Vibe Coding Suffers When AI's Role Expands

Originally published at https://www.unite.ai/vibe-coding-suffers-when-ais-role-expands/

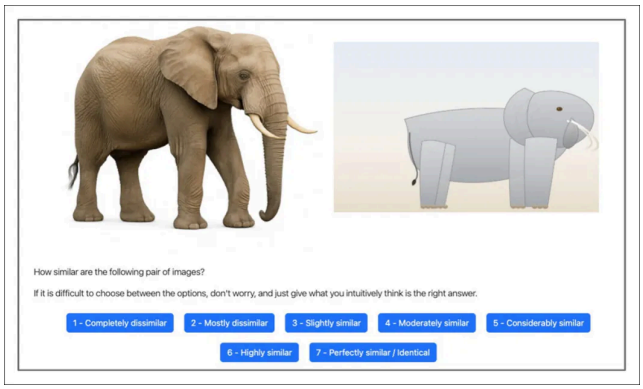
A new study finds vibe coding improves when humans give the instructions, but declines when AI does, with the best hybrid setup keeping humans foremost, with AI as an arbiter or judge.

New research from the United States, examining what happens when AI systems are allowed to steer vibe coding, rather than simply execute human instructions, has found that when Large Language Models (LLMs) take on a larger directional role, the outcomes are almost always worse.

Though the researchers used OpenAI's GPT-5 as the framework for their human/AI collaborative experiments, they later confirmed that both Anthropic's Claude Opus 4.5 and Google Gemini 3 Pro were subject to the same deteriorating curve as responsibilities grew, stating that 'even limited human involvement steadily improves performance':

> *'[Humans] provide uniquely effective high-level guidance across iterations, [whereas] AI guidance often leads to performance collapse. Also, we find that a careful role allocation which keeps humans in charge of direction while offloading evaluation to AI can improve hybrid performance.'*

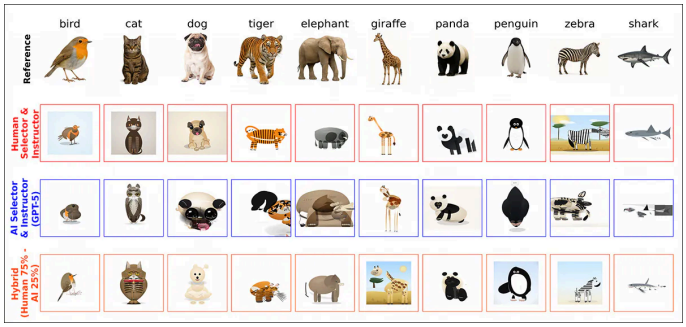
In order to provide a consistent test which could be evaluated equally by humans as by AI, a controlled experimental framework was built around an iterative coding task in which a reference image – featuring a photo of a cat, dog, tiger, bird, elephant, penguin, shark, zebra, giraffe, or panda – had to be recreated using scalable vector graphics (SVG), and that recreation judged against the photo source from which it was derived:



Both human and AI participants were shown a photographic reference image alongside an AI-generated SVG reconstruction, and asked to rate how similar the two were on a seven-point scale. Source

In each round, one agent provided high-level natural language instructions to guide a code generator, and another decided whether to keep the new version or revert to the previous one – a structured loop that mirrors real collaborative workflows.

Across 16 experiments involving 604 participants and thousands of API calls, fully human-led testing rounds were compared directly with fully AI-led rounds, under otherwise identical conditions.



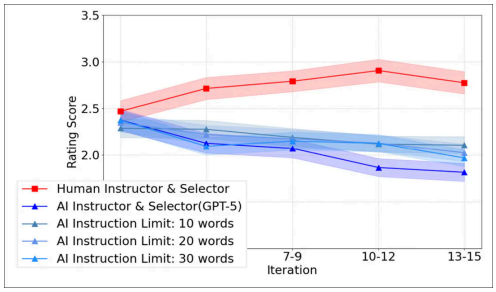
Some of the varied solutions arrived at by different combinations of human/AI collaboration percentages and types (taken from a larger illustration in the source paper, to which we refer the reader).

Though humans and AI performed at similar levels at the baseline start of the tests, over time, their trajectories diverged: when humans provided the instructions and made the selection decisions, similarity scores increased across iterations, with steady cumulative improvement; but when AI systems filled both roles, performance showed no consistent gains, and frequently declined over rounds – even though the same underlying model was used for code generation, and the AI had access to the same information as human participants.

