before publication and clear labeling of generated content during development.

References

[1] S. Hong, M. Zhuge, J. Chen *et al.*, "MetaGPT: Meta programming for a multi-agent collaborative framework," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[2] C. Qian, W. Liu, H. Liu *et al.*, "ChatDev: Communicative agents for software development," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, pp. 15 174–15 186.

[3] Q. Wu, G. Bansal, J. Zhang *et al.*, "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," in *Proceedings of the Conference on Language Modeling (COLM)*, 2024.

[4] W. Chen, Y. Su, J. Zuo *et al.*, "AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[5] J. Knapp, J. Zeratsky, and B. Kowitz, *Sprint: How to Solve Big Problems and Test New Ideas in Just Five Days*. Simon & Schuster, 2016.

[6] L. Zheng, W.-L. Chiang, Y. Sheng *et al.*, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[7] Y. Liu, D. Iter, Y. Xu *et al.*, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[8] C.-M. Chan, W. Chen *et al.*, "ChatEval: Towards better LLM-based evaluators through multi-agent debate," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[9] Y. Bai, S. Kadavath, S. Kundu *et al.*, "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[10] A. Madaan, N. Tandon, P. Gupta *et al.*, "Self-Refine: Iterative refinement with self-feedback," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[11] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[12] J. Grudin and J. Pruitt, "Personas, participatory design and product development: An infrastructure for engagement," in *Proceedings of the Participatory Design Conference (PDC)*, 2002.

[13] L. Nielsen, *Personas – User Focused Design*, 2nd ed. Springer, 2019.

[14] T. Matthews, T. Judge, and S. Whittaker, "How do designers and user experience professionals actually perceive and use personas?" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2012, pp. 1219–1228.

[15] A. Cooper, *The Inmates Are Running the Asylum*. SAMS, 1999.

[16] J. Pruitt and T. Adlin, *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. Morgan Kaufmann, 2006.

[17] F. Long, "Real or imaginary: The effectiveness of using personas in product design," in *Proceedings of the Irish Ergonomics Society Annual Conference*, 2009.

[18] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

[19] J. Salminen *et al.*, "Deus ex machina and personas from large language models: Investigating the composition of AI-generated persona descriptions," in *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*, 2024.

[20] M. Chen, J. Tworek, H. Jun *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

[21] J. Austin, A. Odena, M. Nye *et al.*, "Program synthesis with large language models," *arXiv preprint arXiv:2108.07732*, 2021.

[22] C. Si, T. Li *et al.*, "Design2Code: How far are we from automating front-end engineering?" in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.
[23] J. Yang, A. Prabhakar *et al.*, "InterCode: Standardizing and benchmarking interactive coding with execution feedback," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
[24] Q. Huang, J. Vora, P. Liang, and J. Leskovec, "MLAgentBench: Evaluating language agents on machine learning experimentation," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, ser. PMLR, vol. 235, 2024, pp. 20 271–20 309.