The Prolixity Effect

The results also showed that Human instructions were typically short and action-oriented, focusing on what to change next in the current image; conversely, AI instructions were much longer and heavily descriptive (a factor which was parametrized for GPT-5), detailing visual attributes rather than prioritizing incremental correction.

But, as seen in the graph below, imposing strict word limits on AI instructions did not reverse the pattern; even when constrained to 10, 20 or 30 words, AI-led chains still failed to improve over time:



Similarity ratings across iterations for human-led rounds compared with fully AI-led rounds constrained to 10, 20, or 30-word instructions. Evidently, shortening AI prompts does not prevent the iterative performance decline observed when AI directs both instruction and selection.

Hybrid experiments made the pattern clearer, showing that adding even a little human involvement improved results, compared with fully AI-led setups; yet performance usually declined as the share of AI guidance increased.

When the roles were separated, evaluation and selection could be handed to AI with relatively little loss in quality; but replacing human high-level instruction with AI guidance led to noticeable drops in performance, suggesting that what mattered most was not who *generated the code*, but who *set and sustained the direction* across iterations.

The authors conclude:

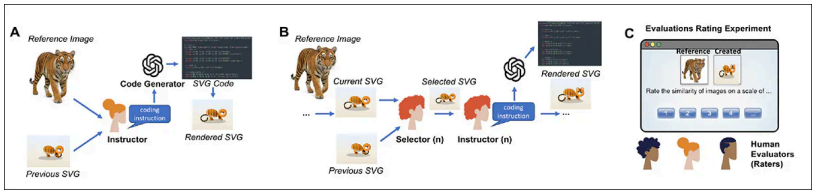
> *'Across multiple experiments, human-led coding consistently improved over iterations, while AI-led coding often collapsed despite access to the same information and similar execution capabilities.*
>
> *'This points to key struggles of today's AI systems in sustaining coherent high-level direction across repeated interactions, of the kind necessary for successful vibe coding'*

The new paper is titled *Why Human Guidance Matters in Collaborative Vibe Coding*, and comes from seven researchers across Cornell University, Princeton University, Massachusetts Institute of Technology, and New York University.

Method

For the experiments, a human instructor looked at a GPT-5-generated animal reference photo, together with the latest associated SVG imitation attempt. It then wrote natural language instructions to guide the code generator toward a closer match.

Thus, the generator would produce a new SVG each round, providing an iterative loop for testing how the effect of guidance accumulates over time. The targets were ten GPT-5 generated animal images, covering a range of shapes and textures so that improvements or mistakes would be easy to detect:



Schema for the vibe coding workflow used in the study. In A), a human instructor views a photographic reference image together with the best SVG produced so fa instructions for the code generator to follow when producing the next SVG; in B), a human selector compares the new SVG with the previous one and chooses wh reference image, before passing the selected SVG forward for the next round of instruction; and in C), independent human evaluators rate how similar each gener image, supplying the scores used to assess overall performance.

A human selector compared each newly generated SVG with the previous one and either accepted or rejected it, which kept the process aligned with the reference image across rounds. In this baseline setup, the same human carried out both roles.

To measure *quality*, independent human evaluators rated how similar each generated SVG was to its reference image. Across sixteen experiments, 120 people produced 4,800 ratings. All experiments were run on the PsyNet framework, a portal designed to accommodate structured interactions between humans and AI systems.

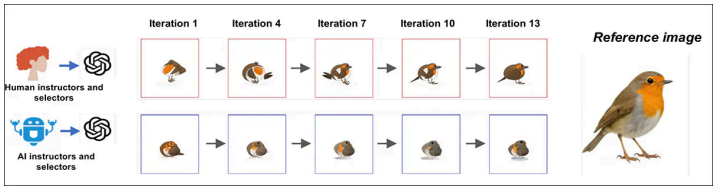
The study would recruit 604 native English speakers, in tests that would burn through 4,800 API calls for code generation, and 5,327 API calls for instruction. Though GPT-5 was the main model used, smaller comparison batches were made with Claude Opus 4.5 and Gemini 3 Pro, which each handled 280 queries.

Results

Thirty vibe-coding rounds were run, each comprising fifteen edits of the core ten reference images. For these, 45 human participants were chosen, each serving as both selector and instructor across ten iterations, in the ‘human-led’ rounds.

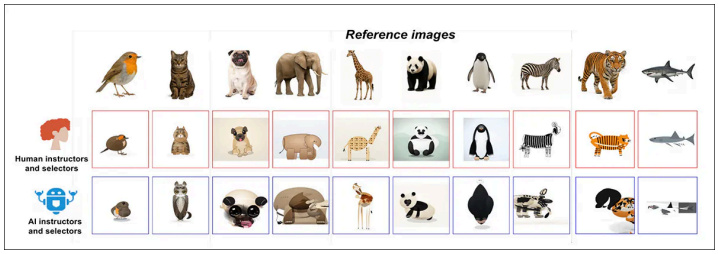
Within each turn, the same participant first chose between the current and previous SVG, and then wrote the next round of instructions. A second version of the test replaced those human decisions with API calls to GPT 5, while keeping the rest of the setup unchanged. In all cases the instructor and selector roles prompted the code generator with plain language.

A representative example of multi-round vibe coding shows how the process diverges over time; when humans acted as both selector and instructor, the SVG output improved steadily across iterations, moving closer to the reference image with each round:



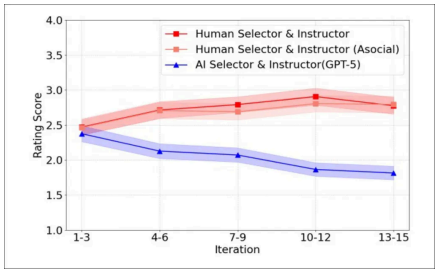
Example progressions for one reference image under human led (top) and AI led (bottom) vibe coding, showing steady improvement across iterations with human roles, and stagnation or drift when both roles are handled by AI.

Conversely, in the AI-led version, early rounds sometimes captured key visual features, but later attempts failed to build on those gains, and in some cases drifted away from the target:



Final outputs from the final iteration, comparing human-led turns (top row) with AI led chains (bottom row), across the same set of reference images. The human-le more closely match the original animals, while the AI-led results demonstrate visible distortions, or loss of key features.

To measure the emerging trends quantitatively, the final images were shown to independent human raters and scored for similarity to the reference pictures. In the early rounds, human-led and AI-led runs scored about the same; but by the fifteenth round, the difference was clear, with the human-chosen images rated much closer to the targets. Over time, the human scores rose steadily, with the biggest relative gain over AI reaching 27.1%.



Average similarity scores across iterations for human-led and AI-led vibe coding, showing steady gains when humans act as both selector and instructor, and a gradual decline when both roles are handled by GPT 5.

To ensure that the emerging trends were not due to the collective power of multiple simultaneous human participants, the researchers recruited ten additional people to work alone, each running three rounds by themselves – and the results improved in the same steady way, demonstrating that the gains were not a fluke of collective effort.

The Big Picture

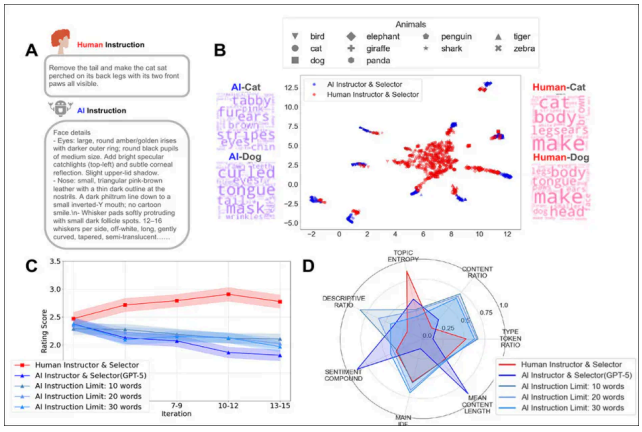
However, if GPT-5 judged the outputs itself, would it admit that the human results were better? Human and AI ratings generally moved in the same direction, so the model could tell good from bad, yet consistently scored AI-generated images *higher than humans did*.

> *'Specifically, we asked whether AI agents would recognize that their own outputs are inferior to those produced by humans, or instead show a preference for their own creations, which would indicate a potential alignment issue.'*

As it transpired, there is indeed an alignment issue*:

> *'AI evaluators assigned higher ratings to AI-generated [outputs]. These findings suggest that the observed performance differences may stem from a misalignment in representations between humans and AI.'*

In examining how humans and AI each phrase their guidance, discrepancies became clear in tests. As seen in the figure below, both focus *and* length are subjects of AI/Human divergence:



A comparison of how humans and AI gave instructions during the coding task. 'A' shows that humans write short, direct instructions, while AI writes long, detailed descriptions. 'B' maps the instructions, revealing that human prompts cluster together, while AI prompts split apart by animal. 'C' tracks how limiting AI instruction length does not fix its poor results over time; and 'D' illustrates that humans give more varied and balanced guidance than AI, even when word limits are imposed.

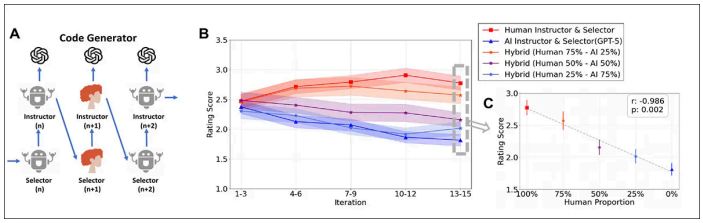
Human instructions tended to be short and to the point, offering clear edits that could be applied generally across targets. AI instructions, on the other hand, were dense with descriptive detail, and often bloated with specifics about shading, textures, lighting, or anatomical minutiae – descriptions that may make sense in isolation, but fail to provide useful next steps for the model (and which will be familiar to those aware of LLMs' issues around context length, i.e., being able to retain 'the big picture' as a project develops and grows).

To see if reduced verbosity would improve performance, GPT-5 was limited to 10, 20, or 30 words per instruction; but even these compressed instructions failed to show any improvement (see bottom-right of graph above).

Joint Endeavors

To test what happens when humans and AI *share* control, the researchers ran coding tasks with different mixes of human and AI input, ranging from *mostly human* to *mostly AI*.

Every hybrid mix outperformed full AI control, so that even a small amount of human guidance improved results:



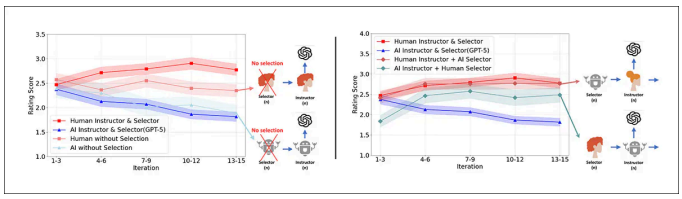
Hybrid coding setups with different human/AI mixes. (A) Shows how humans and AI took turns as instructors and selectors for each coding step; (B) shows that m human involvement led to higher quality results, while greater AI input lowered scores; and (C) depicts a steady drop in final output quality as the share of human participation decreases, confirming that more consistent human direction produced better outcomes.

As AI took over more of the process, performance dropped, with the best results seen when humans led most rounds, and the weakest when AI led most rounds. None of these mixed setups managed to keep improving with each new round, suggesting that human direction works best when it is steady and consistent, rather than occasional.

Role Reversal

The study also explored whether it matters *who does what* in these kinds of tasks, and tested for this. The revised exercise involved two tasks: one participant would dictate how to change the image, and another would choose a preferable version.

When both jobs were done by people, quality was maintained; but when a human gave the instructions and *no-one* picked between versions, quality got worse.:



Tests for role division in vibe coding: in (A), removing the selector role led to worse performance, even when a human provided instructions; in (B), replacing the human selector with an AI reduced quality slightly, but not as severely as skipping selection entirely.

When AI was in charge, skipping the choice step didn't matter, since its outputs stayed consistent either way; but when humans gave the instructions and AI picked between results, quality stayed close to the all-human setup.

The opposite didn't work: having AI give directions while humans chose outputs led to *weaker* results, suggesting that human creative guidance remains essential, while the job of choosing between options can be handed off to AI without much loss.

The paper concludes:

> *'[High-level] idea generation and instruction are the critical human contributions, whereas evaluation and selection can often be delegated to AI without loss in performance.*
>
> *'This suggests a practical design principle for hybrid systems: humans should set direction, while AI can support evaluation and execution.'*

Conclusion

It remains to be seen to what extent improved and/or increased context windows will affect the performance of LLMs in tasks of this kind. The day that 'LLM-amnesia' ceases to be a daily bugbear of human-AI collaboration may be a cause both for celebration and alarm, since the problem that AI is striving to solve, arguably, is *people*.

Nonetheless, the authors' work also makes clear that there are *innate* and critical disagreements between AI and humans regarding *quality*, which may yet be determined, by consumers, to be an irreplaceably human concept.

* *My conversion of the authors' inline citations to hyperlinks.*

First published Friday, February 13, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More